

# SECTION I

## Basics of Laboratory Medicine

- CHAPTER 1** Laboratory Medicine, 2
- CHAPTER 2** Statistical Methodologies in Laboratory Medicine: Analytical and Clinical Evaluation of Laboratory Tests, 10
- CHAPTER 3** Governance, Risk, and Quality Management in the Medical Laboratory,\* 60
- CHAPTER 4** Specimen Collection and Processing, 61
- CHAPTER 5** Preanalytical Variation and Pre-Examination Processes, 80
- CHAPTER 6** Quality Control of the Analytical Examination Process, 129
- CHAPTER 7** Standardization and Harmonization of Analytical Examination Results,\* 164
- CHAPTER 8** Biological Variation and Analytical Performance Specifications,\* 165
- CHAPTER 9** Establishment and Use of Reference Intervals, 166
- CHAPTER 10** Evidence-Based Laboratory Medicine,\* 194
- CHAPTER 11** Biobanking,\* 195
- CHAPTER 12** Laboratory Support of Pharmaceutical, In Vitro Diagnostics, and Epidemiologic Studies,\* 196
- CHAPTER 13** Machine Learning and Big Data in Laboratory Medicine,\* 197
- CHAPTER 14** Laboratory Stewardship and Test Utilization,\* 198
- CHAPTER 15** Principles of Basic Techniques and Laboratory Safety,\* 199

---

Exam questions, case studies, and additional resources are available on ExpertConsult.com.  
\*Full versions of these chapters are available electronically on [www.ExpertConsult.com](http://www.ExpertConsult.com).

# Laboratory Medicine

*Nader Rifai, Rossa W.K. Chiu, Ian Young, Carey-Ann D. Burnham, and Carl T. Wittwer<sup>a</sup>*

## ABSTRACT

### Background

Laboratory medicine is a complex field that measures biomarkers and microorganisms in bodily specimens or tissues to diagnose and manage diseases. It encompasses multiple disciplines including clinical chemistry, hematology and coagulation, clinical microbiology, clinical immunology, molecular diagnostics, and transfusion medicine. Laboratory medicine is driven by technology that helps define the boundaries among its disciplines. Although laboratory medicine specialists are diverse in terms of their education, training, and career paths, their practice of the profession and their adherence to its guiding principles are similar. The goal is to generate relevant chemical, cellular, and molecular data that can be integrated with clinical and other information and interpreted to aid clinical decision making.

## INTRODUCTION

Laboratory medicine is a broad and heterogeneous field that deals with the measurement of chemical, biochemical, cellular, and genetic biomarkers; it encompasses multiple disciplines including clinical chemistry, hematology and coagulation, clinical microbiology (including serology and virology), clinical immunology, molecular diagnostics, and, in certain countries, transfusion medicine. Tissue pathology and cytology, although part of the broad definition of laboratory medicine that includes all testing of human tissue, are not included in this textbook. Although the various fields of laboratory medicine overlap in a continuous dynamic evolution (Fig. 1.1), specific disciplines elicit different images. For clinical chemistry, one thinks of pH measurements or large chemistry analyzers; for hematology or microbiology, microscopic examination is what first comes to mind; and molecular diagnostics conjures up the human genome project, companion diagnostics, and personalized and precision medicine. Whereas clinical chemistry and molecular diagnostics are heavily dependent on technological developments, where the former excels in random access testing and the latter has evolved massively parallel methods, the practice of transfusion medicine and hematology is decidedly clinical. Furthermore, certain disciplines like transfusion medicine are well

### Content

This chapter describes the evolution of laboratory medicine and examines the international practice of the profession, the disciplines it encompasses, academic and postgraduate training, certification, career opportunities, and the skills and roles of laboratory medicine specialists in both clinical laboratory and industry settings. This chapter also discusses the guiding principles of practicing the profession, which include maintaining confidentiality of medical information, using available resources appropriately, abiding by codes of conduct, avoiding conflict of interest, and following ethical publishing rules.

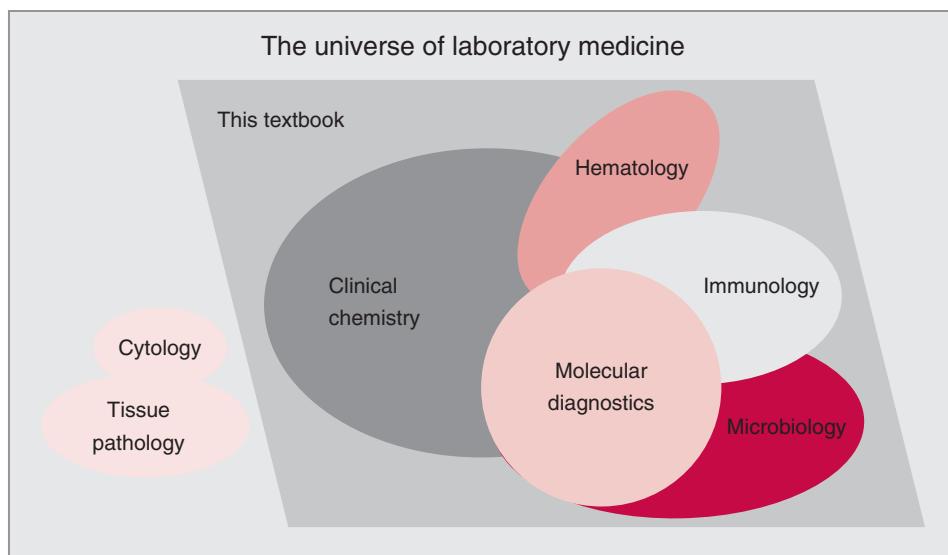
defined and consistently practiced internationally, while others such as clinical chemistry and clinical microbiology may vary in content depending on the country in which they are practiced. According to the definition of the International Federation of Clinical Chemistry and Laboratory Medicine (IFCC), “Clinical Chemistry is the largest subdiscipline of Laboratory Medicine which is a multidisciplinary medical and scientific specialty with several interacting subdisciplines, such as hematology, immunology, clinical biochemistry, and others. Through these activities clinical chemists influence the practice of medicine for the benefit of the public.”<sup>1</sup>

Hospital-based laboratory medicine departments and commercial clinical laboratories provide in vitro testing of a variety of biomarkers in various fluids or tissues of the human body to screen for a disease, confirm or exclude a diagnosis, help to select or monitor a treatment, or assess prognosis. The popular claim that 60 to 70% of clinical decisions are based on laboratory tests cannot be easily justified by objectively measured data.<sup>2,3</sup> Nevertheless, laboratory testing impacts healthcare delivery to virtually every patient.

## LOOKING BACK

The examination of body fluids for the diagnosis of disease is certainly not a modern concept. The Greeks noticed before 400 BC that ants are attracted to “sweet urine.” Laboratory testing, however, was not always appreciated by clinicians; the famous Dublin physician Robert James Graves (1796–1853) once remarked, “Few and scanty, indeed, are the rays of light

<sup>a</sup>The authors gratefully acknowledge the contributions by David E. Bruns, Edward R. Ashwood, Carl A. Burtis, and A. Rita Horvath on which portions of this chapter are based.



**FIGURE 1.1** The interacting disciplines of laboratory medicine. Laboratory medicine encompasses testing and associated activities for the assessment, diagnosis, treatment, management, and prevention of human disease. Although in certain countries tissue pathology and cytology are part of laboratory medicine, their focus on morphology and image analysis sets them apart from other areas of laboratory medicine and they are not considered in this textbook. The largest divisions of laboratory medicine considered within include clinical chemistry, clinical microbiology, clinical immunology, hematology, and molecular diagnostics. These disciplines overlap and evolve over time. *The sizes of the circles are not meant to reflect those of the disciplines.*

which chemistry has flung on the vital mysteries,” and the pioneer Max Josef von Pettenkofer (1818–1901) stated that clinicians use their chemistry laboratory services only when needed for “luxurious embellishment for a clinical lecture.”<sup>4</sup> Such views have changed throughout the years, and laboratory testing has proven to be a useful tool to clinicians who have grown to depend and rely on the clinical laboratory in the routine management of their patients.

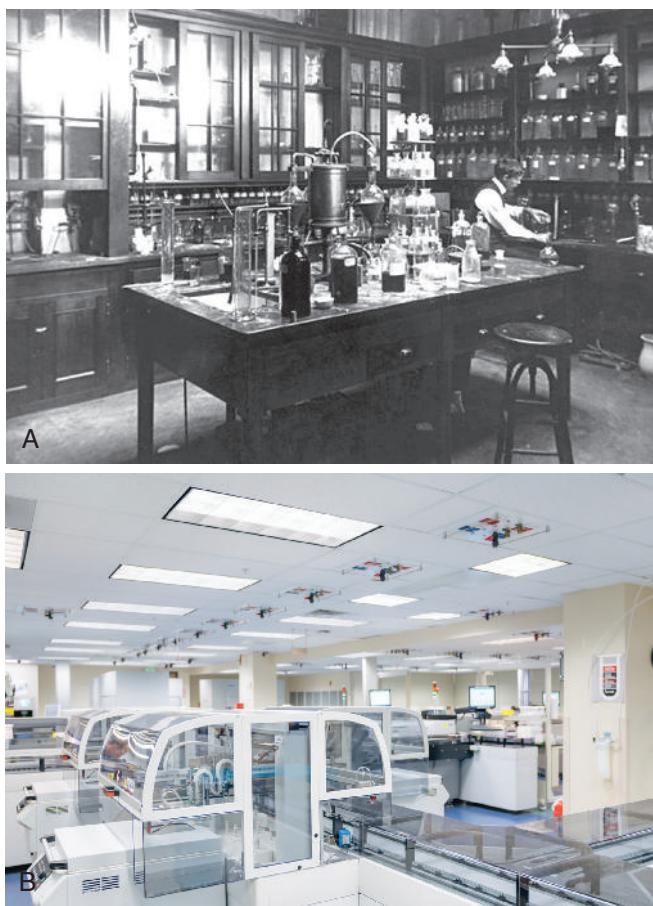
Although it may be difficult to pinpoint the exact date at which the concept of the clinical laboratory was born, a relevant article titled “Hospital Construction” by Francis H. Brown that was published in the *Boston Medical and Surgical Journal*, the precursor of the *New England Journal of Medicine*, in 1861. Dr. Brown stated: “[Every hospital should have] a small room at the end of the ward to serve as a general laboratory ... necessary small cooking might be accomplished here; dishes and other articles washed etc.; and it would serve as a general store-room for brooms, pails, and other articles.” Although Baron Justus von Liebig (1803–1873) once boasted that his clinical laboratory performed more than 400 tests per annum, the average mid- to large-sized laboratory today performs several million tests yearly; the images presented in Fig. 1.2 depict this striking contrast between the legendary Otto Folin in his biochemistry laboratory at McLean Hospital in Boston in 1905 and the University of Utah Clinical Laboratory/ARUP Laboratories more than a century later.

One of the first laboratories attached to a hospital was established in 1886 in Munich, Germany, by Hugo Wilhelm von Ziemssen.<sup>5</sup> In the United States, the first clinical laboratory recorded was The William Pepper Laboratory of Clinical Medicine, established in 1895 at the University of Pennsylvania in Philadelphia.<sup>6</sup> While there may be some uncertainty about the first hospital laboratory, the concept had become sufficiently well established by the late 1880s to enter popular

culture. Arthur Conan Doyle, writing in 1887, set the first meeting of Sherlock Holmes and Dr. Watson in 1881 in the chemical laboratory in St. Bartholomew’s Hospital, London, where Holmes had just discovered a reagent that “is precipitated by haemoglobin, and by nothing else” (*A Study in Scarlet*). Hopefully, the excitement experienced by Holmes at this discovery is still felt by laboratory specialists today.

Basic research usually precedes clinical application. Hematology began with the microscopic observation of red blood cells by Anthony van Leeuwenhoek (1632–1723). The father of microbiology is considered to be Louis Pasteur (1822–1895), who confirmed the germ theory of disease by experimentation. Immunology arose as a combination of the “cellularists,” observing phagocytosis, and the “humoralists,” who observed that immunity could be transferred as a soluble substance (antibodies and/or complement) in the late nineteenth century.

Molecular diagnostics has more recent origins than the other disciplines of laboratory medicine. “Molecular Diagnosis” was first mentioned in 1968 as the title of a *New England Journal of Medicine* editorial, commenting on a new inborn error of metabolism that overproduced oxalic acid, resulting in kidney stones.<sup>7</sup> “Molecular” referred to an enzymatic pathway and the substrates, not nucleic acid variants. Twenty years later, additional articles describing “molecular diagnostics” began to appear. In 1986, molecular diagnostics was defined as, “...the detection and quantification of specific genes by nucleic acid hybridization procedures,” exemplified by speciation of plant nematodes.<sup>8</sup> In 1987, molecular diagnostics was used to describe mapping of antigenic substances by affinity chromatography using immobilized antibodies.<sup>9</sup> In 1988, the term was used to describe methods for detecting gene amplification and rearrangement using Southern blotting.<sup>10</sup> With the advent of polymerase chain reaction (PCR),



**FIGURE 1.2** Early and modern clinical laboratories. The legendary Otto Folin in his biochemistry laboratory at McLean Hospital in Boston in 1905 and the University of Utah Clinical Laboratory/ARUP Laboratories, Salt Lake City, UT, more than a century later. (Image 1 from [http://en.wikipedia.org/wiki/File:1905\\_Otto\\_Folin\\_in\\_biochemistry\\_lab\\_at\\_McLean\\_Hospital\\_by\\_AHFolsom\\_Harvard.png](http://en.wikipedia.org/wiki/File:1905_Otto_Folin_in_biochemistry_lab_at_McLean_Hospital_by_AHFolsom_Harvard.png); Image 2 courtesy ARUP Laboratories.)

the term “molecular diagnostics” became more common, its use doubling in the medical literature every 6 to 7 years.<sup>11</sup> By 1997, commercial real-time PCR instruments solidified “molecular diagnostics” as a branch of laboratory medicine.

## TRAINING IN LABORATORY MEDICINE

Clinical laboratory professionals are individuals with a medical or a doctoral degree (pharmacy, chemistry, biology, biochemistry, microbiology) who are focused on clinical service. In North America, Australia, and Europe, a minimum of 9 years of academic education (a medical or a doctoral degree) and postgraduate professional training (residency and postdoctoral) is required before an individual becomes an independently practicing specialist (Fig. 1.3).<sup>12</sup> The requirements and training in laboratory medicine to become a specialist differ around the world. For example, in the United States, either those with a medical or a doctoral degree can direct a clinical laboratory after obtaining the appropriate board certification. Those with a medical degree usually do a residency in clinical or clinical/anatomical pathology to direct a general clinical laboratory. However, if they chose to direct a discipline-specific laboratory such as clinical chemistry, microbiology,

or transfusion medicine, they may need to complete a fellowship in that specialty. Those with a doctoral degree tend to direct a discipline-specific laboratory and must complete postdoctoral training in that specialty. In the European Union, 40% of laboratory medicine specialists are from medical, 30% are from scientific, and 30% are from pharmacy backgrounds. In some countries such as Austria, Lithuania, Estonia, Malta, and Sweden, only physicians can practice the profession and direct a clinical laboratory. In most other European countries, scientists, pharmacists, and physicians can be laboratory medicine specialists, yet those with a pharmacy degree may not serve as clinical laboratory directors in some of these countries, such as Italy. A pharmacy degree is a “professional” degree (but not equivalent to a PhD) in France.

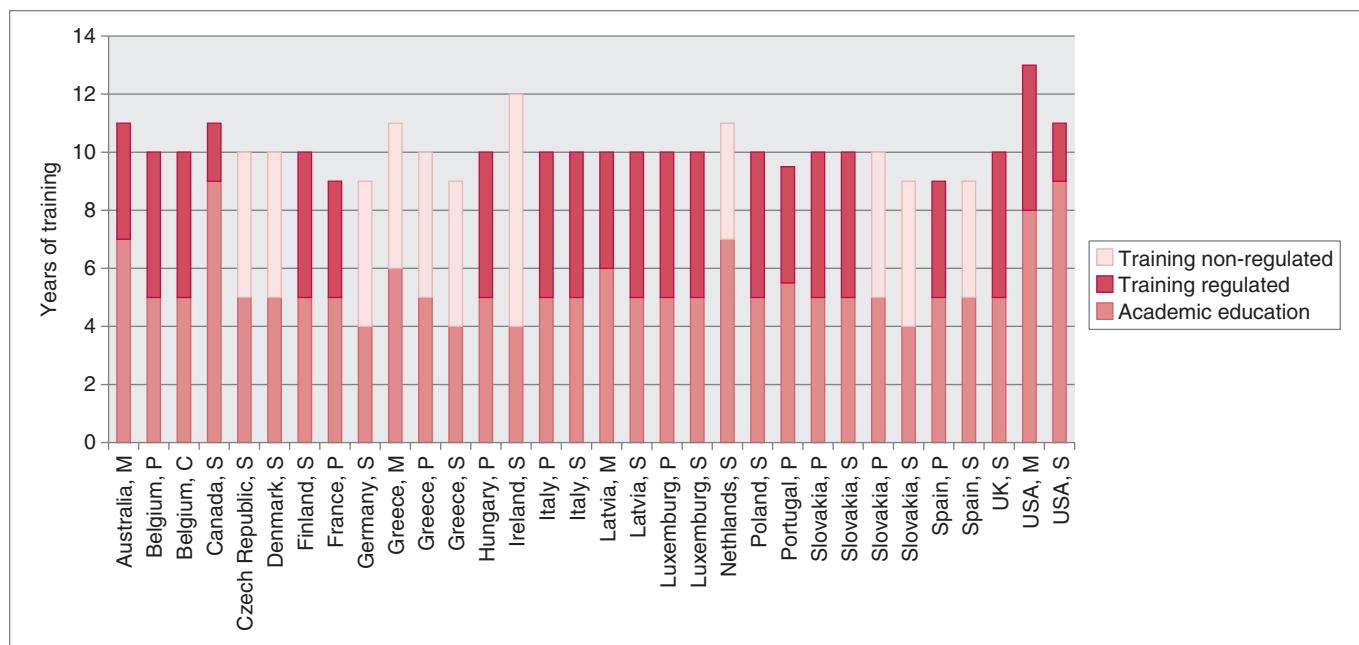
The curriculum used during the training of a clinical laboratory specialist in the European Union varies depending on the country. In the majority of European countries, trainees get exposed to clinical chemistry (45% of the curriculum), hematology (30%), microbiology (15%), and genetics (10%).<sup>1,12</sup> Molecular diagnostics (nucleic acid testing) is considered a technique and is included in all fields. In contrast, in the United Kingdom and Ireland, chemical pathology training is restricted to the traditional subdiscipline of clinical chemistry. This diversity of subspecialties is reflected in the heterogeneity of postgraduate training across countries.<sup>13</sup>

Postgraduate professional training and certification examinations at the end of the training are not mandated in all countries (see Fig. 1.3). The EFLM Register of Specialists in Laboratory Medicine (EuSpLM) (<https://www.eflm.eu/site/page/a/1305>) is attempting to standardize the minimum requirements for education and training for laboratory medicine specialists to facilitate the comparability of their professional training within the European Union.<sup>1,12,13</sup> These issues add to the complexity of defining the qualifications of clinical laboratory directors.

## EXPANDING BOUNDARIES DEFINED BY TECHNOLOGY

The diversity of background, training, and subspecialization has led to heterogeneity in what the profession is called throughout the world. Name designations include clinical chemistry, clinical biochemistry, chemical pathology, hematology, clinical microbiology, transfusion medicine, clinical pathology, laboratory diagnostics, clinical or medical biology, clinical laboratory, laboratory medicine, clinical analysis, and so on. The EC4 Register (now the EuSpLM) adopted the name “specialist in laboratory medicine” to represent clinical laboratorians in Europe.

Everyone, including lay people, knows what a cardiologist is and does; the same is true for an infectious diseases specialist and a surgeon. Within laboratory medicine, the function of certain specialists, such as clinical microbiologists, hematologists, or blood bankers, is also clear. It is more difficult, however, to characterize a clinical chemist. Perhaps, unlike other specialties in laboratory medicine, clinical chemistry is very much influenced and shaped by technology. No discipline in laboratory medicine uses more technologies than clinical chemistry. Technologies that evolved over time not only changed practice but remodeled the boundaries of the traditional clinical chemistry laboratory. For example, with



**FIGURE 1.3** The number of years of education and training required to practice as clinical laboratory specialist in different countries varies from 9 to 13 years. Different training routes include medical (M), pharmacy (P), chemistry (C), and scientific (S). Both academic education (light red bars) and postgraduate training are required. Postgraduate training may be regulated (dark red bars) or nonregulated (pink bars) in different countries and even within the same country. (Modified from EU Directive 2013/55/EU. The recognition of professional qualifications. Proposing a common training framework for specialists in laboratory medicine across the European Union 2013. [http://www.ukipg.org.uk/meetings/international\\_and\\_european\\_forum/ctf\\_e4\\_bid](http://www.ukipg.org.uk/meetings/international_and_european_forum/ctf_e4_bid).)

the emergence of immunochemical techniques in the 1970s, the US Food and Drug Administration approved many tests for the measurement of proteins, small molecule hormones, and drugs, a development that profoundly changed clinical chemistry and its armamentarium of testing at the time. Integrated automated platforms later enabled the measurement of hormones and therapeutic drugs by immunoassays simultaneously with electrolytes, glucose, and other general chemistry tests, thus subsuming the “endocrine lab” and the “drug lab.”

Serologic tests for hepatitis and HIV and assays for the evaluation of autoimmune diseases also moved from their traditional home in microbiology and immunology to chemistry analyzers. Immunoglobulin analysis followed a similar path. In certain countries, coagulation is considered part of clinical chemistry because the measurement of coagulation proteins uses similar instruments to those used in the clinical chemistry laboratory. As a result, the typical clinical chemistry laboratory includes testing for general chemistries, specific proteins and immunoglobulins, therapeutic and abused drugs, blood gases, hormones, biogenic amines, porphyrins, vitamins, and trace elements. Testing for inborn errors of metabolism (such as the measurements of amino acids and organic acids), measurements of coagulation factors, general hematologic testing, and serologic assays can belong either to the clinical chemistry laboratory or to another subspecialty, depending on the institution and country. If amino acids and organic acids are measured in the clinical chemistry laboratory, that does not preclude a biochemical geneticist from providing the clinical interpretation. Similar arguments can be made for coagulation, hematology, and serology testing.

Clinical laboratory professionals have embraced technology over the years and used it effectively to derive answers to clinical questions. In modern clinical laboratories, technologies include spectrophotometry, atomic absorption, cytometry, flame emission photometry, nephelometry, electrochemical, and optical sensor technologies, electrophoresis, and chromatography. The influence of automation, information technology, and miniaturization is evident in today’s clinical laboratory. Mass spectrometry, once thought of as a research tool, is playing an ever-growing role in clinical chemistry for the measurement of both small molecules and peptides and more recently proteins. In fact, matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) mass spectrometry is now routinely used in the identification of microorganisms (including bacteria, mycobacteria and fungi), so it is likely that the evolution in this technology will also bring the clinical chemistry and microbiology laboratories closer. In addition, clinical microbiology laboratories are becoming increasingly automated, with total laboratory automation systems, including fluidic handling and high-resolution digital imaging systems, being adopted with increasing frequency.<sup>14</sup> Molecular diagnostics has forever changed virology and microbiology, introducing faster and more sensitive methods based on nucleic acid detection rather than microbial replication. Nanotechnology, microfluidics, electrical impedance, reflectance spectroscopy, and time-resolved fluorescence are only a few of the technologies used in point-of-care testing for proteins, drugs, DNA, and analysis of metabolites in small samples of whole blood. Point-of-care testing is a disruptive innovation that decentralizes laboratory testing and presents the clinical laboratory specialist with many

challenges and opportunities. Molecular diagnostics in particular impacts diverse specialties, including infectious disease, genetics, and oncology, providing new tools for study at a molecular detail never before considered. In summary, the boundaries of laboratory medicine expand with technology, making the profession vibrant, interesting, and ever evolving.

The scope of the profession is constantly changing for the very same reasons. Scientific and technological developments, medical needs, patient demands, and economic pressures bring various disciplines of medicine closer together, and further integration of diagnostic and therapeutic disciplines is envisaged in the pursuit of more integrated and effective healthcare delivery. For example, companion diagnostics, which help predict therapeutic responses and individualize patient treatment options, bring together pharmacy and medical laboratories. Point-of-care testing and use of biomarker measurements in real time with medical interventions break the walls of laboratories and bring the profession closer to clinicians and patients. Integrated diagnostics (a term coined by the medical device industry), whereby *in vitro* laboratory technology is combined with *in vivo* imaging technology, intends to provide fully coordinated, interpreted, action-oriented results for managing patient conditions, and it places laboratory testing into an integrated patient care pathway (see an example at [http://www.healthcare.siemens.com.au/clinical-specialties/reproductive-endocrinology/integrated-diagnostics](http://www.healthcare.siemens.com.au/clinical-specialities/reproductive-endocrinology/integrated-diagnostics)). New disruptive technologies (e.g., “lab on a chip,” nanotechnology, home monitoring) and movement toward patient empowerment and direct-to-consumer testing bring laboratory testing closer to patients. All of these developments present special challenges to the future generations of clinical laboratory specialists both in terms of how they should be trained and how they will have to practice.

Technology alone is not the answer to more effective clinical practice. There must be meaningful, clinically actionable results as a consequence of the data obtained. The generation of more data does not necessarily lead to better patient management. Some technology platforms are useful discovery tools, but seldom provide cost-effective diagnostic or prognostic information that changes patient care. In the 1960s and 1970s, with the advent of automated clinical analyzers, pathologists reported (and charged for) chemistry panels of 10 to 20 results. Many were later sued for excessive production of data that increased their income without commensurate value to patient care. More recently, dense data from expression arrays, genome-wide association studies, epigenomics, and microRNA analyses excel in discovery research, but translation to clinical practice has been slower than anticipated. The promise of greater clinical significance with larger data sets seems intuitive, but history suggests caution.

Clinical laboratorians in this world of “big data” translate high-quality measurement *data* into clinically relevant *information*. This information—when integrated with clinical history and presentation, clinical signs, and an understanding of pathophysiology—becomes *knowledge*. Knowledge, in the context of the experience and judgment of the clinician, is converted to *wisdom* that translates to clinical action for improved patient outcomes. For example, a 2-week-old boy with a suspected inborn error of metabolism had a suppressed thyroid stimulating hormone and increased free T<sub>4</sub> concentrations. Acting on the basis of the data alone would have suggested treatment with methimazole for thyrotoxicosis.

However, the patient was receiving biotin as part of his treatment for a metabolic disorder, and the biotin interfered with the immunoassays used in the thyroid function tests. Repeat measurement of these parameters with non-biotin-based immunoassays revealed a normal thyroid profile. Another example: A patient presented with Cushingoid appearance and a markedly decreased serum cortisol concentration. A further examination of his clinical history revealed that he was using topical corticosteroids for a skin condition, a treatment that caused adrenal suppression and thus a low cortisol concentration. Yet another example: A 55-year-old woman complained to her primary care physician about long-standing bony aches and pains. All results to exclude musculoskeletal problems came back normal except for a low alkaline phosphatase (ALP) enzyme activity. After excluding potential preanalytical errors [e.g., contamination of sample by K-EDTA (potassium ethylenediaminetetraacetic acid) anticoagulant], the laboratory proposed the diagnosis of hypophosphatasia, and testing for mutations of the tissue nonspecific ALP gene confirmed the diagnosis both in the patient and in her daughter. The world of laboratory medicine is full of such examples that demonstrate the value of acting on information beyond the generated numbers. Knowledge is what we must provide to clinicians to support informed clinical decision making and for achieving improved patient outcomes.

## HOW IS LABORATORY MEDICINE PRACTICED?

Both the training of laboratory medicine professionals and their career paths are heterogeneous. Although the majority of our colleagues choose a career in a clinical laboratory environment, many work in the *in vitro* diagnostics (IVD) and pharmaceutical industries. Clinical laboratorians, by virtue of their training, are translational researchers who are equipped for and capable of developing, evaluating, and validating biochemical, cellular, and genetic assays for clinical use; they develop skills that are essential for new biomarker assays, reagent kits, and companion diagnostics. Laboratory medicine professionals also provide interfaces between researchers, clinicians, the clinical laboratory, and the IVD industry and help to translate biomarker research into clinically meaningful decisions and actions.

The functions of a clinical laboratorian include:

- Develop and validate *de novo* laboratory tests to meet clinical needs.
- Evaluate and characterize the analytical and clinical performance of laboratory tests.
- Present laboratory results to clinicians in an effective manner.
- Provide education and advice on the selection and interpretation of laboratory tests as part of the clinical team.
- Determine the cost-effectiveness and intrinsic value of laboratory tests.
- Participate in the development of clinical testing algorithms and clinical practice guidelines.
- Assure compliance with regulatory requirements.
- Participate in quality assurance and improvement of the laboratory service.
- Teach and train future generations of laboratory specialists.
- Participate in basic or clinical research.

Laboratory medicine specialists practicing in the IVD or the pharmaceutical industry may not need to routinely interact with clinicians or interpret laboratory results, but they

understand and appreciate the clinical utility and relevance of the assays and companion diagnostics they are developing and thus contribute more effectively to the development of diagnostics that improve health. The daily practice of the profession has changed over time. In the 1960s and 1970s, clinical chemists, for example, developed laboratory tests. However, as the profession matured and the instrumentation changed from open systems to “black boxes” that relied on manufacturers for assays, the traditional analytical focus of the profession has significantly diminished. At present, de novo assay development is still active only in certain areas such as chromatography, mass spectrometry, and molecular diagnostics.

Laboratory medicine specialists are now more active in the preanalytical and postanalytical phases of testing and in establishing processes such as how best to select the right test for the right patient and to communicate test results to clinicians in a medically meaningful way, how to build laboratory processes that reduce error, and how to continuously improve the quality of laboratory practice. In today’s healthcare environment, there is increasing emphasis on clinical impact and cost-effectiveness. Laboratories are expected to demonstrate evidence of improved measurable clinical outcomes and the usefulness and added value of tests to clinical decision making. Proving the fact that laboratory testing contributes to improved patient outcomes is challenging because the relationship between testing and clinical outcomes is mostly indirect. Nevertheless, laboratory medicine specialists should move away from being just providers of high-quality data. Transforming laboratory data to information and knowledge requires more skills in information and information management technology, evidence-based medicine, epidemiology, data mining, and translational research. It also requires a shift of thinking from essentialism to consequentialism and from technology-driven to customer-focused and patient-centered laboratory medicine.<sup>15,16</sup>

To summarize, today’s clinical laboratorians are professionals who are trained in pathophysiology and technology. The execution of their daily duties, which are more clinically or technology oriented, is influenced by their training (such as MD vs. PhD), interests, institutional needs, and the country where they practice. Clearly the practice of our profession has evolved over the past half a century, and there are even more challenges on the horizon that will expand and change its scope and role and enhance its diversity.

## GUIDING PRINCIPLES OF PRACTICING THE PROFESSION

As in all branches of medicine, practitioners in the clinical laboratory are faced with ethical issues, often on a daily basis; examples are listed in **Box 1.1**.

### BOX 1.1 Ethical Issues in Laboratory Medicine

- Confidentiality of patient medical information
- Allocation of resources
- Codes of conduct
- Publishing issues
- Conflicts of interest

### Confidentiality of Patient Information

Safeguarding the confidentiality of a patient’s personal and medical information is one of the fundamental ethical principles of the practice of medicine. Upholding of these principles prescribes how some laboratory activities are practiced. The laboratory holds vast amounts of data covering a patient’s identifiers and demographics, as well as health and disease status. The patient’s morbid state and future risks for illnesses and death are conferred by such information. While laboratory information systems are built to facilitate timely access to the data, the data must be stored in a secure format with measures in place to prevent unwarranted access.

On the other hand, development of new tests requires the use of patient samples and access to patient medical information by the laboratory.<sup>17</sup> Ethical judgments are required regarding the type of informed consent that is needed from patients for use of their samples and clinical information. Clinical laboratory physicians and scientists often serve on institutional review boards that examine proposed research on human subjects. In these discussions, ethical concepts such as clinical equipoise (the genuine uncertainty in the expert medical community over whether a particular treatment or test will be beneficial) and preservation of confidentiality of medical information are central to these decisions.

Broad coverage genetic testing is becoming more of a routine affair. Prominent in the news in the first and second decades of this millennium has been the issue of confidentiality of genetic information. Legislation was considered necessary to prevent denial of health insurance or employment to people found by DNA testing to be at risk of disease. The power of DNA information lies in its heritability. Predictions can be made on the phenotypes and traits of a person’s parents, relatives, and offspring based on an individual’s DNA profile. In the event of having identified a clinically significant incidental finding, the right to personal confidentiality against the potential duty to disclose the information to at-risk family members is a current subject of debate among stakeholders. Clinical laboratory professionals are actively participating in the development of such disclosure and clinical management guidelines that will need to adapt to the changing standards of information disclosure or nondisclosure.

### Allocation of Resources

Because resources are finite, clinical laboratory professionals must make ethically responsible decisions about allocation of resources. There is often a trade-off between cost and quality and/or speed (turnaround time). What is best for patients generally? How can the most good be done with the available resources?

### Codes of Conduct

Most professional organizations publish a code of conduct that requires adherence by their members. For example, the American Association for Clinical Chemistry (AACC) has published ethical guidelines that require AACC members to endorse principles of ethical conduct in their professional activities, including (1) selection and performance of clinical procedures, (2) research and development, (3) teaching, (4) management, (5) administration, and (6) other forms of professional service. A similar code of conduct has been developed and approved by the EC4 Register Commission and the European Federation of Clinical Chemistry and Laboratory Medicine.<sup>18</sup>

## Publishing Issues

Publication of documents having high scientific integrity depends on editors, authors, and reviewers all working in concert in an environment governed by high ethical standards.<sup>19</sup>

Editors are responsible for the overall process, including identifying reviewers, evaluating the reviews and the authors' response to them, and making the final decision of whether to accept or reject a manuscript. Editors are also responsible for establishing policies and procedures to assure consistency in the editorial process. Finally, the editor-in-chief is responsible for developing a conflict of interest policy and monitoring it among his or her editors. Publishers, being commercial or scientific societies, should monitor any conflicts of interest of the editor-in-chief.

Authors are responsible for honest and complete reporting of original data produced in ethically conducted research studies. Practices such as fraud, plagiarism (verbatim, mosaic), and falsification or fabrication of data (including image manipulation) are unacceptable. The International Committee of Medical Journal Editors (ICMJE)<sup>20</sup> and the Committee on Publication Ethics (COPE)<sup>21</sup> have published policies that address such behavior. Other practices to be avoided include duplicate publication, redundant publication, and inappropriate authorship credit. In addition, ethical policies require that factors potentially influencing the interpretation of study findings must be revealed, such as (1) the role of the commercial sponsor in the design and conduct of the study, (2) interpretation of results, and (3) preparation of the manuscript. Additional undesirable and harmful practices are publication bias and selective reporting in which only studies with positive findings are reported and authors use "data dredging" and meaningless subanalyses to find positive association rather than reporting the original hypothesis that was negative.<sup>19</sup> These practices inflate the actual value of observations or utility of markers and diminish the quality of meta-analyses. As a result, a comprehensive registry of diagnostic and prognostic studies, similar to the registry of clinical trials, has been advocated.<sup>19,22,23</sup>

To avoid publication of biased study results, reporting guidelines have been published for the main study types on the website of the EQUATOR Network (<http://www.equator-network.org>). For the laboratory profession, the STARD and TRIPOD statements for diagnostic and prognostic studies are probably the most important,<sup>24,25</sup> but reporting guidelines for randomized controlled trials (CONSORT), observational studies (STROBE), systematic reviews (PRISMA), quality improvement studies (SQUIRE), and economic evaluations (CHEERS) are also relevant for the work of laboratory scientists active in research and publication.

Reviewers must provide a timely, fair, and impartial assessment of manuscripts. They must maintain confidentiality and never contact the authors until after the publication of the report. Finally, reviewers must excuse themselves from the review process if they perceive a conflict of interest.

Most journals now require authors to complete conflict of interest forms and delineate each author's contribution. Some journals, including *Clinical Chemistry*, publish this information along with the article for enhanced transparency.

## Conflicts of Interest

The interrelationships between practitioners in the medical field and commercial suppliers of drugs, devices, and equipment

can be positive or negative.<sup>26</sup> Concerns led the National Institutes of Health in 1995 to require official institutional review of financial disclosure by researchers and management in situations when disclosure indicates potential or actual conflicts of interest. In 2009, the Institute of Medicine issued a report<sup>27</sup> that questioned inappropriate relationships between pharmaceutical device companies and physicians and other healthcare professionals.<sup>26</sup> Similarly, the relationship between clinical laboratory professionals and manufacturers and providers of diagnostic equipment and supplies has been scrutinized.

As a consequence of these concerns and as a result of the enactment of various laws designed to prevent fraud, abuse, and waste in Medicare, Medicaid, and other federal programs, professional organizations that represent manufacturers of IVD and other device and healthcare companies have published codes of ethics. For example, the Advanced Medical Technology Association (AdvaMed) has published a revised code of ethics that became effective on January 1, 2020.<sup>28</sup> Topics discussed in this revised code include gifts and entertainment, consulting arrangements and royalties, reimbursement for testing, and education. Similarly, MedTech Europe has recently published a code of ethics.<sup>29</sup> In this document, topics include member-sponsored product training and education, support for third-party educational conferences, sales and promotional meetings, arrangements and consultants, gifts, provision of reimbursements and other economic information, and donations for charitable and philanthropic purposes. Both the AdvaMed and the MedTech Europe documents address demands from regulators while nurturing the unique role that clinical chemists and other healthcare professionals play in developing and refining new technology.<sup>26</sup>

## WHAT IS IN THIS TEXTBOOK?

In this textbook, we have assembled what is essential to effectively practice laboratory medicine. We begin with introductory chapters that describe the basics of laboratory medicine, including statistics, sample handling, preanalytical processes, reference intervals, quality management, quality control, standardization and harmonization, evidence-based laboratory medicine, biobanking, and biomarker and laboratory support for the pharmaceutical and IVD industries, machine learning, test utilization, and laboratory safety. This is followed by a section on analytical techniques and applications, including mass spectrometry and the specialized topics of microfabrication and microfluidics, cytometry, and point-of-care testing. Next, all the major analytes in clinical chemistry, including enzymes, tumor markers, therapeutic drugs, and many others are discussed. Pathophysiology, covering disease states and malfunction of different organ systems that correlate with abnormal laboratory findings follows. A section on genetic metabolic testing discussing newborn screening and inborn error of metabolism is next. This is followed by a section dedicated to molecular diagnostics, perhaps the fastest growing field in laboratory medicine. Then, there is a section discussing automated hematology and white and red blood cell morphologies, as well as hemostasis and coagulation. Following this is coverage of clinical microbiology including antimicrobial stewardship and infection prevention, infectious disease, antimicrobial susceptibility, bacteriology, virology, mycobacteriology, mycology, and parasitology. A

transfusion medicine section then presents blood groups, blood components, indications for blood transfusion, and transfusion reactions. Finally, our last section focuses on clinical immunology including systemic autoimmune disease, transplantations, immunogenetics, allergy testing, immunogenicity of biologics, and primary and secondary immunodeficiencies. An appendix tabulates reference intervals for the clinical laboratory. The online version includes all of the above topics, whereas the print version is more selective to keep the tome manageable.

In addition to the above-mentioned chapters, the online version contains a wealth of other information including biochemical calculations, animation films to illustrate complex mechanisms, clinical cases, numerous atlases, podcasts, important documents, lecture series, and adaptive learning courses.

This is an exciting time to be a laboratory medicine professional. Our aim in this book is to provide current scientific and practical knowledge to support laboratory professionals as a knowledge resource and an interface between science and technology on the one hand and the clinician and the patient on the other.

## POINTS TO REMEMBER

- Laboratory medicine is a heterogeneous field with multiple disciplines including clinical chemistry, hematology and coagulation, clinical microbiology, molecular diagnostics, clinical immunology, and transfusion medicine.
- Laboratory medicine is a profession that has been shaped and defined by technology.
- Training of laboratory medicine specialists is heterogeneous and includes physicians and doctoral scientists in chemistry, pharmacy, biology, biochemistry, and microbiology.
- The role of clinical laboratory specialists evolved over time from analytically and technology focused to customer and patient centered.
- Clinical laboratory specialists are translational researchers who convert laboratory data to clinical knowledge.
- Career paths of clinical laboratory specialists are heterogeneous and include work in clinical laboratories and IVD and pharmaceutical industries.
- Clinical laboratory specialists must adhere to guiding principles of practicing the profession, which include maintaining confidentiality of medical information, using resources appropriately, abiding by codes of conduct, following ethical publishing rules, and managing and disclosing conflict of interest.

## SELECTED REFERENCES

1. McMurray J, Zerah S, Hallworth M, et al. The European Register of Specialists in Clinical Chemistry and Laboratory Medicine: guide to the Register, version 3-2010. *Clin Chem Lab Med* 2010;48:999–1008.
3. Hallworth MJ. The “70% claim”: what is the evidence base? *Ann Clin Biochem* 2011;48:487–8.
6. Young DS, Berwick MC, Jarett L. Evolution of the William Pepper Laboratory. *Clin Chem* 1997;43:174–9.
12. EU Directive 2013/55/EU. The recognition of professional qualifications. Proposing a common training framework for specialists in laboratory medicine across the European Union 2013. <[http://www.ukipg.org.uk/meetings/international\\_and\\_european\\_forum/ctf\\_e4\\_bid](http://www.ukipg.org.uk/meetings/international_and_european_forum/ctf_e4_bid)>; 2013.
13. Jassam N, Lake J, Dabrowska M, Queralto J, Rizos D, Lichtinghagen R, et al. The European Federation of Clinical Chemistry and Laboratory Medicine syllabus for post graduate education and training for specialists in laboratory medicine: version 5-2018. *Clin Chem Lab Med* 2018;56:1846–63.
14. Bailey AL, Ledeboer N, Burnham CAD. Clinical microbiology is growing up: the total laboratory automation revolution. *Clin Chem* 2019;65:634–43.
15. Hallworth MJ, Epner PL, Ebert C, et al. Current evidence and future perspectives on the effective practice of patient-centered laboratory medicine. *Clin Chem* 2015;61(4):589–99.
17. Council of Europe. Additional protocol to the convention for the protection of human rights and dignity of the human being with regard to the application of biology and medicine on biomedical research. *Law Hum Genome Rev* 2004;21:201–14.
18. McMurray J, Zerah S, Hallworth M, et al. The European Register of Specialists in Clinical Chemistry and Laboratory Medicine: code of conduct, version 2—2008. *Clin Chem* 2009;47:372–5.
19. Annesley TM, Boyd JC, Rifai N, et al. Publication ethics: clinical chemistry editorial standards. *Clin Chem* 2009;55:1–4.
20. International Committee of Medical Journal Editors. Uniform requirements for manuscripts submitted to biomedical journals: writing and editing for biomedical publication. <<http://www.icmje.org/recommendations/browse/manuscript-preparation/>>.
21. Graf CWE, Bowman A, Fiack S, et al. Best practice guidelines on publication ethics: a publisher’s perspective. *Int J Clin Pract Suppl* 2007;61:1–26.
22. Rifai N, Bossuyt PM, Ioannidis JP, et al. Registering diagnostic and prognostic trials of tests: is it the right thing to do? *Clin Chem* 2014;60:1146–52.
24. Bossuyt PM, Reitsma JB, Bruns DE, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Clin Chem* 2003;49:7–18.
27. Institute of Medicine. Conflict of interest in medical research, education and practice. <<http://www.nationalacademies.org/hmd/Reports/2009/Conflict-of-Interest-in-Medical-Research-Education-and-Practice.aspx>>.
28. Advanced Medical Technology Association. Code of Ethics on interactions with health care professionals. <https://www.advamed.org/sites/default/files/resource/advamed-code-of-ethics-2020.pdf>.
29. European Diagnostics Manufacturers Association. Part A: interaction with health care professionals. <https://www.medtecheurope.org/resource-library/medtech-europe-code-of-ethical-business-practice/.2019>.

## REFERENCES

1. McMurray J, Zerah S, Hallworth M, et al. The European Register of Specialists in Clinical Chemistry and Laboratory Medicine: guide to the Register, version 3-2010. *Clin Chem Lab Med* 2010;48:999-1008.
2. Forsman RW. Why is the laboratory an afterthought for managed care organizations? *Clin Chem* 1996;42:813-6.
3. Hallworth MJ. The “70% claim”: what is the evidence base? *Ann Clin Biochem* 2011;48:487-8.
4. Rifai N, Annesley T, Boyd J. International year of chemistry 2011: Clinical Chemistry celebrates. *Clin Chem* 2010;56:1783-5.
5. Bruns DE, Ashwood ER, Burtis CA. Clinical chemistry, molecular diagnostics, and laboratory medicine. In: Bruns DE, Ashwood ER, Burtis CA, editors. *Tietz textbook of clinical chemistry and molecular diagnostics*. 5th ed. St Louis: Elsevier; 2012. p. 3-7.
6. Young DS, Berwick MC, Jarett L. Evolution of the William Pepper Laboratory. *Clin Chem* 1997;43:174-9.
7. Molecular diagnosis. *N Engl J Med* 1968;278:276-7.
8. Powers TO, Platzer EG, Hyman BC. Species-specific restriction site polymorphism in root-knot nematode mitochondrial DNA. *J Nematol* 1986;18:288-93.
9. Caliceti P, Fassina G, Chaiken IM. Molecular diagnostics using analytical immuno high performance liquid affinity chromatography. *Appl Biochem Biotechnol* 1987;16:119-28.
10. Fourney RM, Dietrich KD, Paterson MC. Rapid DNA extraction and sensitive alkaline blotting protocol: application for detection of gene rearrangement and amplification for clinical molecular diagnosis. *Dis Markers* 1989;7:15-26.
11. Chiu RW, Lo YM, Wittwer CT. Molecular diagnostics: a revolution in progress. *Clin Chem* 2015;61:1-3.
12. EU Directive 2013/55/EU. The recognition of professional qualifications. Proposing a common training framework for specialists in laboratory medicine across the European Union 2013. <[http://www.ukipg.org.uk/meetings/international\\_and\\_european\\_forum/ctf\\_e4\\_bid](http://www.ukipg.org.uk/meetings/international_and_european_forum/ctf_e4_bid)>; 2013.
13. Jassam N, Lake J, Dabrowska M, Queralto J, Rizos D, Lichtenhagen R, et al. The European Federation of Clinical Chemistry and Laboratory Medicine syllabus for post graduate education and training for specialists in laboratory medicine: version 5-2018. *Clin Chem Lab Med* 2018;56:1846-63.
14. Bailey AL, Leedeboer N, Burnham CAD. Clinical microbiology is growing up: the total laboratory automation revolution. *Clin Chem* 2019;65:634-43.
15. Hallworth MJ, Epner PL, Ebert C, et al. Current evidence and future perspectives on the effective practice of patient-centered laboratory medicine. *Clin Chem* 2015;61(4):589-99.
16. Hofmann BM. Too much technology. *BMJ* 2015;350:h705.
17. Council of Europe. Additional protocol to the convention for the protection of human rights and dignity of the human being with regard to the application of biology and medicine on biomedical research. *Law Hum Genome Rev* 2004;21:201-14.
18. McMurray J, Zerah S, Hallworth M, et al. The European Register of Specialists in Clinical Chemistry and Laboratory Medicine: code of conduct, version 2—2008. *Clin Chem* 2009;47:372-5.
19. Annesley TM, Boyd JC, Rifai N, et al. Publication ethics: clinical chemistry editorial standards. *Clin Chem* 2009;55:1-4.
20. International Committee of Medical Journal Editors. Uniform requirements for manuscripts submitted to biomedical journals: writing and editing for biomedical publication. <<http://www.icmje.org/recommendations/browse/manuscript-preparation/>>.
21. Graf CWE, Bowman A, Fiack S, et al. Best practice guidelines on publication ethics: a publisher’s perspective. *Int J Clin Pract Suppl* 2007;61:1-26.
22. Rifai N, Bossuyt PM, Ioannidis JP, et al. Registering diagnostic and prognostic trials of tests: is it the right thing to do? *Clin Chem* 2014;60:1146-52.
23. Altman DG. The time has come to register diagnostic and prognostic research. *Clin Chem* 2014;60:580-2.
24. Bossuyt PM, Reitsma JB, Bruns DE, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Clin Chem* 2003;49:7-18.
25. Moons KG, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015;162:W1-73.
26. Malone B. Ethics code changes for diagnostics manufacturers. *Clin Lab News* 2009;35.
27. Institute of Medicine. Conflict of interest in medical research, education and practice. <<http://www.nationalacademies.org/hmd/Reports/2009/Conflict-of-Interest-in-Medical-Research-Education-and-Practice.aspx>>.
28. Advanced Medical Technology Association. Code of Ethics on interactions with health care professionals. <https://www.advamed.org/sites/default/files/resource/advamed-code-of-ethics-2020.pdf>.
29. European Diagnostics Manufacturers Association. Part A: interaction with health care professionals. <https://www.medtecheurope.org/resource-library/medtech-europe-code-of-ethical-business-practice/2019>.

**MULTIPLE CHOICE QUESTIONS**

1. In some countries, clinical chemistry encompasses multiple specialties. Which of the following is never included within clinical chemistry?
  - a. Hematology
  - b. Coagulation
  - c. Therapeutic drug monitoring
  - d. Cytology
  - e. Serology
2. Which of the following statements is not part of the professional role of the clinical laboratory specialist?
  - a. Develop and validate de novo laboratory tests to meet clinical needs
  - b. Evaluate and characterize the analytical and clinical performance of laboratory tests
  - c. Decide the pricing of the test and market laboratory services
  - d. Present laboratory results to clinicians in an effective manner
  - e. Determine cost-effectiveness and intrinsic value of laboratory tests
3. Each of the following guiding principles of practicing the profession is correct except?
  - a. Maintaining confidentiality of medical information
  - b. Using resources appropriately
  - c. Establishing strong ties with manufacturers
  - d. Abiding by codes of conduct
  - e. Following ethical publishing rules
4. Each of the following statements is correct except:
  - a. Reviewer must excuse himself if he has a conflict of interest regarding the manuscript
  - b. Reviewer should complete the review in a timely fashion
  - c. Reviewer should contact the author if he has a question
  - d. Reviewer should provide a thorough examination of the manuscript
  - e. Reviewer should provide useful comments to author
5. Molecular diagnostics
  - a. Is as old as clinical chemistry
  - b. Focuses on long polymers of carbohydrates
  - c. Has a long history of providing multiplex assays that translate to clinical practice
  - d. Studies the quantity or sequence of nucleic acids
  - e. Is none of the above
6. The following statements regarding the laboratory medicine director are correct except:
  - a. Must have an MD, PhD, or a pharmacy degree, depending on the country
  - b. Usually has undergone a minimum of 9 years of training
  - c. Determine the strategic direction of the laboratory
  - d. Assist physicians in test utilization
  - e. Must be an expert in computer technology
7. Which of the following is not considered part of the role of a journal editor:
  - a. Establishing conflict-of-interest policy for editors
  - b. Determining the direction of the journal
  - c. Being responsible for the integrity of the overall review process
  - d. Establishing the subscription price
  - e. Developing journal policies
8. Which of the following is not considered a recognized discipline in laboratory medicine?
  - a. Immunology
  - b. Physiology
  - c. Microbiology
  - d. Hematology
  - e. Clinical chemistry
9. Which of the following is not an important driver in transforming laboratory data to information and knowledge?
  - a. Application of evidence-based medicine
  - b. Application of information management technology
  - c. Data mining
  - d. Patient-focused laboratory medicine
  - e. Technology-driven laboratory medicine
10. With respect to handling of patient information, it is inappropriate to
  - a. Store and keep record of patient identifiers
  - b. Publicly disclose without obtaining the patient's consent
  - c. Store securely in the laboratory information system
  - d. Monitor data access by laboratory personnel
  - e. Include genetic information

# Statistical Methodologies in Laboratory Medicine

## *Analytical and Clinical Evaluation of Laboratory Tests*

*Kristian Linnet, Karel G.M. Moons, and James Clark Boyd*

### ABSTRACT

#### Background

The careful selection and evaluation of laboratory tests are key steps in the process of implementing new measurement procedures in the laboratory for clinical use. Method evaluation in the clinical laboratory is complex and in most countries is a regulated process guided by various professional recommendations and quality standards on best laboratory practice.

#### Content

This chapter deals with the statistical aspects of both analytical and clinical evaluations of laboratory assays, tests, or markers. After a short overview on basic statistics, aspects such as accuracy, precision, trueness, limit of detection, and selectivity are considered in the first part. After dealing with

comparison of assays in detail, including using difference plots and regression analysis, the focus is on quantification of the (added) diagnostic value of laboratory assays or tests. First, the evaluation of tests in isolation is outlined, which corresponds to simple diagnostic scenarios, when only a single test result is decisive (e.g., in the screening context). Subsequently, the chapter addresses the more common clinical situation in which a laboratory assay or test is considered as part of a diagnostic workup and thus a test's added value is at issue. This involves use of receiver operating characteristic (ROC) areas, reclassification measures, predictiveness curves, and decision curve analysis. Finally, principles for considering the clinical impact of diagnostic tests on actual decision making and patient outcomes are discussed.

### ASSAY SELECTION OVERVIEW

The introduction of new or revised laboratory tests, markers, or assays is a common occurrence in the clinical laboratory. Test selection and evaluation are key steps in the process of implementing new measurement procedures (Fig. 2.1). A new or revised test must be selected carefully and its analytical and clinical performance evaluated thoroughly before it is adopted for routine use in patient care (see later in this chapter and Chapter 10). Establishment of a new or revised laboratory test may also involve evaluation of the features of the automated analyzer on which the test will be implemented. When a new test is to be introduced to the routine clinical laboratory, a series of technical or analytical evaluations is commonly conducted. Assay imprecision is estimated, and comparison of the new assay versus an existing one is commonly undertaken. The allowable measurement range is assessed with estimation of the lower and upper limits of quantification. Interferences and carryover are evaluated when relevant. Depending on the situation, a limited verification of manufacturer claims may be all that is necessary, or, in the case of a newly developed test or assay, a full validation may be carried out. Subsequent subsections provide details for all these test evaluations. With regard to evaluation of reference intervals or medical decision limits, readers are referred to Chapter 9.

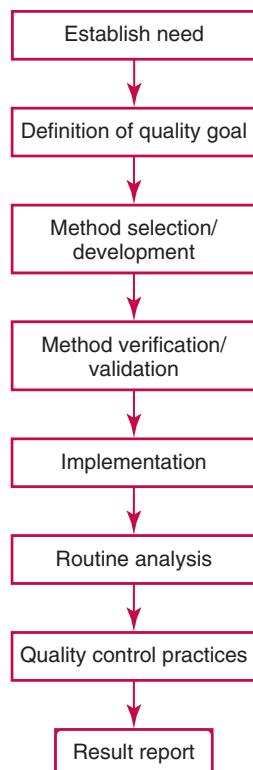
Evaluation of tests, markers, or assays in the clinical laboratory is influenced strongly by guidelines and accreditation or other regulatory standards.<sup>1–3</sup> The Clinical and Laboratory

Standards Institute (CLSI, formerly the National Committee for Clinical Laboratory Standards [NCCLS]) has published a series of consensus protocols (Clinical Laboratory Improvement Amendments [CLIs]) for clinical chemistry laboratories and manufacturers to follow when evaluating methods (see the CLSI website at <http://www.clsi.org>). The International Organization for Standardization (ISO) has also developed several documents related to method evaluation (ISOs). In addition, meeting laboratory accreditation requirements has become an important aspect in the evaluation process with accrediting agencies placing increased focus on the importance of total quality management and assessment of trueness and precision of laboratory measurements. An accompanying trend has been the emergence of an international nomenclature to standardize the terminology used for characterizing laboratory test or assay performance.

This chapter presents an overview of considerations in and methods for the evaluation of laboratory tests. This includes explanation of graphical and statistical methods that are used to aid in the test evaluation process; examples of the application of these methods are provided, and current terminology within the area is summarized. Key terms and abbreviations are listed in Box 2.1.

#### Medical Need and Quality Goals

The selection of the appropriate clinical laboratory assays is a vital part of rendering optimal patient care. Advances



**FIGURE 2.1** A flow diagram that illustrates the process of introducing a new assay into routine use.

in patient care are frequently based on the use of new or improved laboratory tests or measurements. Ascertainment of what is necessary clinically from a new or revised laboratory test is the first step in selecting the appropriate candidate test. Key parameters, such as desired turnaround time and necessary clinical utility for an assay, are often derived by discussions between laboratorians and clinicians. When new diagnostic assays are introduced, for example, reliable estimates of its diagnostic performance (e.g., predictive values, sensitivity and specificity) must be considered. With established analytes, a common scenario is the replacement of an older, labor-intensive test with a new, automated assay that is more economical in daily use. In these situations, consideration must be given to whether the candidate assay has sufficient precision, accuracy, analytical measurement range, and freedom from interference to provide clinically useful results (see Fig. 2.1).

### Analytical Performance Criteria

In evaluation of a laboratory test, (1) trueness (formerly termed accuracy), (2) precision, (3) analytical range, (4) detection limit, and (5) analytical specificity are of prime importance. The sections in this chapter on laboratory test evaluation and comparison contain detailed outlines of these concepts. Estimated test performance parameters should be related to analytical performance specifications that ensure acceptable clinical use of the test and its results. For more details related to the recommended models for setting analytical performance specifications, readers are referred to Chapters 6 and 8. From a practical point of view, the “ruggedness” of the test in routine use is of importance and reliable performance, when used by different operators and with different batches of reagents over long time periods, is essential.

When a new laboratory analyzer is at issue, various instrumental parameters require evaluation, including (1) pipetting, (2) specimen-to-specimen carryover, (3) reagent lot-to-lot variation, (4) detector imprecision, (5) time to first reportable result, (6) onboard reagent stability, (7) overall throughput, (8) mean time between instrument failures, and (9) mean time to repair. Information on most of these parameters should be available from the instrument manufacturer; the manufacturer should also be able to furnish information on what studies should be conducted in estimating these parameters for an individual analyzer. Assessment of reagent lot-to-lot variation is especially difficult for a user, and the manufacturer should provide this information.

### Other Criteria

Various categories of laboratory tests may be considered. New tests may require “in-house” development. (Note: Such a test is also referred to as a laboratory-developed test [LDT].) Commercial kit assays, on the other hand, are ready for implementation in the laboratory, often in a “closed” analytical system on a dedicated instrument. When prospective assays are reviewed, attention should be given to the following:

1. Principle of the test or assay, with original references
  2. Detailed protocol for performing the test
  3. Composition of reagents and reference materials, the quantities provided, and their storage requirements (e.g., space, temperature, light, humidity restrictions) applicable both before and after the original containers are opened
  4. Stability of reagents and reference materials (e.g., their shelf lives)
  5. Technologist time and required skills
  6. Possible hazards and appropriate safety precautions according to relevant guidelines and legislation
  7. Type, quantity, and disposal of waste generated
  8. Specimen requirements (e.g., conditions for collection and transportation, specimen volume requirements, the necessity for anticoagulants and preservatives, necessary storage conditions)
  9. Reference interval of the test and its results, including information on how such interval was derived, typical values obtained in both healthy and diseased individuals, and the necessity of determining a reference interval for one’s own institution (see Chapter 9 for details on how to generate a reference interval of a laboratory test.)
  10. Instrumental requirements and limitations
  11. Cost-effectiveness
  12. Computer platforms and interfacing with the laboratory information system
  13. Availability of technical support, supplies, and service
- Other questions concerning placement of the new or revised test in the laboratory should be taken into account. They include:
1. Does the laboratory possess the necessary measuring equipment? If not, is there sufficient space for a new instrument?
  2. Does the projected workload match the capacity of a new instrument?
  3. Is the test repertoire of a new instrument sufficient?
  4. What is the method and frequency of (re)calibration?
  5. Is staffing of the laboratory sufficient for the new technology?
  6. If training the entire staff in a new technique is required, is such training worth the possible benefits?

### BOX 2.1 Abbreviations and Vocabulary Concerning Technical Validation of Assays

#### Abbreviations

CI	Confidence interval
CV	Coefficient of variation (=SD/x, where x is the concentration)
CV%	= CV × 00%
CV <sub>A</sub>	Analytical coefficient of variation
CV <sub>G</sub>	Between-subject biological variation
CV <sub>I</sub>	Within-subject biological variation
CV <sub>RB</sub>	Sample-related random bias coefficient of variation
DoD	Distribution of differences (plot)
ISO	International Organization for Standardization
IUPAC	International Union of Pure and Applied Chemistry
OLR	Ordinary least-squares regression analysis
SD	Standard deviation
SEM	Standard error of the mean (5SD/√N)
SD <sub>A</sub>	Analytical standard deviation
SD <sub>RB</sub>	Sample-related random bias standard deviation
x <sub>m</sub>	Mean
x <sub>mv</sub>	Weighted mean
WLR	Weighted least-squares regression analysis

#### Vocabulary<sup>a</sup>

**Analyte** Compound that is measured.

**Bias** Difference between the average (strictly the expectation) of the test results and an accepted reference value (ISO 3534-1). Bias is a measure of trueness.<sup>11</sup>

**Certified reference material (CRM)** is a reference material, one or more of whose property values are certified by a technically valid procedure, accompanied by or traceable to a certificate or other documentation that is issued by a certifying body.

**Commutability** Ability of a material to yield the same results of measurement by a given set of measurement procedures.

**Limit of detection** The lowest amount of analyte in a sample that can be detected but not quantified as an exact value. Also called lower limit of detection or minimum detectable concentration (or dose or value).<sup>23</sup>

**Lower limit of quantification (LLOQ)** The lowest concentration at which the measurement procedure fulfills specifications for imprecision and bias (corresponds to the *lower limit of determination* mentioned under *Measuring interval*).

**Matrix** All components of a material system except the analyte.

**Measurand** The “quantity” that is actually measured (e.g., the concentration of the analyte). For example, if the analyte is glucose, the measurand is the concentration of glucose. For

an enzyme, the measurand may be the enzyme activity or the *mass concentration* of enzyme.

**Measuring interval** Closed interval of possible values allowed by a measurement procedure and delimited by the *lower limit of determination* and the *higher limit of determination*. For this interval, the total error of the measurements is within specified limits for the method. Also called the *analytical measurement range*.

**Primary measurement standard** Standard that is designated or widely acknowledged as having the highest metrologic qualities and whose value is accepted without reference to other standards of the same quantity.<sup>73</sup>

**Quantity** The amount of substance (e.g., the concentration of substance). **Random error** Arises from unpredictable variations in influence quantities. These random effects give rise to variations in repeated observations of the measurand.

**Reference material (RM)** A material or substance, one or more properties of which are sufficiently well established to be used for the calibration of a method or for assigning values to materials.

**Reference measurement procedure** Thoroughly investigated measurement procedure shown to yield values having an uncertainty of measurement commensurate with its intended use, especially in assessing the trueness of other measurement procedures for the same quantity and in characterizing reference materials.

**Selectivity or specificity** Degree to which a method responds uniquely to the required analyte.

**Systematic error** A component of error that, in the course of a number of analyses of the same measurand, remains constant or varies in a predictable way.

**Traceability** “The property of the result of a measurement or the value of a standard whereby it can be related to stated references, usually national or international standards, through an unbroken chain of comparisons all having stated uncertainties.”<sup>43</sup> This is achieved by establishing a chain of calibrations leading to primary national or international standards, ideally (for long-term consistency) the Système International (SI) units of measurement.

**Uncertainty** A parameter associated with the result of a measurement that characterizes the dispersion of values that could reasonably be attributed to the measurand. More briefly, *uncertainty* is a parameter characterizing the range of values within which the value of the quantity being measured is expected to lie.

**Upper limit of quantification (ULOQ)** The highest concentration at which the measurement procedure fulfills specifications for imprecision and bias (corresponds to the *upper limit of determination* mentioned under *Measuring interval*).

<sup>a</sup>A listing of terms of relevance in relation to analytical methods is displayed. Many of the definitions originate from Dybkær<sup>12</sup> with statement of original source where relevant (e.g., International Organization for Standardization document number). Others are derived from the Eurachem/Citac guideline on uncertainty.<sup>79</sup> In some cases, slight modifications have been performed for the sake of simplicity.

7. How frequently will quality control (QC) samples be run?
8. What materials will be used to ensure QC?
9. What approach will be used for proficiency testing?
10. What is the estimated cost of performing an assay using the proposed method, including the costs of calibrators, QC specimens, and technologists’ time? Questions applicable to implementation of new instrumentation in a particular laboratory may also be relevant. Does the instrument satisfy local electrical safety guidelines? What are the power, water, drainage, and air conditioning requirements of the instrument? If the instrument is large, does the floor have sufficient load-bearing capacity?

A qualitative assessment of all these factors is often completed, but it is possible to use a value scale to assign points

to the various features weighted according to their relative importance; the latter approach allows a more quantitative test evaluation process. Decisions are then made regarding the assays that best fit the laboratory’s requirements and that have the potential for achieving the necessary analytical quality for clinical use.

## BASIC STATISTICS

In this section, fundamental statistical concepts and techniques are introduced in the context of typical analytical investigations. The basic concepts of (1) populations, (2) samples, (3) parameters, (4) statistics, and (5) probability distributions are defined and illustrated. Two important

probability distributions—Gaussian and Student *t*—are introduced and discussed.

### Frequency Distribution

A graphical device for displaying a large set of laboratory test results is the *frequency distribution*, also called a *histogram*. Fig. 2.2 shows a frequency distribution displaying the results of serum gamma-glutamyltransferase (GGT) measurements of 100 apparently healthy 20- to 29-year-old men. The frequency distribution is constructed by dividing the measurement scale into cells of equal width; counting the number,  $n_i$ , of values that fall within each cell; and drawing a rectangle above each cell whose area (and height because the cell widths are all equal) is proportional to  $n_i$ . In this example, the selected cells were 5 to 9, 10 to 14, 15 to 19, 20 to 24, 25 to 29, and so on, with 60 to 64 being the last cell (range of values, 5 to 64 U/L). The ordinate axis of the frequency distribution gives the number of values falling within each cell. When this number is divided by the total number of values in the data set, the relative frequency in each cell is obtained.

Often, the position of the value for an individual within a distribution of values is useful medically. The *nonparametric* approach can be used to directly determine the *percentile* of a given subject. Having ranked  $N$  subjects according to their values, the  $n$ -percentile,  $\text{Perc}_{n\%}$ , may be estimated as the value of the  $[N(n/100) + 0.5]$  ordered observation.<sup>4</sup> In the case of a noninteger value, interpolation is carried out between neighbor values. The 50th percentile is the median of the distribution.

### Population and Sample

It is useful to obtain information and draw conclusions about the characteristics of the test results for one or more target populations. In the GGT example, interest is focused on the location and spread of the population of GGT values for 20- to 29-year-old healthy men. Thus a working definition of a *population* is the complete set of all observations that might occur as a result of performing a particular procedure according to specified conditions.

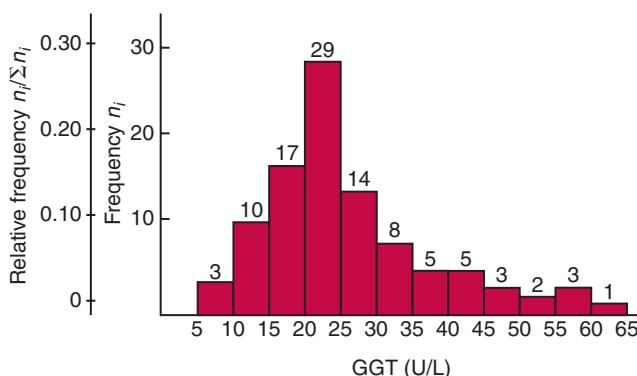
Most target populations of interest in clinical chemistry are in principle very large (millions of individuals) and so are impossible to study in their entirety. Usually a subgroup of observations is taken from the population as a basis for forming conclusions about population characteristics. The group of observations that has actually been selected from the population is called a *sample*. For example, the 100 GGT

values make up a sample from a respective target population. However, a sample is used to study the characteristics of a population only if it has been properly selected. For instance, if the analyst is interested in the population of GGT values over various lots of materials and some time period, the sample must be selected to be representative of these factors, as well as of age, sex, and health factors of the individuals in the targeted population. Consequently, exact specification of the target population(s) is necessary before a plan for obtaining the sample(s) can be designed. In this chapter, a sample is also used as a specimen, depending on the context.

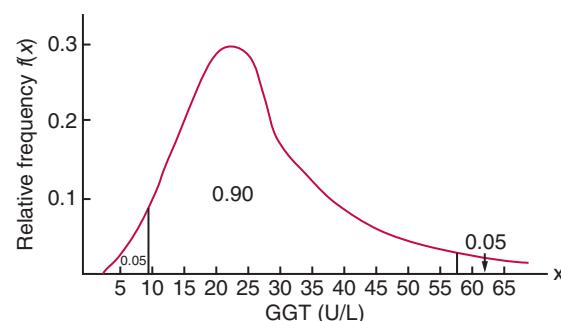
### Probability and Probability Distributions

Consider again the frequency distribution in Fig. 2.2. In addition to the general location and spread of the GGT determinations, other useful information can be easily extracted from this frequency distribution. For instance, 96% (96 of 100) of the determinations are less than 55 U/L, and 91% (91 of 100) are greater than or equal to 10 but less than 50 U/L. Because the cell interval is 5 U/L in this example, statements such as these can be made only to the nearest 5 U/L. A larger sample would allow a smaller cell interval and more refined statements. For a sufficiently large sample, the cell interval can be made so small that the frequency distribution can be approximated by a continuous, smooth curve, similar to that shown in Fig. 2.3. In fact, if the sample is large enough, we can consider this a close representation of the “true” target population frequency distribution. In general, the functional form of the population frequency distribution curve of a variable  $x$  is denoted by  $f(x)$ .

The population frequency distribution allows us to make probability statements about the GGT of a randomly selected member of the population of healthy 20- to 29-year-old men. For example, the probability  $\text{Pr}(x > x_a)$  that the GGT value  $x$  of a randomly selected 20- to 29-year-old healthy man is greater than some particular value  $x_a$  is equal to the area under the population frequency distribution to the right of  $x_a$ . If  $x_a = 58$ , then from Fig. 2.3,  $\text{Pr}(x > 58) = 0.05$ . Similarly, the probability  $\text{Pr}(x_a < x < x_b)$  that  $x$  is greater than  $x_a$  but less than  $x_b$  is equal to the area under the population frequency distribution between  $x_a$  and  $x_b$ . For example, if  $x_a = 9$  and  $x_b = 58$ , then from Fig. 2.3,  $\text{Pr}(9 < x < 58) = 0.90$ . Because the population frequency distribution provides all information related to probabilities of a randomly selected member of the population, it is called the probability distribution of the population. Although the true probability distribution is never exactly known in practice, it can be approximated with a large sample of observations, that is, test results.



**FIGURE 2.2** Frequency distribution of 100 gamma-glutamyltransferase (GGT) values.



**FIGURE 2.3** Population frequency distribution of gamma-glutamyltransferase (GGT) values.

## Parameters: Descriptive Measures of a Population

Any population of values can be described by measures of its characteristics. A *parameter* is a constant that describes some particular characteristic of a population. Although most populations of interest in analytical work are infinite in size, for the following definitions, we shall consider the population to be of finite size  $N$ , where  $N$  is very large.

One important characteristic of a population is its *central location*. The parameter most commonly used to describe the central location of a population of  $N$  values is the *population mean* ( $\mu$ ):

$$\mu = \frac{\sum x_i}{N}$$

An alternative parameter that indicates the central tendency of a population is the *median*, which is defined as the 50th percentile,  $\text{Perc}_{50}$ .

Another important characteristic is the *dispersion* of values about the population mean. A parameter very useful in describing this dispersion of a population of  $N$  values is the *population variance*  $\sigma^2$  (sigma squared):

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

The *population standard deviation* (SD)  $\sigma$ , the positive square root of the population variance, is a parameter frequently used to describe the population dispersion in the same units (e.g., mg/dL) as the population values. For a Gaussian distribution, 95% of the population of values are located within the mean  $\pm 1.96 \sigma$ . If a distribution is non-Gaussian (e.g., asymmetric), an alternative measure of dispersion based on the percentiles may be more appropriate, such as the distance between the 25th and 75th percentiles (the interquartile interval).

## Statistics: Descriptive Measures of the Sample

As noted earlier, clinical chemists usually have at hand only a sample of observations (i.e., test results) from the overarching targeted population. A *statistic* is a value calculated from the observations in a sample to estimate a particular characteristic of the target population. As introduced earlier, the sample mean  $x_m$  is the arithmetical average of a sample, which is an estimate of  $\mu$ . Likewise, the sample SD is an estimate of  $\sigma$ , and the coefficient of variation (CV) is the ratio of the SD to the mean multiplied by 100%. The equations used to calculate  $x_m$ , SD, and CV, respectively, are as follows:

$$\begin{aligned} x_m &= \frac{\sum x_i}{N} \\ \text{SD} &= \sqrt{\frac{\sum (x_i - x_m)^2}{N-1}} = \sqrt{\frac{\sum x_i^2 - (\sum x_i)^2}{N-1}} \\ \text{CV} &= \frac{\text{SD}}{x_m} \times 100\% \end{aligned}$$

where  $x_i$  is an individual measurement and  $N$  is the number of sample measurements.

The SD is an estimate of the dispersion of the distribution. Additionally, from the SD, we can derive an estimate of the uncertainty of  $x_m$  as an estimate of  $\mu$  (see later discussion).

## Random Sampling

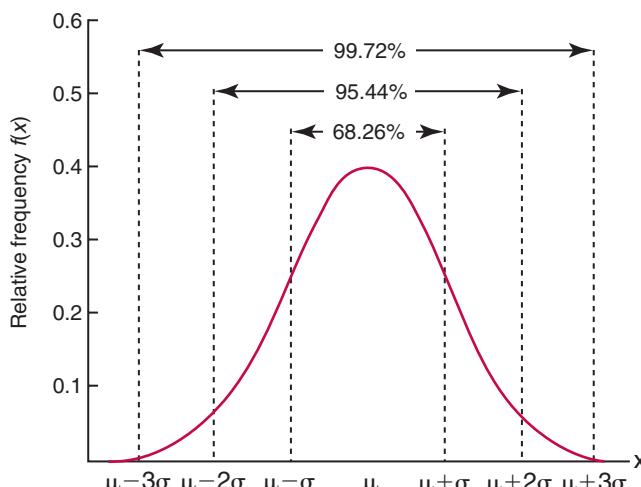
A random sample of individuals from a target population is one in which each member of the population has an equal chance of being selected. A *random sample* is one in which each member of the sample can be considered to be a random selection from the target population. Although much of statistical analysis and interpretation depends on the assumption of a random sample from some population, actual data collection often does not satisfy this assumption. In particular, for sequentially generated data, it is often true that observations adjacent to each other tend to be more alike than observations separated in time.

## The Gaussian Probability Distribution

The *Gaussian probability distribution*, illustrated in Fig. 2.4, is of fundamental importance in statistics for several reasons. As mentioned earlier, a particular test result  $x$  will not usually be equal to the true value  $\mu$  of the specimen being measured. Rather, associated with this particular test result  $x$  will be a particular measurement error  $\epsilon = x - \mu$ , which is the result of many contributing sources of error. Pure measurement errors tend to follow a probability distribution similar to that shown in Fig. 2.4, where the errors are symmetrically distributed, with smaller errors occurring more frequently than larger ones, and with an expected value of 0. This important fact is known as the central limit effect for distribution of errors: if a measurement error  $\epsilon$  is the sum of many independent sources of error, such as  $\epsilon_1, \epsilon_2, \dots, \epsilon_k$ , several of which are major contributors, the probability distribution of the measurement error  $\epsilon$  will tend to be Gaussian as the number of sources of error becomes large.

Another reason for the importance of the Gaussian probability distribution is that many statistical procedures are based on the assumption of a Gaussian distribution of values; this approach is commonly referred to as *parametric*. Furthermore, these procedures usually are not seriously invalidated by departures from this assumption. Finally, the magnitude of the uncertainty associated with sample statistics can be ascertained based on the fact that many sample statistics computed from large samples have a Gaussian probability distribution.

The Gaussian probability distribution is completely characterized by its mean  $\mu$  and its variance  $\sigma^2$ . The notation



**FIGURE 2.4** The Gaussian probability distribution.

$N(\mu, \sigma^2)$  is often used for the distribution of a variable that is Gaussian with mean  $\mu$  and variance  $\sigma^2$ . Probability statements about a variable  $x$  that follows an  $N(\mu, \sigma^2)$  distribution are usually made by considering the variable  $z$ ,

$$z = \frac{x - \mu}{\sigma}$$

which is called the *standard Gaussian variable*. The variable  $z$  has a Gaussian probability distribution with  $\mu = 0$  and  $\sigma^2 = 1$ , that is,  $z$  is  $N(0, 1)$ . The probability that  $x$  is within  $2\sigma$  of  $\mu$  [i.e.,  $\Pr(|x - \mu| < 2\sigma) = ]$  is 0.9544. Most computer spreadsheet programs can calculate probabilities for all values of  $z$ .

### Student *t* Probability Distribution

To determine probabilities associated with a Gaussian distribution, it is necessary to know the population SD  $\sigma$ . In actual practice,  $\sigma$  is often unknown, so we cannot calculate  $z$ . However, if a random sample can be taken from the Gaussian population, we can calculate the sample SD, substitute SD for  $\sigma$ , and compute the value  $t$ :

$$t = \frac{x - \mu}{SD}$$

Under these conditions, the variable  $t$  has a probability distribution called the *Student *t* distribution*. The *t* distribution is really a family of distributions depending on the degrees of freedom (df)  $v (= N - 1)$  for the sample SD. Several *t* distributions from this family are shown in Fig. 2.5. When the size of the sample and the df for SD are infinite, there is no uncertainty in SD, so the *t* distribution is identical to the standard Gaussian distribution. However, when the sample size is small, the uncertainty in SD causes the *t* distribution to have greater dispersion and heavier tails than the standard Gaussian distribution, as illustrated in Fig. 2.5. At sample sizes above 30, the difference between the *t*-distribution and the Gaussian distribution becomes relatively small and can usually be neglected. Most computer spreadsheet programs can calculate probabilities for all values of  $t$ , given the df for SD.

The Student *t* distribution is commonly used in significance tests, such as comparison of sample means, or in testing conducted if a regression slope differs significantly from 1. Descriptions of these tests can be found in statistics textbooks.<sup>5</sup> Another important application is the estimation of confidence intervals (CIs). CIs are intervals that indicate the uncertainty of a given sample estimate. For example, it can be

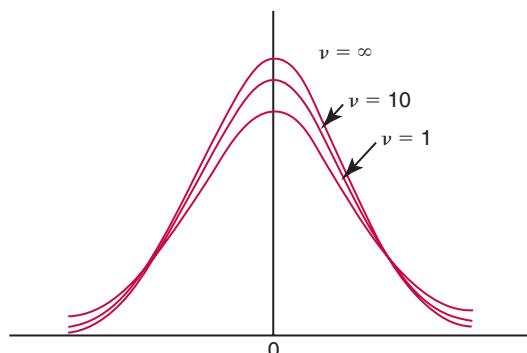


FIGURE 2.5 The *t* distribution for  $v = 1, 10$ , and  $\infty$ .

proved that  $X_m \pm t_{alpha} (\text{SD}/N^{0.5})$  provides an approximate  $2\alpha$ -CI for the mean. A common value for  $alpha$  is 0.025 or 2.5%, which thus results in a 0.95% or 95% CI. Given sample sizes of 30 or higher,  $t_{alpha}$  is ca. 2. ( $\text{SD}/N^{0.5}$ ) is called the standard error (SE) of the mean. A CI should be interpreted as follows. Suppose a sampling experiment of drawing 30 observations from a Gaussian population of values is repeated 100 times, and in each case, the 95% CI of the mean is calculated as described. Then, in 95% of the drawings, the true mean  $\mu$  is included in the 95% CI. The popular interpretation is that for an estimated 95% CI, there is 95% chance that the true mean is within the interval. According to the central limit theorem, distributions of mean values converge toward the Gaussian distribution irrespective of the primary type of distribution of  $x$ . This means that the 95% CI is a robust estimate only minimally influenced by deviations from the Gaussian distribution. In the same way, the *t*-test is robust toward deviations from normality.

### Nonparametric Statistics

Distribution-free statistics, often called nonparametric statistics, provides an alternative to parametric statistical procedures that assume data to have Gaussian distributions. For example, distributions of reference values are often skewed and so do not conform to the Gaussian distribution (see Chapter 9 on reference intervals). Formally, one can carry out a goodness of fit test to judge whether a distribution is Gaussian or not.<sup>5</sup> A commonly used test is the Kolmogorov-Smirnov test, in which the shape of the sample distribution is compared with the shape presumed for a Gaussian distribution. If the difference exceeds a given critical value, the hypothesis of a Gaussian distribution is rejected, and it is then appropriate to apply nonparametric statistics. A special problem is the occurrence of outliers (i.e., single measurements highly deviating from the remaining measurements). Outliers may rely on biological factors and so be of real significance (e.g., in the context of estimating reference intervals or be related to clerical errors). Special tests exist for handling outliers.<sup>5</sup>

Given that a distribution is non-Gaussian, it is appropriate to apply nonparametric descriptive statistics based on the percentile or quantile concept. As stated under the earlier section Frequency Distribution, the  $n$ -percentile,  $\text{Perc}_n$ , of a sample of  $N$  values may be estimated as the value of the  $[N(n/100) + 0.5]$  ordered observation.<sup>4</sup> In the case of a non-integer value, interpolation is carried out between neighbor values. The median is the 50th percentile, which is used as a measure of the center of the distribution. For the GGT example mentioned previously, we would order the  $N = 100$  values according to size. The median or 50th percentile is then the value of the  $[100(50/100) + 0.5 = 50.5]$  ordered observation (the interpolated value between the 50th and 51st ordered values). The 2.5th and 97.5th percentiles are values of the  $[100(2.5/100) + 0.5 = 3]$  and  $[100(97.5/100) + 0.5 = 98]$  ordered observations, respectively. When a 95% reference interval is estimated, a nonparametric approach is often preferable because many distributions of reference values are asymmetric. Generally, distributions based on the many biological sources of variation are often non-Gaussian compared with distributions of pure measurement errors that usually are Gaussian.

The nonparametric counterpart to the *t*-test is the Mann-Whitney test, which provides a significance test for the difference

between median values of the two groups to be compared.<sup>5</sup> When there are more than two groups, the Kruskal-Wallis test can be applied.<sup>5</sup>

### Categorical Variables

Hitherto focus has been on quantitative variables. When dealing with qualitative tests and in the context of evaluating diagnostic testing, categorical variables that only take the value positive or negative come into play. The performance is here given as proportions or percentages, which are proportions multiplied by 100. For example, the diagnostic sensitivity of a test is the proportion of diseased subjects who have a positive result. Having tested, for example, 100 patients, 80 might have had a positive test result. The sensitivity then is 0.8 or 80%. We are then interested in judging how precise this estimate is. Exact estimates of the uncertainty can be derived from the so-called binomial distribution, but for practical purposes, an approximate expression for the 95% CI is usually applied as the estimated proportion  $P \pm 2SE$ , where the SE in this context is derived as:

$$SE = [P(1 - P)/N]^{0.5}$$

where  $P$  is here a proportion and not a percentage.<sup>5</sup> In the example, the SE equals 0.0016 and so the 95% CI is 0.77 to 0.83 or 77 to 83%. The applied approximate formula for the SE is regarded as reasonably valid when  $NP$  and  $N(1 - P)$  both are equal to or higher than 5.

### POINTS TO REMEMBER

- Statistics as means, SDs, percentiles, proportions, and so on are computed from a sample of values drawn from a population and provide *estimates* of the unknown population characteristics.
- Whereas parametric statistics rely on the assumption of a Gaussian population of values, which typically applies for measurement errors, nonparametric statistics is a distribution-free approach that apply to, for example, asymmetric distributions often observed for biologic variables.
- The Gaussian distribution is characterized by the mean and the SD, and other types of distributions are described by the median and the percentile (quantile) values.
- Distributions of categorical variables are characterized by proportions or percentages and their SEs.

### TECHNICAL VALIDITY OF ANALYTICAL ASSAYS

This section defines the basic concepts used in this chapter: (1) calibration, (2) trueness and accuracy, (3) precision, (4) linearity, (5) limit of detection (LOD), (6) limit of quantification, (7) specificity, and (8) others (see Box 2.1 for definitions).

#### Calibration

The calibration function is the relation between instrument signal ( $y$ ) and concentration of analyte ( $x$ ), that is,

$$y = f(x)$$

The inverse of this function, also called the measuring function, yields the concentration from response:

$$x = f^{-1}(y)$$

This relationship is established by measurement of samples with known quantities of analyte<sup>6</sup> (calibrators). One may distinguish between solutions of pure chemical standards and samples with known quantities of analyte present in the typical matrix that is to be measured (e.g., human serum). The first situation applies typically to a reference measurement procedure that is not influenced by matrix effects; the second case corresponds typically to a routine method that often is influenced by matrix components and so preferably is calibrated using the relevant matrix.<sup>7</sup> Calibration functions may be linear or curved and, in the case of immunoassays, may often take a special form (e.g., modeled by the four-parameter logistic curve).<sup>8</sup> This model (logistic in log  $x$ ) has been used for immunoassay techniques and is written in several forms (Table 2.1). An alternative, model-free approach is to estimate a smoothed spline curve, which often is performed for immunoassays; however, a disadvantage of the spline curve approach is that it is insensitive to aberrant calibration values, fitting these just as well as the correct values. If the assumed calibration function does not correctly reflect the true relationship between instrument response and analyte concentration, a systematic error or bias is likely to be associated with the analytical method. A common problem with some immunoassays is the “hook effect,” which is a deviation from the expected calibration algorithm in the high-concentration range. (The hook effect is discussed in more detail in Chapter 26.)

The precision of the analytical method depends on the stability of the instrument response for a given quantity of analyte. In principle, a random dispersion of instrument signal (vertical direction) at a given true concentration transforms into dispersion on the measurement scale (horizontal direction), as is shown schematically (Fig. 2.6). The detailed statistical aspects of calibration are complex,<sup>5,9</sup> but in the following sections, some approximate relations are outlined. If the calibration function is linear and the imprecision of the signal response is the same over the analytical measurement range, the analytical SD ( $SD_A$ ) of the method tends to be constant over the analytical measurement range (see Fig. 2.6). If the imprecision increases proportionally to the signal response, the analytical SD of the method tends to increase proportionally to the concentration ( $x$ ), which means that the *relative* imprecision ( $CV = SD/x$ ) may be constant over the analytical measurement range if it is assumed that the intercept of the calibration line is zero.

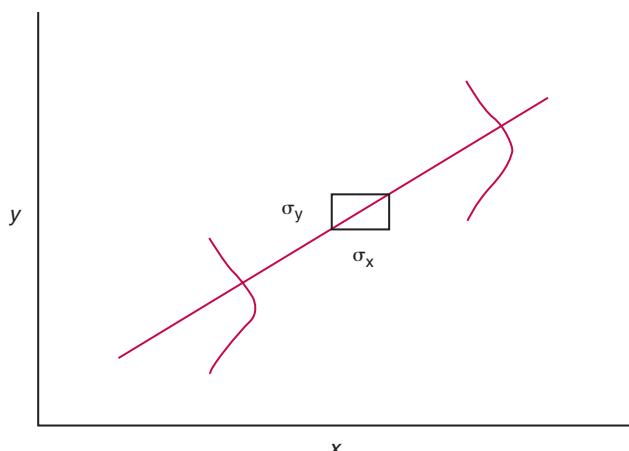
With modern, automated clinical chemistry instruments, the relation between analyte concentration and signal can in some cases be very stable, and where this is the case, calibration is necessary relatively infrequently<sup>10</sup> (e.g., at intervals of

**TABLE 2.1 The Four-Parameter Logistic Model Expressed in Three Different Forms**

Algebraic Form	Variables <sup>a</sup>	Parameters <sup>b</sup>
$y = (a - d)/(1 + (x/c)^b) + d$	( $x, y$ )	$a, b, c, d$
$R = R_0 + K_c/(1 + \exp(-\{a + b \log[C]\}))$	( $C, R$ )	$R_0, K_c, a, b$
$y = y_0 + (y_* - y_0)(x^d)/(b + x^d)$	( $x, y$ )	$y_0, y_*, b, d$

<sup>a</sup>Concentration and instrument response variables shown in parentheses.

<sup>b</sup>Equivalent letters do not necessarily denote equivalent parameters.



**FIGURE 2.6** Relation between concentration ( $x$ ) and signal response ( $y$ ) for a linear calibration function. The dispersion in signal response ( $\sigma_y$ ) is projected onto the  $x$ -axis and is called assay imprecision [ $\sigma_x$  ( $=\sigma_A$ )].

several months). Built-in process control mechanisms may help ensure that the relationship remains stable and may indicate when recalibration is necessary. In traditional chromatographic analysis (e.g., high-performance liquid chromatography [HPLC]), on the other hand, it is customary to calibrate each analytical series (run), which means that calibration is carried out daily.

### Trueness and Accuracy

Trueness of measurements is defined as closeness of agreement between the average value obtained from a large series of results of measurements and the true value.<sup>11</sup>

The difference between the average value (strictly, the mathematical expectation) and the true value is the *bias*, which is expressed numerically and so is inversely related to the trueness. Trueness in itself is a qualitative term that can be expressed, for example, as low, medium, or high. From a theoretical point of view, the exact true value for a clinical sample is not available; instead, an “accepted reference value” is used, which is the “true” value that can be determined in practice.<sup>12</sup> Trueness can be evaluated by comparison of measurements by the new test and by some preselected reference measurement procedure, both on the same sample or individuals.

The ISO has introduced the trueness expression as a replacement for the term *accuracy*, which now has gained a slightly different meaning. *Accuracy* is the closeness of agreement between the result of a measurement and a true concentration of the analyte.<sup>11</sup> Accuracy thus is influenced by both bias and imprecision and in this way reflects the total error. Accuracy, which in itself is a qualitative term, is inversely related to the “uncertainty” of measurement, which can be quantified as described later (Table 2.2).

In relation to trueness, the concepts *recovery*, *drift*, and *carryover* may also be considered. *Recovery* is the fraction or percentage increase in concentration that is measured in relation to the amount added. Recovery experiments are typically carried out in the field of drug analysis. One may distinguish between *extraction recovery*, which often is interpreted as the fraction of compound that is carried through an extraction process, and the recovery measured by the entire analytical procedure, in which the addition of an internal standard

**TABLE 2.2 An Overview of Qualitative Terms and Quantitative Measures Related to Method Performance**

Qualitative Concept	Quantitative Measure
Trueness	Bias
Closeness of agreement of mean value with “true value”	A measure of the systematic error
Precision	Imprecision (SD)
Repeatability (within run)	A measure of the dispersion of random errors
Intermediate precision (long term)	
Reproducibility (inter-laboratory)	
Accuracy	Error of measurement
Closeness of agreement of a single measurement with “true value”	Comprises both random and systematic influences

SD, Standard deviation.

compensates for losses in the extraction procedure. A recovery close to 100% is a prerequisite for a high degree of trueness, but it does not ensure unbiased results because possible nonspecificity against matrix components (e.g., an interfering substance) is not detected in a recovery experiment. *Drift* is caused by instrument or reagent instability over time, so that calibration becomes gradually biased. Assay *carryover* also must be close to zero to ensure unbiased results. Carry-over can be assessed by placing a sample with a known, low value after a pathological sample with a high value, and an observed increase can be stated as a percentage of the high value.<sup>13</sup> Drift or carryover or both may be conveniently estimated by multifactorial evaluation protocols (EPs).<sup>14,15</sup>

### Precision

Precision has been defined as the closeness of agreement between independent replicate measurements obtained under stipulated conditions.<sup>12</sup> The degree of precision is usually expressed on the basis of statistical measures of imprecision, such as SD or CV (CV = SD/ $x$ , where  $x$  is the measurement concentration), which is inversely related to precision. Imprecision of measurements is solely related to the random error of measurements and has no relation to the trueness of measurements.

Precision is specified as follows<sup>11,12</sup>:

**Repeatability:** closeness of agreement between results of successive measurements carried out under the same conditions (i.e., corresponding to within-run precision)

**Reproducibility:** closeness of agreement between results of measurements performed under changed conditions of measurements (e.g., time, operators, calibrators, reagent lots). Two specifications of reproducibility are often used: total or between-run precision in the laboratory, often termed *intermediate precision*, and interlaboratory precision (e.g., as observed in external quality assessment schemes [EQAS]) (see Table 2.2).

The total SD ( $\sigma_T$ ) may be divided into within-run and between-run components using the principle of analysis of variance of components<sup>5</sup> (variance is the squared SD):

$$\sigma^2 T = \sigma^2_{\text{Within-run}} + \sigma^2_{\text{Between-run}}$$

It is not always clear in clinical chemistry publications what is meant by “between-run” variation. Some authors use

the term to refer to the total variation of an assay, but others apply the term *between-run variance component* as defined earlier. The distinction between these definitions is important but is not always explicitly stated.

In laboratory studies of analytical variation, estimates of imprecision are obtained. The more observations, the more certain are the estimates. It is important to have an adequate number so that that analytical variation is not underestimated. Commonly, the number 20 is given as a reasonable number of observations (e.g., suggested in the CLSI guideline for manufacturers).<sup>16</sup> To verify method precision by users, it has been recommended to run internal QC samples for five consecutive days in five replicates.<sup>17</sup> If too few replications are applied, it is likely that the analytical variation will be underestimated.

To estimate both the within-run imprecision and the total imprecision, a common approach is to measure duplicate control samples in a series of runs. Suppose, for example, that a control is measured in duplicate for 20 runs, in which case 20 observations are present with respect to both components. The dispersion of the means ( $x_m$ ) of the duplicates is given as follows:

$$\sigma_{x_m}^2 = \sigma_{\text{Within-run}}^2 / 2 + \sigma_{\text{Between-run}}^2$$

From the 20 sets of duplicates, we may derive the within-run SD using the following formula:

$$\text{SD}_{\text{Within-run}} = \left[ \sum d_i^2 / (2 \times 20) \right]^{0.5}$$

where  $d_i$  refers to the difference between the  $i$ th set of duplicates. When SDs are estimated, the concept df is used. In a simple situation, the number of df equals  $N - 1$ . For  $N$  duplicates, the number of df is  $N(2 - 1) = N$ . Thus both variance components are derived in this way. The advantage of this approach is that the within-run estimate is based on several runs, so that an average estimate is obtained rather than only an estimate for one particular run if all 20 observations had been obtained in the same run. The described approach is a simple example of a *variance component analysis*. The principle can be extended to more components of variation. For example, in the CLSI EP05-A3 guideline,<sup>16</sup> a procedure is outlined that is based on the assumption of two analytical runs per day, in which case within-run, between-run, and between-day components of variance are estimated by a *nested* component of variance analysis approach.

Nothing definitive can be stated about the selected number of 20. Generally, the estimate of the imprecision improves as more observations become available. Exact confidence limits for the SD can be derived from the  $\chi^2$  distribution. Estimates of the variance,  $SD^2$ , are distributed according to the  $\chi^2$  distribution (tabulated in most statistics textbooks) as follows:  $(N - 1) SD^2 / \sigma^2 \approx \chi^2_{(N-1)}$ , where  $(N - 1)$  is the df.<sup>5</sup> Then the two-sided 95% CI is derived from the following relation:

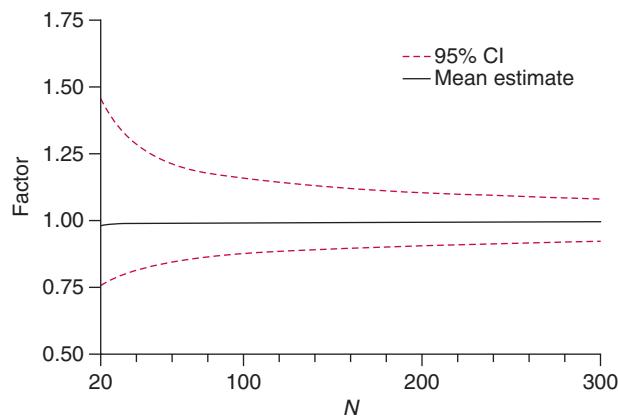
$$\Pr \left[ \chi^2_{97.5\%(N-1)} < (N-1)SD^2 / \sigma^2 < \chi^2_{2.5\%(N-1)} \right] = 0.95$$

which yields this 95% CI expression:

$$SD \times \left[ (N-1) / \chi^2_{2.5\%(N-1)} \right]^{0.5} < \sigma < SD \times \left[ (N-1) / \chi^2_{97.5\%(N-1)} \right]^{0.5}$$

### Example

Suppose we have estimated the imprecision as an SD of 5.0 on the basis of  $N = 20$  observations. From a table of



**FIGURE 2.7** Relation between factors indicating the 95% confidence intervals (CIs) of standard deviations (SDs) and the sample size. The true SD is 1, and the solid line indicates the mean estimate, which is slightly downward biased at small sample sizes.

the  $\chi^2$  distribution, we obtain the following 2.5 and 97.5 percentiles:

$$\chi^2_{2.5\%(19)} = 32.9 \text{ and } \chi^2_{97.5\%(19)} = 8.91$$

where 19 within the parentheses refers to the number of df. Substituting in the equation, we get

$$5.0 \times (19/32.9)^{0.5} < \sigma < 5.0 \times (19/8.91)^{0.5}$$

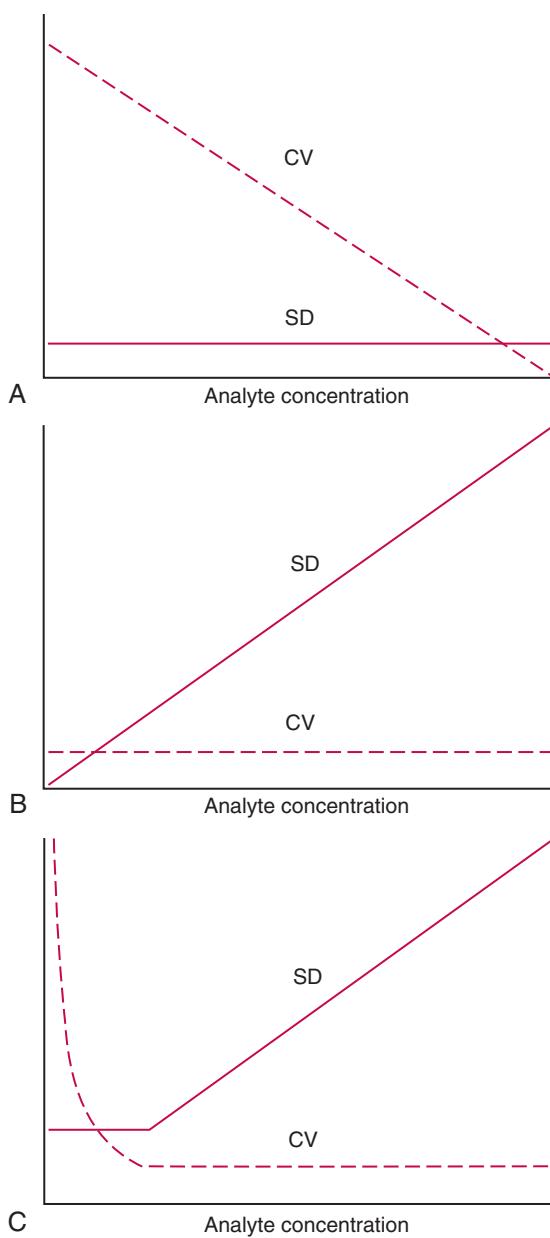
or

$$3.8 < \sigma < 7.3$$

A graphical display of 95% CIs at various sample sizes is shown in Fig. 2.7. For individual variance components, the relations are more complicated.

### Precision Profile

Precision often depends on the concentration of analyte being considered. A presentation of precision as a function of analyte concentration is the precision profile, which usually is plotted in terms of the SD or the CV as a function of analyte concentration (Fig. 2.8). Some typical examples may be considered. First, the SD may be constant (i.e., independent of the concentration), as it often is for analytes with a limited range of values (e.g., electrolytes). When the SD is constant, the CV varies inversely with the concentration (i.e., it is high in the lower part of the range and low in the high range). For analytes with extended ranges (e.g., hormones), the SD frequently increases as the analyte concentration increases. If a proportional relationship exists, the CV is constant. This may often apply approximately over a large part of the analytical measurement range. Actually, this relationship is anticipated for measurement error that arises because of imprecise volume dispensing. Often a more complex relationship exists. Not infrequently, the SD is relatively constant in the low range, so that the CV increases in the area approaching the lower limit of quantification (LLOQ). At intermediate concentrations, the CV may be relatively constant and perhaps may decline somewhat at increasing concentrations. A square root relationship can be used to model the relationship in some situations as an intermediate form of relation between the constant and the proportional case. The relationship between the SD and the concentration is of importance (1) when method specifications over the analytical measurement



**FIGURE 2.8** Relations between analyte concentration and standard deviation (SD)/coefficient of variation (CV). **A**, The SD is constant, so that the CV varies inversely with the analyte concentration. **B**, The CV is constant because of a proportional relationship between concentration and SD. **C**, A mixed situation with constant SD in the low range and a proportional relationship in the rest of the analytical measurement range.

range are considered, (2) when limits of quantification are determined, and (3) in the context of selecting appropriate statistical methods for method comparison (e.g., whether a difference or a relative difference plot should be applied, whether a simple or a weighted regression analysis procedure should be used) (see the “Relative Distribution of Differences Plot” and “Regression Analysis” sections later).

### Linearity

Linearity refers to the relationship between measured and expected values over the analytical measurement range. Linearity may be considered in relation to actual or relative

analyte concentrations. In the latter case, a dilution series of a sample may be examined. This dilution series examines whether the measured concentration changes as expected according to the proportional relationship between samples introduced by the dilution factor. Dilution is usually carried out with an appropriate sample matrix (e.g., human serum [individual or pooled serum] or a verified sample diluent).

Evaluation of linearity may be conducted in various ways. A simple, but subjective, approach is to visually assess whether the relationship between measured and expected concentrations is linear. A more formal evaluation may be carried out on the basis of statistical tests. Various principles may be applied here. When repeated measurements are available at each concentration, the random variation between measurements and the variation around an estimated regression line may be evaluated statistically<sup>18</sup> (by an F-test). This approach has been criticized because it relates only the magnitudes of random and systematic error without taking the absolute deviations from linearity into account. For example, if the random variation among measurements is large, a given deviation from linearity may not be declared statistically significant. On the other hand, if the random measurement variation is small, even a very small deviation from linearity that may be clinically unimportant is declared significant. When significant nonlinearity is found, it may be useful to explore nonlinear alternatives to the linear regression line (i.e., polynomials of higher degrees).<sup>19</sup>

Another commonly applied approach for detecting nonlinearity is to assess the residuals of an estimated regression line and test whether positive and negative deviations are randomly distributed. This can be carried out by a runs test (see “Regression Analysis” section).<sup>20</sup> An additional consideration for evaluating proportional concentration relationships is whether an estimated regression line passes through zero or not. The presence of linearity is a prerequisite for a high degree of trueness. A CLSI guideline suggests procedure(s) for assessment of linearity.<sup>21</sup>

### Analytical Measurement Range and Limits of Quantification

The analytical measurement range (measuring interval, reportable range) is the analyte concentration range over which measurements are within the declared tolerances for imprecision and bias of the method.<sup>12</sup> Taking drug assays as an example, there exist (arbitrary) requirements of a CV% of less than 15% and a bias of less than 15%.<sup>22</sup> The measurement range then extends from the lowest concentration (LLOQ) to the highest concentration (upper limit of quantification [ULOQ]) for which these performance specifications are fulfilled.

The LLOQ is medically important for many analytes. Thyroid-stimulating hormone (TSH) is a good example. As assay methods improved, lowering the LLOQ, low TSH results could be increasingly distinguished from the lower limit of the reference interval, making the test increasingly useful for the diagnosis of hyperthyroidism.

The LOD is another characteristic of an assay. The LOD may be defined as the lowest value that confidently exceeds the measurements of a blank sample. Thus the limit has been estimated on the basis of repeated measurements of a blank sample and has been *reported* as the mean plus 2 or 3 SDs of the blank measurements. In the interval from LOD up to LLOQ, one should report a result as “detected” but not

provide a quantitative result. More complicated approaches for estimation of the LOD have been suggested.<sup>23</sup>

### Analytical Sensitivity

The LLOQ of an assay should not be confused with analytical sensitivity. That is defined as ability of an analytical method to assess small differences in the concentration of analyte.<sup>6</sup> The smaller the random variation of the instrument response and the steeper the slope of the calibration function at a given point, the better is the ability to distinguish small differences in analyte concentrations. In reality, analytical sensitivity depends on the precision of the method. The smallest difference that will be statistically significant equals  $2\sqrt{2} \text{ SD}_A$  at a 5% significance level. Historically, the meaning of the term *analytical sensitivity* has been the subject of much discussion.

### Analytical Specificity and Interference

Analytical specificity is the ability of an assay procedure to determine the concentration of the target analyte without influence from potentially interfering substances or factors in the sample matrix (e.g., hyperlipemia, hemolysis, bilirubin, antibodies, other metabolic molecules, degradation products of the analyte, exogenous substances, anticoagulants). Interferences from hyperlipemia, hemolysis, and bilirubin are generally concentration dependent and can be quantified as a function of the concentration of the interfering compound.<sup>24</sup> In the context of a drug assay, specificity in relation to drug metabolites is relevant, and in some cases, it is desirable to measure the parent drug, as well as metabolites. A detailed protocol for evaluation of interference has been published by the CLSI.<sup>25</sup>

### POINTS TO REMEMBER

- Technical validation of analytical methods focuses on (1) calibration, (2) trueness and accuracy, (3) precision, (4) linearity, (5) LOD, (6) limit of quantification, (7) specificity, and (8) others.
- The difference between the average measured value and the true value is the *bias*, which can be evaluated by comparison of measurements by the new test and by some preselected reference measurement procedure, both on the same sample or individuals.
- The degree of precision is usually expressed on the basis of statistical measures of imprecision, such as SD or CV ( $CV = SD/x$ , where  $x$  is the measurement concentration).
- The measurement range extends from the lowest concentration (LLOQ) to the highest concentration (ULOQ) for which the analytical performance specifications are fulfilled (imprecision, bias).
- Analytical specificity is the ability of an assay procedure to determine the concentration of the target analyte without influence from potentially interfering substances or factors in the sample matrix.

### QUALITATIVE METHODS

Qualitative methods, which currently are gaining increased use in the form of point-of-care testing (POCT), are designed to distinguish between results below and above a predefined cutoff value. Note that the cutoff point should not be confused

with the detection limit. These tests are assessed primarily on the basis of their ability to correctly classify results in relation to the cutoff value.

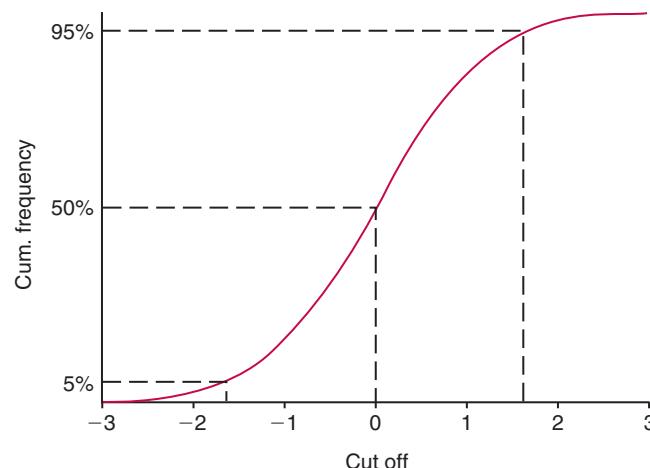
### Diagnostic Accuracy Measures

The probability of classifying a result as positive (exceeding the cutoff) when the true value indeed exceeds the cutoff is called *sensitivity*. The probability of classifying a result as negative (below the cutoff) when the true value indeed is below the cutoff is termed *specificity*. Determination of sensitivity and specificity is based on comparison of test results with a gold standard. The gold standard may be an independent test that measures the same analyte, but it may also be a clinical diagnosis determined by definitive clinical methods (e.g., radiographic testing, follow-up, outcomes analysis). Determination of these performance measures is covered later on in the diagnostic testing part. Sensitivity and specificity may be given as a fraction or as a percentage after multiplication by 100. SEs of estimates are derived as described for categorical variables. The performance of two qualitative tests applied in the same groups of nondiseased and diseased subjects can be compared using the McNemar's test, which is based on a comparison of paired values of true and false-positive (FP) or false-negative (FN) results.<sup>26</sup>

One approach for determining the recorded performance of a test in terms of sensitivity and specificity is to determine the true concentration of analyte using an independent reference method. The closer the concentration is to the cutoff point, the larger the error frequencies are expected to be. Actually, the cutoff point is defined in such a way that for samples having a true concentration exactly equal to the cutoff point, 50% of results will be positive, and 50% will be negative.<sup>27</sup> Concentrations above and below the cutoff point at which repeated results are 95% positive or 95% negative, respectively, have been called the "95% interval" for the cutoff point for that method, which indicates a grey zone where the test does not provide reliable results (Fig. 2.9).<sup>27,28</sup>

### Agreement Between Qualitative Tests

As outlined previously, if the outcome of a qualitative test can be related to a true analyte concentration or a definitive



**FIGURE 2.9** Cumulative frequency distribution of positive results. The x-axis indicates concentrations standardized to zero at the cutoff point (50% positive results) with unit standard deviation.

**TABLE 2.3 2 × 2 Table for Assessing Agreement Between Two Qualitative Tests**

		TEST 1		
		+	-	
Test 2	+	a	b	
	-	c	d	
Total		a + c	b + d	

clinical diagnosis, it is relatively straightforward to express the performance in terms of clinical specificity and sensitivity. In the absence of a definitive reference or “gold standard,” one should be cautious with regard to judgments on performance. In this situation, it is primarily *agreement* with another test that can be assessed. When replacement of an old or expensive routine assay with a new or less expensive assay is considered, it is of interest to know whether similar test results are likely to be obtained. If both assays are imperfect, however, it is not possible to judge which test has the better performance unless additional testing by a reference procedure is carried out.

In a comparison study, the same individuals are tested by both methods to prevent bias associated with selection of patients. Basically, the outcome of the comparison study should be presented in the form of a  $2 \times 2$  table, from which various measures of agreement may be derived (Table 2.3). An obvious measure of agreement is the overall fraction or percentage of subjects tested who have the same test result (i.e., both results negative or positive):

$$\text{Overall percent agreement} = (a + d)/(a + b + c + d) \times 100\%$$

If agreement differs with respect to diseased and healthy individuals, the overall percent agreement measure becomes dependent on disease prevalence in the studied group of subjects. This is a common situation; accordingly, it may be desirable to separate this overall agreement measure into agreement concerning negative and positive results:

$$\text{Percent agreement given test 1 positive: } a/(a + c)$$

$$\text{Percent agreement given test 1 negative: } b/(b + d)$$

For example, if there is a close agreement with regard to positive results, overall agreement will be high when the fraction of diseased subjects is high; however, in a screening situation with very low disease prevalence, overall agreement will mainly depend on agreement with regard to negative results.

A problem with the simple agreement measures is that they do not take agreement by chance into account. Given independence, expected proportions observed in fields of the  $2 \times 2$  table are obtained by multiplication of the fraction's negative and positive results for each test. Concerning agreement, it is excess agreement beyond chance that is of interest. More sophisticated measures have been introduced to account for this aspect. The most well-known measure is kappa, which is defined generally as the ratio of observed excess agreement beyond chance to maximum possible excess agreement beyond chance.<sup>29</sup> We have the following:

$$\text{Kappa} = (I_o - I_e)/(1 - I_e)$$

**TABLE 2.4 2 × 2 Table With Example of Agreement of Data for Two Qualitative Tests**

		TEST 1		
		+	-	Total
Test 2	+	60	20	80
	-	15	40	55
Total		75	60	135

where  $I_o$  is the observed index of agreement and  $I_e$  is the expected agreement from chance. Given complete agreement, kappa equals +1. If observed agreement is greater than or equal to chance agreement, kappa is larger than or equal to zero. Observed agreement less than chance yields a negative kappa value.

### Example

Table 2.4 shows a hypothetical example of observed numbers in a  $2 \times 2$  table. The proportion of positive results for test 1 is  $75/(75 + 60) = 0.555$ , and for test 2, it is  $80/(80 + 55) = 0.593$ . Thus by chance, we expect the ++ pattern in  $0.555 \times 0.593 \times 135 = 44.44$  cases. Analogously, the —pattern is expected in  $(1 - 0.555) \times (1 - 0.593) \times 135 = 24.45$  cases. The expected overall agreement percent by chance  $I_e$  is  $(44.44 + 24.45)/135 = 0.51$ . The observed overall percent agreement is  $I_o = (60 + 40)/135 = 0.74$ . Thus we have

$$\text{Kappa} = (0.74 - 0.51)/(1 - 0.51) = 0.47$$

Generally, kappa values greater than 0.75 are taken to indicate excellent agreement beyond chance, values from 0.40 to 0.75 are regarded as showing fair to good agreement beyond chance, and values below 0.40 indicate poor agreement beyond chance. An SE for the kappa estimate can be computed.<sup>29</sup> Kappa is related to the intraclass correlation coefficient, which is a widely used measure of interrater reliability for quantitative measurements.<sup>29</sup> The considered agreement measures, percent agreement, and kappa can also be applied to assess the reproducibility of a qualitative test when the test is applied twice in a given context.

Various methodological problems are encountered in studies on qualitative tests. An obvious mistake is to let the result of the test being evaluated contribute to the diagnostic classification of subjects being tested (circular argument). This is also termed *incorporation bias*.<sup>30,31</sup> Another problem is partial as opposed to complete verification. When a new test is compared with an existing, imperfect test, a partial verification is sometimes undertaken, in which only discrepant results are subjected to further testing by a perfect test procedure. On this basis, sensitivity and specificity are reported for the new test. This procedure (called *discrepant resolution*) leads to biased estimates and should not be accepted.<sup>30–33</sup> The problem is that for cases with agreement, both the existing (imperfect) test and the new test may be wrong. Thus only a measure of agreement should be reported, not specificity and sensitivity values. In the biostatistical literature, various procedures have been suggested to correct for bias caused by imperfect reference tests, but unrealistic assumptions concerning the independence of test results are usually put forward.

## ASSAY COMPARISON

Comparison of measurements by two assays is a frequent task in the laboratory. Preferably, parallel measurements of a set of patient samples should be undertaken. To prevent artificial matrix-induced differences, fresh patient samples are the optimal material. A nearly even distribution of values over the analytical measurement range is also preferable. In an ordinary laboratory, comparison of two routine assays is the most frequently occurring situation. Less commonly, comparison of a routine assay with a reference measurement procedure is undertaken. When two routine assays are compared, the focus is on observed differences. In this situation, it is not possible to establish that one set of measurements is the correct one and thereby know by how much measurements deviate from the presumed correct concentrations. Rather, the question is whether the new assay can replace the existing one without a systematic change in result values. To address this question, the dispersion of observed differences between paired measurements may be evaluated by these assays. To carry out a formal, objective analysis of the data, a statistical procedure with graphics display should be applied. Various approaches may be used: (1) a frequency plot or histogram of the distribution of differences (DoD) with measures of central tendency and dispersion (DoD plot), (2) a difference (bias) plot, which shows differences as a function of the average concentration of measurements (Bland-Altman plot), or (3) a regression analysis. In the following, a general error model is presented, and some typical measurement relationships are considered. Each of the statistical approaches mentioned is presented in detail along with a discussion of their advantages and disadvantages.

### Basic Error Model

The occurrence of measurement errors is related to the performance characteristics of the assay. It is important to distinguish between pure, random measurement errors, which are present in all measurement procedures, and errors related to incorrect calibration and nonspecificity of the assay. Whereas a reference measurement procedure is associated only with pure, random error, a routine method, additionally, is likely to have some bias related to errors in calibration and limitations with regard to specificity. Whereas an erroneous calibration function gives rise to a systematic error, nonspecificity gives an error that typically varies from sample to sample. The error related to nonspecificity thus has a random character, but in contrast to the pure measurement error, it cannot be reduced by repeated measurements of a sample. Although errors related to nonspecificity for a group of samples look like random errors, for the individual sample, this type of error is a bias. Because this bias varies from sample to sample, it has been called a *sample-related random bias*.<sup>34-36</sup> In the following section, the various error components are incorporated into a formal error model.

### Measured Value, Target Value, Modified Target Value, and True Value

Upon taking into account that an analytical method measures analyte concentrations with some random measurement error, one has to distinguish between the actual, measured value and the average result we would obtain if the given

sample was measured an infinite number of times. If the assay is a reference assay without bias and nonspecificity, we have the following, simple relationship:

$$x_i = X_{\text{True}i} + \varepsilon_i$$

where  $x_i$  represents the measured value,  $X_{\text{True}i}$  is the average value for an infinite number of measurements, and  $\varepsilon_i$  is the deviation of the measured value from the average value. If we were to undertake repeated measurements, the average of  $\varepsilon_i$  would be zero and the SD would equal the analytical SD ( $\sigma_A$ ) of the reference measurement procedure. Pure, random, measurement error will usually be Gaussian distributed.

In the case of a routine assay, the relationship between the measured value for a sample and the true value becomes more complicated:

$$x_i = X_{\text{True}i} + \text{Cal-Bias} + \text{Random-Bias}_i + \varepsilon_i$$

The *cal-bias* term (calibration bias) is a systematic error related to the calibration of the method. This systematic error may be a constant for all measurements corresponding to an offset error, or it may be a function of the analyte concentration (e.g., corresponding to a slope deviation in the case of a linear calibration function). The *random-bias<sub>i</sub>* term is a bias that is specific for a given sample related to nonspecificity of the method. It may arise because of code-determination of substances that vary in concentration from sample to sample. For example, a chromogenic creatinine method codetermines some other components with creatinine in serum.<sup>37</sup> Finally, we have the random measurement error term  $\varepsilon_i$ .

If we performed an infinite number of measurements of a specific sample by the routine method, the random measurement error term  $\varepsilon_i$  would be zero. The cal-bias and the random-bias<sub>i</sub>, however, would be unchanged. Thus the average value of an infinite number of measurements would equal the sum of the true value and these bias terms. This average value may be regarded as the target value ( $X_{\text{Target}i}$ ) of the given sample for the routine method. We have:

$$X_{\text{Target}i} = X_{\text{True}i} + \text{Cal-Bias} + \text{Random-Bias}_i$$

As mentioned, the calibration bias represents a systematic error component in relation to the true values measured by a reference measurement procedure. In the context of regression analysis, this systematic error corresponds to the intercept and the slope deviation from unity when a routine method is compared with a reference measurement procedure (outlined in detail later). It is convenient to introduce a modified target value expression ( $X'_{\text{Target}i}$ ) for the routine method to delineate this systematic calibration bias, so that:

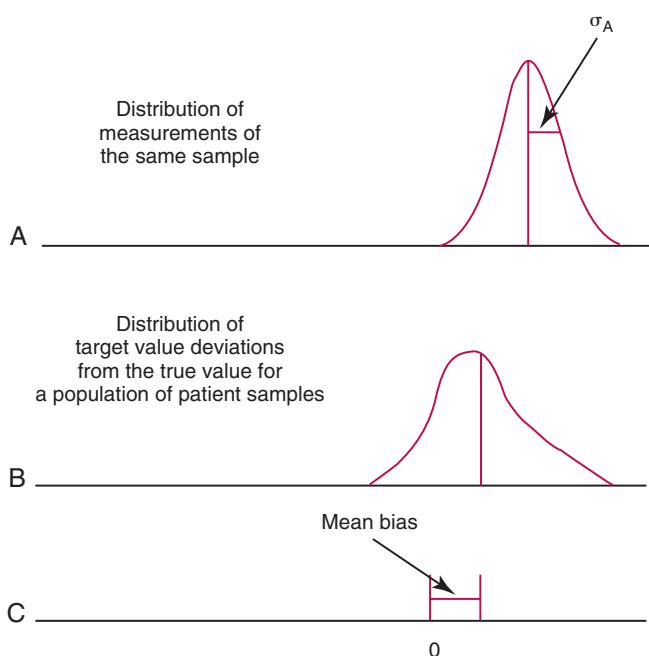
$$X'_{\text{Target}i} = X_{\text{True}i} + \text{Cal-Bias}$$

Thus for a set of samples measured by a routine method, the  $X_{\text{Target}i}$  values are distributed around the respective  $X'_{\text{Target}i}$  values with an SD, which is called  $\sigma_{\text{RB}}$ .

If the assay is a reference method without bias and nonspecificity, the target value and the modified target value equal the true value, that is,

$$X_{\text{Target}i} = X'_{\text{Target}i} = X_{\text{True}i}$$

The error model is outlined in Fig. 2.10.



**FIGURE 2.10** Outline of basic error model for measurements by a routine assay. **A**, The distribution of repeated measurements of the same sample, representing a normal distribution around the target value ( $X_{\text{Target}_i}$ ) (vertical line) of the sample with a dispersion corresponding to the analytical standard deviation,  $\sigma_A$ . **B**, Schematic outline of the dispersion of target value deviations from the respective true values for a population of patient samples. A distribution of an arbitrary form is displayed. The standard deviation equals  $\sigma_{RB}$ . The vertical line indicates the mean of the distribution. **C**, The distance from zero to the mean of the target value deviations from the true values represents the calibration bias (mean bias = cal-bias) of the assay.

### Calibration Bias and Random Bias

For an individual measurement, the total error is the deviation of  $x_i$  from the true value, that is,

$$\text{Total error of } x_i = \text{Cal-Bias} + \text{Random-Bias}_i + \varepsilon_i$$

Estimation of the bias terms requires parallel measurements between the method in question and a reference method as outlined in detail later. With regard to calibration bias, one should be aware of the possibility of lot-to-lot variation in analytical kit sets. The manufacturer should provide documentation on this lot-to-lot variation because often it is not possible for the individual laboratory to investigate a sufficient number of lots to assess this variation. Lot-to-lot variation shows up as a calibration bias that changes from lot to lot.

The previous exposition defines the total error in somewhat broader terms than is often seen. A traditional total error expression is<sup>38</sup>:

$$\text{Total error} = \text{Bias} + 2 \text{ SD}_A$$

which often is interpreted as the calibration bias plus  $2 \text{ SD}_A$ . If a one-sided statistical perspective is taken, the expression is modified to  $\text{Bias} + 1.65 \text{ SD}_A$ , indicating that 5% of results are located outside the limit. If a lower percentage is desired, the multiplication factor is increased accordingly, supposing a normal distribution. Interpreting the bias as identical with the calibration bias may lead to an underestimation of the total error.

Random bias related to sample-specific interferences may take several forms. It may be a regularly occurring additional random error component, perhaps of the same order of magnitude as the analytical error. In this context, it is natural to quantify the error in the form of an SD or CV. The most straightforward procedure is to carry out a method comparison study based on a set of patient samples in which one of the methods is a reference method, as outlined later. Krouwer<sup>34</sup> formally quantified sample-related random interferences in a comparison experiment of two cholesterol methods and found that the CV of the sample-related random interference component exceeded the analytical CV. Another form of sample-related random interference is more rarely occurring gross errors, which typically are seen in the context of immunoassays and are related to unexpected antibody interactions. Such an error usually shows up as an outlier in method comparison studies. A well-known source is the occurrence of heterophilic antibodies. Outliers should not just be discarded from the data analysis procedure. Rather, outliers must be investigated to identify their cause, which may be an important limitation in using a given assay. Supplementary studies may help clarify such random sample-related interferences and may provide specifications for the assay that limit its application in certain contexts (e.g., with regard to samples from certain patient categories).

### Assay Comparison Data Model

Here we consider the error model described earlier in relation to the method comparison situation. For a given sample measured by two analytical methods, 1 and 2, we have

$$\begin{aligned} x_{1i} &= X_{1\text{Target}_i} + \varepsilon_{1i} = X_{\text{True}_i} + \text{Cal-Bias}_1 + \text{Random-Bias}_{1i} + \varepsilon_{1i} \\ x_{2i} &= X_{2\text{Target}_i} + \varepsilon_{2i} = X_{\text{True}_i} + \text{Cal-Bias}_2 + \text{Random-Bias}_{2i} + \varepsilon_{2i} \end{aligned}$$

From this general model, we may study some typical situations. First, comparison of a routine assay with a reference measurement procedure will be treated. Second, comparison of two routine assays is considered.

### Comparison of a Routine Assay With a Reference Measurement Procedure

Assuming that method 1 is a reference method, the bias components disappear by definition, and we have the following situation:

$$\begin{aligned} x_{1i} &= X_{1\text{Target}_i} + \varepsilon_{1i} = X_{\text{True}_i} + \varepsilon_{1i} \\ x_{2i} &= X_{2\text{Target}_i} + \varepsilon_{2i} = X_{\text{True}_i} + \text{Cal-Bias}_2 + \text{Random-Bias}_{2i} + \varepsilon_{2i} \end{aligned}$$

The paired differences become

$$(x_{2i} - x_{1i}) = \text{Cal-Bias}_2 + \text{Random-Bias}_{2i} + (\varepsilon_{2i} - \varepsilon_{1i})$$

We thus have an expression consisting of a systematic error term (calibration bias of method 2) and two random terms. The random-bias<sub>2</sub> term is distributed around cal-bias<sub>2</sub> according to an undefined distribution.  $(\varepsilon_{2i} - \varepsilon_{1i})$  is a difference between two random measurement errors that are independent and, commonly, Gaussian distributed. However, we remind readers that the SD for analytical methods often depends on the concentration, as mentioned earlier. For analytes with a wide analytical measurement range (e.g., some hormones), both sample-related random interferences and analytical SDs are likely to depend on the measurement

concentration, often in a roughly proportional manner. It may then be more useful to evaluate the *relative* differences— $(x_{2i} - x_{1i})/[(x_{2i} + x_{1i})/2]$ —and accordingly express mean and random bias and analytical error as proportions. An alternative is to partition the total analytical measurement range into segments (e.g., three parts) and consider calibration bias, random bias, and analytical error separately for each of these segments. The segments may be divided preferably in relation to important decision concentrations (e.g., in relation to reference interval limits, treatment decision concentrations, or both).

### Comparison of Two Routine Assays

In the comparison of two routine methods, the paired differences become

$$(x_{2i} - x_{1i}) = (\text{Cal-Bias2} - \text{Cal-Bias1}) + (\text{Random-Bias2}_i - \text{Random-Bias1}_i) + (\varepsilon_{2i} - \varepsilon_{1i})$$

The expression again consists of a constant term, the difference between the two calibration biases, and two random terms. The first random term is a difference between two random-bias components that may or may not be independent. If the two field methods are based on the same measurement principle, the random bias terms are likely to be correlated. For example, two chromogenic methods for creatinine are likely to be subject to interference from the same chromogenic compounds present in a given serum sample. On the other hand, a chromogenic method and an enzymatic creatinine method are subject to different types of interfering compounds, and the random bias terms may be relatively independent. In the  $\varepsilon_{2i} - \varepsilon_{1i}$  term, the same relationships as described previously are likely to apply. One may note that the general form of the expressed differences is the same in the two situations. Thus the same general statistical principles actually apply. In the following sections, we will consider the DoD under various circumstances, as well as the measurement relations between methods 1 and 2 on the basis of regression analysis.

### Preliminary Practical Work in Relation to a Method Comparison Study

When a method comparison study is to be conducted, the analytical methods to be examined first should be established in the laboratory according to written protocols and should be stable in routine performance. Reagents are commonly supplied as ready-made analytical kits, perhaps implemented on a dedicated analytical instrument (open or closed system). Technologists performing the study should be trained in the procedures and associated instrumentation. Furthermore, it is important that a QC system is in place to ensure that the methods being compared are running in an in-control state.

### Planning a Method Comparison Study

In the planning phase of a method comparison study, several points require attention, including the (1) number of samples necessary, (2) distribution of analyte concentrations (preferably uniform over the analytical measurement range), and (3) representativeness of the samples. To address the latter point, samples from relevant patient categories should be included, so that possible interference phenomena can be discovered. For example, it may, in a given context, be relevant to include samples from patients with diabetes to

exclude the possibility that aberrations in glucose metabolism may influence test results. Practical aspects related to storage and treatment of samples (e.g., container) and possible artifacts induced by storage (e.g., freezing of samples) and addition of anticoagulants should be considered. Comparison of measurements should preferably be undertaken over several days (e.g., at least 5 days) so that the comparison of methods does not become dependent on the performance of the methods in one particular analytical run. Finally, ethical aspects (e.g., informed consent from patients whose samples will be used) should be considered in relation to existing legislation.

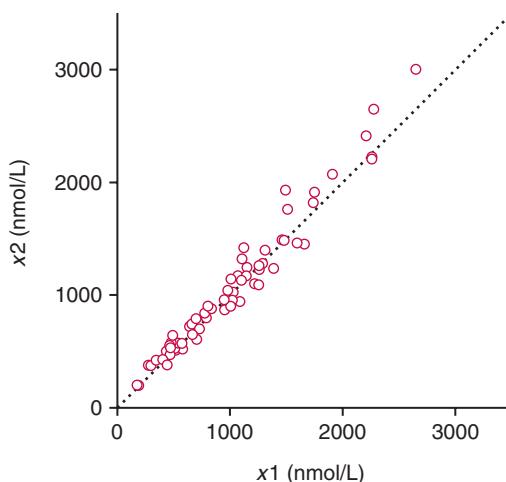
When the comparison protocol is considered, various guidelines may be consulted. The CLSI EP guidelines give advice on various aspects. For example, the CLSI guideline EP-09-A3, "Method Comparison and Bias Estimation Using Patient Samples," suggests measurement of 40 samples in duplicate by each method when a new method is introduced in the laboratory as a substitute for an established one.<sup>39</sup> In addition, it is proposed that a vendor of an analytical test system should have made a comparison study based on at least 100 samples measured in duplicate by each method. The principle of a more demanding requirement for vendors appears reasonable. This initial validation should be comprehensive to disclose the performance of the assay system in detail. Then the requirement for the ordinary user may be more modest. The EP15 guideline "User Verification of Manufacturer's Claims" suggests a more condensed approach based on a bias or difference plot, which does not involve regression analysis and can be carried out using 20 samples.<sup>17</sup> Although these general guidelines on sample size are useful, additional aspects are important. The probability of detecting rarely occurring interferences showing up as outliers should be taken into account when the necessary sample size is considered. Finally, in relation to evaluation of automated methods, special consideration should be given to the sample sequence to evaluate drift, carryover, and nonlinearity (e.g., by a multifactorial design).<sup>14</sup>

### Distribution of Differences Plot

From the end-user viewpoint, it is the differences per se that matter. Thus with regard to the outcome of replacing an established routine method with a new one that perhaps is less expensive or more practical, it is important to focus on the DoD between paired measurements by the old and the new method. A graphic display with assessment of the central tendency and dispersion of differences in the form of an ordinary histogram or frequency polygon plot is useful. The differences may or may not be Gaussian distributed. Because both analytical error components and sample-related random interferences may contribute to the differences, the distribution may be irregular, and outliers may occur. Furthermore, the random dispersion elements may be dependent on analyte concentration. This is also termed the *heteroscedasticity* of the measurement. Therefore a nonparametric approach for interpreting the DoD may generally be preferable as a starting point.

### Nonparametric Approach

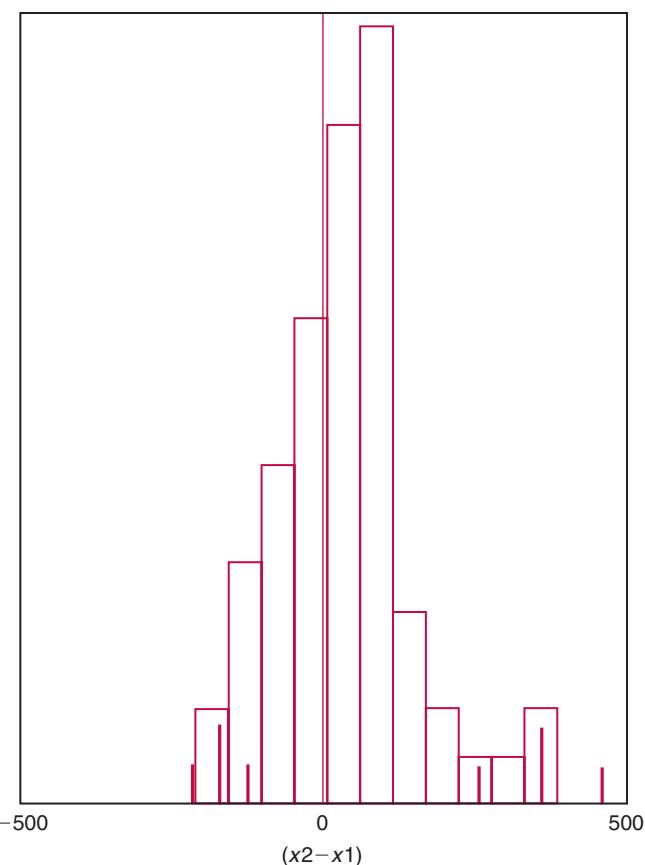
Both the central tendency (median) and extreme percentiles are of interest when the nonparametric approach to the DoD is used. With a traditional 95% level, the 2.5 and 97.5 percentiles



**FIGURE 2.11** A scatter plot of  $n = 65$  ( $x_1$ ,  $x_2$ ) data points for comparison of two drug assays. The *dashed line* is the line of identity.

are considered. A 99% or higher extreme level may be selected, and the related percentiles (0.5 and 99.5 percentiles, or more extreme ones) may then be applied for a description of method differences. Nonparametric estimation of the 2.5 and 97.5 percentiles requires 2.5 times as many observations as the parametric approach to obtain the same uncertainty, which implies that sample sizes cannot be too small.<sup>40</sup> Estimating confidence limits of the percentiles can give an indication of their imprecision. The CIs can be estimated from the ordered observations as described in Chapter 9 in the section on nonparametric estimation of the 95% reference interval. Alternatively, a bootstrap procedure can be applied as described.<sup>40</sup> The advantage of the bootstrap procedure is that SEs can be derived using smaller sample sizes than are used with the simple nonparametric approach.<sup>41</sup>

A method comparison example from the laboratory of one of the authors (K.L.) is considered. Two drug assays developed in house for serum concentrations of the antipsychotic drug clozapine are compared. The established assay (method 1) is an HPLC method based on manual liquid-liquid extraction. The new method (method 2) is an HPLC method with an automated on-column extraction step. An initial impression of the relation between  $x_1$  and  $x_2$  measurements can be obtained from a scatter plot of the 65 measurement sets ( $x_1$ ,  $x_2$ ) with the identity line outlined (Fig. 2.11). The  $x_1$  measurements range from 177 to 2650 nmol/L, and the range of  $x_2$  values is from 200 to 3004 nmol/L (i.e., we have a relatively wide analytical measurement range in the present example). A histogram of the difference ( $x_2 - x_1$ ) is shown in Fig. 2.12. Applying a nonparametric data description, we order the observed differences according to size and derive the median difference as the value of the  $(0.5N + 0.5)$ th ordered observation, here 26 nmol/L. In case the order is a noninteger, interpolation between neighbor-ordered values is carried out. A paired nonparametric test, the Wilcoxon test,<sup>5</sup> shows that the median difference was significantly different from zero ( $P < .02$ ). The 2.5 and 97.5 percentiles correspond to the values of the  $(0.025N + 0.5)$ th and  $(0.975N + 0.5)$ th ordered observations, respectively, as displayed in Table 2.5.<sup>4,40</sup> For a sample size smaller than 120, it is not possible to derive CIs for the percentiles by the simple nonparametric procedure.



**FIGURE 2.12** Distribution of differences plot for comparison of two drug assays: nonparametric analysis. A histogram shows the relative frequency of  $n = 65$  differences with demarcated 2.5 and 97.5 percentiles determined nonparametrically. The 90% confidence intervals of the percentiles are shown. These were derived by the bootstrap technique.

Therefore we also applied the bootstrap procedure to estimate nonparametric percentiles with 90% CIs (see Table 2.5).<sup>40,42</sup> The bootstrap procedure, which is based on computerized random resampling of the observations, provides slightly different percentile estimates, as shown in Table 2.5. In this way, we obtain an estimation of the size of negative and positive differences with uncertainties. The present example shows a considerable range of differences, with the 2.5 percentile being  $-169$  nmol/L (90% CI:  $-214$  to  $-123$ ) and the 97.5 percentile being  $356$  nmol/L (90% CI:  $255$  to  $457$ ). These relatively large differences should be related to the considerable analytical measurement range for the analyte, and an evaluation of *relative* differences may be more relevant for the present example (see later in this chapter).

In the presented examples, no evident outliers were present. However, outliers deserve special attention.<sup>4</sup> Unless they are related to obvious method or apparatus malfunction, discarding of outliers should be considered with caution. Outliers may indicate the presence of large sample-related random interferences, which may be of major clinical importance (e.g., interference by antibodies or degradation products that occur only rarely). Thus a special investigation of outlying results with reanalysis and exploration of the reasons for the outlying observations should be considered.

**TABLE 2.5 Analysis of Distribution of Differences for the Comparison of Drug Assays Example<sup>a</sup>**

Total range of $x_1$ measurements	177 to 2650		
Total range of $x_2$ measurements	200 to 3004		
Total range of differences ( $x_2 - x_1$ )	-210.00 to 437.00		
Test for normality of differences (Anderson-Darling test)	$P < .01$		
Statistical Analysis of Differences	Simple Nonparametric	Bootstrap	Parametric
Median	26.00 ( $P < .02$ )		
Mean			42.00 ( $P < .01$ )
SD			124.42
Coefficient of skewness			+0.83
Coefficient of kurtosis			+1.27
Outlier test (4 SD)			NS
2.5-percentile	-166.00	-169.11	-201.86
97.5 percentile	372.38	355.90	285.86
90% CI for 2.5 percentile		-214.73 to -123.50	-245.24 to -158.47
90% CI for 97.5 percentile		255.03 to 456.77	242.47 to 329.24

<sup>a</sup> $n = 65$  single ( $x_1, x_2$  measurements). The units are nmol/L.

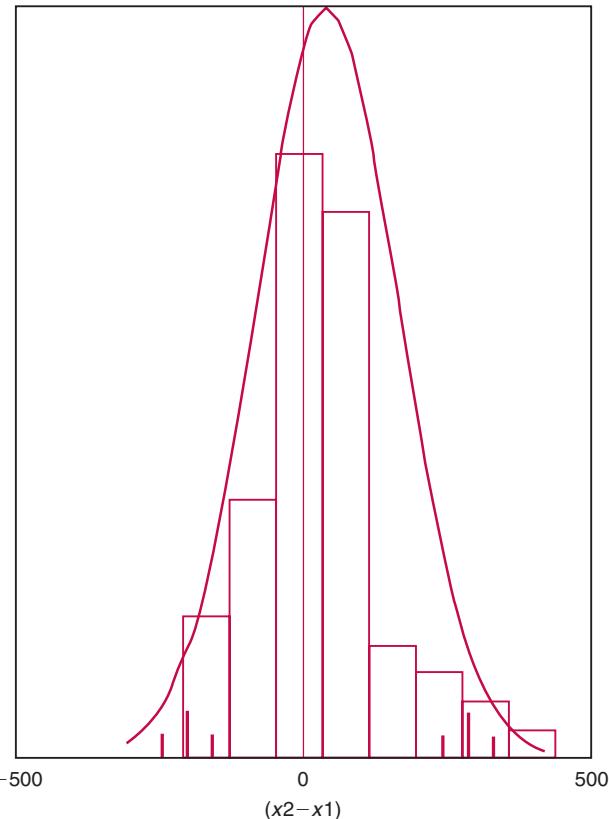
CI, Confidence interval; NS, not significant; SD, standard deviation.

### Parametric Approach

If application of a goodness-of-fit test does not disprove that the DoD is Gaussian distributed, a parametric statistical approach may be undertaken. In the example presented, a significant deviation from normality was present, as assessed by the Anderson-Darling test<sup>43</sup> ( $P < .01$ ); therefore a parametric analysis in principle should not be carried out. However, to demonstrate the procedure, the parametric approach is also carried out (Fig. 2.13 and Table 2.5). The mean and SD ( $SD_{Dif}$ ) of the paired differences ( $x_2 - x_1$ ) are estimated according to standard procedures. A paired  $t$ -test is used to determine whether the mean difference is significantly different from zero ( $P < .01$  in this case). The 2.5 and 97.5 percentiles for the differences are estimated as the mean  $\pm t_{0.025(N-1)} SD_{Dif}$ . An SE for the percentiles ( $SE_{perc}$ ) may be computed, and the 90% CI limits are then derived as  $\pm 1.65 SE_{perc}$  around the percentiles (see Fig. 2.13 and Table 2.5). The parametrically derived 2.5 and 97.5 percentiles (-202 and 286 nmol/L) differ somewhat from the nonparametrically derived percentiles, which in the present context with proven non-normality may be regarded as the most reliable estimates.

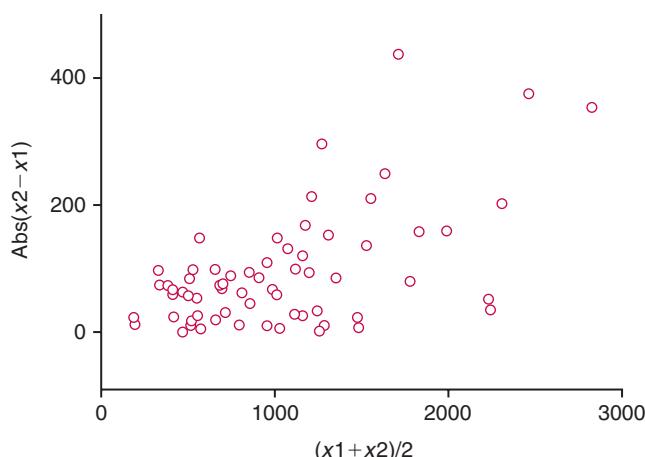
### Relative Distribution of Differences Plot

In some cases in which a wide analytical measurement range (i.e., corresponding to 1 or several decades) is used, the random error components depend on the concentration, as previously mentioned. Analytical SDs may be approximately proportional to the concentration over the major part of the analytical measurement range. In the present example, the initial scatter plot of ( $x_1, x_2$ ) values suggests that the random error of the differences increases with the concentration (see Fig. 2.11). A formal test for this possible relation is to compute the correlation coefficient between the average concentration and the absolute value of the differences. This correlation coefficient,  $r$ , is +0.57, which is significantly different from zero ( $P < .001$ ), and it confirms the relationship of scatter increasing with concentration, which also can be visualized in a scatter plot of the absolute differences against the average concentration (Fig. 2.14). A natural next step is to

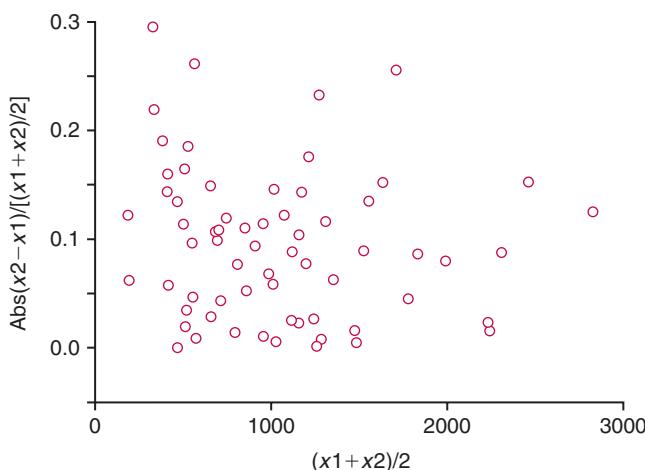


**FIGURE 2.13** Distribution of differences plot for comparison of two drug assays: parametric analysis. A histogram shows the relative frequency of  $n = 65$  differences with the estimated Gaussian density distribution. Parametrically estimated 2.5 and 97.5 percentiles are shown with 90% confidence intervals.

assess the *relative* differences in relation to the average concentration. The correlation coefficient between the absolute values of the relative differences [ $|x_2 - x_1|/(x_1 + x_2)/2$ ] and the average concentration [ $(x_1 + x_2)/2$ ] was not significantly different ( $P > .05$ ) from zero ( $r = -0.15$ ); a scatter plot also



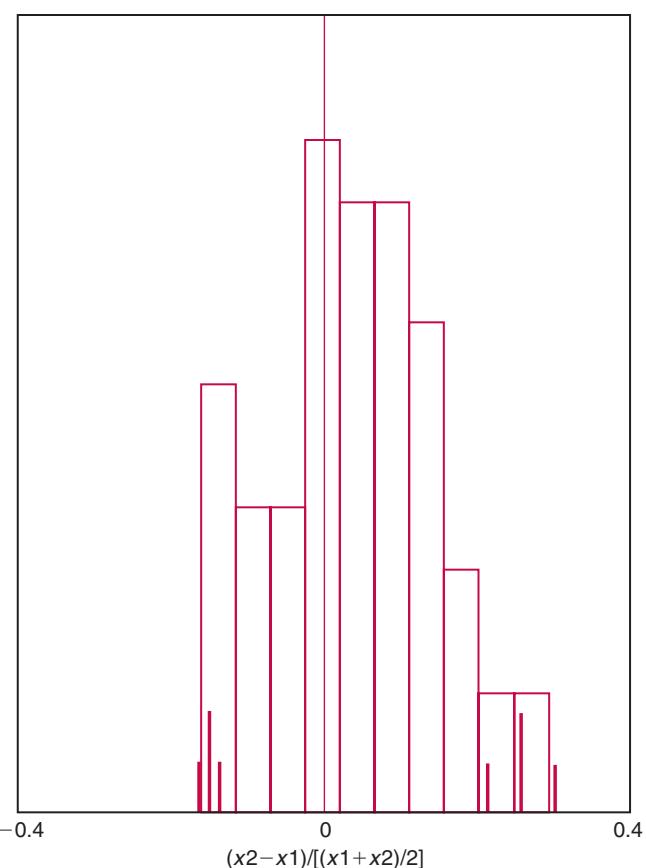
**FIGURE 2.14** Plot of absolute differences (ordinate) against average concentration (abscissa) for the comparison of drug assays example. The scatter increases with the average concentration ( $r = +0.57$ ).



**FIGURE 2.15** Plot of absolute relative differences (ordinate) against average concentration (abscissa) for the comparison of drug assays example. The scatter is not significantly correlated with the average concentration ( $r = -0.15$ , not significant).

suggests a more homogeneous dispersion (Fig. 2.15). In this situation, it is more reasonable to deal with *relative* differences or percentage differences [ $\{(x_2 - x_1)/(x_1 + x_2)/2\} \times 100\%$ ]. The same nonparametric descriptive measures as used earlier may be applied for the central tendency and the dispersion (Fig. 2.16). The median relative difference amounts to 0.042, or 4.2%, which is significantly higher than zero ( $P < .01$ ; Wilcoxon test; Table 2.6). The 2.5 and 97.5 percentiles are  $-0.15$  and  $0.26$ , respectively. The 90% CIs derived by the bootstrap procedure were  $-0.16$  to  $-0.14$  and  $0.21$  to  $0.30$ , respectively. Thus from this analysis, we may conclude that the 95% interval for percentage differences ranges from about  $-15$  to  $+26\%$ .

Finally, we may consider a parametric analysis of the relative differences (Fig. 2.17 and Table 2.6). A goodness-of-fit test (Anderson-Darling test;  $P > .05$ ) showed that the relative differences did not depart significantly from a normal distribution, which in this case supports the parametric approach (Fig. 2.18). The parametric 2.5 and 97.5 percentiles were  $-0.18$  and  $0.26$ , respectively. The mean was 0.042, and the SD



**FIGURE 2.16** Relative distribution of differences plot for comparison of two drug assays: nonparametric analysis. A histogram shows the relative frequency of relative differences with demarcated 2.5 and 97.5 percentiles determined nonparametrically. The 90% confidence intervals (bootstrap) of the percentiles are shown.

of the relative differences was 0.11. Thus we may conclude that there is an average bias of about 4%, which might rely on a calibration difference (an estimate of  $[cal\_bias2 - cal\_bias1]$ ), and a random error corresponding to a CV of 11%. The random error CV of 11% is an estimate of the combined dispersion of  $[(Random\_Bias2_i - Random\_Bias1_i) + (\epsilon_2 - \epsilon_1)]$ . If we ascribe the random variation equally to the two assays, it corresponds to a random error level of  $11\%/\sqrt{2} = 7.8\%$  for each assay. In the present example, the average bias of 4% and the estimated random variation of differences between the two assays were considered acceptable in relation to the clinical use of the assay, and it was decided to replace the manual assay with the new, automated assay.

### Verification of Distribution of Differences in Relation to Specified Limits

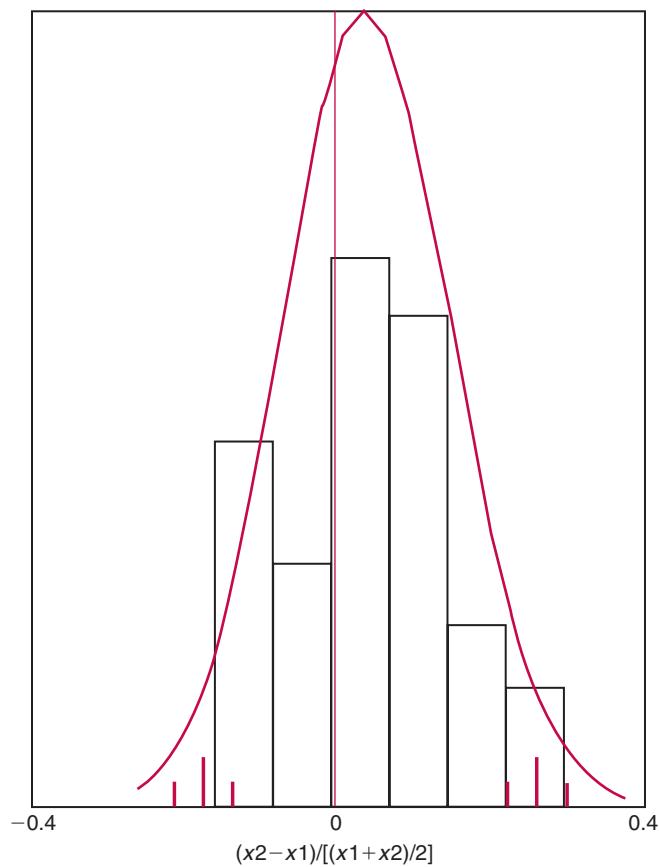
In situations in which a field method is being considered for implementation, it may be desired primarily to *verify* whether the differences in relation to the existing method are located within given specified limits rather than *estimating* the DoD. For example, one may set limits corresponding to  $\pm 15\%$  as clinically acceptable and may desire that a majority (e.g., 95% of differences) are located within this interval.

By counting, it may be determined whether the expected proportion of results is within the limits (i.e., 95%). One may accept percentages that do not deviate significantly from the

**TABLE 2.6 Analysis of Distribution of *Relative Differences* for the Comparison of Drug Assays Example;  $n = 65$**

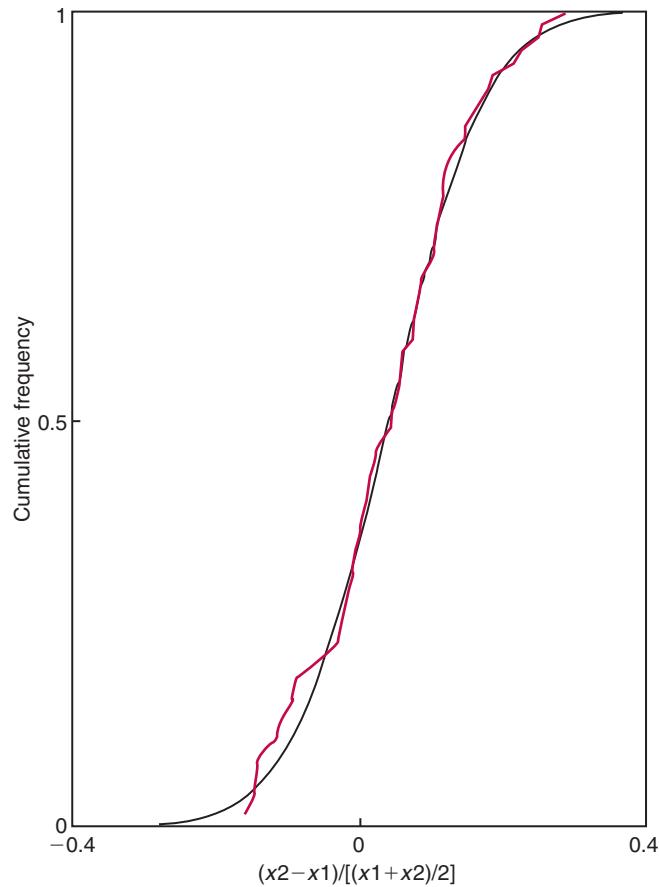
Statistical Analysis	Simple Nonparametric	Bootstrap	Parametric
Total range of relative differences	–0.1598 to 0.2953		
Test for normality (Anderson-Darling test)	NS		
Median	0.0467 ( $P < .01$ )		
Mean			0.0418 ( $P < .01$ )
SD			0.1109
Coefficient of skewness			+0.05
Coefficient of kurtosis			–0.60
Outlier test			NS
2.5 percentile	–0.1487	–0.1492	–0.1754
97.5 percentile	0.2607	0.2570	0.2591
90% CI for 2.5 percentile		–0.1627 to –0.1357	–0.2141 to –0.1368
90% CI for 97.5 percentile		0.2135–0.3005	0.2204 to 0.2978

CI, Confidence interval; NS, not significant; SD, standard deviation.



**FIGURE 2.17** Distribution of *relative differences* plot for comparison of two drug assays: parametric analysis. A histogram shows the relative frequency of relative differences with the estimated Gaussian density distribution. Parametrically estimated 2.5 and 97.5 percentiles are shown with 90% confidence intervals.

supposed percentage at the given sample size derived from the binomial distribution (Table 2.7). For example, if 50 paired measurements have been performed in a method comparison study, and if it is observed that 46 of these results (92%) are within specified limits (e.g.,  $\pm 15\%$ ), the study supports that the achieved goal has been reached because the



**FIGURE 2.18** Cumulative frequency distribution of relative differences for the comparison of drug assays example. The black curve indicates the Gaussian cumulative frequency distribution curve. In accordance with the test for normality, good agreement is observed.

lower boundary for acceptance is 90%. It is clear that a reasonable number of observations should be obtained for the assessment to have acceptable power. If very few observations are available, the risk is high of falsely concluding that at least 95% of the observations are within specified limits in case it is not true (i.e., committing a type II error).

**TABLE 2.7 Lower Bounds (One-Sided 95% Confidence Interval) of Observed Proportions (%) of Results Being Located Within Specified Limits for Paired Differences That Are in Accordance With the Hypothesis of at Least 95% of Differences Being Within the Limits**

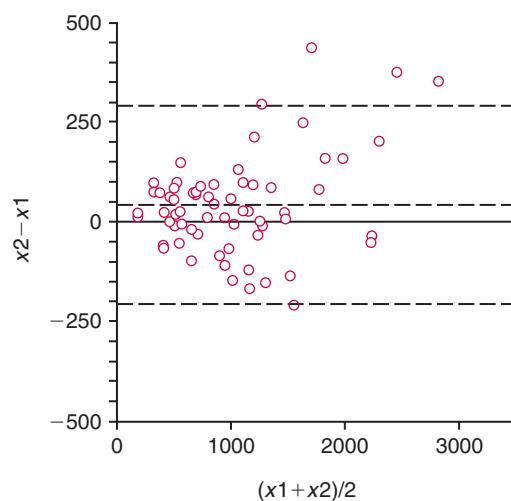
<i>n</i>	Observed Proportions
20	85
30	87
40	90
50	90
60	90
70	90
80	91
90	91
100	91
150	92
200	93
250	93
300	93
400	93
500	93
1000	94

### Difference (Bland-Altman) Plot

The difference plot suggested by Bland and Altman is widely used for evaluating method comparison data.<sup>44,45</sup> The procedure was originally introduced for comparison of measurements in clinical medicine, but it has also been adopted in clinical chemistry.<sup>46–48</sup> The Bland-Altman plot is usually understood as a plot of the differences against the average results of the methods. Thus the difference plot in this version provides information on the relation between differences and concentration, which is useful in evaluating whether problems exist at certain ranges (e.g., in the high range) caused by nonlinearity of one of the methods. It may also be of interest to observe whether differences tend to increase proportionally with the concentration or whether they are independent of concentration. In some situations, particular interest may be directed toward the low-concentration region. Information on the relation between differences and concentration is useful in the context of how to adjust for an irregularity (e.g., by changing the method to correct for nonlinearity, by restricting the analytical measurement range).

The basic version of the difference plot requires plotting of the differences against the average of the measurements. Fig. 2.19 shows the plot for the drug assay comparison data. The interval  $\pm 2$  SD of the differences is often delineated around the mean difference (i.e., corresponding to the mean and the 2.5 and 97.5 percentiles considered in the parametric distribution of differences plot [DoD plot]<sup>45</sup>). To assess whether the bias is significantly different from zero, the SE of the mean difference is estimated as the SD divided by the square root of the number of paired measurements ( $SE = SD/N^{0.5}$ ) and tested against zero by a *t*-test ( $t = [\text{Mean} - 0]/SE$ ).

Nonparametric limits may also be considered. The distribution of the differences as measured on the *y*-axis of the

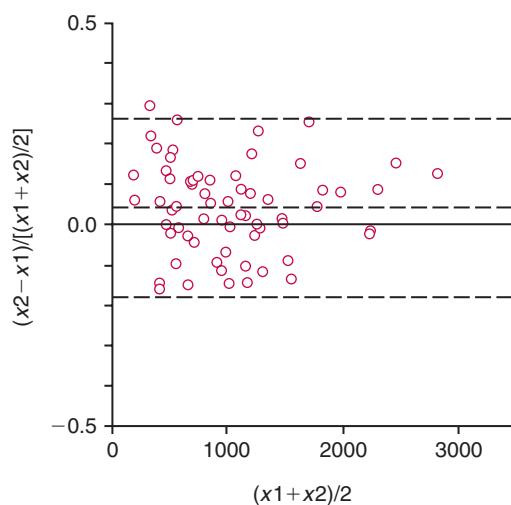


**FIGURE 2.19** Bland-Altman plot of differences for the drug comparison example. The differences are plotted against the average concentration. The mean difference (42 nmol/L) with  $\pm 2$  standard deviation of differences is shown (dashed lines).

coordinate system corresponds to the relations outlined for the DoD plot, which represents a projection of the differences on the *y*-axis. A constant bias over the analytical measurement range changes the average concentration away from zero. The presence of sample-related random interferences increases the width of the distribution. If the calibration bias depends on the concentration, if the dispersion varies with the concentration, or if both occur, the relations become more complex, and the interval mean  $\pm 2$  SD of the differences may not fit very well as a 95% interval throughout the analytical measurement range.

The displayed Bland-Altman plot for the drug assay comparison data (see Fig. 2.19) shows a tendency toward increasing scatter with increasing concentration, which is a reflection of increasing random error with concentration, as considered in detail in previous paragraphs. Thus a plot of the relative differences against the average concentration is of relevance (Fig. 2.20). This plot has a more homogeneous dispersion of values, agreeing with the estimated limits for the dispersion, that is, the relative mean difference  $\pm t_{0.025(N-1)} SD_{RelDif}$  equal to  $0.042 \pm 1.998 \times 0.11$  corresponding to  $-0.18$  and  $0.26$ , analogous to the situation with the relative DoD plot considered earlier.

Use of *relative* differences in situations with a proportional random error relationship prevents very large differences in the high-concentration range from dominating the analysis and making a balanced interpretation difficult. In the low range, the proportional relationship may not necessarily hold true, and sometimes the relative difference plot overcompensates for lack of proportionality in this region. It is then possible to truncate the proportional relationship at some lower limit and assume a constant SD for differences below this limit<sup>49</sup> (i.e., corresponding to the relationship in Fig. 2.3C). In the actual drug example (see Fig. 2.20) with a slightly negative correlation coefficient between relative differences and average concentration, a tendency toward this pattern is seen. An alternative to the relative difference plot is to plot the logarithm of the differences against the average concentration, but this type of plot is more difficult to interpret, because the scale is changed.

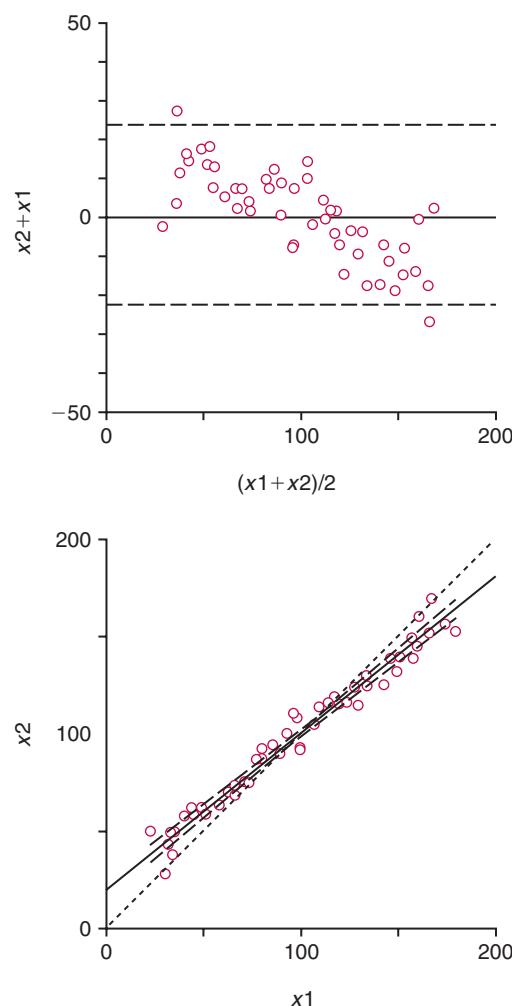


**FIGURE 2.20** Bland-Altman plot of relative differences for the drug comparison example. The differences are plotted against the average concentration. The mean relative difference (0.042) with  $\pm 2$  standard deviation of relative differences is shown (dashed lines).

Although it is customary to display the *estimated* limits for the differences (often, mean  $\pm 2$  SD<sub>dif</sub>), one may, as an alternative, display specification limits considered reasonable, as mentioned for the DoD plot.<sup>47</sup> It may then be assessed whether the observed differences conform to these limits, as discussed earlier (see Table 2.7). Application of the difference plot in various specific contexts has been considered.<sup>50,51</sup> It has also been suggested to estimate a regression line for the differences as a function of the average measurement concentration.<sup>52</sup>

### A Caution Against Incorrect Interpretation of Paired t-Tests in Method Comparison Studies

In association with the difference plot, the paired *t*-test is usually applied as described earlier,<sup>44</sup> but one should be careful with regard to the interpretation. For example, consider the case shown below, in which method 2 ( $x_2$ ) measurements tend to exceed method 1 ( $x_1$ ) measurements in the low range and vice versa at high concentrations (Fig. 2.21A). This corresponds to a positive calibration bias in the low range, changing to a negative calibration bias in the high range. In this situation, the overall averages of both sets of measurements are nearly equal, and the paired *t*-test yields a nonsignificant result because the average paired difference (i.e., the overall bias) is close to zero (Table 2.8). This does not mean that the measurements are equivalent. Subjecting the data to Deming regression analysis (see the next section) clearly discloses the relation (Fig. 2.21B).<sup>53</sup> Results of the regression analysis confirm the existence of both a systematic constant error (intercept different from zero) and a systematic proportional error (slope different from 1). Therefore as pointed out previously, the statistical significance revealed by the paired *t*-test cannot be used to indicate whether measurements are equivalent. The paired *t*-test is just a test for the average bias; it does not say anything about the equivalency of measurements throughout the analytical measurement range.



**FIGURE 2.21** Simulated example with positive and negative differences in the low and high ranges, respectively. A, Bland-Altman plot. B, x-y Plot with diagonal (dotted straight line) and estimated Deming regression line (solid line) with 95% confidence curves (dashed lines).

**TABLE 2.8 Comparison of Paired t-Test Results and Deming Regression Results for a Simulated Method Comparison Example With Positive Intercept ( $a_0 = 20$ ) and Slope Below Unity ( $b = 0.80$ ),  $n = 50$  ( $x_1, x_2$ ) Measurements**

	Paired t-Test	Regression Analysis (Deming)
Mean difference (SEM)	0.78 (1.63)	
$t = \text{mean difference}/\text{SEM}$	$0.78/1.63 = 0.48$ (NS)	
Slope ( $b$ ) [SE( $b$ )]		0.80 (0.027)
$t = (b - 1)/\text{SE}(b)$		$-7.4 (P < .001)$
Intercept ( $a_0$ ) [SE( $a_0$ )]		20.3 (2.82)
$t = (a_0 - 0)/\text{SE}(a_0)$		7.2 ( $P < .001$ )

NS, Not significant; SEM, standard error of the mean.

## Regression Analysis

Regression analysis is commonly applied in comparing the results of analytical method comparisons. Typically, an experiment is carried out in which a series of paired values is collected when a new method is compared with an established method. This series of paired observations ( $x_{1i}, x_{2i}$ ) is then used to establish the nature and strength of the relationship between the tests. This discussion outlines various regression models that may be used, gives criteria for when each should be used, and provides guidelines for interpreting the results.

Regression analysis has the advantage that it allows the relation between the target values for the two compared methods to be studied over the full analytical measurement range. If the systematic difference between target values (i.e., the calibration bias between the two methods, or the systematic error) is related to the analyte concentration, such a relationship may not be clearly shown when the previously mentioned types of difference plots are used. Although nonlinear regression analysis may be applied, the focus is usually on linear regression analysis. In linear regression analysis, it is assumed that the systematic difference between target values can be modeled as a constant systematic difference (intercept deviation from zero) combined with a proportional systematic difference (slope deviation from unity), usually related to a discrepancy with regard to calibration of the methods. In situations when random errors have a constant SD, unweighted regression procedures are used (e.g., Deming regression analysis). For cases with SDs that are proportional to the concentration, the weighted Deming regression procedure is preferred.

## Error Models in Regression Analysis

As outlined previously, we distinguish between the measured value ( $x_i$ ) and the target value ( $X_{\text{Target}i}$ ) of a sample subjected to analysis by a given method. In linear regression analysis, we assume a linear relationship between values devoid of random error of any kind.<sup>54,55</sup> Thus to operate with a linear relationship between values without random measurement error and sample-related random bias, we have to introduce modified target values:<sup>1</sup>

$$X1_{\text{Target}i} = X1'_{\text{Target}i} + \text{Random-Bias}_1$$

$$X2_{\text{Target}i} = X2'_{\text{Target}i} + \text{Random-Bias}_2$$

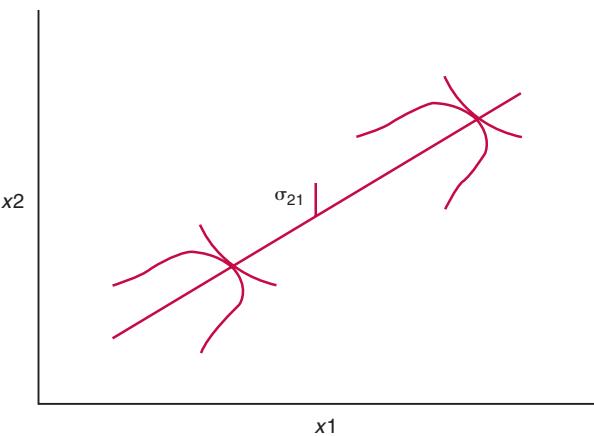
where we now assume a linear relationship between these modified target values:

$$X2'_{\text{Target}i} = \alpha_0 + \beta X1'_{\text{Target}i}$$

In this model,  $\alpha_0$  corresponds to a constant difference with regard to calibration, and  $(\beta - 1)$  is a proportional deviation. Thus the systematic error or calibration difference between the measurements corresponds to

$$X2'_{\text{Target}i} - X1'_{\text{Target}i} = \alpha_0 + (\beta - 1)X1'_{\text{Target}i}$$

Because of sample-related random interferences and measurement imprecision (of the type that can be described by a Gaussian distribution, e.g., caused by pipetting variability, signal variability), individually measured pairs of values ( $x_{1i}, x_{2i}$ ) will be scattered around the line expressing the relationship between  $X1'_{\text{Target}i}$  and  $X2'_{\text{Target}i}$ . Fig. 2.22 outlines schematically



**FIGURE 2.22** Outline of the relation between  $x_1$  and  $x_2$  values measured by two assays subject to random errors with constant standard deviations over the analytical measurement range. A linear relationship between the modified target values ( $X1'_{\text{Target}i}$ ,  $X2'_{\text{Target}i}$ ) is presumed. The  $x_{1i}$  and  $x_{2i}$  values are Gaussian distributed around  $X1'_{\text{Target}i}$  and  $X2'_{\text{Target}i}$ , respectively, as schematically shown.  $\sigma_{21}$  ( $\sigma_{xy}$ ) is demarcated.

how the random distribution of  $x_1$  and  $x_2$  values occurs around the regression line. We have

$$x_{1i} = X1'_{\text{Target}i} + \varepsilon_{1i} = X1'_{\text{Target}i} + \text{Random-Bias}_1 + \varepsilon_{1i}$$

$$x_{2i} = X2'_{\text{Target}i} + \varepsilon_{2i} = X2'_{\text{Target}i} + \text{Random-Bias}_2 + \varepsilon_{2i}$$

The random error components may be expressed as SDs, and generally we can assume that sample-related random bias (SD  $\sigma_{\text{RB}}$ ) and analytical imprecision (SD  $\sigma_A$ ) are independent for each analyte, yielding the relations

$$\sigma_{\text{ex1}}^2 = \sigma_{\text{RB1}}^2 = \sigma_{A1}^2$$

$$\sigma_{\text{ex2}}^2 = \sigma_{\text{RB2}}^2 = \sigma_{A2}^2$$

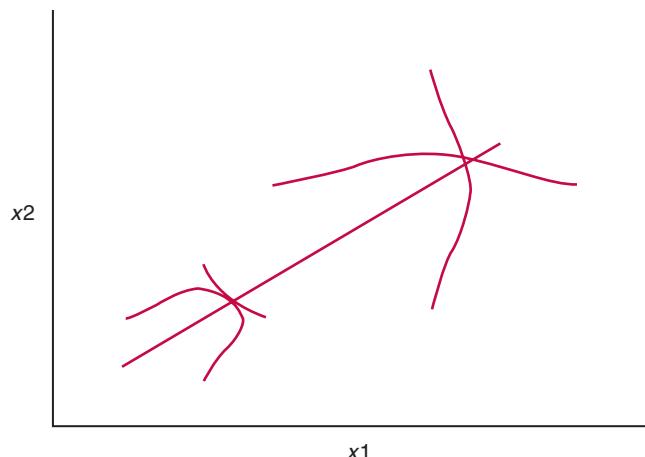
$\sigma_{\text{ex1}}$  and  $\sigma_{\text{ex2}}$  are the total SDs of the distributions of  $x_{1i}$  and  $x_{2i}$  around their respective modified target values,  $X1'_{\text{Target}i}$  and  $X2'_{\text{Target}i}$ . The sample-related random bias components for methods 1 and 2 may not necessarily be independent. They also may not be Gaussian distributed, contrary to the analytical components. Thus when a regression procedure is applied, the explicit assumptions to take into account should be considered. In situations without random bias components of any significance, the relationships simplify to

$$\sigma_{\text{ex1}}^2 = \sigma_{A1}^2$$

$$\sigma_{\text{ex2}}^2 = \sigma_{A2}^2$$

In this situation, it usually can be assumed that the error distributions are Gaussian, and estimates of the analytical SDs may be available from QC data.

Another methodologic problem concerns the question of whether the dispersion of sample-related random bias and the analytical imprecision are constant or change with the analyte concentration, as considered previously in the difference plot sections. In cases with a considerable range (i.e., a decade or longer), this phenomenon should also be taken into account when a regression analysis is applied. Fig. 2.23 schematically shows how dispersions may increase proportionally with concentration.



**FIGURE 2.23** Outline of the relation between  $x_1$  and  $x_2$  values measured by two assays subject to proportional random errors. A linear relationship between the modified target values is assumed. The  $x_{1,i}$  and  $x_{2,i}$  values are Gaussian distributed around  $X_{1,\text{Target},i}$  and  $X_{2,\text{Target},i}$ , respectively, with increasing scatter at higher concentrations, as is shown schematically.

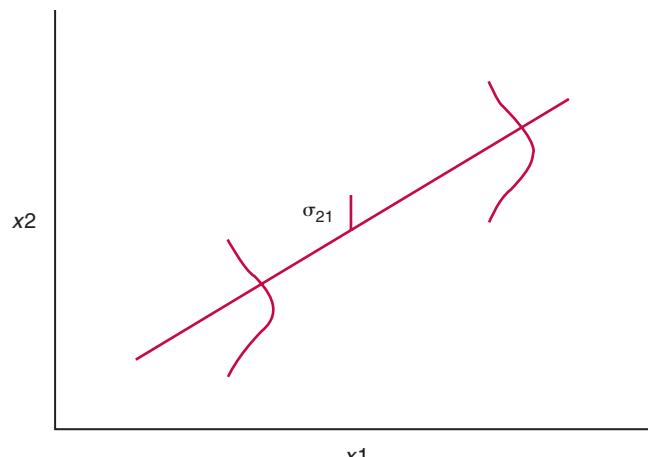
### Deming Regression Analysis and Ordinary Least-Squares Regression Analysis (Constant Standard Deviations)

To reliably estimate the relationship between modified target values (i.e.,  $a_0$  for  $\alpha_0$  and  $b$  for  $\beta$ ), a regression procedure taking into account errors in both  $x_1$  and  $x_2$  is preferable<sup>6</sup> (i.e., Deming approach; see Fig. 2.22). Although the ordinary least-squares (OLR) procedure is commonly used in method comparison studies, it does not take errors in  $x_1$  into account but is based on the assumption that only the  $x_2$  measurements are subject to random errors (Fig. 2.24). In the Deming procedure, the sum of squared distances from measured sets of values ( $x_{1,i}, x_{2,i}$ ) to the regression line is minimized at an angle determined by the ratio between SDs for the random variations of  $x_1$  and  $x_2$ . It can be proven theoretically that, given Gaussian error distributions, this estimation procedure is optimal. It should here be noted that it is the error distributions that should be Gaussian, not the dispersion of values over the measurement range. This is often misunderstood. In Fig. 2.25, the symmetric case is illustrated with a regression slope of 1 and equal SDs for the random variations of  $x_1$  and  $x_2$ , in which case the sum of squared distances is minimized orthogonally in relation to the line.

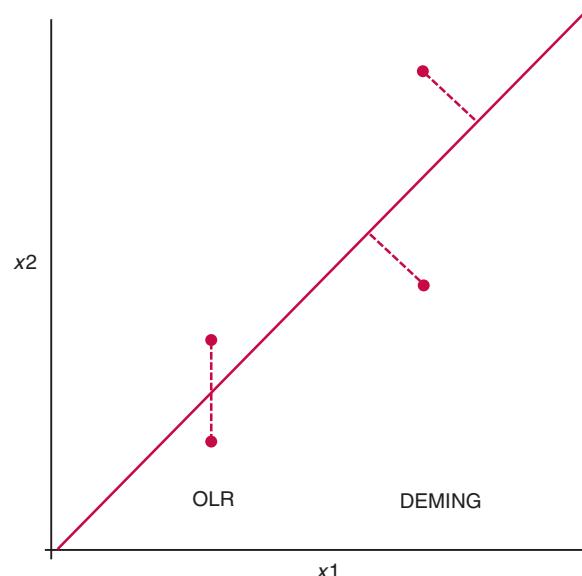
OLR regression is not recommended except in special situations. In OLR, the sum of squared distances is minimized in the vertical direction to the line (see Fig. 2.25). It can be proven theoretically that neglect of the random error in  $x_1$  induces a downward biased slope estimate

$$\beta' = \beta \left[ \sigma_{x_{1,\text{target}}}^2 / \left( \sigma_{x_{1,\text{target}}}^2 + \sigma_{x_1}^2 \right) \right] = \beta / \left[ 1 + \left( \sigma_{x_1} / \sigma_{x_{1,\text{target}}} \right)^2 \right]$$

where  $\sigma_{x_{1,\text{target}}}$  is the SD of  $X_{1,\text{target}}$  values.<sup>5</sup> The magnitude of the bias depends on the ratio between the SD for the random error in  $x_1$  and the SD of the  $X_{1,\text{target}}$  values. Fig. 2.26 shows the bias as a function of the ratio of the random error SD to the SD of the  $X_{1,\text{target}}$  value dispersion. For a ratio up to 0.1, the bias is less than 1%. At a ratio of 0.33, the bias amounts to 10%; it increases further for increasing ratios. In a given case, one can take the analytical SD (e.g., from QC data)



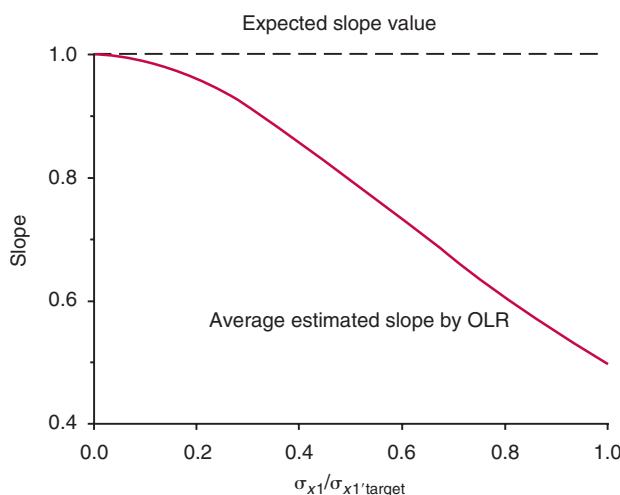
**FIGURE 2.24** The model assumed in ordinary least-squares regression. The  $x_2$  values are Gaussian distributed around the line with constant standard deviation over the analytical measurement range. The  $x_1$  values are assumed to be without random error.  $\sigma_{21}$  ( $\sigma_{yx}$ ) is shown.



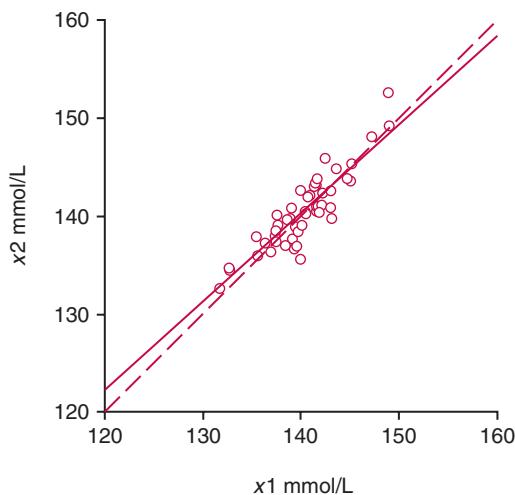
**FIGURE 2.25** In ordinary least-squares regression (OLR), the sum of squared deviations from the line is minimized in the vertical direction. In Deming regression analysis, the sum of squared deviations is minimized at an angle to the line, depending on the random error ratio. Here the symmetric case is displayed with orthogonal deviations. (From Linnet K. The performance of Deming regression analysis in case of a misspecified analytical error ratio. *Clin Chem* 1998;44:1024–31.)

and divide by the SD of the measured  $x_1$  values, which approximately equals the SD of  $X_{1,\text{target}}$  values. As an example, a typical comparison study for two serum sodium methods may be associated with a downward directed slope bias of about 10% (Fig. 2.27).

In the example presented previously, the ratio of the analytical SD to the SD of the target value distribution is large because of the tight physiologic regulation of electrolyte concentrations, which means that the biological variation is limited. Most other types of analytes exhibit wider distributions, and the ratio of error to target value distribution is



**FIGURE 2.26** Relations between the true (expected) slope value and the average estimated slope by ordinary least-squares regression (OLR). The bias of the OLR slope estimate increases negatively for increasing ratios of the standard deviation (SD) random error in  $x_1$  to the SD of the  $X_1$  target value distribution.



**FIGURE 2.27** Simulated comparison of two sodium methods. The solid line indicates the average estimated ordinary least-squares regression (OLR) line, and the dotted line is the identity line. Even though no systematic difference is evident between the two methods, the average OLR line deviates from the identity line corresponding to a downward slope bias of about 10%.

smaller. For example, for analytes with a distribution of longer than 1 decade and an analytical error corresponding to a CV of 5% at the middle of the analytical measurement range, the OLR slope bias amounts to about -1%.

### Computation Procedures for Ordinary Least-Squares Regression and Deming Regression

Assuming no errors in  $x_1$  and a Gaussian error distribution of  $x_2$  with constant SD throughout the analytical measurement range, OLR is the optimal estimation procedure, as proved by Gauss in the eighteenth century. Given errors in both  $x_1$  and  $x_2$ , the Deming approach is the method of choice.<sup>56</sup> It should be noted for these parametric procedures that only the error distributions must be Gaussian or normal. The least-squares principle does not require normality to be

applied, but it is optimal under normality conditions, and the nominal type I errors for associated statistical tests for slope and intercept hold true under this assumption. The procedures are generally robust toward deviations from normality, but they are sensitive to outliers because of the squaring principle. Finally, the distribution of the  $x_1$  and  $x_2$  values over the measurement range does not have to be normal. A uniform distribution over the analytical measurement range is generally of advantage, but the distribution in principle may take any form. For both procedures, we may evaluate the SD of the dispersion in the vertical direction around the line (commonly denoted  $SD_{yx}$  and here given as  $SD_{21}$ ). We have

$$SD_{21} = \left[ \sum (x_{2i} - \bar{x}_{2\text{Targetest}})^2 / (N-1) \right]^{0.5}$$

Further discussion regarding the interpretation of  $SD_{21}$  will be given later.

To compute the slope in Deming regression analysis, the ratio between the SDs of the random errors of  $x_1$  and  $x_2$  is necessary, that is,

$$\lambda = (\sigma_{RB1}^2 + \sigma_{AI}^2) / (\sigma_{RB2}^2 + \sigma_{A2}^2)$$

$SD_{AS}$  can be estimated from duplicate sets of measurements as

$$SD_{AI}^2 = (1/2N) \left[ \sum (x_{12i} - \bar{x}_{11i})^2 \right]$$

$$SD_{A2}^2 = (1/2N) \left[ \sum (x_{22i} - \bar{x}_{21i})^2 \right]$$

or they may be available from QC data. The latter is a practical approach that avoids the need for duplicate measurements by each measurement procedure.

If a specific value for  $\lambda$  is not available and the two routine methods that are compared are likely to be associated with random errors of the same order of magnitude,  $\lambda$  can be set to 1. The Deming procedure is generally relatively insensitive to a misspecification of the  $\lambda$  value.<sup>57</sup>

Formulas for computing slope ( $\beta$ ), intercept ( $\alpha_0$ ), and their SEs are available from other sources<sup>5,49,56</sup> and are not provided here.

### Evaluation of the Random Error Around an Estimated Regression Line

The estimated slope and intercept provide an estimate of the systematic difference or calibration bias between two methods over the analytical measurement range. Additionally, an estimate of the random error is important. It is common to consider the dispersion around the line in the vertical direction, which is quantified as  $SD_{yx}$  (here denoted  $SD_{21}$ ).  $SD_{21}$  was originally introduced in the context of OLR, but it also can be considered in relation to Deming regression analysis.

### Interpreting $SD_{yx}$ ( $SD_{21}$ ) With Random Errors in Both $x_1$ and $x_2$

With regard to  $SD_{21}$ , we have here without sample-related random interferences

$$\sigma_{21}^2 = \beta^2 \sigma_{AI}^2 + \sigma_{A2}^2$$

Thus  $\sigma_{21}$  reflects the random error both in  $x_1$  (with a rescaling) and in  $x_2$ . Often  $\beta$  is close to unity, and in this case,  $\sigma_{21}^2$  becomes approximately the sum of the individual squared SDs. This relation holds true for both Deming and OLR

analyses. Frequently, OLR is applied in situations associated with random measurement error in both  $x_1$  and  $x_2$ , and in these situations,  $\sigma_{21}$  reflects the errors in both.

The presence of sample-related random interferences in both  $x_1$  and  $x_2$  gives the following expression:

$$\sigma_{21}^2 = (\beta^2 \sigma_{A1}^2 + \sigma_{A2}^2) + (\beta^2 \sigma_{RB1}^2 + \sigma_{RB2}^2)$$

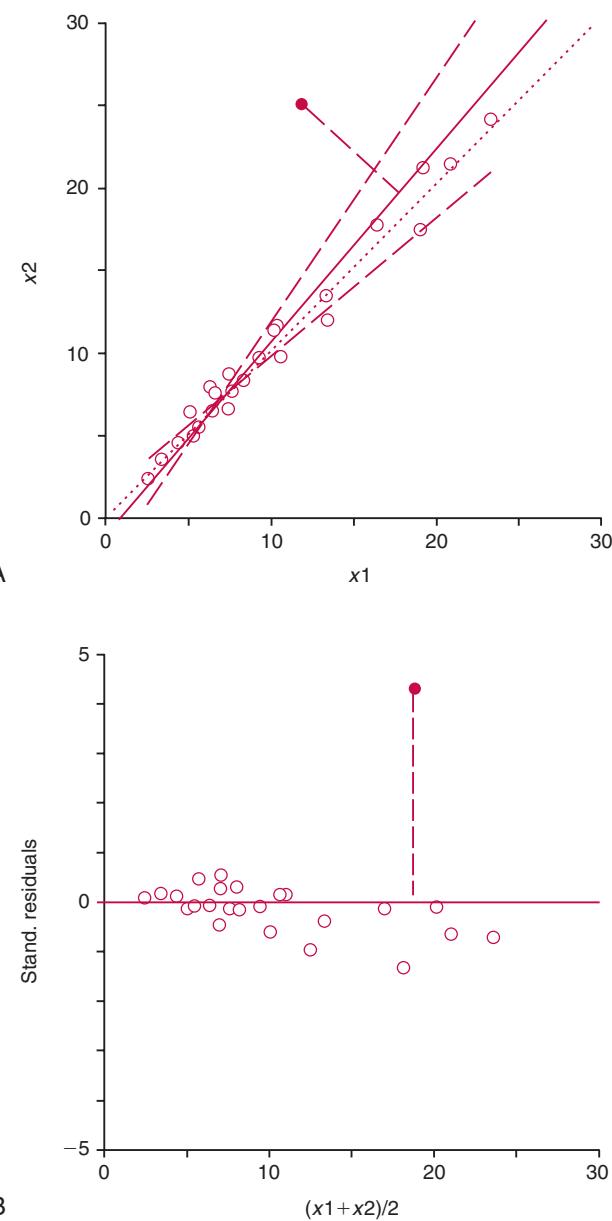
Thus the  $\sigma_{21}$  value is influenced by the slope value and the analytical error components  $\sigma_{A1}$  and  $\sigma_{A2}$  (grouped in the first bracket) and  $\sigma_{RB1}$  and  $\sigma_{RB2}$  (grouped in the second bracket). In many cases, the slope is close to unity, in which case we have simple addition of the components. As mentioned earlier, the sample-related random interferences may not be independent. In this case, simple addition of the components is not correct because a covariance term should be included. However, in a real case, we can estimate the combined effect corresponding to the bracket term. Information on the analytical components is usually available from duplicate sets of measurements or from QC data. On this basis, the combined random bias term in the second bracket can be derived by subtracting the analytical components from  $\sigma_{21}$ . Overall, it can be judged whether the total random error is acceptable or not. The systematic difference can be adjusted for relatively easily by rescaling one of the sets of measurements. However, if the random error term is very large, such a rescaling does not ensure equivalency of measurements with regard to individual samples. Thus it is important to assess both the systematic difference and the random error when deciding whether a new routine method can replace an existing one.

### Assessment of Outliers

The principle of minimizing the sum of squared distances from the line makes the described regression procedures sensitive to outliers, and an assessment of the occurrence of outliers should be carried out routinely. The distance from a suspected outlier to the line is recorded in SD units, and the outlier is rejected if the distance exceeds a predetermined limit (e.g., 3 or 4 SD units). In the case of OLR, the SD unit equals  $SD_{21}$ , and the vertical distance is considered. For Deming regression analysis, the unit is the SD of the deviation of the points from the line at an angle determined by the error variance ratio  $\lambda$ . A plot of these deviations, a so-called residuals plot, conveniently illustrates the occurrence of outliers.<sup>54</sup> Fig. 2.28 A illustrates an example of Deming regression analysis with occurrence of an outlier and the associated residuals plot (B), which clearly shows the outlier pattern. In this example, the residuals plot was standardized to unit SD. Use of an outlier limit of 4 SD units in this example led to rejection of the outlier, and a reanalysis was undertaken. In this example, rejection of the outlier changed the slope from 1.14 to 1.03. With regard to outliers, these measurements should not be rejected automatically; the reason for their presence should be investigated as a method limitation (e.g., possibly a non-specificity for the analyte).

### The Correlation Coefficient

Now that the random error components related to regression analysis have been outlined, some comments on the correlation coefficient may be appropriate. The ordinary correlation coefficient,  $\rho$ , also called the Pearson product moment



**FIGURE 2.28** A, A scatter plot with the Deming regression line (solid line) with an outlier (filled point). The dotted straight line is the diagonal, and the curved dashed lines demarcate the 95% confidence region. B, Standardized residuals plot with indication of the outlier.

correlation coefficient, is estimated as  $r$  from sums of squared deviations for  $x_1$  and  $x_2$  values as follows:

$$r = p/(uq)^{0.5}$$

where

$$P = \sum (x_{1i} - x_{1m})(x_{2i} - x_{2m})$$

$$u = \sum (x_{1i} - x_{1m})^2 \quad \text{and} \quad q = \sum (x_{2i} - x_{2m})^2$$

and

$$x_{1m} = \sum x_{1i}/N \quad \text{and} \quad x_{2m} = \sum x_{2i}/N$$

A look at the theoretical model reveals that  $\rho$  is related to the ratio between the SDs of the distributions of target values

( $\sigma_{X1\text{target}}$  and  $\sigma_{X2\text{target}}$ ) and the associated independent total random error components<sup>58</sup> ( $\sigma_{\text{ex}1}$  and  $\sigma_{\text{ex}2}$ ):

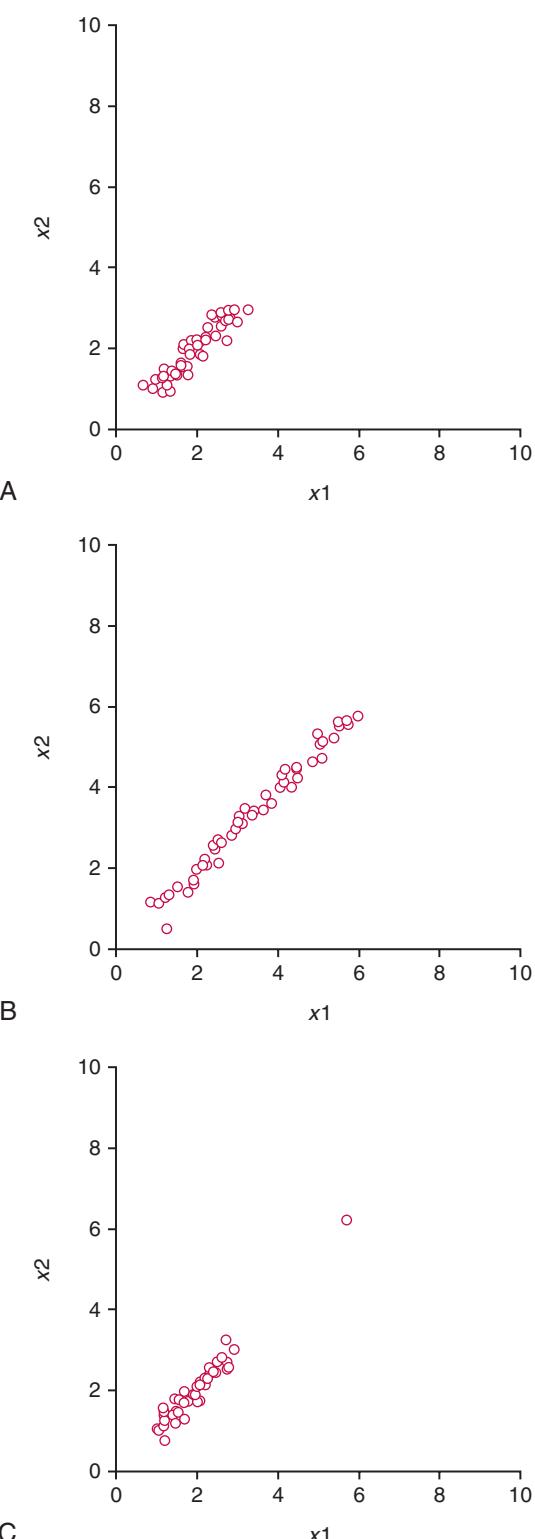
$$\rho = \sigma_{X1\text{target}} \sigma_{X2\text{target}} / \left[ \left( \sigma_{X1\text{target}}^2 + \sigma_{\text{ex}1}^2 \right) \left( \sigma_{X2\text{target}}^2 + \sigma_{\text{ex}2}^2 \right) \right]^{0.5}$$

The total random error components comprise both imprecision error and sample-related random interferences (i.e.,  $\sigma_{\text{ex}1}^2 = \sigma_{A1}^2 + \sigma_{RB1}^2$  and  $\sigma_{\text{ex}2}^2 = \sigma_{A2}^2 + \sigma_{RB2}^2$ ). Thus  $\rho$  is a relative indicator of the amount of dispersion around the regression line. If the numeric interval of values is short,  $\rho$  tends to be low and vice versa for a long range of values. For example, consider simulated examples, where the random errors of  $x_1$  and  $x_2$  are the same but the width of the distributions of measured values differs (Fig. 2.29A and B). In A, the target values are uniformly distributed over the range 1 to 3, and in B, the range is 1 to 6. The random error SD is presumed constant, and it is set to 0.15 for both  $x_1$  and  $x_2$ , corresponding to a CV of 5% at the value 3. Given sets of 50 paired measurements, the correlation coefficient is 0.93 in case A and 0.99 in case B. Furthermore, a single point located outside the range of the rest of the observations exerts a strong influence (Fig. 2.29C). In C, 49 of the observations are distributed within the range 1 to 3, with a single point located apart from the others around the value 6, other factors being equal. The correlation coefficient here takes an intermediate value, 0.97. Thus a single point located away from the rest has a strong influence (a so-called influential point). Note that it is not an outlying point, just an aberrant point with regard to the range.

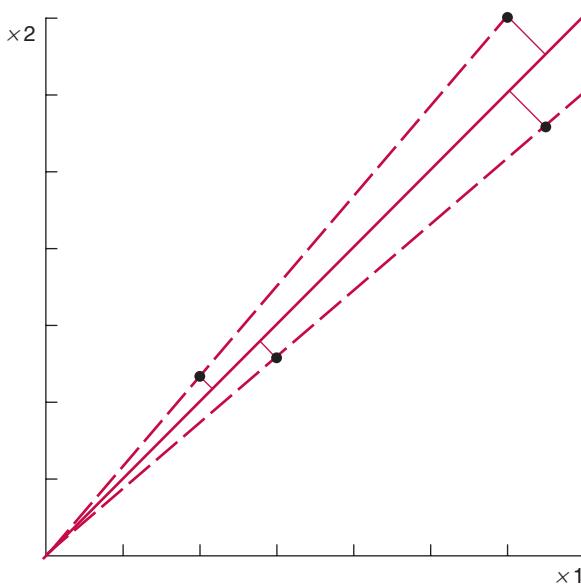
Although  $\sigma_{21}$  is the relevant measure for random error in method comparison studies,  $\rho$  is still incorrectly used as a supposed measure of agreement between two methods. It should be noted that a systematic difference due to a difference with regard to calibration is not expressed through  $\rho$  but solely in the form of an intercept ( $\alpha_0$ ) deviation from zero or a slope ( $\beta$ ) deviation from unity. Thus even though the correlation coefficient is very high, a considerable calibration bias may be noted between the measurements of two methods.

### Regression Analysis in Cases of Proportional Random Error

As discussed in relation to the precision profile, for analytes with extended ranges (e.g., 1 or several decades), the  $SD_A$  is seldom constant. Rather, a proportional relationship may apply. This may also be true for the random bias components. In this situation, the regression procedures described previously may still be used, but they are not optimal because the SEs of slope and intercept become larger than is the case when a weighted form of regression analysis is applied. The optimal approaches are weighted forms of regression analysis that take into account the relationship between random error and analyte concentration.<sup>49,54</sup> Given a proportional relationship, a weighted procedure assigns larger weights to observations in the low range; low-range observations are more precise than measurements at higher concentrations that are subject to larger random errors. More specifically, weights are applied in the computations that are inversely proportional to the squared SDs (variances) that express the random error. In the weighted modification of the Deming procedure, distances from  $(x_{1i}, x_{2i})$  to the line are inversely weighted according to the squared SDs at a given concentration



**FIGURE 2.29** Scatter plots illustrating the effect of the range on the value of the correlation coefficient  $\rho$ . **A**, Target values are uniformly distributed over the range 1 to 3 with random errors of both  $x_1$  and  $x_2$  corresponding to a standard deviation (SD) of 5% of the target value at 3 (constant error SDs). **B**, The range is extended to 1 to 6 with the same random error levels. The correlation coefficient equals 0.93 in A and 0.99 in B. **C**, The effect of a single aberrant point is shown. Forty-nine of the target values are distributed over the range 1 to 3, with a single point at 6. The correlation coefficient is 0.97.



**FIGURE 2.30** Distances from data points to the line in weighted Deming regression assuming proportional random errors in  $x_1$  and  $x_2$ . The symmetric case is illustrated with equal random errors and a slope of unity yielding orthogonal projections onto the line. (Modified from Linnet K. Necessary sample size for method comparison studies based on regression analysis. *Clin Chem* 1999;45:882–94. Used with permission.)

(Fig. 2.30). The regression procedures are most conveniently performed using dedicated software.

### Testing for Linearity

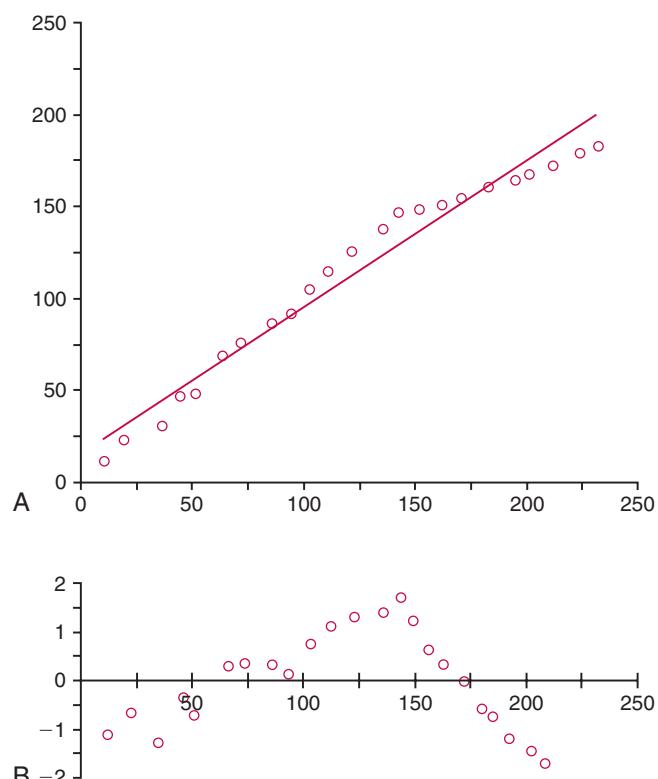
Splitting of the systematic error into a constant and a proportional component depends on the assumption of linearity, which should be tested. A convenient test is a runs test, which in principle assesses whether negative and positive deviations from the points to the line are randomly distributed over the analytical measurement range. The term *run* here relates to a sequence of deviations with the same sign. Consider, for example, the situation with a downward trend of  $x_2$  values at the upper end of the analytical measurement range (Fig. 2.31A). The SDs from the line (i.e., the residuals) will tend to be negative in this area instead of being randomly distributed above and below the line (Fig. 2.31B).<sup>20</sup> Given a sufficient number of points, such a sequence will turn out to be statistically significant in a runs test.

### Nonparametric Regression Analysis (Passing-Bablok Procedure)

The slope and the intercept may be estimated by a nonparametric procedure, which is robust to outliers and requires no assumptions of Gaussian error distributions.<sup>59,60</sup> The method takes measurement errors for both  $x_1$  and  $x_2$  into account, but it presumes that the ratio between random errors is related to the slope in a fixed manner:

$$\lambda = \left( \text{SD}_{\text{RB1}}^2 + \text{SD}_{\text{Al}}^2 \right) / \left( \text{SD}_{\text{RB2}}^2 + \text{SD}_{\text{A2}}^2 \right) = 1/\beta^2$$

Otherwise, a biased slope estimate is obtained.<sup>49,60</sup> The procedure may be applied both in situations with random errors with constant SDs and in cases with proportional SDs. The



**FIGURE 2.31** A, Scatter plot showing an example of nonlinearity in the form of downward-deviating  $x_2$  values at the upper part of the range. B, Plot of residuals showing the effects of nonlinearity. At the upper end of the analytical measurement range, a sequence (run) of negative residuals is present.

method is not as efficient as the corresponding parametric procedures<sup>49</sup> (i.e., Deming and weighted Deming procedures). Slope and intercept with CIs are provided, together with Spearman's rank correlation coefficient. A software program is required for the procedure.

### Interpretation of Systematic Differences Between Methods Obtained on the Basis of Regression Analysis

A systematic difference between two methods is identified if the estimated intercept differs significantly from zero or if the slope deviates significantly from 1. This is decided on the basis of *t*-tests:

$$t = (a_0 - 0)/\text{SE}(a_0)$$

$$t = (b - 1)/\text{SE}(b)$$

The *t*-tests can be supplemented with 95% CIs.

$\text{SE}(a_0)$  and  $\text{SE}(b)$  are the SEs of the estimated intercept  $a_0$  and the slope  $b$ , respectively. SEs can be derived by a computerized resampling principle called *the jackknife procedure*, which in practice can be carried out using appropriate software.<sup>61</sup> Having estimated  $a_0$  and  $b$ , we have the estimate of the systematic difference between the methods,  $D_c$ , at a selected concentration,  $X1'_{\text{Targetc}}$ :

$$D_c = X2'_{\text{Targetc}} - X1'_{\text{Targetc}} = a_0 + (b - 1) X1'_{\text{Targetc}}$$

$X2'_{\text{Targetc}}$  is the estimated  $X2'$  target value at  $X1'_{\text{c}}$ . Note that  $D_c$  refers to the *systematic* difference (i.e., the difference between modified target values corresponding to a calibration

difference). The SE of  $D_c$  can be derived by the jackknife procedure using a software program. By evaluating the SE throughout the analytical measurement range, a confidence region for the estimated line can be displayed. If method comparison is performed to assess the calibration to a reference measurement procedure, correction of a significant systematic difference  $\Delta_{c}$  will often be performed by recalibration [ $x_{2,\text{rec}} = (x_1 - a_0)/b$ ]. The associated standard uncertainty is the SE of  $\Delta_{c}$ . Even though the intercept and the slope are not significantly different from zero and 1, respectively, the combined expression  $\Delta_{c}$  may be significantly different from zero.

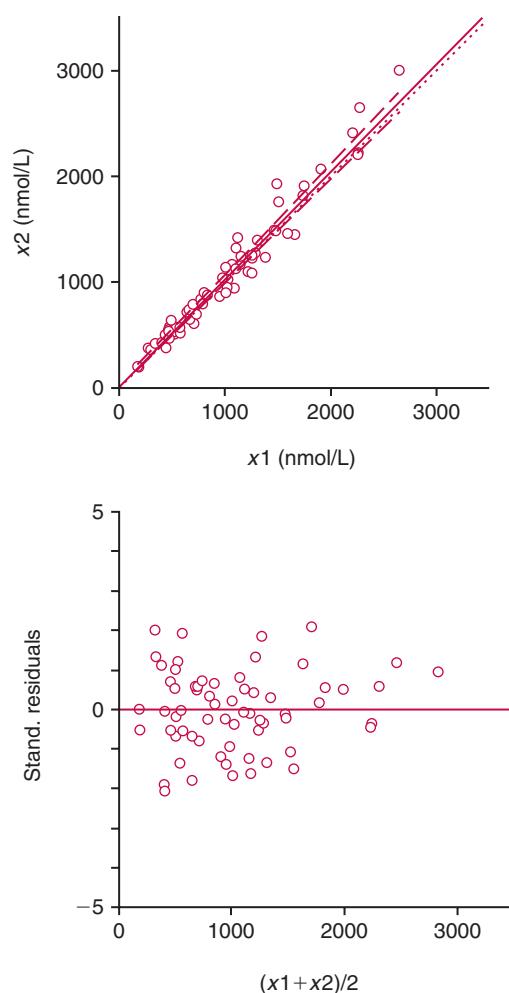
### Example of Application of Regression Analysis (Weighted Deming Analysis)

Application of weighted Deming regression analysis may be illustrated by the comparison of drug assays example [ $N = 65$  ( $x_1, x_2$ ) single measurements]. As outlined in the section on the Bland-Altman plot (see Fig. 2.15), in this example the random error of the differences increases with the concentration, suggesting that the weighted form of Deming regression analysis is appropriate. Fig. 2.32 shows (A) the estimated regression line with 95% confidence bands and (B) a plot of normalized residuals. The nearly homogeneous scatter in the residuals plot supports the assumed proportional random error model and the assumption of linearity. The slope estimate (1.014) is not significantly different from 1 (95% CI: 0.97 to 1.06), and the intercept is not significantly different from zero (95% CI: -6.7 to 47.4) (Table 2.9). A runs test for linearity does not contradict the assumption of linearity. The amount of random error is quantified in the form of the  $SD_{21}$  proportionality factor equal to 0.11, or 11%. In the present example, with a slope close to unity and two routine methods with assumed random errors of about the same magnitude, we divide the random error by the square root of 2 and get  $CV_{x_1} = CV_{x_2} = 7.8\%$ . QC data in the laboratory have provided  $CV_{AS}$  of 6.1% and 7.2% for methods 1 and 2, respectively. Thus in this example, the random error may be attributed largely to analytical error. The assay principle for both methods is HPLC, which generally is a rather specific measurement principle; considerable random bias effects are not expected in this case.

In Table 2.9, estimated systematic differences at the limits of the therapeutic interval (300 and 2000 nmol/L) are displayed (24.6 and 48.9 nmol/L, respectively). This corresponds to percentage values of 8.2% and 2.4%, respectively. Estimated SEs by the jackknife procedure yield the 95% CIs, as shown in the table. At the low concentration, the difference is significant (95% CI: 5.7 to 44 nmol/L; does not include zero), which is not the case at the high level (95% CI: -19 to 117 nmol/L). Even though the intercept and slope estimates separately are not significantly different from the null hypothesis values of 0 and 1, respectively, the combined difference  $\Delta_{c}$  is significant at low concentrations in this example. If the difference is considered of medical importance and both methods are to be used simultaneously in the laboratory, recalibration of one of the methods might be considered.

### Discussion of Application of Regression Analysis

Generally, it is recommended that Deming or weighted Deming regression analysis should be used to operate with a type of regression analysis that is based on a correct error model.



**FIGURE 2.32** An example of weighted Deming regression analysis for the comparison of drug assays. **A**, The solid line is the estimated weighted Deming regression line, the dashed curves indicate the 95%-confidence region, and the dotted line is the line of identity. **B**, A plot of residuals standardized to unit standard deviation. The homogeneous scatter supports the assumed proportional error model and the assumption of linearity.

**TABLE 2.9 Results of Weighted Deming Regression Analysis for the Comparison of Drug Assays Example,  $n = 65$  Single ( $x_1, x_2$ ) Measurements**

	Estimate	SE	95% CI
Slope ( $b$ )	1.014	0.022	0.97 to 1.06
Intercept ( $a_0$ )	20.3	13.5	-6.7 to 47.4
Weighted correlation coefficient	0.98		
$SD_{21}$ proportionality factor	0.11		
Runs test for linearity	NS		
$\Delta_{c} = X_2 - X_1$ at $X_c = 300$	24.6	9.5	5.72 to 43.6
$\Delta_{c} = X_2 - X_1$ at $X_c = 2000$	48.9	34.2	-19.3 to 117

CI, Confidence interval; NS, not significant; SD, standard deviation; SE, standard error.

Most published method evaluations are based on unweighted regression analysis; here the use of unweighted analysis is considered in the setting of proportional random errors.

Basically, the Deming procedure provides unbiased estimates of slope and intercept when the SDs vary, provided that their ratio is constant throughout the analytical measurement range. This aspect is important and means that generally the estimates of slope and intercept are reliable in this frequently encountered situation. However, application of the unweighted Deming analysis in cases of proportional SD<sub>As</sub> is less efficient than applying the weighted approach. For uniform distributions of values with range ratios from 2 to 100, 1.2 to 3.7 times as many samples are necessary to obtain the same uncertainty of the slope estimated by the unweighted compared with the weighted approach.<sup>61</sup> Thus the larger the range ratio, the more inefficient is the unweighted method.

### POINTS TO REMEMBER

- Comparison of two analytical methods is usually based on parallel measurement of a suitable number of patient samples (e.g., 40 in a laboratory and 100 for a vendor of analytical kit methods).
- Data analysis can be based on either a difference plot or regression analysis, the latter providing more details.
- Differences between measurement results may rely on calibration differences, random measurement errors, and biologically based bias sources.
- The optimal regression technique takes measurement errors by both methods into account (e.g., the parametric Deming approach or the nonparametric Passing-Bablok procedure).

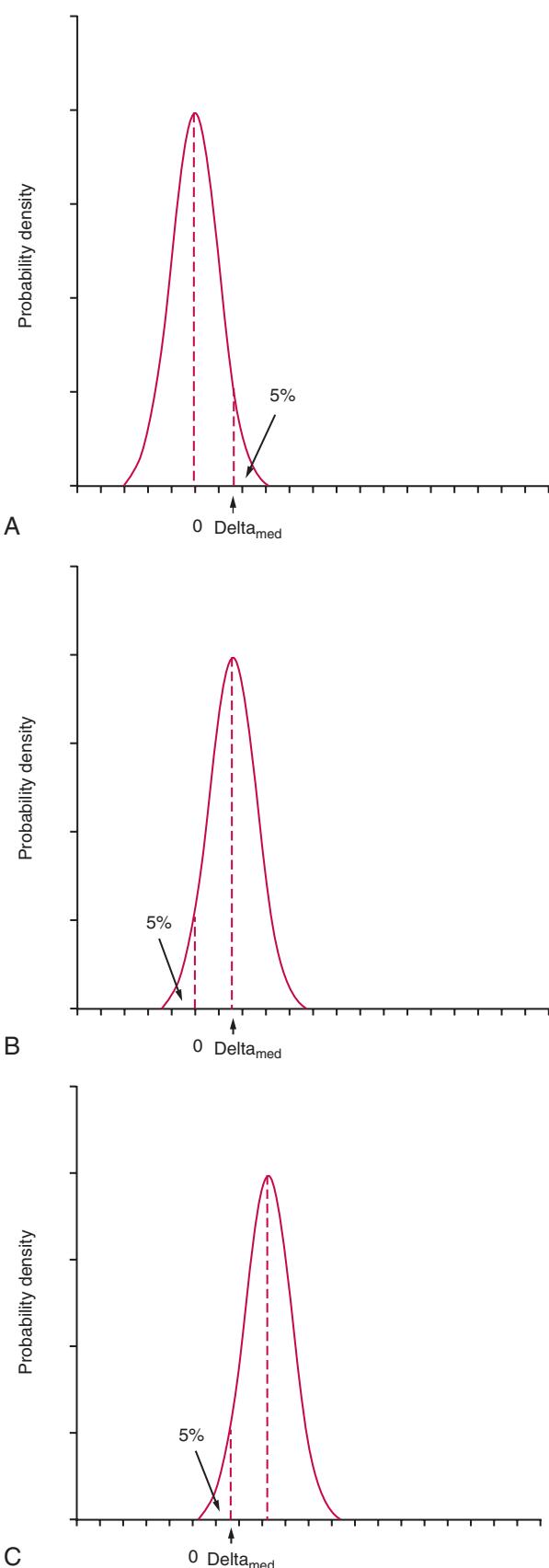
### MONITORING SERIAL RESULTS

An important aspect of clinical chemistry is monitoring of disease or treatment (e.g., tumor markers in cases of cancer, drug concentrations in cases of therapeutic drug monitoring). To assess changes in a rational way, various imprecision components have to be taken into account. Biologic within-subject variation ( $SD_I$ ) and preanalytical ( $SD_{PA}$ ) and analytical variation ( $SD_A$ ) all have to be recognized.<sup>62</sup> We assume in the following discussion that preanalytical variation is already included in the estimated within-subject variation SD, which often is the case. On this basis, using the principle of adding squared SDs (variances), a total SD ( $SD_T$ ) can be estimated as follows:

$$SD_T^2 = SD_{\text{Within } B}^2 + SD_A^2$$

The limit for statistically significant changes then is  $k\sqrt{2} SD_T$ , where  $k$  depends on the desired probability level. Considering a two-sided 5% level,  $k$  is 1.96. The corresponding one-sided factor is 1.65. If a higher probability level is desired,  $k$  should be increased.

Limits for statistically significant changes ( $\Delta_{\text{stat}}$ ) may be related to changes that are considered of medical importance by clinicians<sup>63</sup> (i.e., action limits [ $\Delta_{\text{med}}$ ]). Here we will consider a one-sided situation in which an increase is of importance and a 5% significance level is selected (i.e.,  $\Delta_{\text{stat}} = 1.65\sqrt{2} SD_T = 1.65 SD_{\Delta}$ ). Suppose as a starting point that the true change ( $\Delta_{\text{true}}$ ) for a patient is zero (Fig. 2.33A). If  $\Delta_{\text{stat}}$  is less than  $\Delta_{\text{med}}$ , the frequency of FP alarms will be



**FIGURE 2.33** The monitoring situation. A, Distribution of observed changes given a true change of zero. B, A true change equal to  $\Delta_{\text{med}}$ . C, A true change of  $(\Delta_{\text{med}} + 1.65 SD_{\Delta})$ .  $\Delta_{\text{stat}} (=1.65 SD_{\Delta})$  equals  $\Delta_{\text{med}}$  in these examples.

less than 5%. If, on the other hand,  $\Delta_{\text{stat}}$  exceeds  $\Delta_{\text{med}}$ , the frequency of FP alarms will exceed 5% (i.e., medical action will be taken too frequently). Fig. 2.33A illustrates the situation with  $\Delta_{\text{stat}}$  equal to  $\Delta_{\text{med}}$ . We now consider the situation with a true change equal to the medically important change (i.e.,  $\Delta_{\text{true}} = \Delta_{\text{med}}$ ) (see Fig. 2.33B), where exactly 50% of observed changes exceed the medically important limit. If  $\Delta_{\text{stat}}$  is less than or equal to  $\Delta_{\text{med}}$ , fewer than 5% of patients will exhibit an observed delta value in the opposite direction of the true change (an obviously misleading trend). If the condition is not met, more than 5% will have a misleading change. Finally, when the true change equals the sum of  $\Delta_{\text{med}}$  and  $\Delta_{\text{stat}}$  (see Fig. 2.33C), more than 95% of observed changes exceed the medically important change, and appropriate action will be taken for most patients.

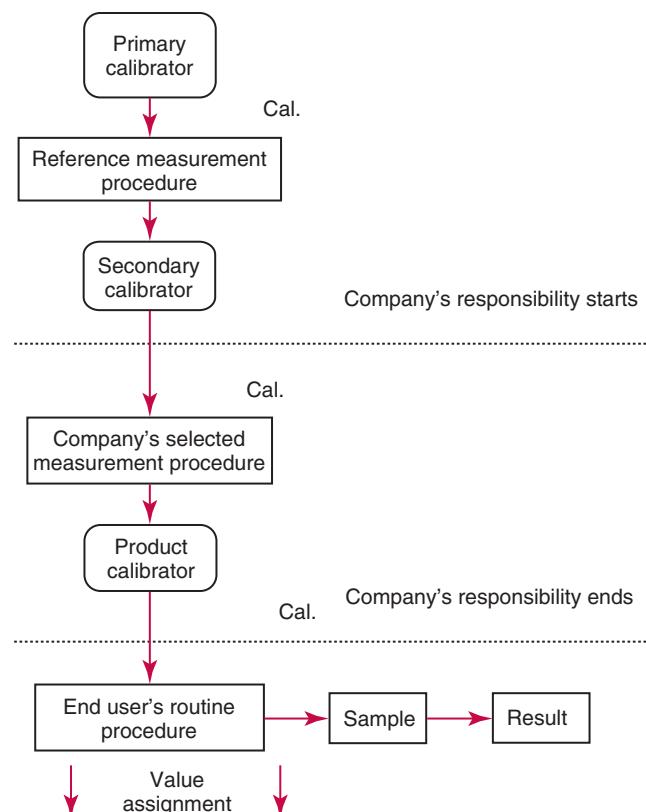
The outline presented previously illustrates that in the monitoring situation, not only the requirement for statistical significance (i.e., the type I error problem concerning false alarms) but also the type II error problem or the risk of overlooking changes should be addressed; the latter is an aspect that often is overlooked.<sup>64</sup> Provided that  $\Delta_{\text{stat}}$  is small relative to  $\Delta_{\text{med}}$ , both type I and type II errors can be kept small. On the other hand, if  $\Delta_{\text{stat}}$  equals or exceeds  $\Delta_{\text{med}}$ , the relative importance of type I and type II errors may be weighed against each other. If the consequences of overlooking a medically important change are serious, one should keep the type II error small and accept a relatively large type I error (i.e., accept the occurrence of false alarms). On the contrary, if overlooking changes only gives rise to minor or transient problems, the priority may be to keep the type I error small. In addition to simple evaluation of a shift between two measurements, as considered here, sequential results may be analyzed using more refined time-series models.<sup>65</sup>

## TRACEABILITY AND MEASUREMENT UNCERTAINTY

As outlined previously in the error model sections, laboratory results are likely to be influenced by systematic and random errors of various types. Obtaining agreement of measurements between laboratories or agreement over time in a given laboratory often can be problematic.

### Traceability

To ensure reasonable agreement between measurements of routine methods, the concept of traceability comes into focus. (See also the “Calibration Traceability to a Reference System and Commutability Considerations” section in Chapter 7.) Traceability is based on an unbroken chain of comparisons of measurements leading to a known reference value (Fig. 2.34). A hierarchical approach for tracing the values of routine clinical chemistry measurements to reference measurement procedures was proposed by Tietz<sup>66</sup> and has been adapted by the ISO. For well-established analytes, a hierarchy of methods exists with a *reference measurement procedure* at the top, *selected measurement procedures* at an intermediate level, and finally *routine measurement procedures* at the bottom.<sup>66–68</sup> A reference measurement procedure is a fully understood procedure of highest analytical quality containing a complete uncertainty budget given in Système Internationale (SI) units.<sup>12,69</sup> Reference procedures are used to measure the analyte concentration in *secondary reference*



**FIGURE 2.34** The calibration hierarchy from a reference measurement procedure to a routine assay. The uncertainty increases from top to bottom. *Cal.*, Calibration.

materials, which typically have the same matrix as samples that are to be measured by routine procedures (e.g., human serum). Secondary reference materials are usually of high analytical quality, and certified secondary reference materials must be validated for commutability with clinical samples if they are intended for use as trueness controls for routine methods.<sup>70,71</sup> Otherwise, their use is restricted to selected measurement procedures for which they are intended. The certificate of analysis should state the methods for which the secondary reference materials have been validated to be commutable with clinical samples. When no information is given for commutability, it must be assumed that the reference material is not commutable with clinical samples, and the user has the responsibility to validate commutability for the methods of interest.<sup>72</sup> Uncertainty of the measurement procedure results in increases from the top level to the bottom. ISO guidelines (15193 and 15194) address requirements for reference methods and reference materials.<sup>69,73</sup>

The measurement uncertainty down the traceability chain can end up being too high. By repetition of independent measurements at each step, it may be possible to reduce the overall uncertainty so that it becomes acceptable in relation to analytical performance specification (e.g., those based on biological variation).<sup>74</sup> For more detailed information on analytical performance specifications see “Performance of a Measurement Procedure for Its Intended Medical Use” section of Chapter 7 and the “Analytical Performance Specifications Based on Biological Variation” section of Chapter 8.

Using cortisol as an example, the primary reference material is crystalline cortisol with a chemical analysis for impurities<sup>71</sup>

(National Institute of Standards and Technology [NIST] standard reference material 921, cortisol [hydrocortisone]). A primary calibrator is then a cortisol preparation with a stated mass fraction (purity) (e.g., 0.998 and a 95% CI of  $\pm 0.001$ ). The reference measurement procedure is an isotope-dilution gas chromatography–mass spectrometry method that is calibrated with the primary calibrator. A panel of individual frozen serum samples that have values assigned by the primary reference measurement procedure is available from the Institute for Reference Materials and Measurements (IRMM) as secondary reference materials (European Reference Material [ERM]-DA451/International Federation of Clinical Chemistry and Laboratory Medicine [IFCC]). A manufacturer's *selected measurement* procedure is calibrated with the secondary reference materials and is used for measurement of the quantity in the manufacturer's *product calibrator*, which is the calibrator used for the routine method in clinical laboratories.

At the time of writing, the Joint Committee for Traceability in Laboratory Medicine (JCTLM) Database (<http://www.bipm.org/jctlm>) lists reference materials, either pure or matrix matched, for more than 95 measurands, of which over 50 are traceable to the SI (list I). SI traceable measurands include electrolytes, some metabolites (e.g., glucose, creatinine, urea, uric acid), drugs, metals, steroids, and thyroid and other hormones. Analytes listed in the database that are not traceable to the SI (list II) include 13 plasma proteins (e.g., albumin,  $\alpha_1$  acid glycoprotein,  $\alpha_1$  antitrypsin, transferrin, transerythrin), using the ERM-DA470k/IFCC. The database also lists reference measurement procedures that define the top of the traceability chain for eight serum enzymes. There are also analytes traceable to higher-order international reference preparations not listed in the JCTLM database, which have been produced by National Measurement Institutes (e.g., pH,  $PCO_2$ ,  $PO_2$ , and ammonia) or by the World Health Organization (<http://www.who.int/bloodproducts/catalogue/en>) (e.g., for allergens, various proteins and antigens, coagulation factors, and various hormones); however, these are not traceable to the SI. With protein hormones, the existence of heterogeneity or microheterogeneity complicates the problem of traceability.<sup>75,76</sup>

In case a reference measurement procedure exists for an analyte (measurand), comparable results among measurement procedures can be achieved as described earlier, so-called standardization. When reference measurement procedures are not available, so-called harmonization refers to the process of establishing comparable results among measurement procedures for the given analyte.<sup>77</sup> Harmonization is typically based on distribution among laboratories of commutable secondary reference materials with arbitrarily set target values. For more information on harmonization, readers are referred to the “Calibration Traceability to a Reference System and Commutability Considerations” section of Chapter 7.

Harmonization and standardization are especially important when disease is defined by clinical biochemistry results. This pertains to, for example, the diagnosis of diabetes based on plasma glucose determinations. The analytical quality in this instance becomes very critical for a correct evaluation. An analytical bias results in misclassification of subjects into diseased and nondiseased groups. With regard to imprecision, repeat testing may partly circumvent classification errors.<sup>78</sup>

## The Uncertainty Concept

To assess in a systematic way errors associated with laboratory results, the *uncertainty* concept has been introduced into laboratory medicine.<sup>79,80</sup> According to the ISO's “Guide to the Expression of Uncertainty in Measurement” (GUM), *uncertainty* is formally defined as “a parameter associated with the result of a measurement that characterizes the dispersion of the values that could reasonably be attributed to the measurand.” In practice, this means that the uncertainty is given as an interval around a reported laboratory result that specifies the location of the true value with a given probability (e.g., 95%). In general, the uncertainty of a result, which is traceable to a particular reference, is the uncertainty of that reference together with the overall uncertainty of the traceability chain.<sup>79</sup> Updated information on traceability aspects is available on the website of the Joint Committee on Traceability in Laboratory Medicine ([www.bipm.org/jctlm/](http://www.bipm.org/jctlm/)).

## The Standard Uncertainty ( $u_{st}$ )

The uncertainty concept is directed toward the end user (clinician) of the result, who is concerned about the total error possible and who is not particularly interested in the question of whether the errors are systematic or random. In the outline of the uncertainty concept, it is assumed that any known systematic error components of a measurement method have been corrected, and the specified uncertainty includes uncertainty associated with correction of the systematic error(s).<sup>80</sup> Although this appears logical, one problem may be that some routine methods have systematic errors dependent on the patient category from which the sample originates. For example, kinetic Jaffe methods for creatinine are subject to positive interference by 2-OXO compounds and to negative interference by bilirubin and its metabolites, which means that the direction of systematic error will be patient dependent and not generally predictable.

In the theory on uncertainty, a distinction between type A and B uncertainties is made. Type A uncertainties are frequency-based estimates of SDs (e.g., an SD of the imprecision). Type B uncertainties are uncertainty components for which frequency-based SDs are not available. Instead, uncertainty is estimated by other approaches or by the opinion of experts. Finally, the total uncertainty is derived from a combination of all sources of uncertainty. In this context, it is practical to operate with *standard uncertainties* ( $u_{st}$ ), which are equivalent to SDs. By multiplication of a standard uncertainty with a *coverage factor* ( $k$ ), the uncertainty corresponding to a specified probability level is derived. For example, multiplication with a coverage factor of two yields a probability level of approximately 95%, given a Gaussian distribution. When the total uncertainty of an analytical result obtained by a routine method is considered ( $u_{st}$ ), preanalytical variation ( $u_{PAs}$ ), method imprecision ( $u_{As}$ ), sample-related random interferences ( $u_{RBst}$ ), and uncertainty related to calibration and bias corrections (traceability) ( $u_{Tract}$ ) should be taken into account. In expressing the uncertainty components as standard uncertainties, we have the following general relation:

$$u_{st} = \left[ u_{PAs}^2 + u_{As}^2 + u_{RBst}^2 + u_{Tract}^2 \right]^{0.5}$$

Uncertainty can be assessed in various ways; often a combination of procedures is necessary. In principle, uncertainty can be judged *directly* from measurement comparisons (“top down”)<sup>81,82</sup> or *indirectly* from an analysis of individual error sources according to the law of error propagation (“error budget,” bottom up).<sup>80</sup> Measurement comparison may consist of a method comparison study with a reference method based on patient samples according to the principles outlined previously or by measurement of commutable certified matrix reference materials (CRMs). This approach demonstrates the actual uncertainty, which is an advantage. The indirect procedure, on the other hand, builds on an assumed error model, which may or may not be correct and so the uncertainty estimate. It may depend on the circumstances which procedure is feasible. Below, examples of the two types of approaches are outlined.

### Example of Direct Assessment of Uncertainty on the Basis of Measurements of a Commutable Certified Reference Material

Suppose a CRM is available that was validated to be *commutable* with patient samples for a given routine method with a specified value 10.0 mmol/L and a standard uncertainty of 0.2 mmol/L. Ten repeated measurements in independent runs give a mean value of 10.3 mmol/L with SD 0.5 mmol/L. The SE of the mean is then  $0.5/\sqrt{10} = 0.16$  mmol/L. The mean is not significantly different from the assigned value [ $t = (10.3 - 10.0)/(0.2^2 + 0.16^2)^{0.5} = 1.17$ ]. The total standard uncertainty with regard to traceability is then  $u_{\text{Trac st}} = (0.16^2 + 0.2^2)^{0.5} = 0.26$  mmol/L. If the bias had been significant, one might have considered making a correction to the method, and the standard uncertainty would then be the same at the given concentration. Thus measurements of the CRM provide an estimate of the uncertainty related to traceability, *given the assumption of commutability with patient samples*. The other components have to be estimated separately. Concerning method imprecision, long-term imprecision (e.g., observed from QC measurements) should be used rather than the short-term SD observed for CRM material. Here we suppose that the long-term  $SD_A$  is 0.8 mmol/L. Data on preanalytical variation can be obtained by sampling in duplicates from a series of patients or can be a matter of judgment (type B uncertainty) based on literature data or data on similar analytes. We here suppose that  $SD_{PA}$  equals half the analytical SD (i.e., 0.4 mmol/L). Finally, we lack data on a possible sample-related random bias component, which we may choose to ignore in the present example. The standard uncertainty of the results then becomes

$$\begin{aligned} u_s &= \left[ u_{PAst}^2 + u_{Ast}^2 + u_{RBst}^2 + u_{Tracst}^2 \right]^{0.5} \\ &= (0.4^2 + 0.8^2 + 0.26^2)^{0.5} \\ &= 0.93 \text{ (mmol/L)} \end{aligned}$$

In this case, the major uncertainty component is the long-term imprecision in the laboratory. To attain a reasonably precise uncertainty estimate, estimated SDs should be based on an appropriate number of repetitions. In the subsection on method precision, it can be seen that  $N = 30$  repetitions provides SD estimates with 95% CIs extending from about 20% below to 35% above an estimated value (see Fig. 2.7), which may be regarded as reasonable.

### Example of Direct Assessment of Uncertainty on the Basis of a Method Comparison Study With a Reference Measurement Procedure Using Patient Samples

Suppose a set of patient samples has been measured by a routine method ( $X_2$ ) in parallel with a reference measurement procedure ( $X_1$ ) and that a linear relationship exists between measurements. We want to assess a possible calibration bias and evaluate the standard uncertainty of results of the routine method on the basis of regression analysis results and information on standard uncertainty related to the traceability of reference method results. The imprecision of the reference method is 2.5% or, as a fraction (used in the following), 0.025 (= $CV_{A1}$ ), and the component related to the uncertainty of the traceability chain for the reference method is 0.020 (= $u_{\text{Trac st}}$ ). Proportional measurement errors are assumed for both methods, and a weighted form of Deming regression analysis is applied. The error variance ratio  $\lambda$  is not known exactly, but the reference method is devoid of sample-related random bias, so it is assumed that the random error is about half that of the routine method (i.e.,  $\lambda$  is set to  $\lambda^2 = \frac{1}{4}$ ). At a decision point ( $X_1'_{\text{Targtc}}$ ) (e.g., corresponding to the upper limit of the 95% reference interval), the systematic difference between methods ( $D_c = a_0 + [b - 1] X_1'_{\text{Targtc}}$ ) is estimated with SE (see section on regression):

$$D_c = X_2'_{\text{Targtc}} - X_1'_{\text{Targtc}} = 20 \text{ mg/L with } \text{SE}(D_c) = 1.0 \text{ mg/L}$$

corresponding to a relative  $SE(D_c)$  of 0.050 (=[1.0 mg/L]/[20 mg/L]). For the Deming procedures, the SE can be conveniently computed by the jackknife procedure. We observe that the difference is highly significant and decide to recalibrate the routine method in relation to the reference method using the estimated slope and intercept (i.e., the recalibrated  $x_2$  values equals  $[x_2 - a_0]/b$ ). Having done this, the routine method is assumed to have no systematic error in relation to the reference method, but when the uncertainty of the results is considered, we have to add the standard uncertainty of the bias correction. The uncertainty related to traceability for the routine method is now obtained as the uncertainty inherent to the reference method and the comparison step, that is,

$$u_{\text{Tracst}} = (0.020^2 + 0.050^2)^{0.5} = 0.054$$

We are now further interested in deriving estimates of random error components for the routine method from regression analysis results. Both analytical error (e.g., estimated from QC data) and sample-related random bias should be assessed, and it should be recognized that the observed total random error is the result of contributions from both measurement methods. Suppose that  $CV_{21}$  of the regression analysis has been calculated to be 0.10 ( $CV_{21}$  is analogous to  $SD_{21}$  or  $SD_{yx}$ ), given constant measurement errors over the analytical measurement range (i.e., an expression for the random error in the vertical direction in the  $x$ - $y$  plot). From the regression section, we have

$$CV_{21}^2 = \left[ CV_{A1}^2 + CV_{A2}^2 \right] + \left[ CV_{RB1}^2 + CV_{RB2}^2 \right]$$

By substituting  $CV_{A1} = 0.025$ ,  $CV_{RB1} = 0$ , and  $CV_{21} = 0.10$ , we derive

$$CV_{RB2}^2 + CV_{A2}^2 = 0.009375$$

and get

$$\left[ CV_{RB2}^2 + CV_{A2}^2 \right] = 0.0968$$

Thus the total random error of the routine method corresponds to a CV of 0.097. If we had measured samples in duplicate in the method comparison experiment or had available QC data, we could split the total random error into its components.  $CV_{A2}$  was here determined to be 0.035 from QC data, which gives 0.090 corresponding to  $CV_{RB2}$ . We may here note that the assumed error ratio  $\lambda$  of  $(\frac{1}{2})^2$  is not quite correct. According to our results,  $\lambda$  should be  $(0.025/0.0968)^2$ .<sup>2</sup> Although the Deming regression principle is rather robust toward misspecified  $\lambda$  values, we could choose to carry out a reanalysis with the more correct  $\lambda$  value—a process that could be iterated. Finally, assuming a value of 0.03 for the preanalytical CV, we derive a total standard uncertainty estimate of

$$u_s = \left[ u_{PAsf}^2 + u_{Asf}^2 + u_{RBsf}^2 + u_{Tract}^2 \right]^{0.5}$$

$$(0.03^2 + 0.0968^2 + 0.054^2)^{0.5} = 0.115$$

At the given decision level of 20 mg/L and with a coverage factor of 2, we obtain the 95% uncertainty interval of a single routine measurement as

$$20 \text{ mg/L} \pm (2 \times 0.115 \times 20) \text{ mg/L} = 15.4 - 24.6 \text{ mg/L}$$

Having estimated the uncertainty as outlined, additional uncertainty sources should be considered. If the comparison was undertaken within a short time period, one might consider adding an additional long-term imprecision component as a variance component to the standard uncertainty expression.

When the two approaches briefly outlined are compared, the latter is the more informative. Using a series of patient samples instead of a pooled sample, individual random bias components are included in the uncertainty estimation, assuming that the patient samples are representative. Also, natural patient samples are preferable to a stabilized pool that perhaps is distributed in freeze-dried form, which may introduce artefactual errors into some analytical systems. Using a commutable CRM, on the other hand, is more practical and in many situations is the only realistic alternative.

Care is necessary in estimating the uncertainty when it is derived from a comparison study of patient samples. First, it is important to correctly estimate the SE of the difference at selected decision points or at points covering the analytical measurement range (i.e., at the lower limit, in the middle part, and at the upper limit). From the expression of the estimated difference [ $D_c = a_0 + (b - 1) X_1' T_{Targetc}$ ], initially, one might estimate the SE (standard uncertainty) by adding (squared) the SEs of the intercept and the slope. However, simple squared addition of SEs is correct only when the independence of estimates is given (see later). Estimates of intercept and slope in regression analysis are negatively correlated, which implies that simple squared addition of SEs leads to an overestimation of the total standard uncertainty.<sup>83</sup> Rather, a direct estimation procedure for the SE should be applied, as mentioned earlier.

As mentioned earlier a method comparison study based on genuine patient samples represents a real assessment of traceability. In Fig. 2.34, the focus is on the calibration aspect

intended to mediate traceability. One should recognize that the matrix of product calibrators for practical reasons often is artificial (e.g., the matrix of a calibrator may be bovine albumin instead of human serum). Many routine methods are matrix sensitive, which implies that calibrators and patient samples are not commutable. To ensure traceability in this situation, the assigned concentration of a calibrator has to be different from the real concentration.

### Indirect Evaluation of Uncertainty by Quantification of Individual Error Source Components

On the basis of a detailed quantitative model of the analytical procedure, the standard approach is to assess the standard uncertainties associated with individual input parameters and combine them according to the law of propagation of uncertainties.<sup>79,84</sup> The relationship between the combined standard uncertainty  $u_c(y)$  of a value  $y$  and the uncertainty of the *independent* parameters  $x_1, x_2, \dots, x_n$ , on which it depends, is

$$u_c \left[ y(x_1, x_2, \dots) \right] = \left[ \sum c_i^2 u(x_i)^2 \right]^{0.5}$$

where  $c_i$  is a sensitivity coefficient (the partial differential of  $y$  with respect to  $x_i$ ). These sensitivity coefficients indicate how the value of  $y$  varies with changes in the input parameter  $x_i$ . If the variables are not independent, the relationship becomes

$$u_c \left[ y(x_1, x_2, \dots) \right] = \left[ \sum c_i^2 u(x_i)^2 + \sum c_i c_k u(x_i, x_k)^2 \right]^{0.5}$$

where  $u(x_i, x_k)$  is the covariance between  $x_i$  and  $x_k$ , and  $c_i$  and  $c_k$  are the sensitivity coefficients. The covariance is related to the correlation coefficient  $\rho_{ik}$  by

$$u(x_i, x_k) = u(x_i)u(x_k)\rho_{ik}$$

This is a complex relationship that usually will be difficult to evaluate in practice. In many situations, however, the contributing factors are independent, thus simplifying the picture. Below, some simple examples of combined expressions are shown.<sup>84</sup> The rules are presented in the form of combining SDs or CVs given *independent* input components.

$q = x + y$	$SD(q) = [SD(x)^2 + SD(y)^2]^{0.5}$
$q = x - y$	$SD(q) = [SD(x)^2 + SD(y)^2]^{0.5}$
$q = ax$	$SD(q) = a SD(x)$ and $CV(q) = CV(x)$
$q = x^p$	$CV(q) = pCV(x)$
$q = xy$	$CV(q) = [CV(x)^2 + CV(y)^2]^{0.5}$
$q = x/y$	$CV(q) = [CV(x)^2 + CV(y)^2]^{0.5}$

The formulas shown may be used, for example, to calculate the combined uncertainty of a calibrator solution from the uncertainties of the reference compound, the weighting, and dilution steps (see later).

The SD for certain non-Gaussian distributions may also be of relevance for uncertainty calculations (type B uncertainties) (Table 2.10). For example, if the uncertainty of a CRM value is given with some percentage, it may be understood as referring to a rectangular probability distribution. In relation to calibration of flasks, the triangular distribution is often assumed.

**TABLE 2.10 Relations Between Standard Deviation and Range for Various Types of Distributions**

Normal Distribution	Rectangular Distribution	Triangular Distribution
SD = Half width of 95% interval/ $t_{0.975}(v) \approx$ Half width of 95% interval/2	SD = Half width $\sqrt{3}$	SD = Half width $\sqrt{6}$

SD, Standard deviation.

It has been suggested to apply the standard uncertainty estimate as the smallest analyte reporting interval.<sup>85</sup> Using a coverage factor of two, the uncertainty of a result becomes twice the smallest reporting interval, and the reference change value (RCV) becomes approximately three times ( $2\sqrt{2}[2]$ ) the smallest reporting interval.

**Example.** Briefly, computation of the standard uncertainty of a calibrator solution will be outlined. The concentration  $C$  equals the mass  $M$  divided by the volume  $V(C = M/V)$ . We will here express the standard uncertainties as relative values and will derive the approximate total standard uncertainty by squared addition of the individual contributions. Starting with the mass, the purity is stated on the certificate as  $99.4 \pm 0.4\%$ . Assuming a rectangular distribution, the relative SD becomes  $0.004/\sqrt{3} = 0.0023$ . The uncertainty of the weighing process is known in the laboratory to have a CV of 0.1%, or 0.0010. Thus the relative standard uncertainty of the mass becomes

$$u_{Mst} = (0.0023^2 + 0.0010^2)^{0.5} = 0.0025$$

The certificate of the flask (50 mL at 20 °C) indicates  $\pm 0.1$  mL as uncertainty. Assuming here a triangular distribution, we derive the standard uncertainty as  $0.10 \text{ mL}/\sqrt{6} = 0.0408$  mL, which is converted to a relative value of 0.000816. The temperature expansion coefficient is given as 0.020 mL per degree change of temperature. Assuming a variability of  $20 \pm 4$  °C, this contribution amounts to  $\pm 0.080$  mL. Assuming here a rectangular distribution, we get an SD of  $0.080/\sqrt{3}$  mL, or 0.00092 as a relative SD. The repeatability of the volume dispensing process in the laboratory has been assessed to 0.020 mL expressed as an SD, which corresponds to a relative value of 0.00040. The total standard uncertainty of the volume dispensing process becomes

$$u_{Vst} = (0.000816^2 + 0.00092^2 + 0.00040^2)^{0.5} = 0.0013$$

The total standard uncertainty of the calibrator solution is

$$\begin{aligned} u_{Calst} &= \left( u_{Mst}^2 + u_{Vst}^2 \right)^{0.5} \\ &= (0.0025^2 + 0.0013^2)^{0.5} \\ &= 0.0028, \text{ or } 0.28\% \end{aligned}$$

Generally, when squared CVs are added, minor contributions in practice can be ignored (e.g., CVs less than a third or a quarter of the other components).<sup>79</sup>

The indirect procedure is mainly of relevance for relatively simple procedures. For closed, automated clinical chemistry procedures, it is often not possible to discern the individual error elements. Furthermore, the correlation aspect is difficult to take into account in practice. In these cases, the direct

procedure of measurement comparison is preferable. However, the indirect procedure has been applied in clinical chemistry.<sup>86,87</sup>

In some situations, a simulation model of a complex analytical method may be established to estimate the combined uncertainty of the method on the basis of input uncertainties.<sup>88,89</sup> Farrant and Frenkel<sup>89</sup> investigated Monte Carlo simulations using Microsoft Excel for the calculation of uncertainties building on functional relationships and taking into account uncertainties in empirically derived constants. In this way, complex relationships can be evaluated relatively easily and a resulting standard uncertainty estimate of an analytical result estimated. This procedure is useful for generating standard uncertainties of derived expressions, such as the estimated glomerular filtration rate or the expression for the anion gap.

### Uncertainty in Relation to Traditional Systematic and Random Error Classifications

As mentioned previously, systematic errors are not included in the uncertainty expression because it is assumed that they have been corrected. Therefore the uncertainty of the correction procedure should be taken into account. Otherwise, systematic errors have been added linearly or squared in error propagation models.<sup>90,91</sup> One may further consider that the distinction between systematic effects and random effects may be a matter of the reference frame. For example, a systematic error over time may turn into a random error because a bias may change over time. Lot-to-lot reagent effects may be interpreted as systematic or random errors. When a laboratory changes from an old to a new lot, a shift in measurement values may occur. Initially, this will be considered a systematic change. However, over a long time period involving several lots of reagents, the recorded shifts typically will be up and down and will be regarded as a long-term random error component. Additionally, a bias in a particular laboratory may be viewed as a random error component when dealing with a whole group of laboratories because individual laboratory biases appear randomly distributed and are quantified as the interlaboratory SD. Thus there are arguments for using the uncertainty concept as outlined earlier to end up with one overall uncertainty expression directed toward the end user of the laboratory result. Still, as mentioned previously, systematic errors linked to samples from specific patient subcategories may constitute a problem because a general correction is not possible. A way to quantify this error contribution is to include samples from all patient subgroups in a balanced way in a method comparison study so that this error type is incorporated into the uncertainty component related to traceability. Another problem with systematic errors is that they often depend on the analyte concentration. Thus if a commutable CRM is measured at a particular concentration, one should consider whether a bias correction is valid only at the given concentration or generally over the analytical measurement range. Furthermore, the occurrence of outliers caused by rarely occurring interference (e.g., heterophilic antibodies in relation to immunoassays) constitutes a problem.<sup>92</sup> If the uncertainty estimation is based on parametric statistics (standard uncertainty expanded by a coverage factor), inclusion of gross outliers may increase the standard uncertainty considerably and make the uncertainty specification useless. A solution

here might be to omit the outliers in the first hand, compute the 95% uncertainty interval, and then finally add a special note with regard to the probability of occurrence of outliers in the uncertainty specification.

Although it may appear complicated to specify the uncertainty in a detailed manner, a rough estimate may be obtained by adding the squares of CVs corresponding to essential uncertainty elements (e.g., grouped as factors outside the laboratory) (derived from the traceability chain), the analytical factors inside the laboratory (intermediate precision), and the preanalytical elements.<sup>93</sup> In estimating uncertainty, it is important to include relevant elements, but one must be careful to avoid counting the same elements twice. Application of the uncertainty concept and the pros and cons of “top-down” versus “bottom-up” approaches in the field of clinical chemistry are subject to some discussion.<sup>81,92,94</sup> Further reading and case studies with worked example calculations can be found in freely downloadable resources.<sup>81,82</sup>

### POINTS TO REMEMBER

- For well-established analytes, a hierarchy of methods exists with a *reference measurement procedure* at the top, *selected measurement procedures* at an intermediate level, and finally *routine measurement procedures* at the bottom.
- The uncertainty is given as an interval around a reported laboratory result that specifies the location of the true value with a given probability (e.g., 95%).
- The uncertainty of a result, which is traceable to a particular reference, is the uncertainty of that reference together with the overall uncertainty of the traceability chain.
- The uncertainty can be judged *directly* from measurement comparisons (“top down”) or *indirectly* from an analysis of individual error sources according to the law of error propagation (“error budget,” bottom up”).

### DIAGNOSTIC ACCURACY OF LABORATORY TESTS<sup>a</sup>

Application of diagnostic assays or tests represents a form of medical intervention and therefore requires systematic evaluation before the tests are put into clinical use. We here consider the basic steps for evaluation of the clinical accuracy of laboratory tests, although it applies to any type of diagnostic test, including imaging or electrophysiologic tests. In diagnostic accuracy studies, the measurements or results of one (or more) laboratory test under evaluation (i.e., the so-called index test) are compared with the results of a reference standard or method. This reference is the best prevailing test or strategy that is used to establish the presence or absence of the disease of interest (i.e., the so-called target disease that is to be detected or excluded by the index tests). This reference standard is conducted and its results interpreted as blindly for and independently from the index test(s) results as possible. Test accuracy studies show the concordance in results of the index test(s) with the presence or absence of disease as

defined by the reference standard results.<sup>98–100</sup> These studies provide information regarding the frequency of types of errors (i.e., FP and FN test results) by the index test in relation to the reference standard.

### Diagnostic Accuracy of a Test in Isolation

#### Diagnostic Accuracy, Sensitivity, and Specificity

A systematic and unbiased evaluation and comparison of tests is important.<sup>101–103</sup> The basic approach for any diagnostic accuracy study is one in which the results of the index test are compared with those of a reference test in the same individuals, all of whom are suspected to have the target disease. The simplest situation is a comparison of a single index test, with only two result categories (i.e., a dichotomous or binary index test) to a reference standard (i.e., a single-test accuracy study). The ideal dichotomous index test correctly identifies all individuals as diseased or nondiseased with an error rate of zero. A zero error rate is only possible when there is no overlap between index test results in the diseased and nondiseased individuals. However, when there is overlap in index test results, some individuals are classified wrongly as shown below in an example concerning the diagnosis of deep venous thrombosis (DVT) using a D-dimer index test. When using a quantitative (continuous) index test to classify individuals as diseased or nondiseased, a cutoff value needs to be chosen to estimate these error rates. This results in a so-called dichotomized index test.

Values of the dichotomous or dichotomized index test that exceed the cutoff in individuals having the target disease are classified as true positives (TP) (Fig. 2.35). Similarly, index test results lower than the cutoff in nondiseased individuals are true negatives (TN). Accordingly, index test results below the cutoff in truly diseased subjects are FN, and correspondingly, index test results exceeding the cutoff in truly nondiseased subjects are FP. Based on the frequencies of FN and FP results, an overall error rate or nonerror rate can be derived. The overall diagnostic accuracy of an index test is then defined as the fraction of true classifications out of all classifications:

$$\text{Diagnostic accuracy} = (\text{TB} + \text{TP}) / (\text{TN} + \text{TP} + \text{FP} + \text{FN})$$

This is an overall nonerror rate that can be subdivided into the nonerror rate of the nondiseased individuals, which is the specificity of the test, and the nonerror rate of diseased individuals, which is the sensitivity of the test

$$\text{Specificity} = \text{TN} / (\text{TN} + \text{FP})$$

$$\text{Sensitivity} = \text{TP} / (\text{TP} + \text{FN})$$

Whereas a very specific test provides negative results for all or almost all subjects who are free of the target disease, a very sensitive test detects all or almost all diseased subjects.

	Disease status	
Test result	Diseased	Nondiseased
<b>Positive</b>	TP	FP
<b>Negative</b>	FN	TN

**FIGURE 2.35** The basic  $2 \times 2$  table for estimating the diagnostic accuracy of a dichotomized quantitative test result. Positive test results are divided into true positives (TPs) and false positives (FPs) and negative results into true negatives (TNs) and false negatives (FNs). (From Linnet K, Bossuyt PM, Moons KG, Reitsma JB. Quantifying the accuracy of a diagnostic test or marker. *Clin Chem* 2012;58:1292–301.)

<sup>a</sup>This section relies on three published papers.<sup>95–97</sup>

**TABLE 2.11 Relationship Between Sample Size and 95% Confidence Intervals of a Proportion (e.g., a Sensitivity or Specificity): Selected Examples of Proportions of 0.05 and 0.8**

Sample Size	95% CI of a Proportion of 0.05	95% CI of a Proportion of 0.80
20	0.00–0.25	0.56–0.94
60	0.01–0.14	0.68–0.90
100	0.02–0.11	0.71–0.87
500	0.03–0.07	0.76–0.83
1000	0.04–0.07	0.77–0.82

CI, Confidence interval.

### Confidence Intervals of Diagnostic Accuracy, Sensitivity, and Specificity

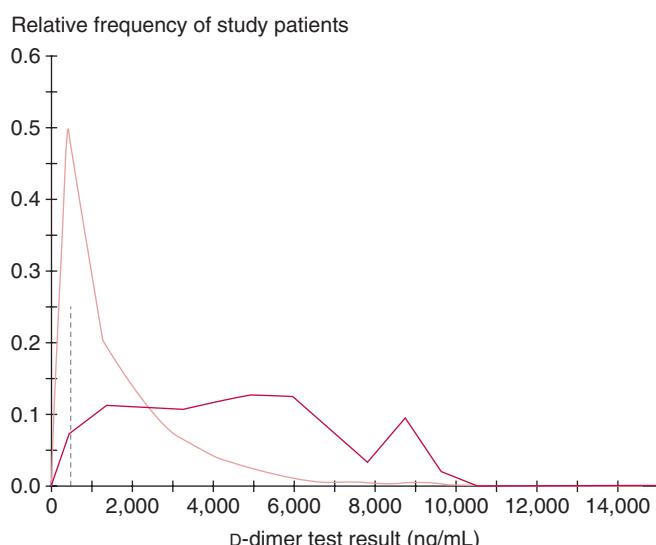
To assess the (im)precision of these estimates, CIs for either the estimates or the SEs should be specified. If the cutoff value of a quantitative index test is fixed and has not been estimated from the results obtained in the study, the binomial distribution can be applied. Given random sampling, the 95% CI of a proportion can be derived from tables or by applying simple computer programs. An approximation of the binomial to the normal distribution is often used for estimation of the 95% CI of proportions such as a sensitivity and specificity, that is,  $\pm 2 \text{ SE}(P)$ , where  $\text{SE}(P) = [P(1 - P)/N]^{0.5}$  ( $P$ , proportion;  $N$ , sample size).

The normal approximation does not work well with small sample sizes or proportions close to 0 or 1. Both situations occur frequently in diagnostic accuracy research. The method of Wilson is an alternative.<sup>104</sup> Table 2.11 displays the widths of the 95% CIs at various sample sizes of 20 to 1000 for two selected proportions, corresponding to either a sensitivity or a specificity (for that threshold) of an index test. For example, at a sample size of 20, the 95% CI extends from 0.56 to 0.94 for a proportion of 0.80. Thus rather wide estimates of specificity or sensitivity are obtained for small samples. Bachmann and colleagues<sup>105</sup> reported that for 43 nonscreening studies on diagnostic accuracy of tests, the median sample size was 118 (interquartile range, 71 to 350). For the diseased group, the median sample size was only 49 (interquartile range, 28 to 91), but for the nondiseased group, it was 76 (interquartile range, 27 to 209). The specificity and sensitivity of two tests applied in the same study subjects can be statistically compared using the McNemar's test, which is based on a comparison of paired values of true and FP or FN results.<sup>26</sup>

### Clinical Example: Accuracy of D-Dimer Test in Diagnosis of Deep Venous Thrombosis

We illustrate the concepts using some of the empirical data of a previously published study in primary care patients suspected of having DVT, the target disease (Fig. 2.36).<sup>106,107</sup> The data given here are used for illustration purposes only and not to quantify the true diagnostic accuracy of the index test for this clinical situation.

The study consisted of 2086 patients suspected of DVT, where DVT was defined as present in patients manifesting at least one of the following symptoms or signs: presence of swelling, redness, or pain in the leg. All patients were given a



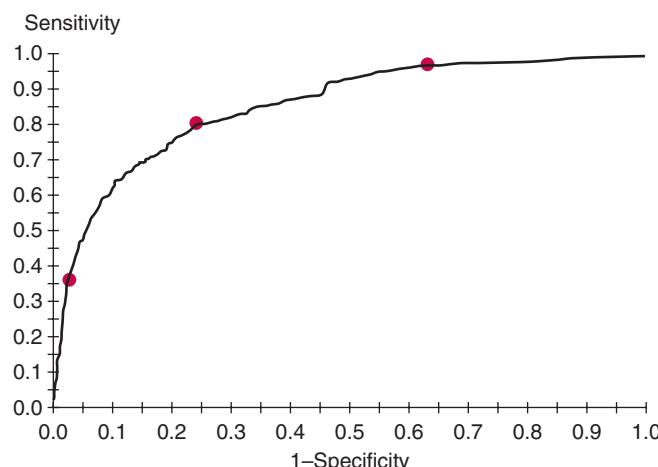
**FIGURE 2.36** Distribution of the quantitative d-dimer values for deep venous thrombosis (DVT) and non-DVT subjects in the example study. Light red line, non-DVT; red line, DVT. The dashed line indicates the commonly used cutoff value of 500 µg/L. (From Linnet K, Bossuyt PM, Moons KG, Reitsma JB. Quantifying the accuracy of a diagnostic test or marker. *Clin Chem* 2012;58:1292–301.)

standardized diagnostic workup, including medical history; clinical examination; and testing for d-dimer, the (quantitative) index test. The reference procedure consisted of repeated compression ultrasonography tests and was performed in all patients, blinded to and independent of the index test results. A total of 416 (20%) of the 2086 included patients had DVT. It should be noted that although the reference test is applied currently, it may not be infallible. The potential consequences of applying imperfect reference tests and how to cope with this problem are very important, but these aspects are beyond the scope of this chapter, and for this, we refer to the literature.<sup>108–113</sup>

Applying a commonly used cutoff of 500 µg/L or greater for the (originally) quantitative d-dimer assay (dashed line in Fig. 2.36), the sensitivity was 0.97 (i.e., 3% of the subjects with DVT had a value <500 µg/L). The specificity was only 0.37. The resulting overall diagnostic accuracy was 0.50. Whereas the test displayed good sensitivity at this threshold, detecting all but 3% of those having DVT, its specificity at this test threshold was relatively low, resulting in many FP results. The sample size was high enough to provide precise estimates of specificity and sensitivity. The SEs were 0.012 for the specificity and 0.008 for the sensitivity, resulting in CIs of 0.356 to 0.402 and 0.955 to 0.987, respectively.

### Receiver Operating Characteristic Curves

As said, for a quantitative index test, the specificity and sensitivity depend on the selected cutoff point. A plot of the sensitivity and specificity pairs for all possible cutoff values over the measurement range provides the so-called ROC curve, which is shown in Fig. 2.37 for the d-dimer example.<sup>114–117</sup> Usually, sensitivity ( $y$ ) is plotted against  $(1 - \text{specificity})$  ( $x$ ) at each possible cutoff value. The better the performance of the test, the higher the ROC curve is located in the left, upper region of the plot. With use of the ROC curve, an appropriate combination of specificity and sensitivity, or rather for an



**FIGURE 2.37** Receiver operating characteristic curve of the d-dimer assay result for diagnosis of deep venous thrombosis in our example study. The red markers correspond to various cutoff choices (from left to right, 5435, 2133, and 500  $\mu\text{g}/\text{L}$ ). (From Linnet K, Bossuyt PM, Moons KG, Reitsma JB. Quantifying the accuracy of a diagnostic test or marker. *Clin Chem* 2012;58:1292–301.)

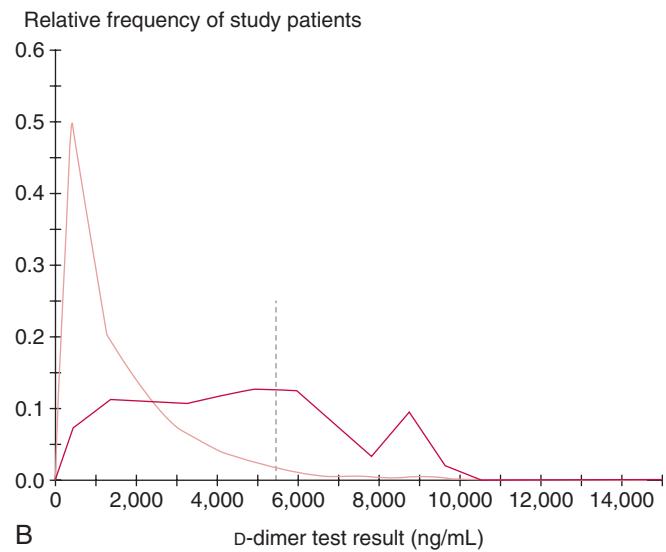
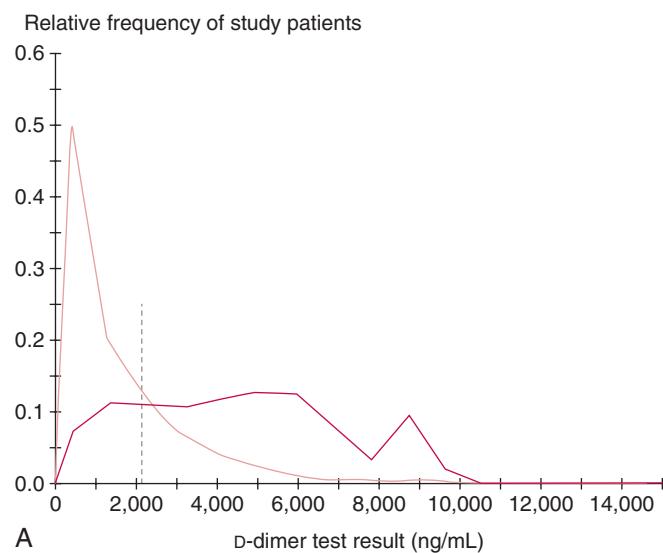
acceptable FN and FP proportion, may be chosen, and the corresponding cutoff then selected. For the d-dimer test, in the example given, the commonly used cutoff of 500  $\mu\text{g}/\text{L}$  corresponds to a sensitivity of 0.97 and (1 – specificity) of 0.63.

An area under the ROC curve (i.e., the ROC area or so-called concordance or c-index) can be assessed statistically. The approach used for assessment should emulate the approach used to estimate the ROC curve, either parametric or nonparametric, of which the latter approach generally is preferable. Standard computer programs to perform these calculations are widely available. Given an SE of the ROC area or c-index, it is possible to test whether the area significantly exceeds 0.5, which would demonstrate that the index test performs better than chance. A worthless test has an area of 0.5. Furthermore, using the SE also a 95% CI can be derived for the ROC area or c-index. For the d-dimer test in the earlier example, the area under the ROC curve was 0.86 (SE, 0.011), with a 95% CI of 0.84 to 0.88.

The ROC area provides an overall measure of an index test's diagnostic ability. It can be shown that the area under the ROC curve indicates, for all possible pairings of individuals, one with and one without the target disease, the proportion of pairs in which diseased individuals have a higher (more severe) index test result than individuals without the disease.<sup>114–117</sup> Although ROC curve evaluation has various advantages, it also has some drawbacks.<sup>117–119</sup> The ROC curve does not reflect directly the index test performance for a given cutoff value but can be used for this purpose depending on the desired sensitivity and specificity, or rather the acceptable FN and FP proportions; Fig. 2.37 also displays the sensitivity and specificity at various d-dimer cutoff points (including 500  $\mu\text{g}/\text{L}$ , as well as the cutoff values used in Fig. 2.38 below).

#### Selection of Cutoff Value in Case of Quantitative Index Tests

The specificity and sensitivity determined for an index test usually vary inversely over the range of possible cutoffs. One may select the cutoff point that provides the maximum of the



**FIGURE 2.38** Alternative cutoffs to 500  $\mu\text{g}/\text{L}$  in the d-dimer example. A, Cutoff (2133  $\mu\text{g}/\text{L}$ ) giving maximum value of the sum of the specificity and sensitivity. B, Cutoff (5435  $\mu\text{g}/\text{L}$ ) providing a high specificity (0.975). Light red line, non–deep venous thrombosis (DVT); red line, DVT. The dashed line indicates the cutoff value. (From Linnet K, Bossuyt PM, Moons KG, Reitsma JB. Quantifying the accuracy of a diagnostic test or marker. *Clin Chem* 2012;58:1292–301.)

sum of the specificity and sensitivity. In the d-dimer example, this cutoff would be close to 2000  $\mu\text{g}/\text{L}$ , yielding a specificity of 0.76 and a sensitivity of 0.80 (see Fig. 2.38A). However, this method of cutoff selection is commonly not recommended. The selection should rather be based on the intended purpose of the index test. If an index test is applied primarily to rule out the presence of disease (e.g., in the case of the d-dimer assay for exclusion of DVT), the cutoff point should be at the lower end of the distribution of values of diseased individuals (see Fig. 2.36) (e.g., a cutoff of 500  $\mu\text{g}/\text{L}$ ). At this cutoff, the sensitivity approaches 1.0. But attaining such a high sensitivity is at the cost of a loss of specificity. How low the specificity becomes depends on the extent of overlap of test values in the diseased and nondiseased individuals.

Conversely, when FP results are judged unacceptable, the cutoff should be toward the upper limit of the distribution of values for the nondiseased group. For the d-dimer test example, a cutoff value corresponding to the 97.5 percentile of the distribution of values for those not having DVT (5435 µg/L) resulted in a specificity of 0.975, but now the sensitivity was only 0.36 (i.e., nearly the opposite of the situation with a cutoff of 500 µg/L) (see Fig. 2.38B).

The estimation of an optimal cutoff point can be biased when the cutoff value is selected in the same study in which sensitivity and specificity of the index test have been estimated.<sup>26,120</sup> A good rule is to use independent samples for estimation of the optimal diagnostic cutoff value of the index test and for estimating the diagnostic accuracy measures. Evaluation of the index test in an independent sample also gives an indication of the robustness of the index test.

### Posterior Probabilities (Predictive Values)

A straightforward question arising after the application of a diagnostic index test is what is the probability that the target disease is present given the index test value ( $P[D|T_{\text{pos}}]$ )? The sensitivity and specificity of a test do not directly relate to this question. The probability of presence of target disease given the index test result is an example of a so-called posterior disease probability, where the prior probability corresponds to the prevalence of the disease in the given situation. The prevalence of disease ( $P[D]$ ) in the study sample is the a priori (pretest) probability of disease.

Given a positive test result ( $T_{\text{pos}}$ ), the posterior disease probability is estimated as the fraction of TP out of all test result positives:

$$P(D|T_{\text{pos}}) = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Analogously for a negative result ( $T_{\text{neg}}$ ), the probability that the given disease is absent is

$$P(\text{Non-}D|T_{\text{neg}}) = \frac{\text{TN}}{\text{TN} + \text{FN}}$$

Just as with sensitivity and specificity values, these posterior disease probabilities depend on the selected cutoff point for a quantitative test. In case of a dichotomous or dichotomized index test, these posterior probabilities are also called predictive values.<sup>121</sup> They are highly dependent on the disease prevalence.

From the Bayes rule, the following relations exist:

$$P(D|T_{\text{pos}}) = \frac{[\text{Sensitivity} \times P(D)]}{[\text{Sensitivity} \times P(D)] + (1 - \text{Specificity})(1 - P(D))}$$

$$P(\text{Non-}D|T_{\text{neg}}) = \frac{[\text{Sensitivity} \times (1 - P(D))]}{[\text{Sensitivity} \times (1 - P(D))] + P(D) \times (1 - \text{Specificity})}$$

### Likelihood Ratios and Odds Ratios

Besides the above parameters, one may also estimate the so-called diagnostic likelihood ratio (LR) for index test results. From relative frequency distributions for results of the index test in the nondiseased and diseased groups, one may calculate the LR of an index test result ( $X$ ) as the ratio between the heights of the relative frequency ( $f$ ) distributions at that specific test value.<sup>122</sup> We get:

$$\text{LR}(X) = f_D(X)/f_{\text{Non-}D}(X)$$

In case the relative frequency of the distribution of diseased individuals is higher than that of the nondiseased individuals, the ratio exceeds 1. This indicates that disease is more likely than nondisease given this particular index test result. More formally, the ratio can be used to calculate posterior disease probabilities given specific values of the index test ( $X$ ) and the disease prevalence ( $D$ ):

$$P(D|X) = P(D) \times \text{LR}(X)/[P(D) \times \text{LR}(X) + (1 - P(D))]$$

or a simpler calculation can be carried out using odds instead of probabilities:

$$\text{Odds}(D|X) = \text{Odds}(D) \times \text{LR}(X)$$

based on the relation

$$\text{Odds} = P(1 - P)$$

Odds is an alternative way of expressing probabilities commonly used in betting games in Anglo-Saxon countries. For example, a probability of 0.80, or 80%, corresponds to an odds value of 4 according to the formula above. The higher the odds, the closer a probability is to one. From the equation, the posterior odds are equal to the prior odds multiplied by the diagnostic LR for the result  $X$ .

For a dichotomous or dichotomized index test, the following relationships apply:

$$\text{LR}(\text{pos}) = \text{Sensitivity}/(1 - \text{Specificity})$$

$$\text{LR}(\text{neg}) = (1 - \text{Sensitivity})/\text{Specificity}$$

Although the LR approach has been tried in various situations, generally the application of diagnostic LRs has been limited in clinical chemistry. Specific conditions are required for the concept to be applied in a practical and reliable way. A simple way of achieving the posttest probability of disease from the prevalence (pretest probability of disease) and the diagnostic LR is to use the Fagan nomogram.<sup>123</sup> A recent example is the estimation of the probability of DVT from testing for d-dimer.<sup>124</sup> Finally, it can be noted that the diagnostic LR of a result  $X$  equals the slope of the ROC curve at that index test value.

### Comparison of Diagnostic Accuracy of Two Tests in Isolation

The diagnostic accuracy—that is, the ability to detect or exclude the target disease as determined by the reference method—of a new diagnostic index test is usually compared with another, established, index test. We here focus on the pure performances of the tests without consideration of other tests (i.e., we consider each test in isolation). When comparing the accuracy of two or more diagnostic index tests, a paired design is generally preferable for reasons of both validity and efficiency. In the target disease-suspected patients, the two index tests under comparison and the reference standard are performed on all subjects, again independently and blinded with regard to each other's test results. Because both index tests are applied to the same nondiseased and diseased individuals (as classified by the same reference standard), any bias effects caused by differences in disease spectrum or comorbidity are automatically balanced.

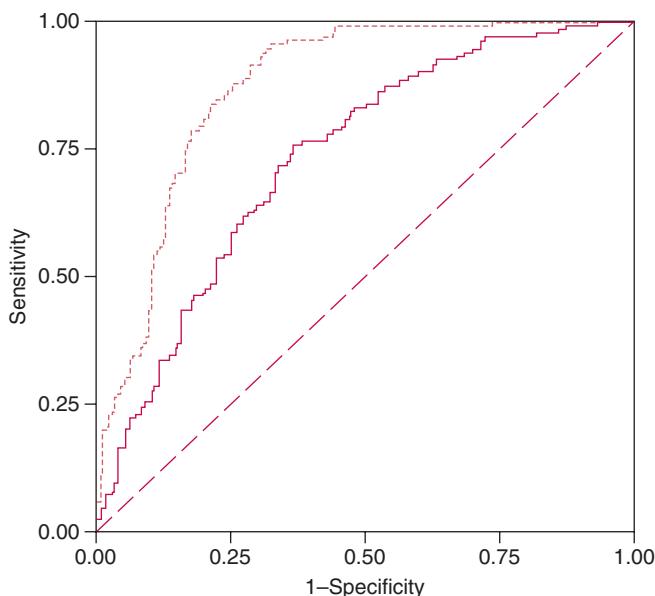
A paired comparison of, for example, the sensitivities or specificities for two dichotomous or dichotomized index tests can be evaluated using the McNemar's test.<sup>26,125</sup> The principle

of this statistical procedure is that the number of preferences for index test A (cases detected by index test A but not by index test B) is compared with the number of preferences for index test B, and if the difference exceeds some critical value, one index test is found to be superior to the other.

Receiving operating characteristic curve areas may also be compared. Here, a paired comparison should also be undertaken when the index tests have been applied in the same groups of individuals. An example of a paired comparison is displayed in Fig. 2.39. Parametric and nonparametric statistical procedures exist that usually are performed by computer programs.<sup>116,126</sup> Overall, the index test having the largest area under the ROC curve represents the best test, although this assessment becomes more difficult if the ROC curves of tests cross each other.<sup>119</sup> Preferably, CIs of areas and differences of areas should be provided.

### Shortcomings of Diagnostic Accuracy Studies of Tests in Isolation

The accuracy of a diagnostic test highly depends on the context. The estimated diagnostic accuracy measures of an index test (posterior probabilities, sensitivity, specificity, LR, or ROC area) preferably obtained from data of a cohort of target disease–suspected patients are not constant; they vary across other index test results, patient characteristics, disease prevalence, or disease severities.<sup>127–129</sup> We illustrate this for our d-dimer example in Table 2.12. The overall sensitivity and specificity for the 500 µg/L threshold were 0.97 and 0.37, respectively (upper row). However, when estimating these measures for patient subgroups within the study sample defined by other test results from patient history and physical



**FIGURE 2.39** Comparison of the receiver operating characteristic curves of two hypothetical index tests for the same target disease undertaken in the same individuals. The *dotted, red curve* represents a superior diagnostic test, both with regard to sensitivity and specificity over all possible cutoff points. The *dashed diagonal* represents a worthless test, with equal probability of a false-positive ( $1 - \text{Specificity}$ ) and false-negative ( $1 - \text{Sensitivity}$ ) result across all cutoff values (i.e., flipping a coin test). (From Linnet K, Bossuyt PM, Moons KG, Reitsma JB. Quantifying the accuracy of a diagnostic test or marker. *Clin Chem* 2012;58:1292–301.)

**TABLE 2.12 Variations in the Sensitivity and Specificity (at Cutoff Values 500 ng/mL and 1000 ng/mL) and the Receiver Operating Characteristic Area of the d-Dimer Test According to Various Other Test Results or Patient Characteristics.**

	d-DIMER > 500 ng/mL		d-DIMER > 1000 ng/mL		d-Dimer (Continuous)
	Sensitivity	Specificity	Sensitivity	Specificity	AUC (CI)
Overall	0.97	0.37	0.89	0.55	0.86 (0.84;0.88)
Previous lung embolism					
Yes ( $n = 173$ )	1.00	0.37	0.84	0.53	0.82 (0.75;0.90)
No ( $n = 1913$ )	0.97	0.37	0.89	0.55	0.86 (0.84;0.88)
Malignancy					
Yes ( $n = 115$ )	0.95	0.25	0.95	0.44	0.86 (0.79;0.93)
No ( $n = 1971$ )	0.97	0.38	0.89	0.55	0.84 (0.83;0.87)
Recent surgery					
Yes ( $n = 278$ )	0.96	0.22	0.90	0.38	0.84 (0.78;0.90)
No ( $n = 1808$ )	0.97	0.39	0.89	0.57	0.86 (0.84;0.88)
Leg trauma					
Yes ( $n = 344$ )	0.96	0.32	0.85	0.48	0.79 (0.72;0.87)
No ( $n = 1742$ )	0.97	0.38	0.89	0.56	0.86 (0.84;0.89)
Pitting edema					
Yes ( $n = 1301$ )	0.97	0.32	0.88	0.50	0.84 (0.82;0.87)
No ( $n = 785$ )	0.97	0.46	0.90	0.62	0.87 (0.84;0.91)
Pregnancy					
Yes ( $n = 45$ )	1.00	0.28	1.00	0.55	0.98 (0.00;1.00)
No ( $n = 2041$ )	0.97	0.37	0.89	0.55	0.85 (0.83;0.88)

AUC, Area under the receiver operating characteristic curve; CI, 95% confidence interval.

examination, substantial differences appear with regard to specificity, especially for the malignancy, recent surgery, and pitting-edema subgroups. At a higher threshold (1000 µg/L) variations in sensitivity also occur, for example, in the pregnancy and previous embolism subgroups. The last column of Table 2.12 reveals that this variation in single-test accuracy measures also holds true for non-threshold-dependent measures such as the ROC area. The ROC area was from 0.79 to 0.98, with 0.86 for the total study group. Although all these differences should not be overinterpreted, one must always be careful when judging a single test's diagnostic accuracy measures. A diagnostic laboratory test should always be considered in relation to a specific clinical situation and its results judged within the diagnostic pathway (i.e., in view of the results of other usually applied tests) in which the test under study is to be applied.<sup>128–130</sup> How to do so is covered in the next section.

The Standards for Reporting of Diagnostic Accuracy Studies (STARD) initiative, originally launched in 2003<sup>101</sup> and updated in 2015,<sup>102</sup> aims to improve the quality and reporting of diagnostic accuracy studies. A checklist guides investigators regarding what information to report on patient selection, the order of test application, and the number of individuals undergoing the test under evaluation, the reference test, or both, and other characteristics important for unbiased study design (Box 2.2).<sup>102</sup> Similarly, the so-called Quality Assessment Tool for Diagnostic Accuracy Studies (QUADAS-2) tool to critically appraise and assess risk of bias in primary diagnostic accuracy studies has been developed to assist systematic reviews of diagnostic accuracy studies (<http://www.quadas.org>).<sup>131</sup> For more information on STARD and QUADAS and the Diagnostic Test Accuracy Review Group of the Cochrane Collaboration, see Chapter 10 on evidence-based laboratory medicine.

### POINTS TO REMEMBER

- The diagnostic accuracy of a test indicates the frequency and type of errors that a test will produce when differentiating between patients with and without the target disease.
- The cohort design based on patients suspected of the diseases targeted by the index test is generally preferable for evaluating diagnostic accuracy.
- It is not meaningful to regard estimates of diagnostic performance as properties of the test itself but rather to interpret them as depending on the setting in which the index test was applied and dependent on other tests that are commonly used in that setting.

### Diagnostic Accuracy of a Test in the Clinical Context

The diagnostic process in practice begins with a patient having particular symptoms or signs. These symptoms and signs may direct the suspicion toward several possible diseases (the differential diagnosis). The diagnostic workup is often primarily targeted to include or exclude a particular disease or disorder, the so-called target disease, among several possible differential diagnoses.<sup>99,101,132–135</sup> For example, a woman showing up with a red, swollen leg may be suspected of having DVT; a man with blood in his stool may be suspected of having colon carcinoma; and a child with convulsions may be suspected of having

bacterial meningitis. The target disorder in question can be the most severe disorder of the differential diagnoses ("the one not to miss") but also the most probable one.

The diagnostic process commonly consists of a series of sequential steps in which much diagnostic information (i.e., diagnostic test results) is acquired. After each step, the physician intuitively judges the probability of the target disease being present. The initial step always consists of patient history and physical signs. If uncertainty about the presence and type of disease remains, subsequent tests are performed, often in another stepwise fashion. These supplementary tests may consist of simple blood or urine tests or be imaging, electrophysiology, or genetic tests or even later in the process more invasive testing such as biopsy, angiography, or arthroscopy. The supplementary information of each subsequent test is implicitly added to the yet collected diagnostic information, and the target disease probability is constantly updated. This process continues until the target disease can be included or excluded with sufficient certainty and some therapeutic management can be started, including the decision to refrain from treatment.

Hardly any diagnosis is based on a single test; for example, information from the history and the physical examination are almost always collected before any laboratory test is applied. Rather, the diagnostic context involves a multivariable (multiple-test) and phased process in which physicians decide whether the next test will add information to what is already established.<sup>129,136,137</sup>

Investigations of diagnostic laboratory tests should incorporate this multivariable clinical context in their studies. Laboratory tests should not be evaluated in isolation; rather, their studies should reflect the steps in the diagnostic process so that the added value of such tests in excess of the information that is already present can be assessed. Depending on the situation, studies may reveal that the diagnostic information of any subsequent test is already supplied by the simpler previous test results. When regarded in isolation such subsequent test or marker may indeed show diagnostic accuracy or value, but when assessed in the overall diagnostic workup, it does not. Such a case can arise because different tests may gauge the same underlying pathologic processes to varying degrees and thus provide related diagnostic information. From a statistical point of view, the various test values, whether obtained from patient history, physical signs, or subsequent testing, are to varying degrees mutually correlated.<sup>127,128,130</sup> The main point in diagnostic accuracy assessment, therefore is not what the diagnostic accuracy of a particular (laboratory) test is, as covered in the previous section, but rather whether it is going to improve the diagnostic accuracy of the existing setup beyond what is present from the already acquired diagnostic information.

In the following, we focus on the extent to which a certain laboratory test adds information to test results that have already been obtained. How much the new test adds in terms of improved discrimination between the presence or absence of the target disease in relation to a reference standard is of interest in this section.

### Clinical Example: Added Value of d-Dimer Testing in the Diagnosis of Suspected Deep Venous Thrombosis

The concept of assessing the added value of a subsequent diagnostic test will be illustrated by the same DVT case study

## BOX 2.2 STARD 2015

### An Updated List of Essential Items for Reporting Diagnostic Accuracy Studies

#### Title or Abstract

1. Identification as a study of diagnostic accuracy using at least one measure of accuracy (e.g., sensitivity, specificity, predictive values, 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 (e.g., symptoms, results from previous tests, 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 or 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 (e.g., 95% CIs)
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

AUC, Area under the curve; CI, confidence interval; STARD, Standards for Reporting of Diagnostic Accuracy Studies.

From Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD Group. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. Clin Chem 2015;61(12):1446–52.

described earlier.<sup>106,107</sup> In short, 2086 patients were suspected of DVT, having at least one of the following symptoms: swelling, redness, or pain in the leg. All patients had a standardized diagnostic workup consisting of index tests from medical history taking, physical examination, and quantitative D-dimer testing. The reference standard was repeated compression ultrasonography, according to current clinical practice. This reference test was carried out in all patients independent of the results of the index tests and blinded with regard to all preceding collected index test results. In total, 416 of the 2086

included patients (20%) had DVT confirmed by ultrasonography. In this section, we focus on estimating the added value of D-dimer testing to the information provided by history taking and physical examination (Table 2.13).

Table 2.13 displays the relationship between each diagnostic test result and the presence or absence of DVT. The values in fact correspond to single-test accuracy values, as discussed in the preceding section. It would be difficult, if not impossible, to select the most promising index tests from these single-test accuracy values. None of the history and physical

**TABLE 2.13 Distribution and Accuracy of Each Diagnostic Variable Compared With the Reference Standard Outcome (Deep Venous Thrombosis Present or Absent Based on Repeated Compression Ultrasonography)**

	DEEP VENOUS THROMBOSIS					
	Yes (n = 416)		No (n = 1670)		NPV (%) (95% CI)	ROC Area <sup>a</sup> (95% CI)
	n	Sens (%) (95% CI)	PPV (%) (95% CI)	n	Spec (%) (95% CI)	
Male gender	194	47 (42–51)	25 (22–29)	569	66 (64–68)	83 (81–85)
Mean age in years (SD)	62 (17)	—	—	59 (18)	—	—
Presence of malignancy	40	10 (7–13)	35 (27–44)	75	96 (94–96)	81 (79–83)
Recent surgery	76	18 (15–22)	27 (22–33)	202	88 (86–89)	81 (79–83)
Absence of recent leg trauma	47	89 (85–91)	21 (19–23)	297	18 (16–20)	86 (82–90)
Vein distension	115	28 (24–32)	28 (24–32)	302	82 (80–84)	82 (80–84)
Pain on walking	344	83 (79–86)	21 (19–23)	1325	21 (19–23)	83 (79–86)
Swelling whole leg	247	59 (55–64)	26 (23–29)	699	58 (56–60)	85 (83–87)
Mean difference in calf circumference in cm (SD)	3 (2)	—	—	2 (2)	—	—
Mean d-dimer in ng/mL (SD)	4549 (2665)	—	—	1424 (1791)	—	0.86 (0.84–0.88)

<sup>a</sup>A receiving operator characteristic (ROC) area lower than 0.5 means that overall this test result was better for excluding than including deep venous thrombosis (DVT) presence.

NPV, Negative predictive value, the proportion of subjects labeled no DVT by the diagnostic test with true absence of DVT; PPV, positive predictive value, the proportion of subjects labeled DVT by the diagnostic test with true DVT; Sens, sensitivity, the proportion of subjects with true DVT who are labeled as DVT by the diagnostic test; Spec, specificity, the proportion of subjects with true absence of DVT who are labeled as no DVT by the diagnostic test.

examination tests was pathognomonic for DVT. Some tests or investigations had a high sensitivity but a low specificity (e.g., absence of leg trauma and pain on walking), but other tests exhibited a high specificity and low sensitivity (e.g., presence of malignancy or recent surgery). Some tests would serve better for exclusion, others for inclusion. The ROC areas for the continuous tests, age and difference in calf circumference (but also for the d-dimer test), were all below 1 and above 0.5. One questions whether combinations of history and physical examination test results have better accuracy compared with their individual accuracy values and whether the d-dimer biomarker has incremental accuracy.

A multivariable statistical approach is needed to assess the diagnostic accuracy of combined index test results. Given a dichotomous outcome (DVT present or not), multivariable logistic regression modeling is the most appropriate approach. Logistic regression models express the probability of DVT (on the logit scale) as a linear function of the included index test results. Note that index test results may be included as binary, categorical, or even continuous results. The latter two do by no means need to be dichotomized first. Indeed, this is even contraindicated because it may often lose diagnostic value to the index test. Table 2.14 (model 1) shows the results from history and physical examination test results that were significantly related to DVT in the multivariable analysis, here defined as a multivariable odds ratio significantly ( $P < .05$ ) different from 1 (no association).

To quantify whether the quantitative d-dimer assay value has added diagnostic value beyond the history and physical examination results combined, the basic model 1 was simply extended by including the index test d-dimer value, resulting in model 2 (see Table 2.14). After the inclusion of

the d-dimer assay result, the regression coefficients of most history and physical tests in model 2 are found to be different from those in model 1: They now express the contribution of the corresponding test results, given a specific d-dimer result. This change reveals that the history and physical and the d-dimer results are indeed correlated and partly provide the same diagnostic information regarding whether DVT is present or not. The trend of lower regression coefficients of most findings can be interpreted as follows: A portion of the information supplied by the history and physical items is now replaced by the d-dimer assay result. Notice that the influence of the variable, recent surgery, has completely disappeared after the addition of the d-dimer biomarker.

### Diagnostic Accuracy of Combinations of Diagnostic Tests: Receiver Operating Characteristic Area

The multivariable diagnostic model, which is based on a combination of diagnostic index tests, as exemplified in models 1 and 2 in Table 2.14, can actually be considered as a single (overall or combined) quantitative index test, consisting of a composite of individual index tests. The test result of this “combined index test model” for each study patient is simply the calculated posterior probability of DVT presence given the observed pattern of the individual index test results in that patient. (See the footnote to Table 2.14 on how to calculate this probability of disease presence.) Note that this posterior probability is now the probability of DVT based on combination of multiple index test results rather than of a single index test result.

As for single continuous index tests described earlier, also for “test combinations” combined into a single multivariable model, one can calculate the ROC area ( $c$ -statistic) to indicate

**TABLE 2.14 The Basic and Extended Multivariable Diagnostic Model to Discriminate Between Deep Venous Thrombosis Presence versus Absence<sup>a</sup>**

	MODEL 1 (BASIC MODEL)			MODEL 2 (BASIC MODEL + D-DIMER)		
	Regression Coefficient (SE)	OR (95% CI)	P	Regression Coefficient (SE)	OR (95% CI)	P
(Intercept)	-3.70 (0.26)	—	<.01	-4.94 (0.32)	—	<.01
Presence of malignancy	0.62 (0.22)	1.9 (1.2–2.9)	<.01	0.22 (0.26)	1.2 (0.7–2.1)	.41
Recent surgery	0.44 (0.16)	1.6 (1.1–2.1)	<.01	0.003 (0.19)	1.0 (0.7–1.5)	.99
Absence of leg trauma	0.75 (0.18)	2.1 (1.5–3.0)	<.01	0.67 (0.20)	2.0 (1.3–2.9)	<.01
Vein distension	0.48 (0.13)	1.6 (1.1–2.1)	<.01	0.25 (0.16)	1.3 (0.9–1.8)	.12
Pain on walking	0.41 (0.15)	1.5 (1.1–2.0)	<.01	0.46 (0.18)	1.6 (1.1–2.3)	.01
Swelling whole leg	0.36 (0.12)	1.4 (1.1–1.8)	<.01	0.47 (0.14)	1.6 (1.2–2.1)	<.01
Difference in calf circumference (per cm)	0.36 (0.04)	1.4 (1.3–1.5)	<.01	0.29 (0.04)	1.3 (1.2–1.4)	<.01
d-Dimer (per 500 ng/mL)	NA	NA	NA	0.29 (0.02)	1.3 (1.3–1.4)	<.01

<sup>a</sup>Exp(regression coefficient) is the odds ratio (OR) of a diagnostic test result. For example, an odds ratio of 2 for absence leg trauma (model 2) means that a suspected patient without a recent leg trauma has a two times higher chance of having deep venous thrombosis (DVT) than a patient with a recent leg trauma (because in the latter the leg trauma would more likely be the cause of the presenting symptoms and signs). Similarly, an odds ratio of 1.3 for calf difference in cm (model 2) means that for every centimeter increase in calf circumference difference, a patient has a 1.3 times (or 30%) higher chance of having DVT.

A diagnostic model can be considered as a single overall or combined test consisting of different test results, with the probability of DVT presence as its test result. For example, for a male subject without malignancy, recent surgery, or leg trauma but with vein distension and a painful not swollen leg when walking with a calf difference of 6 cm the formula is (model1):

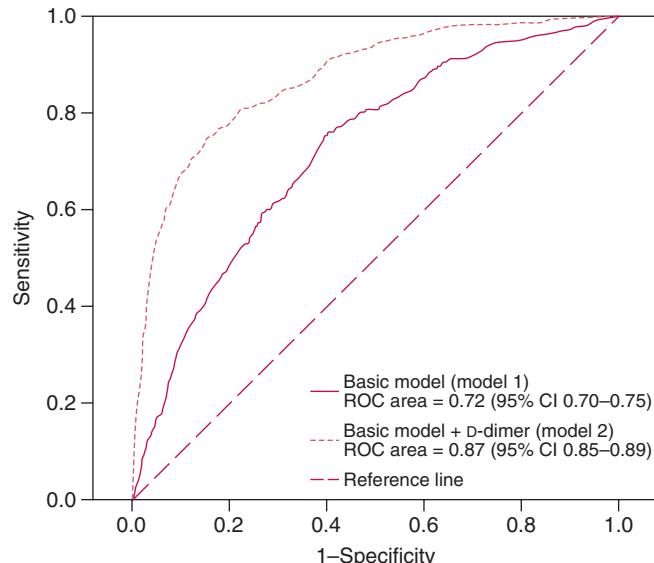
$$Z = -3.70 + 0.62*0 + 0.44*0 + 0.75*0 + 0.48*1 + 0.41*1 + 0.36*0 + 0.36*6 = -0.65.$$

The probability for this patient of the presence of DVT based on the basic model then is  $\exp(-0.65)/[1 + \exp(-0.65)] = 34\%$ .

the ability of this “test combination” to discriminate between the presence versus absence of the target disease (here DVT).<sup>115,138</sup> Here the ROC area expresses the proportion out of all possible pairs of patients with and without DVT for which the patient with DVT has a higher estimated probability by the model than the patient without DVT. Fig. 2.40 shows the ROC curves and areas for models 1 and 2, which is not much different from the comparison of two continuous index tests in isolation described earlier except that in this section, all tests are not considered in isolation but in combination or within the diagnostic pathway.

Fig. 2.40 displays how adding the quantitative d-dimer assay to model 1 mediated an increase in the ROC area from 0.72 to 0.87, a considerable and statistically significant gain ( $P < .01$ ),<sup>115,116</sup> which can be estimated using the same method described earlier for comparing two quantitative index tests in isolation. This implies that the overall diagnostic accuracy of the information from patient history and physical examination can be improved substantially by addition of the d-dimer test.

The use of the difference in ROC area to express the added value of a new test or biomarker has been subject to criticism.<sup>139–141</sup> First, the area under the receiver operating characteristic curve (AUC) is a summary measure of discrimination and has no direct clinical implication in terms of correct or incorrect diagnostic classifications or absolute patient numbers. Various investigators have noticed that the increase in AUC commonly may be relatively small by adding new but still relevant biomarkers, particularly when the AUC of the baseline model already is large.<sup>140–143</sup> Several alternative measures have been suggested to quantify the added value of a novel test or biomarker to circumvent these limitations.



**FIGURE 2.40** Receiver operating characteristic (ROC) curves for the combination of history and physical examination tests before and after addition of the d-dimer assay result. *CI*, Confidence interval. (From Moons KG, de Groot JA, Linnet K, Reitsma JB, Bossuyt PMM. Quantifying the added value of a diagnostic test or marker. *Clin Chem* 2012;58:1408–17.)

### Reclassification Measures

To handle the problems associated with the difference in ROC areas, reclassification analysis has been proposed.<sup>144</sup> The reclassification table displays how many patients actually are regrouped by adding a new test to (a combination of) existing tests after defining a threshold for a given posterior probability for presence of disease. The reclassification table with

**TABLE 2.15 Reclassification Table From The Basic and Extended (With d-Dimer) Model at an Arbitrary Deep Venous Thrombosis Probability Threshold of 25%<sup>a</sup>**

<b>DVT Yes (n = 416)</b>		<b>Model 2 with d-Dimer</b>		
		≤25%	>25%	Total
Model 1 without d-dimer	≤25%	92	123	215
	>25%	26	175	201
	Total	118	298	416
<b>DVT No (n = 1670)</b>		<b>Model 2 with d-Dimer</b>		
		≤25%	>25%	Total
Model 1 without d-dimer	≤25%	1223	116	1339
	>25%	227	104	331
	Total	1450	220	1670

<sup>a</sup>A patient with a model's probability of greater than 25% is considered high probability of having deep venous thrombosis (DVT) and is further worked up or managed for DVT.

a threshold at 25% is shown in **Table 2.15**. This means, in this theoretical example, that patients with a calculated posterior probability of 25% or higher are considered to be at high risk of having DVT and need to be referred for reference testing, but those of less than 25% receive no reference testing.

**Table 2.15** displays the reclassification of patients according to model 2 instead of model 1 at a DVT probability threshold of 25%. For example, in patients with DVT, 36% (123/416 + 26/416) were reclassified by model 2 compared with model 1. For patients without DVT, this percentage was 21% (227/1670 + 116/1670).

The simple change in classification of individuals to different posterior probability categories of DVT presence, however, is not satisfactory for assessing improvement in diagnostic accuracy by a new test or biomarker; the changes should also be in the right direction. Otherwise, an increase in posterior probability categories for subjects with DVT implies improved diagnostic classification, and any movement in the other direction implies worse diagnostic categorization. The picture is opposite for individuals without DVT present.<sup>145</sup>

The overall improvement in diagnostic reclassification can be expressed in various ways depending on the selected denominators, but commonly it is quantified as the difference between two differences. This is done by first calculating the difference between the proportions of individuals moving up and the proportion of subjects moving down for those with DVT, computing the corresponding difference in proportions for those without DVT, and then taking the difference of these two differences. This measure has been proposed as the net reclassification improvement (NRI).<sup>146</sup> The NRI is thus estimated as follows:

$$\text{NRI} = [P(\text{up}|D=1) - P(\text{down}|D=1)] - [P(\text{up}|D=0) - P(\text{down}|D=0)]$$

where  $P$  is the proportion of patients, upward movement (up) is defined as a change into a higher probability of disease presence category based on model 2, and downward movement (down) is defined as a change in the opposite direction. D denotes the disease classification (in this case, DVT), present (1) or absent (0).

The NRI results for addition of d-dimer assay to the combination of history and physical examination using the numbers displayed in **Table 2.15** were  $(0.30 - 0.06) - (0.07 - 0.14) = 0.31$  (95% CI, 0.24 to 0.36). For 123 of 416 (i.e., 0.30) of patients who experienced DVT events, classification improved with the model with d-dimer, and for 26 of 416 (0.06) people, it became worse, resulting in a net gain in reclassification proportion of 0.24. In subjects who did not have DVT, 116 of 1670 (0.07) individuals were reclassified worse by the model with the d-dimer, and 227 of 1670 (0.14) were reclassified better, resulting in a net gain in reclassification proportion of 0.07. The overall net gain in reclassification proportion therefore was  $0.24 + 0.07 = 0.31$ . This estimate was significantly different from 0 ( $P < .001$ ). The 95% CI around the NRI estimate was computed as suggested by Pencina and colleagues.<sup>146</sup>

Most investigators use three or four categories. But the NRI is clearly highly dependent on what probability threshold(s) are selected. Different thresholds may result in very different NRIs for the same added test result. To circumvent this problem of arbitrary cutoff choices, another possibility is to compute the so-called integrated discrimination improvement (IDI), which determines the magnitude of the reclassification probability improvements or deteriorations by a new test or biomarker over all possible categorizations or probability thresholds.<sup>145–147</sup>

The IDI is calculated as follows:

$$\text{IDI} = [(P_{\text{extended}}|D=1) - (P_{\text{basic}}|D=1)] - [(P_{\text{extended}}|D=0) - (P_{\text{basic}}|D=0)]$$

In this equation,  $P_{\text{extended}}|D=1$  and  $P_{\text{extended}}|D=0$  are the means of the predicted DVT probability by the extended model 2 (see **Table 2.14**) for, respectively, the patients with DVT and the patients without DVT, and  $P_{\text{basic}}|D=1$  and  $P_{\text{basic}}|D=0$  are the means of the predicted DVT probability by model 1 (see **Table 2.14**) for, respectively, the patients with DVT and the patients without DVT. The 95% CI around the NRI estimate again was calculated as outlined by Pencina and colleagues.<sup>146</sup>

The IDI for the DVT example was  $(0.49 - 0.13) - (0.28 - 0.18) = 0.26$  (95% CI, 0.23 to 0.28).

This implies that adding d-dimer to history and physical examination increased the difference in mean predicted probability between patients with DVT and patients without DVT by 0.26. This can also be interpreted as corresponding to the increase in mean sensitivity given an unchanged specificity.<sup>146</sup>

Although very popular and increasingly requested in reports on added value estimations, the NRI and IDI are only measures of discrimination between disease and nondiseased, as is also the case for ROC area. They give no information about whether the diagnostic probabilities calculated with a diagnostic model are in agreement with the observed disease prevalence (i.e., whether the models' DVT probabilities are over- or underestimated compared with the observed DVT prevalence), nor do they account in any way for the

consequences of diagnostic misclassifications when a diagnostic biomarker or test is added.<sup>148,149</sup> The following methods better address these issues.

### Predictiveness Curve

The predictiveness curve<sup>147,150</sup> is a graphic outline of the distribution of the predictive disease probabilities. Accordingly, the predictive probabilities of model 1 (without d-dimer) are ordered from lowest to highest and then plotted (Fig. 2.41).

The x-axis delineates the cumulative percentage over all individuals in the study; the y-axis shows the probabilities according to model 1. Looking first at the results only for model 1, we observe for the DVT example that if individuals who have a posterior risk (after history and physical examination) of more than 25% are selected for further investigation (regarded as positive), then 74% of patients will actually be negative and 26% will be positive (vertical dividing line in Fig. 2.41).

The four areas defined by the vertical dividing line represent, respectively, the TNs (64%), FPs (16%), FNs (10%), and TPs (10%).

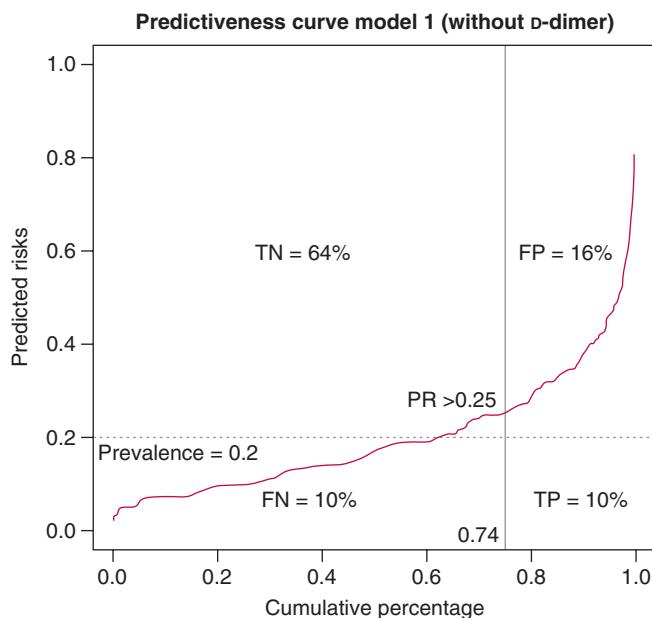
In this example (threshold of >25%), the sensitivity becomes  $TP/\text{Prevalence} \times 100 = 0.10/0.2 \times 100 = 50\%$ .

The specificity becomes

$$TN/(1 - \text{Prevalence}) \times 100 = 0.64/0.8 \times 100 = 80\%$$

The graph thus displays the estimated probabilities associated with the history and physical examination mode when applied to the source population from which the study patients theoretically originated.<sup>151</sup>

The graph may also be used to assess the two different diagnostic models and, accordingly, the added value of the d-dimer test for correct estimation of the probability of DVT



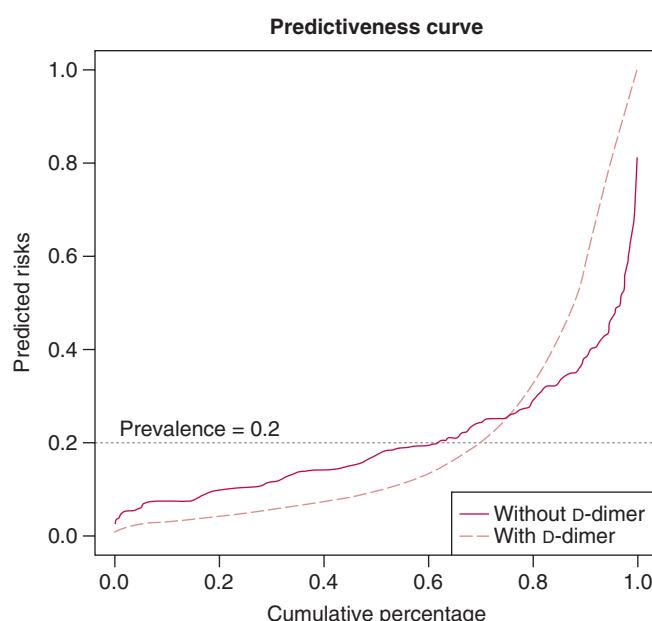
**FIGURE 2.41** Predictiveness curve for model 1 (without d-dimer) showing the distribution of positive and negative patients at a posterior risk (PR) cutoff of 0.25. FN, False-negative; FP, false-positive; TN, true negatives; TP, true positives. (From Moons KG, de Groot JA, Linnet K, Reitsma JB, Bossuyt PMM. Quantifying the added value of a diagnostic test or marker. *Clin Chem* 2012;58:1408–17.)

presence. The predictiveness curve for model 1 is substantially inferior to that of the more comprehensive model including the d-dimer results (Fig. 2.42). For example, if we set less than 0.1 as a cutoff for low risk and greater than 0.4 as a threshold for high risk (on the y-axis), we observe that  $90 - 20 = 70\%$  of the predictions of model 1 are in the equivocal zone between these thresholds, but only  $85 - 50 = 35\%$  fall between these thresholds for the predictions of model 2. Thus model 2 performs much better with regard to classifying patients into low ( $<0.1$ ) versus high ( $>0.4$ ) risk, as can be directly seen from the difference in steepness or slope of the predictiveness curve of the model with d-dimer (steeper) as opposed to the model without.<sup>151</sup>

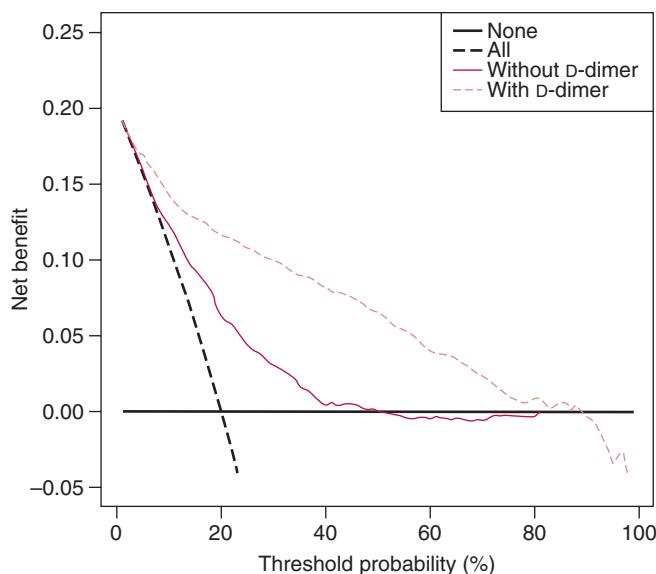
### Decision Curve Analysis

Decision curve analysis, according to Vickers and Elkin,<sup>152</sup> is a procedure that focuses on the explicit quantification of the clinical usefulness of a new index test when added to established ones in the intended clinical context. As opposed to the NRI, which is based on a single predefined probability threshold, this approach allows each professional (or even patient) to select his or her individual threshold to determine whether to take further steps such as referral for supplemental diagnostic investigations or for treatment initiation in the context of the intended use of the index test or index model. Accordingly, the corresponding net benefits can be considered without explicitly assigning weights or utilities to the wrong diagnostic classifications.

As displayed in Fig. 2.43, a posterior probability threshold of 50% would indicate that an incorrect referral (FP) is equivalent in consequences to a missed thrombosis (FN). To reduce the risk, the physician or patient might prefer reference testing or further management at a lower posterior probability threshold. This would be of particular relevance



**FIGURE 2.42** Comparison of predictiveness curves for the two models of Table 2.16 with and without d-dimer. (From Moons KG, de Groot JA, Linnet K, Reitsma JB, Bossuyt PMM. Quantifying the added value of a diagnostic test or marker. *Clin Chem* 2012;58:1408–17.)



**FIGURE 2.43** Decision curve analysis showing the net benefit of referring none of the patients for reference testing, referring all patients for reference testing, the basic prediction model, and the extended prediction model, in relation to the selected probability threshold for referral. (From Moons KG, de Groot JA, Linnet K, Reitsma JB, Bossuyt PMM. Quantifying the added value of a diagnostic test or marker. *Clin Chem* 2012;58:1408–17.)

if the further testing were simple, noninvasive, or inexpensive or if the given treatment had relatively low risk of adverse reactions. In such a situation, the risk of DVT could be, for example, 20%. Such a lower cutoff for referral would lead to acceptance of a larger percentage of incorrect referrals (FPs) rather than missing a diseased DVT subject (FN) (i.e., implicitly a higher weight is assigned to FN cases than to FP cases). On the contrary, another one might pay more attention to the costs or burden of further testing or initiation of treatment. This might be the case if the subsequent test was very invasive or complicated or therapy implied a heavy risk of adverse reactions. In this situation, a higher probability threshold of, for example, 70% might be relevant, implying a higher weight to incorrect referral of patients without the disease (FPs) and less weight to missed cases (FNs) with disease (DVT).

The graph displays the whole range of probability thresholds for further management on the x-axis and the net benefit of the diagnostic strategies or models on the y-axis.<sup>153</sup> To compute the net benefit, the proportion of all patients who are FP is subtracted from the proportion of all patients who are TP, weighting by the relative cost of an FP and an FN classification.<sup>152</sup> A numerical example shows the relations.

Table 2.16, presenting results for this empirical study on DVT, shows for the above-mentioned threshold of 20% that the TP count for model 1 was 263, and the FP count was 528. The total number of patients ( $n$ ) was 2086. The net benefit for model 1 at the threshold of 20% was  $(263/2086) - (528/2086) \times (0.2/0.8) = 0.06$ . The net benefit ratio  $0.2/0.8$  explicitly reveals that less weight is now assigned to the FPs compared with the FNs, as considered earlier. For model 2, the net benefit at the cutoff of 20% was  $(319/2086) - (301/2086) \times (0.2/0.8) = 0.12$ , two times larger. Although model 2 outperforms model 1, it should be noted that when the extended model is used at this threshold, 97 of 413 known

**TABLE 2.16 The Relationship Between True Deep Venous Thrombosis Status and the Result of the Basic and Extended Prediction Model With Thresholds of 20% and 70% Predicted Probability**

	DVT ( $n = 2086$ )	
	Present	Absent
Basic model (model 1):	Yes	263 528
probability of DVT $\geq 20\%$	No	153 1142
Extended model (model 2):	Yes	319 301
probability of DVT $\geq 20\%$	No	97 1369
Basic model (model 1):	Yes	3 6
probability of DVT $\geq 70\%$	No	413 1664
Extended model (model 2):	Yes	123 31
probability of DVT $\geq 70\%$	No	293 1639

DVT, Deep venous thrombosis.

cases would be missed and thus not treated or referred for additional testing. This implies that the overall diagnostic performance of our shown model 2 is relatively poor for this theoretically selected threshold.

At the threshold of 70%, the net benefit of model 1 was  $(3/2086) - (6/2086) \times (0.7/0.3) = -0.01$ , and for model 2, it was  $(123/2086) - (31/2086) \times (0.7/0.3) = 0.02$ .

The net benefit of model 1 of 0.06 for a threshold of 20% can be expressed as: "Compared with the case of no referrals, referral by model 1 is the equivalent of a strategy that correctly refers 6 patients with DVT of 100 suspected patients without having any unnecessary (i.e., FP) referrals."

The important point of the decision curve (see Fig. 2.43) is to observe which diagnostic strategy provides the best net benefit given the doctor's or patient's individual choice for a probability cutoff. The horizontal black line along the x-axis in Fig. 2.43 presupposes that no patients will be referred to reference testing. Because this strategy refers 0 patients, the net benefit of this strategy becomes 0 (i.e., corresponding to incorrect handling of all patients with DVT). The grey steep declining line in Fig. 2.43 shows the net benefit of the strategy of simply subjecting all individuals to reference testing. This line intersects the x-axis at the threshold probability of 20% (i.e., the prevalence in the study). Accordingly, model 2 has the greatest net benefit (i.e., it is the highest line) for all threshold probabilities. Thus we can conclude that, irrespective of the applied probability thresholds, the extended model with D-dimer added is better than the basic model.

### POINTS TO REMEMBER

- Focus has been on approaches and measures for quantification of the diagnostic accuracy of combinations of index tests and of the added value of a new diagnostic test beyond existing diagnostic tests.
- Reporting on the increase in discrimination and in (re)classification is important to gain insight into the full clinical value of a biomarker.
- Decision-curve analysis implicitly accounts for the consequences of the FP and FN classifications, complementing the information from the other measures.

## Test Evaluation Beyond Diagnostic Accuracy

Despite the methods we have presented in this chapter for the evaluation of the analytical performance and diagnostic accuracy of laboratory tests, increasingly more information is demanded about the actual impact laboratory tests and test results have on medical decision making and indeed on patient outcomes.<sup>97,137,154–161</sup> For instance, government health policymakers and private insurers in the United States now want to see empirical evidence that testing quantifiably improves actual patient outcomes in relevant patient populations or that it enhances healthcare quality, efficiency, and cost-effectiveness before recommending diagnostic tests and markers for clinical use and before deciding on their reimbursement.<sup>162</sup> In Europe, a visionary document was recently issued stressing the importance of evaluating medical technology, including diagnostic devices, on their ability to actually improve medical care and patient outcomes.<sup>163</sup> When making decisions and considering recommendations about diagnostic tests, clinicians and other decision makers have to consider the (cost-)effectiveness of the test use.

Restricting ourselves to the above explained traditional diagnostic accuracy study designs and statistical measures (e.g., sensitivity, specificity, predictive values, ROC curves), we cannot easily infer on a test's actual impact on patient health or healthcare. Addressing these challenges ideally requires comparative studies, wherein the use of a certain (new) test is examined in the clinical context compared with the current best alternatives of care, such as other form(s) of testing or no testing at all. Downstream effects of testing on clinical decision making and patient care and patients' health outcomes can be compared between both strategies. Furthermore, such studies allow for in-depth examination whether the diagnostic test improves patient outcomes and healthcare quality, efficiency, and cost-effectiveness at large.<sup>97,137,154–159,161</sup> The terms *clinical utility* or *clinical effectiveness* and *impact* of a diagnostic test are often used to express the extent to which diagnostic testing or diagnostic test results improve decision making, patient outcomes, or (cost-)effectiveness of healthcare. A more detailed discussion of clinical effectiveness and the overall impact of diagnostic tests is provided in Chapter 10 on evidence-based laboratory medicine.

## How Does Testing Yield Health(care) Benefits?

Tests, including laboratory tests that are diagnostic, prognostic, monitoring, or screening tests, do not by themselves alleviate diseases, symptoms, or signs directly but rather *indirectly*.<sup>99,154,155,158,159,163</sup> A test provides information to a user (e.g., healthcare professional or patient), which in turn indicates subsequent actions or interventions, such as therapies or lifestyle changes. For example, a test provides information (test results) that can be used to better identify patients who will and who will not benefit from helpful downstream management actions, such as administration of effective interventions or actions in individuals with certain (positive) test results and alternative or no treatment for those with other (negative) results. These interventions or actions, if beneficial, yield benefits in terms of improved health outcomes of individuals or patients. Diagnostic tests may affect patient outcomes and healthcare cost-effectiveness by improving the selection of the most effective treatment modalities or methods; examples include companion diagnostics, molecular imaging devices, and imaging devices to

guide surgery. Moreover, a certain (new) diagnostic test may be beneficial because it allows for less invasive or less costly detection of disorders. A new screening or monitoring test that leads to early detection makes it possible to administer the appropriate treatment at an early stage. Also, knowing certain diagnostic test results may affect the cognition, behavior, and lifestyle of individuals, which in turn may affect their health outcomes. Finally, besides intended effects (benefits) of testing, some tests may also lead to unintended (side) effects.

## The Working Pathway

When thinking about approaches to evaluate the impact of diagnostic tests on medical decision making, patient outcomes, and healthcare at large, it is useful to describe the pathways through which benefits (and risks) of using the test are likely to occur. This so-called *working pathway* provides a framework (Fig. 2.44) to explain how a given test leads to benefits or risks for patients' health or healthcare. Such working pathways include:

1. The anticipated technical or analytical capabilities of the test
2. The unintended and intended results and effects of the test when applied in the targeted context
3. In whom these effects are likely to occur (e.g., in the targeted patients or in the care providers)
4. The anticipated mechanisms through which these potential effects will occur
5. Existing care in the targeted context and individuals
6. The expected time frame in which potential risks and benefits might occur

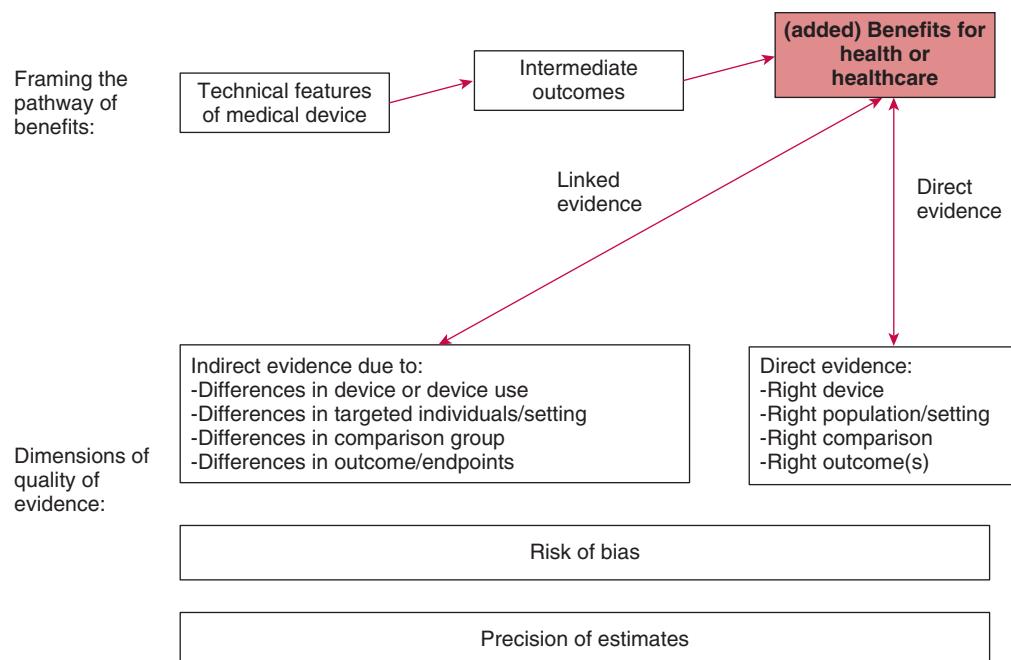
A clear description of the working pathway of a new test can determine the current benefits (and risks) of prevailing care in the intended medical context. It also helps determine what added value or benefits the new test must provide to improve existing care and what evidence is necessary to quantify whether these (added) benefits are indeed achieved at what risks or costs.

Having a detailed description of the working pathway at an early stage is particularly useful for invasive and costly (new) tests. A detailed description of the working pathway can also be used when evidence is interpreted from different studies (e.g., technical, safety, and clinical studies). For instance, if analytical performance studies fail to provide evidence for the intended technical capabilities of a test, further studies are unnecessary. If safety and analytical performance studies of a new version (modification) of an existing test show that its safety and analytical performance is similar to the preceding version but that it is less burdensome or cheaper to use, then subsequent studies may not be needed.

## Comparative Tests: Treatment Studies to Quantify the Impact of Tests

One could design a longitudinal study to validly quantify whether a certain (new) test has impact on patient's health or healthcare beyond what is achieved by current practice.<sup>156,157,161,163–168</sup> This is in fact a similar approach used to evaluate the *effectiveness* (not the efficacy) of medical drugs: so-called comparative or pragmatic randomized trial. The study should:

- Investigate the (new) test in the same targeted individuals (e.g., professionals and patients) as it is intended to be used in practice.



**FIGURE 2.44** Relationship between the pathway through which devices may lead to benefits or added benefits for health or healthcare and the three dimensions of quality for evidence (indirectness of evidence, risk of bias, precision of estimates). (From KNAW. Evaluation of new technology in health care. In need of guidance for relevant evidence. Amsterdam: KNAW, 2014.)

- Investigate the test in the clinical pathway in which it is intended to be used in practice.
- Study the use of the test in combination with any subsequent management (e.g., therapeutic) actions indicated by the test or its results.
- Compare the test and subsequent management actions to the best prevailing care. Ideally, this would be a randomized comparison: Individuals are allocated randomly to either the new test use or to the comparative strategy.
- Measure all outcomes or endpoints relevant for the targeted individuals, professionals, and ideally for society at large, including unintended outcomes, intended health outcomes (for patient and users), burden and ease of test use, speed of administering subsequent management, and even costs of test use.
- Be sufficiently long to investigate the long-term healthcare effects.
- Be sufficiently large to obtain precise estimates of the safety and health benefits of the test use.

The two comparison groups are thus created randomly. In the index group, the new test is used in combination with subsequent management or therapeutic actions, with the prevailing tests and management being applied in the comparison (control) group. Provided the two groups are large enough, any observed differences between the two groups in terms of benefits (and risks) can then be ascribed to the difference in tests plus subsequent management. Such randomized studies compare the use of the index test or tests combined with subsequent actions, directly with the best alternative strategy in the right population, measuring all relevant outcomes (for patients, users, and healthcare) in the short and long terms. Such studies thus generate the most *direct* and *valid evidence* as to whether

the test use will indeed produce the intended relevant benefits at an acceptable level of risks and costs compared with prevailing care.

This randomized comparative study approach is not always feasible or possible.<sup>156,157,161,168</sup> Numerous alternative comparative study approaches may also produce (direct) evidence of the (added) benefits of test use for the relevant health outcomes in the intended medical context. These approaches may include alternative randomized designs as described.<sup>169,170</sup> Furthermore, there are also many nonrandomized comparative study approaches, ranging from quasi-randomized studies to controlled before-and-after studies and comparative cohort or even case-control studies.<sup>133,171,172</sup> These nonrandomized comparative studies are more prone to bias because of differences in demographic and clinical characteristics between the two groups being compared. Fortunately, there are various approaches to controlling or adjusting for such biases, for which we refer to the literature.<sup>133,171,172</sup>

### Linked-Evidence Approaches to Quantify the Impact of Tests

Besides technical performance and cross-sectional diagnostic test accuracy studies described earlier, many clinical studies of (new) laboratory tests focus on measuring intermediate effects or outcomes along the working pathway for a given test. Each of these studies by themselves often do not allow for inferring on the desired longer-term (added) benefits and risks of the test use. However, it is possible to use the so-called quantitative linked-evidence approach in which the evidence of these different types of diagnostic test studies are quantitatively combined to estimate a test's effect on relevant health or healthcare outcomes.<sup>158,159,173–177</sup>

An example is the assessment of the analytical performance requirements for glucose monitoring devices to achieve clinically useful glucose control. Rather than conducting large-scale longitudinal comparative, expensive randomized studies using many different measuring devices as comparators, computers using a variety of underlying modeling schemes have been used to generate simulated glucose results from glucose measurement devices having varying amounts of bias and imprecision. The patient data used in these studies were derived from physiologic models of glucose metabolism or hospitalized patients.<sup>178</sup> The results of such linked-evidence modeling approaches can be used to evaluate the effects of measurement bias and imprecision on quantifiable intermediate results such as the percentage of the time glucose falls within the desired therapeutic range, the frequency and duration of hypoglycemic episodes, and the within-patient variability of glucose. Such intermediate results are known to have direct relationships with the rates of clinical complications in individuals with poor glucose control. For the evaluation of the impact or utility of (new) laboratory tests on patient and healthcare outcomes in the intended medical context, linked-evidence approaches offer an attractive alternative when direct evaluation of a test's benefits on long-term, patient-relevant outcomes is difficult

or impossible. For the glucose example, it would be difficult to gain ethical approval for a randomized trial in which yet imprecise glucose analyzers were evaluated for their effects when used for patient glucose monitoring. The validity of a linked-evidence approach is dependent on how predictive the existing study results and evidence are and the relation of these intermediate outcomes with the long-term, relevant health or healthcare outcomes.

Linked-evidence studies can be particularly useful for laboratory test studies when there is evidence from cross-sectional studies on the diagnostic accuracy of a test and results of therapeutic studies provide a link to health outcomes. Simple modeling approaches might be used to link both types of evidence and to actually quantify the benefits (and risks) of the test use on these health outcomes. Such linked-evidence models might include various sensitivity analyses, for example, to account for the risks of using various types of evidence taken from different sources.<sup>158,159,173–176</sup> For example, a so-called Markov model was used to combine evidence from analytical performance studies, cross-sectional diagnostic accuracy studies, and long-term management studies to quantify the long-term cost-effectiveness of point-of-care D-dimer tests compared with the use of central laboratory D-dimer tests to rule out DVT in primary care.<sup>179</sup>

## POINTS TO REMEMBER

- It is important to assess information concerning the actual impact or utility of the use of a diagnostic test on patient outcomes or healthcare at large.
- The impact of diagnostic tests on medical decision making and patient outcomes can be considered by describing the pathways through which benefits (and risks) of using the test are likely to occur (the so-called *working pathway*).
- *Direct and valid evidence* as to whether the new test will indeed produce the intended relevant benefits can be assessed

by randomized studies comparing the outcome of use of the new test with that of the index test.

- A supplementary approach is the so-called quantitative linked-evidence procedure in which the evidence of different types of diagnostic test studies are quantitatively combined to estimate a test's effect on relevant health or healthcare outcomes.

## SELECTED REFERENCES

5. Snedecor GW, Cochran WG. Statistical methods. 8th ed. Ames, Iowa: Iowa State University Press; 1989. p. 75, 121, 140–2, 170–4, 177, 237–8, 279.
12. Dybkaer R. Vocabulary for use in measurement procedures and description of reference materials in laboratory medicine. Eur J Clin Chem Clin Biochem 1997;35:141–73.
35. Krouwer JS. Setting performance goals and evaluating total analytical error for diagnostic assays. Clin Chem 2002; 48:919–27.
40. Linnet K. Nonparametric estimation of reference intervals by simple and bootstrap-based procedures. Clin Chem 2000;46: 867–9.
44. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;i:307–10.
49. Linnet K. Evaluation of regression procedures for methods comparison studies. Clin Chem 1993;39:424–32.
53. Linnet K. Limitations of the paired t-test for evaluation of method comparison data. Clin Chem 1999;45:314–15.
54. Linnet K. Estimation of the linear relationship between the measurements of two methods with proportional errors. Stat Med 1990;9:1463–73.
60. Passing H, Bablok W. Comparison of several regression procedures for method comparison studies and determination of sample sizes. J Clin Chem Clin Biochem 1984;22:431–45.
71. Vesper HW, Thienpont LM. Traceability in laboratory medicine. Clin Chem 2009;55:1067–75.
95. Linnet K, Bossuyt PM, Moons KG, et al. Quantifying the accuracy of a diagnostic test or marker. Clin Chem 2012;58:1292–301.
96. Moons KG, de Groot JA, Linnet K, et al. Quantifying the added value of a diagnostic test or marker. Clin Chem 2012;58:1408–17.
97. Bossuyt PMM, Reitsma JB, Linnet K, et al. Beyond diagnostic accuracy: the clinical utility of diagnostic tests. Clin Chem 2012;58:1636–43.
102. Bossuyt PM, Reitsma JB, Bruns DE, et al. An updated list of essential items for reporting diagnostic accuracy studies. Clin Chem 2015;in press.
106. Oudega R, Moons KG, Hoes AW. Ruling out deep venous thrombosis in primary care. A simple diagnostic algorithm including D-dimer testing. Thromb Haemost 2005;94:200–5.

118. Obuchowski NA, Lieber ML, Wians FH Jr. ROC curves in clinical chemistry: uses, misuses, and possible solutions. *Clin Chem* 2004;50:1118–25.
137. Moons KG. Criteria for scientific evaluation of novel markers: a perspective. *Clin Chem* 2010;56:537–41.
143. Pencina MJ, D'Agostino RB, Vasan RS. Statistical methods for assessment of added usefulness of new biomarkers. *Clin Chem Lab Med* 2010;48:1703–11.
177. Horvath AR, Lord SJ, StJohn A, et al. From biomarkers to medical tests: the changing landscape of test evaluation. *Clin Chim Acta* 2014;427:49–57.
179. Hendriksen JMT, Geersing GJ, van Voorthuizen SC, et al. The cost-effectiveness of point-of-care D-dimer tests compared with a laboratory test to rule out deep venous thrombosis in primary care. *Expert Rev Mol Diagn* 2015; 15:125–36.

## REFERENCES

1. Directive 98/79/EC of the European Parliament and of the Council of 27 October on in vitro diagnostic medical devices. *Off J Eur Comm* 1998;L331:1–37.
2. U.S. Department of Health and Human Services. Medicare, Medicaid, and CLIA programs: regulations implementing the Clinical Laboratory Improvement Amendments of 1988 (CLIA). Final rule. *Fed Regist* 1992;57:7002–186.
3. U.S. Department of Health and Human Services. Medicare, Medicaid, and CLIA programs: laboratory requirements relating to quality systems and certain personnel qualifications. Final rule. *Fed Regist* 2003;68:3640–714. Available as CMS-2226-F.pdf at: <<http://www.cdc.gov/clia/chronol.aspx>>; [accessed 03.11.15].
4. David HA. Order statistics. New York: Wiley; 1981. p. 80–2.
5. Snedecor GW, Cochran WG. Statistical methods. 8th ed. Ames, Iowa: Iowa State University Press; 1989. p. 75, 121, 140–2, 170–4, 177, 237–8, 279.
6. Currie LA. Nomenclature in evaluation of analytical methods including detection and quantification capabilities (IUPAC recommendations 1995). *Pure Appl Chem* 1995;67:1699–723.
7. Prichard FE, Day JA, Hardcastle WA, et al. Quality in the analytical chemistry laboratory. Chichester, United Kingdom: Wiley; 1995. p. 136–43, 169.
8. Rodbard D, McClean SW. Automated computer analysis for enzyme-multiplied immunological techniques. *Clin Chem* 1977;23:112–15.
9. Shukla GK. On the problem of calibration. *Technometrics* 1972;14:547–53.
10. Powers DM. Establishing and maintaining performance claims. *Arch Pathol Lab Med* 1992;116:718–25.
11. International Organization for Standardization (ISO). Guide 99. International vocabulary of metrology: basic and general concepts and associated terms (VIM). Geneva: ISO, 2007.
12. Dybkær R. Vocabulary for use in measurement procedures and description of reference materials in laboratory medicine. *Eur J Clin Chem Clin Biochem* 1997;35:141–73.
13. Armbruster DA, Alexander DB. Sample to sample carryover: a source of analytical laboratory error and its relevance to integrated clinical chemistry/immunoassay systems. *Clin Chim Acta* 2006;373:37–43.
14. Krouwer JS. Multifactor protocol designs IV: how multifactor designs estimate the total error by accounting for protocol-specific biases. *Clin Chem* 1991;37:26–9.
15. CLSI. Preliminary evaluation of quantitative clinical laboratory measurement procedures; approved guideline, 3rd edition. CLSI document EP10-A3-AMD. Wayne, PA: Clinical and Laboratory Standards Institute, 2014.
16. CLSI. Evaluation of precision performance of quantitative measurement methods; approved guideline, 3rd edition. CLSI document EP05-A3. Wayne, PA: Clinical and Laboratory Standards Institute, 2014.
17. CLSI. User verification of performance for precision and trueness; approved guideline, 3rd edition. CLSI document EP15-A3. Wayne, PA: Clinical and Laboratory Standards Institute, 2014.
18. Hald A. Statistical theory with engineering applications. New York: Wiley; 1952. p. 534–5, 551–7.
19. Emancipator K, Kroll MH. A quantitative measure of nonlinearity. *Clin Chem* 1993;39:766–72.
20. Draper NR, Smith H. Applied regression analysis. 3rd ed. New York: Wiley; 1998. p. 192–8.
21. CLSI. Evaluation of the linearity of quantitative measurement procedures: a statistical approach; approved guideline. CLSI document EP06-A. Wayne, PA: Clinical and Laboratory Standards Institute, 2003.
22. Shah VP, Midha KK, Findlay JWA, et al. Bioanalytical method validation: a revisit with a decade of progress. *Pharm Res* 2000;17:1551–7.
23. CLSI. Protocols for determination of limits of detection and limits of quantitation; approved guideline. CLSI document EP17-A2. Wayne, PA: Clinical and Laboratory Standards Institute, 2012.
24. Glick MR, Ryder KW. Analytical systems ranked by freedom from interferences. *Clin Chem* 1987;33:1453–8.
25. CLSI. Interference testing in clinical chemistry; approved guideline, 2nd edition. CLSI document EP07-A2. Wayne, PA: Clinical and Laboratory Standards Institute, 2005.
26. Linnet K, Brandt E. Assessing diagnostic tests once an optimal cutoff point has been selected. *Clin Chem* 1986;32:1341–6.
27. European Committee for Clinical Laboratory Standards. Guidelines for the evaluation of diagnostic kits. Part 2. General principles and outline procedures for the evaluation of kits for qualitative tests. Lund, Sweden: ECCLS, 1990.
28. CLSI. User protocol for evaluation of qualitative test performance; approved guideline, 2nd edition. CLSI document EP12-A2. Wayne, PA: Clinical and Laboratory Standards Institute, 2008.
29. Fleiss JL. Statistical methods for rates and proportions. 2nd ed. New York: Wiley; 1981 [Chapter 13].
30. Whiting P, Rutjes AW, Reitsma JB, et al. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. *Ann Intern Med* 2004;140:189–202.
31. Schmidt RL, Factor RE. Understanding sources of bias in diagnostic accuracy studies. *Arch Pathol Lab Med* 2013;137:558–65.
32. Lipman HB, Astles JR. Quantifying the bias associated with use of discrepant analysis. *Clin Chem* 1998;44:108–15.
33. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999;282:1061–6.
34. Krouwer JS. Estimating total analytical error and its sources. *Arch Pathol Lab Med* 1992;116:726–31.
35. Krouwer JS. Setting performance goals and evaluating total analytical error for diagnostic assays. *Clin Chem* 2002;48:919–27.
36. Lawton WH, Sylvester EA, Young-Ferraro BJ. Statistical comparison of multiple analytic procedures: application to clinical chemistry. *Technometrics* 1979;21:397–409.
37. International Organization for Standardization (ISO). Capability of detection. Part 2. Methodology in the linear calibration case (11843-2). Geneva: ISO, 2000.
38. Westgard JO, Hunt MR. Use and interpretation of common statistical tests in method-comparison studies. *Clin Chem* 1973;19:49–57.
39. CLSI. Method comparison and bias estimation using patient samples; approved guideline, 3rd edition. CLSI document EP09-A3. Wayne, PA: Clinical and Laboratory Standards Institute, 2013.
40. Linnet K. Nonparametric estimation of reference intervals by simple and bootstrap-based procedures. *Clin Chem* 2000;46:867–9.
41. CLSI. Defining, establishing, and verifying reference intervals in the clinical laboratory; approved guideline, 3rd edition. CLSI document EP28-A3C. Wayne, PA: Clinical and Laboratory Standards Institute, 2010.

42. Efron B. An introduction to the bootstrap. London: Chapman and Hall; 1993.
43. Linnet K. Two-stage transformation systems for normalization of reference distributions evaluated. *Clin Chem* 1987;33:381–6.
44. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;i:307–10.
45. Bland JM, Altman DG. Comparing methods of measurement: why plotting difference against standard method is misleading. *Lancet* 1995;346:1085–7.
46. Linnet K, Bruunshuus I. HPLC with enzymatic detection as a candidate reference method for serum creatinine. *Clin Chem* 1991;37:1669–75.
47. Petersen PH, Stöckl D, Blaabjerg O, et al. Graphical interpretation of analytical data from comparison of a field method with a reference method by use of difference plots. *Clin Chem* 1997;43:2039–46.
48. Pollock MA, Jefferson SG, Kane JW, et al. Method comparison: a different approach. *Ann Clin Biochem* 1992;29:556–60.
49. Linnet K. Evaluation of regression procedures for methods comparison studies. *Clin Chem* 1993;39:424–32.
50. Clarke WL, Cox D, Gonder-Frederick LA, et al. Evaluating clinical accuracy of systems for self-monitoring of blood glucose. *Diabetes Care* 1987;10:622–8.
51. Stöckl D. Beyond the myths of difference plots [letter]. *Ann Clin Biochem* 1996;36:575–7.
52. Thienpont LM, Van Nuwenborg JE, Stöckl D. Intrinsic and routine quality of serum total potassium measurement as investigated by split-sample measurement with an ion chromatography candidate reference method. *Clin Chem* 1998;44:849–57.
53. Linnet K. Limitations of the paired t-test for evaluation of method comparison data. *Clin Chem* 1999;45:314–15.
54. Linnet K. Estimation of the linear relationship between the measurements of two methods with proportional errors. *Stat Med* 1990;9:1463–73.
55. Mandel J. The statistical analysis of experimental data. New York: Wiley; 1964. p. 290–1.
56. Cornbleet PJ, Gochman N. Incorrect least-squares regression coefficients in method-comparison analysis. *Clin Chem* 1979;25:432–8.
57. Linnet K. The performance of Deming regression analysis in case of a misspecified analytical error ratio. *Clin Chem* 1998;44:1024–31.
58. Bookbinder MJ, Panosian KJ. Using the coefficient of correlation in method-comparison studies. *Clin Chem* 1987;33: 1170–6.
59. Passing H, Bablok W. A new biometrical procedure for testing the equality of measurements from two different analytical methods. *J Clin Chem Clin Biochem* 1983;21:709–20.
60. Passing H, Bablok W. Comparison of several regression procedures for method comparison studies and determination of sample sizes. *J Clin Chem Clin Biochem* 1984;22:431–45.
61. Linnet K. Necessary sample size for method comparison studies based on regression analysis. *Clin Chem* 1999;45:882–94.
62. Simundic AM, Kackov S, Miler M, et al. Terms and symbols used in studies on biological variation: the need for harmonization. *Clin Chem* 2015;61(2):438–9. doi:10.1373/clinchem.2014.233791.
63. Skendzel LP, Barnett RN, Platt R. Medically useful criteria for analytic performance of laboratory tests. *Am J Clin Pathol* 1985;83:200–5.
64. Linnet K. Choosing quality control systems to detect maximum medically allowable analytical errors. *Clin Chem* 1989;35:284–8.
65. Harris EK, Boyd J. Statistical bases of reference values in laboratory medicine. New York: Marcel Dekker; 1995. p. 238–50.
66. Tietz NW. A model for a comprehensive measurement system in clinical chemistry. *Clin Chem* 1979;25:833–9.
67. Büttner J. Reference materials and reference methods in laboratory medicine: a challenge to international cooperation. *Eur J Clin Chem Clin Biochem* 1994;32:571–7.
68. International Organization for Standardization (ISO). In vitro diagnostic medical devices. Measurement of quantities in biological samples. Metrological traceability of values assigned to calibrators and control materials (17511). Geneva: ISO, 2003.
69. International Organization for Standardization (ISO). In vitro diagnostic medical devices. Measurement of quantities in samples of biological origin. Requirements for content and presentation of reference measurement procedures (15193). Geneva: ISO, 2009.
70. Vesper HW, Miller WG, Myers GL. Reference materials and commutability. *Clin Biochem Rev* 2007;28:139–47.
71. Vesper HW, Thienpont LM. Traceability in laboratory medicine. *Clin Chem* 2009;55:1067–75.
72. CLSI. Characterization and qualification of commutable reference materials for laboratory medicine; proposed guideline. CLSI document EP30-A. Wayne, PA: Clinical and Laboratory Standards Institute, 2010.
73. International Organization for Standardization (ISO). In vitro diagnostic medical devices. Measurement of quantities in samples of biological origin. Requirements for certified reference materials and the content of supporting documentation (15194). Geneva: ISO, 2009.
74. Bais R, Armbruster D, Jansen RTP, et al. Defining acceptable limits for the metrological traceability of specific measurands. *Clin Chem Lab Med* 2013;51:973–9.
75. Sturgeon CM, Berger P, Bidart J-M, et al. Differences in recognition of the 1st WHO international reference reagents for hCG-related isoforms by diagnostic immunoassays for human chorionic gonadotropin. *Clin Chem* 2009;55:1484–91.
76. Thienpont L, Van Uytfanghe K, De Leenheer AP. Reference measurement systems in clinical chemistry. *Clin Chim Acta* 2002;323:73–87.
77. Gantzer ML, Miller WG. Harmonisation of measurement procedures: how do we get it done? *Clin Biochem Rev* 2012;33:95–100.
78. Petersen PH, Klee GG. Influence of analytical bias and imprecision on the number of false positive results using guideline-driven medical decision limits. *Clin Chim Acta* 2014;430:1–8.
79. Ellison SLR, Rosslein M, Williams A, editors. Eurachem/Citac guide: quantifying uncertainty in analytical measurement. 2nd ed. Berlin: Eurachem; 2000. p. 4, 5, 9, 17.
80. International Organization for Standardization (ISO). Guide 98-1. Uncertainty of measurement. Part 1. Introduction to the expression of uncertainty in measurement. Geneva: ISO, 2009.
81. Barwick V. Evaluating measurement uncertainty in clinical chemistry. Case studies. LGC/R/2010/17. UK National Measurement System March 2012; pp38 <[https://biowsearch-cdn.azureedge.net/assetsv6/Clinical\\_worked\\_examples\\_report\\_Final.pdf](https://biowsearch-cdn.azureedge.net/assetsv6/Clinical_worked_examples_report_Final.pdf)>.
82. National Pathology Accreditation Advisory Council Requirements for the Estimation of Measurement Uncertainty. 2007; pp37 [accessed 20.09.20] <[https://www.health.gov.au/internet/main/publishing.nsf/Content/B1074B732F32282DCA257BF001FA218/\\$File/dhaeou.pdf](https://www.health.gov.au/internet/main/publishing.nsf/Content/B1074B732F32282DCA257BF001FA218/$File/dhaeou.pdf)>.

83. Davis RB, Thompson JE, Pardue HL. Characteristics of statistical parameters used to interpret least-squares results. *Clin Chem* 1978;24:611–20.
84. Farrance I, Frenkel R. Uncertainty of measurement: a review of the rules for calculating uncertainty components through functional relationships. *Clin Biochem Rev* 2012;33:49–75.
85. Badrick T, Hawkins RC. The relationship between measurement uncertainty and reporting interval. *Ann Clin Biochem* 2015;52:177–9.
86. Inal BB, Koldas M, Inal H, et al. Evaluation of measurement uncertainty of glucose in clinical chemistry. *Ann N Y Acad Sci* 2007;1100:223–6.
87. Linko S, Örnemark U, Kessel R, et al. Evaluation of uncertainty of measurement in routine clinical chemistry: application to determination of the substance concentration of calcium and glucose in serum. *Clin Chem Lab Med* 2002;40:391–8.
88. Aronsson T, deVerdier C, Groth T. Factors influencing the quality of analytical methods: a systems analysis, with computer simulation. *Clin Chem* 1974;20:738–48.
89. Farrance I, Frenkel R. Uncertainty in measurement: a review of Monte Carlo simulation using Microsoft Excel for the calculation of uncertainties through functional relationships, including uncertainties in empirically derived constants. *Clin Biochem Rev* 2014;35:37–61.
90. Petersen PH, Stöckl D, Westgard JO, et al. Models for combining random and systematic errors: assumptions and consequences for different models. *Clin Chem Lab Med* 2001;39:589–95.
91. Taylor JR. An introduction to error analysis. Oxford: Oxford University Press; 1982.
92. Krouwer JS. Critique of the guide to the expression of uncertainty in measurement methods of estimating and reporting uncertainty in diagnostic assays. *Clin Chem* 2003;49:1818–21.
93. Kristiansen J. Description of a generally applicable model for the evaluation of uncertainty of measurement in clinical chemistry. *Clin Chem Lab Med* 2001;39:920–31.
94. Kristiansen J. The guide to expression of uncertainty in measurement approach for estimating uncertainty: an appraisal. *Clin Chem* 2003;49:1822–9.
95. Linnet K, Bossuyt PM, Moons KG, et al. Quantifying the accuracy of a diagnostic test or marker. *Clin Chem* 2012;58:1292–301.
96. Moons KG, de Groot JA, Linnet K, et al. Quantifying the added value of a diagnostic test or marker. *Clin Chem* 2012;58:1408–17.
97. Bossuyt PMM, Reitsma JB, Linnet K, et al. Beyond diagnostic accuracy: the clinical utility of diagnostic tests. *Clin Chem* 2012;58:1636–43.
98. Knottnerus JA, Frank Buntinx F, editors. The evidence base of clinical diagnosis: theory and methods of diagnostic research. 2nd ed. Oxford: BMJ Books; 2008.
99. Grobbee DE, Hoes AW. Clinical epidemiology. Jones & Bartlett; 2014.
100. Hoes AW, Grobbee DE. Chapter 3 Diagnostic research. Clinical epidemiology. Sudbury (MA): Jones & Bartlett Publishers; 2008.
101. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Clin Chem* 2003;49:1–6.
102. Bossuyt PM, Reitsma JB, Bruns DE, et al. An updated list of essential items for reporting diagnostic accuracy studies. *Clin Chem* 2015;61:1446–52.
103. Linnet K. A review on the methodology for assessing diagnostic tests. *Clin Chem* 1988;34:1379–86.
104. Brown LD, Cai TT, DasGupta A. Interval estimation for a binomial proportion. *Stat Sci* 2001;16:101–33.
105. Bachmann LM, Puhan MA, ter Riet G, et al. Sample sizes of studies on diagnostic accuracy: literature survey. *MBJ* 2006;332:1127–9.
106. Oudega R, Moons KG, Hoes AW. Ruling out deep venous thrombosis in primary care. A simple diagnostic algorithm including D-dimer testing. *Thromb Haemost* 2005;94:200–5.
107. Toll DB, Oudega R, Bulten RJ, et al. Excluding deep vein thrombosis safely in primary care. *J Fam Pract* 2006;55:613–18.
108. Rutjes AW, Reitsma JB, Di NM, et al. Evidence of bias and variation in diagnostic accuracy studies. *CMAJ* 2006;174:469–76.
109. Rutjes AW, Reitsma JB, Coomarasamy A, et al. Evaluation of diagnostic tests when there is no gold standard. A review of methods. *Health Technol Assess* 2007;11:iii, ix–51.
110. de Groot JA, Janssen KJ, Zwinderman AH, et al. Correcting for partial verification bias: a comparison of methods. *Ann Epidemiol* 2011;21:139–48.
111. de Groot JA, Bossuyt PM, Reitsma JB, et al. Verification problems in diagnostic accuracy studies: consequences and solutions. *BMJ* 2011;343:d4770.
112. de Groot JA, Janssen KJ, Zwinderman AH, et al. Correcting for partial verification bias: a comparison of methods. *Ann Epidemiol* 2011;21:139–48.
113. de Groot JA, Dendukuri N, Janssen KJ, et al. Adjusting for differential-verification bias in diagnostic-accuracy studies: a Bayesian approach. *Epidemiology* 2011;22:234–41.
114. Metz CE, Goodenough DJ, Rossmann K. Evaluation of receiver operating characteristic curve data in terms of information theory, with applications in radiography. *Radiology* 1973;109:297–303.
115. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;143:29–36.
116. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* 1983;148:839–43.
117. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin Chem* 1993;39:561–77.
118. Obuchowski NA, Lieber ML, Wians FH Jr. ROC curves in clinical chemistry: uses, misuses, and possible solutions. *Clin Chem* 2004;50:1118–25.
119. Moons K, Stijnen T, Michel BC, et al. Application of treatment thresholds to diagnostic-test evaluation: an alternative to the comparison of areas under receiver operating characteristic curves. *Med Decis Making* 1997;17:447–54.
120. Leeflang MM, Moons KG, Reitsma JB, et al. Bias in sensitivity and specificity caused by data-driven selection of optimal cut-off values: mechanism, magnitude, and solutions. *Clin Chem* 2008;54:729–37.
121. Vecchio TJ. Predictive value of a single diagnostic test in unselected populations. *NEJM* 1966;274:1171–3.
122. Albert A. On the use and computation of likelihood ratios in clinical chemistry. *Clin Chem* 1982;28:1113–19.
123. Fagan TJ. Letter: Nomogram for Bayes theorem. *NEJM* 1975;293:257.
124. Geersing G-J, Toll DB, Janssen KJM, et al. Diagnostic accuracy and user-friendliness of 5 point-of-care D-dimer tests

- for the exclusion of deep vein thrombosis. *Clin Chem* 2010;56:1758–66.
125. Altman DG. Practical statistics for medical research. 1st ed. London, UK: Chapman & Hall; 1991. p. 258.
  126. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating curves: a nonparametric approach. *Biometrics* 1988;44:837–45.
  127. Hlatky MA, Pryor DB, Harrell FE Jr, et al. Factors affecting sensitivity and specificity of exercise electrocardiography. Multivariable analysis. *Am J Med* 1984;77:64–71.
  128. Moons KG, van Es GA, Deckers JW, et al. Limitations of sensitivity, specificity, likelihood ratio, and Bayes' theorem in assessing diagnostic probabilities: a clinical example. *Epidemiology* 1997;8:12–17.
  129. Moons KG, van Es GA, Michel BC, et al. Redundancy of single diagnostic test evaluation. *Epidemiology* 1999;10:276–81.
  130. Leeflang MM, Bossuyt PM, Irwig L. Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. *J Clin Epidemiol* 2009;62:5–12.
  131. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011;155:529–36.
  132. Pepe MS. The statistical evaluation of medical tests for classification and prediction. Oxford: Oxford University Press; 2003.
  133. Sackett DL, Haynes RB, Guyatt GH, et al. Clinical epidemiology. 2nd ed. Boston (MA): Little, Brown and Company; 1991.
  134. Biesheuvel CJ, Vergouwe Y, Oudega R, et al. Advantages of the nested case-control design in diagnostic research. *BMC Med Res Methodol* 2008;8:48.
  135. Lord SJ, Staub LP, Bossuyt PM, et al. Target practice: choosing target conditions for test accuracy studies that are relevant to clinical practice. *BMJ* 2011;343:d4684.
  136. Moons KG, Grobbee DE. Diagnostic studies as multivariable, prediction research. *J Epidemiol Community Health* 2002;56:337–8.
  137. Moons KG. Criteria for scientific evaluation of novel markers: a perspective. *Clin Chem* 2010;56:537–41.
  138. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996;15:361–87.
  139. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* 2007;115:928–35.
  140. Pepe MS, Janes H, Longton G, et al. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol* 2004;159:882–90.
  141. Ware JH. The limitations of risk factors as prognostic tools. *N Engl J Med* 2006;355:2615–17.
  142. Tzoulaki I, Liberopoulos G, Ioannidis JP. Assessment of claims of improved prediction beyond the Framingham risk score. *JAMA* 2009;302:2345–52.
  143. Pencina MJ, D'Agostino RB, Vasan RS. Statistical methods for assessment of added usefulness of new biomarkers. *Clin Chem Lab Med* 2010;48:1703–11.
  144. Cook NR, Ridker PM. Advances in measuring the effect of individual predictors of cardiovascular risk: the role of reclassification measures. *Ann Intern Med* 2009;150:795–802.
  145. Steyerberg EW, Pencina MJ, Lingsma HF, et al. Assessing the incremental value of diagnostic and prognostic markers: a review and illustration. *Eur J Clin Invest* 2011;42:216–28.
  146. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, et al. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med* 2008;27:157–72.
  147. Pepe MS, Feng Z, Gu JW. Comments on 'Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond'. *Stat Med* 2008;27:173–81.
  148. Greenland S. The need for reorientation toward cost-effective prediction: comments on 'Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond' by M. J. Pencina et al. *Statistics in Medicine*. *Stat Med* 2008;27:199–206.
  149. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010;21:128–38.
  150. Huang Y, Sullivan PM, Feng Z. Evaluating the predictiveness of a continuous marker. *Biometrics* 2007;63:1181–8.
  151. Pepe MS, Feng Z, Huang Y, et al. Integrating the predictive ness of a marker with its performance as a classifier. *Am J Epidemiol* 2008;167:362–8.
  152. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* 2006;26:565–74.
  153. Steyerberg EW, Vickers AJ. Decision curve analysis: a discussion. *Med Decis Making* 2008;28:146–9.
  154. Bossuyt PM, McCaffery K. Additional patient outcomes and pathways in evaluations of testing. *Med Decis Making* 2009;29:E30–8.
  155. Ferrante di Ruffano L, Hyde CJ, McCaffery KJ, et al. Assessing the value of diagnostic tests: a framework for designing and evaluating trials. *BMJ* 2012;344:e686.
  156. Lijmer JG, Bossuyt PM. Various randomized designs can be used to evaluate medical tests. *J Clin Epidemiol* 2009;62:364–73.
  157. Lord SJ, Irwig L, Bossuyt PMM. Using the principles of randomized controlled trial design to guide test evaluation. *Med Decis Making* 2009;29:E1–12.
  158. Koffijberg H. Cost-effectiveness analysis of diagnostic tests. *Neurosurgery* 2013;73:E558–60.
  159. Koffijberg H, van Zaane B, Moons KGM. From accuracy to patient outcome and cost-effectiveness evaluations of diagnostic tests and biomarkers: an exemplary modelling study. *BMC Med Res Methodol* 2013;13:12.
  160. Hlatky MA, Greenland P, Arnett DK, et al. Criteria for evaluation of novel markers of cardiovascular risk: a scientific statement from the American Heart Association. *Circulation* 2009;119:2408–16.
  161. Biesheuvel CJ, Grobbee DE, Moons KG. Distraction from randomization in diagnostic research. *Ann Epidemiol* 2006;16:540–4.
  162. Neumann PJ, Tunis SR. Medicare and medical technology—the growing demand for relevant outcomes. *N Engl J Med* 2010;362:377–9.
  163. KNAW. Evaluation of new technology in health care. In need of guidance for relevant evidence. Amsterdam: KNAW; 2014.
  164. Clarke J, Wentz R. Pragmatic approach is effective in evidence based health care. *BMJ* 2000;321:566–7.
  165. Hunink MG, Krestin GP. Study design for concurrent development, assessment, and implementation of new diagnostic imaging technology. *Radiology* 2002;222:604–14.
  166. MacRae KD. Pragmatic versus explanatory trials. *Int J Technol Assess Health Care* 1989;5:333–9.
  167. Roland M, Torgerson DJ. What are pragmatic trials? *BMJ* 1998;316:285.

168. Bossuyt PM, Lijmer JG, Mol BW. Randomised comparisons of medical tests: sometimes invalid, not always efficient. *Lancet* 2000;356:1844–7.
169. Friedman LM, Furberg CD, DeMets DL. Fundamentals of clinical trials. 4th ed. Springer; 2010.
170. Meinert CL. Clinical trials. 2nd ed. Oxford University Press; 2012.
171. Rothman KJ, Greenland S. Modern epidemiology. 2nd ed. Philadelphia: Lippincott-Raven Publishers; 1998.
172. Rosenbaum PR. Design of Observational Studies. 2nd ed. Springer; 2002.
173. Bluhm R. From hierarchy to network: a richer view of evidence for evidence based medicine. *Perspect Biol Med* 2005;8:535–47.
174. Walach H, Falkenberg T, Fonnebo V, et al. Circular instead of hierarchical: methodological principles for the evaluation of complex interventions. *BMC Med Res Methodol* 2006;6:6–29.
175. Merlin T, Lehman S, Hiller JE, et al. The “linked evidence approach” to assess medical tests: a critical analysis. *Int J Technol Assess Health Care* 2013;29:343–50.
176. Schaafsma JD, van der Graaf Y, Rinkel GJ, et al. Decision analysis to complete diagnostic research by closing the gap between test characteristics and cost-effectiveness. *J Clin Epidemiol* 2009;62:1248–52.
177. Horvath AR, Lord SJ, StJohn A, et al. From biomarkers to medical tests: the changing landscape of test evaluation. *Clin Chim Acta* 2014;427:49–57.
178. Boyd JC, Bruns DE. Performance requirements for glucose assays in intensive care units. *Clin Chem* 2014;60:1463–5.
179. Hendriksen JMT, Geersing GJ, van Voorthuizen SC, et al. The cost-effectiveness of point-of-care D-dimer tests compared with a laboratory test to rule out deep venous thrombosis in primary care. *Expert Rev Mol Diagn* 2015;15:125–36.

## MULTIPLE CHOICE QUESTIONS

1. Which statement applies to a sample of values drawn from a Gaussian distribution?
  - a. The central location is best described by the median.
  - b. The dispersion is best described by the interquartile range.
  - c. The distribution of the values is likely to be asymmetric.
  - d. The t-distribution is useful for estimation of the 95% confidence interval for the mean value.
  - e. A goodness of fit test is likely to provide a significant deviation.
2. Which statement concerning the technical validity of an analytical assay is correct?
  - a. The precision expressed by the standard deviation (SD) is always constant over the measurement range.
  - b. The limit of detection expresses the analytical specificity of an analytical assay.
  - c. The four-parameter logistic calibration model is useful for linear calibration curves.
  - d. For a qualitative analytical assay, the cutoff point is defined as the true concentration at which there are 50% positive and 50% negative results.
  - e. The measurement range extends from the limit of detection (LOD) to the upper limit of quantification (ULOQ)
3. Two analytical methods are to be compared by analysis in parallel of a suitable number of patient samples. Which statement is true?
  - a. Ordinary least-squares regression analysis is the most appropriate data analysis approach
  - b. It is generally recommended that the manufacturer uses 40 samples for comparison and the user laboratory 100 samples.
  - c. A calibration difference is most typically disclosed by an intercept estimate significantly different from zero obtained by regression analysis.
  - d. A weighted Deming regression analysis approach should not be considered when the analytical CV%<sub>s</sub> are constant over the measurement range.
  - e. In case of constant CV%, the Bland-Altman difference plot shows an increasing scatter of the measured differences at increasing measurement values.
4. When assessing the relation between a set of paired measurements performed by two analytical assays on a group of subjects, the following applies?
  - a. The slope and intercept of linear regression analysis fully describe the relationship.
  - b. Outliers are usually without significance.
  - c. The correlation coefficient describes the relationship adequately.
  - d. The correlation coefficient tends to be high given a short measurement range.
  - e. A plot of residuals may illustrate both nonlinearity and random error.
5. The traceability chain extends downward from the reference measurement procedure to the routine analytical method. Which statement is true?
  - a. A reference measurement procedure is sensitive to matrix effects.
  - b. The uncertainty decreases down the traceability chain from the reference measurement procedure to the routine analytical method.
  - c. The standard uncertainty indicates a 95% uncertainty interval.
  - d. The reference measurement procedure is always more precise than the routine analytical method.
  - e. Harmonization of laboratory measurements does not presuppose traceability to a reference measurement procedure.
6. The diagnostic accuracy of a test is assessed on a number of subjects suspected of having a given target disease. Which statement is true?
  - a. The diagnostic accuracy is characteristic for the test and is not influenced by the actual setting in which it is evaluated.
  - b. The receiver operating characteristic area provides a measure of the diagnostic accuracy, which is not dependent on a selected cut-off value.
  - c. When the cut-off value of a quantitative test is increased, the specificity declines and the sensitivity increases.
  - d. The cut-off value should generally be selected so that the sum of specificity and sensitivity is maximized.
  - e. To rule out the presence of disease, it is important that the specificity is high.
7. It is considered to add a new test to an existing set of diagnostic procedures. Which statement is true?
  - a. The diagnostic accuracy of the new test is the most important point to consider.
  - b. A multivariate data treatment based on logistic regression analysis presupposes quantitative test results.
  - c. The net reclassification improvement (NRI) of an added test indicates the improvement in the difference between the proportions of individuals moving up and the proportion of subjects moving down for those with the target disease.
  - d. In decision curve analysis, the graph displays the probability thresholds for further management on the x-axis and the net benefit of the diagnostic approaches on the y-axis.
  - e. It is unlikely that results from several tests are correlated.
8. What study design is preferred when the aim is to quantify the diagnostic accuracy of a test in its clinical context?
  - a. A Randomized trial.
  - b. A case control study.
  - c. A routine care database study.
  - d. A cohort study.
  - e. A registry study.

9. Why is it important that in diagnostic accuracy studies the results of the reference standard are interpreted and classified without knowledge of the index tests under study?
- Otherwise the results of the index test are incorporated in the classification of the reference standard results, potentially leading to inflated estimates of the accuracy of the index test.
  - Otherwise the results of the index test are incorporated in the classification of the reference standard results, potentially leading to inflated estimates of the accuracy of the reference standard.
  - Otherwise the results of the reference standard are incorporated in the classification of the index test results, potentially leading to inflated estimates of the accuracy of the index test.
  - Otherwise the results of the reference standard are incorporated in the classification of the index test results, potentially leading to inflated estimates of the accuracy of the reference standard.
  - Otherwise the results of the index test are incorporated in the classification of the reference standard results, potentially leading to deflated estimates of the accuracy of the index test.
10. What is the main threat if the study data at hand are used to select the optimal threshold for disease presence or absence?
- It will yield inflated estimates of the diagnostic accuracy of the test.
  - It will yield deflated estimates of the diagnostic accuracy of the test.
  - It will yield selection bias.
  - It will yield too wide confidence intervals of the diagnostic accuracy estimates of the test.
  - It will yield a biased estimate of the optimal threshold.

# Governance, Risk, and Quality Management in the Medical Laboratory\*

*Leslie Burnett, Mark Mackay, and Tony Badrick*

## ABSTRACT

The aspirational goal of diagnostic pathology laboratories is to always provide the right result for the right test on the right patient at the right time and with the right support. To deliver this outcome systematically, reliably, and safely for the patient, has required the long-term development of an organizational culture and framework by the international laboratory medicine community.

### Background

Quality management (QM) is a set of principles for coordinating management and improvement activities to ensure that an organization continuously meets the requirements of its customers (users), even as those needs change. There are many approaches to and tools for deploying these principles, some tools focusing only on selected principles. It is widely regarded as best management practice when these principles are fully adopted as an organizational framework.

Quality management systems (QMS) are consensus-driven structured frameworks for ensuring consistency in the quality of products and services to meet customer needs. Both QM and QMS evolved from technical quality, namely, quality control (QC) and external quality assessment/proficiency testing (EQA/PT) activities, and these elements still play important roles. The International Organization for Standardization (ISO) has developed ISO 15189 as an internationally accepted QMS standard suitable for accreditation of medical laboratories. In some countries, alternative or additional standards or accreditation requirements may apply.

Further evolution and development of this framework is continuing. Laboratory medicine is different to many other

industries because it includes the patient, and thus carries additional ethical and legal responsibilities for maintaining patient safety and optimizing and improving patient care—this is addressed under the banner of clinical governance. There is also increasing recognition of the need to prioritize and focus efforts on those organizational activities most vulnerable to failure or with the greatest consequences—this is addressed through risk management. And layered on top of these frameworks and standards is the recognition that leadership and accountability are required to drive and maintain quality improvement.

### Content

The most widely adopted QMS used internationally (ISO 15189) is described in some detail and is compared with an alternative or supplementary framework (Clinical and Laboratory Standards Institute [CLSI] QMS01). The role of standards and guidelines in self-assessment and in the accreditation and regulation of medical laboratories, and the differences between accreditation and certification are explained. The principles of clinical governance and risk management are described and their links to QMS are explored. Finally, the philosophy and principles of QM are described, and within this, the rationale is explained for a QMS as a means of maintaining and controlling existing operations, as well as systematically improving the quality of test results and organizational services. Various pathways taken by laboratories to implementing QM are considered, and various structured quality improvement tools and management approaches are described.

\*The full version of this chapter is available electronically on [ExpertConsult.com](http://ExpertConsult.com).

## INTRODUCTION

### Purpose and Scope of a Medical Laboratory

The services of a medical laboratory play a key role in the practice of modern medicine. In different countries, and indeed within individual countries, there are different views of what constitutes the scope of a medical laboratory service and its purpose. International Organization for Standardization (ISO) 15189<sup>1</sup> is the international standard for quality and competence in the medical laboratory.

ISO 15189 defines a medical laboratory as a “laboratory for the ... examination of materials derived from the human body for the purpose of providing information for the diagnosis, management, prevention, and treatment of disease in, or assessment of the health of, human beings, and which may provide a consultant advisory service covering all aspects of laboratory investigation.” The standard itself details a specific requirement for advisory services that includes the provision “of advice on the choice of examinations and use of services including repeat frequency and required type of sample” and “where appropriate, interpretation of the results of examinations.”

Note that the definition also states: “Facilities which only collect or prepare samples, or act as a mailing or distribution center, are not considered to be medical or clinical laboratories, although they may be part of a larger laboratory or system.”

### Defining Quality in the Medical Laboratory

Discussions of quality in relation to the medical laboratory often invoke phrases such as “fit for purpose” or, in the case of a laboratory test or examination, as being “fit for its intended use.” The importance of these phrases is that they link the quality requirement to its purpose.

The ISO definitions of quality have evolved over a number of years with ISO 9000<sup>2</sup> defining quality as the “degree to which a set of inherent characteristics fulfills requirements,” where requirement is a “need that is either stated, generally implied or obligatory.” The bias and imprecision of a measurement process are examples of inherent characteristics (i.e., inbuilt features, as distinct from, say, the price charged for a particular test, which is an assigned characteristic).

Quality, as defined, is judged by the ability to reliably meet needs and expectations. This concept of quality as not being an absolute, but rather being matched to specific needs, has important implications. For example, a patient with suspected hypoglycemia would have a different requirement for the timeliness of a blood glucose result compared with an asymptomatic person being screened for diabetes. A test might have inherent characteristics that meet appropriate analytical and biological goals and yet still not be fit for purpose because it was not delivered in a timely manner and thus failed to fulfill the stated requirements.

In later chapters, statistical techniques are described for ensuring that analytical processes remain in control (see Chapter 6) and for ensuring that analytical results meet biological requirements (see Chapter 8).

It should also be noted that, with rapid change in all of medical practice, scientific advancement, and in broader society, even after optimization to ensure that all laboratory services are of appropriate quality, changing circumstances require ongoing and constant monitoring and adjustment

of laboratory performance to ensure that it remains correctly matched to the needs of patients, users, and customers. That is, a systematic continuous improvement process is essential.

### Evolution of the Approach to Quality Management

Quality management (QM) had its origins in the 1930s with Shewhart’s work on the understanding and control of analytical variation. This changed the quality process from “inspect for errors after production” to “control the process to reduce the errors produced,” which became the classic quality control (QC) approach and is these days formalized in statistical process control (SPC) (see Chapter 6). Shewhart also introduced a cyclic method for making improvements, which has become known as “Plan-Do-Check-Act” (PDCA cycle). This was important because it introduced a systematic approach to improvement, which has since spawned a vast array of improvement methodologies, including the very successful “Define, Measure, Analyze, Improve, Control” (DMAIC) methodology of Six Sigma.<sup>3</sup>

Within the field of laboratory medicine two key themes have driven major improvement: firstly, the introduction of EQA/PT programs, a form of benchmarking, highlighted great variability in laboratory performance first in technical quality and later service quality; and secondly, the introduction of QMS brought a systematic approach to managing, giving greater control to enable improvements to be made and retained in place.

There has always been parallel development of different approaches to the broader management of quality, including governance and risk management outside healthcare. Laboratory medicine has adapted those techniques that work in healthcare. The result is that quality can be seen from several viewpoints, each providing valuable tools for managing laboratory quality.

Although some laboratories see their role as primarily the delivery of analytically correct test results, this is a very narrow view and one that overlooks the reality that an analytically correct result may still be of poor quality because of not meeting, for example, turnaround time, cost, result delivery method, or any of a host of user-specified criteria or expectations. A quality laboratory service is one that knows at all times the users’ criteria and needs and has internal systems that can ensure these are met. Whenever user requirements change, a quality laboratory will have procedures to detect these changes, even if unannounced, and will adapt itself to respond appropriately.

To have the organizational sophistication to be able to achieve these degrees of responsiveness may take time. Although some laboratories see “meeting accreditation standards” as an end in itself, such laboratories will lose the benefit of the culture of organization-wide customer focus and systematic continuous improvement that comes from the use of a QMS to drive overall quality, rather than being merely a tool for achieving the minimum acceptable standard. Laboratory organizations typically evolve or mature through various stages in their understanding and adoption of QMS, and this rate of maturity is generally driven by management leadership. Governments may seek to accelerate this process by mandating that some of these stages are a condition of operation.

The rest of this chapter is presented as four main topics:

- Governance (including clinical governance), benchmarking, and risk management,
- QMS, including internationally applicable standards and guidelines and the key components of such systems
- Accreditation and regulation of medical laboratories, including an international model that includes self-assessment through accreditation and regulation
- Approaches to QM, integrating its basic principles

## CLINICAL GOVERNANCE, BENCHMARKING, AND RISK MANAGEMENT

### Governance

Historically, governance, risk management, and compliance are three related facets that aim to assure an organization reliably achieves its objectives, addresses uncertainty, and acts with integrity. Regulatory compliance means conforming with stated requirements. At an organizational level, it is achieved through management processes which identify the applicable requirements (defined, e.g., in laws, regulations, contracts, strategies, and policies) including ethical and prudential expectations, assess the state of compliance, assess the risks and potential costs of noncompliance against the projected expenses to achieve compliance, and hence prioritize, fund, and initiate any corrective actions deemed necessary or appropriate.

Governance and management of risk and compliance are not new, and most organizations including laboratories have been governed and their risk and compliance managed in the past, but an understanding of the interlinked nature of these activities is more recent.

Healthcare organizations, both public and private, have traditionally been controlled by boards or equivalent structures, as corporate entities with a focus on the strategic direction of the organization as a whole and ensuring that the operations, finances, and risk management systems are in place to meet statutory and prudential requirements. It has become apparent that errors and omissions probably occur more frequently in healthcare than in other industries or fields of activity, and certainly at a higher rate than is acceptable to the public<sup>4</sup>; from these findings has emerged acceptance there is a need to have more accountability for clinical care delivery. The term governance describes the overall management approach through which senior executives direct and control the entire organization, using a combination of management information and hierarchical management control structures.

In healthcare settings, (clinical) governance of healthcare or more broadly, governance of clinical effectiveness and clinical performance, has become increasingly important.<sup>5</sup> The approach to this challenge has been the integration of QM and risk management under an umbrella of clinical governance.

The term “clinical governance” emerged in health during the 1990s and is used to describe the framework through which health organizations are accountable for continuously improving the quality of their services and safeguarding high standards of care. It is an extension of self-regulation that acknowledges the complexities of the delivery of modern healthcare in an integrated, organization-wide context and has been defined as “a framework through which ... organizations are accountable for continuing to improve the quality of the service and safeguarding high standards of care by creating an

**TABLE 3.1 The Main Components of Clinical Governance<sup>7</sup>**

Clear lines of responsibility and accountability for the overall quality of clinical care.
• a designated senior clinician, ideally at board level, is responsible for ensuring that systems of clinical governance are in place and are monitoring their continued effectiveness
• an annual report on clinical governance is to be produced which is received centrally and open to public scrutiny
A comprehensive program of quality improvement clinical activities.
• full participation by all doctors in clinical audit programs
• ensure clinical standards are implemented
• workforce planning and development
• continual professional development, including clinical leadership
• consultant appraisal leading to revalidation
Clear policies aimed at managing risk, for example development of risk management strategy
Procedures for all professional groups to identify and remedy poor performance.
• critical incident reporting to ensure that adverse incidents are identified, openly investigated, and lessons are learned
• professional performance procedures that take effect at an early stage before patients are harmed and which help individuals to improve their performance whenever possible
• all staff supported in their duty to report any concerns about colleagues' professional conduct and performance.

environment in which excellence in clinical care would flourish.”<sup>6</sup> Clinical governance can be viewed as a whole system of cultural change that provides the means of developing the organizational capability to deliver sustainable, accountable, patient-focused, and quality-assured healthcare. The main components of clinical governance are described in Table 3.1.<sup>7</sup>

Quality initiatives to facilitate clinical governance in the laboratory include accreditation to an external standard such as ISO 15189, and benchmarking.

### Benchmarking

Benchmarking is the process of measuring products, services, and practices against peers and leaders in a field, allowing the identification of best practices that might lead to sustained and improved performance.<sup>8</sup> Performance can be compared either in a generic way, in which there is a comparison of a process regardless of the industry, or in a functional way, in which there are comparisons within the same industry. Laboratories seek to benchmark for four main reasons: quality improvement, for accreditation requirements, to meet funder requirements, and to improve competitive advantage.<sup>9</sup>

It has been shown that laboratories that regularly report on a particular aspect of quality tend to perform better on that quality measure than facilities in which this regular monitoring is not taking place<sup>10</sup> and that benchmarking quality over time is also associated with improved performance.<sup>11</sup>

There are a number of benchmarking schemes available to clinical laboratories including The Benchmarking Partnership (UK)<sup>12</sup> and the College of American Pathologists’ (CAP) Q-Probes scheme.<sup>13</sup> Variation in clinical practice between laboratories may be identified through benchmarking and this may lead to an agreement over what should be best or preferred practice.

## **POINTS TO REMEMBER**

Governance describes the overall management approach through which senior executives direct and control the entire organization, using a combination of management information and hierarchical management control structures.

Clinical governance requires the following:

- Clear lines of responsibility and accountability for the overall quality of clinical care.
  - A comprehensive program of quality improvement clinical activities.
  - Clear policies aimed at managing risk, for example, development of risk management strategy.
  - Procedures for all professional groups to identify and remedy poor performance.

Benchmarking is the process of measuring products, services, and practices against leaders in a field, allowing the identification of best practices that will lead to sustained and improved performance.

## Risk Management

Risk management is defined by the European Foundation for Quality Management<sup>14</sup> as: "The systematic use of organization-wide processes to identify, assess, manage, and monitor risks—such that aggregated information can be used to protect, release, and create value." Risk includes not only clinical

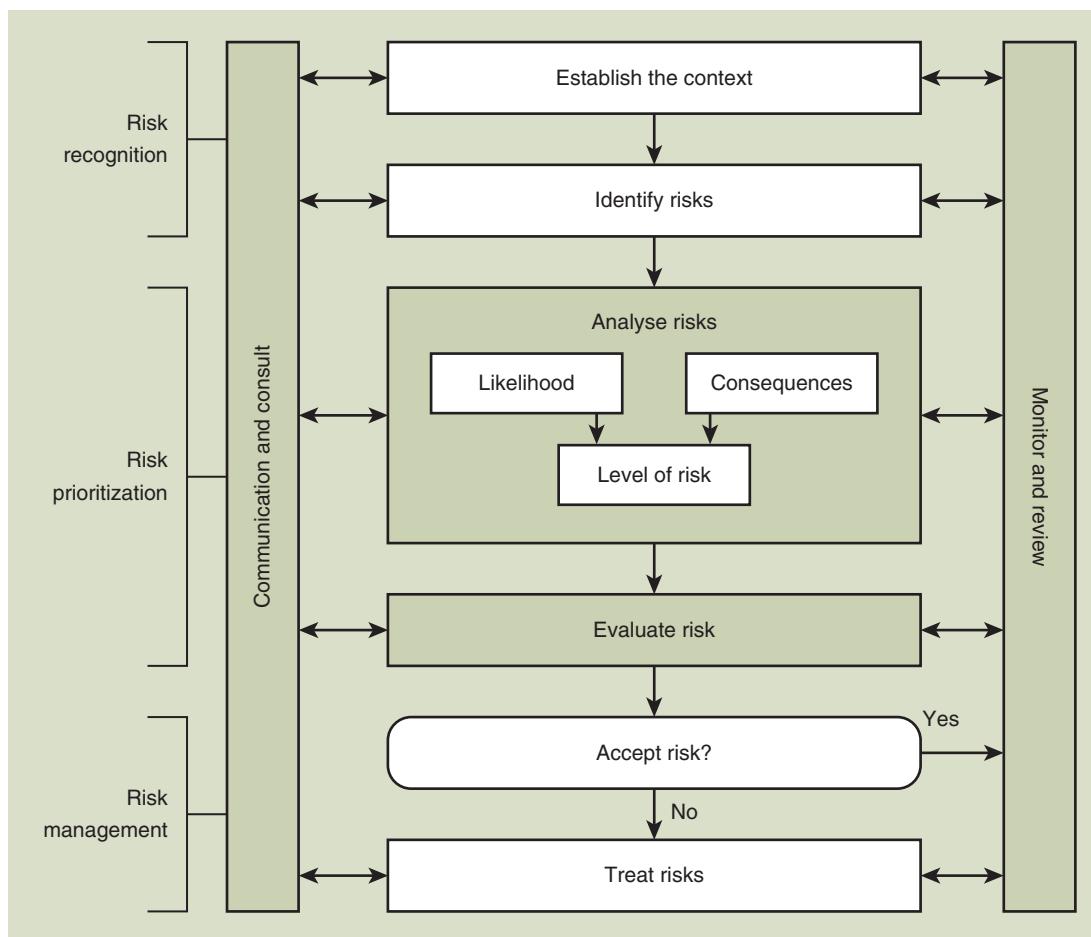
risk but financial risk, facilities, workplace health and safety, and reputational risk. There are three types of risks: predictable risks that organizations know they face; the risks which an organization knows it might run but which are caused by chance; and the risks that organizations do not know they are running. The response to risks typically depends on their perceived gravity and involves controlling, avoiding, accepting, or transferring them to a third party.

The ISO standard that describes risk management is ISO 31000:2018, Risk management—Guidelines, which provides principles, framework, and a process for managing risk. The following diagram (Fig. 3.1) is modified from this standard and describes the risk management process and the interconnections with monitoring and review, which are related to QM, and communication and consultation which is a component of governance.<sup>15</sup>

Using this approach as our model we will start with the contextualization of risk in a pathology laboratory. The context refers to the question of what is at risk, and the potential sources of risk associated with the preanalytical, analytical, and postanalytical processes, but should also consider risks associated with operations, such as billing, human resources, security, privacy, and safety.

## Risk Identification

The assessment task is to understand what is at risk and what events could potentially cause or contribute to harm or



**FIGURE 3.1** Model of the Risk Management Framework modified from ISO 31000:2018, risk management—guidelines.

benefits, their associated causes, and their potential consequences. There are a number of risks that are easily identified but the value of the risk identification process is the recognition of unexpected threats.<sup>16</sup> Table 3.2 lists the risk points that must be present.<sup>17</sup>

**TABLE 3.2 Risk Points that Must be Present in Clinical Laboratory Service<sup>17</sup>**

Governance structure
Competence of staff
Patient identification
Specimen integrity
Specimen traceability
Specimen analysis
Verification, validation, and documentation of methods
Quality assurance
Reporting results
Turnaround time
Communication
Sendaway (referral) tests
Reviews of incidents
IT algorithms, systems, and interfaces
Logistics
Skilled workforce
Critical results
Accreditation failure
Business and service continuity
Financial risk
Health and safety risk
Reputational risk

Other tools may be useful to identify risk in a complex process. One of these tools is the process map.<sup>18</sup>

### Analysis of Risk

Once a risk has been identified there are a wide range of possible responses an organization may take. Deciding what to do will be a management decision based on many factors including cost, understanding of the risk, and a balance between the perceived cost/liability of insurance and the cost/liability of risk. To inform this decision-making process with some objective process is risk analysis, which is based on likelihood and consequence of a particular risk. Likelihood depends on the probability of occurrence and the frequency of activity.

The most used quantitative risk model is Failure Mode and Effects Analysis (FMEA).<sup>19</sup> This model provides a means to systematically analyze postulated component failures and identify the resultant effects on system operations. There are three factors that are used in the model to calculate a “risk”; occurrence, severity, and detection.<sup>20</sup>

The full model used in industry usually includes all three factors, which are ranked on a scale from 1 to 10 (low risk to high risk), and then multiplied together to give a Risk Priority Number (RPN) to prioritize risks. In healthcare applications, it is common to use a two-factor model that considers only the probability of occurrence and severity of harm, then ranks each factor on scales from 1 to 5 or from 1 to 3, calculates criticality (using RPN), or uses a graphical risk acceptability matrix to describe and prioritize risks. Once risk has been assessed, the next step is to mitigate the effects by preventing or reducing occurrence, improving detection, and

**TABLE 3.3 Risk Matrix to Categorize Risk Criticality From a Combination of Severity and Occurrence**

	<b>1 Negligible</b>	<b>2 Minor</b>	<b>3 Serious</b>	<b>4 Critical</b>	<b>5 Catastrophic</b>
5 Frequent	5 low	10 medium	15 high	20 very high	25 very high
4 Probable	4 low	8 medium	12 medium	16 high	20 very high
3 Occasional	3 low	6 low	9 medium	12 medium	15 high
2 Remote	2 low	4 low	6 low	8 medium	10 medium
1 Improbable	1 low	2 low	3 low	4 low	5 low

#### Legend for Risk Matrix Scales

##### Occurrence Levels

- 5 Frequent—once per week
- 4 Probable—once per month
- 3 Occasional—once per year
- 2 Remote—once every few years
- 1 Improbable—once in the lifetime of the measuring system

##### Severity Levels

- 5 Catastrophic—could result in patient death
- 4 Critical—could result in permanent or life-threatening injury
- 3 Serious—could result in injury or impairment requiring professional medical intervention
- 2 Minor—could result in temporary injury or impairment requiring professional medical intervention
- 1 Negligible—could result in inconvenience or temporary discomfort

##### Risk Management Category

- Very high—a dangerous level of risk that is required to be controlled immediately
- High—an unacceptable level of risk that should be controlled immediately
- Medium—manage by specific monitoring or audit procedures
- Low—manage by routine procedures

##### Example of Potential Response Matched to Risk Category if Incident Were to Occur

- Formal investigation and root cause analysis (RCA) with recommendations reported to board and externally
- Formal investigation with recommendations reported to the Executive
- Departmental investigation and improvement summarized in report to operations
- Local monitoring and KPI

reducing harm by corrective or preventive actions and disclosure of information for safety.<sup>21</sup>

The next step is risk evaluation where risks are evaluated against an appropriate risk-acceptance criterion to give a ranking, for example.

- low (tolerable—These risks are generally considered acceptable. Manage by routine procedures. Monitor and review effectiveness throughout),
- medium (Manage by specific monitoring or audit procedures),
- high (Unacceptable level of risk that should be controlled immediately),
- very high (Dangerous level of risk that is unacceptable and required to be controlled immediately).

To decide whether risk is acceptable, a risk acceptability matrix is used, as shown in [Table 3.3](#). The matrix shows the ranking for severity of harm across the columns and the ranking for probability of occurrence of harm down the rows. The matrix defines certain row-column combinations as acceptable and others as unacceptable, generally separated by a diagonal from the top left to the lower right. The risk acceptability matrices in ISO 14971<sup>22</sup> and Clinical and Laboratory Standards Institute (CLSI) EP23A<sup>23</sup> are not identical, even although the EP23A document quotes ISO 14971 as its source. That situation is an indicator of the subjectivity of risk evaluation, which together with the ranking scales should alert the laboratory to the qualitative nature of this methodology.

### Risk Mitigation

The next stage in the process of risk management is the management of the risks which have been identified and prioritized. Management of risks can again be categorized in different ways. Perhaps the simplest of these is the “four Ts”:

- terminate—cease activities related to the risk (e.g., giving up smoking avoids associated health risks).
- treat—add control measures or contingency plans to manage the likelihood and consequence of events (e.g., wearing a hard hat reduces the consequences of being hit by a falling object): additional control measures or contingency plans become part of the management system.
- tolerate—accept the risk; and
- transfer—move the impact of risks to another entity (e.g., insurance).

### Integrating Risk and Quality Management

Quality and risk management have developed in parallel though with much in common and effective risk management is dependent on a robust QM program. The success of both systems requires collection and analysis of data on reportable incidents which are defined as any event inconsistent with routine procedure, policy, or practice that could adversely affect or threaten the well-being of a patient, staff, or member of the public or could seriously impact or compromise the organization.

There are three major ways in which QM experience can assist the implementation of risk management.<sup>15</sup>

1. in distinguishing between risks that can, and that cannot, be treated statistically;
2. through knowledge and experience in managing key processes; and
3. in implementing major organizational and cultural change.

On the other hand, a key aspect of risk management is the prioritization of risks based on their likely occurrence and

effects. If a similar approach could be used to prioritize quality risks, then the traditional focus of QM (solving the problem of too much inspection and waste) may be changed to analysis of faults and elimination of causes through improved product and process design.

### POINTS TO REMEMBER

Risk management is the systematic use of organization-wide processes to identify, assess, manage, and monitor risks—such that aggregated information can be used to protect, release, and create value.

There are three types of risks: predictable risks that organizations know they face; the risks that an organization knows it might run but which are caused by chance; and the risks that organizations do not know they are running.

The way risks are managed can be categorized in different ways: terminate—cease activities related to the risk; treat—add control measures or contingency plans to manage the likelihood and consequence of events; tolerate—accept the risk; and transfer—move the impact of risks to another entity (e.g., insurance).

Risk management involves three key steps: risk identification, risk analysis, and risk mitigation.

The most used quantitative risk model used is FMEA. This model provides a means to systematically analyze postulated component failures and identify the resultant effects on system operations. There are three factors that are used in the model to calculate a “risk”: occurrence, severity, and detection.

## QUALITY MANAGEMENT SYSTEMS

### Standards and Guidelines for the Medical Laboratory

The concept of a management system<sup>24</sup> is common in many ISO standards including ISO 15189, ISO 9001, ISO 14001, and ISO 17025. The common elements are given in [Table 3.4](#) but are discussed in more detail in the later section “Approaches to Quality Management.”

A management system is a way in which an organization manages the inter-related parts of its business to achieve its objectives. These objectives can relate to several different topics, including product or service quality, operational efficiency, environmental performance, health and safety in the workplace, and many more.

The level of complexity of the system will depend on each organization’s specific context. For some organizations, especially smaller ones, it may simply mean having strong leadership from the business owner, providing a clear definition of what is expected from each individual employee and how they contribute to the organization’s overall objectives, without the need for extensive documentation. More complex businesses operating, for example, in highly regulated sectors, may need extensive documentation and controls to fulfill their legal obligations and meet their organizational objectives.

ISO management system standards help organizations improve their performance by specifying repeatable steps that organizations can consciously implement to achieve their goals and objectives and to create an organizational culture that reflexively engages in a continuous cycle of self-evaluation, correction, and improvement of operations and

**TABLE 3.4 The Common Elements of a Management System**

Customer focus	The primary focus of quality management is to meet customer requirements and strive to exceed customer expectations.
Leadership	Leaders at all levels establish unity of purpose and direction and create the conditions in which people are engaged in achieving the organization's quality objectives.
Engagement of people	Competent, empowered, and engaged people at all levels throughout the organization are essential to enhance its capability to create and deliver value.
Process approach	Consistent and predictable results are achieved more effectively and efficiently when activities are understood and managed as interrelated processes that function as a coherent system.
Improvement	Successful organizations have an ongoing focus on improvement.
Evidence-based decision making	Decisions based on the analysis and evaluation of data and information are more likely to produce desired results.
Relationship management	For sustained success, an organization manages its relationships with interested parties, such as suppliers.

processes through heightened employee awareness and management leadership and commitment.

The benefits of an effective management system to an organization include:

- More efficient use of resources and improved financial performance,
- Improved risk management and protection of people and the environment, and
- Increased capability to deliver consistent and improved services and products, thereby increasing value to customers and all other stakeholders.

In considering documented standards for medical laboratories, it is important to distinguish between those written into governmental regulations or promulgated by accreditation bodies and those written by independent standards organizations.

Of prime importance are various standards prepared by ISO. The International Standard ISO 15189 Medical laboratories—Requirements for quality and competence, which incorporates QMS requirements, is internationally recognized for use in the accreditation of laboratories (see later).

The guidelines prepared by the CLSI in the United States, although not recognized for accreditation purposes, are a valuable aid to laboratories seeking to implement a QMS. The elements of both these standards and guidelines are compared and described below.

In addition to these materials, the National Pathology Accreditation Advisory Council (NPAAC), which is managed by the Department of Health of the Australian Government, has developed and published various standards and resources, which are available free of charge online. Examples include Requirements for the Estimation of Measurement Uncertainty, for the Packaging and Transport of Pathology Specimens

and Associated Materials, and for the Retention of Laboratory Records and Diagnostic Material.

The requirements of ISO 15189 are organized under two main clause headings, Management Requirements and Technical Requirements. Although the content of the subclauses is well formulated, the arrangement of the subclauses does not facilitate an understanding of the way they should interact (see Quality Management Systems and the Process Model later).

Other ISO standards of importance to medical laboratories are fully discussed and referenced in A Practical Guide to ISO 15189 in Laboratory Medicine<sup>25</sup> and include ISO 22870 Point-of-Care (POCT)—requirements for quality and competence,<sup>26</sup> to be used in conjunction with ISO 15189, and ISO/IEC 17043 Conformity assessment—general requirements for proficiency testing (external quality assessment).<sup>27</sup>

### Clinical and Laboratory Standards Institute—Quality Management Systems Guidelines

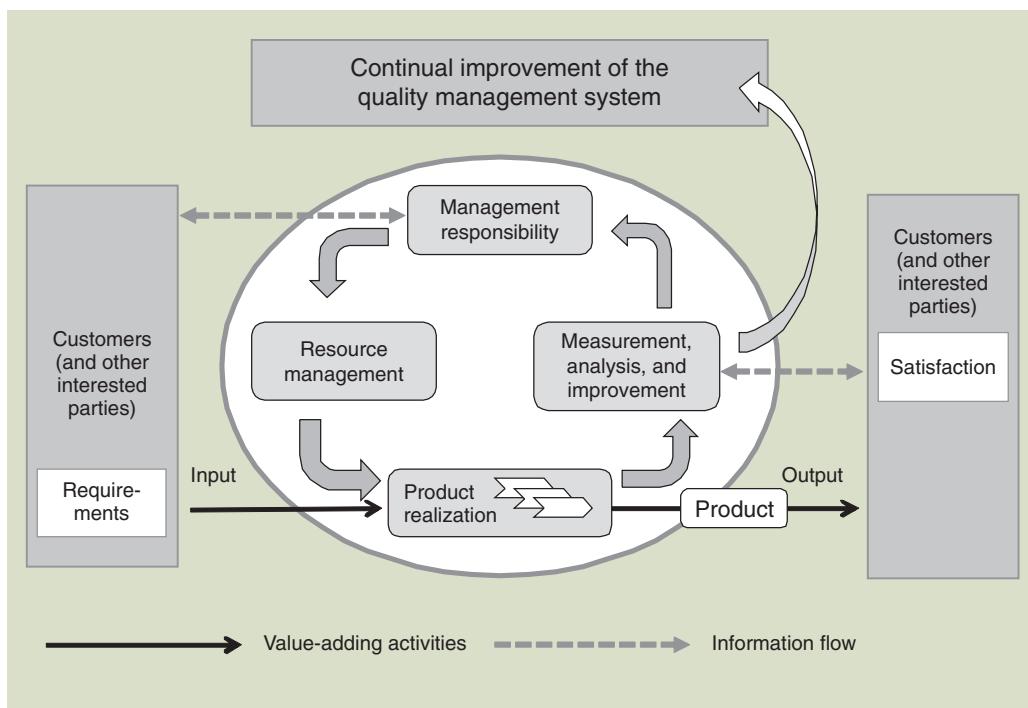
The mission statement of the CLSI (formerly known as the National Committee for Clinical Laboratory Standards) is to “develop best practices in clinical and laboratory testing and promote their use throughout the world, using a consensus-driven process that balances the viewpoints of industry, government, and the healthcare professions.” Its flagship QM guideline is QMS01 titled Quality Management System—A Model for Laboratory Services—Approved Guidelines 4th Edition.<sup>28</sup> The document describes 12 quality system essentials (QSEs) and embraces a QMS model with the emphasis on the path of workflow concept, including consultative functions (pre-examination, examination, and postexamination activities). The QMS guideline is supplemented by a series of QM documents such as QMS03 and QMS02 Quality Management System: Development and Management of Laboratory Documents, Approved Guideline—Sixth Edition. Discussion of and reference to many of these documents can be found in A Practical Guide to ISO 15189 in Laboratory Medicine.

### Quality Management Systems and the Process Model

Before describing and comparing the subclauses of ISO 15189 with the QSE of CLSI’s QMS01 guideline, it is useful to present the ISO 15189 subclauses in the context of a “process-based” approach to QM. The introduction to ISO 9001 promotes the adoption of “a process approach when developing, implementing, and improving the effectiveness of a quality management system” to “enhance customer satisfaction by meeting customer requirements.”<sup>2</sup> ISO 15189 defines a process as “a set of interrelated or interacting activities which transforms inputs into outputs.”<sup>1</sup> Inputs into a process are generally the outputs from other processes. Within any organization (e.g., a medical laboratory), there are numerous interrelated or interacting processes, and a process approach involves the identification and interactions of these processes and their management. Fig. 3.2, which was modified from ISO 9001, represents a model of such a system or series of interacting processes.

Given that the overall diagram (see Fig. 3.2) represents a QMS, it is helpful to translate some of the terms used in this process-based model into language that is more familiar to medical laboratory professionals. It can be viewed in two different ways.

The first view involves the totality of the model in which laboratory management (referred to in ISO 9001 as management responsibility) creates a quality system (QMS) and



**FIGURE 3.2** Model of the International Organization for Standardization (ISO) 9001 process-based quality management system. (Modified with permission from Burnett D. *A practical guide to ISO 15189 in laboratory medicine*. London: ACB Venture Publications; 2013.)

uses resources, staff, equipment, and so on (resource management) to carry out pre-examination, examination, and postexamination processes (product realization) to fulfill the requirements of the users. Depending on whether their requirements have been met or not, users may be defined as satisfied or dissatisfied. The pre-examination, examination, and postexamination processes are continually evaluated and improvements made as appropriate (measurement, analysis, and improvement). Evaluation and continual improvement activities include, for example, assessment of users' needs and requirements, internal audit of the examination processes, and review of participation in external quality assessment schemes.

The second view of the model focuses on the "product realization" processes. The users or customers have requirements or input that are formulated in consultation with laboratory management (management responsibility), and the laboratory responds by carrying out pre-examination, examination, and postexamination processes (product realization) to produce a product or report or output for the user. Depending on whether their requirements have been met or not, users may be defined as satisfied or dissatisfied.

It is only a short step from Fig. 3.2 to redraw the ISO 9001 process-based model and place the subclauses of ISO 15189 under headings more familiar to the medical laboratory professional, thus reordering the clauses of ISO 15189 into a practical interrelationship (Fig. 3.3).

Three subclauses (4.5, 4.7, and 5.6) occur in two places in the figure because they have content applicable in two places.

### ISO 15189 and Clinical and Laboratory Standards Institute—Quality System Essential Compared

Table 3.5 presents the CLSI-QSE (column 2) against the subclauses of ISO 15189 (column 3) divided into main sections

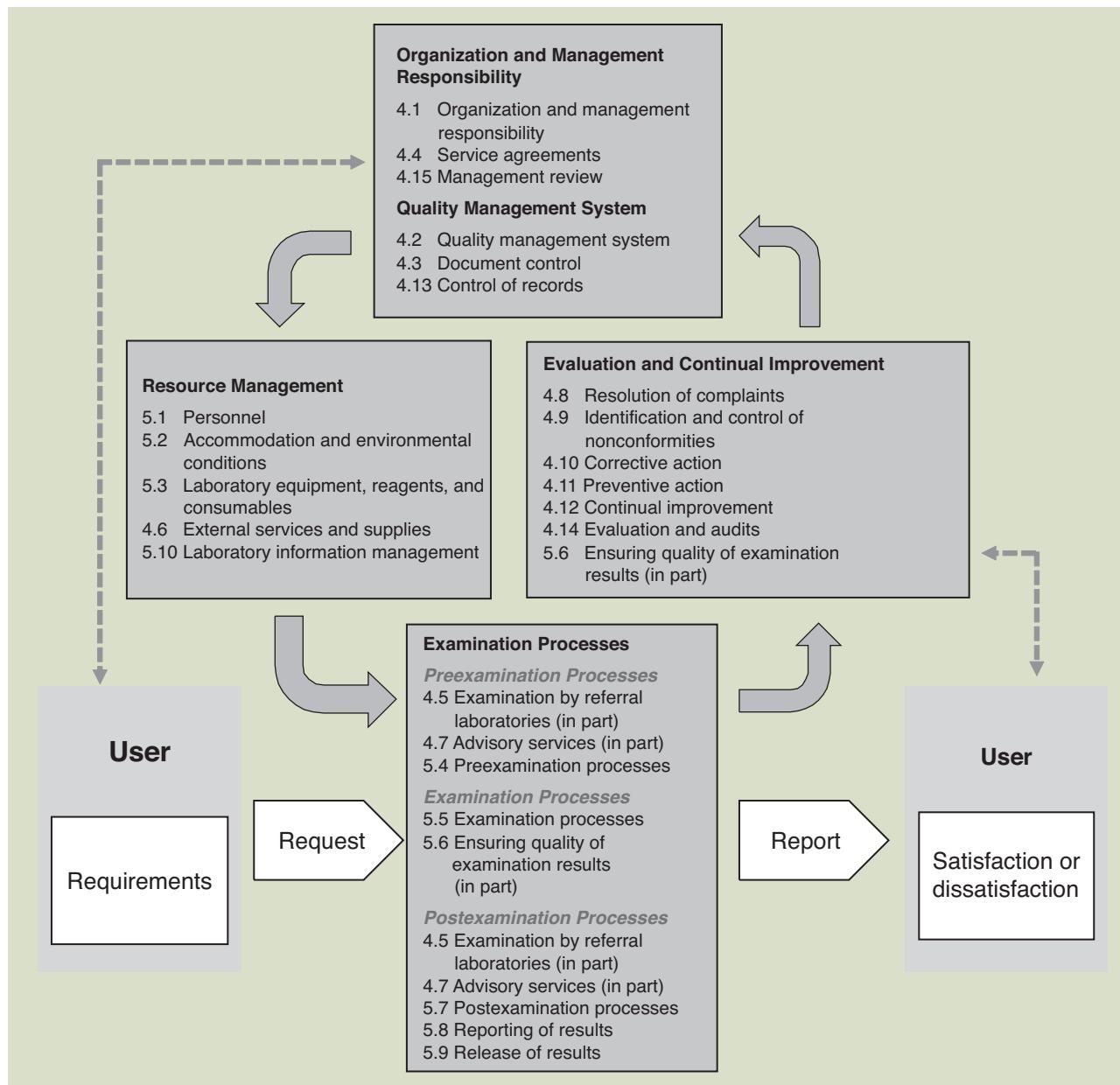
(column 1) that are the same as those presented in the process model shown in Fig. 3.3. The comparable titles in ISO 15189 are shown in column 3. Because of the differing content of the QSE and subclauses, such a comparison can only be an approximation. However, by using this table, the QSE can be placed in the appropriate sections of Fig. 3.3 to provide a pictorial representation of the QSE in a process-based model.

### The Key Components of Quality Management Systems for Medical Laboratories

The ISO defines a QMS as a "management system to direct and control an organization with regard to quality,"<sup>2</sup> and ISO 15189 states, "The laboratory shall establish, document, implement and maintain a QMS and continually improve its effectiveness" that "shall provide for the integration of all processes required to fulfill its quality policy and objectives and meet the needs and requirements of the users."<sup>1</sup> These statements are important in that they underline the fact that all processes within the laboratory form an integral part of the QMS. These key processes include both "core processes"—pre-examination, examination, or postexamination activities (or as CLSI terms them, "path of workflow")—and "support processes"—management responsibility, resource management and evaluation, and continual improvement (see Fig. 3.3).<sup>28</sup> The key components of both core and support activities are discussed below. This summary account is based on ISO 15189 terminology.

### Management Responsibility

Laboratory management is responsible for developing and implementing a QMS and continually improving its effectiveness.



**FIGURE 3.3** The main subclauses of International Organization for Standardization (ISO) 15189 reordered into a “process-based” model of a quality management system. (Modified with permission from Burnett D. *A practical guide to ISO 15189 in laboratory medicine*. London: ACB Venture Publications; 2013.)

### POINTS TO REMEMBER

Quality means consistent and uniform delivery of product and service that fulfills the needs and requirements of customers/users.

ISO 15189 is an international standard developed specifically for medical laboratories.

- Addresses both QMS and technical requirements
- Process focus: connects inputs (referring clinicians, test requests, and orders) with outputs (test results, reports, and interpretations) by managing and controlling internal operations
- Clauses in the standard can be mapped to equivalent elements of alternative frameworks (e.g., CLSI QMS01)

### Accreditation or Certification?

Laboratories can seek recognition in a few ways, but an apparent choice is between accreditation and certification. In common English usage,<sup>29</sup> the two words have similar meanings; to accredit means to “certify or guarantee someone or something as meeting required standards,” and to certify means “to endorse or guarantee that certain required standards have been met.” In practice, if laboratories seek certification, the standard applied would be ISO 9001, which examines the QMS; however, accreditation examines both the QMS and the professional and technical competence of the laboratory.

The importance of accreditation rather than certification has been endorsed by professional bodies<sup>30</sup> in far-reaching

**TABLE 3.5 Comparison of the Clinical and Laboratory Standards Institute Quality System Essentials With Subclauses of International Organization for Standardization (ISO) 15189 and Main Clauses of ISO 9001**

Headings From Fig. 3.3 (ISO9001 Main Clause Titles)	QMS01-A4 (QSE) and Path of Workflow Concept	ISO 15189 Subclauses
Organization and management responsibility (Management responsibility)	Organization	4.1 Organization and management responsibility 4.4 Service agreements 4.15 Management review
Quality management system (Quality management system)	Organization Documents and records	4.2 Quality management system 4.3 Document control 4.13 Control of records
Resource management (Resource management)	Personnel Facilities and safety Equipment Purchasing and inventory Information management	5.1 Personnel 5.2 Accommodation and environmental conditions 5.3 Laboratory equipment, reagents, and consumables 4.5 Examination by referral laboratories (in part) 4.6 External services and supplies 5.10 Laboratory information management
Examination processes (Product realization)		
Pre-examination	Customer focus Path of workflow—pre-examination activities	4.7 Advisory services (in part) 5.4 Pre-examination processes
Examination	Path of workflow: examination activities	5.5 Examination processes 5.6 Ensuring the quality of examination results (in part)
Postexamination	Purchasing and inventory Customer focus Path of workflow: Postexamination activities. Consultation on application of examination results to patient care	4.5 Examination by referral laboratories (in part) 4.7 Advisory services (in part) 5.7 Postexamination processes 5.8 Reporting of results 5.9 Release of results
Evaluation and continual improvement (Measurement analysis and improvement)	Nonconformance management Assessments Continual improvement Process management	4.8 Resolution of complaints 4.9 Identification and control of nonconformities 4.10 Corrective action 4.11 Preventive action 4.14 Evaluation and audits 4.12 Continual improvement 5.6 Ensuring the quality of examination results (in part)

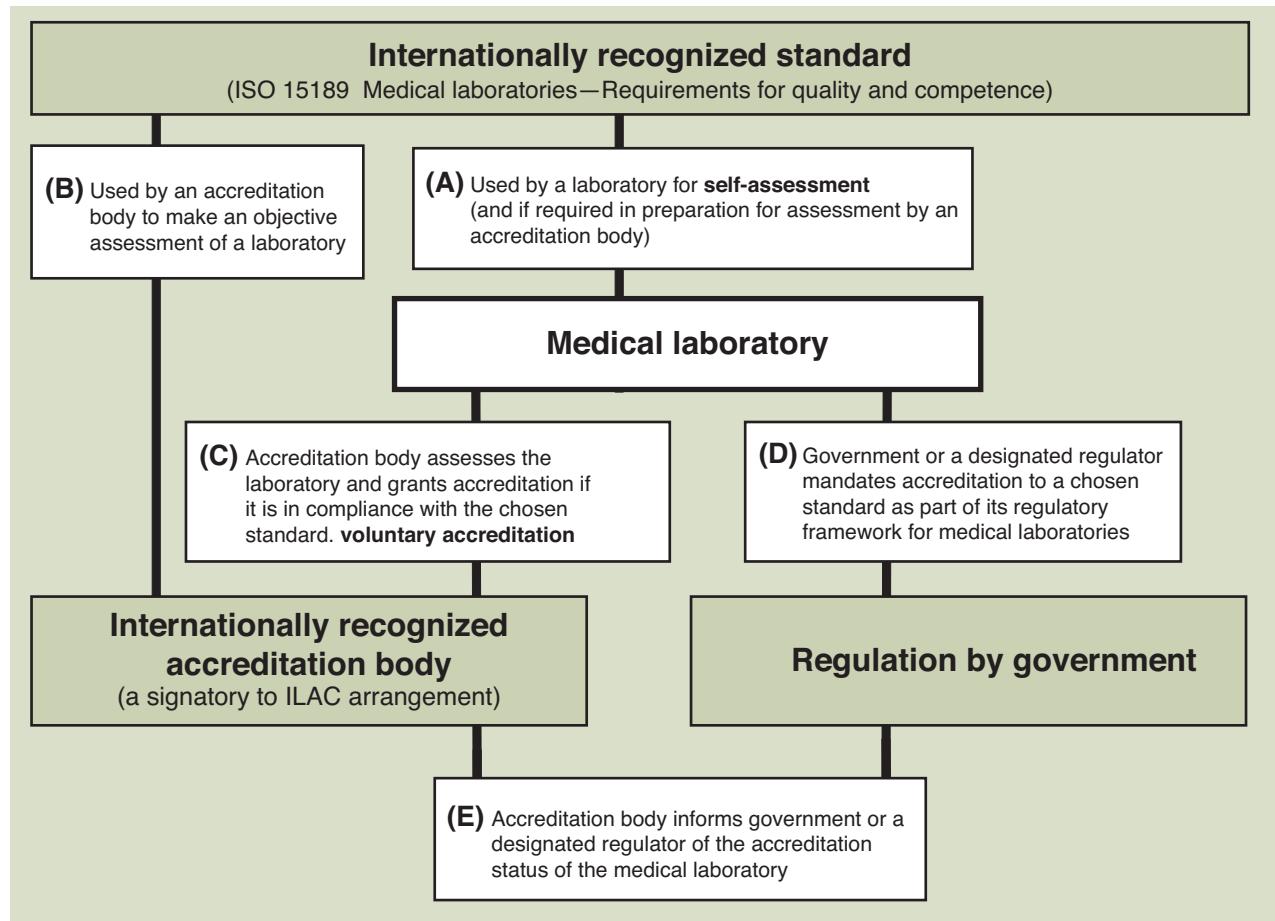
ISO, International Organization for Standardization; QSE, quality system essentials.

terms: “It is in the interests of patients, of society, and of governments that clinical laboratories operate at high standards of professional and technical competence” and “it is in the interests of competent laboratories that their competence is verified through a process of inspection and comparison against appropriate standards, as a confirmation of their good standing.” The primary objective of laboratory accreditation is to ensure that when a patient undergoes laboratory examinations, the results of those examinations are comparable irrespective of where they are performed, and consultation

is available regarding the choice and interpretation of such examinations.

### Accreditation and Regulation of Medical Laboratories An International Model

Although ISO 15189 is now widely used by governments as a means for mandating and assuring defined levels of quality in laboratory operations, it can also be used as a tool by individual laboratories for improving their operations even where this is not a regulatory requirement. Fig. 3.4 summarizes the



**FIGURE 3.4** An international model of accreditation and regulation of medical laboratories. Stages A to C can use either self-assessment or an accreditation body, but the result is voluntary accreditation. Stages D to E involve regulated or mandated accreditation (either by government or an accreditation body). ILAC, International Laboratory Accreditation Cooperation. (Modified with permission from Burnett D. *A practical guide to ISO 15189 in laboratory medicine*. London: ACB Venture Publications; 2013.)

relationship of ISO 15189 to a medical laboratory, from its use by the laboratory as a “self-assessment” tool to its use by governments in the regulation and licensing of medical laboratories.<sup>31</sup> This model (see Fig. 3.4) has three elements and five stages, A to E.

The first element is an “internationally recognized standard,” and ISO 15189 is recognized by the International Federation of Clinical Chemistry and Laboratory Medicine (IFCC) and the International Laboratory Accreditation Co-operation (ILAC) for use by accreditation bodies in confirming or recognizing the competence of medical laboratories.

The second element is that accreditation be undertaken by an “internationally recognized accreditation body,” and the introduction to ISO 15189 states that “if the laboratory seeks accreditation it should select an accrediting body which operates to appropriate international standards and which takes into account the particular requirements of medical laboratories.” Those criteria can be met by an accreditation body that is a signatory to the so-called “ILAC mutual recognition arrangement.” ILAC is an international cooperation among the various laboratory accreditation schemes operating throughout the world for facilitating trade by promotion of the acceptance of accredited test and calibration results between countries. National accreditation bodies (NABs) are

required to operate an appeal mechanism if a laboratory is dissatisfied with the service provided.

The third element is regulation by government. This regulation can be by government or by a designated regulator.

The five stages in the relationship between the use of ISO 15189 and the laboratory are:

- Stage A, ISO 15189, which is used by the medical laboratory for “self-assessment” and, if required, in preparation for assessment by an accreditation body.
- Stage B, ISO 15189, which is used by an accreditation body to make an objective assessment of the laboratory.
- Stage C, in which the laboratory is assessed as working in compliance with ISO 15189 and is granted accreditation, a process that can be termed “voluntary accreditation.”
- Stage D, in which government mandates accreditation to ISO 15189 as part of its regulatory framework.
- Stage E, in which the accreditation body informs government or the designated regulator of the accreditation status of the laboratory to provide evidence of fulfillment of a regulatory requirement.

Reaching stage E is often described as “mandatory accreditation,” but it is important to recognize that whereas government “mandates” accreditation, accreditation bodies provide assessment and grant accreditation under that mandate.

Accreditation bodies do not have the authority to mandate accreditation. The number of countries that mandate regulations citing ISO 15189 as the standard of choice is increasing worldwide.

Although the individual elements of the model are interdependent, the conceptual intention is that the elements function independently. In other words, those who write standards, those who accredit, and those who regulate have defined and independent roles. This provides the opportunity for government to set regulatory requirements, which might include a requirement for a laboratory to operate in conformity with the International Standard ISO 15189 but at the same time to remain “at arms’ length” from the assessment and accreditation process. In countries where there is no government regulation, it may still be explicit or implicit that medical laboratories be accredited in order, for example, to obtain contracts for the provision or reimbursement of services. Note that in some countries, some of the roles of the standard-setter and the accreditation agency are held by the same organization.

### Alternative Models

In the United States, accreditation is mandatory and the standards to be met are defined in the Clinical Laboratory Improvement Amendments (CLIA) Act.<sup>32</sup> Inspection is

### POINTS TO REMEMBER

Introduction of a QMS leads to progressive improvement in organizational quality over time.

Governments can mandate minimum standards for medical laboratory testing. The two standards most widely used are ISO 15189 (used in many countries) and CLIA (used in the United States and some other countries).

While laboratories may choose to meet only these minimum accreditation standards, and even in countries that do not yet mandate the use of such standards, QMS may provide laboratory managers with a systematic framework to support laboratory improvements in meeting customer requirements in both analytical and nonanalytical laboratory operations.

The standard setting body (e.g., ISO, government) and the accrediting bodies (varies by country) should be separate and “at arm’s length” from each other.

#### Accreditation

- Performed by a recognized accreditation agency, which is itself subject to accreditation by international agencies.
- Examines both the QMS and the professional competence of the laboratory.
- Uses a standard (such as ISO 15189) either on its own or supplemented by additional government or professional standards, as the basis for assessing the laboratory’s performance.

#### Certification

- Performed by a recognized certification agency, which is sometimes also subject to external agency oversight.
- Examples include ISO standards such as ISO 9001 and ISO 14001.
- Focuses on the QMS of the organization; does not address the technical competence of the medical laboratory.

carried out by governmental organizations or organizations recognized under the act as “deemed authorities,” such as the CAP. CAP sets its own standards over and above the requirements of CLIA and has recently been offering an inspection that additionally involves the requirements of ISO 15189. There are also several accreditation bodies in the United States that fit the international model, such as the American Association for Laboratory Accreditation (A2LA), and that are signatories to the ILAC arrangement. A2LA offers accreditation to ISO 15189 and CLIA or a combination of both.

In Europe, the situation is variable. In the United Kingdom, health is a devolved function of government, with England, Northern Ireland, Scotland, and Wales each having independent regulators of health and social care services. For example, the Care Quality Commission (CQC) in England uses accreditation to ISO 15189 as part of its regulatory framework for laboratories. In other countries, such as France, “according to the order of 13 January 2010 and Law No. 2013-442 of 30 May 2013 on a general reform of medical biology, all French medical laboratories have to be accredited in accordance with ISO 15189 standard in order to be able to carry out their activities.”<sup>33</sup>

In Australia, accreditation is also mandatory and the standards to be met are defined by the “Requirements” of the NPAAC. The NPAAC includes recognition of ISO 15189. The National Association of Testing Authorities, Australia (NATA)<sup>34</sup> is the sole accrediting organization.

### APPROACHES TO QUALITY MANAGEMENT

Having described the key elements of the quality management system, we now consider how they can be combined to support an organization-wide integrated approach to QM and quality improvement.

#### Principles of Quality Management

There are many diverse approaches to the implementation of QM and the introduction and maintenance of the QMS, but underlying these approaches are some basic principles. ISO 9000 defines QM as the “coordinated activities to direct and control an organization (the laboratory) with regard to quality,” and an accompanying note indicates the activities involved are “direction and control with regard to quality generally, including establishment of the quality policy and quality objectives, quality planning, QC, quality assurance, and quality improvement.”<sup>1</sup>

QM includes everything an organization does to meet the needs and requirements of its users or customers in addition to complying with regulatory requirements, technical requirements, and business needs. It is not just QC of technical work, quality assurance of processes, or compliance with minimum regulatory standards. Rather, it includes the approach taken by the organization, the way resources are deployed, the results achieved, and any improvement to the approach to fine tune the effort and maintain focus. Achieving these goals requires a governance model, executive leadership support, specific goals, resources, and defined time frames. Standards documents such as ISO 15189 provide a backbone for the effort and specify the minimum elements that need to be included. However, although international standards recognize the

likely existence of and need to meet specific local user or customer or regulatory requirements, they cannot realistically specify what the organizational culture should be, the exact path to take, or the tools to use along the journey. These things require an understanding of the principles behind QM and of the successful approaches and models others have taken and the tools used. One such QMS model is described in the CLSI guideline QMS01.<sup>25</sup>

The ISO document titled “Quality Management Principles”<sup>24</sup> provides standardized descriptions of seven principles that underpin all QM activities (see Table 3.4). It is intended that management use these principles to guide their organizations toward improvement.

While agreeing with the above list, some authors<sup>35–39</sup> have noted that there are additional key principles that can be regarded as components of a factual approach to decision making, the major ones being:

1. Understanding variation.
2. Cost of quality (including waste).
3. Nonconforming events (NCEs), and risk management.

Note that the order of numbering of these principles is not intended to imply that lower numbered principles are less important.

Although these basic principles of QM have application to all the components of a QMS (outlined in Table 3.4), there are clear linkages between certain ISO 15189 QMS components and the broader QM principles. For example:

- Management responsibility: customer focus and leadership.
- QMS: system approach to management and process approach.
- Resource management: involvement of people and mutually beneficial supplier relationships.
- Examination processes: all principles.
- Evaluation and continual improvement: continual improvement and a factual approach to decision making.

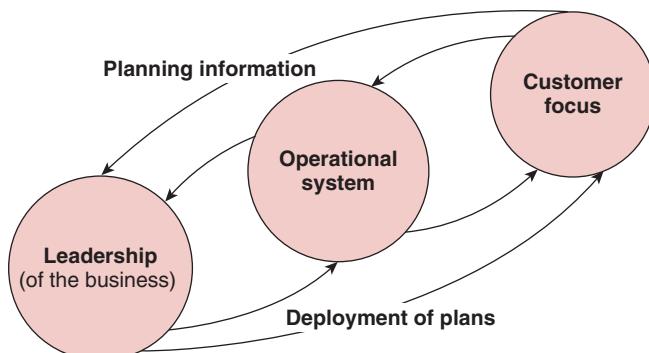
In the sections below, each principle is dealt with by providing the standardized description followed by a discussion of their importance in the context of QM in a medical laboratory and ideas for implementation.

### **Customer Focus**

The standardized ISO description of this principle is that “Organizations depend on their customers and therefore should understand current and future customer needs, should meet customer requirements, and strive to exceed customer expectations.”

A key principle of QM is to create the organization-wide recognition that customers can define quality. In this framework, the concept of a customer includes not only the external users or customers of the service but also includes a complete network of “internal customers,” that is, people or functional units who have an internal operational relationship within the company, such as employees and managers, or interactions between departments or units frequently as close as the next step in the analytical process. Delivery of quality product and service to the ultimate (external) customer requires consistently and reliably meeting the quality needs of the immediate (internal) customers.

An important approach is to identify “key” customers using multiple criteria (e.g., high work volume, medical importance, or urgency) to determine and confirm their needs



**FIGURE 3.5** Relationship among the three key elements of quality management: leadership, systems, and improvement driven by customer focus.

(by survey, customer interview, operational staff observation, or otherwise). Workflow process mapping can be used to identify internal customer needs and critical control points. Operational interfaces with customers are often key in determining customer needs and the laboratory’s performance in meeting them. After all needs have been identified and specified, one then designs a process and appropriate controls to ensure that all these needs can be consistently met (Fig. 3.5). This performance should be measured and controlled using statistical tools (see Chapters 2 and 6).

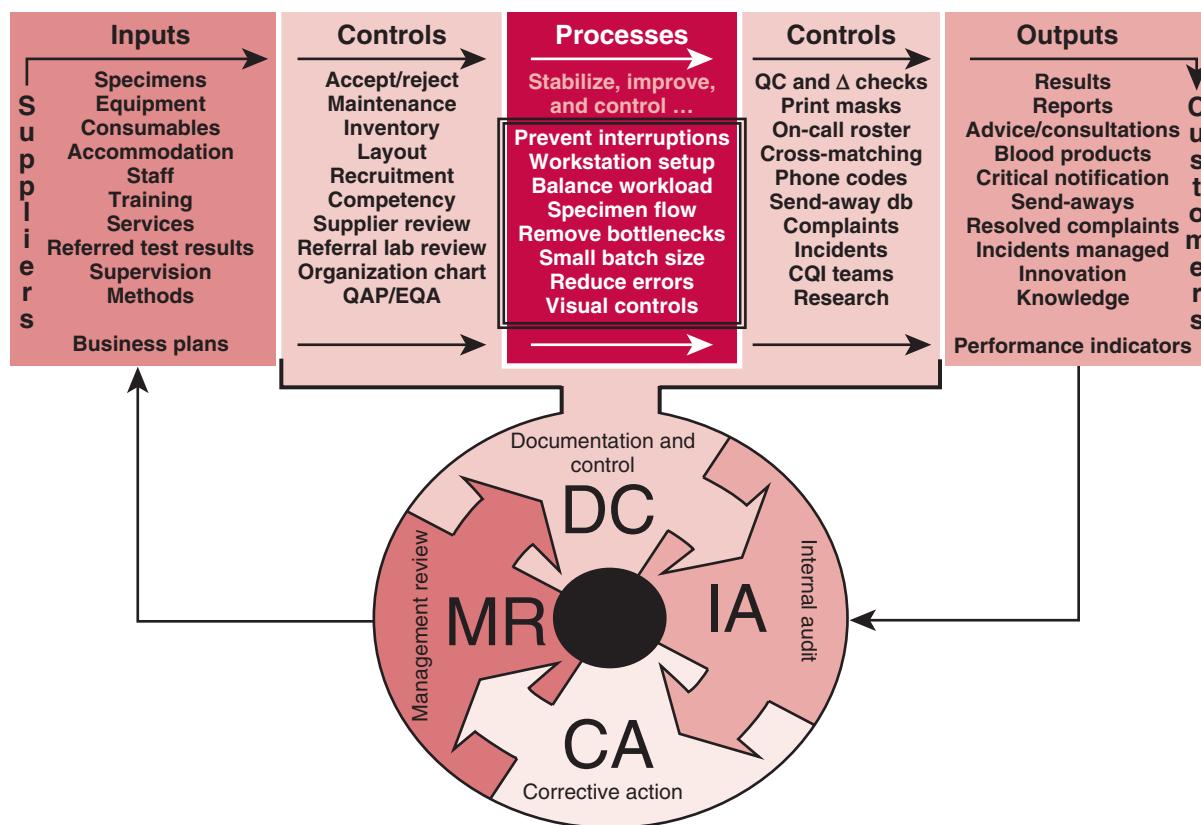
### **Process Approach**

The standardized ISO description of this principle is that “A desired result is achieved more efficiently when activities and related resources are managed as a process.”

ISO 9001 defines a process as “a set of interrelated or interacting activities which transform inputs into outputs” and the related principle of a “system approach to management” involves managing the interrelationship between processes, where the output of one process becomes the input of another. For example, the collection of a blood sample is followed by the separation of serum from that blood sample.

A diagram that identifies suppliers, inputs, processes, outputs, and customers (SIPOC) is of value in implementing a process approach and the associated system approach. A SIPOC diagram (see Fig. 3.6) describes what comes into the laboratory (inputs) and who provides those things (suppliers), what activities occur (processes), what the laboratory produces or delivers (outputs), and who receives those (users or customers). This diagram also identifies basic metrics that must be understood to run the business, such as money, staffing, inventory, requests or orders, and reports.

An important modification to the traditional SIPOC diagram is to identify what controls need to be in place for each key input and output so that intended quality and service delivery meet customer requirements. These controls need to be documented and to describe areas where quality indicators are appropriate. Internal audit is used to confirm that actual practice corresponds with that intended, and correction actions (including both preventive and corrective action) is taken to address and fix identified problems and nonconformities. Preventive action and risk management techniques can be used to identify and resolve potential problems and nonconformities. Finally, all the above feed into a reporting and review cycle, called management review.



**FIGURE 3.6** Suppliers, inputs, processes, outputs, and customers (SIPOC) diagram for a medical laboratory. This SIPOC diagram shows the laboratory as a process with matching controls on its inputs, processes, and outputs; four key quality management system elements control the whole system. It illustrates how the various elements in standards (International Organization for Standardization [ISO] 15189, QMS01), business plans, and performance indicators all fit into day-to-day running of the laboratory. CQI, Continuous quality improvement; CA, corrective action; DC, documentation and control; IA, internal audit; MR, management review; QAP/EQA, quality assurance program/external quality assurance; QC, quality control. (Diagram courtesy Mark Mackay and Collin Sheppard. Note that some specific procedures referred to in the diagram have used slightly different terms from those referred to in this chapter; this is a matter of local laboratory usage.)

Thus quality systems provide a structured framework to ensure the system's processes work as intended. As part of this framework, improvement actions require one to look for the underlying "root cause" of the problem and to show evidence that the proposed solution has fixed the problem before the solution can sign off or "close" the improvement action.

The SIPOC diagram (Fig. 3.6) can be compared with Figs. 3.2 and 3.3 to assist in understanding where standards such as ISO 15189 or guidelines such as QMS01 overlay onto the overall business model.

### Systems Approach to Management

The standardized ISO description of this principle is that "Identifying, understanding, and managing interrelated processes as a system contributes to the organization's effectiveness and efficiency in achieving its objectives."

As described earlier, processes work together as part of a system. A medical laboratory is a system with many complex interlinked processes with positive and negative feedback loops, including workload management, supply ordering, reflexive testing, critical risk result notification, and consultation. Note that some processes compete with others for the

same key resources: optimization of one process may result in degradation of performance of some other process. In QM, one seeks to optimize the overall system to maximize the quality delivered to the ultimate user or customer, through optimizing the quality of each process in delivering to those internal customers who serve the ultimate customer.

By redrawing laboratory workflow as a SIPOC diagram and then adding key supporting processes and the path of materials and data (e.g., supply ordering and inventory, the flow of paper, data to and from the laboratory information system, critical risk result notification), one can generate a map of the key processes that are essential for optimal laboratory performance. For each process, one identifies what is important to perform correctly in each step and at each handover or interface. These become the critical control points with definitions of what is expected and acceptable, which in turn leads to identification of the priority areas for quality improvement.

### Continual Improvement

The standardized ISO description of this principle is that "Continual improvement of the organization's overall performance should be a permanent objective of the organization."

ISO 9000 defines continual improvement as “a recurring activity to increase the ability to fulfill requirements,” and the clause in ISO 9001 provides some explanation of how this might be achieved—the organization shall continually improve the effectiveness of the quality management system, through the use of its quality policy, quality objectives, audit results, analysis of data, corrective and preventive actions, and management review.”

QM is a continuous systematic improvement process.<sup>37,40,41</sup> Many tools and techniques are available<sup>42–45</sup> to identify improvement opportunities and to discover how to address them. In healthcare environments, a common technique has been the PDCA cycle. In recent years, the Six Sigma improvement method<sup>44–48</sup> which uses a five-step cycle of “define, measure, analyze, improve, and control” (DMAIC) is becoming widely used. A number of case studies<sup>49–54</sup> have shown that these can be applied to medical laboratories as readily as to other fields of endeavor.

As an organization’s QMS matures, it moves from reacting to problems to preventing problems from happening. For example, it goes from being reactive to a nonconformity, to becoming proactive through management review of performance and adopting preventive actions and risk management strategies to error-proof processes.

### Factual Approach to Decision Making

The standardized ISO description of this principle is that “Effective decisions are based on the analysis of data and information.”

QM is a culture that is data rich and based on facts; it uses data to drive decision making.

**Measures and quality indicators.** Measures<sup>49</sup> should cover organizational performance, key customer expectations, departmental or divisional indicators, indicators on critical steps in the process, and failures or errors in processes or outcomes. Reasons for collecting data include understanding how processes actually work, examining if they can meet customer needs, monitoring ongoing performance, identifying improvement opportunities, and aggregating into overall organizational performance.

Laboratory quality indicators should provide evidence of a laboratory’s contribution to critical healthcare domains (i.e., patient safety, effectiveness, equity, patient-centeredness, timeliness, and efficiency).<sup>50</sup> More often, quality indicators are focused on laboratory resources, processes, or outcomes rather than the patient. However, it is possible to combine the approaches.

Indicators on the efficiency of resource utilization come from ratios of the important metrics for running the laboratory, which we suggest can be easily identified using a SIPOC diagram (see Fig. 3.6). The efficiency of a system of production always involves the cost of poor quality (i.e., avoiding all types of waste and rework along the process). It is also important to ensure that patient-centeredness (e.g., patient needs, preferences, education, and support) and patient safety (risk) are emphasized at each stage and interface along the laboratory total testing process by asking what is important for each type of patient and for an individual patient in addition to asking what is important for operating the laboratory. One needs to start by targeting what is important to get right and what can or does go wrong, determine the risk and approach to risk, then accept/reject criteria, and then implement real-time controls and be able to measure nonconformities and failures (see Critical Control Points, Figs. 3.7 and 3.8).

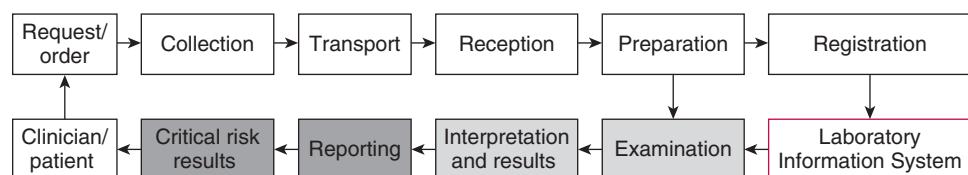
The ability to stratify indicator data by patient type or stream (e.g., by ward, clinic, or source) is important because it contributes to assessing the equity of patient care provided by the laboratory. This is a major reason why many laboratory indicators are often best recorded in the laboratory information system, which has easy access to patient data that can be de-identified and aggregated for summary reporting and analysis.

It is recommended to use only a few generic and a few specific measures at the relevant level (e.g., local department, process, or customer requirement). The more specific the measure, the more it will be restricted to that department or process. To aggregate multiple measures, one can convert them into percentage compliance with the defined target.

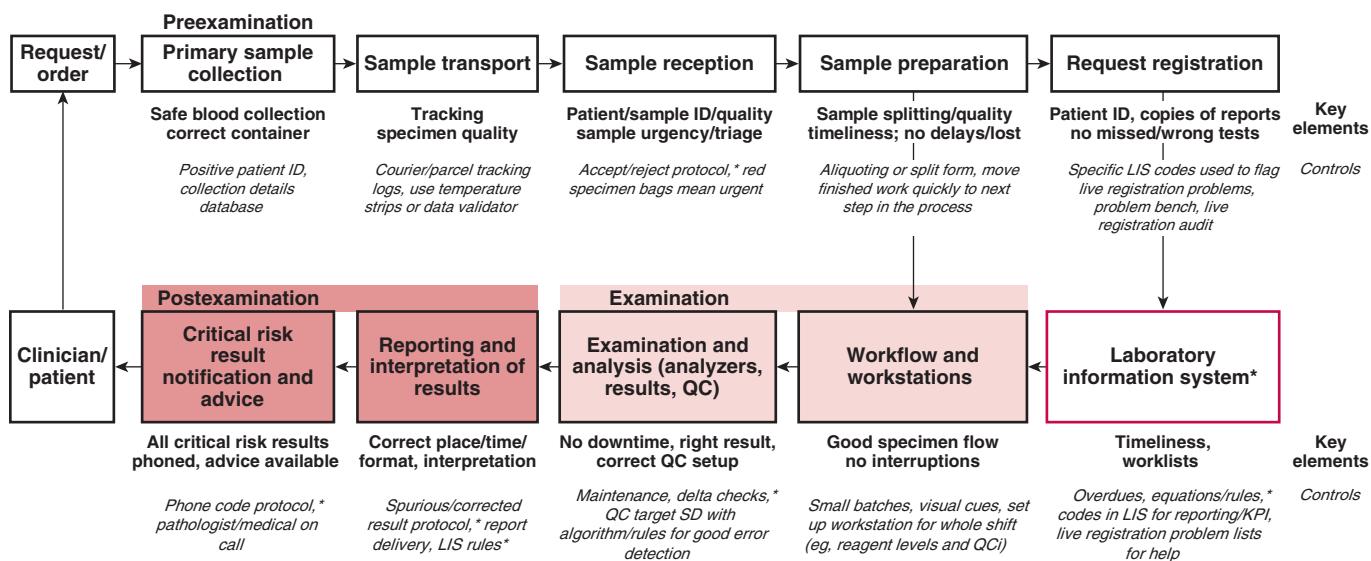
Although percent compliance indicators are outcome measures, effectiveness appears to be the weakest area of laboratory quality indicators because often indicators are not linked to patient outcomes.<sup>50</sup> Some studies go to the considerable trouble of making this link to patient outcomes.<sup>51,52</sup>

Before implementing potential quality indicators, their importance, scientific soundness, and feasibility need to be considered. Ideally, they require stakeholder interest and should target a clinically important area of healthcare that is amenable to improvement and that can be influenced. In addition, the measures should be reliable, valid, easily understood, and have explicit specifications with data available that are accessible in a timely way at justifiable cost for the potential improvement in patient care.<sup>50</sup>

It is unlikely that any single laboratory indicator satisfies all conditions. Therefore a distinction should be made among



**FIGURE 3.7** Basic laboratory workflow before adding critical controls and control points. A flowchart of materials and data is used to (1) identify the critical success factors at each physical step and interface in the process and then (2) define appropriate controls and to confirm that they are in place, they are documented, and staff are trained in their use. The laboratory information system can easily be set up as a process control (e.g., entering codes for problems which can generate work lists) and as a data repository used as a reporting tool for problems or errors, performance data, and key performance indicators. Compare with Fig. 3.8. Preanalytical steps are shown in white, analytical steps shaded light gray, and postanalytical steps are dark gray.



**FIGURE 3.8** Basic laboratory workflow after adding critical controls and control points. This is an example showing a general flow, key things to get right (key elements), and matching controls. Compare with Fig. 3.7 Many controls can operate in the laboratory information system (highlighted by an asterisk). Each key customer workstream should be examined for critical control points, including the needs and safety of individual patients, and subjected to end-to-end operational audits on performance and on compliance with customer requirements. KPI, Key performance indicator(s); LIS, laboratory information systems; QC, quality control. (Diagram courtesy Mark Mackay and Collin Sheppard.)

internal measures for control, standardized measures for benchmarking, and quality indicators that can drive improvement in patient care.

In addition to internally derived indicators, it is possible to participate in a number of national and international programs on quality indicators or laboratory performance for benchmarking. Under the auspices of IFCC, great strides have been made in proposing a set of standardized laboratory indicators<sup>53–57</sup> and other jurisdictions have also proposed indicators.<sup>58,59</sup> The CAP offers Q-Probes<sup>60,61</sup> and Q-Track<sup>62</sup> programs. The Royal College of Pathologists of Australasia (RCPA) Quality Assurance Programs offer the Key Incident Monitoring and Management Systems (KIMMS)<sup>63</sup> for pre- and postanalytical issues. KIMMS provides both benchmarked frequency of recorded laboratory problems and an analysis based on risk using a modified FMEA approach.

The most effective measures for improvement are simple, robust, and often in the hands of the front-line operators because staff who own their data are driven to maximize the benefits. The data can come from many sources, such as the laboratory information system, or can be measured manually. Some measures are designed for immediate “go/no go” decisions (e.g., specimen arrival flux to allow for workload rebalancing, overdue specimens), and others are used to track performance against organizational goals (e.g., QC failures, unscheduled maintenance or downtime, proportions of specimens meeting turn-around time targets). It is common practice to deploy a range of these controls and indicators, each deployed where most applicable in the laboratory.

**Understanding variation.** All systems and processes have variation. It is much easier to control and optimize a system in which variation has been reduced. Variation can increase along a system of successive processes because unstable inputs to a process make it harder to control that process, affecting its outputs that then feed the next step. Consequently,

systems are harder to control than processes because one must balance multiple processes within the overall system.

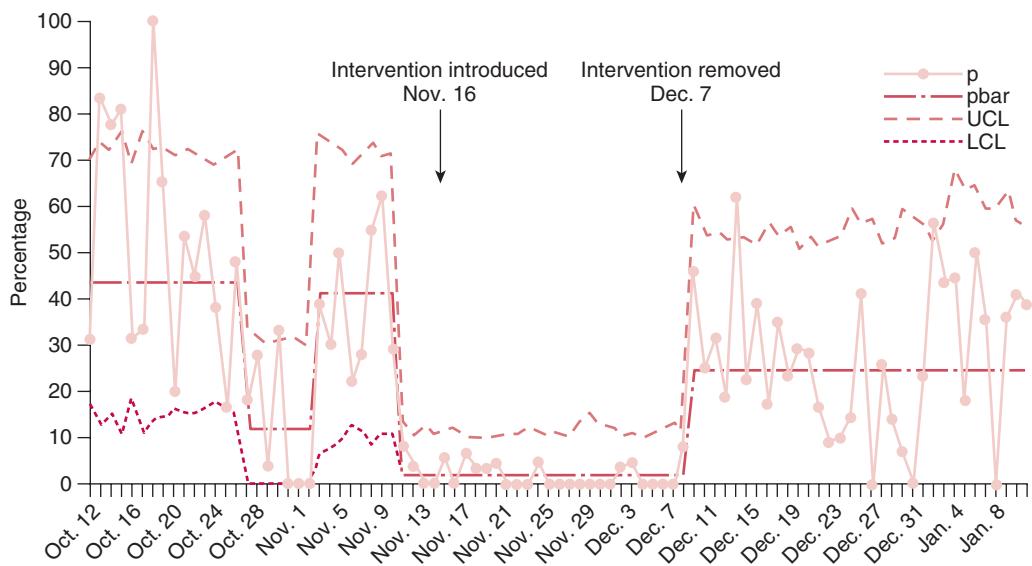
Causes of variation can be classified into two types: common cause and special cause. Whereas common cause variation is due to the performance characteristics of the underlying process, special cause variation is due to external factors impinging on the process.

SPC<sup>39,40,64,65</sup> (see also Chapter 6) can be used to control and improve processes. By adding statistical limits to the measurement of a process, a run chart is converted into control chart, which then easily highlights out-of-control situations. This allows one to react to variation appropriately so as to minimize the likelihood of harmful tampering with an already statistically stable system, which would make the variation worse.<sup>35</sup>

Control limits are a statistical way to show the natural variation in a process. Individual results can be compared with customer expectations (converted into the language and definitions of laboratory specifications) to see if they pass or fail, but the aim is to have the process control limits well within such specifications to ensure that the process can reliably meet customer expectations. If average performance just meets customer specifications, then 50% of results will fail to meet customer expectations.

Control charts (see also Chapter 6) facilitate analysis of the causes of variation. Special causes make a process go out of control. Common cause is the variation between control limits. It is generally easier to first control special cause variation before then tackling common cause variation. This is analogous to improving analytical performance by firstly addressing imprecision before addressing bias.

The critical control points identified in the SIPOC workflow analysis are used to monitor the adequacy of performance of these key processes. For analytical processes, the critical control points would correspond to the QC rules



**FIGURE 3.9** Example of statistical process control rules applied to nonanalytical processes in the laboratory. This example shows the effect of providing requesting clinicians with an aid (in this case, a simple tool to avoid the need to reenter basic invariant self-identification data) to facilitate providing complete information on a pathology request. The resultant quality of output was measured using a Shewhart P control chart (65) to track the proportion of incomplete requests submitted to the laboratory. *LCL*, Lower control limit; *p*, proportion; *pbar*, average proportion; *UCL*, upper control limit ( $P \approx .997$ ).<sup>69</sup>

(see Chapter 6); similar QC rules can be applied to nonanalytical processes throughout the organization (e.g., Fig. 3.9; additional examples are in the cited case studies).<sup>49,66–70</sup>

**Cost of quality (including waste).** “The costs connected with both attaining and failing to attain the desired level of quality in a service or product are considered quality costs. Quality costs include the cost of preventing problems; the cost of measuring, controlling, and/or inspecting quality levels; and the cost of failing to accomplish the desired quality levels.”<sup>71</sup> The lowest cost to provide a service or product is to do it right the first time, every time. When there is unnecessary or additional work in the process it adds cost, often viewed as waste, and results in a lower level of quality.

Signals of poor quality are delays, interruptions, rework, waste, complaints, and so on. The cost of poor quality has often been estimated at about 25% of total cost, so there is dramatic potential for improvement by focusing on this area.<sup>72</sup> The places to look for improvement include:

- Design: being proactive and preventing errors in the first place
- Performance: natural variation of the process may be too large to meet customer or desirable specifications
- Conformance: where marginal quality can be accepted, and failures can be both detected and missed

Each of these places can contribute to higher risk and extra work that detract from the efficiency of the process. A process can be redesigned to make it error proof and reduce the natural variation. All of this can be done using QM tools.

Lean Six Sigma<sup>3</sup> is a combination of two continual improvement methodologies that are together used to improve process quality. “Lean” focuses on reducing waste. There are seven commonly identified types of waste: waiting, overproduction, rework, motion, processing, inventory, and transportation. In recent years within the quality community, one additional type of waste is commonly added to the list, human

potential. “Six Sigma” focuses on reducing variation in a process. Organizations generally start improvement efforts by identifying and removing process waste, which often also results in reduction in process variation. When waste and variation in a process are eliminated or minimized, the result is increased capacity through better organization, not extra effort.

**Nonconforming events and risk management.** When things go wrong,<sup>73,74</sup> it is termed a nonconformity, regardless of whether a patient has been affected. ISO 15189 defines a nonconformity as “nonfulfillment of a requirement.” NCEs must be managed using risk; that is, the actions taken must be proportional to the risk, both in preventing a situation and addressing the NCE. This includes appropriate escalation and direct notification of high-risk events. When a patient has been harmed or there was a potential for patient harm, it is a requirement to conduct a detailed investigation, identify the root cause(s), implement corrective action, and document all activities.

A laboratory may classify risk by assessing the severity level and probability of occurring again. Although there are many factors to consider in each individual NCE, the use of a classification matrix can provide general guidance (see Table 3.3).<sup>75</sup> These risk matrices often use color coding to assist prioritizing risk classifications (e.g., red: critical severity and high probability of recurrence; yellow: intermediate severity and probability of recurrence; green: minor severity and low probability of recurrence). The communication expectations of NCEs vary by organization and country and should be defined in the NCE procedure. For example, regions on the risk matrix colored red may require a detailed investigation to identify and eliminate the root cause(s), and yellow regions may indicate that although immediate correction of the problem may be required, more detailed investigation might not be.

FMEA is another tool that can be used to assess risk along a process by examining and rating each way a step can fail. In addition to severity and consequence, FMEA adds the ability

to detect a problem as a factor because if a problem goes undetected, the severity can compound. The big advantage is that FMEA can be used in before-and-after mode to assess the reduced risk from proposed interventions (see Risk section).

### Mutually Beneficial Supplier Relationships

The standardized ISO description of this principle is that “An organization and its suppliers are interdependent, and a mutually beneficial relationship enhances the ability of both to create value.”

Because variation can increase along a series of processes, it is important that laboratory inputs are controlled and stable and that they meet laboratory needs. Key laboratory inputs include all of staff, money, specimens, equipment, supplies, and inventory management, and each of these needs to be optimized as a key input into the overall laboratory system. Effective communication between an organization

and its suppliers allows for joint sharing of expertise and resources,<sup>76</sup> improving inventory utilization and management processes, and planning for future needs.<sup>77</sup>

### Summary of Useful Quality Management Tools and Resources

There are many QM approaches and tools. Some were developed from principles or theories, but many were practical implementations developed by organizations that found new ways to improve. Over time, these improvement methods were shared with other organizations and adopted when it was a good fit for their culture, environment, and improvement needs. Tools are selected based on the type of improvement needed and the expected outcome. A sample of commonly available approaches, resources, and tools organized by the eight QM principles that have proved useful in clinical laboratories is summarized in Table 3.6.

**TABLE 3.6 Suggested Approach to Using Quality Management Principles and Core Tools<sup>6</sup>**

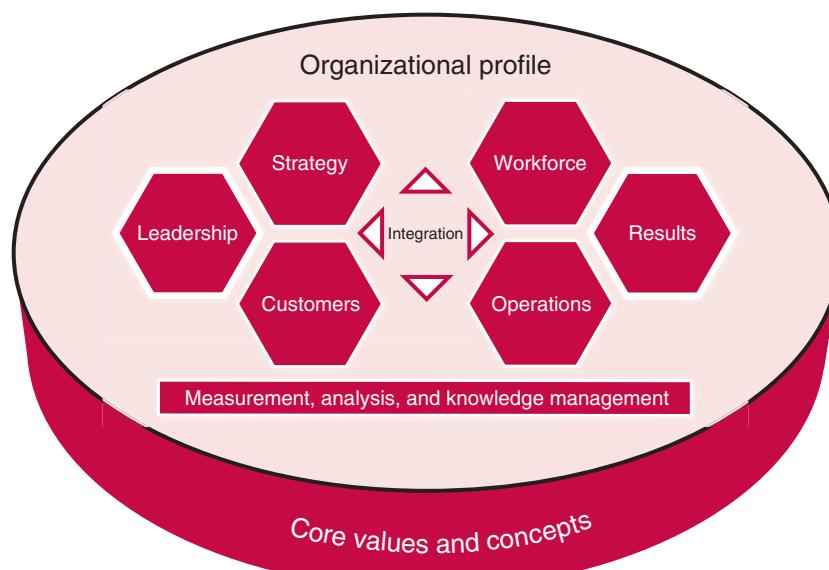
Principle	Improvement Approach, Resources, or Tools
Leadership	<ul style="list-style-type: none"> <li>Establish an organizational chart with roles and responsibilities, and a planning process</li> <li>Gap analysis of organizational performance and culture</li> <li>Use QM principles to generate a planning process and quality plan based on what is needed to meet customer requirements in addition to business, technical and regulatory requirements</li> <li>Ensure adequate resources for the improvement effort are made available</li> <li>Recognize teamwork, innovation, and milestones</li> <li>Management review feeds the planning cycle</li> </ul>
Involvement of people	<ul style="list-style-type: none"> <li>Establish job descriptions and a competency system</li> <li>Establish improvement teams involving local staff on key issues, scoped, with team goals, time frames, and teamwork guidelines</li> <li>Teamwork tools</li> <li>Recognition for teamwork, innovation, and milestones</li> </ul>
Systems approach to management	<ul style="list-style-type: none"> <li>Establish a document control system that holds procedures, including the planning process, how controls work, and how performance is reported and reviewed</li> <li>Internal audit process to check compliance with procedures</li> <li>Use SIPOC to identify customers, suppliers, controls, metrics, and ensure processes are cooperative, not competing</li> <li>Turnaround time</li> </ul>
Process approach	<ul style="list-style-type: none"> <li>Generic path of workflow (key or core process diagram)</li> <li>Add generic critical control points and requirements for patient needs and safety; then add accept/reject criteria</li> <li>Create real-time visual controls</li> <li>Document critical controls and real-time visual controls and train staff in their use</li> </ul>
Cost of quality (including scrap and waste)	<ul style="list-style-type: none"> <li>A practical focus on reducing waste, bottlenecks, and delays using what is at hand and not waiting for capital injections</li> <li>Laboratory layout (spaghetti diagram)</li> <li>Workstation decluttering, setup, handovers, maintenance (Lean 5S methodology)</li> <li>Include reagent and supply usage and inventory system (see supplier relationships)</li> <li>Identify process bottlenecks and capacity (Lean value stream mapping)</li> </ul>
Customer focus	<ul style="list-style-type: none"> <li>Key customer questions: Who are the users or customers? What do they need? How well do users or customers think their needs are being met? How well does the laboratory think it currently performs?</li> <li>Obtain voice of the customer; confirm user or customer requirements with user or customer</li> <li>Identify critical controls along customer path of workflow and perform walk through improvement audit</li> <li>Overlay patient safety on path of workflow</li> <li>Establish critical customer performance indicators</li> </ul>
Factual approach to decision making	<ul style="list-style-type: none"> <li>Indicators from SIPOC, accept/reject, critical control points, nonconforming events, critical customer performance indicators</li> <li>Aim for a small set of focused indicators, including leading indicators (e.g., process), not just outcome indicators</li> </ul>
Understanding variation	<ul style="list-style-type: none"> <li>Statistical process control</li> <li>General capability and Six Sigma</li> </ul>

**TABLE 3.6 Suggested Approach to Using Quality Management Principles and Core Tools—Cont'd**

Continual improvement	<ul style="list-style-type: none"> <li>Improvement teams focus on important gaps in performance using basic QM tools (e.g., process maps or flow charts, cause-and-effect diagrams, histograms, Pareto charts, control charts, scatter diagram, check, or data sheets)</li> <li>Teams follow improvement step-by-step model (e.g., PDCA, DMAIC)</li> <li>Small step improvements plus innovation</li> <li>Preventive action</li> <li>Management review</li> </ul>
Nonconforming events and risk management	<ul style="list-style-type: none"> <li>Establish an incident reporting system that complies with local regulations and organizational governance and meshes with the local system of clinical governance covering patients</li> <li>Establish a safety program in which actions are appropriate to risk</li> <li>Nonconforming event management</li> <li>Corrective action</li> <li>Root cause analysis</li> <li>FMEA</li> </ul>
Mutually beneficial relationships with suppliers	<ul style="list-style-type: none"> <li>Data from SIPOC</li> <li>Workstation stock and reagent usage</li> <li>Inventory control management</li> <li>Supplier relationships</li> </ul>

<sup>a</sup>The quality management (QM) principles and core tools align with QM systems elements to improve and control laboratory performance to meet customer needs.

DMAIC, Define, measure, analyze, improve, and control; FMEA, failure mode effects analysis; PDCA, plan, do, check, act; SIPOC, suppliers, inputs, processes, outputs, and customers.



**FIGURE 3.10** Example of a business excellence framework (Modified from Baldrige Performance Excellence Program. 2015–2016 *Baldrige excellence framework: a systems approach to improving your organization's performance [health care]*. Gaithersburg, MD: U.S. Department of Commerce, National Institute of Standards and Technology. <<http://www.nist.gov/baldrige>>; 2015.)

### Organizational Quality and Business Excellence Frameworks

To recognize and reward outstanding organizations, many countries have introduced annual national quality awards. Among the best known of these are the Deming Prize (established in 1951 in Japan) and the Malcolm Baldrige National Quality Award (established in 1988 in the United States).<sup>78</sup>

Award recipients are required to demonstrate outstanding performance across a full range of activities within a QM framework (Fig. 3.10). Performance is rated using a

gap-analysis scorecard, which has predefined criteria and milestones in each dimension of this framework. For each of these dimensions, the organization is rated on its approach, deployment, results, and further cycles of improvement (ADRI), which is an organization-wide version of Plan, Do, Check, Act.

To assist aspiring organizations that wish to improve their own performance, these scorecards can be applied<sup>49</sup> to less mature organizations to highlight shortcomings in existing management practices, which can then be targeted for improvement.

## POINTS TO REMEMBER

QM applies to the entire organization, not only to the laboratory's technical components.

Long-term and sustained success and maintenance of QM requires establishment of a quality culture by the executive leadership of the organization.

The eight ISO QM principles relate to:

- Customer focus.
- Leadership.
- Involvement of people.
- Process approach.
- System approach to management.
- Continual improvement.

- Factual approach to decision making.

- Mutually beneficial supplier relationships.

Some authorities have noted that there are additional key principles, the three major ones being:

- Understanding variation.
- Cost of quality (including waste).
- NCEs and risk management.

A QMS uses a systematic and process-focused approach to meet the needs of its customers, defined in the quality goals and objectives.

A variety of tools and techniques are available for organizations to adapt to their needs and local organizational culture. Many of these tools are adaptations of the scientific method.

## SELECTED REFERENCES

1. ISO 15189:2012 Medical Laboratories – Requirements for quality and competence; 3rd edition.
8. Freedman DB. Clinical governance- bridging management and clinical approaches to quality in the UK. *Clin Chim Acta* 319 (2002) 133–141.
14. European Foundation for Quality Management (2005), EFQM Framework for Risk Management, European Foundation for Quality Management, Brussels.
21. Perspectives on Quality Control, Risk Management, and Analytical Quality Management. Westgard JO. *Clinics in Laboratory Medicine*, 2013, 33(1); 1-14.
25. Burnett D. A Practical guide to ISO 15189 in laboratory medicine. ACB Venture Publications; 2013. (ISBN 978-0-902429-49-9).
28. CLSI. Quality management system: a model for laboratory services; approved guideline-fourth edition. CLSI document QMS01-A5. Wayne, PA: Clinical and Laboratory Standards Institute; 2019.
36. Walton M. The Deming management method. Perigree. 1986.
37. Imai M. Gemba Kaizen. A commonsense low-cost approach to a continuous improvement strategy. 2nd ed. McGraw-Hill; 2012.
39. Deming WE. Out of the crisis. MIT Press; 2000. (ISBN 9780262541152).
40. Juran JM, De Feo JA. Juran's quality handbook. 6th ed. McGraw-Hill; 2010.
42. Brassard M, Ritter D, Oddo F, et al. The Memory Jogger II healthcare edition: a pocket guide of tools for continuous improvement and effective planning. 2008.
43. Sholtes PR, Joiner BL, Streibel BJ. The team handbook. 3rd ed. Released 2003. (ISBN-13: 978-1884731266).
46. George M, Rowlands D, Price M, et al. Lean Six Sigma pocket toolkit. McGraw-Hill; 2002.
47. George M, Rowlands D, Kastle B. What is Lean Six Sigma? McGraw-Hill; 2004.
56. Plebani M, Astion ML, Barth JH, et al. Harmonization of quality indicators in laboratory medicine. A preliminary consensus. *Clin Chem Lab Med* 2014;52(7):951-8.
59. CLSI. development and use of quality indicators for process improvement and monitoring of laboratory quality; approved guideline. CLSI document QMS12-A. Wayne, PA: Clinical and Laboratory Standards Institute; 2010. (ISBN: 1-56238-738-3).
65. Grant E, Leavenworth R. Statistical quality control. 7th ed. McGraw Hill.
66. Banning J, Brown J, Hooper L, et al. Reduction of errors in laboratory test reports using Continuous Quality Improvement techniques. *Clin Lab Manag Rev* 1993;7:424-6.
75. CLSI. Nonconforming event management: second edition, CLSI document QMS11-A2. Wayne, PA: Clinical and Laboratory Standards Institute; 2015 (Appendix D, Table D4).
78. Baldrige Performance Excellence Program. 2019–2020 Baldrige Excellence framework: a systems approach to improving your organization's performance (health care). Gaithersburg, MD: U.S. Department of Commerce, National Institute of Standards and Technology; 2019 <https://www.nist.gov/baldrige/publications/baldrige-excellence-framework/health-care>.

## REFERENCES

1. ISO 15189:2012 Medical Laboratories – Requirements for quality and competence; 3rd edition.
2. ISO 9000:2015 Quality Management Systems – Fundamentals and Vocabulary.
3. Pande PS, Neuman RP, Cavanagh RR. The Six Sigma way. 2nd ed. McGraw-Hill; 2014.
4. Institute of Medicine (US) Committee on Quality of Health Care in America; Kohn LT, Corrigan JM, Donaldson MS, editors. To Err is Human: Building a Safer Health System. Washington (DC): National Academies Press (US); 2000. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK225182/doi:10.17226/9728>.
5. Quality, Risk Management and Governance in Mental Health: An Overview. Tom Callaly, Dinesh Arya, Harry Minas. Austral Psychiatry. 2005;13(1):16-20.
6. Donaldson EJ, Gray JAM. Clinical governance—a quality duty for health organizations. Quality in Healthcare 1998;7: S37 – 44, Suppl.
7. Swage T. Clinical governance in health care practice. 1st ed. Oxford: Butterworth-Heinemann, 2000.
8. Freedman DB. Clinical governance- bridging management and clinical approaches to quality in the UK. Clin Chim Acta 319 (2002) 133–141.
9. Benchmarking Laboratory Quality. Valenstein P, Schneider F. Lab Medicine 39(2) 108-112) 2008.
10. Howanitz PJ. Quality assurance measurements in departments of pathology and laboratory medicine. Arch Pathol Lab Med. 1990;114:1131-1135.
11. Zarbo RJ, Jones BA, Friedberg RC, et al. Q-Tracks. A College of American Pathologists program of continuous laboratory monitoring and longitudinal performance tracking. Arch Pathol Lab Med. 2002;126:1036-1044.
12. The Benchmarking Partnership. [www.thebenchmarkingpartnership.com](http://www.thebenchmarkingpartnership.com) (accessed 18 January 2020).
13. College of American Pathologists. Q-Probes. <https://www.cap.org/laboratory-improvement/quality-management-programs> (accessed 18 January 2020).
14. European Foundation for Quality Management (2005), EFQM Framework for Risk Management, European Foundation for Quality Management, Brussels.
15. Quality and risk management: what are the key issues? Williams R, Bertsch B, Dale B, van der Wiele T, van Iwaarden J, Smith M, Visser R. TQM 18(1); 2006: 67-86.
16. Teare EL, Masterton RG. Risk management in pathology. J Clin Pathol. 2003;56(3):161-163. doi:10.1136/jcp.56.3.161.
17. NPAAC (National Pathology Accreditation Advisory Council), Australian Government Department of Health. <<http://www.health.gov.au/internet/main/publishing.nsf/Content/health-npac-publication.htm>>.
18. Dias S, Saraiva PM. (2004). Use Basic Quality Tools To Manage Your Processes. Quality Progress, 37(8), 47-53. [http://sharptinkers.info/articles/Basic\\_Tools.pdf](http://sharptinkers.info/articles/Basic_Tools.pdf).
19. System Reliability Theory: Models, Statistical Methods, and Applications, Marvin Rausand, Arnljot Hoylan, Wiley Series in probability and statistics—second edition 2004.
20. An Improved Failure Mode Effects Analysis for Hospitals, Krouwer JS. Arch Path Lab Med 2004 128:6, 663-667.
21. Perspectives on Quality Control, Risk Management, and Analytical Quality Management. Westgard JO. Clinics in Laboratory Medicine, 2013, 33(1); 1-14.
22. ISO 14971:2019. Medical devices—application of risk management to medical devices. Geneva (Switzerland):
23. CLSI EP23A Laboratory Quality Control based on risk management. <https://clsi.org/standards/products/method-evaluation/documents/ep23/>. (accessed 19 January 2020).
24. Quality management principles. <[http://www.iso.org/iso/qmp\\_2012.pdf](http://www.iso.org/iso/qmp_2012.pdf)>.
25. Burnett D. A Practical guide to ISO 15189 in laboratory medicine. ACB Venture Publications; 2013. (ISBN 978-0-902429-49-9).
26. ISO 22870:2016 Point-of-care testing (POCT) – Requirements for quality and competence.
27. ISO/IEC 17043:2010 (R2015) Conformity assessment – General requirements for proficiency testing.
28. CLSI. Quality management system: a model for laboratory services; approved guideline-fourth edition. CLSI document QMS01-A5. Wayne, PA: Clinical and Laboratory Standards Institute; 2019.
29. Collins English dictionary, Harper Collins; 1995. (ISBN 0 00 470678-1).
30. Principles of clinical laboratory accreditation, A policy statement. IFCC and WASPaLM, 1999. <[http://www.acibademylabmed.com/tr/ifcc\\_accreditation.pdf](http://www.acibademylabmed.com/tr/ifcc_accreditation.pdf)>.
31. Burnett D. Standards and accreditation for medical laboratories post Carter. Bulletin RCPPath 2009;146:166-9.
32. Clinical Laboratory Improvement Amendments (CLIA) <<https://www.cdc.gov/clia/>>; [Accessed 07.10.15].
33. European Diagnostic Manufacturers Association. Supporting information to the Suppliers' Charter on Clinical Laboratories seeking accreditation to EN ISO 15189:2012. <[http://www.edma-ivd.eu/uploads/PositionPapers/2014\\_10\\_07\\_EDMA\\_SuppliersCharter\\_SupportingInformationDoc\\_ENISO15189\\_2012\\_PUB.pdf](http://www.edma-ivd.eu/uploads/PositionPapers/2014_10_07_EDMA_SuppliersCharter_SupportingInformationDoc_ENISO15189_2012_PUB.pdf)>.
34. The National Association of Testing Authorities, Australia <<http://www.nata.com.au>>.
35. Deming WE. The new economics for industry, government and education. 2nd ed. MIT Press; 1994.
36. Walton M. The Deming management method. Perigree. 1986.
37. Imai M. Gemba Kaizen. A commonsense low-cost approach to a continuous improvement strategy. 2nd ed. McGraw-Hill; 2012.
38. Bonini P, Plebani M, Ceriotti F, et al. Errors in laboratory medicine. Clin Chem 2002;48:691-8.
39. Deming WE. Out of the crisis. MIT Press; 2000. (ISBN 9780262541152).
40. Juran JM, De Feo JA. Juran's quality control handbook. 6th ed. McGraw-Hill; 2010.
41. De Feo JA, Gruna FM. Juran's quality management and analysis. 6th ed. McGraw-Hill; 2015.
42. Brassard M, Ritter D, Oddo F, et al. The Memory Jogger II healthcare edition: a pocket guide of tools for continuous improvement and effective planning. 2008.
43. Sholtes PR, Joiner BL, Streibel BJ. The team handbook. 3rd ed. Released 2003. (ISBN-13: 978-1884731266).
44. Arthur J. Breakthrough improvement with QI Macros and Excel. Finding the low-hanging fruit. McGraw-Hill; 2014.
45. Arthur J. Lean Six Sigma demystified. Place: McGraw-Hill; 2007.
46. George M, Rowlands D, Price M, et al. Lean Six Sigma pocket toolbook. McGraw-Hill; 2002.
47. George M, Rowlands D, Kastle B. What is Lean Six Sigma? McGraw-Hill; 2004.
48. George ML. Lean Six Sigma. Combining Six Sigma quality with Lean speed. McGraw-Hill; 2002.

49. Burnett L. Applying the Australian Quality Awards criteria to a clinical chemistry department. *Quality Management in Healthcare* 1994;3(1):1-15.
50. Shahangian S, Snyder SR. Laboratory Medicine quality indicators—a review of the literature. *Am J Clin Pathol* 2009;131:418-31.
51. Georgiou A, Vecellio E, Toouli G, et al. The impact of the implementation of electronic ordering on hospital pathology services. Report to Commonwealth of Australia, Department of Health and Ageing, Quality Use of Pathology Committee. Sydney: Australian Institute of Health Innovation, University of New South Wales; 2012. (ISBN: 978-0-7334-3194-4).
52. Vecellio E, Li L, Mackay M, et al. A benchmark study of the frequency and variability of haemolysis reporting across pathology laboratories – the implications for quality use of pathology and safe and effective patient care. Report to Royal College of Pathologists Australasia. Sydney: Australian Institute of Health Innovation, Macquarie University; 2015.
53. Plebani M. Quality indicators to detect pre-analytical errors in laboratory testing. *Clin Biochem Rev* 2012;33:85-8.
54. Sciacovelli L, Sonntag O, Padoan A, et al. Monitoring quality indicators in laboratory medicine does not automatically result in quality improvement. *Clin Chem Lab Med* 2012;50(3):463-9.
55. Plebani M, Chiozza ML, Sciacovelli L. Towards harmonization of quality indicators in laboratory medicine. *Clin Chem Lab Med* 2013;51(1):187-95.
56. Plebani M, Astion ML, Barth JH, et al. Harmonization of quality indicators in laboratory medicine. A preliminary consensus. *Clin Chem Lab Med* 2014;52(7):951-8.
57. Plebani M, Sciacovelli L, Aita A, et al. Performance criteria and quality indicators for the pre-analytical phase. *Clin Chem Lab Med* 2015;53(6):943-8.
58. The Royal College of Pathologists. Key performance indicators in pathology: recommendations from the Royal College of Pathologists. 2013. <[http://www.rcpath.org/Resources/RCPPath/Migrated%20Resources/Documents/K/KPI%20review%20v3%202011%20final%20TH%20%20web%20v\(2\).pdf](http://www.rcpath.org/Resources/RCPPath/Migrated%20Resources/Documents/K/KPI%20review%20v3%202011%20final%20TH%20%20web%20v(2).pdf)>.
59. CLSI. development and use of quality indicators for process improvement and monitoring of laboratory quality; approved guideline. CLSI document QMS12-A. Wayne, PA: Clinical and Laboratory Standards Institute; 2010. (ISBN: 1-56238-738-3).
60. Howanitz PJ, Perrotta PL, Bashleben CP, et al. Twenty-five years of accomplishments of the College of American Pathologists Q-Probes program for clinical pathology. *Arch Pathol Lab Med* 2014;138:1141-9.
61. Volmar KE, Wilkinson DS, Wagar EA, et al. Utilization of stat test priority in the clinical laboratory. A College of American Pathologists Q-Probes study of 52 institutions. *Arch Pathol Lab Med* 2013;137:220-7.
62. College of American Pathologists (CAP). Quality management quality indicator monitoring guidance. <[http://www.cap.org/apps/docs/laboratory\\_accreditation/qim.pdf](http://www.cap.org/apps/docs/laboratory_accreditation/qim.pdf)>.
63. RCPA Quality assurance programs Pty Ltd. KIMMS. <<https://www.rcpaqap.com.au/products/kimms/>> Accessed 24 march 2021.
64. Shewhart WA. Statistical method from the viewpoint of quality control. Dover Publications; 1986 (originally published in 1939 by the Graduate School of the Department of Agriculture, Washington, D.C.).
65. Grant E, Leavenworth R. Statistical quality control. 7th ed. McGraw Hill.
66. Banning J, Brown J, Hooper L, et al. Reduction of errors in laboratory test reports using Continuous Quality Improvement techniques. *Clin Lab Manag Rev* 1993;7:424-6.
67. Burnett L, Chesher D. Application of CQI tools to the reduction in risk of needlestick injury. *Infect Control* 1995;16(9): 503-5.
68. Burnett L, Chesher D, Burnett JR. Optimizing the availability of 'stat' laboratory tests using Shewhart 'c' control charts. *Ann Clin Biochem* 2002;39:140-4.
69. Burnett L, Chesher D, Mudaliar Y. Improving the quality of information on pathology request forms. *Ann Clin Biochem* 2004;41:53-6.
70. Burnett L, Chesher D. Application of CQI tools to the reduction in risk of needlestick injury. *Infect Control Hosp Epidemiol* 1995;16:503-5.
71. CLSI QMS20-R Understanding the cost of quality in the laboratory; a report.
72. Berry TH. Managing the total quality transformation. ASQC Quality Press. McGraw-ill Inc. 1990.
73. Khoury M, Burnett L, Mackay MA. Error rates in Australian chemical pathology laboratories. *Med J Aust* 1996;165: 128-30.
74. Bryant S. Ensuring quality in all aspects of the pathology cycle. *MJA* 1996;165:125-6.
75. CLSI. Nonconforming event management: second edition, CLSI document QMS11-A2. Wayne, PA: Clinical and Laboratory Standards Institute; 2015 (Appendix D, Table D4).
76. European Diagnostic Manufacturers Association. Suppliers' charter on clinical laboratories seeking accreditation to EN ISO 15189:2012. <[http://www.edma-ivd.eu/uploads/Position-Papers/2014\\_10\\_07\\_Suppliers%20Charter\\_2014\\_V02\\_PUB.pdf](http://www.edma-ivd.eu/uploads/Position-Papers/2014_10_07_Suppliers%20Charter_2014_V02_PUB.pdf)>.
77. Proos A, Sutton D, Burnett L, et al. Taming the supplies and inventory process at Westmead Hospital. *Qual Lett Healthc Lead* 1993;6:83-6.
78. Baldrige Performance Excellence Program. 2019–2020 Baldrige Excellence framework: a systems approach to improving your organization's performance (health care). Gaithersburg, MD: U.S. Department of Commerce, National Institute of Standards and Technology; 2019 <https://www.nist.gov/baldrige/publications/baldrige-excellence-framework/health-care>.

**MULTIPLE CHOICE QUESTIONS**

1. Which of the following laboratory standards is the most widely used around the globe for medical laboratory accreditation?
  - a. ISO 9001
  - b. ISO/IEC Guide 25
  - c. ISO 17025
  - d. ISO 15189
  - e. CLSI QMS01-A4
2. Which of the following is included in accreditation but not in certification?
  - a. Uses one or more Standard documents
  - b. Examines both the QMS and the professional technical competence of the laboratory
  - c. Examines the QMS of the organization
  - d. Examines the professional technical competence of the laboratory
  - e. Performed by a recognized agency, which itself is subject to external agency oversight
3. Which of the following is the best definition of quality?
  - a. To meet or exceed desirable specifications for QC
  - b. To meet or exceed desirable specifications based on biological variation
  - c. To achieve the standard mandated by government regulation (e.g., CLIA)
  - d. To be matched to the requirements of a referring clinician/physician
  - e. To meet the requirements of all stakeholders
4. What is the typical length of time required for a laboratory to achieve consistent quality?
  - a. 12 to 18 months, which is the frequency of accreditation cycles
  - b. Initially 3 months, and thereafter 1 month, which is the frequency of proficiency tests
  - c. One to 3 years, which is the frequency of certification agency cycles
  - d. It varies depending on the stakeholders' quality requirements
  - e. It is indefinite because a laboratory needs to continually monitor and adapt to changing stakeholder requirements
5. Which of the following is the correct technique or tool to use for quality improvement?
  - a. There is no one single correct tool to use; there is more than one tool that can be used
  - b. PDCA Cycle, DMAIC, and FMEA
  - c. Root cause analysis
  - d. Statistical process control
  - e. SIPOC analysis
6. What are the components used to calculate risk in the FMEA calculation?
  - a. occurrence, severity, and failed QC
  - b. occurrence, severity, and failed external quality assurance
  - c. occurrence, severity, and detection
  - d. occurrence, detection, and risk
  - e. occurrence, severity and risk
7. What are the four "T's" of risk mitigation?
  - a. take, terminate, tolerate, talk
  - b. terminate, tolerate, test, talk
  - c. terminate, treat, talk, transfer
  - d. terminate, treat, tolerate, transfer
  - e. test, terminate, tolerate, transfer
8. Which of these is NOT a requirement of clinical governance?
  - a. Clear lines of responsibility and accountability for the overall quality of clinical care
  - b. External accreditation
  - c. A comprehensive program of quality improvement clinical activities
  - d. Clear policies aimed at managing risk, for example development of risk management strategy
  - e. Procedures for all professional groups to identify and remedy poor performance

# Specimen Collection and Processing

*Khushbu Patel and Patricia M. Jones*

## ABSTRACT

### Background

Proper specimen collection and processing are critical to avoiding common preanalytical errors and ensuring accurate test results. Specific steps, recommendations, and procedures are designed to protect both the patient and the individual collecting the specimen.

### Content

This chapter addresses in detail the issues related to specimen collection. The most common types of specimens collected are discussed with the collection method(s) outlined, and some caveats for special populations, such as pediatric patients,

are included. Details on collection devices and preservatives, and their appropriate use for individual test requests are outlined with attention to how to recognize when an incorrect sample is submitted for testing. The chapter concludes with the equally important details on proper specimen processing, handling, and transport to the testing facility. It is stressed that specimen collection and handling must be done in a manner that is validated for the tests that will be performed. The information provided is designed to assist laboratorians in mitigating preanalytical errors associated with specimen collection and ensure accurate and quality results.

## INTRODUCTION

Proper collection, processing, storage, and transport of common sample types associated with requests for diagnostic testing are critical to the provision of quality test results. Each of the steps involved, as well as factors associated with the patient from whom the sample is being collected, can be the source of errors that cause inaccurate results. Minimizing these errors through careful adherence to the concepts discussed here and to individual institutional policies will result in more reliable information for use by healthcare professionals in providing quality patient care.

This chapter provides a review of the most common specimen types and discusses how they are (1) collected, (2) identified, (3) processed, (4) stored, and (5) transported. Body fluids other than blood and urine are covered in detail elsewhere (see Chapter 45) as are additional preanalytical factors (see Chapter 5). Attention to the differences between adult and pediatric collection are also discussed.

## PATIENT IDENTIFICATION

Before any specimen is collected, the phlebotomist must confirm the identity of the patient. Two or three items of identification should be used (e.g., full name, medical record number, date of birth, telephone number, or other person-specific identifier). The Joint Commission, a US hospital accreditation body, requires at least two of these unique identifiers be used to properly identify the patient.<sup>1</sup> In specialized situations, such as paternity testing or other tests of

medicolegal importance, establishment of a chain of custody for the specimen may require that additional patient identification, such as a photograph, be provided as part of the identification process or taken to confirm the identity of the patient.

Identification must be an active process. When possible, the patient should state his or her full name and date of birth or other identifier, and the phlebotomist should verify information on the patient's wrist band if the patient is hospitalized. If the patient is an outpatient, the phlebotomist should ask the patient to state his or her full name and date of birth and should confirm the information on the test requisition form with identifying information provided by the patient. In the case of pediatric patients, the parent or guardian should be present and should provide active identification of the child such as: "Please tell me the name of your child." Parents with young children are often distracted or worried about the upcoming procedure and may answer without paying attention to the question, so the question should always be posed in a manner to prevent a yes or no answer. Strict adherence to institutional policies is required.

## TYPES OF SPECIMENS

Types of biologic specimens that are analyzed in clinical laboratories include (1) whole blood; (2) serum; (3) plasma; (4) urine; (5) stool; (6) saliva; (7) other body fluids such as spinal, synovial, amniotic, pleural, pericardial, and ascitic fluids; and (8) cells and various types of solid tissue. The

**TABLE 4.1 Clinical and Laboratory Standards Institute Documents Related to Specimen Collection, Processing, and Transport**

Document Name	Document Number
Accuracy in patient and sample identification, 2nd ed.	GP33
Blood collection on filter paper for newborn screening programs: approved standard, 6th ed.	NBS01-A6
Body fluid analysis for cellular composition: approved guideline, 1st ed.	H56-A
Collection, transport, and processing of blood specimens for testing plasma-based coagulation assays and molecular hemostasis assay: approved guideline, 5th ed.	H21-A5
Collection, transport, preparation, and storage of specimens for molecular methods: approved guideline, 1st ed.	MM13-A
Ionized calcium determinations: precollection variables, specimen choice, collection, and handling: approved guideline, 2nd ed.	C31-A2
Procedures and devices for the collection of diagnostic capillary blood specimens: approved standard, 6th ed.	GP42-A6
Collection of diagnostic venous blood specimens	GP41
Tubes and additives for venous and capillary blood specimen collection: approved standard, 6th ed.	GP39-A6
Procedures for the handling and processing of blood specimens for common laboratory tests: approved guideline, 4th ed.	GP44-A4
Protection of laboratory workers from occupationally acquired infections: approved standard, 4th ed.	M29-A4
Quality management system: qualifying, selecting and evaluating a referral laboratory: approved guideline, 2nd ed.	QMS05-A2
Sweat testing: sample collection and quantitative chloride analysis, 4th ed	C34-A4

World Health Organization<sup>2</sup> and the Clinical and Laboratory Standards Institute (CLSI) have published several guidelines for collecting many of these specimens under standardized conditions (Table 4.1). In addition, the CLSI has published documents related to sample collection and analysis for specialized tests such as sweat chloride collection and testing (CLSI C34-A4, see Table 4.1).

## Blood

Blood for analysis may be obtained from veins, arteries, or capillaries. Venous blood is usually the specimen of choice, and venipuncture is the method for obtaining this specimen. Arterial puncture is used mainly for blood gas analyses. In young children and for many point-of-care tests, skin puncture is frequently used to obtain capillary blood. The process of collecting blood is known as phlebotomy (from *phleb*, which means vein, and *tome*, to cut or incise) and should always be performed by a trained phlebotomist.

## Venipuncture

In the clinical laboratory, venipuncture is defined as all of the steps involved in obtaining an appropriate and identified blood specimen from a patient's vein (CLSI GP41, see Table 4.1).

**Preliminary steps.** Before venipuncture is started the patient should be asked about latex allergies. If latex allergy is present and if latex gloves or a latex tourniquet may be used, the phlebotomist should secure an alternative tourniquet and put on gloves that are latex free. Finally, for some specialized tests such as testing for genetic diseases, the performing laboratory may request the use of a special requisition. When these are required, in general they should be provided by the requesting physician and be brought by the patient to the collection.

Before collection of a specimen, the phlebotomist should dress in personal protective equipment (PPE), such as an impervious gown and gloves applied immediately before

approaching the patient, and adhere to standard precautions against potentially infectious material; the goal is to limit the spread of infectious disease from one patient to another and to promote the safety of the patient and phlebotomist. Because small children are often frightened of anyone in a white coat or gown, pediatric phlebotomists often dress in bright, cheerful colors, including colored PPE rather than standard white. Pediatric drawing stations are also often brightly colored with lots of distractors for the patient. If the phlebotomist must collect a specimen from a patient in isolation in a hospital, the phlebotomist must put on a clean gown and gloves and a face mask and goggles before entering the patient's room. The face mask limits the spread of potentially infectious droplets, and the goggles limit the possible entry of infectious material into the eye. The extent of the precautions required varies with the nature of the patient's illness and the institution's policies and bloodborne pathogen plan to which a phlebotomist must adhere. For example, if airborne precautions are indicated, the phlebotomist must wear an N95 tuberculosis respirator in the United States.

If required, the phlebotomist should verify that the patient has fasted, identify what medications are being taken or have been discontinued as required, and determine any other relevant information required. Chapter 5 describes in more detail the effects of diet and fluid intake and the recommended steps for patient preparation, including fasting, before phlebotomy. The patient should be comfortable, seated or supine (if sitting is not feasible), and should have been in this position for as long as possible before the specimen is drawn. The correct interpretation of certain tests (e.g., aldosterone, renin, plasma metanephrides) requires that the patient be in a supine position for at least 30 minutes before venipuncture. (For details on the effects of position, refer to Chapters 5 and 53.) For an outpatient, it is generally recommended that patients be seated before completion of the identification process to maximize their relaxation. At no time should venipuncture be performed on a standing patient.



**FIGURE 4.1** Holding a child for venipuncture. (Modified from World Health Organization. *WHO guidelines on drawing blood: best practices in phlebotomy. Pediatric and neonatal blood sampling*. Geneva: World Health Organization; 2010. <http://www.ncbi.nlm.nih.gov/books/NBK138647/>.)

Infants and young children may need to be held in order to restrain them and prevent movement. Young children may be held sitting upright in a parent's lap with the parent helping to support and hold the patient and arm still (Fig. 4.1).<sup>2</sup> Infants' blood is often drawn with the infant in a supine position, and the infant may be swaddled in a blanket, or a paupoose board may be used to restrain movement. Occasionally, the parents will be more anxious than the child or will wish not to be associated with a procedure which causes the child pain, and the phlebotomist will need to make the decision to request help from a colleague phlebotomist to properly and safely perform the collection.<sup>3,4</sup>

Either of the patient's arms should be extended in a straight line from the shoulder to the wrist. An arm with an inserted intravenous (IV) line should be avoided, as should an arm with extensive scarring or a hematoma at the intended collection site. If a woman has had a mastectomy, arm veins on that side of the body should not be used because the surgery may have caused lymphostasis (blockade of normal lymph node drainage), affecting the blood composition. If a woman has had a double mastectomy, blood should be drawn from the arm of the side on which the first procedure was performed. If the surgery was done within 6 months on both sides, a vein on the back of the hand or at the ankle should be used.

Before performing a venipuncture, the phlebotomist should estimate the volume of blood to be drawn and should select the appropriate number and types of tubes for the blood, plasma, or serum tests requested. In many settings, this is facilitated by computer-generated collection recommendations and should be designed to collect the minimum amount necessary for testing. Estimating volume of blood to be drawn is especially critical in a pediatric setting. An average-weight newborn infant has a total blood volume of approximately 350 mL. Collecting too much blood from an infant in a hospital setting will eventually result in the need to give the infant blood back in the form of a transfusion, risking exposure to bloodborne pathogens. Blood collection in the pediatric population should not exceed recommended volumes for the pediatric patient's weight.<sup>5</sup> The later sections on "Order of Draw for Multiple Collections" and "Collection with Evacuated Blood Tubes" discuss in greater detail the recommended order in which to draw multiple specimens and the types of tubes to be used. Careful consideration

should also be taken in the case of an adult patient, as one study showed that on average, every 100 mL of phlebotomy was associated with a decrease in hemoglobin of 7.0 g/L and hematocrit of 1.9%.<sup>6</sup> Such iatrogenic blood loss can lead to the same possible unnecessary required blood transfusion and an increased risk of exposure to bloodborne pathogens in adults as in children.

In addition to tubes, an appropriate needle must be selected. The most commonly used sizes for adults are 19 to 22 gauge (the larger the gauge number, the smaller the bore). The usual choice for an adult with normal veins is 20 gauge; if veins tend to collapse easily, a size 21 is preferred. For volumes of blood from 30 to 50 mL, an 18-gauge needle may be required to ensure adequate blood flow. In pediatric patients, 23- to 25-gauge needles are most commonly used, with 23-gauge being the preferred size. Venipuncture on infants and children younger than 2 years old is often performed on dorsal hand veins rather than arm veins, and the veins in either place are very small in this age group. Even for larger volumes of blood, rarely will a needle larger than a 21 gauge be used because it will not fit into the vein easily. A needle is typically 1.5 inches (3.7 cm) long, but 1-inch (2.5-cm) needles, usually attached to a winged or butterfly collection set, are also used and are common in pediatrics. All needles must be sterile and sharp and without barbs. If blood is drawn for trace element measurements, the needle should be stainless steel and should be known to be free from contamination.

Finally, the phlebotomist should ensure that all postdraw safety devices are in place. These include (for the person drawing) quick, convenient, and safe access to proper disposal devices for all (now) contaminated needles and associated devices and (for the patient) the appropriate post–blood draw supplies (gauze and bandage) are in place to ensure no adverse events might affect the patient.

**Timing.** The time at which a specimen is obtained is important for blood constituents that undergo marked diurnal variation (e.g., corticosteroids, iron), for those for which a fasting sample has been requested, and for those used to monitor drug therapy. In each case, the timing should match the conditions under which reference intervals or clinical decision points were determined (see Chapter 9). Furthermore, timing is important in relation to specimens for alcohol or drug measurements in association with medicolegal considerations.

**Location.** The median cubital vein in the antecubital fossa, or crook of the elbow, is the preferred site for collecting venous blood in adults because the vein is large and is close to the surface of the skin (CLSI GP41, see *Table 4.1*). Veins on the back of the hand or at the ankle may be used, although these are less desirable and should be avoided in people with diabetes and other individuals with poor circulation. However, in infants and children younger than 2 years old, collection from superficial veins is recommended, and these sites may be preferred over the median cubital vein. In the inpatient setting, it is appropriate to collect blood through a cannula that is inserted for long-term fluid infusions at the time of first insertion to avoid the need for a second stick. This method of collection may increase the chances of a hemolyzed sample and contamination of the collected sample with fluids being infused. Careful adherence to withdrawal of a discard volume and discussions with the clinical team on alternative site for phlebotomy can greatly reduce these preanalytical variables. For severely ill individuals and those requiring many IV injections, an alternative blood-drawing site should be chosen. Selection of a vein for puncture is facilitated by palpation. An arm containing a cannula or an arteriovenous fistula should not be used without consent of the patient's physician. If fluid is being infused intravenously into a limb, the fluid should be shut off for at least 3 minutes (with clinician consent) before a specimen is obtained and a suitable note made in the patient's chart and on the result report form and the recommencement of the infusion must be ensured. Specimens obtained from the opposite arm are preferred. Specimens below the infusion site in the same arm may be satisfactory for most tests, except for analytes that are contained in the infused solution (e.g., glucose, electrolytes).

**Preparation of the site.** The area around the intended puncture site should be cleaned with whatever cleanser is approved for use by the institution. Three commonly used materials are a prepackaged alcohol swab, a gauze pad saturated with 70% isopropanol, and a benzalkonium chloride solution (e.g., Zephiran chloride solution, 1:750). Cleaning of the puncture site should be done with a circular motion from the site outward. The skin should be allowed to dry in the air. No alcohol or cleanser should remain on the skin because traces may cause hemolysis and invalidate test results. After the skin has been cleaned, it should not be touched until after the venipuncture has been completed.

**Venous occlusion.** After the skin is cleaned, a blood pressure cuff or a tourniquet is applied 4 to 6 inches (10 to 15 cm) above the intended puncture site (distance for adults). This obstructs the return of venous blood to the heart and distends the veins (venous occlusion). When a blood pressure cuff is used as a tourniquet, it is usually inflated to approximately 60 mm Hg (8.0 kPa). Tourniquets typically are made from precut soft rubber strips or from Velcro. If a dorsal hand vein is being accessed in infants and young children, no tourniquet is used. The phlebotomist applies enough pressure with the hand holding the patient's wrist and hand to occlude and distend the vein.

It is rarely necessary to leave a tourniquet in place for longer than 1 minute after venous access is secured and the tourniquet is removed, but even within this short time, the composition of blood changes, and adherence to institutional policies must be followed. Although the changes that occur in 1 minute are

slight, marked changes have been observed after 3 minutes for some chemistry analytes. The composition of blood drawn first—that is, the blood closest to the tourniquet—is most representative of the composition of circulating blood and the least affected by fluid shifts where protein bound components and other large molecules will be concentrated; water-soluble smaller molecules such as electrolytes may be less affected. The first-drawn specimen should therefore be used for analytes such as calcium and other analytes that are both protein bound and pertinent to critical medical decisions and that may be affected by the collection process.<sup>7,8</sup> A uniform procedure for the order of draw for tests should therefore be established (see later discussion). If it is only possible to collect a small volume of blood, the priority of which tests to perform should be established.

Two special notes on the collection process: Pumping of the fist before venipuncture should be avoided because it causes an increase in plasma potassium, phosphate, and lactate concentrations. The lowering of blood pH by accumulation of lactate also causes the plasma ionized calcium concentration to increase.<sup>7</sup> The ionized calcium concentration reverts to normal 10 minutes after the tourniquet is released.<sup>8</sup> Importantly, the stress associated with blood collection and/or hospitalization can have effects on patients at any age. As a consequence, plasma concentrations of analytes affected by stress, such as cortisol, thyroid-stimulating hormone, and growth hormone, may increase. Stress occurs particularly in young children who are frightened, struggling, and held in physical restraint. Collection under these conditions may cause adrenal stimulation, leading to an increased plasma glucose concentration, or may create increases in the serum activities of enzymes that originate in skeletal muscle.

**Order of draw for multiple blood specimens.** In a few patients, backflow from blood tubes into veins occurs owing to a decrease in venous pressure. The dangerous consequences of this occurrence are prevented by using only sterile tubes for collection of blood. Backflow is minimized if the arm is held downward and blood is kept from contact with the stopper during the collection procedure. When collecting multiple specimens with an evacuated tube system, one of the primary concerns is to prevent cross-contamination between tubes. For example, potassium ethylenediaminetetraacetic acid (EDTA) contamination can cause an erroneously reported hyperkalemia or hypocalcemia when an inappropriate tube type is used.<sup>9</sup> To minimize problems if backflow occurs and to optimize the quality of specimens by preventing cross-contamination with anticoagulants, blood should be collected into tubes in the order outlined in *Table 4.2*, which generally follows a process of no anticoagulant to mild anticoagulant to strong anticoagulant. This table also provides the recommended number of inversions for each tube type because it is critical that complete mixing of any additive with the blood collected be accomplished as quickly as possible. In addition, completing a blood collection within 2 minutes of starting, and getting the tubes mixed correctly as soon as possible, helps to prevent clotting in anticoagulated tubes. The order of collection when multiple tubes are drawn from a skin puncture is different than when an evacuated tube system is used (see the later section on skin puncture).

**Collection with evacuated blood tubes.** Evacuated blood tubes are usually considered to be safer, less expensive, more convenient, and easier to use than syringes and thus are the

**TABLE 4.2 Recommended Order of Draw for Multiple Blood Specimen Collection**

Stopper Color	Contents	Inversions
Yellow	Sterile media for blood culture	8
Royal blue	No additive	0
Clear	Nonadditive; this is a discard tube if no royal blue is collected, used to fill collection set spaces prior to collecting coagulation (sodium citrate) tube	0
Light blue	Sodium citrate	3–4
Gold/red	Serum separator tube	5
Red/red, orange/yellow, royal blue	Serum tube, with or without clot activator, with or without gel	5
Green	Heparin tube with or without gel	8
Tan (glass)	Sodium heparin	8
Royal blue	Sodium heparin, sodium EDTA (trace metal free)	8
Lavender, pearl white, pink/pink, tan (plastic)	EDTA tubes, with or without gel	8
Gray	Glycolytic inhibitor	8
Yellow (glass)	ACD for molecular studies and cell culture	8

ACD, Acid citrate dextrose; EDTA, ethylenediaminetetraacetic acid.

Modified from information in Clinical and Laboratory Standards Institute. *Tubes and additives for venous blood specimen collection: CLSI-approved standard GP39-A6*. 6th ed. Wayne, PA: Clinical and Laboratory Standards Institute; 2010; and Garza D, Becan-McBride K. Venipuncture procedures. In: Garza D, Becan-McBride K, editors. *Phlebotomy handbook: blood specimen collection from basic to advanced*. 10th ed. Upper Saddle River, NJ: Pearson Prentice Hall; 2019. p. 308–70.

collection device of choice in many institutions. Evacuated blood tubes may be made of soda-lime or borosilicate glass or plastic (polyethylene terephthalate). Because of the decreased likelihood of breakage and subsequent exposure to infectious materials, many, if not most, laboratories have converted from glass to plastic tubes. Several types of evacuated tubes may be used for venipuncture collection. They vary by the type of additive added and the volume of the tube. The different types of additives are identified by the color of the stopper used. Color coding of specimen collection tubes is not yet harmonized and may vary according to manufacturers. Table 4.3 presents the most common forms of color codes of various tube types. Serum or plasma separator tubes are available that contain an inert, polymer gel material with a specific gravity of approximately 1.04. Aspiration of blood into the tube and subsequent centrifugation displaces the gel, which settles like a disk between cells and supernatant when the tube is centrifuged. A minimum relative centrifugal force (RCF) of  $1100 \times g$  is required for gel release and barrier formation in most tubes. Release of intracellular components into the supernatant is prevented by the barrier for several hours or, in many cases, for 7 days or more, allowing for additional testing (“add-ons”) from samples collected at a specific time in the patient’s care. However, all laboratories need to review the specific manufacturers’ recommendations of what may be allowed based on provided data or perform their own validation studies. Most importantly, these separator tubes may be used as primary containers from which serum or plasma can be directly aspirated by a number of analytical instruments, avoiding aspiration of red blood cells (RBCs) or possible errors of patient or sample identification during aliquoting. Additional tubes, not listed, are sold for special applications, such as RNA isolation. As with all specimen collection containers, these less common tubes must be validated by each laboratory before use if not approved by the manufacturer for the specific analysis to be conducted.

Stoppers may contain zinc, invalidating the use of evacuated blood tubes for zinc measurement, and tris(2-butoxyethyl)

phosphate (TBEP), also a constituent of the stopper, which may interfere with the measurement of certain drugs. With time, the vacuum in evacuated tubes is lost and their effective draw diminishes. The silicone coating also decays with age. Therefore the stock of these tubes should be rotated and careful attention paid to the expiration date. Blood collected into a tube containing one additive should never be transferred into other tubes because the first additive may interfere with tests for which a different additive is specified. Additionally, cross-contamination of additives from one tube to another during multiple tube draws should be minimized (or adverse effects reduced) through strict adherence to recommendations for order of tube use (see Table 4.2).

Typical systems for collecting blood<sup>10</sup> are shown in Fig. 4.2. Single-use devices incorporate a cover that is designed to be placed over the needle when collection of the blood is complete, thereby reducing the risk of puncture of the phlebotomist by the now contaminated needle. A needle or winged (butterfly) set is screwed into the collection tube holder, and the tube is then gently inserted into this holder. The tube should be gently tapped to dislodge any additive from the stopper before the needle is inserted into a vein; this prevents aspiration of the additive into the patient’s vein.

After the skin has been cleaned, the needle should be guided gently into the patient’s vein; when the needle is in place, the tube should be pressed forward into the holder to puncture the stopper and release the vacuum. As soon as blood begins to flow into the tube, the tourniquet should be released without moving the needle (see earlier discussion on venous occlusion). The tube is filled until the vacuum is exhausted. It is critically important that the evacuated tube be filled completely. Many additives, particularly for coagulation testing, are provided at concentrations in the tube based on a specified volume requirement; both short and too-full draws can be a source of preanalytical error because they can significantly affect the established testing parameters that are based on a properly collected sample. Therefore a vacuum tube should always be filled using the vacuum that is designed

**TABLE 4.3 Coding of Stopper Color to Indicate Additive in Evacuated Blood Tube**

<b>Tube Type</b>	<b>Additive</b>	<b>Stopper Color</b>	<b>Alternative</b>
Gel separation tubes	Polymer gel/silica activator	Red/black	Gold
Serum tubes (nonadditive)	Polymer gel/silica activator/lithium heparin	Green/gray	Light gray
Silicone-coated interior		Red	None
Uncoated interior		Red	Pink
Serum tubes (with additives)	Thrombin (dry additive)	Gray/yellow	Orange
Particulate clot activator		Yellow/red	Red
Thrombin (dry additive)		Light blue	Light blue
Whole blood/plasma tubes	K <sub>2</sub> EDTA (dry additive)	Lavender	Lavender
K <sub>3</sub> EDTA (liquid additive)		Lavender	Lavender
Na <sub>2</sub> EDTA (dry additive)		Lavender	Lavender
Citrate, trisodium (coagulation)		Light blue	Light blue
Citrate, trisodium (erythrocyte sedimentation rate)		Black	Black
Sodium fluoride (antiglycolic agent)		Gray	Light/gray
Heparin, lithium (dry or liquid additive)		Green	Green
Potassium oxalate/sodium fluoride		Light gray	Light gray
Lithium heparin/iodoacetate		Light gray	Light gray
<b>Specialty Tubes (Microbiology)</b>			
Blood culture	Sodium polyanethol sulfonate (SPS)	Light yellow	Light yellow
<b>Specialty Tubes (Chemistry)</b>			
Lead	Heparin, potassium (liquid additive)	Tan	Tan
	Heparin, sodium (dry additive)	Royal blue	Royal blue
Trace elements	Silicone-coated interior (serum tube)	Royal blue	Royal blue
Stat chemistry	Thrombin	Gray/yellow	Orange
<b>Specialty Tubes (Molecular Diagnostics)</b>			
Plasma	K <sub>2</sub> EDTA (dry additive)/polymer gel/silica activator ACD solution A (Na <sub>3</sub> citrate, 22.0 g/L; citric acid, 8.0 g/L; dextrose, 24.5 g/L)	Opalescent white Bright yellow	Opalescent white Bright yellow
	ACD solution B (Na <sub>3</sub> citrate, 13.2 g/L; citric acid, 4.8 g/L; dextrose, 14.7 g/L)	Bright yellow	Bright yellow
Mononuclear cell preparation tube	Sodium citrate with density gradient polymer fluid Sodium heparin with density gradient polymer fluid	Blue/black Green/red	Blue/black Green/red

ACD, Acid citrate dextrose; EDTA, ethylenediaminetetraacetic acid.

Modified from information in Clinical and Laboratory Standards Institute. *Tubes and additives for venous blood specimen collection: CLSI-approved standard GP39-A6*. 6th ed. Wayne, PA: Clinical and Laboratory Standards Institute; 2010; and Becton Dickinson. <http://www.bd.com>.



**FIGURE 4.2** Assorted venipuncture collection devices.

to fill it correctly. These tubes should never be opened and filled from a syringe or other source. After the tube is filled completely, it should be withdrawn from the holder, mixed gently by inversion, and replaced by another tube if necessary. Other tubes may be filled using the same technique with the holder in place.

**Blood collection with a syringe.** Syringes are customarily used for patients with difficult veins, including very small veins, and for blood gas analysis. If a syringe is used, the needle is placed firmly over the nozzle of the syringe, and the cover of the needle is removed. If the syringe has an eccentric nozzle, the needle should be arranged with the nozzle downward but the bevel of the needle upward. The syringe and the needle should be aligned with the vein to be entered and the needle pushed into the vein at an angle to the skin of approximately 15 degrees. When the initial resistance of the vein wall is overcome as it is pierced, forward pressure on the syringe is eased, and the blood is withdrawn by very gently pulling back the plunger of the syringe. If a second syringe is necessary, a gauze pad may be placed under the hub of the

needle to absorb the spill; the first syringe is then quickly disconnected, and the second is put in place to continue the blood draw.

After filling the syringe and completing the collection, if the sample needs to be transferred to an evacuated tube, a transfer device should be used to puncture the cap of the tube. Use of transfer devices prevents having to puncture an evacuated tube with a needle and risking a needle-stick injury. The tube should be allowed to fill passively using its vacuum; uncapping the evacuated tube is not recommended for the reasons stated earlier. Vigorous withdrawal of blood into a syringe during collection or forceful transfer from the syringe to the receiving vessel may cause hemolysis of blood and will likely make the sample not valid for testing. Communication of this common preanalytical error to those not trained in routine sample collection is the responsibility of all laboratory directors and the experts, the phlebotomy team. Although safe use and disposal of sharps is important with any collection device, this is particularly important with the use of a needle and syringe. The phlebotomist must ensure an appropriate sharps disposal bin is available at the point of collection, that the location is free of interference or distractions that may increase the risk of a needle-stick injury, and that he or she has been trained in all procedures.

**Completion of collection.** When blood collection is complete and the needle withdrawn, the patient should be instructed to hold a dry gauze pad tightly over the puncture site with the arm raised to stop residual bleeding and promote the clotting process. The pad should then be held in place firmly by a bandage or by a nonadhesive strap (which avoids pulling hairs on the arm when it is removed); these may be removed after 15 minutes. With a collection device, such as that shown in Fig. 4.2, the needle is covered, and the needle and the tube holder are immediately discarded into a sharps container that should be conveniently and safely positioned. In the event that a winged (butterfly) set is used, the wings are pushed forward to cover the needle, or with newer available equipment, a button is pressed, releasing a spring that retracts the needle. If a syringe was used, the needle should not be removed because of the danger of a needlestick on the part of the phlebotomist. All used supplies should be discarded in a hazardous waste receptacle.

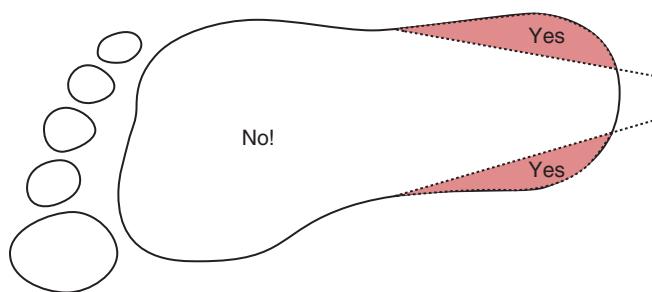
All tubes should then be labeled per institutional policy. Most institutions have a written procedure prohibiting the advance labeling of tubes because this is seen as providing the potential for mislabeling, one of the most common sources of preanalytical error. Collectors should ensure that the correct labels are applied. Incorrect labels which have been mistakenly filed in patient notes, files, or are in the room of a hospitalized patient, are a major source of mislabeling and preanalytical error. Many US institutions recommend showing the labeled tube to the patient to further confirm correct identification. At the conclusion of this process, gloves should be discarded in a hazardous waste receptacle if visibly contaminated or in uncontaminated trash if not visibly contaminated. Before applying new gloves and proceeding to the next patient, and depending on institutional policy, all caregivers including phlebotomists should use an alcohol-based cleanser or soap and water to wash their hands.

**Venipuncture in children.** The techniques for venipuncture in children and adults are similar. However, children are likely to make unexpected movements, and assistance in holding them still is often desirable. Although a syringe or an evacuated

blood tube system may be used to collect specimens, a syringe with a winged butterfly collection set is more commonly used in younger children. The pressure on a syringe can be more easily controlled by the phlebotomist than the vacuum pressure in an evacuated tube, thus preventing the pressure from pulling small veins closed instead of drawing blood from them. A syringe should be the tuberculin type or should have a 3-mL capacity, except when a large volume of blood is required for analysis. A 21- to 23-gauge needle or a 20- to 23-gauge butterfly needle with attached tubing is appropriate to collect specimens. In the pediatric population, alternative collection through skin puncture is often used.

### Skin Puncture

Skin puncture is an open collection technique in which the skin is punctured by a lancet and a small volume of blood is collected into a micro device. In practice, skin puncture is used in situations in which (1) sample volume is limited (e.g., pediatric applications), (2) repeated venipunctures have resulted in severe vein damage, or (3) patients have been burned or bandaged and veins therefore are unavailable for venipuncture. This technique is also commonly used when the sample is to be applied directly to a testing device in a point-of-care testing situation or to filter paper. It is most often performed on the tip of the finger or the heels of infants. For example, in an infant younger than 6 months, the lateral or medial plantar surface of the foot should be used for skin puncture; suitable areas are illustrated in Fig. 4.3. These areas are the fleshiest part of the foot in an infant, with the most distance between the skin surface and the underlying bone. Risk of inadvertently hitting bone and causing a bone infection is the lowest when these areas are used for a heel stick. For this reason, the back of the heel and the toes should not be used.<sup>11</sup> Blood collection from anywhere on the foot should be avoided on ambulatory patients; thus when an infant starts walking, a heel stick should no longer be performed. In addition, devices are made specifically for a heel stick or a finger stick, and they should not be used interchangeably. These devices have different tip lengths and thus make a shallower or a deeper puncture (Table 4.4). The finger-stick procedure should not be performed on infants younger than 6 months old because no commercially available device punctures shallow enough to avoid bones. The complete procedure for collecting blood from infants using skin puncture is described in a CLSI document (CLSI GP42-A6, see Table 4.1).



**FIGURE 4.3** Acceptable sites for skin puncture to collect blood from an infant's foot. (Modified from Blumenfeld TA, Turi GK, Blanc WA. Recommended site and depth of newborn heel punctures based on anatomical measurements and histopathology. *Lancet* 1979;1:230–33. Reprinted with permission from Elsevier.)

**TABLE 4.4 Tip Lengths in Finger-and Heel-Stick Devices**

Collection Type and Age of Use	Tip Length (mm)
Heel stick: premature infants	0.85
Heel stick: term infants to 6 months old	1.0
Heel stick: 6 months to 1 year	1.25
Finger stick: walking to 8 years	1.5
Finger stick: >8 years	1.75–2.0

To collect a blood specimen by skin puncture, the phlebotomist first thoroughly cleans the skin with a gauze pad saturated with an approved cleaning solution, as outlined earlier for venipuncture. If an alcohol swab is used, the alcohol must be allowed to evaporate from the skin so that hemolysis does not occur. When the skin is dry, it is quickly punctured by a sharp stab with a lancet. The depth of the incision should be less than 2.0 mm to prevent contact with bone. To minimize the possibility of infection, a different site should be selected for each puncture. The finger should be held in such a way that gravity assists collection of blood at the fingertip and the lancet held to make the incision as close to perpendicular to the fingernail as possible.<sup>11</sup> Massaging of the finger to stimulate blood flow should be avoided because it causes the outflow of debris and tissue fluid, which does not have the same composition as plasma. To improve circulation of the blood, the finger (or the heel in the case of a heel stick) may be warmed by application of a warm, wet washcloth or a specialized device, such as a heel warmer, for 3 minutes before the lancet is applied. Warming the heel or finger properly will not only cause the capillary blood to be free flowing and improve the ability to collect the sample, but the analytes in the sample will also approach arterial blood values in a properly warmed heel stick. In a cold heel, the values more approximate venous blood. Warming thus may be especially important for capillary blood gas collections. The first drop of blood is wiped off, and subsequent drops are transferred to the appropriate collection tube by gentle contact. Filling should be done rapidly to prevent clotting, and introduction of air bubbles should be prevented.

As the name suggests, blood is collected into capillary blood tubes by capillary action. A variety of collection tubes are commercially available (Fig. 4.4). Containers are available that contain different anticoagulants, such as sodium and lithium heparin, and some are available in brown glass for collection of light-sensitive analytes, such as bilirubin (see later section on anticoagulants). As with evacuated blood tubes and to prevent the possibility of breakage and the spread of infection, capillary devices frequently are plastic or coated with plastic. A disadvantage of some of the collection devices shown in Fig. 4.4 is that blood tends to pool in the mouth of the tube and must be flicked down the tube, creating a risk of hemolysis. Drop-by-drop collection and scooping along the skin with the edge of the tube to collect the blood should be avoided because both practices increase hemolysis. The correct order of filling of these devices is different than evacuated blood tubes because the concerns are different.<sup>11,12</sup> These samples are collected by dripping or capillary action from the puncture site into the small tubes



**FIGURE 4.4** Microcollection tubes. (From Flynn JC. *Procedures in phlebotomy*. 3rd ed. St. Louis, MO: Saunders; 2005.)

that hold less than 1 mL each. There is less chance of cross-contamination with anticoagulant between tubes; however, the flow of blood also clots quickly in a heel or finger stick, and platelet levels drop quickly as the clots form. Thus the anticoagulant tubes, especially the EDTA tube for the complete blood count (CBC), are drawn first rather than last. The serum tubes are drawn last (Table 4.5) because it does not matter if clots are formed in the sample.<sup>13,14</sup>

For collection of blood specimens on filter paper for molecular genetic testing and neonatal screening, the skin is cleaned and punctured as described previously. The first drop of blood should be wiped away. Then the filter paper is gently touched against a large drop of blood that is allowed to soak into the paper to fill the marked circle. Only a single application per circle should be made to prevent non-uniform analyte concentration. The paper is examined to verify that there has been complete penetration of the paper. The procedure is repeated to fill all the circles. As with all skin puncture collections, avoid milking or squeezing the finger or foot because this procedure contributes tissue fluids to the sample. The filter papers should be air dried (generally for 2 to 3 hours and horizontally placed to prevent mold or bacterial overgrowth and possible separation of blood components, respectively) before storage in a properly labeled envelope. Blood should never be transferred onto filter paper after it has been collected in non-anticoagulated capillary tubes because partial clotting may have occurred, compromising the quality of the specimen. However, blood collected into any type of tube containing an anticoagulant may be applied directly to the filter paper. Dried blood spots

**TABLE 4.5 Order of Draw for Skin Puncture: Capillary Blood**

Usage or Additive	Tube Top Color
Blood gases (heparin)	Microhematocrit tubes
EDTA	Lavender
Heparin	Green
Other additives	Light blue, gray
Nonadditives	Red, tiger, yellow

EDTA, Ethylenediaminetetraacetic acid.

on filter paper are a convenient way to store a sample for possible future molecular testing (with patient consent). These blood spots are handled in the same manner as neonatal screening specimens, with air drying and storage in a dry protected environment.

### Arterial Puncture

Arterial puncture requires considerable skill and is usually performed only by physicians or specially trained technicians or nurses. Preferred sites of arterial puncture are, in order, the (1) radial artery at the wrist, (2) brachial artery in the elbow, and (3) femoral artery in the groin. Because leakage of blood from the femoral artery tends to be greater, especially in older adults, sites in the arm are used most often. The proper technique for arterial puncture has been described.<sup>15,16</sup>

In neonates, an indwelling catheter in the umbilical artery is best to obtain specimens for blood gas analysis. In older children and adults in whom it is impossible to perform an arterial puncture, a capillary puncture may be performed to obtain arterialized capillary blood. Such a specimen yields acceptable values for pH and  $PCO_2$  but not always for  $PO_2$  unless the site is properly warmed. In children and adults, the preferred puncture site for arterialized capillary blood is the finger; in infants, it is the heel. Capillary blood specimens are particularly inappropriate when blood circulation is poor and thus should be avoided when a patient has reduced cardiac output, hypotension, or vasoconstriction or has a condition of fluid overload. For each capillary puncture, the skin should be warmed first with a hot, moist towel to improve the circulation. The puncture itself should be performed as described previously; a free flow of blood is essential. Heparinized capillary tubes containing a small metal bar can be used to collect the blood. Tubes should be sealed quickly and the contents mixed well by using a magnet to move the metal bar up and down in the tube so that a uniform specimen is available for analysis.

### Anticoagulants and Preservatives for Blood

Serum is defined as that portion of blood that remains after coagulation has occurred and the cells have been removed. Serum is the specimen of choice for many analyses, including viral and antibody screening and protein electrophoresis. Samples are collected into tubes with no additive or with a clot activator and must be allowed to complete the coagulation process before further processing. Plasma is defined as the noncellular component of anticoagulated whole blood after the cellular components have been removed. There are multiple ways to produce a plasma sample as detailed later. Heparinized plasma is increasingly being used for routine chemistry testing to decrease turnaround time, because it is not necessary to wait for the blood to clot. Sometimes considerable differences may be observed between the concentrations of analytes in serum and in plasma, as shown in Table 4.6. For molecular diagnostics, anticoagulated whole blood or plasma is more likely to be the specimen of choice for either genomic DNA isolation from the white blood cells (WBCs) still intact from a whole blood collection or from plasma that will yield viral identification and quantification. A number of anticoagulants are available, including heparin, EDTA, sodium fluoride (NaF), citrate, acid citrate dextrose (ACD), oxalate, and iodoacetate, which are covered in detail later.

**TABLE 4.6 Differences in Composition Between Heparin Plasma and Serum<sup>a,b</sup>**

<b>Plasma Value &gt; Serum Value (%)</b>	<b>No Difference Between Serum and Plasma Values</b>		<b>Plasma Value &lt; Serum Value (%)</b>
	Lactate dehydrogenase	Bilirubin	
Total protein	4.0	Cholesterol	Phosphorus
		Creatinine	Potassium

<sup>a</sup>To estimate the probable effect of a factor on results, relate the percent increase or decrease shown in the table to analytical variation ( $\pm$  % coefficient of variation) routinely found for analytes.

<sup>b</sup>This list includes only differences that are of clinical significance and may need to be annotated with a comment on patient results that general (plasma) reference intervals may not apply.

Modified from Ladenson JH, Tsai LMB, Michael JM, Kessler G, Joist JH. Serum versus heparinized plasma for eighteen common chemistry tests. *Am J Clin Pathol* 1974;62:545–52. Copyright 1974 by the American Society of Clinical Pathologists. Reprinted with permission.

For any assay provided for clinical use, manufacturers specify the appropriate sample type(s) for which they have validated the assay. Use of different sample types is acceptable only if the laboratory has validated the alternate type(s). For example, care should be taken with gel tubes because they may vary among tube manufacturers. Acceptability of a wide range of sample types can be advantageous because it can reduce the need for recollections if the preferred tube is not provided.

**Heparin.** Heparin is the most widely used anticoagulant for chemistry testing. It is a mucoitin polysulfuric acid and is available as sodium, potassium, lithium, and ammonium salts, all of which can adequately prevent coagulation. This anticoagulant accelerates the action of antithrombin III, which neutralizes thrombin and thus prevents the formation of fibrin from fibrinogen. Most blood tubes are prepared with approximately 0.2 mg of heparin for each milliliter of blood (18 units/mL) to be collected. The heparin is usually present as a dry powder that is hygroscopic and dissolves rapidly assuming that the tube of blood is correctly mixed (see Table 4.2). Heparin is a naturally occurring anticoagulant and has the disadvantages of high cost and a more temporary action of anticoagulation than is attained by the chemicals discussed later. It produces a blue background in blood smears that are stained with Wright's stain. In addition, heparin can interfere with the binding of calcium to EDTA in analytical methods for calcium involving complexing with EDTA. Heparin, which is negatively charged, binds calcium and can reduce results for ionized calcium measurements. Thus either serum tubes are required or blood gas syringes with either low heparin concentrations or so-called "balanced heparin" with added calcium to block the binding of further calcium are used. Of course, the use of lithium or ammonium heparin is unacceptable for lithium and ammonia measurements, respectively, because the tube contains an amount similar to that found in treated patients (lithium) or can elevate the clinically actionable value (ammonia). Heparin tubes are also unsuitable for protein electrophoresis as the

fibrinogen in the unclotted sample causes an interfering band. This fibrinogen is also the cause of a higher total protein in heparin tubes than in serum tubes.

It should be noted that heparin is unacceptable for most tests performed using polymerase chain reaction (PCR) because of inhibition of the polymerase enzyme by this large molecule. In some special circumstances, a heparin tube can be shared with a molecular diagnostic laboratory if a non-heparinized tube is not available. DNA can be extracted from heparinized samples, but amplification may be reduced, and the effect of heparin on any molecular diagnostic assay should be assessed as part of a method validation study.

**Ethylenediaminetetraacetic acid.** EDTA is a chelating agent of divalent cations such as  $\text{Ca}^{2+}$  and  $\text{Mg}^{2+}$  that is particularly useful for (1) hematologic examinations including transfusion medicine applications, (2) measurement of intracellular drugs such as cyclosporine or tacrolimus, (3)  $\text{HbA}_{1c}$  analysis, (4) isolation of genomic DNA, and (5) qualitative and quantitative virus determinations by molecular techniques because it preserves the cellular components of blood. It is used as the disodium, dipotassium, or tripotassium salt, the last two being more soluble with the tripotassium salt commonly provided as a liquid in the collection tube. It is effective at a final concentration of 1 to 2 g/L of blood. Higher concentrations hypertонically shrink the RBCs. EDTA prevents coagulation by binding calcium, which is essential for the clotting mechanism. EDTA tubes are also available with a gel barrier to separate plasma from cells when EDTA plasma is required (white tubes; see Table 4.3). In blue/black tubes, incorporation of a density gradient allows recovery of nucleated cells after centrifugation, thus increasing the yield of DNA.

Because it chelates calcium, magnesium, and iron, EDTA is unsuitable for specimens for these analyses using the most common photometric or titrimetric techniques. For many laboratories, an undetectable calcium value or a critical potassium value in an otherwise stable patient is often used as a flag that an inappropriate specimen type has been submitted for just this reason. Additionally, EDTA, probably by chelation of required metallic cofactors, inhibits alkaline phosphatase and creatine kinase activities. As an anticoagulant, it has little effect on other clinical tests, although the concentration of cholesterol has been reported to be decreased by 3% to 5%.

**Sodium fluoride.** NaF is a weak anticoagulant that is often added as a preservative for blood glucose and lactate.<sup>17,18</sup> As a preservative, together with another anticoagulant such as potassium oxalate, it is effective at a concentration of approximately 2 g/L blood. It exerts its preservative action by inhibiting enolase, a downstream enzyme in the glycolysis pathway. Therefore the inhibition is not immediate,<sup>17</sup> and a certain amount of degradation occurs during the first 1 to 2 hours after collection. Most specimens are then stable at 25 °C for at least 24 hours or at 4 °C for 48 hours. Without an antiglycolytic agent, the blood glucose concentration decreases approximately 10 mg/dL (0.56 mmol/L) per hour at 25 °C. The rate of decrease is faster in newborns because of the increased metabolic activity of their erythrocytes and in leukemic patients because of the high metabolic activity of the WBCs. NaF is poorly soluble, and blood must be well mixed for NaF to be effective (see Table 4.2). Because of the delay in onset of action, recent protocols recommend placing the tube on ice until the sample can be separated to ensure

accurate glucose measurements. Newer NaF combination tubes include tubes containing NaF–citrate buffer– $\text{Na}_2\text{EDTA}$ , NaF–Na–heparin, NaF–K<sub>2</sub>oxalate, NaF–citrate, and NaF– $\text{Na}_2\text{EDTA}$ . Acidification of specimens by citrate and addition of EDTA to a fluoride tube immediately inhibit glycolysis and preserve glucose in the specimen for at least 24 hours. These various tubes have all been found to be suitable for glucose preservation,<sup>18</sup> but care must be taken to ensure use according to manufacturers' guidelines when using different tube types.<sup>19</sup>

If NaF is used alone for anticoagulation, three to five times greater concentrations than the usual 2 g/L are required. This high concentration and inhibition of the glycolytic cycle are likely to cause fluid shifts and a change in the concentration of some analytes. Fluoride is also a potent inhibitor of many serum enzymes and in high concentrations also affects urease, which is used to measure urea nitrogen in many analytical systems.

**Citrate.** Sodium citrate solution, at a concentration of 34 to 38 g/L in a ratio of 1 part to 9 parts of blood, is widely used for coagulation studies. The correct ratio of blood to anticoagulant is critical (refer to the earlier discussion concerning proper filling of vacuum blood tubes) to achieve proper coagulation measurements because the anticoagulant effect is reversed by addition of standard amounts of  $\text{Ca}^{2+}$  that are based on a proper collection volume. Because citrate chelates calcium, it is unsuitable as an anticoagulant for specimens for measurement of this element. It also inhibits aminotransferases and alkaline phosphatase but stimulates acid phosphatase when phenylphosphate is used as a substrate. Because citrate complexes molybdate, it decreases the color yield in phosphate measurements that involve molybdate ions and produces low results.

**Acid citrate dextrose.** As indicated previously, the collection of specimens into EDTA is often used for isolation of genomic DNA from the patient. However, additional and complementary diagnostic tests, such as cytogenetic testing, may be requested at the same time. For this reason, samples for molecular diagnostics with an accompanying cytogenetic request are often collected into ACD anticoagulant so as to preserve both the form and the function of the cellular components. There are two ACD tube designations: ACD A and ACD B. These differ only by the concentrations of the additives (see Table 4.3). Both enhance the vitality and recovery of WBCs for several days after collection of the specimen; thus they are suitable for both molecular diagnostic testing and cytogenetic testing.

Whereas solution A is used for an 8.5-mL blood draw (10 mL total volume), solution B is used for a 3- or a 6-mL blood draw (7 mL total volume). The specific test(s) requested will determine the size of tube necessary for specimen collection.

**Oxalates.** Sodium, potassium, ammonium, and lithium oxalates inhibit blood coagulation by forming rather insoluble complexes with calcium ions. Potassium oxalate ( $\text{K}_2\text{C}_2\text{O}_4 \cdot \text{H}_2\text{O}$ ), at a concentration of approximately 1 to 2 g/L of blood, is the most widely used oxalate. At concentrations of greater than 3 g oxalate per liter, hemolysis is likely to occur.

Combined ammonium and/or potassium oxalate does not cause shrinkage of erythrocytes. However, other oxalates have been known to cause shrinkage by drawing water into the plasma. Reduction in hematocrit may be as much as

10%, causing a reduction in the concentration of plasma constituents of 5%. As fluid is lost from the cells, an exchange of electrolytes and other constituents across the cell membrane occurs. Oxalate inhibits several enzymes, including acid and alkaline phosphatases, amylase, and lactate dehydrogenase (LDH), and may cause precipitation of calcium as the oxalate salt.

**Iodoacetate.** Sodium iodoacetate at a concentration of 2 g/L is an effective antiglycolytic agent (with the caveats mentioned earlier) and a substitute for NaF. Because it has no effect on urease, it can be used when glucose and urea tests are performed on a single specimen. It inhibits creatine kinase but appears to have no notable effects on other clinical tests.

### Influence of Site of Collection on Blood Composition

Blood obtained from different sites differs in composition. In general, skin puncture blood is more similar to arterial blood than venous blood, depending on the collection condition as described earlier. Thus there are no clinically significant differences between freely flowing capillary blood and arterial blood in pH,  $\text{PCO}_2$ ,  $\text{PO}_2$ , and oxygen saturation. The  $\text{PCO}_2$  of venous blood is up to 6 to 7 mm Hg (0.8 to 0.9 kPa) higher. Venous blood glucose is as much as 7 mg/dL (0.39 mmol/L) less than capillary blood glucose.

Blood obtained by skin puncture is contaminated to some extent with interstitial and intracellular fluids. The major differences between venous serum and capillary serum are illustrated in Table 4.7.

### Collection of Blood From Intravenous or Arterial Lines

When blood is collected from a central or peripheral venous catheter or arterial line, it is necessary to ensure that the composition of the specimen is not affected by the fluid that is infused into the patient. With clinical approval, fluid may be shut off using the stopcock on the catheter, and 10 mL of blood is aspirated through the stopcock and discarded before the specimen for analysis is withdrawn. In pediatric patients, 10 mL of blood going to waste is often not feasible, so lesser volumes are aspirated, although the goal is still to aspirate roughly three times the dead space of the line before collecting the sample for testing. Any infused fluid contamination may affect basic biochemical tests such as electrolytes, lactate, or glucose. Aspirating this blood and clearing the lines is

equally important for molecular diagnostics and coagulation testing because the stopcock is often heavily saturated with heparin to prevent clotting. Collection of samples for therapeutic drug monitoring should not be done from the line used for the infusion irrespective of the time since infusion or amount of blood aspirated because the drug may adhere to the line and leak into the collected sample, causing false elevations. Additionally, blood collected from a heparinized catheter line cannot be used for coagulation testing as it will significantly affect results.<sup>20</sup>

In theory, blood may be collected from the veins of an arm below an IV line without interference from the fluid being infused because retrograde blood flow does not occur in the veins and the fluid that is infused must first circulate through the heart and return to the tissue before it reaches the sampling site. However, as stated previously, collection from the arm without the IV line is strongly recommended.

### Hemolysis

Hemolysis is defined as the disruption of the RBC membrane, resulting in the release of hemoglobin, and may be the consequence of intravascular events (*in vivo* hemolysis) or may occur subsequent to or during blood collection (*in vitro* hemolysis). In *vitro* hemolysis is perhaps the single biggest preanalytical factor affecting test results and has many causes. Mechanical disruption of the cells may occur from improper mixing of sample and anticoagulant after collection. Improper collection may occur, including applying high shear forces on the sample from too rapid collection of the sample into the collection device or performing the collection through an inadequately cleared IV line. Higher rates of hemolysis are commonly seen when collection is done through IV lines and in samples collected in the emergency department. This may be due to a combination of rapid collection in a high-stress environment, poor collection technique causing contamination between tubes, and improper mixing. Low-volume evacuated tubes are available which have reduced vacuum and cause lower shear force on the RBCs. These tubes also collect less blood, but can be used to help reduce the percentage of hemolyzed samples being collected. In pediatrics and especially with neonates, *in vitro* hemolysis is a considerable and constant problem. The shear forces generated on the cells when collecting a blood sample are exacerbated by the small-bore needles necessary to get into tiny veins. Additionally, most samples from neonates are collected via heel-stick skin puncture, and this collection technique commonly results in hemolyzed samples. Finally, neonates have larger, more fragile RBCs with a shorter half-life that are prone to hemolyze. Again, low-volume evacuated tubes can help reduce hemolysis rates, as can continuous education on proper collection techniques for heel sticks and careful handling of the collected samples.

Serum and plasma show visual evidence of hemolysis when the hemoglobin concentration exceeds 50 mg/dL (7.7  $\mu\text{mol/L}$ ). When the level exceeds 150 to 200 mg/dL (23 to 31  $\mu\text{mol/L}$ ), the plasma will appear bright red to most observers. Slight hemolysis has little effect on most but not all test values. A clinically significant interpretation of results at this lower concentration may be observed on those constituents that are present at a higher concentration in erythrocytes than in plasma. Thus plasma activities or concentrations of LDH, aspartate transaminase (AST), potassium, magnesium,

TABLE 4.7 Difference in Composition of Capillary and Venous Serum<sup>a</sup>

No Difference Between Capillary and Venous Values			
Capillary Value > Venous Value (%)	Capillary and Venous Values	Capillary Value < Venous Value (%)	
Glucose 1.4	Phosphate	Bilirubin 5.0	
Potassium 0.9	Urea	Calcium 4.6	
		Chloride 1.8	
		Sodium 2.3	
		Total protein 3.3	

<sup>a</sup>To estimate the probable effect of a factor on results, relate the percent increase or decrease shown in the table to analytical variation ( $\pm$  % coefficient of variation) routinely found for analytes. From Kupke IR, Kather B, Zeugner S. On the composition of capillary and venous blood serum. *Clin Chim Acta* 1981;112:177–85.

and phosphate are particularly increased by even a slight degree of hemolysis and may need to be explained to the care team to determine whether or not the result(s) represent *in vivo* or *in vitro* hemolysis and determine the implication(s) of the test result. Most manufacturers provide data on the effects of hemolysis on the analytical performance of individual tests, and this should be evaluated in the selection of individual methods. Each laboratory must define at what level of hemolysis results should be held and not reported to prevent poor clinical action on such unreliable test results.

Although the amount of free hemoglobin could be measured and a calculation made to correct test values affected by hemoglobin, this practice is undesirable because factors other than hemoglobin could contribute to the altered test values, and it would be impossible to assess their impact. Hemolysis may affect many un-blanked or inadequately blanked analytical methods.<sup>21</sup> Currently, most large chemistry analyzers have the capability of measuring the amount of hemolysis, icterus, and lipemia (the HIL indices) in samples placed on the instrument. These are the three main interferences present in patient samples, and the instrument can be programmed to prevent release of results affected by a specific level of each of these. For some analytes, the effect of hemolysis is time dependent (e.g., the degradation of insulin by released intracellular enzymes), and any hemolysis may give falsely low results. For more details about HIL indices and their assessment and management in the laboratory, refer to Chapter 5.

In molecular diagnostic testing, hemoglobin may interfere with the amplification reaction, particularly when reverse transcriptase (RT)-PCR is the first step in the analysis of RNA. In some situations, the isolation of nucleic acid is sufficiently selective that free hemoglobin from the ruptured cells is removed and will not cause a problem. However, with hemolyzed blood, alternative or additional extraction methods are usually needed to ensure that RNA is fully and accurately transcribed and that the greatest amplification of DNA is achieved.

### Urine

The type of urine specimen to be collected is dictated by the tests to be performed. Untimed or random specimens are suitable for only a few chemical tests; usually, urine specimens must be collected over a predetermined interval of time, such as 4, 12, or 24 hours. A clean, early morning, fasting specimen is usually the most concentrated specimen and thus is preferred for microscopic examinations and for detection of constituents, such as proteins, especially albumin, or human chorionic gonadotropin. The clean timed specimen is one obtained at specific times of the day or during certain phases of the act of micturition. Whereas bacterial examination of the first 10 mL of urine voided is most appropriate to detect urethritis, the midstream specimen is best for investigating bladder disorders. The double-voided specimen is the urine excreted during a timed period after complete emptying of the bladder; it is used, for example, to assess glucose excretion during a glucose tolerance test. Its collection must be timed in relation to the ingestion of glucose. Similarly, in some metabolic disorders, urine must be collected during or immediately after symptoms of the disease appear (see Chapters 41 and 61).

When used for testing alcohol and drugs of abuse content, urine specimens are collected under rigorous conditions.

Such collections may begin with a requirement for formal identification such as a driver's license or other picture identification as discussed earlier. Patients are asked to leave all personal belongings outside of the restroom facility to prevent substitution of the patient's urine with urine brought from outside. Many locations that routinely perform these collections have the capacity to turn running water off to prevent dilution or put a coloring agent in the toilet water so such dilution attempts are easily identified; the temperature of the urine is often recorded as well to detect such attempts. Finally, patients are often asked to put their initials on the urine cup and sign the paperwork, accepting the sample as theirs. This chain of custody documentation is then sealed in a transport bag with the sample (also sealed) and transported to the testing laboratory directly, where the chain of custody documentation will be continued throughout the testing process. It is necessary for laboratories to be aware of the relevant legal requirements for this type of testing and plan in advance how these samples and the supporting paperwork will be collected and handled. Institutional policies should be in place and adhered to in all cases.

Catheter specimens are used for microbiologic examination in critically ill patients and in those with urinary tract obstruction but should not normally be obtained just for examination of chemical constituents. The suprapubic tap specimen is a useful alternative because the tap is unlikely to cause infection. After appropriate cleaning of the skin over the full bladder, a 22-gauge spinal needle is passed through a small wheal made by a local anesthetic. The bladder is penetrated and the urine withdrawn into the syringe.

Even though tests in the clinical laboratory are not usually affected by lack of sterile collection procedures, the patient's genitalia should be cleaned before each voiding to minimize the transfer of surface bacteria to the urine. Cleansing is essential if the true concentration of WBCs is to be obtained.

Currently, urine is an uncommon specimen type in the molecular diagnostic laboratory for genomic testing, although some laboratories use urine samples for bladder cancer screening and monitoring of therapy for bladder cancer. However, urine is frequently used for molecular testing for infectious agents, such as *Chlamydia*, a common sexually transmitted organism, or BK virus (also known as polyomavirus hominis 1), associated with potential rejection or failure of transplanted kidneys. Because most requests involve a specific organism, an untimed or random urine specimen collected into a sterile container with no preservative is usually acceptable.

### Timed Urine Specimens

The collection period for timed specimens should be long enough to minimize the influence of short-term biologic variations. When specimens are to be collected over a specified period of time, the patient's close adherence to instructions is critically important, and a common source of a preanalytical variable. The bladder must be emptied at the time the collection is to begin and this urine discarded. Thereafter, all urine must be collected until the end of the scheduled time, including emptying of the bladder at the end of the collection period. If a patient has a bowel movement during the collection period, precautions should be taken to prevent fecal contamination of the urine. If the collection has to be made over several hours, urine should be passed into a separate container at each voiding and then emptied into a larger container for the

complete specimen. This two-step procedure prevents the danger of patients splashing themselves with a preservative, such as acid that may be included in the timed collection device. The large container generally should be stored at 4 °C during the entire collection period.

Before beginning a timed collection, a patient should be given written instructions with regard to diet or drug ingestion, if appropriate, to avoid interference of ingested compounds with analytical procedures. For example, instructions for collection of specimens for 5-hydroxyindoleacetic acid measurements should specify avoidance of avocados, bananas, plums, walnuts, pineapples, eggplant, acetaminophen, and cough syrups containing glyceryl guaiacolate (guaifenesin). These dietary components are sources of 5-hydroxytryptamine and should be avoided for this reason; the other compounds interfere with certain analytical procedures but may not interfere with highly specific analytical methods. Each laboratory should determine its own requirements. See also specimen information for specific analytes in the respective chapters.

For 2-hour specimens, a prelabeled 1-L bottle is generally adequate. For a 12-hour collection, a 2-L bottle usually suffices; for a 24-hour collection, a 3- or 4-L bottle is appropriate for most patients. A single bottle allows adequate mixing of the specimen and prevents possible loss of some of the specimen if a second container does not reach the laboratory. Urine should not be collected at the same time for two or more tests requiring different preservatives. Aliquots for an analysis such as a microscopic examination should not be removed while a 24-hour collection is in process. Removal of aliquots during collection is not permissible even when the volume removed is measured and corrected because excretion of most compounds varies throughout the day, and test results will be affected. Appropriate information regarding the collection, including warnings with respect to handling of the specimen, should appear on the bottle label.

When a timed collection is complete, the specimen should be delivered without delay to the clinical laboratory, where the volume should be measured. This may be done by using graduated cylinders or by weighing the container and the urine when preweighed or uniform containers are used. The mass in grams may be reported as if it were the volume in milliliters. There is rarely a need to measure the specific gravity of a weighed specimen because errors in analysis usually exceed the error arising from failure to correct the volume of urine for its mass.

Before a specimen is transferred into small containers for each of the ordered tests, it must be thoroughly mixed to ensure homogeneity because the specific gravity, volume, and composition of the urine all may vary throughout the collection period. The small container into which an aliquot is transferred should not be a plastic bottle if toluene or another organic compound has been used as a preservative; metal-free containers must be used for trace metal analyses. See the later discussion on appropriate labeling of such secondary containers.

### Collection of Urine From Children

Collection of any type of urine specimen from an infant is difficult, with timed collections being the most problematic. Fortunately, timed collections are rarely required in neonates. The approved method for collecting a random urine



**FIGURE 4.5** Urine collection device used in children.

specimen from an infant involves a process known as bagging. The scrotal or perineal area is cleaned and dried first, and any natural or applied skin oils are removed. Then a plastic bag (e.g., U-Bag, Hollister, Chicago, IL, or Tink-Col, C.R. Bard, Murray Hill, NJ) is placed around the infant's genitalia and held in place by a mild adhesive (Fig. 4.5). The baby's diaper is reapplied to help hold the bag in place. As soon as voiding has occurred, the bag containing the urine sample is removed and emptied into a regular urine collection cup. The mild adhesive on the bag will often fail when it becomes wet with urine, so the infant should be monitored and the bag removed as soon as the urine sample is collected. Even a random or spot urine collection for something as simple as a urinalysis may require either catheterization or a suprapubic tap collection in an infant, especially if there is difficulty getting the infant to urinate or in collecting the sample when the infant does urinate. In infants and very young children requiring a rare 24-hour urine collection, hospitalization and catheterization are often required to obtain a complete collection.

### Urine Preservatives

The most common preservatives and the tests for which preservatives are required are listed in Table 4.8. Preservatives have different roles but usually are added to reduce bacterial action or chemical decomposition or to solubilize constituents that otherwise might precipitate out of solution. Another application is to decrease atmospheric oxidation of unstable compounds. Some specimens should not have *any* preservatives added because of the possibility of interference with analytical methods.

One of the most acceptable forms of preservation of urine specimens is refrigeration immediately after collection; it is even more successful when combined with chemical preservation. Urinary preservative tablets that contain a mixture of chemicals, such as potassium acid phosphate, sodium benzoate, benzoic acid, hexamethylene tetramine, and sodium bicarbonate have been used for chemical and microscopic examination. Because these tablets contain sodium and potassium salts, among others, they should not be used for analysis of these analytes. The preservative tablets act mainly by lowering the pH of the urine and by releasing formaldehyde. Formalin has also been used for preserving specimens, but in large amounts it precipitates urea and inhibits certain reactions (e.g., the dipstick esterase test for leukocytes). Acidification to below a pH of 3 is widely used to preserve 24-hour specimens and is particularly useful for specimens for determination of calcium, steroids, adrenaline, noradrenaline, and vanillylmandelic acid. However, precipitation of

**TABLE 4.8 Commonly Used Urine Preservatives**

Preservative	Concentrations or Volumes	Common Usage (Analytes)
HCl	6 mol/L; 30 mL per 24-h collection	Acidification/common preservative for 24-h urine collections
Acetic acid	50%; 25 mL per 24-h collection	Acidification/preservative for 24-h urine collections (Aldosterone, Catecholamines, Serotonin, 5-hydroxyindoleacetic acid [5-HIAA], homovanillic acid/vanillylmandelic acid [HVA/VMA])
Na <sub>2</sub> CO <sub>3</sub>	5 g per 24-h collection	Alkalization/ preservative for 24-h urine collections (porphyrins, urobilinogen, uric acid)
HNO <sub>3</sub>	6 mol/L; 15 mL 24-h hour collection	Acidification/preservative for 24-h urine collections (used sometimes for trace metal analysis)
Boric acid	10 g per 24-h collection	Acidification/preservative for 24-h urine collections (Urea, glucose, cortisol, aldosterone, other corticosteroids)
Toluene	30 mL per 24-h collection	Bacteriostatic agent/preservative for 24-h urine collections (oxalate, cystine, lysine, ornithine, arginine)
Thymol	10% in isopropanol; 10 mL per 24-h collection	Microscopy (not commonly used)

Modified from information provided in Clinical and Laboratory Standards Institute. *Routine urinalysis and collection, transportation, and preservation of urine specimens: CLSI-approved guideline GP16-A3*. Wayne, PA: Clinical and Laboratory Standards Institute; 2009.

urates will occur, thereby rendering a specimen unsuitable for measurement of uric acid.

Sulfamic acid (10 g/L urine) has also been used to reduce pH. Boric acid (5 mg/30 mL) has been used, but it also causes precipitation of urates. Thymol and chloroform were widely used in the past to preserve specimens for chemical and microscopic urinalysis, and thymol is still used in some cases. For many analytes, it is now recognized that specimens should be analyzed immediately and that the addition of preservatives is both largely ineffective and a source of interference with several analytical methods. Toluene is the only organic solvent that is still regularly used as a preservative. When present in a large enough amount, it acts as a barrier between the air and the surface of the specimen. Toluene, however, does not prevent the growth of anaerobic microorganisms and, because of its flammable nature, is a safety hazard. A mild base, such as sodium bicarbonate or a small amount of sodium hydroxide, is used to preserve porphyrins, urobilinogen, and uric acid. A sufficient quantity should be added to adjust the pH to between 8 and 9.

### Stool

Small aliquots of stool are frequently analyzed to detect the presence of “hidden” blood, so-called occult blood. Occult blood screening is included as part of many periodic health examinations. Guaiac-based occult blood tests are subject to many interferences (aspirin, vitamin C, steroids, various drugs, red meat, alcohol), causing both false-positive and false-negative results, and should be interpreted cautiously unless the intent is to identify current bleeding in any portion of the gastrointestinal (GI) tract in an emergent patient. Newer immunochemical-based tests, immunochemical fecal occult blood (iFOB) tests, have decreased the interfering effects of food intake greatly for the purpose of identifying a lower GI tract bleed that may indicate the presence or possibility of malignant growth. In either case, tests for occult blood should be done on aliquots of excreted stools rather

than on material obtained on the glove of a physician doing a rectal examination because this procedure may cause enough bleeding to produce a positive result. Conversely, the small amount of stool present on the glove may not be representative of the whole, so bleeding may not be recognized.

In newborns, the first specimen from the bowel (meconium) may be used for detection of maternal drug use during the gestational period, which requires specific attention to the details of collection and identification similar to the chain of custody procedure for urine collection discussed earlier. Stool from infants and children may be screened for tryptic activity or for increased fecal fat concentrations, both of which can be indicators of cystic fibrosis. Fecal material is also commonly collected in childhood for the detection of parasites (ova and parasites [O & P]), enteric disease organisms such as *Salmonella* and *Shigella*, and viruses, all of which are useful in sorting out the differential diagnosis of diarrhea. Fecal testing is also used for helping to determine causes of malabsorption. In infants, fecal material for these tests is usually recovered from the diaper.

In adults and children, measurement of fecal nitrogen and fat in 72-hour specimens is used to assess the severity of malabsorption; measurement of fecal porphyrins is occasionally required to characterize the type of porphyria. Usually, no preservative is added to the stool, but the container should be kept refrigerated throughout the collection period, and care should be taken to prevent contamination from urine. When the collection is complete, the container and stool are weighed, and the mass of excreted stool is calculated. The specimen is homogenized and aliquoted so that the amount of fat or nitrogen excreted per day and the proportion of dietary intake excreted can be calculated.

For metabolic balance studies, collections of stool are usually made over a 72-hour period. Many balance studies are carried out in conjunction with research on the metabolism of such elements as calcium. It is important for such studies that a patient be on a controlled diet for a sufficiently long

time before commencement of the study, so that a steady state has been attained.

DNA isolated from fecal samples is representative of the genetic composition of the colonic mucosa at the time of stool collection. The analysis of stool DNA is recommended by the American Cancer Society as a sensitive and specific biomarker useful for the detection of colorectal cancer.<sup>22</sup>

### Other Body Fluids

Specimens may be collected for analysis from a range of different body fluids. These include cerebrospinal fluid (CSF), pleural fluid, ascitic fluid, pericardial fluid, amniotic fluid, synovial fluid, and others.<sup>23–25</sup> Readers are referred to Chapter 45 for a complete discussion of collection, analysis, and interpretation for these specimens.

### Bronchoalveolar Lavage

Bronchoalveolar lavage (BAL) samples are another type of fluid sample that may be received in a laboratory. BAL is performed by a skilled clinician and involves passing a bronchoscope into a part of the lung, squirting a small amount of saline into that section, and then aspirating it back for examination. BAL is the most common method of sampling the internal lung milieu in the lower respiratory tract and is especially useful in patients with cystic fibrosis; immunocompromised patients; patients with pneumonia on ventilators; and patients with lung diseases, including cancers. Tests that are commonly ordered on BAL samples include cell count and WBC differential; cytopsin and various histology slides for staining; and aerobic and anaerobic bacterial, mycobacterial, fungal, and viral cultures. Respiratory viral panels by PCR are also commonly ordered on BAL samples, as are acid-fast bacilli culture and stain and *Mycobacterium tuberculosis*-specific PCR testing.

### Chorionic Villus Sampling

Chorionic villus sampling (CVS) testing can be performed at a gestation period of 10 to 12 weeks, but with amniotic fluid, testing generally is not performed until week 15 to 20 of gestation. CVS is the technique of inserting a catheter or needle into the placenta and removing some of the chorionic villi, or vascular projections, from the chorion. This tissue mostly has the same chromosomal and genetic makeup as the fetus and can be used to test for disorders that may be present in the fetus. When chorionic villus is sampled, ultrasonography is performed to assess the placenta and determine its position. The sample of the placenta is obtained through the vagina or through the abdomen, depending on the location of the placenta.

Maternal cell contamination testing is used to definitively identify the source of isolated cells in an amniotic fluid sample and in CVS. Such confirmation of the source of the sample is strongly recommended for any prenatal diagnostic testing and may be required as a quality monitor in some laboratories. Of note, measurement of circulating nucleic acids in maternal plasma has become the method of choice and has led to a worldwide reduction of such invasive procedures for prenatal diagnosis of genetic and chromosomal abnormalities (see Chapter 72).

### Buccal Cells

Collection of buccal cells (cells of the oral cavity of epithelial origin) has been identified as providing an excellent source of

genomic DNA. Collection of buccal cells is often viewed as less invasive than collection of blood. It is particularly useful for collecting cells with the patient's genomic DNA when the patient has had blood transfusions and thus has blood with another person's (or persons') DNA. Similarly, it is useful after bone marrow transplantation when the circulating blood cells are derived wholly or partially from the donor of the bone marrow. Two methods are used commonly to collect buccal cells: rinsing with mouthwash and using swabs or cytobrushes.

Rinsing of the oral cavity generally provides a higher yield of cells than can be obtained by using swabs. For these collections, the patient is provided with a small amount of mouthwash and is instructed to rinse well for a minimum of 60 seconds; then the patient returns the mouthwash to a collection tube. There is no harm in doing this longer than 60 seconds, but shortening the time may decrease the yield of buccal cells. Mouthwash solutions high in phenol and ethanol are destructive to recovered cells and should be avoided. It is necessary for each laboratory to validate a list of acceptable solutions.

Swabs or cytobrushes have also been used to collect buccal cells for molecular genetics testing. For swabs, a sterile Dacron or rayon swab with a plastic shaft is preferred because calcium alginate swabs or swabs with wooden sticks may contain substances that inhibit PCR-based testing. After collection, the swab or cytobrush should be stored in an air-tight plastic container or immersed in liquid, such as phosphate-buffered saline or viral transport medium. In general, the yield of cells and nucleic acid is lower with physical scraping using swabs or cytobrushes than with rinsing.

Although collection of buccal cells from children is rare except in the situation of identification of paternity or maternity, the same process is followed.

### Solid Tissue

Traditionally, the solid tissue most often analyzed in the clinical laboratory was malignant tissue from the breast for estrogen and progesterone receptors. During surgery, at least 0.5 to 1 g of tissue is removed and trimmed of fat and nontumor material. This tissue is quickly frozen, within 20 minutes, preferably snap frozen in liquid nitrogen or in a mixture of dry ice and alcohol. A histologic section should always be examined at the time of analysis of the specimen to confirm that the specimen is indeed malignant tissue. Another traditional use of solid tissue analysis is measurement of liver iron or copper to assist with the diagnosis of hemochromatosis or Wilson's disease, respectively.

The same procedure may be used to obtain and prepare solid tissue for elemental or toxicologic analysis; however, when trace element determinations are to be made, all materials used in the collection or handling of the tissue should be made of plastic or materials known to be free of contaminating trace elements.

Somatic gene analysis such as T and B cell clonality and the identification of possible clinically actionable mutations in malignant tissue (*KRAS* mutations, *MGMT/MLH* methylation status) are now proving to be of increasing importance for clinicians in both diagnosis and direction of appropriate therapeutic options for the patient. For these studies, the molecular diagnostic laboratory often receives tissue that has been formalin-fixed and paraffin-embedded (FFPE) rather

than fresh tissue because the request for further testing is generally made after the pathologist's diagnosis of the particular malignancy. In general, neutral buffered formalin, containing no heavy metals, will not interfere with amplification reactions. However, recovery of nucleic acids is greatly decreased if the tissue has been over-fixed or if a decalcification process has been applied to, for example, bone marrow samples. DNA can still be extracted from tissue embedded in paraffin, but the DNA will be degraded to low-molecular-weight fragments. In most cases, segments of DNA will amplify in a PCR reaction, but Southern blot methods are problematic because most require high-molecular-weight DNA.

Tissue structure and better recovery of DNA can be retained without permanent fixation by freezing specimens in an optimal cutting temperature compound (OCT). OCT is a mixture of polyvinyl alcohol and polyethylene glycol that surrounds but does not infiltrate the tissue. The sample is then frozen at about  $-80^{\circ}\text{C}$ , and sections are prepared for review by a pathologist. OCT is fully water soluble and should be completely removed from a tissue specimen before it is used as a source of DNA. In general, DNA of higher molecular weight can be extracted from OCT-fixed tissues compared with that extracted from FFPE samples.

### Hair and Nails

Hair and fingernails or toenails have been used for trace metal and drug analyses with the potential advantage of timing of exposure if separate segments of longer hair are analyzed, although no current standards for such testing currently exist. For the latter examinations, clear labeling of the follicular end of the sample is required. However, collection procedures have been poorly standardized, and quantitative measurements are better obtained from blood or urine. Use of such samples requires each laboratory to validate the processes because, again, there are no published standards for this unusual specimen type. Currently, the use of hair or nails in molecular diagnostics is limited to forensic analysis (genomic DNA identification).

## HANDLING OF SPECIMENS FOR ANALYSIS

Steps that are important for obtaining a valid specimen for analysis include (1) identification, (2) preservation, (3) separation and storage, and (4) transport.

### Maintenance of Specimen Identification

Although the collection of an acceptable specimen is a key aspect of excellent testing, proper identification of the specimen must be maintained at each step of the testing process to ensure that the correct result is reported for the correct patient at all times. The minimum information on any label associated with a specific specimen should include the patient's name, location, and identifying number, as well as the date and time of collection. Many institutions also require the collecting person's initials or some means of identifying the person who collected the sample be included on the label. All labels should conform to the laboratory's stated requirements to facilitate proper processing of specimens. In the United States, no specific labeling should be attached to specimens from patients with infectious diseases that are submitted to the routine laboratory to suggest that these specimens should be handled with special care. Universal precautions

should always be used, meaning that all specimens should be treated as if they are potentially infectious with the following caveat. The exception will be samples from patients suspected to have known, high-risk pathogens (e.g., hemorrhagic viruses such as Ebola) that must have separate, pre-prepared sample handling protocols that may or may not involve the routine laboratory. Proactive procedures must be in place including by whom and where such samples might be analyzed unless the sample can be rendered safe (e.g., by heat treatment), in which case again standard universal precautions should be used.

In practice, every specimen container must be adequately labeled even if the specimen must be placed in ice or if the container is so small that a label cannot be placed along the tube, as might happen with a capillary blood tube. Direct labeling of a capillary blood tube by folding the label like a flag around the tube is preferred or recommended by most laboratories. For small volumes of urine submitted in screw-cap urine cups and any specimen submitted in a screw-cap test tube or cup, the label should be placed on the cup or tube directly, not just on the cap.

It is critical that samples be positively identified through all steps of processing and analysis, and this is especially important in pediatrics because the samples are often collected in small tubes that cannot be sampled from directly by an automated instrument. Aliquoting a sample from the primary collection container to one or more other containers configured for the instrumentation requires close attention to proper labeling and tracking of the sample identifiers to ensure samples are not switched. Good work practice includes "piece work" in which only a single patient's samples are in the work area at one time, the area is clean with no old labels present, and the worker is not disturbed. Although this may not be possible in a large laboratory facility, training that emphasizes the criticality of this function to achieve best patient care and adherence to all policies can be equally effective. Because the majority of samples received from pediatric patients are in microtubes, many samples need to be aliquoted; poured into a microsampling device; or hand entered into instruments that use whole blood, such as hematology instruments, because these systems are not made to deal with such small tubes. Additionally, many times bar codes do not fit these tubes. Bar code readers in instruments cannot be used unless the sample is aliquoted into a larger tube. This extra handling of the specimens offers more opportunities for error and thus requires stricter attention to detail and analysis of the possible risks during design of the process. Special attention should be placed on molecular diagnostics, forensic specimens, and transfusion medicine specimens as applicable.

### Preservation of Specimens

The practitioner must ensure that specimens are collected into the correct container and are properly labeled; in addition, specimens must be properly treated both during transport to the laboratory and from the time the serum, plasma, or cells have been separated until analysis. For some tests, it is crucial that the sample be analyzed immediately from the time the blood is drawn to minimize metabolism and degradation of sample components. Examples are specimens for ammonia and blood gas determinations, such as  $\text{PCO}_2$ ,  $\text{PO}_2$ , and blood pH, which should be analyzed within 30 minutes

after collection. Specimens for lactate, pyruvate, and certain hormone tests (e.g., adrenocorticotrophic hormone [ACTH], gastrin and renin activity) are also time sensitive and may require transport of the specimen on ice after collection. A notable decrease in pyruvate and increase in lactate concentration occurs within a few minutes at ambient temperature. Information on sample stability is generally provided by kit manufacturers, however this often does not cover all circumstances, for example, stability in whole blood prior to centrifugation, and the manufacturer may not have provided information for the required time period. Laboratories should gain extra information from the literature or conduct local studies as needed.

For all test constituents that are thermally labile, serum and plasma should be separated from cells in a refrigerated centrifuge. Specimens for bilirubin or carotene and for some drugs, such as methotrexate, may need to be protected from both daylight and fluorescent light to prevent photodegradation, although the use of plastic rather than glass tubes has decreased this preanalytical variable.

For molecular diagnostic laboratories, a substantial challenge is the recovery of RNA from transported specimens. Depending on the tissue source, RNA yields vary, primarily because of the amount of RNA present at the time of collection. Specimens from the liver, spleen, and heart have larger amounts of RNA than specimens from skin, muscle, and bone. Increasingly, creative solutions to this issue continue to be produced with collection kits that contain stabilizers and even the first reagents required for extraction, all of which have the effect of maximizing the recoverable nucleic acid. Tissue samples should be frozen immediately. Alternatively, a blood specimen should never be frozen before separation of the cellular elements because of hemolysis and released hemoglobin that may interfere with subsequent amplification processes. For tissue samples, it is critical to choose the disruption method best suited for the specific type of tissue. Thorough cellular disruption is critical for high RNA quality and yield. RNA that is trapped in intact cells is often removed with cellular debris by centrifugation.<sup>26</sup>

For specimens that are collected in a remote facility with infrequent transportation by courier to a central laboratory, proper specimen processing must be done in the remote facility so that appropriately separated and preserved plasma or serum is delivered to the laboratory. This necessitates that the remote facility has ready access to appropriately calibrated centrifuges, all commonly used preservatives, and wet ice.

### Add-on Requests

In the interest of preventing additional phlebotomies and to assess a clinical situation from a specimen collected at a specific time, many physicians request an “add-on” test, that is, for the laboratory to perform a test on a sample already in the laboratory and processed. This is especially true in specimens collected from pediatric patients, in whom more blood may not be able to be collected promptly, and for patients from an emergency department, where additional testing from the time of presentation with specific symptoms may be needed after a clinical diagnosis has been made or narrowed by the clinician. Each laboratory must establish its own guidelines for what will be allowed in what time frame. For example, evaporation of small or even routine samples with requests for volatile compounds such as ethanol or methanol can

make them unsuitable for additional testing, so storage conditions and time in the laboratory are important considerations. Also, most samples are stored at refrigerated temperatures after initial analysis; this makes them unacceptable for LDH analysis later but does not affect, for example, alkaline phosphatase or electrolyte analysis, provided that evaporation and air exposure has been kept to a minimum.

### Separation and Storage of Specimens

Plasma or serum should be separated from cells as soon as possible and certainly within 2 hours<sup>27</sup> for some but not all analytes. Premature separation of serum, however, may permit continued formation of fibrin, which can clog sampling devices in testing equipment. If it is impossible to centrifuge a blood specimen within 2 hours, the specimen should be held at room temperature rather than at 4 °C to decrease any effect on potassium measurement caused by leakage from the RBCs by inhibition of the Na/K ATPase pump. For most plasma samples used for molecular diagnostics, the plasma should be removed from the primary tube promptly after centrifugation and held at –20 °C in a freezer capable of maintaining this temperature. In all instances of freezing a sample, frost-free freezers should be avoided because they have a wide temperature swing during the freeze–thaw cycle. Although changes in concentration of test constituents have been observed when serum or plasma is stored in a gel separator tube in a refrigerator for 24 hours, these changes do not appear to be large enough to be of clinical significance.

Primary specimen tubes should always be centrifuged with the original cap in place. Such containment reduces evaporation, which occurs rapidly in a warm centrifuge with the air currents set up by centrifugation. Caps on the original tube also prevent aerosolization of infectious particles and thus provide a further safeguard for laboratorians. Specimen tubes with requested test for volatiles, such as ethanol, *must* have the initial cap in place while they are spun to prevent release of the volatile compound and result in an artificially reduced measurement of ethanol, methanol, or such compounds. Centrifuging specimens with the cap in place also maintains anaerobic conditions, which are important in the measurement of carbon dioxide and ionized calcium. Removal of the stopper before centrifugation allows loss of carbon dioxide and an increase in blood pH. Control of pH is especially important for the accurate measurement of ionized calcium.

Cryopreservation of WBC and DNA is one method to store and maintain samples for extended periods of time. Whole blood specimens can be centrifuged and WBCs removed and cryopreserved at –20 °C in a temperature-controlled freezer until these cells are required for DNA extraction. For even longer periods of storage, isolated DNA can be stored at –70 °C, although 4 °C may be adequate for most purposes. The extracted DNA should not be exposed to repetitive cycles of freezing and thawing because this can lead to shearing of the DNA. After these extracted DNA samples have completely thawed, it is important to fully mix the sample to ensure a homogeneous specimen.

### Transport of Specimens

Hemolysis may occur in pneumatic tube systems unless the tubes are completely filled and movement of the blood tubes inside the specimen carrier is prevented.<sup>28</sup> The pneumatic

tube system should be designed to eliminate sharp curves and sudden stops of specimen carriers because these factors are responsible for much of any in vitro hemolysis that may occur. With many systems, however, the plasma hemoglobin concentration may be increased, and the serum activity of RBC enzymes, such as LDH and AST, may also be increased (see the earlier discussion on the effect of hemolysis). Nonetheless, the amount of hemolysis from transport issues is usually so small that it can be ignored. In special cases, such as a patient undergoing chemotherapy whose cells are fragile or leukemia patients with fragile leukocytes, samples should be centrifuged before they are placed in the pneumatic tube system or identified as “messenger delivery only” and delivered rapidly to the laboratory. There are also occasional tests that cannot be transported to the laboratory via a pneumatic tube system because of the effect the transport has on the test results. For example, sending blood gas samples through the tube system has been shown to adversely affect  $\text{PO}_2$  results,<sup>29</sup> and samples for thromboelastography and rotational thromboelastometry or platelet function testing may also require hand delivery to the laboratory.

Although the remaining discussion uses the specific example of referral laboratory testing by another laboratory, many of the issues discussed, such as regulations related to shipping, are also relevant to a laboratory that receives specimens from outlying clinics via a courier service, which may be laboratory owned or operated. This may involve validating specific transport or storage conditions that are not specified in or in conflict with existing CLSI recommendations.<sup>30</sup> For example, a laboratory may have a clinic that provides sweat chloride collection and sends the sweat samples to the main laboratory for chloride analysis through a courier service. In all cases, the appropriate transport parameters for these samples must be validated.

Before a referral laboratory is used for any tests, the quality of its work should be verified by the referring laboratory. Guidelines for selection and evaluation of a referral laboratory have been published (QMS05-A2, see Table 4.1). For laboratories accredited by the College of American Pathologists (CAP), it is a requirement that the referring laboratory validate that the referral laboratory is CLIA'88 certified by obtaining a copy of the Clinical Laboratory Improvement Act (CLIA) certificate before specimens are shipped. For molecular diagnostic testing, this is of particular importance because often the latest genetic test being requested by a physician has not yet been moved from research interest status to patient care status and may not be available in a CLIA-certified laboratory.

Specimen type and quantity and specimen handling requirements of the referral laboratory must be observed, and in laboratories operating under CLIA'88 regulations, test results reported by a referral laboratory must be identified as such when they are filed in a patient's medical record. The director of a referring laboratory has the responsibility to ensure that specimens will be adequately transported to the referral laboratory. Also, the director should determine the benefits of different services and should keep in mind that the fastest service may be the most expensive. The director should also know that specimens should not be sent to a referral laboratory at the end of the week or in a holiday period because more delays in transit occur during these times than during the working week, and deterioration of specimens is more likely.

It should be assumed that transport from a referring laboratory to a referral laboratory may take as long as 72 hours.

Under optimal conditions, a referring laboratory should retain enough of the original specimen for retesting in case an unanticipated problem arises during shipment, although this essentially never happens in pediatric laboratories where sample volume is at a minimum. Most reference laboratories have lower minimum volume requirements for pediatric specimens than for adult specimens, but these lower minimums generally preclude being able to retest the sample if there is a problem with the initial analysis. The tube and transport condition for the specimen should be constructed such that the contents do not escape if the container is exposed to extremes of heat, cold, or sunlight. Reduced pressure of 0.50 atmosphere (50 kPa) may be encountered during air transport, together with vibration, and specimens should be protected from these adverse conditions by a suitable container. Variability in temperature is a significant factor causing instability of test constituents.

Polypropylene and polyethylene containers are usually suitable for specimen transport. Glass should be avoided. Polystyrene is unsuitable because it may crack when frozen. Containers must be leak-proof and should have a Teflon-lined screw cap that does not loosen under the variety of temperatures to which the container may be exposed. The materials of both stopper and container must be inert and must not have any effect on the concentration of the analyte.

In situations in which sample delivery for molecular analysis will be delayed, extracted nucleic acid, usually DNA only, can be transported in a buffer solution or water, or it can be dried down and shipped as a loose powder. With either method, DNA should be transported at ambient temperatures and should not be exposed to extremely high temperatures for an extended period of time because it will begin to degrade and testing may be compromised. Because dried blood spot samples are so easy to store and transport, and with an increasing number of DNA tests being developed using dried blood spots (e.g., PCR testing for cystic fibrosis and severe combined immunodeficiency), such samples may become one of the best ways to collect, store, and ship samples for DNA testing.

The shipping or secondary container used to hold one or more specimen tubes or bottles must be constructed to prevent the tubes from contact with another specimen. Corrugated, fiberboard, or Styrofoam boxes designed to fit around a single specimen tube are commonly used. A padded shipping envelope provides adequate protection for shipping single specimens. When specimens are shipped as drops of blood on filter paper (e.g., for neonatal screening), the paper should be enclosed in a paper envelope to ensure that the sample remains dry. The initial paper envelope can be placed in a shipping envelope and transported to the testing facility; rapid shipping is rarely required for dried blood on paper.

For transport of frozen or refrigerated specimens, a Styrofoam container should be used. The container walls should be 1 inch (2.5 cm) thick to provide effective insulation. The container should be vented to prevent buildup of carbon dioxide under pressure and a possible explosion. Solid carbon dioxide (dry ice) is the most convenient refrigerant material for keeping specimens frozen, and temperatures as low as  $-70^{\circ}\text{C}$  can be achieved. The amount of dry ice required in a container depends on the size of the container, the efficiency of its insulation, and the length of time for which the specimens must be kept frozen. One piece of solid dry ice (about 3 inches  $\times$  4 inches  $\times$  1 inch) in a container with 1-inch Styrofoam walls and a volume of 125 cubic inches (2000  $\text{cm}^3$ )

will maintain a single specimen frozen for 48 hours. More commonly, smaller pieces of the solid will be used and it is critical that the specimen be buried rather than sitting on top of this refrigerant.

Various laws and regulations apply to the shipment of biologic specimens. Although such regulations theoretically apply only to etiologic agents (known infectious agents), all specimens should be transported as if the same regulations apply. In many countries, airlines have rigid regulations covering the transport of specimens. Airlines deem dry ice a hazardous material; therefore the transport of most clinical laboratory specimens is affected by the regulations and those who package the specimens should be trained in the appropriate regulations, such as those put forth by the US International Air Transport Association.

The various modes of transport of specimens influence the shipping time and cost, and each laboratory needs to make its own assessment as to adequate service. The objective is to ensure that the properly collected, processed, and identified specimen arrives at the testing facility in time and under the correct storage conditions so that the analytical phase can then proceed.

## CONCLUSION

Accurate test results (i.e., the right result for the right patient) begin and end with the integrity of the sample being tested. Integrity can only be assured by proper preparation of the patient; choice of sample container; collection of the sample; and finally transport, processing, and storage of the collected sample with each step maintaining proper identification. Every step of the process affects the quality of the end result. For these reasons, best laboratory practice demands attention to detail and following appropriate protocols. It is incumbent on laboratories and laboratory professionals to fully delineate their processes in complete policies and procedures that not only cover the routine and correct procedures but also cover the unusual. These should include how to handle the process when the system breaks down and steps are not properly performed and may need to be addressed case-by-case by the laboratory director. Finally, laboratories should fully validate protocols that may not be covered under normal procedures or that may deviate from local regulatory guidelines (e.g., the US Food and Drug Administration or CLIA). Preanalytical variables can be lessened or even avoided if these steps are followed.

## POINTS TO REMEMBER

- Proper identification of the patient is essential, and the sample should be properly labeled at all steps, including when separated from the primary collection container.
- Policies designed to ensure the safety of both the patient and the person collecting any sample should always be followed.
- Collection of all samples must be in the correct primary container, and those collecting specimens should understand the biochemical or chemical actions of any additive and possible implications on the test result.
- Attention to the details related to processing of the collected sample (time, temperature, special handling) should always follow validated local policies.
- The accurate result for any patient's sample that will be acted on by the clinician depends on adherence to all policies and procedures, and personnel collecting specimens bear a tremendous responsibility to ensure that they do not contribute to errors that may impact patient care.

## SELECTED REFERENCES

1. The Joint Commission. National Patient Safety Goals. <https://www.jointcommission.org/en/standards/national-patient-safety-goals/hospital-2020-national-patient-safety-goals/> [accessed 2/10/2020].
2. WHO Guidelines on Drawing Blood: Best practices in phlebotomy. Pediatric and neonatal blood sampling. World Health Organization; 2010 [Chapter 6] <<http://www.ncbi.nlm.nih.gov/books/NBK138647/>>; [accessed 2/10/2020].
3. Renoie BW, McDonald JM, Ladenson JH. The effects of stasis with and without exercise on free calcium, various cations, and related parameters. *Clin Chim Acta* 1980;103:91–100.
4. McNair P, Nielsen SL, Christiansen C, et al. Gross errors made by routine blood sampling from two sites using a tourniquet applied at different positions. *Clin Chim Acta* 1979;98:113–18.
5. Cornes MP, Ford C, Gama R. Spurious hyperkalemia due to EDTA contamination: common and not always easy to identify. *Ann Clin Biochem* 2008;45:601–3.
6. Mikesh LM, Bruns DE. Stabilization of glucose in blood specimens: mechanism of delay in fluoride inhibition of glycolysis. *Clin Chem* 2008;54:930–2.
7. Fokker M. Stability of glucose in plasma with different anticoagulants. *Clin Chem Lab Med* 2014;52:1057–60.
8. Laessig RH, Indriksons AA, Hassemer DJ, et al. Changes in serum chemical values as a result of prolonged contact with the clot. *Am J Clin Pathol* 1976;66:598–604.
9. Farnsworth CW, Webber D, Budelius M, Bartlett N, Gronowski AM. Parameters for validating a hospital pneumatic tube system. *Clinical Chemistry*. 2019;65(5):694–702.
10. Victor PJ, Patole S, Fleming JJ, et al. Agreement between paired blood gas values in samples transported either by a pneumatic tube system or by human courier. *Clin Chem Lab Med* 2011; 49:1303–9.
11. Haverstick DM, Brill LB, Scott MG, et al. Preanalytical variables in measurement of free (ionized) calcium in lithium heparin-containing blood collection tubes. *Clin Chim Acta* 2009;403: 102–4.

## REFERENCES

1. The Joint Commission. National Patient Safety Goals. <https://www.jointcommission.org/en/standards/national-patient-safety-goals/hospital-2020-national-patient-safety-goals/> [accessed 2/10/2020].
2. WHO Guidelines on Drawing Blood: Best practices in phlebotomy. Pediatric and neonatal blood sampling. World Health Organization; 2010 [Chapter 6] <<http://www.ncbi.nlm.nih.gov/books/NBK138647/>>; [accessed 2/10/2020].
3. Garza D, Becan-McBride K. Pediatric and geriatric procedures. In: Garza D, Becan-McBride K, editors. Phlebotomy handbook: blood specimen collection from basic to advanced. 10th ed. Upper Saddle River, NJ.: Pearson Prentice Hall; 2019. p. 423–457.
4. Hoeltke LB. Caring for the pediatric patient. In: Hoeltke LB, editor. The complete textbook of phlebotomy. 3rd ed. Clifton Park, NY: Thomas Delmar Learning; 2006. p. 249–2645.
5. Jones PM, Patel K. Pediatric Clinical Biochemistry: Why is it Different? In: Dietzen D, Bennett MJ, Wong E (eds), Biochemical and Molecular Basis of Pediatric Disease, 5th Edition, Cambridge MA, Elsevier; 2020. In press
6. Thavendiranathan P, Bagai A, Ebidia A, et al. Do blood tests cause anemia in hospitalized patients? *J Gen Intern Med* 2005;20:520–4.
7. Renoe BW, McDonald JM, Ladenson JH. The effects of stasis with and without exercise on free calcium, various cations, and related parameters. *Clin Chim Acta* 1980;103:91–100.
8. McNair P, Nielsen SL, Christiansen C, et al. Gross errors made by routine blood sampling from two sites using a tourniquet applied at different positions. *Clin Chim Acta* 1979;98:113–18.
9. Cornes MP, Ford C, Gama R. Spurious hyperkalemia due to EDTA contamination: common and not always easy to identify. *Ann Clin Biochem* 2008;45:601–3.
10. Garza D, Becan-McBride K. Venipuncture Procedures. In: Garza D, Becan-McBride K, editors. Phlebotomy handbook: blood specimen collection from basic to advanced. 10th ed. Upper Saddle River, NJ.: Pearson Prentice Hall; 2019. p. 308–370.
11. Garza D, Becan-McBride K. Capillary or dermal blood specimens. In: Garza D, Becan-McBride K, editors. Phlebotomy handbook: blood specimen collection from basic to advanced. 10th ed. Upper Saddle River, NJ.: Pearson Prentice Hall; 2019. p. 371–395.
12. McCall RE. Capillary puncture equipment, principles and procedures. In: McCall RE. Phlebotomy Essentials, 7<sup>th</sup> ed. Philadelphia, PA.: Wolters Kluwer; 2020. p. 303–335.
13. Green AMI, Gray J. Neonatology & laboratory medicine. London, UK: ACB Venture Publications; 2003.
14. Hoeltke LB. The challenge of phlebotomy. In: Hoeltke LB, editor. The complete textbook of phlebotomy. 3rd ed. Clifton Park, NJ: Thomas Delmar Learning; 2006. p. 227–48.
15. Garza D, Becan-McBride K. Blood cultures, arterial, intravenous (IV) and special collection procedures. In: Garza D, Becan-McBride K, editors. Phlebotomy handbook: blood specimen collection from basic to advanced. 10th ed. Upper Saddle River, NJ.: Pearson Prentice Hall; 2019. p. 473–506.
16. McCall RE. Arterial puncture procedures. In: McCall RE. Phlebotomy Essentials, 7<sup>th</sup> ed. Philadelphia, PA.: Wolters Kluwer; 2020. p. 435–452.
17. Mikesh LM, Bruns DE. Stabilization of glucose in blood specimens: mechanism of delay in fluoride inhibition of glycolysis. *Clin Chem* 2008;54:930–2.
18. Fokker M. Stability of glucose in plasma with different anticoagulants. *Clin Chem Lab Med* 2014;52:1057–60.
19. Ridefelt P, Akerfeldt T, Helmersson-Karlqvist J. Increased plasma glucose levels after change of recommendation from NaF to citrate blood collection tubes. *Clin Biochem* 2014; 47:625–8.
20. Jeon H, Han A, Kang H, Lee KH, Lee JH, Lee JH. A comparison of coagulation test results from heparinized central venous catheter and venipuncture. *Blood Coagul Fibrinolysis*. 2020; 31(2):145–151.
21. Young DS. Effects of preanalytical variable on clinical laboratory tests. 3rd ed. Washington, DC: AACC Press; 2007.
22. Wolf AMD, Fontham ETH, Church TR, Flowers CR, Guerra CE, LaMonte SJ, Etzioni R, McKenna MT, Oeffinger KC, Shih Y-CT, et al. Colorectal cancer screening for average-risk adults: 2018 guideline update from the American Cancer Society. *CA* 2018;68(4):250–281.
23. AACC Lab Tests Online. Body Fluid Testing. <https://labtestsonline.org/tests/body-fluid-testing> [accessed February 14, 2020]
24. Natsugoe S, Tokuda K, Matsumoto M. Molecular detection of free cancer cells in pleural lavage fluid from esophageal cancer patients. *Int J Mol Med* 2003;12:771–5.
25. Carroll T, Raff H, Findling JW. Late-night salivary cortisol for the diagnosis of Cushing's syndrome: a meta-analysis. *Endocr Pract* 2009;6:1–17.
26. Groszbach A. Nucleic acid preparation. Presented at: 4th Annual University of Connecticut Molecular Review Symposium, 26th Annual Meeting of the Association of Genetic Technologists, May 30, 2001, Minneapolis, Minn.
27. Laessig RH, Indriksons AA, Hassemer DJ, et al. Changes in serum chemical values as a result of prolonged contact with the clot. *Am J Clin Pathol* 1976;66:598–604.
28. Farnsworth CW, Webber D, Budelius M, Bartlett N, Gronowski AM. Parameters for validating a hospital pneumatic tube system. *Clinical Chemistry*. 2019;65(5):694–702.
29. Victor PJ, Patole S, Fleming JJ, et al. Agreement between paired blood gas values in samples transported either by a pneumatic tube system or by human courier. *Clin Chem Lab Med* 2011;49: 1303–9.
30. Haverstick DM, Brill LB, Scott MG, et al. Preanalytical variables in measurement of free (ionized) calcium in lithium heparin-containing blood collection tubes. *Clin Chim Acta* 2009;403: 102–4.

**MULTIPLE CHOICE QUESTIONS**

1. Identify the incorrect scenario related to specimen collection below.
  - a. Ask a parent to name a pediatric patient.
  - b. Ask an adult patient to state his or her full name and date of birth.
  - c. Ask an adult patient to state his or her full name and telephone number.
  - d. Show the patient or parent the labeled tube after collection is complete.
  - e. Check the room and bed number of the patient to be collected and proceed with collection.
2. Which of the following is true about specimen identification and labeling?
  - a. One patient identifier is adequate before specimen collection.
  - b. Aliquot tubes do not need labeling.
  - c. The Joint Commission requires at least three unique identifiers
  - d. Prelabeling tubes before collection is acceptable.
  - e. Date and time of collection should be included on a properly labeled specimen.
3. Identify the correct statement regarding tube additives and biochemical action with clinical use below:
  - a. A NaF tube helps to prevent glycolysis and is used for glucose measurement.
  - b. A sodium citrate tube with chelation of calcium allows effective coagulation testing.
  - c. A no additive, no gel plasma sample does not need to clot before use.
  - d. An EDTA anticoagulated specimen can be used for calcium determination.
  - e. Lithium heparin tubes are widely used for chemistry tests and require clotting before use.
4. Patients should be cautioned against pumping their fists after the tourniquet is applied because,
  - a. it may or may not cause venous stasis, which is not under the control of the collector.
  - b. it will affect the chemical interaction between the additive in the tube and the patient's blood.
  - c. metabolic activity of the blood and muscles will affect analytes such as potassium and calcium through effects on pH and possibly creatine kinase.
  - d. it will affect the distribution of analytes between plasma and serum.
  - e. it will make it more difficult to find the patient's vein.
5. Identify the correct order of tubes in which blood should be collected for multiple specimens:
  - a. Yellow (no additive), Gold/ red (serum), Green (heparin), Lavender (EDTA), Gray (Sodium fluoride)
  - b. Gold/ red (serum), Yellow (no additive), Green (heparin), Lavender (EDTA), Gray (Sodium fluoride)
  - c. Gray (Sodium fluoride), Lavender (EDTA), Green (heparin), Gold/ red (serum), Yellow (no additive)
  - d. Yellow (no additive), Gold/ red (serum), Green (heparin), Gray (Sodium fluoride), Lavender (EDTA)
  - e. Green (heparin), Yellow (no additive), Gold/ red (serum), Gray (Sodium fluoride), Lavender (EDTA)
6. For urine collections, which of the following is true?
  - a. Samples should be kept at room temperature for timed collections.
  - b. Proper collection technique on an infant utilizes "bagging" with a small plastic collection bag.
  - c. Preservatives in a 24-hour collection jug do not need to be identified to the patient.
  - d. A 24-hour collection includes the first morning urine sample both at the start and the finish of collection.
  - e. A "clean catch" is not necessary for urine culture collections.
7. Which of the following analytes would be most affected for arterialized capillary blood sample if the site is not appropriately warmed?
  - a. pH
  - b.  $PCO_2$
  - c.  $PO_2$
  - d.  $PO_2$  and  $PCO_2$
  - e. pH and  $PO_2$
8. Which of the following analytes does not exhibit a difference in composition between heparin plasma and serum?
  - a. Cholesterol
  - b. Total protein
  - c. Glucose
  - d. Lactate dehydrogenase
  - e. Potassium
9. Which of the following analytes exhibits the greatest difference in composition between capillary and venous serum?
  - a. Phosphate
  - b. Urea
  - c. Glucose
  - d. Potassium
  - e. Calcium
10. Additional testing (add-ons) performed on a previous sample requires:
  - a. any sample that was collected within the appropriate time frame.
  - b. a sample that has been sitting at room temperature.
  - c. sufficient sample volume left on a sample of the correct type for the add-on assay.
  - d. using an unlabeled aliquot of the original sample.
  - e. an additional sample be collected for the test.

# Preanalytical Variation and Pre-Examination Processes

*Ana-Maria Simundic, Lora Dukić, and Vanja Radisic Biljak*

## ABSTRACT

### Background

The preanalytical phase has long been recognized as a source of substantial variability in laboratory medicine. Laboratory errors, mostly due to some defect in the preanalytical phase, may lead to diagnostic errors. Understanding preanalytical variation and reducing errors in the pre-examination phase of the testing process are therefore important for improved safety and quality of laboratory services delivered to patients.

### Content

There are numerous preanalytical factors that may affect the concentration of the analyte, the measurement procedure, or the test result. These factors may be divided into two major groups: influencing and interference factors. Influencing factors are effects on laboratory results of biological origin that most commonly occur *in vivo* but can also be derived from the sample *in vitro* during transport and

storage. Biological influence factors lead to changes in the quantity of the analyte in a method-independent way. Interference factors (interferences) are defined as mechanisms and factors that lead to falsely increased or decreased results of laboratory tests of a defined analyte. Interference factors and their mechanisms differ with respect to the intended analyte and analytical method. Interference factors do not affect the concentration of the analyte. On the contrary, they alter the test result for a specific analyte after the sample has been collected. They are different from the measured analyte and interfere with the analytical procedure. Therefore their effect is method dependent and may thus be reduced or eliminated by selecting a more specific method. This chapter describes the most common preanalytical sources of variability (influences and interferences) and provides recommendations on how to deal with them in everyday practice.

## INTRODUCTION

The incidence of premature patient deaths associated with some kind of preventable medical error has been estimated to be 98,000 per year.<sup>1</sup> More recent data indicate that the actual mortality caused by preventable medical errors is fourfold higher.<sup>2</sup> According to the European Commission (EC) and World Health Organization (WHO), 1 in 10 patients is being harmed while receiving hospital care in developed countries.<sup>3,4</sup> Errors in laboratory medicine can lead to increased health care expenditure, cause patient harm to various degrees, and lead to different diagnostic errors (i.e., missed diagnosis, misdiagnosis, and delayed diagnosis).<sup>5</sup>

It has been suggested that laboratory test results affect approximately 70% of medical decisions, and this clearly explains why laboratory errors have a large contribution to the overall error frequency in health care.<sup>6,7</sup> Almost 40% of diagnostic errors are attributed to some error that has occurred within the area of radiology or laboratory medicine, and the majority of those laboratory errors are due to some defect in the preanalytical phase of the total testing process (TPP).<sup>8</sup>

### Historical Perspective

The preanalytical or, according to ISO 15189 terminology, *pre-examination* phase of the laboratory testing process has been recognized since the early 1970s as an important source

of variability, and it still represents one of the greatest challenges for specialists in laboratory medicine.<sup>9,10</sup>

In the second half of the 20th century, when quality assurance programs were introduced for the analytical processes, laboratories became aware that some factors outside the analytical phase also significantly impacted laboratory results.<sup>11</sup> Results that did not correspond with the patient's clinical condition have often been called "laboratory errors." It also became clear that these variables could not be standardized or controlled by analytical quality assurance programs. In the late 1970s, Statland and Winkel defined the phase prior to analysis as the "preinstrumental phase,"<sup>12</sup> which was later changed to the "preanalytical phase."<sup>13</sup>

Even before the preanalytical phase was recognized as an important issue in laboratory medicine, some experts from different areas of laboratory medicine defined these variables as *influencing* and *interference* factors,<sup>14,15</sup> which were not immediately recognized as important sources of "laboratory errors." It took some time for laboratory medicine professionals to gather knowledge about their causes and mechanisms and acknowledge their importance.

After years of discussion within several national and international expert groups in the 1960s and 1970s,<sup>12–16</sup> the term *biological influence factor* was introduced and distinguished from interference factors. This led to the definitions established in the 1980s,<sup>17,18</sup> which are still valid today.<sup>19</sup>

## The Preanalytical Phase Today

The preanalytical phase is recognized as the most vulnerable part of the TTP, and it accounts for two-thirds of all laboratory errors.<sup>20</sup> Preanalytical errors can occur at any step of the preanalytical phase—for example, during test requesting, patient preparation, sample collection, sample transport, handling, and storage.<sup>21</sup> This high frequency of preanalytical errors may be attributed to various reasons. Many preanalytical steps are performed outside the laboratory and are not under the direct supervision of laboratory staff. Furthermore, many individuals are involved in various preanalytical steps, and those individuals have different levels of education and professional background. Finally, safe practice standards for many activities and procedures are either not available, or are available but not evidence-based, or the level of compliance with those standards is low.

The ISO 15189 accreditation standard clearly defines that medical laboratories are responsible for the management and quality of the pre-examination phase.<sup>21</sup> It is the role of the laboratorian that the right sample be taken from the right patient at the right time, and that correct test results are provided to the requesting physician in a timely manner. If the quality of the specimen is compromised to a degree where the expected effect is larger than the allowable error, thus causing clinically significant bias, the sample should be rejected for analysis. Our guiding principle should be “No result is always better than a wrong result.” Patient benefit should always be the top priority.

## Influencing and Interference Factors

### Influencing Factors

Influencing factors are the effects on laboratory results of biological origin that most commonly occur *in vivo* but can also be derived from the sample *in vitro* during transport and storage. Biological influence factors lead to changes in the quantity of the analyte to be measured in a defined matrix. They modify the concentration of the measured (affected) analyte in a method-independent way.

These factors are either present in the healthy individual, like circadian rhythms, or they appear as side effects of a disease and its treatment. Influencing factors may be modifiable, such as diet, time of the day, or time of the year (season), or unmodifiable, such as gender, race, ethnicity, genetic background, and so on. Some modifiable biological influence factors can be controlled by patient action—for example, diet—whereas others—for example, age—are not controllable. Particular care should be taken with the influencing factors whose effects may be reduced through standardization of preanalytical conditions.

As already mentioned, modification of the concentration of certain analytes can also occur *in vitro*. For example, glucose concentration will decrease during prolonged storage of unseparated blood due to cell metabolism, whereas potassium concentration will increase if blood is kept at lower temperatures or refrigerated (+4 °C). Such increase in potassium will occur even without visible hemolysis.

### Interference Factors

Interferences are defined as mechanisms and factors that lead to falsely increased or decreased laboratory test results for a defined analyte. Interferences may be endogenous (i.e., biological constituents of the sample) or exogenous. Exogenous

interferences occur in the preanalytical phase due to the action of some external factors or conditions that are not normally present in properly collected, transported, handled, and stored specimens.<sup>22</sup> Interference factors and their mechanisms differ with respect to the intended analyte and analytical method, and they alter the result of a sample constituent after the specimen has been collected. They are different from the measured analyte and interfere with the analytical procedure. Therefore their effect is method dependent and may thus be reduced or eliminated by selecting a more specific method.<sup>14</sup>

Possible interferents include the following:

1. Biological constituents of the sample (e.g., free hemoglobin, lipids, bilirubin, paraproteins, fibrin, fibrin clots, etc.)
2. Exogenous molecules present in the sample (e.g., drugs, herbal supplements, contrast media)
3. Exogenous molecules added to the sample during sampling or after the sampling procedure (e.g., anticoagulants, tube additives, intravenous infusions, etc.)

Because interference factors are analyte and method specific, they may be eliminated or at least reduced by changing the measurement method.

Although exogenous preanalytical interferences are not rare, they are often neglected and overlooked in everyday routine work. If they go undetected, preanalytical interferences may cause unnecessary harm to the patient and increase health care-related costs. Some examples of harmful results due to erroneous immunoassay findings are listed in Table 5.1.

It is very important that laboratory staff has a thorough knowledge and understanding of the assays and instruments in use in their laboratory and potential interferents which may affect laboratory measurements. This chapter provides an overview of the most common preanalytical sources of variability (influences and interferences) and provides recommendations on how to deal with them in practice.

## POINTS TO REMEMBER

### Influencing and Interference Factors

- Influencing factors lead to changes in the quantity of the analyte in a method-independent way.
- Influencing factors may be changeable (e.g., diet, time of sample collection during the day) or unchangeable (e.g., gender, race, ethnicity, genetic background).
- The effect of influencing factors may be reduced through standardization of preanalytical conditions.
- Interferences are mechanisms and factors that lead to falsely increased or decreased results of laboratory tests.
- Interference factors and their mechanisms differ with respect to the intended analyte and analytical method and may be reduced or eliminated by selecting a more specific method.

## INFLUENCING FACTORS

The effect of most modifiable influencing factors can be either minimized or even entirely eliminated by standardization of preanalytical processes. Several local and international guidelines provide recommendations for efficient standardization of patient preparation and sample collection.<sup>24,25</sup> These documents provide guidance on timing of

**TABLE 5.1 Effects of Laboratory Errors on Patient Outcome**

Wrong Result	Consequence
<b>Falsely Increased Concentration</b>	
High human chorionic gonadotrophin indicating gonadal tumor	Unnecessary surgery, chemotherapy
High calcitonin indicating medullary thyroid cancer	Unnecessary fine-needle aspiration
High prolactin	Misdiagnosis of prolactinoma
High urine free cortisol	Unnecessary diagnostic follow-up
High testosterone in women	Unnecessary diagnostic follow-up
High luteinizing hormone and follicle-stimulating hormone	Unnecessary diagnostic follow-up
<b>Falsely Decreased Concentration</b>	
Low 25-hydroxyvitamin D result despite replacement therapy	Incorrect diagnosis of hypovitaminosis D
Negative human chorionic gonadotropin result	Missed diagnosis of choriocarcinoma
Low digoxin	Wrong treatment (overdosing with digoxin, risk of digoxin toxicity)
Low insulin	Missed diagnosis of insulinoma
Negative troponin result	Missed diagnosis of myocardial infarction

Modified from Jones, A.M. & Honour, J.W. Unusual results from immunoassays and the role of the clinical endocrinologist. *Clin Endocrinol (Oxf)* 2006;64:234–44.

sampling, diet and activities before sampling, body position and disinfection during sampling, and regulations regarding documentation of these variables for diagnostic and/or therapeutic purposes.

On the other hand, since the effects of unmodifiable factors cannot be eliminated by standardization, they are addressed by assigning appropriate reference intervals (e.g., gender-specific, age-specific, etc.).

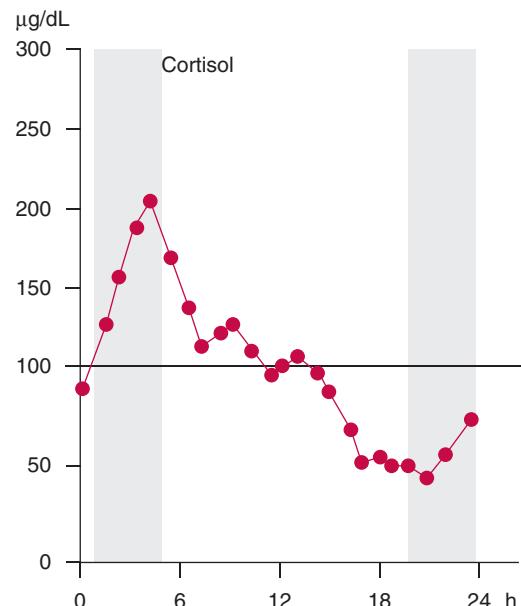
### Controllable Variables

#### Time of Sampling

Time of sampling matters for all analytes which are subject to substantial biological variation. Changes of the concentration of an analyte due to biological variation may significantly affect the given result of a particular laboratory test and their nature can be either linear or cyclic. Linear changes occur in a chronological order, while cyclic changes are of repetitive nature, such as seasonal changes, or changes due to the menstrual cycle. Knowledge about the time of the sample collection is therefore necessary for correct test result interpretation and as such is an important preanalytical factor which should be carefully considered.

Several analytes tend to fluctuate in terms of their plasma concentration over the course of a day, and for this reason reference intervals are preferentially defined for sampling between 7 and 9 am.<sup>26</sup> For example, the concentration of potassium is lower in the afternoon than in the morning, whereas that of cortisol decreases during the day and increases at night (Fig. 5.1).<sup>29</sup> Furthermore, the cortisol circadian rhythm may well be responsible for the poor results obtained from oral glucose tolerance testing in the afternoon.

For many years, it was believed that iron has a substantial circadian variation, with an early morning peak and a decrease in the afternoon; that was the main reason most of the blood collections for iron were done in the morning. It was only recently demonstrated that iron concentration is quite sustainable throughout the day, up until 3 pm, when it starts to decrease, whereas the lowest iron concentrations were observed with collection times past 4 pm.<sup>27</sup>



**FIGURE 5.1** Daily variation of plasma concentrations of cortisol (shaded area = sleep period). (Modified from: Evans K, Laker MF. Intraindividual factors affecting lipid, lipoprotein and apolipoprotein measurement: a review. *Ann Clin Biochem* 1995;32:261–80.)

In some cases, seasonal influences also have to be considered. For example, total triiodothyronine (T3) is 20% lower in the summer than in the winter,<sup>28</sup> whereas 25 OH-cholecalciferol exhibits higher serum concentrations in the summer than in the winter.<sup>29</sup>

Some analytes can exhibit significant changes due to the hormone biological variations that occur during menstruation cycle. For example, aldosterone concentration in plasma is twice as high before ovulation than in the follicular phase, concentration of renin is increased preovulatory, cholesterol exhibits a significant decrease during ovulation, while the concentration of phosphate and iron decreases during menstruation.<sup>28</sup>

## Influence of Diagnostic and Therapeutic Procedures

Time of sampling is not only important to eliminate confounding effects of biological variation, but is also extremely important for patients receiving some diagnostic and/or therapeutic procedures which may cause some *in vivo* (influencing effect, very frequent) or *in vitro* (interference effect, much less common) effects on laboratory tests.<sup>30,31</sup> Some examples of these diagnostic and/or therapeutic procedures are listed below:

- Surgical operations
- Infusions and transfusions
- Punctures, injections, biopsies, palpations, whole-body massage
- Endoscopy
- Dialysis
- Physical stress (e.g., ergometry, exercise, ECG)
- Function tests (e.g., oral glucose tolerance test)
- Immunoscintigraphy
- Contrast media
- Drugs, herbal supplements and over-the-counter medicines
- Mental stress
- Ionizing radiation

Plateletpheresis procedure is used in Transfusion Medicine to obtain platelets needed for treatment of thrombocytopenia. Citrate used in this procedure has chelating effect on ionized calcium and magnesium and along with decreases in some hematology parameters, can also lead to acute ionized hypocalcemia and hypomagnesemia.<sup>32,33</sup> Citrate chelation can also lead to decreased ionized calcium concentration in critically ill patients with high risk of bleeding and acute renal failure who are subjected to citrate-anticoagulated continuous veno-venous hemofiltration. In this situation, hypocalcemia reflects citrate overdose. Due to citrate effect, lower concentrations of ionized calcium and magnesium have also been observed in patients receiving high volumes of transfused blood.

Administration of hypertonic saline in patients having severe head trauma with therapeutic target sodium concentration of 155 mmol/L (or mEq/L) and should not be misinterpreted as contamination with saline intravenous fluid.<sup>34</sup>

Iodinated contrast media used in computed tomography (CT) have high osmolality and high iodine content.<sup>35</sup> Application of these iodinated contrast media for imaging purposes in the pediatric population exceeds normal daily intake of iodine, carries a risk of thyroid dysfunction, and prompts close monitoring of pediatric patients after exposure.<sup>36</sup>

Contrast media may also affect some coagulation and inflammatory parameters. For example, ioxaglate and iodixanol, radiographic contrast media used in diagnostic and therapeutic angiography, may inhibit generation of thrombin in studies performed on platelet-poor and platelet-rich plasma (PRP).<sup>37</sup> The effect of contrast media during coronary angiographic procedure on the inflammatory markers interleukin-6 (IL-6) and soluble (s) receptors (R) for tumor necrosis factor alpha (TNF $\alpha$ ) sTNFR $\alpha$ 1 and sTNFR $\alpha$ 2 has also been reported. Ioxaglate causes the increase in inflammatory markers after contrast media administration, and this effect is more pronounced after the administration of ionic (ioxaglate) compared to nonionic (iohexol and iodixanol) contrast media.<sup>38</sup>

Many drugs, herbal supplements, and over-the-counter medicines may cause various *in vivo* changes of the composition

of the blood and subsequently influence the measurement of many laboratory parameters. For example, long-term treatment with proton pump inhibitors leads to the increase in the concentration of the neuroendocrine tumor marker chromogranin A by stimulating enterochromaffine-like cells. In order to avoid unnecessary diagnostic procedures, it is therefore advised to cease proton pump inhibitors at least 2 weeks before chromogranin A testing.<sup>39</sup>

Another example of the drug which exerts *in vivo* influencing effect is azithromycin. In patients on azithromycin, there is a risk of the occurrence of drug-induced immune thrombocytopenia.<sup>40</sup> Alemtuzumab infusion to patients with active relapsing-remitting multiple sclerosis has remarkable effect on several hematology and biochemistry tests.<sup>41</sup> Cross-sectional data analysis of the Rotterdam study on 9820 participants has demonstrated the significant association of thiazide diuretics use with the increased risk of hypomagnesemia.<sup>42</sup> Trimethoprim, a drug commonly prescribed together with other antibiotics for urinary tract infection treatment, may cause reversible increase in serum creatinine concentration; this increase affects calculation of estimated glomerular filtration rate. Lack of information about this *in vivo* effect of trimethoprim can cause misinterpretation and erroneous clinical decision.<sup>43</sup>

The influencing effect of herbal supplements is exerted through toxicity or enzyme induction. Since herbal supplements are categorized as dietary supplements, they are not subject to strict regulations like drugs. Such permissive regulation carries a significant risk due to the uncontrolled use of herbal supplements alone or in combination with other supplements or drugs.<sup>44-46</sup> Obviously, there is a need to increase the level of awareness about potential risks associated with the use of herbal supplements.<sup>47</sup>

Kava is a traditional medicinal substance used in the Pacific region. It has relaxing effect and is consumed to treat anxiety, as an aqueous nonalcoholic drink made of kava rhizome. Cases of heavy kava consumption are associated with the 70- and 60-fold increase of alanine aminotransferase (ALT) and aspartate aminotransferase (AST) activity, as well as largely increased alkaline phosphatase (ALK),  $\gamma$ -glutamyltransferase (GGT), lactate dehydrogenase (LD), and total and conjugated bilirubin concentration and in extreme cases even with fulminant hepatic failure.<sup>48</sup>

The most probable underlying mechanism of hepatotoxicity is related to metabolic interaction of alcohol with kava, although multiple factors, involving genetic defects in hepatic metabolism, contribute to development of extreme reactions.<sup>49</sup> Some other dietary supplements such as LipoKinetix and Centella asiatica which are largely used for weight loss, may also cause hepatotoxicity and even lead to fulminant hepatic failure associated with extreme increase of liver enzymes, which can be resolved after discontinuation of consumption.<sup>50,51</sup>

Kelp, a kind of seaweed used in Asia as a selenium supplement, is rich in iodine. Patients taking kelp commonly have high serum and urine iodine concentration even if on low-iodine diet, which is mandatory for radioiodine therapy.<sup>52,53</sup>

To prevent the confounding effect of these factors, samples should always be collected before any diagnostic or therapeutic procedure with potential influencing or interfering effects. Likewise, drugs exerting influencing or interfering effects should be administered exclusively after collecting a

blood sample, if not advised differently by the requesting physician (Note: time of sampling for therapeutic drug monitoring is discussed in Chapter 42).

For all the above-mentioned reasons, and given the fact that in some circumstances a sample taken at the wrong time might be worse than taking no sample, exact time of sample collection always needs to be provided to the laboratory.

For effective standardization of the time of sampling, laboratories should ensure that:

- the best time of sampling for each analyte is known (taking into account how the concentration of the particular analyte changes over time),
- blood is always taken at the recommended time,
- the exact time of sampling is known for each sample and is recorded into the laboratory information system (LIS) by the health care staff,
- patients and the health care staff are educated about when blood samples should be collected for laboratory testing, as well as about the importance of the time of blood sampling and its effect on laboratory test results.

### Diet

Prior to blood sampling, the confounding influences of food and fluid intake should be excluded. Diet and fluid intake substantially affect the composition of plasma. Differences in serum composition may occur respective to the source of nutrients, number of meals, and proportion of nutrients in a diet. Moreover, malnutrition or obesity, prolonged fasting, starvation, and vegetarianism may also influence plasma composition. The effects from diet can be divided into long-term and acute effects.

**Long-term effects of diet.** It is well known that changes in protein intake that occur over a couple of days may affect the composition of nitrogenous components of plasma and the excretion of end products of protein metabolism. Creatinine is an important example of the effect of diet on the composition of plasma. It has been shown that an increase of up to 20% of plasma creatinine concentration (measured by kinetic Jaffe method) is observed after ingesting cooked meat.<sup>54</sup> Protein-rich food affects not only the concentration of serum creatinine but also the concentration of urea and urate in serum.

A diet rich in fat leads to increased serum triglyceride concentration, reduced serum urate, and a depletion of the body's nitrogen pool. The nitrogen pool is affected because excretion of ammonium ions is required to maintain acid-base homeostasis.<sup>55,56</sup> The relative ratio in which various dietary fats are consumed closely relates to serum lipid concentrations. A diet rich in monounsaturated and polyunsaturated fats causes a reduction of low-density lipoprotein (LDL) and high-density lipoprotein (HDL) cholesterol concentrations,<sup>57</sup> although in some situations HDL cholesterol may be increased.

A diet rich in carbohydrates decreases serum protein and lipid concentrations (triglycerides, and total and LDL cholesterol).<sup>58</sup> It should be emphasized that not only the proportion but also the source of nutrients in the diet affect the composition of serum. For example, some early studies have shown that serum ALP and LD activities are higher, whereas AST and ALT activities are lower in individuals who consume carbohydrates rich in sucrose or starch rather than other sugar types.<sup>59</sup> Moreover, total, LDL, and HDL cholesterol concentrations tend to be much lower in those who consume

the same amount of food in many small meals throughout the day than in individuals who eat three meals per day.<sup>55</sup>

Compared to omnivorous subjects, vegetarians tend to have lower concentrations of plasma cholesterol, triglycerides, and creatinine, with reduced urinary excretion of creatinine and a higher urinary pH as a result of reduced intake of precursors of acid metabolites.<sup>55</sup> In malnourished individuals, the activity of most of the commonly measured proteins and enzymes is reduced.<sup>60,61</sup> Most of the above-described changes normalize following the restoration of good nutrition.

**Acute effects of diet and other influencing factors.** While it is clear that many analytes are affected by acute ingestion of food, the direction and magnitude of the change still remain largely unclear, mainly due to substantial differences in the design of the original studies published so far. Some of the methodological aspects of these studies which might be responsible for the observed differences are: time of blood collection (morning, afternoon, etc.), time intervals at which blood collection was done (i.e., length of time after the meal), sample type (serum, plasma), assay (method principle, manufacturer), measurement equipment, baseline concentration of the analyte, measurement unit (i.e., mmol/L, mg/dL), type of the meal, food composition, and other patient-related characteristics (e.g., health status, age, gender, ethnicity, physical activity, smoking status, and consumption of alcohol and coffee).<sup>62,63</sup> Moreover, while some postprandial effects are a result of the in vivo physiologic changes, some effects occur due to the interfering effect of sample turbidity caused by the increase of triglyceride (chylomicrons) concentration, and as such are method- and instrument-dependent. Table 5.2 shows the maximal postprandial effects observed anywhere within 1 to 4 hours after a meal on some most common chemistry analytes and hormones.

Certainly, not all changes are clinically significant, but for those analytes for which postprandial changes are clinically significant, fasting prior to blood collection is recommended to overcome this problem. The most pronounced change after a recent meal among chemistry tests is observed in triglycerides. Triglyceride concentrations in serum increase almost twofold during the absorptive phase, within 1 to 2 hours after the meal, and the magnitude of the increase obviously depends on the type of meal and time of sample collection after the meal.

Whereas the nature of the change for most analytes is mostly unidirectional, some analytes, like phosphate, exhibit a characteristic bi-phase change. Concentration of phosphorus initially drops for -2.7 to -8% within 1 hour after the meal and is followed by an increase of up to 12.6% 4 hours after the meal.<sup>65,67</sup>

It is noteworthy to point out that not only chemistry tests but also some hormones (e.g., thyroid stimulating hormone [TSH], free thyroxine [fT4], cortisol, insulin) are significantly affected by acute food ingestion.

Acute food ingestion also affects hematology and coagulation parameters. Tables 5.3 and 5.4 show the maximal postprandial effects observed anywhere within 1 to 8 hours after a meal on complete blood count (CBC) and coagulation parameters.

There is an evident postprandial increase in neutrophil count, along with a decrease in red blood cell (RBC) count, as well as some RBC indices. Variations in RBC count and

**TABLE 5.2 Results of Four Original Studies Demonstrating the Postprandial Effects (% Change) on the Concentration of Some Most Common Chemistry Analytes and Hormones**

	<b>Guder, 2009<sup>64</sup></b>	<b>Lima-Oliveira, 2012<sup>65</sup></b>	<b>Kackov, 2013<sup>66</sup></b>	<b>Bajana, 2019<sup>67</sup></b>
Type of meal	Standard meal	Light meal, containing standardized amounts of carbohydrates, protein, and lipids.	Standardized High-calorie meal (823 kcal)	Andean breakfast, containing standardized amounts of carbohydrates, protein, and lipids.
Blood collection time points	Baseline and 2 h after the meal	Baseline and 1 h, 2 h, 4 h after the meal	Baseline and 3 h after the meal	Baseline and 1 h, 2 h, 4 h after the meal
<b>Analyte</b>				
Triglycerides	78%	28%	71.4%	85%
CRP	NA	25%	-5%	6%
Urea	0% (no change)	-4%	NA	26%
Creatinine	NA	-2.2%	NA	33%
AST	25%	14%	NA	5%
ALT	5.5%	18%	NA	4.3%
Albumin	1.8%	3.4%	NA	4.4%
Bilirubin (total)	16%	-16%	NA	-29%
Bilirubin (direct)	NA	-24%	NA	-29%
Calcium	1.6%	3.5%	NA	4%
Magnesium	NA	3.4%	NA	9%
Iron	NA	10%	NA	-35%
Potassium	5.2%	5.8%	NA	3%
Uric acid	NA	-5%	NA	-3.6%
TSH	NA	NA	NA	27%
fT4	NA	NA	NA	6.6%
Cortisol	NA	NA	NA	-29%

Note: Presented are maximal deviations (% change) that occur anywhere within the observed period of time. Red fields show parameters with an increase, light red fields show parameters with a decrease.

ALT, Alanine aminotransferase; AST, aspartate aminotransferase; CRP, C-reactive protein; fT4, free thyroxine; TSH, thyroid stimulating hormone.

**TABLE 5.3 Results of the Four Original Studies Demonstrating the Postprandial Effects (% Change) on Complete Blood Count Parameters**

	<b>van Oostrom, 2003<sup>68</sup></b>	<b>Lippi, 2010<sup>69</sup></b>	<b>Kościelniak, 2017<sup>70</sup></b>	<b>Arredondo, 2019<sup>71</sup></b>
Type of meal	Standardized oral-fat loading test.	Light meal, containing standardized amounts of carbohydrates, protein, and lipids.	Light meal, containing standardized amounts of carbohydrates, protein, and lipids (300–700 kcal).	Chilean breakfast, containing standardized amounts of carbohydrates, protein, and lipids.
Blood collection time points	Baseline and 2 h, 4 h, 6 h, 8 h after the meal	Baseline and 1 h, 2 h, 4 h after the meal	Baseline and 1 h, 2 h after the meal	Baseline and 1 h, 2 h, 4 h after the meal
<b>Analyte</b>				
WBC	NA	NA	16%	16.9%
Neutrophils	42%	7.6%	37%	27.4%
Lymphocytes	42%	-18.7%	-12%	15.9%
Monocytes	No change	-6.9%	No change	25.0%
Eosinophils	NA	-23.2%	No change	No change
RBC	No change	-3.3%	-7%	-3.4%
Hgb	NA	No change	-8%	-2.7%
Hct	NA	-3.9%	-6%	-4.4%
MCV	NA	No change	No change	-2.1%
MCH	NA	1.6%	No change	No change
Plt	NA	No change	-6%	6.9%
MPV	NA	-2.3%	No change	-8.5%

Note: Presented are maximal deviations (% change) that occur anywhere within the observed period of time. Red fields show parameters with an increase, light red fields show parameters with a decrease.

Hct, Hematocrit; Hgb, hemoglobin; MCH, mean cell hemoglobin; MCV, mean corpuscular volume; MPV, mean platelet volume; Plt, platelets; RBC, red blood cells; WBC, white blood cells.

**TABLE 5.4 Results of the Two Original Studies Demonstrating the Postprandial Effects (% Change) on Coagulation Parameters**

Analyte	Lima-Oliveira, 2014 <sup>72</sup>	Arredondo, 2019 <sup>10</sup>
Activated partial thromboplastin time (aPTT)	-6.2	-4.5
Fibrinogen	No change	-3.1
Antithrombin III	3.7	1.8
Type of meal	Light meal, containing standardized amounts of carbohydrates, protein, and lipids (563 kcal).	Chilean breakfast, containing standardized amounts of carbohydrates, protein, and lipids.
Blood collection time points	Baseline and 1 h, 2 h after the meal	Baseline and 1 h, 2 h, 4 h after the meal

Note: Presented are maximal deviations (% change) that occur anywhere within the observed period of time. *Red fields* show parameters with an increase, *light red fields* show parameters with a decrease.

indices are most likely attributable to hemodilution caused by the ingestion of food and fluids. Postprandial neutrophil increase is suggested to play a role in the pathogenesis of atherosclerosis.<sup>68</sup> The effect of acute ingestion of food and fluid on lymphocyte and platelet count is less clear and is most likely instrument-dependent.

Postprandial triglyceridemia causes a transient increase of the plasma levels of the activated factor VII (FVIIa) and plasminogen activator inhibitor (PAI-1); mechanism of this phenomenon is still not completely understood.<sup>73,74</sup> As FVIIa is the first enzyme of the blood coagulation system, postprandial phase fluctuations can trigger the coagulation cascade and significantly change some of the plasma coagulation parameters.<sup>75</sup> It should be emphasized that activated partial thromboplastin time (aPTT) is shortened in the postprandial state and this is why monitoring of unfractionated heparin could be jeopardized if samples are taken in a nonfasting state. Considering the above, a period of fasting is required before hemostasis testing.

Finally, the human body experiences a mild postprandial metabolic alkalosis in response to a meal. This alkalois occurs due to the secretion of the hydrochloric acid in the parietal cells of the stomach, which is followed by extraction of chloride from the plasma and the release of bicarbonate into the plasma in order to maintain electrical neutrality. Thus venous blood leaving the stomach is enriched with bicarbonates, and this phenomenon is responsible for postprandial metabolic alkalosis (i.e., the alkaline tide) with concomitant increase of pCO<sub>2</sub> and a subsequent reduction of ionized calcium by 0.2 mg/dL (0.05 mmol/L).<sup>55</sup>

To avoid any misinterpretation due to the above-described effects, blood collection should preferably be done after an overnight (12 hours) fast (discussed in more details later in the section entitled Preparing for Blood Sampling).

### Effects of Fluid Intake Before Sampling

Whereas drinking coffee or small amounts of alcohol is largely seen as part of normal life and therefore not worth reporting to the physician, one should be aware of the influence of the intake of various fluids on the concentration of different analytes. Ingestion of various fluids may also exert acute and chronic effects.

**Caffeinated beverages.** Many beverages, such as tea, coffee, and cola drinks, contain caffeine. Caffeine stimulates the adrenal cortex and medulla, leading to the subsequent increase of the concentration of catecholamines and their

metabolites, as well as free cortisol, 11-hydroxycorticoids, and 5-hydroxyindoleacetic acid (5-HIAA) in serum. These hormonal changes are followed by the increase in plasma glucose concentration. Plasma renin activity may also be increased following caffeine ingestion.<sup>55,76</sup> Caffeine induces diuresis and inhibits the reabsorption of electrolytes, thus leading to a transient increase in their excretion and this effect is dose-dependent. Total urine output of water and electrolytes (calcium, magnesium, sodium, chloride, potassium) increases within 2 hours following caffeine ingestion, and caffeine-induced urinary loss of calcium and magnesium is therefore largely attributable to a reduction of the renal reabsorption of calcium and magnesium.<sup>77</sup> Caffeine also has a marked effect on lipid metabolism. Ingestion of coffee increases the rate of lipid catabolism, thus leading to an increase of plasma free fatty acids, glycerol, and lipoproteins.<sup>78,79</sup> Finally, caffeine is a strong stimulant of gastrin release and gastric acid secretion and also induces the secretion of pepsin.<sup>80</sup>

**Alcohol.** Alcohol consumption, depending on its duration and extent, may affect a number of analytes. Among alcohol-related changes, acute and chronic effects should be considered separately. The decrease of plasma glucose and increase of lactate are the acute effects that occur within 2 to 4 hours of ethanol consumption. Ethanol is metabolized to acetaldehyde and then to acetate. This increases hepatic formation of uric acid<sup>81</sup> and inhibits renal urea excretion, thus causing an increase of uric acid in plasma.<sup>82</sup> Together with lactate, acetate decreases plasma bicarbonate, resulting in mild to severe metabolic acidosis, depending on the amount of ingested alcohol.

Acute alcohol ingestion increases the activity of serum GGT and some other enzymes (e.g., isocitrate dehydrogenase, ornithine carbamoyltransferase).<sup>83</sup> Chronic effects of ethanol ingestion include the increase in serum triglyceride concentration due to decreased plasma triglyceride breakdown and an increase in the serum activity of many enzymes (GGT, AST, and ALT).

Moreover, chronic alcohol consumption affects pituitary and adrenal function and is associated with numerous biochemical abnormalities.<sup>84,85</sup> It affects lipid metabolism and inhibits the sialylation of transferrin that leads to increased serum concentration of carbohydrate-deficient forms of transferrins (CDT).<sup>86</sup> Increased mean corpuscular volume (MCV) is related to the direct toxic effect of alcohol on erythropoietic cells or a deficiency of folate.<sup>87</sup> Increased urine ethanol excretion leads to a decreased formation of vasopressin

with increasing diuresis. Enhanced diuresis is followed by increased secretion of renin and aldosterone.<sup>88</sup>

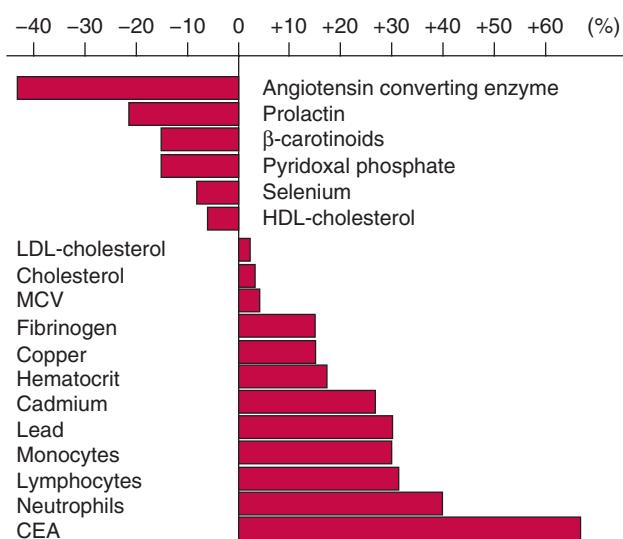
To assess the effect of alcoholic drinks on test results and to avoid misinterpretation of laboratory results, it is recommended that the history of alcohol intake (i.e., the ingested amount and frequency/time of ingestion) be documented in clinical records.

### Smoking Tobacco

Smoking tobacco leads to a number of acute and chronic changes in analyte concentrations, with the chronic changes being rather modest. Smoking increases the serum concentrations of fatty acids, epinephrine, free glycerol, aldosterone, and cortisol.<sup>26</sup> These changes occur within 1 hour of smoking a cigarette. Through adrenal gland stimulation, nicotine causes the increase of the concentration of epinephrine in the plasma and the urinary excretion of catecholamines and their metabolites.<sup>89</sup> Smoking leads to the acute increase in serum triglyceride, and total and LDL cholesterol concentrations.<sup>90</sup> Glucose metabolism is also dramatically affected by nicotine. Within only 10 minutes of smoking a single cigarette, glucose concentration increases by up to 10 mg/dL (0.56 mmol/L). This increase may persist for 1 hour.

Alterations in analytes induced by chronic smoking affect numerous blood components such as CBC, some enzymes, lipoproteins, carboxyhemoglobin, hormones, vitamins, tumor markers, and heavy metals (Fig. 5.2). These changes are induced by nicotine and its metabolites and reflect pathophysiological responses to toxic effects. To avoid a risk of misinterpretation of laboratory test results, smoking habits should be documented in clinical records.

In heavy smokers blood leukocyte count may be increased by as much as 30%, with a proportional increase of



**FIGURE 5.2** Chronic effects of smoking. Deviation (%) of blood analyte concentrations between current smokers and nonsmokers. *CEA*, Carcinoembryonic antigen; *HDL-cholesterol*, high density cholesterol; *LDL-cholesterol*, low density cholesterol; *MCV*, mean corpuscular volume. (Reproduced from Guder WG, Narayanan S, Wisser H, & Zawta B. *Diagnostic Samples: From the Patient to the Laboratory*. 4th updated ed. Weinheim: Wiley-Blackwell; 2009, with permission by Wiley-VCH-Verlag, Weinheim, Germany.)

the lymphocyte count.<sup>55</sup> For carcinoembryonic antigen (CEA), different reference limits should be applied for smokers and nonsmokers due to the large differences between the two groups. The higher concentration found in smokers is caused by an increased synthesis and secretion of CEA in the colon. Tobacco smokers have higher carboxyhemoglobin concentration. To compensate for the impaired capacity for oxygen transport in heavy smokers, there is also an increase in RBC count. Partial pressure of oxygen ( $pO_2$ ) is lower in tobacco smokers than in nonsmoking individuals by about 5 mm Hg (0.7 kPa).<sup>55</sup> Like caffeine, nicotine is also a very potent stimulant of the secretion of gastric juice and an inhibitor of duodenal bicarbonate secretion.<sup>91</sup> These effects may be observed within 1 hour of smoking several cigarettes. Smoking also affects the body's immune response and male fertility by affecting the sperm count, morphology, and motility.<sup>55,92</sup> The effect of smoking may persist even after smoking cessation. It usually takes 5 years, or even longer, for most parameters to normalize (e.g., C-reactive protein [CRP] and fibrinogen concentrations, hematocrit). Interestingly, for some parameters (e.g., white blood cell [WBC] count), it may take up to 20 years to return to baseline value.<sup>93</sup>

### Body Position and Tourniquet

Body posture influences blood constituent concentrations. This is caused by the net capillary filtration (i.e., the net result of the differences in the membrane permeability, hydrostatic pressure, colloid osmotic pressure of plasma, and interstitial fluid). Capillary filtration is especially increased in the lower extremities when changing from the supine to the upright position. The change in body posture from the supine to sitting and from sitting to the upright position leads to a significant decrease in plasma volume with a subsequent increase in the concentration of all constituents that usually do not pass the capillary filtration barrier (e.g., blood cells, large molecular weight molecules). Although this effect is observed in healthy and diseased individuals, the degree of the change is usually greater in some disease states—for example, in cardiac insufficiency.

Variations in the plasma volume subsequent to the change of the body position alter blood cell count (RBC, WBC, and platelets), concentrations of hemoglobin, and hematocrit; a short period of 10 minutes is usually enough for the vascular volumes to re-equilibrate and to adapt to the new posture.<sup>94-96</sup> It was also demonstrated that patient posture might have a significant impact on results of routine hemostasis testing, decreasing the prothrombin time (PT) values, and increasing the fibrinogen concentration when patient position is changed from supine to sitting.<sup>97</sup> Finally, net capillary filtration effect due to the change in body posture also affects small molecular weight molecules which are transported in blood bound to proteins. For example, while the concentration of free calcium is not affected, total calcium concentration increases by 5 to 10%<sup>98</sup> when changing from the supine to the upright position. To minimize the effect of this pre-analytical source of potential bias, reference intervals should ideally be obtained under identical conditions with regard to body posture. Blood sampling should be performed after at least 15 minutes of rest in a supine or sitting position.<sup>25</sup>

A similar mechanism occurs when a tourniquet is applied to facilitate finding appropriate veins for venipuncture. The higher pressure obtained in veins leads to the loss of water

and low molecular weight substances, increasing the concentration of proteins and analytes bound to them, cells, hemoglobin concentration, and hematocrit.<sup>99,100</sup> This becomes clinically significant after 1 to 2 minutes of tourniquet application.<sup>55</sup> Prolonged venous stasis can also cause a significant increase of fibrinogen and a shortening of APTT and PT.<sup>101</sup> Therefore the tourniquet should be released 1 minute after it has been applied.

### Muscular Activity

Physical activity of varying duration and intensity may lead to substantial changes in the plasma composition, and the extent of this change depends on several factors, such as training status, intake of fluid, electrolytes and carbohydrates, and even the ambient temperature.<sup>102,103</sup> For example, even a mild physical effort, like clenching the fist during venous blood sampling, can increase the concentration of potassium and should therefore be avoided.<sup>104</sup> This occurs due to the release of potassium from skeletal muscles and even without a tourniquet.

Intensive exercise is associated with transient increases in cardiac biomarkers, markers of muscle damage, platelet aggregation, tissue-plasminogen activator, activation of the fibrinolytic system, and a decrease in the ability of the blood to clot and generate thrombin, as well as with leukocytosis.<sup>105–108</sup> Cardiac troponin (cTn) rises after a maximal bicycle stress test.<sup>109</sup> The majority of changes are of transient nature and most of the parameters return to baseline within 3 hours after the exercise, although it was observed that some hematologic indices, such as red cell distribution width (RDW), continue to increase after the half-marathon run, reaching a peak 20 hours after the run.<sup>110</sup> Furthermore, it has been demonstrated that in individuals who are physically active more than 12 hours per week, concentrations of creatine kinase (CK), Creatine kinase MB (CK-MB), ALT, and LD are increased for a prolonged period of time.<sup>111</sup>

Due to such substantial changes in plasma composition, in professional athletes (e.g., marathon runners), a large proportion of laboratory results may fall outside the usual reference intervals.<sup>112</sup>

Intensive physical activity (within 12 hours before blood sampling) may also affect homeostasis for numerous hormones including catecholamines and their derivatives, epinephrine, norepinephrine, dopamine, corticotropin (ACTH) and vasopressin, gastrin, TSH, prolactin, growth hormone, aldosterone, cortisol, testosterone, human chorionic gonadotropin (hCG), insulin, glucagon, and β-endorphin.<sup>113–116</sup>

### Preparing for Blood Sampling

Because food, fasting time, circadian rhythm, muscular activity, smoking, drugs, and ethanol consumption can affect the concentration of numerous analytes, standardization of all those controllable variables is highly recommended. Proper standardization of controllable variables leads to significant reduction of preanalytical variability. In the past, there has been a great heterogeneity in the definition of *fasting state* used for different analytes by different health care facilities and in the literature. To facilitate the agreement on the definition of *fasting state* and encourage uniform and consistent compliance the European Federation for Clinical Chemistry and Laboratory Medicine (EFLM) Working Group for Preanalytical Phase WG-PRE has published a recommendation for the definition

of fasting requirements as a guiding framework for harmonization of this important preanalytical aspect.<sup>117</sup>

According to these recommendations, the following general requirements should be applied to all blood tests:

1. Blood should be drawn preferably in the morning between 7 am and 9 am
2. Fasting should last for 12 hours, during which only water consumption is permitted.
3. Alcohol should be avoided for 24 hours before blood sampling.
4. In the morning before blood sampling, patients should refrain from cigarette smoking and caffeine-containing drinks (tea, coffee, etc.).

Professional associations and laboratories worldwide are encouraged to adopt, implement, and disseminate the EFLM WG-PRE recommendation for the definition of *fasting*. Moreover, laboratories worldwide should have policies for sample acceptance criteria related to fasting samples. Blood samples for routine testing should not be taken if a patient has not been appropriately prepared for sample collection.

### Noncontrollable Variables

Various unavoidable biological factors can lead to changes in analyte concentration and can therefore only be considered during interpretation with the respective knowledge. Table 5.5 summarizes some of these factors and their respective effects. These factors should be considered when interpreting laboratory results because their influence cannot be prevented by preanalytical standardization.

### Age and Gender

Due to dramatic physiologic changes associated with growth and development, the reference intervals for many analytes differ substantially with respect to an individual's age and gender (see Chapter 9 and the Appendix). In newborn subjects, the body fluids reflect the trauma of birth and early postnatal events related to the adaptation of the baby to new extrauterine life. Immediately after birth, infants usually experience a mild metabolic acidosis of transient nature, due to the accumulation of lactates. This acid-base disturbance is usually normalized within the first day after birth.<sup>55</sup> The CALIPER study is an excellent source of reference intervals in childhood (see the Appendix).<sup>126</sup> In the early hours of extrauterine life, the concentration of some biochemical markers (AST, direct bilirubin, total bilirubin, creatinine, CRP, GGT, immunoglobulin G [IgG], LD, magnesium, phosphate, rheumatoid factor, uric acid) is increased, thus reflecting the maternal concentrations, but it then declines within the first 2 weeks of life.<sup>127</sup> Concentrations of other markers (e.g., amylase, transferrin, antistreptolysin O [ASO], cholesterol, IgA, IgM) are very low in the neonatal period and gradually increase within the first 2 weeks of extrauterine life. This upward trend in analyte concentrations continues over time from birth to 18 years. Most of the biochemistry parameters (albumin, ALP, AST, total bilirubin, creatinine, IgM, iron, lipase, transferrin, HDL cholesterol, and uric acid) exert differences between genders during the early childhood years. However, these changes are most significant during puberty (age 14 to 18 years), due to the strong influence of sexual development and growth.<sup>127</sup>

Hemoglobin concentration, hematocrit, and the other RBC indices follow a similar pattern, showing the gradual

**TABLE 5.5 Unavoidable Influences on Laboratory Results**

Influence (Reference)	Examples of Analyte Concentrations Changed	Remarks
Age <sup>118–120</sup>	ALP, LDL cholesterol, hormones, creatinine, total WBC count, WBC subpopulations, RBC, hemoglobin, hematocrit, RBC indices, VWF, AT, PC, PS, plasminogen.	Provide age-dependent reference intervals
Race <sup>119–123</sup>	CK higher in black than in white males. Creatinine higher in black than in white males. Granulocytes higher in white than in black males. Hematocrit, hemoglobin, and MCV lower in African Americans than Caucasians. Hematocrit, hemoglobin, MCH, MCHC, and MPV lower in Asians than Caucasians.	Provide race-specific reference intervals
Gender <sup>118,124,120</sup>	ALT, γ-GT, creatinine, hemoglobin, hematocrit, RBC, WBC, PLT	Provide gender-specific reference intervals
Pregnancy <sup>26,64,125</sup>	Triglycerides ↑, homocysteine ↓, WBC ↑, d-dimers ↑, PT ↑, fibrinogen ↑	Document months of pregnancy with laboratory results
Altitude <sup>64,120</sup>	CRP, hemoglobin ↑, hematocrit ↑, RBC ↑, transferrin ↓	Consider weeks of adaptation, when coming from or going to high altitude

ALP, Alkaline phosphatase; ALT, alanine aminotransferase; AT, antithrombin; CK, creatine kinase; CRP, C-reactive protein; γ-GT, gamma-glutamyltransferase; LDL, low density lipoprotein; MCH, mean corpuscular hemoglobin; MCHC, mean corpuscular hemoglobin concentration; MCV, mean corpuscular volume; MPV, mean platelet volume; PC, protein C; PLT, platelets; PS, protein S; PT, prothrombin time; RBC, red blood cell; VWF, von Willebrand Factor; WBC, white blood cell.

increase during the first 10 years of life. First gender differences are observed at the age of 10 years, when values in boys show a sharp increase during puberty and adolescence. Concentrations in females are much lower, but they also slowly increase throughout puberty. Such gender differences are related to the lower metabolic demand, decreased muscle mass, and lower iron stores in females.<sup>128</sup>

Concentration of thrombopoietin peaks shortly after birth and then slowly decreases. Subsequent to the change of thrombopoietin concentration, immediately after birth there is a peak in platelet count, followed by a decline during childhood and into adulthood. The WBC count is also higher in the early extrauterine days and throughout the first couple of years of childhood; values decline in older children. Females have slightly higher platelet count than males during adolescence and adulthood.<sup>128</sup>

Although bone marrow cellularity decreases with age, in the absence of disease WBC, hemoglobin, platelets, and differential are maintained within adult reference intervals in individuals older than 65 years.<sup>120,129–131</sup>

Hemostasis develops during fetal development and changes with gestational age. In neonates, the concentrations of the proteins of the prothrombin and contact factor groups are lower than in adults, due to liver immaturity, and reach adult values only after 6 months of age.<sup>120</sup>

## INTERFERENCE FACTORS

As mentioned earlier, interference factors have the ability to interfere with the analytical procedure and alter the test results. The effect of interference factors depends on the method—that is, the same interferent may not necessarily affect two different methods used to measure the same analyte. Common interference factors are hemolysis, lipemia, icterus, drugs, paraproteins, and various sample contaminants such as gels, tube additives, and fibrin clots.

Interfering factors are considered clinically relevant when the bias caused by their interference is greater than the maximum allowable deviation of a measurement procedure. How this “maximum allowable deviation” should be established is still debated. The Clinical Laboratory Standards Institute (CLSI) EP7-A2 guideline, for example, sets this criterion at ±10% as a rule of thumb. Others would argue that the degree of allowable deviation caused by interfering factors should be derived (I) from data on the biological variation of the analyte, (II) by simulation modeling based on the effect of preanalytical and analytical performance on clinical decisions or patient outcomes, or (III) from information on the state-of-the-art.<sup>132</sup> The choice of the method for determining the maximum allowable deviation for a certain analyte not only depends on the medical use of the test but also on the national and international regulations in use.

Interferences can be endogenous and exogenous. Endogenous interferences originate from the substances present in

## POINTS TO REMEMBER

### Influencing Factors

- Samples should be taken before any therapeutic and diagnostic procedures that have a potential influencing effect.
- Tobacco smoking leads to several acute and chronic changes in the concentrations of numerous analytes.
- Even within only 1 hour of smoking one to five cigarettes, there is an increase in serum concentration of fatty acids, epinephrine, free glycerol, aldosterone, and cortisol.
- Diet substantially affects the composition of plasma. The effects of diet can be long term and acute.
- Physical activity of varying duration and intensity leads to changes in the plasma composition of many analytes. The extent of this change depends on training status, intake of liquid, electrolytes and carbohydrates, and even the ambient temperature.
- Most of the reference interval data for children are obtained from the CALIPER study.

the patient sample, whereas exogenous interferences relate to the effect of various substances added to the patient sample, such as separator gels, anticoagulants, surfactants, and so on, all of which may cause significant interference.<sup>133,134</sup>

## Hemolysis

### Definition and Background

Hemolysis is defined as a process of membrane disruption of erythrocytes and other blood cells, accompanied by the subsequent release of cell components into the plasma and red coloration of the serum (or plasma) to various degrees after centrifugation.<sup>135,136</sup> Though hemoglobin is the most abundant protein in RBC, hemolysis is not necessarily always associated with the release of hemoglobin into the surrounding extracellular fluid. For example, if the blood sample is stored at a low temperature, low molecular intracellular components like electrolytes diffuse from the cells, but hemoglobin will not. Furthermore, efflux of cell components due to cell lysis affects all blood cells (i.e., platelets and WBC) and not only erythrocytes. Therefore it is important to remember that red coloration of the serum or plasma can never accurately predict the concentration of blood cell components.

Hemolysis is the most common preanalytical error and the most common cause of sample rejection. It occurs with a frequency of up to 30%<sup>137,138</sup> and accounts for almost 60% of unsuitable specimens.<sup>139</sup> The frequency of hemolysis largely depends on the collection facility, characteristics of the patient population, and the type of professional who is doing the phlebotomy. The highest frequency of hemolysis has been observed in samples from emergency departments, pediatric departments, and intensive care units, whereas hemolysis has proven to be the least frequent in outpatient phlebotomy centers, where blood sampling is done by specialized laboratory staff.<sup>140,141</sup> These differences are due to the level of knowledge and skills of the staff who perform the blood collection.<sup>24</sup> One large study in Australia of five hospitals from October 2009 to September 2013 found that the hemolysis rate is much higher in emergency departments (up to 8.73%, depending on the triage category) than in other inpatient settings (<4%). Interestingly, the hemolysis rate was highest in patients who were triaged in the most urgent category. Also, the hemolysis rate was higher if the phlebotomy was done by the clinical staff than by laboratory phlebotomists.<sup>142</sup>

The two major sources of hemolysis are in vivo hemolysis and in vitro hemolysis. In vivo hemolysis is a result of a pathologic condition and occurs within the body before the blood has been drawn. It may occur as a result of numerous biochemical (enzyme deficiencies, erythrocyte membrane defects, hemoglobinopathies), physical (prolonged marching, drumming, prosthetic heart valves), chemical (ethanol, drug overdose, toxins, snake venom), or immunologic (autoantibodies) mechanisms, and infections (babesiosis, malaria). In vivo hemolysis can further be categorized as intravascular and extravascular, depending on the site of the destruction of RBC. Intravascular hemolysis occurs as a direct and immediate disruption of RBC due to the cell injury within the vasculature, whereas in extravascular hemolysis, RBC membranes are damaged by the reticuloendothelial system, primarily in the spleen.<sup>143</sup> The most common causes of in vivo hemolysis are reaction to incompatible transfusion and autoimmune hemolytic anemia.

In vivo hemolysis is not very common and accounts for only 3% of all hemolyzed samples.<sup>144</sup> Nevertheless, in vivo hemolysis is of great clinical importance because it reflects an underlying pathologic process in a patient. Laboratories should therefore have a procedure in place for distinguishing in vivo and in vitro hemolysis. In vivo hemolysis should always be suspected when patient blood is hemolyzed over a longer period after different types of samples (e.g., citrate, serum, and heparinized tube) are hemolyzed or repeated blood sampling, even after special care has been taken to avoid hemolysis.

Common findings associated with in vivo hemolysis which may help in distinguishing in vivo from in vitro hemolysis:

- dark brown serum/plasma and urine,
- ↓↓ serum/plasma haptoglobin,
- hemoglobinuria (free hemoglobin in urine) and methemoglobinuria,
- ↑ indirect bilirubin concentration in serum/plasma,
- ↑ reticulocyte count (compensatory bone marrow response),
- normal potassium concentration in serum/plasma,
- ↑↑ LDH in serum/plasma,

Decreased concentrations of haptoglobin in serum and free hemoglobin in urine are the most pronounced and specific laboratory signs of in vivo hemolysis. Haptoglobin is a protein that binds free hemoglobin in the circulation to prevent oxidative damage induced by hemoglobin.<sup>145</sup> Once released from the erythrocyte into the plasma, hemoglobin forms complexes with haptoglobin, and those complexes are removed from the circulation by macrophages. In more pronounced cases of in vivo hemolysis, haptoglobin in serum can be undetectable (i.e., below the detection range), whereas its concentration in cases of in vitro hemolysis remains unchanged.<sup>146,147</sup> When in vivo hemolysis is confirmed, the laboratory should not reject hemolyzed samples for analysis, because parameters in hemolyzed samples reflect the actual patient condition and are extremely relevant for adequate patient care (diagnosis, therapy management, monitoring).

In vitro hemolysis occurs outside the patient at many steps of the preanalytical phase: blood sampling, sample handling and delivery to the laboratory, and sample storage. Causes of in vitro hemolysis are described in Chapter 4.

### Mechanisms of Hemolysis Interference

Hemolysis is an endogenous interference that causes clinically relevant bias of patient results through the several distinct mechanisms described in the following.

**Spectrophotometric interference.** Spectrophotometric interference of hemolysis occurs due to the ability of hemoglobin to absorb light at 415-, 540-, and 570-nm wavelengths.<sup>148</sup> This characteristic of hemoglobin causes optical interference that can lead to either falsely increased or decreased concentrations of the measured parameters. The direction and degree of the interference largely depend on the analyte and the method.

**Release of the cell components into the sample.** Some components are present in blood cells in concentrations that are several times higher than those in the extracellular space (i.e., plasma or serum). Table 5.6 shows some of the most pronounced differences between intracellular and extracellular concentration in RBC.<sup>149–152</sup>

From this it follows that there is a dramatic increase in the concentration of the listed analytes measured in hemolyzed

**TABLE 5.6 Ratio Between Intracellular and Extracellular Concentration of Various Analytes in Red Blood Cells**

Analyte	Intracellular Concentration (Compared to Extracellular)
Lactate dehydrogenase	↑ 160×
Inorganic phosphate	↑ 100×
Potassium	↑ 40×
Aspartate aminotransferase	↑ 40×
Folic acid	↑ 30×
Alanine aminotransferase	↑ 7×
Magnesium	↑ 3×

plasma (or serum) due to the efflux of those substances from erythrocytes into the sample. The most pronounced effect of hemolysis is seen for LD. LD activity may be increased by over 20% in mildly hemolyzed samples (at a concentration of only 0.27 g/L of free hemoglobin), by over 60% at 0.75 g/L of free hemoglobin, and up to over 350% in grossly hemolyzed samples with 3.34 g/L of free hemoglobin.<sup>153</sup>

Because intracellular components may also escape from platelets during clotting, there is a marked difference in the potassium concentration between serum and plasma. The mean estimated difference in the concentration of potassium in serum and plasma is  $0.36 \pm 0.18$  mmol/L, and this difference is positively associated with the platelet count.<sup>154</sup> Plasma is therefore the recommended sample type for the accurate measurement of potassium.

**Sample dilution.** Some analytes are present in much higher concentrations in plasma than in blood cells like albumin, bilirubin, glucose, sodium, and a few others.<sup>150</sup> For those parameters, hemolysis will cause a dilution effect, and their concentrations will be lower in hemolyzed samples. The effect of sample dilution causes clinically significant bias only at higher degrees of hemolysis. For example, glucose is negatively affected by severe hemolysis ( $-8.3\%$ ) only at the concentration of 3.34 g/L of free hemoglobin if measured by the Beckman Coulter chemistry analyzer and reagents (Olympus AU2700, Beckman Coulter, O'Callaghan's Mills, County Clare, Ireland).<sup>155</sup>

**Chemical interference.** Various blood cell components may affect the analyte measurement procedure by directly or indirectly modifying the analyte (Table 5.7).

An example of the direct interference through competition is the effect caused by the enzyme adenylate kinase, which is present in both erythrocytes and platelets.<sup>155</sup> Adenylate kinase (EC 2.7.4.3) is an enzyme that catalyzes the reversible conversion of ATP and AMP to two ADP molecules and maintains the adenine nucleotide cell content.<sup>156</sup> When released from the cells during hemolysis, adenylate kinase may compete for ADP with CK in a CK assay if inhibitors are not supplied in the reaction mixture.<sup>157</sup>

Hemoglobin released from erythrocytes during hemolysis may interfere with various assays through its pseudo-peroxidase activity. Pseudo-peroxidase activity of free hemoglobin released from erythrocytes interferes in the assay for measurement of bilirubin concentration through the inhibition of the formation of diazonium salt.<sup>158</sup>

**TABLE 5.7 Chemical Mechanisms of Hemolysis Interference**

Direct mechanism	Indirect mechanism
<ul style="list-style-type: none"> <li>• Competition for a substrate or any other component of the assay reaction mixture (e.g., creatine kinase assay)</li> <li>• Inhibition of the assay (e.g., pseudo-peroxidase activity of free hemoglobin)</li> </ul>	<ul style="list-style-type: none"> <li>• Complexation</li> <li>• Proteolysis (cathepsin E)</li> <li>• Precipitation of the analyte or any other component of the assay reaction mixture</li> </ul>

Hemolysis may cause a clinically significant interference on a wide range of analytes in immunochemistry assays. This interference is caused by modifying the reaction analytes (antigens and antibodies) by the proteolytic action of cathepsin E, the major proteolytic enzyme in mature erythrocytes. Proteolytic enzymes released from erythrocytes may mask or potentially enhance epitope recognition in various immunoassays. Interference caused by proteolytic activity may cause measurement bias of various degrees and various directions, depending on the assay. For example, current cTn assays have variable susceptibility to hemolysis interference.

Hemolysis has been shown to cause negative interference with concentrations of cTnT, insulin, cortisol, testosterone, and vitamin B12, and false-positive increases for prostate-specific antigen (PSA) and cTnI in a concentration-dependent manner.<sup>159–161</sup> However, the degree and direction of bias are analyte and method dependent. For example, hemolysis causes falsely decreased concentrations of cTnT assayed with the Roche hs cTnT assay on the Elecsys E170 immunochemistry analyzer, whereas concentrations of cTnI measured using the Ortho Clinical Diagnostics TnI ES assay on the Vitros 5600 Integrated System (Ortho Clinical Diagnostics, Rochester, NY) are falsely increased in hemolyzed samples.<sup>162</sup> Abbott Architect TnI assay appears to be more robust against interference from hemolysis.<sup>163</sup> The microparticle enzyme immunoassay for cTnI (Abbott Laboratories, Abbott Park, IL) is not affected by moderate hemolysis and exerts clinically relevant bias only for grossly hemolyzed samples.<sup>164</sup>

### Lipemia

Lipemia is defined as a turbidity of the sample visible to the naked eye. Turbidity of the sample is caused by the light scattering due to the presence of large lipoprotein particles (chylomicrons). The increase in concentration of lipoproteins in blood most commonly occurs due to postprandial triglyceride increase, parenteral lipid infusions, or some lipid disorders. Not all lipoproteins have equal contribution to the sample turbidity. The effect of lipoprotein particles on the sample turbidity depends on the size of the particles. Chylomicrons and very low-density lipoproteins (VLDL), the largest lipoprotein particles in the circulation, have the greatest contribution to the sample turbidity. To avoid postprandial lipemia, patients are therefore requested to fast for 12 hours before the blood sampling.<sup>117</sup>

### Mechanisms of Interference Caused by Lipemia

Lipemia is an important endogenous interference that may cause clinically relevant bias of patient results through the several mechanisms described below.

**Spectrophotometric interference.** Lipemia causes interference by light absorbance and light scattering. The lipemic sample absorbs light, causing a decrease in the intensity of the light passing through the sample. The ability of lipoprotein particles to absorb light is manifested in the range of wavelengths (300 to 700 nm). Sample absorbance rises with the decreasing wavelengths and is maximal in the ultraviolet range. That is why many enzymatic methods in which the end product is measured at 340 nm (NAD[P] or NADP[H]) are strongly affected by lipemia.

Lipemic samples also cause light scattering. Light scattering occurs in all directions, and its intensity depends on the number and size of lipoprotein particles and the wavelength of measurement.<sup>165</sup> For this reason, light scattering of lipoprotein particles causes significant interference with turbidimetry and nephelometry. In methods where the transmittance of light is inversely proportional to the concentration of the analyte, in the absence of the sample blank, sample turbidity causes positive bias. However, in some competitive assays where the transmittance of light is directly proportional to the concentration of the analyte, sample turbidity will cause negative bias.

**Interference caused by the volume depletion effect.** Plasma in healthy individuals in the fasting state consists of only minor portion of lipids (<10% of the total plasma volume). The rest of the plasma is water. The increase in the concentration of lipoprotein particles leads to an increase in the plasma volume occupied by lipids. Particles that are not lipid soluble are displaced by the lipids to the water part of the plasma. Therefore lipemia leads to a false decrease in the concentration of the measured analyte in all methods in which the concentration of respective analyte is measured in the total plasma volume.

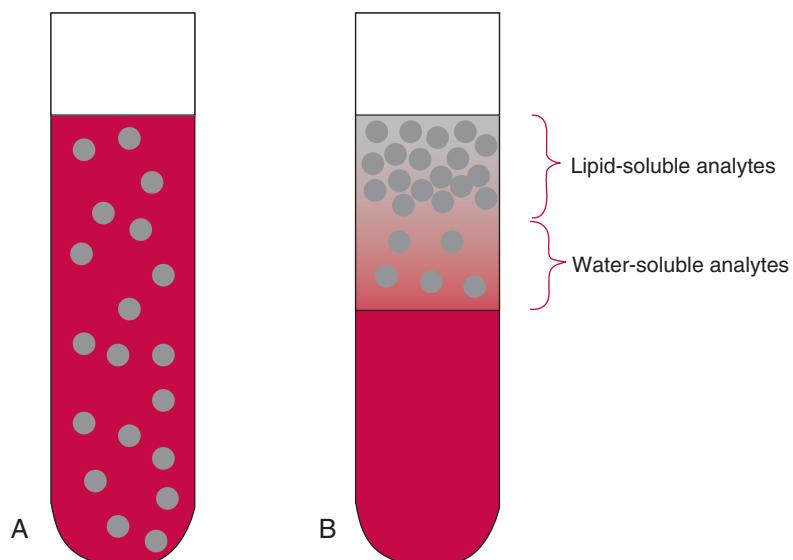
One example of interference caused by the volume depletion effect is the bias in electrolyte measurement, leading to so-called pseudo-hyponatremia. This type of interference affects electrolytes only if measured by flame photometry and by indirect measurement using ion-selective electrodes

(ISEs) but not in direct potentiometry (for more details, see Chapters 17 and 37). However, it must be noted that the volume displacement effect of the lipemic sample will affect the electrolyte measurement only in grossly lipemic samples with concentrations of triglycerides greater than 17 mmol/L (1504 mg/dL).<sup>166</sup>

**Interference caused by partitioning of the sample.** Upon centrifugation of a lipemic sample, lipoproteins are not homogeneously distributed in the serum or plasma due to the lipid gradient (Fig. 5.3). Water-soluble analytes are more concentrated in the lower layer of the plasma or serum, whereas lipids and lipid-soluble analytes, such as drugs and some lipid-soluble hormones, are more concentrated in the top lipid-rich layer. This is especially important in automated chemistry analyzers with fixed path lengths of the sample probe. Test results may differ for those analytes that are not evenly distributed between the lipid and water portion of the sample, depending on the part of the sample from which the sample probe is taking the sample for analysis.

**Interference caused by physicochemical mechanisms.** An excess of lipoproteins in the blood may interfere in electrophoretic and chromatographic methods by causing abnormal peaks. Increased concentrations of triglycerides and lipoprotein particles may disturb the electrophoretic pattern and morphology, as well as falsely increase the relative percentage of the prealbumin, albumin, and  $\alpha_1$ - and  $\alpha_2$ -globulin regions.<sup>167,168</sup> Moreover, lipemia may even affect some immunochemistry assays by masking the binding sites on antigens and antibodies and thus physically interfering with antigen–antibody binding.<sup>169</sup>

One additional complication of excessive lipemia is the increased sample susceptibility to hemolysis leading to the specific turbid and reddish appearance of the sample (the so-called “strawberry milk” appearance). This effect is most probably caused by the increased fragility of the erythrocyte membranes due to the alterations in the content of the phospholipid membrane layer and is more pronounced with the increase in lipid (particularly triglycerides) concentrations.<sup>170</sup>



**FIGURE 5.3** (A) Lipids are distributed evenly in whole blood prior to centrifugation. (B) Lipid gradient in centrifuged sample. Top lipid-rich layer contains lipid-soluble analytes. Lower plasma layer contains water-soluble analytes.

### Removal of Lipids From the Sample

In the hospital environment, lipemic samples are not infrequent. They most often originate from emergency departments, intensive care units, and endocrinology and gastrointestinal clinics from patients suffering from conditions that include acute pancreatitis, acute or chronic kidney failure, thyroid or lipid disorders, and diabetes mellitus. Lipemic samples quite commonly require immediate results. Unlike hemolysis, the interference caused by lipemia can be fully eliminated, or at least reduced, by removing the excess of lipids from the sample. Still, even if lipids have been successfully removed from the sample, any visible turbidity of a sample should be documented and reported with the test results because it offers clinically useful information about the patient.<sup>166</sup> Moreover, lipid testing and testing for lipid-soluble drugs (e.g., benzodiazepines) and hormones (e.g., thyroid hormones) should always be done on the native sample, before delipidation. Methods for lipid removal include ultracentrifugation, high-speed centrifugation, and some lipid-clearing agents.

#### Lipid removal by ultracentrifugation and high-speed centrifugation.

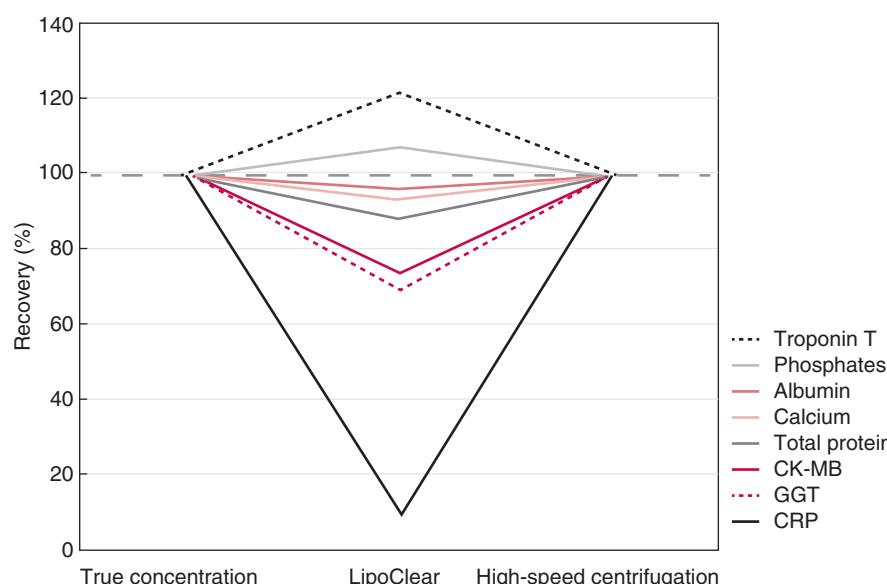
According to the CLSI C56-A standard for Hemolysis, Icterus, and Lipemia/Turbidity Indices as Indicators of Interference in Clinical Laboratory Analysis, ultracentrifugation is the recommended method for the removal of the excess of lipids in the sample.<sup>171</sup> Ultracentrifuges use the centrifugation force of almost 200,000 g and are very effective in clearing lipemic sera by separating lipids, especially chylomicrons (top layer) from the aqueous part (lower layer) of the sample. After centrifugation, the infranatant (lower part of the sample) can be transferred into the clean tube and analyzed. It should be kept in mind that by removing the upper lipid layer, one also removes lipid-soluble analytes like drugs and hormones. Results reported from ultracentrifuged samples or samples from which lipids have been removed in any other

way should be appropriately annotated to ensure clinicians are aware that the sample has been manipulated to obtain the reported results.

Though considered a gold standard, ultracentrifugation is not widely available in many laboratories. High-speed centrifugation using the microcentrifuge with a maximum centrifugation speed of up to 20,000 g may therefore serve as an acceptable alternative and is the method of choice for most laboratories.<sup>172</sup> The effectiveness of high-speed centrifugation depends on the concentration of lipids in the lipemic sample. However, it must be emphasized that ultracentrifugation is superior to high-speed centrifugation for grossly lipemic samples. By using the ultracentrifuge, triglyceride concentration may be reduced 7-fold (from 59.2 to 8.1 mmol/L; or 5239 to 717 mg/dL), whereas the high-speed centrifuge may achieve only 3.4-fold reduction.<sup>173</sup>

**Lipid removal by lipid-clearing agents.** Lipid-clearing agents are widely used in many laboratories due to their low cost, convenience, and ease of use. Those agents (cyclodextrin, polyethylene glycol, dextran sulphate, hexane, and others) may vary in their ability to extract lipids from a lipemic sample and may also lead to reduction of a significant amount of protein from the sample.<sup>174</sup> It is therefore extremely important for laboratories to verify the performance of such reagents before their routine use because they may not be appropriate for a wide range of analytes due to their low recovery. For example, LipoClear spin columns (Iris International Inc., Westwood, MA) may lead to serious underestimation of CRP (−92%), CK-MB (−25%), and GGT (−30%) and overestimation of cTnT (+20%) and phosphates (+7%) (Fig. 5.4).<sup>175</sup>

Lipid removal takes time and may cause delays in reporting the results. It is up to each individual laboratory to establish its own procedure for managing lipemic samples, bearing in mind to ensure the highest possible accuracy of results and



**FIGURE 5.4** Recoveries for several chemistry assays after lipid removal from a lipemic sample with a lipid clearing agent (LipoClear, Iris International Inc., Westwood, MA) and high-speed centrifugation in Eppendorf Mini Spin centrifuge (Eppendorf, Hamburg, Germany) at 12,100 g for 5 minutes.

CK-MB,

Creatine kinase MB; CRP, C-reactive protein; GGT, gamma-glutamyl transferase. (Data from Saracevic A, Nikolac N, Simundic AM. The evaluation and comparison of consecutive high-speed centrifugation and LipoClear reagent for lipemia removal. *Clin Biochem* 2014;47:309–14.)

speed. To minimize the prolongation of the turnaround time and subsequent delays in reporting the results for grossly lipemic samples, laboratories may consider analyzing electrolytes using the blood gas analyzers (direct ion selective electrode methodology) while manipulating the rest of the sample to remove the lipids.

### Intravenous Lipid Emulsion Therapy as an Antidote to Drug Overdose

Lipid emulsions were introduced in 2006 as a remedy for systemic toxic effects caused by local anesthetic and are used increasingly today in emergency settings to treat patients who have overdosed on antiepileptic, cardiovascular, or psychotropic drugs.<sup>176,177</sup> Their use is recommended in patients suffering from severe systemic cardiovascular toxic effects who have not otherwise responded to conventional resuscitation protocol and antidotal therapies.<sup>176</sup> The American College of Medical Toxicology recommends it as a reasonable therapeutic option in circumstances where there is serious hemodynamic or other instability from a lipid-soluble drug, even if the patient is not in cardiac arrest.<sup>178</sup> The exact mechanism of action of lipid emulsions is not known. In patients treated with large doses of lipid emulsions, possible side effects include severe hypertriglyceridemia, pancreatitis, lipemia, and numerous interferences in laboratory assays.<sup>179</sup> To avoid compromising patient outcome caused by reporting of incorrect results and delays in reporting of critical results, it is important that blood samples in such cases are collected prior to initiating the intravenous lipid emulsion therapy whenever possible.<sup>180</sup> If intravenous lipid emulsion therapy has already been initiated before the blood sampling, the laboratory should make an effort to remove the lipids and ensure acceptable accuracy of results and turnaround time. Good communication between the laboratory and the clinical staff in such cases of life-threatening toxicity from lipophilic drugs is of paramount importance.

### Icterus

The normal concentration of bilirubin in human plasma (or serum) is up to 20 µmol/L (1.2 mg/dL). Change in the color of the serum (or plasma) becomes detectable when bilirubin concentration exceeds 34 µmol/L (2 mg/dL). Bilirubin concentrations above 100 µmol/L (5.9 mg/dL) are clinically defined as *icterus*. Icteric plasma is commonly seen in patients from intensive care units, gastroenterology centers, and pediatric clinics. Bilirubin interferes with numerous chemistry tests such as enzymes (ALT, ALK, CK, lipase), electrolytes, metabolites (urea, creatinine, glucose), lipids (cholesterol, triglycerides), proteins (albumin, total proteins, IgG), hormones (estradiol, beta-HCG, free triiodothyronine [FT3]), and even some drugs (gentamicin, phenobarbital, theophylline, tobramycin).<sup>181–183</sup>

Just as with hemolysis and lipemia, interference caused by bilirubin differs among instruments and assays. For example, bilirubin exerts interference of different magnitudes (strong, moderate, or negligible) and directions (positive and negative interferences), both with Jaffe and enzymatic methods for the measurement of serum creatinine from different manufacturers. While some methods are not affected by bilirubin at all, others may exhibit strong interference by bilirubin, causing clinically significant bias for creatinine measurement and compromising the adequate management of patients with

kidney disease.<sup>184,185</sup> It has been demonstrated that even if two methods have identical reagents, differences in their susceptibility to interference by bilirubin may still occur. These differences may be due to the different incubation times and temperatures and some other parameters related to the assay setup.<sup>186</sup> Interestingly, although enzymatic methods are often considered the method of choice due to being less susceptible to various interferences, bilirubin has been reported to cause greater interference in some enzymatic creatinine assays than in Jaffe methods.<sup>187</sup>

Bilirubin is present in the blood in several distinct forms: as unconjugated and conjugated (mono- and diglucuronide conjugates). Unconjugated bilirubin is not soluble in water and is therefore transported in blood bound to albumin. Bilirubin conjugates are soluble in water. Additionally, bilirubin photoisomers may be found in the blood of neonates.<sup>188</sup> All these different molecular forms of bilirubin have different physical and chemical properties and behave differently in different chemistry assays. The total amount of measured bilirubin in the patient is a mixture of these different forms. Different forms of bilirubin cause interference to various degrees with different laboratory methods, and the same forms of bilirubin can act differently with the same assays on different instruments.

Most interference studies performed by manufacturers and most original studies published by different authors were done on commercially available forms of unconjugated bilirubin that may not correspond to what is found in the blood. This is why sometimes data obtained by interference studies does not mimic real scenarios in human blood and cannot be extrapolated to define rules for adequate detection and management of icteric samples.

Unfortunately, laboratories cannot do much about removing or minimizing the effect of icteric interference. Bilirubin oxidase and blanking procedures have been recommended.<sup>19</sup> Possible options are dilution of the sample (possible only for analytes present at high enough concentrations in the blood) and testing the requested analytes with a different method or on a different instrument for which icterus does not cause clinically significant interference. For maximal patient benefit, laboratories may consider having special protocols (dilutions or different methods) for some critical analytes in icteric samples to avoid unnecessary sample rejections.

### Mechanisms of Interference Caused by Icterus

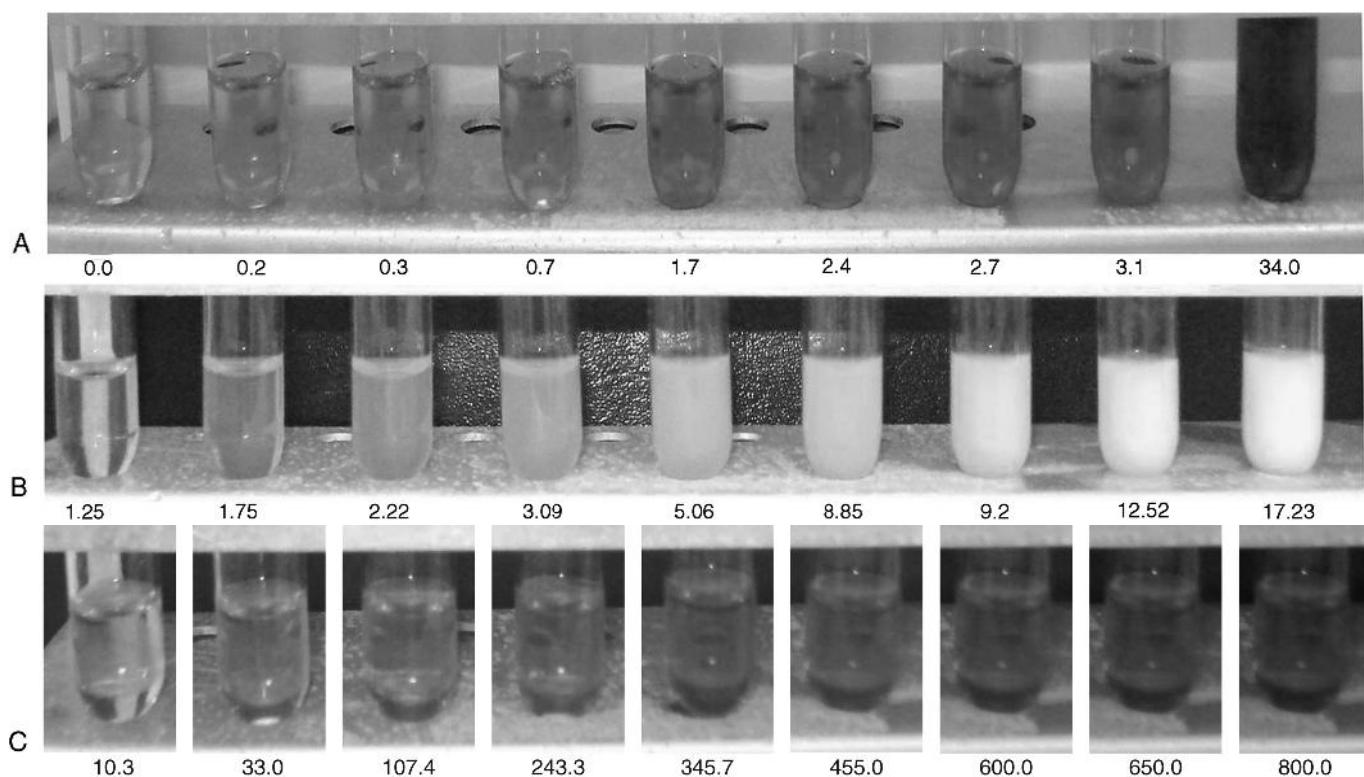
Icterus interferes through two mechanisms: spectrophotometric interference and by interfering with chemical reaction. It is important to recognize that both mechanisms may occur simultaneously in one sample.

**Spectrophotometric interference of bilirubin.** Bilirubin causes spectrophotometric interference due to its ability to absorb light in the wide range of wavelengths between 400 and 540 nm.

**Chemical interference of bilirubin.** Bilirubin produces negative bias on assays that involve H<sub>2</sub>O<sub>2</sub> as an intermediate reaction (e.g., cholesterol, glucose, uric acid, triglycerides).

### Detection of Hemolytic, Icteric, and Lipemic Samples

Hemolysis becomes visible at the concentration of 0.3 to 0.5 g/L of free hemoglobin, and the intensity of the red color of the serum or plasma further increases with the increase in



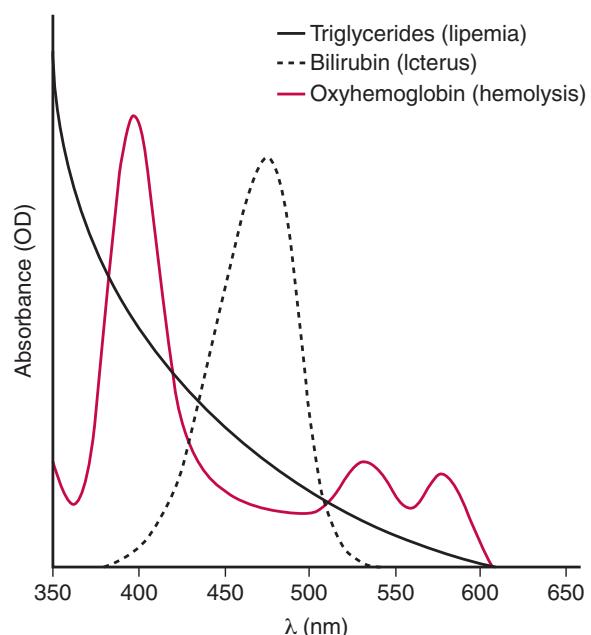
**FIGURE 5.5** (A) Hemolysis: the intensity of the red color of the serum and corresponding concentrations of free serum hemoglobin (in g/L). (B) Lipemia: the degree of turbidity and corresponding concentrations (in mmol/L) of triglycerides. (C) Icterus: the intensity of the yellow color of the serum and corresponding concentrations of bilirubin (in  $\mu\text{mol}/\text{L}$ ). (Color standard scales provided by Clinical Institute of Chemistry, University Hospital Center "Sestre milosrdnice," Zagreb, Croatia. Please see the online version of this figure for full color.)

concentration of free serum hemoglobin (Fig. 5.5A). Lipemia causes sample turbidity, which approximately corresponds to the concentration of serum triglycerides (Fig. 5.5B). Increased concentrations of serum bilirubin lead to yellow to orange coloration of the serum, and the change of the color correlates to the increasing concentration of the bilirubin in the serum (Fig. 5.5C).

Free hemoglobin, triglycerides, and bilirubin have characteristic absorption peaks in a wide wavelength range of 300 to 600 nm. This is also the range where sample absorbance is measured in spectrophotometric methods, and that is why hemolysis, icterus, and lipemia cause spectral interferences. Fig. 5.6 presents the characteristic absorption curves of oxyhemoglobin, triglycerides, and bilirubin. Serum indices may be detected by visual inspection and by the use of automated detection systems.

### Visual Detection of Serum Indices

Although detection of the degree of hemolysis, icterus, or lipemia has historically been done by visual inspection, such an approach is highly unreliable.<sup>189</sup> Laboratory personnel are not able to accurately assess the degree of hemolysis, icterus, or lipemia in serum, even if well trained and when using a color standard for comparison.<sup>190</sup> Moreover, there is a poor inter-rater agreement (reproducibility) in estimating the degree of serum indices between different members of laboratory staff,



**FIGURE 5.6** Absorption curves of oxyhemoglobin in serum with characteristic peaks at 415, 540, and 570 nm (red line); triglycerides absorption curve covers wide range of wavelengths, with a maximum in the lower part of the spectrum (dotted black line); bilirubin has one distinct peak at 460 nm (black line).

reflecting the substantial interindividual differences in visual sensitivity to different colors.<sup>191</sup> For example, it has been demonstrated that visual inspection of the degree of hemolysis is influenced by the sample type (serum or plasma) and the test requested, thus leading to either over- or underestimation of the actual degree of hemolysis, depending on the expected effect of hemolysis on the measured analyte.<sup>192</sup> The ability to detect hemolysis by visual inspection is further impaired in samples that are both hemolyzed and icteric.<sup>193</sup> This is especially important in neonatal samples, where increased bilirubin concentrations are quite common.

Other substances, such as medical contrast media, may also influence the human ability to detect not only hemolysis but also icterus and lipemia.<sup>194</sup> One such example is Patent Blue dye, which is commonly used for sentinel lymph node biopsy in breast cancer patients. The presence of this dye in serum negatively affects the ability of laboratory personnel to reliably detect hemolysis, as well as icterus and lipemia. For the above-mentioned reasons, visual detection of the degrees of hemolysis, lipemia, and icterus is not recommended and should be replaced with automated detection systems whenever and wherever possible.

### Automated Serum Indices

Today, most mainstream chemistry analyzers can detect serum indices by the use of semiquantitative, spectrophotometric measurement and grading the interfering substances into categories. The serum index is automatically reported for every sample and can be used to determine the degree of interference and its effect on the requested parameters. Where an automated detection system for serum indices is not available, grading of interference factors by visual inspection is still a practical alternative used by many laboratories.<sup>195–197</sup>

Automated serum index detection systems have numerous advantages over visual detection (**Table 5.8**).

Such systems are highly reproducible and provide an objective and standardized way to screen for common interferences and manage specimen rejection via built-in rules. Moreover, their implementation improves laboratory turn-around time, leads to an increase in laboratory efficiency, and minimizes waste by reducing the number of rejected samples.<sup>198</sup> However, there are still some problems and challenges

associated with the automated detection of serum indices on various analytical platforms, which are detailed below.

**Variability between different analytical platforms.** There is a large variability across different chemistry analyzers in analytical characteristics of their serum index measurement. Different analyzers have different sensitivities, measurement ranges of, and decision thresholds for hemolysis, icterus, and lipemia. Moreover, they differ in the sample volume necessary for the estimation of serum indices and the type of solution used (saline, sample diluent, etc.). Different analyzers measure sample absorbance at various wavelengths and use different algorithms for determining the degree of serum indices. Finally, different manufacturers have employed different reporting systems to report the results of serum indices. Some are reporting qualitative results using the ordinal scale, whereas others are reporting semiquantitative results using actual concentrations of the interferent.<sup>199,200</sup>

**Necessity to verify manufacturers' claims.** It is the responsibility of the manufacturers of in vitro diagnostic systems and reagents to validate the analytical performance characteristics of their reagents and provide this information to the customer. The instructions for use must particularly contain the information about the effect of all known relevant interferences (e.g., serum indices) on laboratory assays. The CLSI EP7-A2 standard for interference testing in clinical chemistry<sup>201</sup> recommends that validation of the effect of an interferent on clinical chemistry assay be done at two concentrations of an analyte and at five concentrations of an interferent. The maximum concentration of an interferent must reflect the maximum expected concentration of that interferent in the clinical laboratory on patient samples. Moreover, the acceptance limits for allowable interference should be derived whenever possible from biological variability or clinically established thresholds. Due to financial constraints and the lack of time and staff, laboratories often rely on the information provided by the manufacturers, and only a minority of laboratories verifies manufacturer declarations.<sup>195</sup> However, manufacturers do not always comply with the recommended procedure for testing interferences, and their claims are often not accurate, reproducible, or reliable.<sup>202,203</sup> It is therefore a good practice for a laboratory to perform its own verification of serum indices. Alternatively, laboratories may rely on the evidence from the literature if it exists and only if the evidence is of adequate quality.

**Necessity to implement a systematic approach to internal and external quality control of serum indices.** Like all other laboratory methods, analytical quality of the method for detection of serum indices should be continuously monitored by using appropriate internal quality control (IQC) and through participation in an external quality assessment program (EQA). The ISO 15189:2012 International Standard for medical laboratories states, “EQA programs should, as far as possible, provide clinically relevant challenges that mimic patient samples and have the effect of checking the entire examination process, including pre- and postexamination procedures.”<sup>21</sup> EQA for serum indices may be conducted by sending samples with varying degrees of lipemia, hemolysis, and icterus to laboratories, and then participants provide their serum indices value and report results as they would for a patient sample.<sup>204,205</sup> Unfortunately, IQC and EQA are not widely available for serum indices. EQA for serum indices is currently available only from few providers (WEQAS and

**TABLE 5.8 Advantages and Disadvantages of Visual and Automated Detection of Serum Indices**

Visual detection	Automated serum indices
<ul style="list-style-type: none"> <li>• Subjective</li> <li>• Time-consuming</li> <li>• Requires available staff</li> <li>• Low reproducibility</li> <li>• Unreliable</li> </ul>	<ul style="list-style-type: none"> <li>• Objective</li> <li>• Reduces turn-around time (compared to visual detection)</li> <li>• Reproducible</li> <li>• Reliable</li> <li>• Potential problems:           <ul style="list-style-type: none"> <li>• Presence of paraproteins</li> <li>• Cross-reactivity of serum indices</li> <li>• Interference caused by drugs, contrast media, and other interferents</li> </ul> </li> </ul>

RCPA-QAP), and IQC material for serum indices has been recently made commercially available by a single manufacturer.<sup>206</sup> To overcome this problem, laboratories are encouraged to produce their own in-house prepared IQC materials and manage them in the same way as all other conventional laboratory IQC procedures.<sup>207</sup> For additional discussion on IQC and EQC, refer to Chapter 6.

#### Potential sources of interferences affecting serum indices.

Some medical contrast media are known to interfere with serum indices and impair the accurate determination of hemoglobin, bilirubin, and sample turbidity. It is important that laboratory staff be aware of this issue and that each sample be visually checked whenever serum index measurements raise suspicion or do not match the sample appearance or clinical condition of the patient. Patent Blue dye, which negatively affects the ability to detect changes in the sample's color, has also been found to interfere with serum indices measurement on the Roche Modular Pre-Analytics system and the Abbott Architect chemistry analyzer.<sup>208,209</sup> Patent Blue exerts positive interference on the lipemia index and a negative interference on hemolytic and icteric indices in a linear, dose-responsive fashion.

Rose Bengal has a peak absorbance at 562 nm and is a component of a drug that is used for intralesional therapy in patients with refractory cutaneous or subcutaneous metastatic melanoma.<sup>210</sup> Used in a treatment trial for severe melanoma lesions, it was found to cause false-positive interference on the hemolysis index on Roche Modular D in a sample with a red/pink tinge collected 20 minutes after the injection of a drug.<sup>211</sup>

Monoclonal proteins may also give an abnormal reading of serum lipemic index in apparently clear serum.<sup>212</sup> Markedly increased serum lipemia indices in clear sera were also quite frequently observed in patients with high concentrations of polyclonal immunoglobulins.<sup>213</sup> Nevertheless, unusually high lipemia indices in otherwise clear sera do not occur in all patients with monoclonal or biclonal peaks.

One serum index may also adversely affect the other when two or three HIL indices are abnormal in the same sample (e.g., serum hemolyzed and icteric, hemolyzed and lipemic). In these cases, the magnitude and direction of the bias of one index on the measurement of the other will vary greatly among different instruments and will depend on the respective wavelengths used.<sup>171</sup>

#### Management of Hemolytic, Icteric, and Lipemic Samples

Laboratories should be aware of the effect of preanalytical interferences on their assays. When there is a significant deviation from the true value of the analyte caused by the presence of cell compounds released by sample hemolysis, bilirubin, or increased concentration of serum lipids, such a result is a threat to patient safety. Biased and inaccurate results may cause diagnostic errors and affect patient management. To ensure the accuracy of their results, laboratories should have procedures in place to systematically detect the presence of potential interferences and how to address them. Unfortunately, there is a large discrepancy among the ways results are reported from samples with interferences, among different countries, institutions, and even individuals (e.g., analyze and report all components, reject the sample and not analyze anything, or analyze only selected components that

are not affected by the interferent).<sup>142,171,196,197</sup> There is room for improvement and harmonization in this respect.

When interferences from hemolysis, icterus, and lipemia are causing unacceptable bias and results are clinically inaccurate, such results should not be reported and sample redraw should be requested.<sup>214,215</sup> Such a test report should always be accompanied with comments informing the clinical staff about the reasons for not reporting the originally requested test results. It is also important that a laboratory notifies the medical staff when sample appearance (color, turbidity) deviates from a normal state by including a comment on a test report (e.g., sample hemolyzed, icteric, lipemic, or turbid), even if the tests are not affected by this change in appearance. Such comments provide useful information to the clinicians. Comments should also indicate if the sample has been treated in any way to minimize the effect of interfering substances (e.g., delipidation).

Unfortunately, each time a redraw is requested, there is a delay in providing the requested test results, and this leads to delays in patient management. In a previously mentioned study by Vecellio and colleagues, the length of stay (LOS) in emergency departments of five large Australian hospitals was on average 18 minutes longer if one or more samples for a particular patient were hemolyzed.<sup>142</sup> To avoid such delays, a laboratory should make a thorough investigation of the causes of the unsuitable specimens and be actively engaged in process improvement to reduce the frequency of errors that affect the quality of the sample.

#### POINTS TO REMEMBER

##### Hemolysis, Lipemia, and Icterus

- Visual assessment of the degree of hemolysis, lipemia, and icterus is not reliable and leads to errors.
- Hemolysis is the most common preanalytical error and most common cause of sample rejection.
- Hemolysis may cause clinically relevant bias through spectrophotometric and chemical interference, sample dilution, and release of the cell components into the sample.
- Lipemia causes interference by spectrophotometric interference (light absorption and light scattering), the volume depletion effect, partitioning of the sample, and physicochemical mechanisms (e.g., disturbance of the electrophoretic pattern).
- Laboratories should verify the performance of lipid removal reagents before their routine use because they may not be appropriate for a wide range of analytes due to their low recovery.
- Different forms of bilirubin cause varying degrees of interference with different laboratory methods, and the same forms of bilirubin act differently with the same assays on different instruments.

#### Interferences Caused by Paraproteins

Paraprotein interferences are not uncommon. The frequency of paraprotein interference has been estimated to be as high as 3 or 4% in hospitals,<sup>216</sup> and it has been reported to affect numerous laboratory assays (Table 5.9). Laboratory staff should carefully review every case in which a measured concentration of an analyte does not correlate with the clinical condition of the patient after all potential sources of errors have been investigated.

**TABLE 5.9 Paraprotein Interference With Different Assays**

Group of Analytes	Molecule
Enzymes	Alkaline phosphatase <sup>217</sup> Gamma-glutamyl transferase <sup>218</sup> Lactate dehydrogenase <sup>217</sup>
Electrolytes, minerals, and microelements	Calcium <sup>219</sup> Inorganic phosphorus <sup>220-224</sup> Iron <sup>225</sup>
Metabolites	Bilirubin <sup>226-229</sup> Cholesterol <sup>226,230,231</sup> Creatinine <sup>230</sup> Glucose <sup>218</sup> Urea <sup>232</sup> Uric acid <sup>217</sup>
Proteins	C-reactive protein <sup>230,233,234</sup> IgA, IgG <sup>235</sup>
Hormones	Thyroid-stimulating hormone <sup>236,237</sup> Human chorionic gonadotropin <sup>236</sup>
Drugs	Gentamicin <sup>238</sup> Vancomycin <sup>238-241</sup> Valproic acid <sup>238</sup> Phenytoin <sup>242</sup>
Cardiac markers	Troponin I <sup>236</sup>
Tumor markers	$\alpha$ -Fetoprotein <sup>236</sup> CA-125 <sup>236</sup>

Paraprotein interferences have been observed on different analytical instruments, and they appear to be methodology and concentration dependent. They affect not only measurements by turbidimetry and nephelometry but also some common chemistry assays with spectrophotometric detection. The likelihood of paraprotein-caused interferences increases with an increasing paraprotein concentration.<sup>229</sup>

### Mechanisms of Paraprotein Interference

Paraprotein interference may affect chemistry assays through several distinct mechanisms, including precipitation, volume displacement, and change of sample viscosity.

**Precipitation of the paraprotein.** A case of paraprotein interference has been reported in a 93-year-old female with severe dementia who presented with cellulitis on the leg and sepsis. Gentamicin was not measurable due to the paraprotein interference caused by the IgM monoclonal protein.<sup>238</sup> The IgM concentration was 18.9 g/L (reference interval: 0.4 to 2.3 g/L). Paraprotein interference was observed on a Beckman Dx-C600 general chemistry analyzer with a particle-enhanced turbidimetric inhibition immunoassay method (Beckman Coulter, Brea, CA), and it was a result of the persistent high blank absorbance readings. Gentamicin concentration for that patient was successfully measured (3.3 mg/L) on a Roche Cobas system (Roche Diagnostics, Mannheim, Germany) where interference was not present.

Paraprotein interference can be detected by reviewing the reaction curve on the instrument for that specific patient. The reaction curve for the sample without interference shows that precipitation of the gentamicin occurs when precipitating reagent is added to the reaction mixture. The reaction curve for the affected sample shows that IgM precipitation

has occurred in the blanking phase, even before precipitating reagent is added to the reaction mixture. This phenomenon was also observed in a sample that was diluted with a normal saline at a ratio of 1:20 (Fig. 5.7A–C). The manufacturer's instructions for the gentamicin assay states that there is no interference by IgM up to 5 g/L; IgM concentration in this patient was fourfold higher.

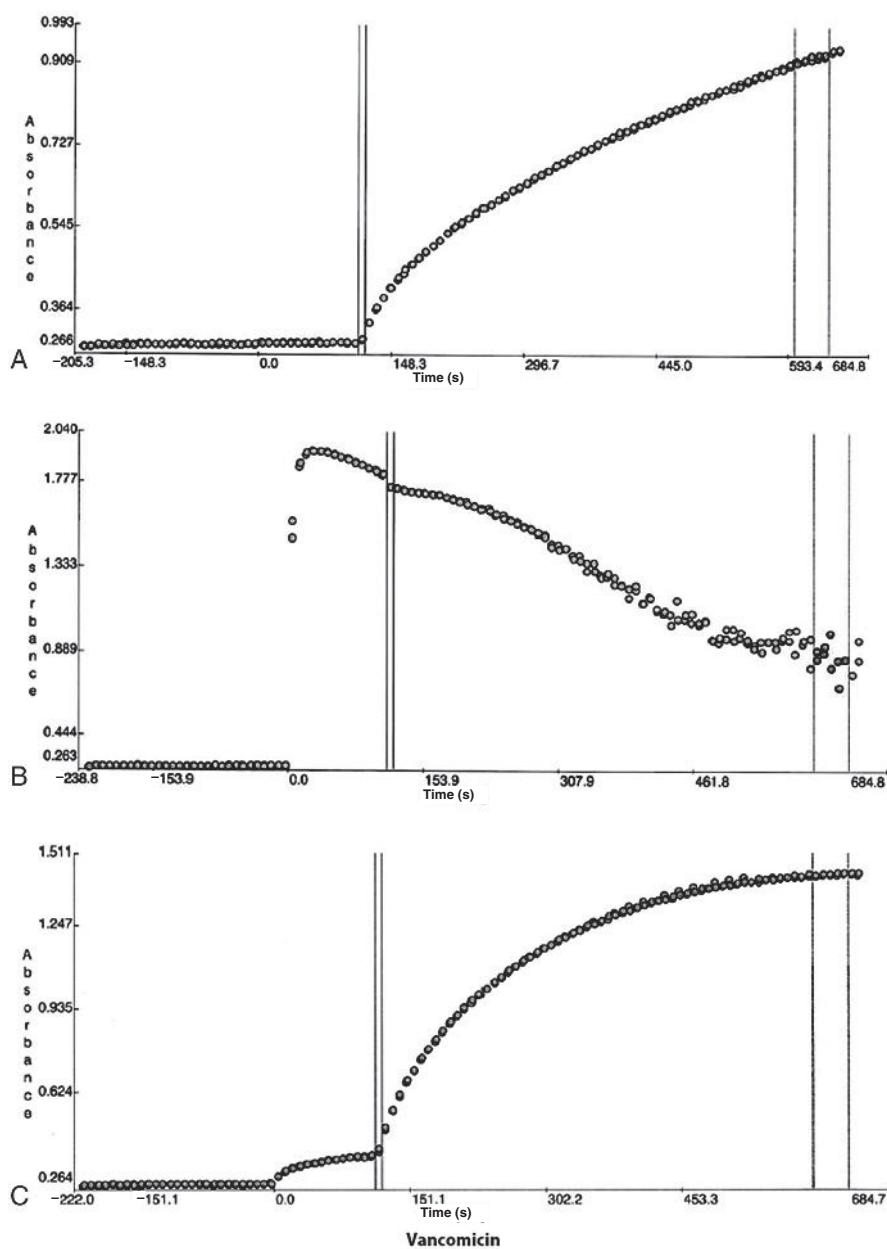
Precipitation depends on various assay parameters, such as reaction components, presence of assay additives such as preservatives and surfactants, ionic strength, pH (protein precipitation can occur at both very low and very high pH), and the physicochemical characteristics of the paraprotein. This explains why some assays are affected and others are not by the same paraprotein on the same instrument.

Monoclonal proteins have been reported to appear in serum with a concentration of up to 104.1 g/L.<sup>228</sup> Manufacturers should improve the way in which they test and report data from interference studies. Paraprotein interference should be studied in the whole range of expected concentrations of paraproteins. Laboratories must carefully read the declarations provided by manufacturers and perform their own interference studies to verify the absence or presence of paraprotein interference.

Paraprotein interference may vary according to the type of specimen or the choice of anticoagulant used in sample collection. For example, IgM interfered with a hexokinase method for glucose and a Szasz method for GGT using the Hitachi Modular D and P systems (Roche Diagnostics GmbH).<sup>218</sup> Glucose in lithium heparinized plasma was extremely low in that patient, but the interference was not present in the actual collection tube. Moreover, the interference was absent when glucose and GGT were tested with dry chemistry on a Vitros 950 analyzer (Ortho Clinical Diagnostics). Precipitation of a paraprotein has occurred due to fibrinogen precipitation resulting from the action of heparin. In this case, the best solution is to request a serum sample for that patient or to run the tests affected by the interference using a different method.

Precipitation of paraprotein may occur due to the reaction of paraprotein and the solubilizing agent in the reagent used for the measurement of the concentration for total bilirubin. This mechanism of paraprotein interference has been reported to affect bilirubin measurement by Hitachi Modular P random access autoanalyzer using Roche test kits (Roche Diagnostics GmbH). Such interference leads to a false increase of bilirubin concentration in serum with otherwise normal (anicteric) appearance and in the absence of hemolysis or lipemia.<sup>228,243</sup>

**Binding of paraprotein to assay components.** Paraproteins may bind to the analyte or any other component of the reaction mixture. The effect of such interference depends on the component to which the paraprotein is bound. Binding of an IgM paraprotein to latex particles resulted in high CRP and ASO values in a young Japanese female myeloma patient.<sup>234</sup> The IgM concentration was grossly increased at 70.0 g/L (reference interval: 1.31 to 2.83 g/L). Concentrations of CRP, ASO, IgG, IgA, and IgM were measured by a Behring nephelometer II automated analyzer (Behringwerke AG, Marburg, Germany) that used latex particles coated with anti-CRP rabbit antibody, streptolysin-O antigen, and anti-human IgG, IgA, and IgM rabbit antibody (Behringwerke AG), respectively. When measured with another method, CRP and ASO concentrations were within the reference interval.

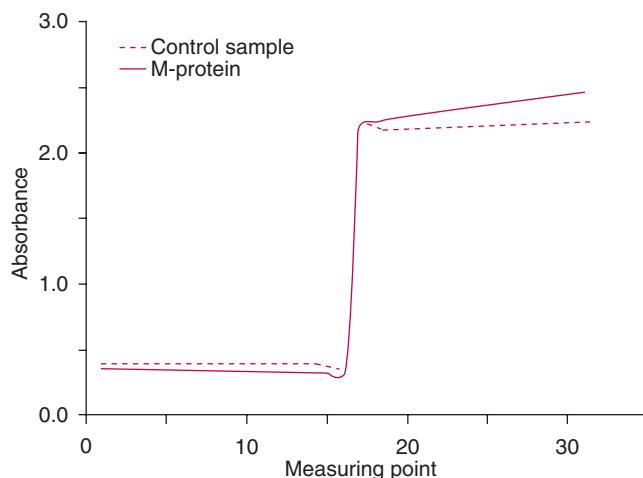


**FIGURE 5.7** Reaction curve for the gentamicin particle-enhanced turbidimetric inhibition immunoassay method on the Beckman DxC600 general chemistry analyzer in a patient in whom gentamicin was not measurable due to the paraprotein interference caused by the IgM monoclonal protein. (A) Precipitation of the analyte occurs when precipitating reagent is added to the reaction mixture. The reaction curve for the sample is affected by the paraprotein interference (IgM). (B) Precipitation of the analyte occurs in the blanking phase (unexpectedly high blank absorbance readings), even before the precipitating reagent is added to the reaction mixture. (C) The change in the reaction curve is visible even in a sample that has been diluted with a normal saline in a ratio of 1:20. (From Dimeski G, Bassett K, Brown N. Paraprotein interference with turbidimetric gentamicin assay. *Biochem Med* 2015;25:117–24, with permission by the Croatian Society of Medical Biochemistry and Laboratory Medicine.)

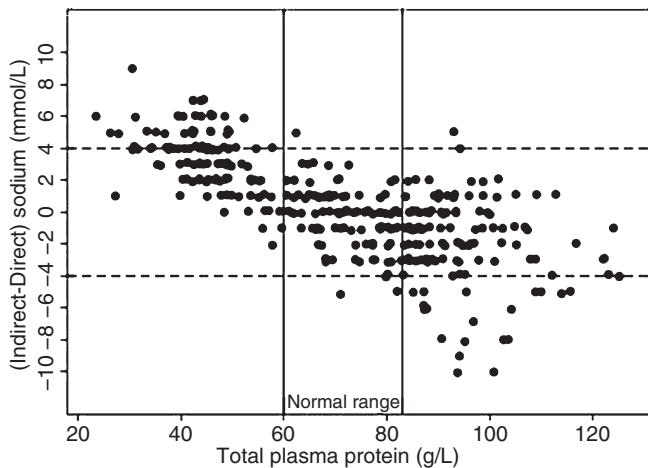
This kind of interference does not cause sample turbidity. Reaction kinetics for the unaffected sample and a sample with an interfering paraprotein are very similar, and therefore this type of interference cannot be detected by reviewing the reaction curve on the instrument (Fig. 5.8).<sup>244</sup>

**Paraprotein interference due to volume displacement.** Paraproteins affect chemistry assays by the same mechanism as

lipemia—that is, due to the volume displacement effect. Most pronounced are changes in serum electrolytes and especially in serum sodium measurement by indirect ISE technology (ISE). A high concentration of paraproteins leads to false hyponatremia if sodium is measured by indirect ISE. As a general rule, in samples with total protein concentration greater than 100 g/L or less than 40 g/L, electrolytes (and



**FIGURE 5.8** An example of IgM (7.45 g/L) paraprotein interference with measurement of ferritin concentration on the Roche/Hitachi 911 analyzer, caused by binding of the paraprotein to the components of the reaction mixture. The ferritin concentration was 492 mg/L on the Roche/Hitachi 911 analyzer and six times lower (80 mg/L) when measured with a different assay on another instrument. (From Bakker AJ, Mücke M. Gammopathy interference in clinical chemistry assays: Mechanisms, detection and prevention. *Clin Chem Lab Med* 2007;45:1240–43, with permission by Walter de Gruyter.)



**FIGURE 5.9** The association of the difference between direct and indirect ion-selective electrode measurement of plasma sodium relative to plasma protein concentration. Vertical lines demonstrate the reference interval for total plasma protein. Dashed line shows disagreement of 4 mmol/L or greater in sodium measurements. (From Dimeski G, Morgan TJ, Presneill JJ, Venkatesh B. Disagreement between ion selective electrode direct and indirect sodium measurements: Estimation of the problem in a tertiary referral hospital. *J Crit Care* 2012;27:326.e9–16, with permission from Elsevier.)

especially sodium) should be measured by direct ISE technology.<sup>245,246</sup> Otherwise, clinically significant bias is possible and may lead to adverse patient outcomes.

The concentration of plasma sodium, measured by indirect ISE methods, is inversely proportional to the concentration of plasma proteins. The greater the concentration of plasma proteins, the lower the concentration of plasma sodium (Fig. 5.9).

### Paraprotein interference due to change in sample viscosity.

Paraproteins may exert their interference simply by affecting the viscosity of the sample. Viscosity is much higher in samples with very high paraprotein concentration or in refrigerated samples in which a gel has been formed (e.g., in the case of cryoglobulinemia). Sample viscosity affects the volume of the sample pipetted in the reaction mixture. Many instruments are able to detect such changes in sample viscosity and trigger an alarm if viscosity is not within predefined limits. In such cases, a rerun in a dilution mode is the recommended corrective action, provided the analyte can be accurately measured in the diluted specimen. In instruments where this feature is not available, increased sample viscosity may lead to incorrect sample volume and cause falsely increased or decreased results, depending on the assay format.

### How to Detect and Deal With Paraprotein Interference.

Laboratories can put safeguards in place to automatically detect interference caused by paraproteins:

- Have the instrument sound an alarm if the difference in absorbance between selected points is unexpectedly greater than some predetermined value
- In cases of interference of paraproteins on bilirubin concentration, have the instrument flag the results for confirmation when the bilirubin concentration in a sample is higher than some predetermined value and the icterus index is normal
- Have the instrument flag and block all test results that are preceded by a minus sign
- Have the instrument block all test results for patients who have erroneous results, such as higher direct bilirubin than total bilirubin or higher albumin than total proteins

Hopefully, future chemistry analyzers will be able to monitor analytical reactions in real time and automatically screen for potential interferences similarly to the detection of serum indices.

Laboratories may also apply various approaches to eliminate paraprotein interference<sup>173</sup>:

- Samples can be analyzed on an alternative instrument with a different method
- Proteins in the sample can be precipitated by a blocking agent, ethanol, ammonium sulphate, or polyethylene-glycol, while the analyte of interest remains in the supernatant
- Serial dilutions of the sample may be performed
- The sample can be filtered to remove the proteins

### Exogenous Interferences

Exogenous interferent is defined as a substance originating outside of the body (e.g., a drug or its metabolites, a specimen preservative, or a sample contaminant) that causes interference with the analysis of another substance in the specimen.<sup>247</sup> Exogenous interferences are quite complex and difficult to identify and deal with.<sup>173,201</sup> Interferences can be introduced into the sample via several different sources:

- prescribed medication used for patient treatment
- supportive medical therapy, including parenteral emulsions, contrast media agents, and infusion solutions
- natural preparations (herbal and animal remedies) and dietary supplements
- accidental exposures and poisonings
- contamination during sample handling (from rubber tube stoppers, lubricants, anticoagulants, or surfactants) or sample analysis (antibiotics used for reagent and buffer stability)

## Mechanisms of Interference

Drugs interfere in laboratory testing in two ways: *in vivo*, through pharmacologic effect of the drug on a specific analyte, and *in vitro* during the actual analysis. The comprehensive database of the physiologic (*in vivo*) and analytical (*in vitro*) drug effects is available as an on-line resource.<sup>248</sup>

Analytical (chemical) *in vitro* interference is caused when the presence of the drug directly or indirectly leads to falsely increased or decreased concentration of a measured analyte.

There are different mechanisms of drug *in vitro* interference:

- the parent drug or its metabolite cross-reacts with the substrate, due to the structural similarity with the tested analyte
- the parent drug or its metabolite interferes with the chemical reaction by accelerating or inhibiting it
- the parent drug or its metabolite affects sample turbidity or absorbance and causes overlap of drug and chromogen absorption peaks<sup>249,250</sup>
- some drugs can interfere with the integrity of the sample by changing sample density (viscosity) and cause obstruction problems on analytical systems (e.g., iodine-based contrast media)

Chemical mechanisms of interference are discussed in more detail later in this chapter.

## Manufacturer Claims Regarding Drug Interferences

The impact of exogenous interferent on patient condition could range from harmless to fatal. Manufacturers of laboratory reagents are responsible to conduct extensive interference testing and provide all relevant information about any possible interference susceptibility of their assays to their end-users. According to the CLSI EP07-ED3:2018 standard, interference testing is composed of the following steps:

- selection of appropriate analyte concentrations for interference testing (concentrations at medical decision points)
- selection of substances which could interfere with assay
- selection of appropriate tested interferent concentrations
- assurance of reliability of testing procedures during interference evaluation
- paired testing of sample without and with added interferent drug in concentration three times the highest recorded concentration during therapeutic drug monitoring
- if difference in concentration of tested analyte does not exceed defined acceptance criteria, there is no need for additional interference testing
- if acceptance criteria have been exceeded, dose-response experiment has to be performed
- when dose-response experiment is completed, interference should be reported.<sup>247</sup>

Laboratory professionals are responsible for verifying manufacturer interference claims, investigating any discrepant result that might be caused by an interfering substance, and providing feedback to the manufacturers.<sup>201</sup> End-users are also responsible for reporting any inconsistencies in the use of diagnostic products which could impact patient safety to the local agencies responsible for postmarket surveillance. Moreover, it is important that laboratory professionals have safe procedures in place for effective detection of new potential sources of exogenous interferences and the knowledge about the type and quantity of potential exogenous interferents used in patient management in order to prevent possible negative outcome.

Exogenous interferences are not easy to detect and it is quite unlikely that a laboratory will have an error-proof system, consistent policy, and procedures in place to safeguard from exogenous interferences. Most commonly, a suspicion is raised when test results do not match the clinical condition of the patient or when there is a sudden change in patient test results. Different strategies are suggested for investigation of suspected exogenous interference. Table 5.10 provides a list of some most common exogenous interferents and measures which should be undertaken in case of their occurrence.

Whenever there is a clinical suspicion, due to the sudden change in laboratory test results, that some other tube components might have caused some interference, it is recommended to recollect the sample using different tube type and/or to perform the analysis on a different analytical platform and/or method.

Besides interferences of endogenous components that are listed in manufacturers' claims (lipids, hemoglobin, bilirubin, proteins), manufacturers usually include the results of interference testing for drugs that could potentially affect the measurement. The drug list is adjusted for any specific analyte. Drugs and metabolites that are likely to interfere on the basis of their chemical and physical properties and those most often prescribed in the patient population for which the test is ordered should be tested.<sup>201</sup> Tested drug concentrations should cover the entire range of the expected blood drug concentrations. To minimize the risk of unrecognized drug effect due to the low tested concentration, the highest expected drug concentration should be tested at least three times. A list of proposed tested drug concentrations is provided in CLSI document EP7-A2, *Interference Testing in Clinical Chemistry*.<sup>201</sup> Table 5.11 presents an example of drug interferences listed in the reagent insert sheet provided with the Abbott Hemoglobin A1c enzymatic assay (Abbott Laboratories, Lake Bluff, IL).

Unfortunately, information about patient medication is often not available to laboratory staff. Hospital electronic medical records can help identify whether any potentially interfering drug is prescribed to the patient. However, in most cases there is no way of knowing if the drug concentration in patient blood exceeds the allowable threshold. For most of the drugs listed in Table 5.11, concentration can be measured only in specialized laboratories. Therefore many drug interferences remain undetected and only a clear discrepancy between the test result and clinical information prompts the laboratory staff to suspect drug interference and proceed to further investigation, as already described above.<sup>251</sup>

Drug interference in laboratory tests is a major cause of concern. Lack of information on influence and interference of specific drugs carries risk of reporting of erroneous results that could lead to faulty patient management. Unfortunately, these interactions are still largely unrecognized and only a minority of drug labels contain information on drug-laboratory test interaction. The way forward to solve this problem would be to establish an on-line database, continuously updated, with information on possible impacts of drugs and other exogenous interferents on laboratory measurements.<sup>252,253</sup>

## Prescribed Medication

Prescribed medication can often interfere with the measurement and cause erroneous laboratory results. There are numerous reports describing the interference of prescribed

**TABLE 5.10 A Summary of Common Exogenous Interferences and How to Identify and Investigate Them**

Contaminant	Common Potentially Affected Analytes (Not an Exhaustive List)	When to Suspect	Point of Occurrence of the Interference	Initial Investigations
Drugs/herbal remedies	All analytes could be affected	Sudden change in value Clinical suspicion	Patient	Literature search Recollect the sample after removal of medication
Skin disinfectants	Electrolytes, LD, AST, bilirubin, phosphate, folate, urate	Hemolysis Sudden change in value Clinical suspicion	Phlebotomy	Recollect the sample Check indices values
EDTA	Potassium, calcium, zinc, magnesium, iron, ALP	Sudden change in value Clinical suspicion Potassium very high	Phlebotomy	Check divalent cations Measure EDTA Recollect the sample
Sodium citrate	Sodium	Sudden change in value Clinical suspicion Sodium very high	Phlebotomy	Check chloride Check osmolar gap Recollect the sample
Heparin	Sodium, analytes measured by immunoassays, thyroid Function tests	Sudden change in value Clinical suspicion	Patient or phlebotomy	Recollect the sample once source of heparin has been removed
Fluoride oxalate	Hematocrit, electrolytes	Sudden change in value Clinical suspicion	Phlebotomy	Recollect the sample
Gel separators	Hydrophobic drugs, drugs and hormones measured by mass Spectrometry, e.g. testosterone	Sudden change in value Clinical suspicion Interfering peaks	Phlebotomy	Recollect the sample using different tube type
Contrast media	Sample potentially unsuitable for all analytes measured by spectrophotometric assays Divalent cations Indices value	Sudden change in value Clinical suspicion Gel barrier incorrectly positioned after centrifugation Instrument flags	Patient	Recollect the sample when source has been removed
Antibody therapies	Analytes measured by immunoassays	Sudden change in value Clinical suspicion Instrument flags	Patient	Try different analytical method/platform Remove treatment
IV infusion	Potassium, sodium, glucose Dilution of analytes not being infused	Sudden change in value Clinical suspicion Reagent lots of low analyte concentrations (with normal sodium if saline)	Patient	Recollect the sample from other arm, or when infusion is finished
Analyzer problems or Contamination	All analytes could be affected	Sudden change in value across multiple patients Clinical suspicion	Laboratory	Run samples on another analyzer Recalibrate analyzer Maintain analyzer

ALP, Alkaline phosphatase; AST, aspartate aminotransferase; LD, lactate dehydrogenase.

Modified from Cornes MP. Exogenous sample contamination. Sources and interference. *Clin Biochem* 2016;49:1340–5.

medication in the measurement of various analytes in different matrices, such as whole blood, serum, plasma, and urine samples. The type and magnitude of the interference is dependent on a number of factors such as drug dose, timing, route of administration, duration of drug application, concomitant use of other drugs and herbal supplements, method of analyte determination, reagent composition, analyzer used, age of detecting components, sample type, etc.

Although all laboratory methods can potentially be affected by drug interferences, specific enzymatic chemistry methods are usually not as susceptible as colorimetric tests. Creatinine is one of the most commonly measured parameters in a clinical chemistry laboratory. Laboratories worldwide use either the Jaffe colorimetric assay with alkaline

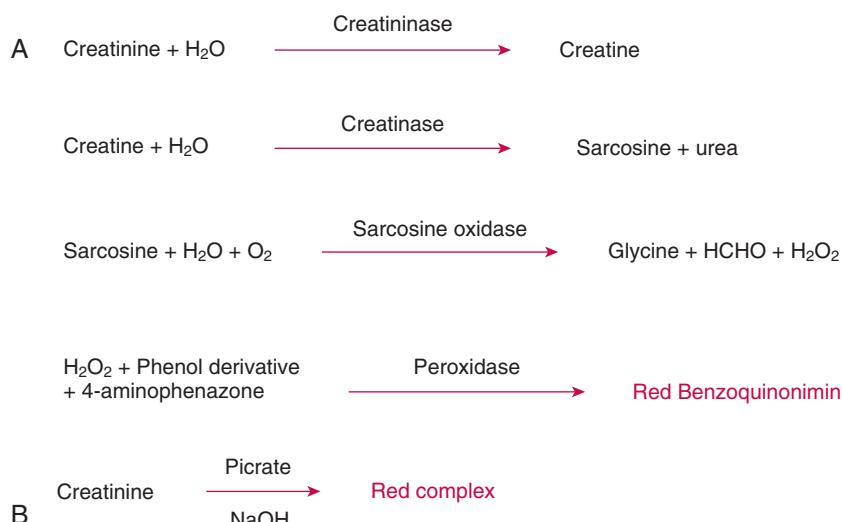
picate or an enzymatic assay for creatinine measurement (Fig. 5.10).

Some commonly prescribed antibiotics and analgesic drugs can interfere with creatinine measurement. For example, cephalosporin antibiotics cause falsely increased creatinine concentrations (measured by the Jaffe method) in sera drawn shortly after intravenous administration of cefpirome.<sup>254,255</sup> Falsely increased creatinine concentration measured by the Jaffe method has also been observed after adding subtherapeutic, therapeutic, and toxic concentrations of acetaminophen, acetylsalicylic acid, or metamizole to the serum. Although the Jaffe method has always been considered to be more susceptible to interference compared to the enzymatic assay, it is now known that even the enzymatic assay is not free from the interfering

**TABLE 5.11 List of Interfering Drugs Tested for Potential Interference With HbA<sub>1c</sub> Enzymatic Assay**

Interfering Substance	HIGH TEST LEVEL		% INTERFERENCE	
	Conventional Units	SI Units	6.0–7.0% HbA <sub>1c</sub> Value	≥8.0% HbA <sub>1c</sub> Value
Acarbose	50 mg/dL	0.77 mmol/L	0.0	0.0
Acetaminophen	200 µg/mL	1324 µmol/L	-1.5	-1.1
Acetylsalicylate	50.8 mg/dL	2.82 mmol/L	0.0	0.0
Atorvastatin	600 µg Eq/L	600 µg Eq/L	0.0	0.0
Captopril	0.5 mg/dL	23 µmol/L	-1.5	-1.1
Chlorpropamide	74.7 mg/dL	2.7 mmol/L	0.0	0.0
Cyanate	50 mg/dL	6.16 mmol/L	0.0	1.1
Furosemide	6.0 mg/dL	181 µmol/L	0.0	0.0
Gemfibrozil	7.5 mg/dL	300 µmol/L	0.0	0.0
Ibuprofen	0.5 mg/mL	2524 µmol/L	0.0	0.0
Insulin	450 µU/mL	450 µU/mL	0.0	0.0
Losartan	5 mg/dL	0.11 mmol/L	0.0	0.0
Metformin	5.1 mg/dL	310 µmol/L	0.0	0.0
Nicotinic acid	61 mg/dL	4.95 mmol/L	0.0	0.0
Propranolol	0.2 mg/dL	7.71 µmol/L	0.0	0.0
Repaglinide	60 ng/mL	132.57 nmol/L	0.0	0.0

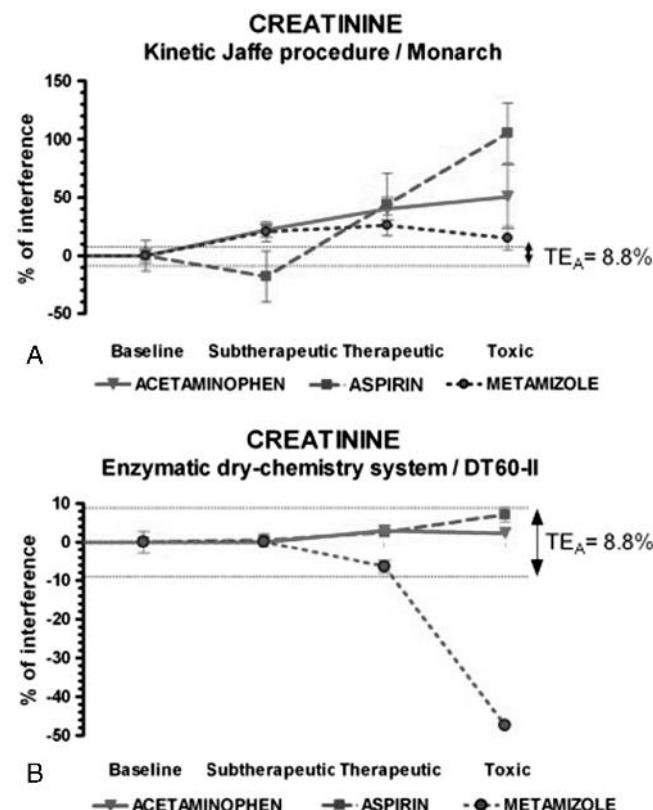
From reagent insert sheet provided with the Abbott HbA<sub>1c</sub> reagent (REF 4P52-21). From Abbott Laboratories, Lake Bluff, IL.



**FIGURE 5.10** Methods for the measurement of creatinine. (A) Enzymatic assay: the absorption at 546 nm of the red-colored product of the indicator reaction is proportional to creatinine concentration. (B) Jaffe colorimetric assay: creatinine reacts with picrate in alkaline media and forms a red complex. Change in absorption at 509 nm is proportional to creatinine concentration.

effects of various drugs. An *in vitro* study has demonstrated that the addition of toxic concentrations of metamizole to serum pool causes falsely decreased creatinine concentrations measured by the enzymatic method (Fig. 5.11).<sup>256</sup> It has also been demonstrated that metamizole causes significant negative

interference with several isotopic dilution mass spectrometric traceable enzymatic creatinine methods and some other routine biochemistry analytes.<sup>257,258</sup> To avoid the interference of metamizole, it has been suggested that creatinine measurements are done at least 3 hours after drug injection.<sup>257</sup>



**FIGURE 5.11** Analgesic drug interference with creatinine measurement. (A) Interference of acetaminophen, acetylsalicylic acid (aspirin), and metamizole on creatinine measurement by Jaffe method. All tested drugs cause interference error over acceptance criteria based on biological variations (dashed line; total error = 8.8%). (B) Interference of acetaminophen, acetylsalicylic acid (aspirin), and metamizole on creatinine measurement by enzymatic method. Only toxic concentrations of metamizole cause a falsely decreased concentration of creatinine. (Reproduced from Luna-Záízar H, Virgen-Montelongo M, Cortez-Álvarez CR, et al. In vitro interference by acetaminophen, aspirin, and metamizole in serum measurements of glucose, urea, and creatinine. *Clin Biochem* 2015;48:538–44, with permission by Elsevier.)

Ethamsylate is a hemostatic drug indicated in cases of capillary bleeding.<sup>259</sup> This drug causes a significant decrease in creatinine concentration (approximately 50%) measured by enzymatic assay. Interestingly, the Jaffe method is not influenced by ethamsylate. The interference is probably caused by the presence of a hydrochinone structure in the ethamsylate molecule. Hydrochinone interferes in the last reaction of creatinine quantification (see Fig. 5.10A). Therefore the other methods that are using the same indicator reaction are also affected by the administration of this drug. There is evidence that ethamsylate also causes significant false decrease in the concentrations of cholesterol (9.2%), triglycerides (15.6%), and uric acid (15.4%).<sup>260,261</sup> A similar problem with the enzymatic creatinine assay was found in the presence of high concentrations of therapeutically administered catecholamines. Dopamine, dobutamine, epinephrine, and norepinephrine cause strong negative interference with the Creatinine Plus enzymatic (Roche Diagnostics GmbH, Penzberg, Germany) assay. The Vitros (Ortho Clinical Diagnostics) enzymatic creatinine assay demonstrates slight negative interference, while

the i-STAT enzymatic (Abbott) and Jaffe methods (Roche Diagnostics and Siemens [Siemens Healthcare Diagnostics, Munich, Germany]) are unaffected by the presence of therapeutically administered catecholamines.<sup>262</sup>

Due to its hepatotoxicity, eltrombopag, a thrombopoietin receptor agonist used for treatment of immune thrombocytopenia, requires careful monitoring of liver function tests.<sup>263</sup> It has been shown that eltrombopag is a colored drug and as such has the potential to interfere with many routine chemistry spectrophotometric assays. The color of eltrombopag is pH dependent. In alkaline pH, the color of eltrombopag is reddish brown, while it is yellow in acidic pH. Therefore this drug may cause interference with varying degree on different instruments, depending on the assay design and pH of the reaction solution. For example, the pH-dependent color change interferes with the total bilirubin assay and some other routine chemistry tests with varying degree on several chemistry instruments (Beckman Coulter DxC, Roche Cobas, Siemens Advia and Abbott Architect), whereas total bilirubin assay on Beckman Coulter AU analyzer is resistant to this interference due to the existence of separate total and blank channels with identical pH reagents.<sup>263</sup>

Ionized calcium is the best indicator of calcium homeostasis in patients with suspected or known derangements of calcium metabolism.<sup>264</sup> Direct potentiometric determination of ionized calcium is known to be susceptible to the interference caused by several drugs. Degree of positive or negative drug interference is dependent on half-life of drug, specific manufacturer design, and age of the electrode. One such example is the active metabolite of leflunomide (teriflunomide), a synthetic isoxazole-derivative drug with immunosuppressive and antiviral properties, which exerts interference with ionized calcium. In kidney transplant patients treated with leflunomide, low ionized calcium concentrations are sometimes observed without any clinical signs of hypocalcemia. This interference is caused by an active metabolite of leflunomide. Shortly after oral administration, leflunomide is rapidly converted into its active metabolite teriflunomide, which is responsible for this falsely decreased ionized calcium concentration. The interference is dependent on the type of the blood gas analyzer used for the measurement of ionized calcium; it was found to affect the Rapidlab-1265 (Siemens Diagnostics) and i-Stat point-of-care analyzer (Abbott), but not the ABL800-FLEX blood gas analyzer (Radiometer, Copenhagen, Denmark).<sup>265</sup> Because this drug is widely used not only in kidney transplant patients but also in patients with rheumatoid arthritis, every hypocalcemia in laboratories using RAPIDlab or i-Stat point-of-care analyzers should be interpreted with caution and compared with the patient's clinical condition. Unfortunately, this is not the only reported interference for ionized calcium measurement. Sodium perchlorate, which is used as an oral drug to treat hyperthyroidism, causes a significant false decrease in calcium results when using Radiometer and Siemens RAPIDlab POCT instruments for measuring ionized calcium.<sup>266</sup>

Another example of interference with indirect potentiometry is related to electrolyte measurement. Measurement of chloride concentration is important in the management of critically ill patients, where acid-base balance is compromised and information on anion gap is essential for correct patient management. It has been demonstrated that bicarbonates interfere with chloride measured by ISE on Roche cobas 6000

c501 analyzer; the chloride concentration increase is proportional to the bicarbonate concentration.<sup>267</sup>

Thiopental is a barbiturate used for the treatment of increased intracranial pressure. The central laboratory analyzer Dimension Vista (Siemens), which uses the V-LYTE Integrated Multisensor Technology system for electrolyte measurement, produces falsely increased sodium concentrations in the presence of thiopental. However, when the sodium is measured on a point-of-care analyzer using direct potentiometry (RAPIDlab 1200, Siemens), there is no evidence of thiopental interference.<sup>268</sup>

Drug interferences are also observed in coagulation testing. Dabigatran etexilate, a new oral direct thrombin inhibitor, is used as an anticoagulant drug. Application of this drug causes significant bias in many coagulation tests. Dabigatran causes dose-dependent prolongation of PT and aPTT, and thus significantly changes the results of various other coagulation tests, including antithrombin III and coagulation factors II, V, VII, VIII, IX, X, XI, XII, and XIII. Falsely decreased factor activities and falsely positive misdiagnosis of lupus anticoagulant are observed in the presence of dabigatran.<sup>269</sup>

Another example of the interference in coagulation testing is telavancin, an antibiotic used for the treatment of methicillin-resistant *Staphylococcus aureus* (MRSA). Telavancin was found to interact with the artificial phospholipid reagent Dade Behring Innovin for PT measurement, causing falsely decreased values.<sup>270</sup>

Carbohydrate-deficient transferrin (CDT) is a biomarker for long-term alcohol consumption. The evidence shows that the sensitivity of the test for its measurement can be affected by the use of several drugs like amlodipine (calcium channel blocker), perindopril (ACE inhibitor), atorvastatin (statin), isosorbide mononitrate (nitrate), carvedilol (beta blocker), ticlopidine (inhibitor of platelet aggregation), and pantoprazole (gastric acid pump inhibitor).<sup>271</sup> It is still unclear whether the false-negative measurement is a result of polytherapy where unexpected metabolic pathways can be activated. Even if some or all of the drugs are listed as potential interfering substances, manufacturers usually do not declare what effects can occur *in vivo* when multiple drugs are combined.

The most common drug interferences are reported for urine matrix and most common among them are interferences related to screening for drugs of abuse and dipstick tests. This is why urine drugs of abuse screening for employment or employee control demands confirmatory testing with more reliable methods.<sup>272</sup> Regular use of substances like bupropion, an antidepressant and an aid for smoking cessation, could have an impact on urine drug testing. Bupropion metabolite at concentrations higher than 500 ng/mL cross-reacts with two enzyme-linked immunosorbent assays for urine amphetamine, giving false positive results.<sup>273</sup> Hydrochloroquine, a drug used for rheumatoid arthritis treatment, was also shown to interfere with drugs of abuse urine tests.<sup>274</sup> Long-term, high-dose morphine therapy in hospice patients could cause an appearance of low concentration of hydro-morphine in urine, rather as morphine metabolite, than as an indicator of opioid abuse.<sup>275</sup> Quinolone antibiotics and drugs derived from quinine interfere with urine test strip Multistix 10 SG protein detection and with quantitative urine protein test method using pyrogallol red-molibdate.<sup>276</sup> Ciprofloxacin, quinine, and chloroquine in supratherapeutic concentrations and chloroquine in therapeutic concentration

give spurious proteinuria results when using Multistix 10 SG urine tests strips.<sup>277</sup>

Even highly specific and sensitive methods employed for measurement specific analytes in 24-hour urine sample, like high-performance liquid chromatography with electrochemical detection (HPLC/ECD) and liquid chromatography with tandem mass spectrometry (LC-MS/MS), are susceptible to drug interference. Sulfasalazine interference caused falsely high normetanephrine concentration in 24-hour urine sample.<sup>278</sup> Determination of free cortisol concentration in 24-hour urine by liquid chromatography coupled to atmospheric pressure ionization tandem mass spectrometry (LC-ESI-MS/MS) was shown to be affected by the antibiotic piperacillin; one way to circumvent this problem is to adjust urine sampling according to the kinetics of the interfering drug.<sup>279</sup>

Electrophoresis-based methods are also susceptible to this interference and may give false-positive findings, if unrecognized.<sup>280</sup> False positive electrophoresis and serum immunofixation electrophoresis results showing monoclonal IgG-kappa were reported in patients taking ofatumumab, a monoclonal antibody used in the treatment of patients with Waldenstrom macroglobulinemia.<sup>281</sup> Another example is siltuximab, a monoclonal antibody directed to IL-6, intended for treatment of malignancies in clinical trials, which is known to interfere with serum protein electrophoresis. Interference was observed in patients with IgDκ multiple myeloma, who had spurious finding of IgG heavy chain immunoglobulin.<sup>282</sup>

Last, but not the least, interference of monoclonal antibodies used for therapeutic purposes are now receiving increasing attention due to the fast-growing production of antigen-directed monoclonal antibodies which enable targeted, effective, personalized patient treatment. Monoclonal antibodies are used for the treatment of malignancies, autoimmune diseases, transplant-related conditions, and inflammatory diseases. They have a significant impact on different laboratory tests that may lead to diagnostic errors. Most compromised are those methods which utilize monoclonal antibodies such as immunoassays, cytometric analysis, immunohistochemistry assays, etc. One such example is alemtuzumab, therapeutic immunoglobulin which interferes with flow cytometry, giving false positive finding of light chain clonality, thus raising the risk of misdiagnosis of B-cell neoplasms.<sup>283</sup> Another example is a monoclonal anti-CD47 which has been recognized as a promising treatment option for hematologic and solid malignancies. It was observed that it interferes in pretransfusion erythrocyte and platelet compatibility testing.<sup>284</sup> There are also studies reporting the interference of daratumumab, a monoclonal antibody used for the treatment of multiple myeloma, in electrophoretic and immunofixation testing<sup>285–287</sup> and in immunohematologic assays.<sup>288</sup> Other therapeutic monoclonal antibodies have also been known to interfere with serum protein electrophoresis and several strategies for resolving this type of interference have been suggested.<sup>289–291</sup>

It should be pointed out again that this interference depends on the assay and instrument. This has been nicely demonstrated by the College of American Pathologists (CAP) survey aimed to explore the impact of the therapeutic monoclonal antibody omalizumab, intended for treatment of asthma patients, on IgE immunoassay performance. Plasma specimens

of atopic patients incubated with omalizumab were sent to 159 laboratories—participants of Allergy Proficiency CAP Survey—and the results have shown that all manufacturers except Pharmacia ImmunoCAP total and specific IgE assays are susceptible to interference by omalizumab.<sup>292</sup>

Potential drug interferences are numerous, and not all of them can be recognized or predicted. The largest available online source for analytical interferences is the *Effects on Clinical Laboratory Tests* series, edited by Young and colleagues.<sup>293,294</sup> This database has compiled the largest body of evidence from the published literature and is the most extensive source of analytical interferences. For example, the database lists 307 results of potential drug interferences for creatinine.

### Supportive Medical Therapy

Medical contrast media are used during medical imaging procedures to enhance the contrast of organs and fluids. Iodine-based compounds (iohexol, iodixanol, and ioversol) are mostly used for the x-ray methods, while gadolinium contrast agents (ionic, neutral, albumin-bound, or polymeric) are typically used in magnetic resonance imaging.<sup>194</sup> Due to their specific chemical characteristics, effects of contrast media on laboratory tests cover a broad spectrum, some of which are listed below:

- gel displacement in blood collection tubes
- abnormal peak in electrophoresis
- chemical interference (contrast media interfere with the chemical reaction in various ways)
- chelating effect.

The type and magnitude of interference of the medical contrast media on laboratory tests depend on the agent, its concentration and half-life, mode of application, assay used, and instrument on which specific analyte is measured.

Blood collection from a patient who has recently received iodinated contrast media such as iohexol or iodixanol may cause some serious problems in sample processing, such as needle blockage, if serum/plasma gel separator tubes are used; iodine molecules have high density and can prevent proper formation of the barrier in the serum gel separator tubes. Upon centrifugation, instead of positioning between cell and serum/plasma layer, gel remains on the top and blocks the access of the needle to the serum/plasma.<sup>295–297</sup> This problem may cause considerable disruption in workflow and damage in highly automated laboratories, where visual inspection of the sample, after centrifugation, is not done.

Also, some specific interferences of the medical contrast media with chemistry tests have been observed. Iopromide is used as a contrast media agent in coronary angiography. A false increase in cTnI concentration measured by Opus Magnum reagent (Opus cTnI immunoassay system; Behring Diagnostics, Siemens) is detected if the sample is taken immediately after the procedure. This interference was deemed to be reagent-specific since no similar finding was observed when using cTn assay by a different manufacturer (ACCESS cTnI immunoassay; Beckman Coulter, Tokyo, Japan).<sup>298</sup>

Gadolinium (Gd) contrast agents act as chelators, and that seems to be the main mechanism of their interference with clinical laboratory assays. Gd contrast agents interfere with many clinical chemistry assays such as angiotensin-converting enzyme, total iron-binding capacity, zinc, magnesium, bilirubin, and creatinine.<sup>299–301</sup> Interestingly, calcium measurement is one of the parameters, which is most affected by

the presence of Gd-based compounds. Colorimetric calcium assays (o-cresol-phthalein complexone or methylthymol blue) are affected by the presence of gadodiamide, while ion selective electrode and inductively coupled plasma-atomic emission spectroscopy assays can reliably determine calcium concentration.<sup>302–304</sup>

Inductively coupled mass spectrometry (ICP-MS) is often used for the measurement of trace elements (see Chapter 44). There is ample evidence that Gd interferes with ICP-MS measurement of selenium (Se) and causes false increases in patients undergoing magnetic resonance imaging. Ryan and colleagues published a case report of a 30-year-old man with no history of Se exposure or toxicity symptoms who had lethal concentrations of plasma Se measured by ICP-MS.<sup>305</sup> This patient had undergone magnetic resonance imaging with a Gd contrast agent prior to measurement of Se concentration. It was postulated that the Gd<sup>2+</sup> isotope was causing interference with <sup>78</sup>Se<sup>+</sup>, the isotope used for the Se measurement, by having an identical mass-to-charge ratio. If this interference is not recognized, potential Se exposure can be suspected for these patients, and the patient could be misdiagnosed and mistreated. To avoid this interference, another Se isotope (<sup>82</sup>Se) that has a different mass-to-charge ratio, and thus is not affected by Gd, should be measured, or pure hydrogen should be used in the collision cell.<sup>306,307</sup>

Any chemical substance that absorbs light at the same wavelength as the peptide bond (200 nm) has the potential to interfere with protein detection in capillary electrophoresis (CE). Iodinated contrast media may mimic abnormal peaks in  $\alpha$ 2- or  $\beta$ -globulin fraction on CE, because their absorbance (200 to 275 nm) overlaps with that of protein absorbance (200 nm).<sup>308</sup> The magnitude of the interfering effect largely depends on the compound, its half-life, concentration, time of sampling, and patient kidney function. Due to their short plasma half-life, most contrast media are rapidly eliminated by the kidney in individuals with normal glomerular filtration rate. Therefore in most patients with normal kidney function, interference caused by iodinated contrast media is usually cleared from the serum/plasma 24 hours after the contrast infusion.<sup>309</sup>

To avoid possible interferences caused by contrast media, it is recommended to perform blood/urine sampling before administration of contrast media. If blood collection has to be done after the imaging, care should be taken to allow that sufficient time has elapsed from the application of the contrast media, taking into account the half-life of contrast used and patient renal function. Also, as a precautionary measure, whenever an additional narrow peak on the CE is detected, especially in a patient with impaired kidney function, care should be taken to exclude contrast media as a potential interfering cause. Another electrophoretic evaluation may be done after 24 hours to exclude or confirm this interference.<sup>310</sup>

### Natural Preparations

Today, it is becoming more common to use natural preparations for self-medication or supportive therapy. Patients consume herbal and other dietary products, but they fail to report the usage to their doctors or to the laboratory staff during phlebotomy.<sup>311</sup> The influence of these products on laboratory tests is not fully appreciated. Herbal medicines can cause direct interference with immunoassays due to cross-reactivity. Due to their structural similarity to the

tested analyte, active compounds that are present in herbal products can react with the antibody in the assay and, based on the structure and design of the immunoassay, result in both falsely increased or decreased analyte concentration. The concept of cross-reactivity in immunoassays is described in more detail in Chapter 26.

Another potential problem is the unexpected reactions since the exact content of these preparations is not always known.<sup>312</sup> Preparations used in Chinese medicine, like Chan Su, can contain some physiologically highly active molecules, like bufadienolides (bufalin, cinobufagin, and resibufogenin) that are extracted from the glands of Chinese toads. This preparation is used for the treatment of a variety of conditions, such as tonsillitis, sore throat, furuncle, and heart palpitations.<sup>313</sup> The structural similarity of bufadienolides and digoxin is responsible for both cardiotoxicity and interference with immunoassays. Ingestion of this medicine can cause both false-positive and false-negative results for digoxin, depending on the assay format. The most affected assays are those using polyclonal antibodies, like fluorescence polarization immunoassay or microparticle enzyme immunoassay; the assays using monoclonal antibodies are also susceptible to interference by bufadienolides.<sup>314,315</sup> Herbal supplements that are used widely throughout the world, like ginseng, can also interfere with digoxin measurement, even though some more recently introduced chemiluminescent microparticle assays seem to be free of such interference.<sup>316,317</sup> Labeling of herbal products may not accurately reflect their content, and adverse events or interactions attributed to a specific herb may be due to misidentification of plants, contamination of plants with pharmaceuticals or heavy metals, or quality control problems. For example, it has been recognized that some Chinese herbal remedies may contain steroids, with the potential of interfering with some assays and causing suppression of the hypothalamic-pituitary-adrenal axis.<sup>318</sup> Similarly, it is also well documented that some Ayurvedic herbal medicine products may be contaminated with lead, with the potential to cause toxicity.<sup>319</sup>

Besides being used for therapeutic purposes in patients with multiple sclerosis and metabolic disorders, biotin is increasingly being used as an over-the-counter dietary supplement to treat hair loss and problems with nails and skin.<sup>320</sup> Supplements are available in doses which are much higher than the recommended daily intake. This increasing trend in high-dose biotin supplementation poses a significant risk for interference in immunoassays. The body of evidence related to this interference is growing and numerous case reports and in vivo and in vitro studies are now available in the literature.<sup>321</sup> The mechanism of biotin interference depends on the format of the immunoassay. In sandwich immunoassays, biotin causes false negative results, whereas in competitive immunoassays it causes false positive results.<sup>322</sup>

Upon ingestion, biotin absorption is rather fast and plasma concentrations peak within 1 to 2 hours. Low-dose elimination half-life is 2 hours and biotin is cleared almost completely from the circulation of healthy individuals within 8 hours. Regular daily biotin supplementation leads to the accumulation of biotin in the body and steady state is usually achieved on the third day. In patients on high doses of biotin supplementation, half-life may be as long as 18 to 19 hours.<sup>323</sup>

In order to prevent unfavorable outcomes, in vitro diagnostics manufacturers were encouraged to conduct additional

interference studies and update their instructions for use with warnings about potential biotin interference and biotin threshold for interference, while reformulation of the assays was suggested as a long-term solution of the problem.<sup>324</sup>

Several approaches have been suggested when biotin interference is suspected<sup>323,325</sup>:

- to re-analyze the sample using a different, biotin-free immunoassay
- to dilute the sample, with the dedicated assay diluent
- to remove the excess of biotin by the use of streptavidin-coated beads
- to measure the concentration of biotin in the sample (if such method is available)
- to repeat the analysis after the wash-out period (minimum of 8 hours in patients taking 5 to 10 mg biotin daily, and  $\geq 72$  hours in patients who are prescribed a high-dose biotin therapy, i.e.,  $\geq 100$  mg/day). Patients with impaired kidney function would need a longer wash-out period.

### Accidental Exposures and Poisonings

In the case of accidental poisonings with herbs, household cleaning products, or any other exogenous compounds, interferences with laboratory test results can potentially delay the diagnostic procedures in acute patient care and cause harm to the patient. Due to their structural similarity to the digoxin molecule, cardiotonic glycosides can interfere with the digoxin measurement. Numerous cases of poisonings by cardiac glycoside-containing plants like lily of the valley (*Convallaria majalis*) or oleander (*Nerium oleander*) are reported. Ingestion of oleander is potentially fatal due to the cardiotoxicity of its active component, oleandrin. Positive and negative interferences of oleandrin and oleander extract on a Loci digoxin assay using the Vista 1500 analyzer have been reported.<sup>326</sup> Convallatoxin is a glycoside extracted from *Convallaria majalis*. Due to the significant cross-reactivity between convallatoxin and digoxin, the digoxin assay was suggested to be used as a screening tool for the detection of convallatoxin ingestion.<sup>327</sup>

Accidental intoxications often occur as a result of children ingesting cleaning products commonly found in the home. Miniature racing cars run on nitromethane, which has been shown to interfere with creatinine measurement, producing a falsely increased concentration.<sup>328–330</sup> Nitromethane is also used in racing cars to enhance combustion. Extreme creatinine concentration (8270  $\mu\text{mol/L}$ ; 93.6 mg/dL) without evident renal failure has been observed in a suicide attempt in which Blue Thunder fuel containing nitromethane was ingested.<sup>331</sup> In a Jaffe method, nitromethane also reacts with alkaline picrate and forms a red chromophore with absorbance similar to the creatinine picrate chromophore. This reaction causes falsely increased creatinine concentration. When an enzymatic assay is used (as opposed to Jaffe), creatinine can be accurately measured in the presence of nitromethane.<sup>332,333</sup>

Laboratory professionals should be alerted by any unexpected result and discuss it with the clinical staff. If the source of suspected interference cannot be determined, laboratories should try to involve manufacturers of the reagents to identify the potential interfering substance and quantify its effect.

### Sample Contaminants

Blood samples can be contaminated by several different exogenous substances during phlebotomy, sample handling, or

even test measurement. Several components of the blood collection systems (i.e., blood tube, needle, holder) can interfere with the measurement and influence laboratory results, including lubricants, needles, surfactants, separator gels, clot activators, and anticoagulants.<sup>133</sup>

**Tube additives.** Needles are made of stainless steel, which contains a minimum of 11% of chromium, to prevent oxidative needle corrosion. Due to the leaching of chromium into the sample, falsely increased chromium concentrations may be observed.<sup>334</sup> Lubricants like silicone oils and glycerol facilitate insertion and removal of the stoppers. Glycerol should not be used for the lubrication of stoppers in tubes that will be used for triglyceride measurement because glycerol interferes with most triglyceride assays.<sup>335</sup> Silicone-based lubricants are less likely to interfere with assays, although silicone can falsely increase ionized magnesium and T3 concentrations.<sup>336</sup> Additional peaks in mass spectrometry in the presence of silicone-based lubricants can interfere with interpretation of results.<sup>337</sup> Plastic tubes require clot activators to ensure rapid clot formation. Some clot activators based on silica particles affect the measurement of some analytes like lithium<sup>338</sup> and testosterone.<sup>339</sup> Bovine thrombin in some serum tubes causes falsely decreased parathyroid hormone concentration.<sup>340</sup> Silicone surfactants used to decrease nonspecific adsorption of components on tube walls may interfere with measurement of vitamin B12 and cancer antigen 15-3.<sup>336</sup>

**Separator gels.** Separator gels are used to ensure rapid and prolonged separation of serum and plasma from clotted blood and cells, respectively. Sample separation is enabled by the specific gravity of the gel (1.03–1.06 kg/L), its ability to undergo a temporary change in viscosity during centrifugation, and its ability to lodge between the packed cells and the top serum/plasma layer.<sup>134</sup>

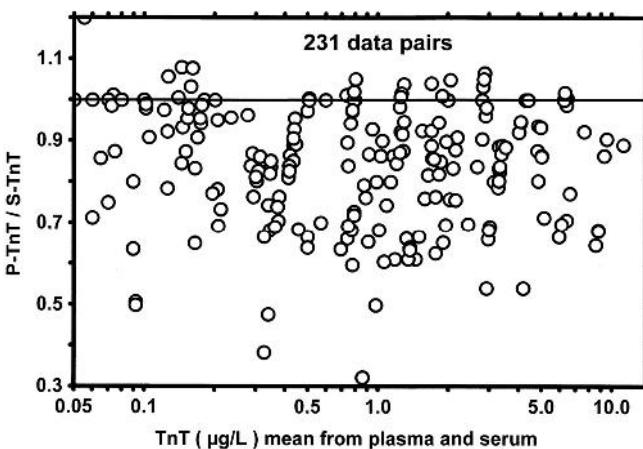
Hydrophobic compounds may bind to the gel, which is why tubes containing separator gels are not appropriate for some hydrophobic drugs and hormones such as the following<sup>341–344</sup>:

1. Drugs: phenytoin, phenobarbital, carbamazepine, tricyclic antidepressants, quinidine, lidocaine
2. Hormones: testosterone, estradiol, cortisol, fT4, fT3.

Due to differences in the gel composition among different manufacturers, it is quite possible that one manufacturer's gel tube may be used for a particular analyte but not another.

Moreover, if kept under improper storage conditions (time and temperature), the gel may degrade and release small particles or globules into the supernatant. These particles may affect instrument performance by interfering with the sample probe, coating the inner surface of the reaction cuvettes, and causing interference in immunoassays.<sup>133</sup> It is therefore important to strictly follow recommendations provided by the tube manufacturers on the appropriate storage and handling of gel tubes.

**Anticoagulants.** Although serum is still the predominant sample type in clinical chemistry testing worldwide, more laboratories are transitioning from the use of serum to plasma because of the shorter plasma separation time, greater plasma yield, and the ability to eliminate the problem of fibrin clots formation.<sup>345</sup> Particularly for urgent or stat testing, such as for cardiac markers, plasma has always been the preferred specimen.<sup>346</sup> Nevertheless, due to the interfering effect of some plasma additives (EDTA, heparin, sodium citrate) with some cardiac markers assays, serum is an acceptable alternative.



**FIGURE 5.12** Differences (ratio) in the concentration of cardiac troponin T (cTnT) in plasma versus serum (P-TnT/S-TnT) relative to the average troponin concentrations in plasma and serum (log transformed and presented as natural logarithm). (From Gerhardt W, Nordin G, Herbert AK, Burzell BL, Isaksson A, Gustavsson E, et al. Troponin T and I assays show decreased concentrations in heparin plasma compared with serum: Lower recoveries in early than in late phase of myocardial injury. *Clin Chem* 2000;46:817–21, with permission by American Association for Clinical Chemistry.)

Heparin interferes with numerous immunoassays by affecting the binding of antibody and antigen and thus affecting the rate of reaction.<sup>133</sup> Heparinized plasma has been documented to cause significant negative bias (up to 30%) in cTn results with some earlier-generation cTn assays from different manufacturers.<sup>347,348</sup> The observed bias did not correlate with the concentration of cTn (Fig. 5.12). Heparin interference occurs due to its negative charge and its binding to positively charged cTn. Binding of heparin and cTn leads to the conformational change of cTn and affects the antibody-antigen interaction. This interference was neutralized in the fourth-generation cTnT assay by adding cationic heparin blocking agent to the assay's mixture, although in certain cases there is still poor comparability between serum and plasma values.<sup>349</sup> For this reason, it is essential that the sample type for cTn testing remain consistent within a given patient.<sup>350</sup>

EDTA is a commonly used additive, especially in hematology and in some countries in endocrinology, because it offers increased stability of cells and analytes.<sup>351</sup> Most hormones (except adrenocorticotrophic hormone or ACTH) are stable for up to 5 days in EDTA plasma if they are kept refrigerated at 4 °C.<sup>352</sup> The main action of EDTA is chelation of cations (e.g., calcium, magnesium, and zinc). If EDTA is present in higher concentrations in the sample (when tubes are underfilled), its chelating activity is enhanced. This may lead to interferences in some chemiluminescence immunoassays that use conjugated ALP as a secondary enzyme in their reactions. For example, it has been shown that underfilling the EDTA tubes by half or more causes clinically significant bias (the reported concentration was <75% of the true value) in the measurement of intact parathyroid hormone with the DPC IMMULITE assay.<sup>353</sup>

Potassium oxalate also acts as calcium chelating anticoagulant and is often combined with antiglycolytic agents (sodium fluoride and sodium iodoacetate). As with EDTA, oxalate can also inhibit some enzymes (e.g., amylase, LD, ALP) by chelating bivalent cations that are necessary for their activity.<sup>134</sup>

**Intravenous fluids.** Contamination of blood sample with intravenous fluids is often caused by inadequate volume of discarded blood prior to sampling from a catheter or canula. This could cause a considerable effect on results of hematology, biochemistry, and coagulation tests, depending on the type of administered intravenous fluid. One example is a trisodium citrate solution (Citra-Lock solution), commonly used in dialysis catheters to prevent infection and thrombosis between dialysis sessions, which contains 35 times higher sodium concentration than serum. Sample contamination with a trisodium citrate solution may lead to gross hypernatremia with sodium results up to five times higher than the upper reference interval. To avoid the risk of contamination, first draw should be discarded when collecting a sample from a vascular access device (VAD).<sup>354</sup>

### Other Mechanisms of Interference

#### Formation of Fibrin Clots

Under optimal clotting conditions, serum is considered to be free of fibrin, fibrinogen, and cells, and it is a preferred matrix for most immunoassays. To allow complete clot formation, serum tubes should be allowed to stand for a minimum of 30 minutes. This delay, which is due to clotting time, is a major shortcoming for serum use, especially in emergency settings. However, with new tube types containing clot activators (thrombin-based clotting agent), serum clotting time is substantially reduced (on average <2.5 minutes) without compromising the sample quality and stability for most chemistry analytes.<sup>355,356</sup>

Blood from patients who are receiving heparin therapy may require a longer time to completely clot, and there is a greater likelihood for latent post-centrifugation clot formation. Insoluble fibrin has been found in both serum and plasma.<sup>357–360</sup> Insoluble fibrin, fibrin strands, and microclots formed as a result of delayed and latent clotting may affect instrument performance and cause interferences. Some analyzers have the ability to detect clots and flag such samples for rerun, but if this feature is not available on the instrument being used, clots may interfere with assays and cause erroneous results and unnecessary delays.

If fibrin is aspirated for analysis and goes undetected, there is a high likelihood of getting false-positive results for that sample.<sup>357</sup> This is manifested by duplicate measurement errors (i.e., unacceptable deviations in two measurements on the same sample). The false-positive result is caused by the nonspecific binding of insoluble fibrin strands present in the sample. Duplicate errors due to latent clotting or incomplete fibrin removal during centrifugation have been repeatedly reported for cTnI measurements (Fig. 5.13), and laboratories have been implementing different strategies to minimize the risk of reporting erroneous cTnI results due to fibrin clots.<sup>361,362</sup> A possible approach is to recentrifuge all positive cTnI samples (e.g., cTnI concentration >0.1 mg/L, measured on Dxi 800 analyzer, Beckman Coulter) at a high speed (6700 g for 5 min) and then repeat the analysis.<sup>363</sup> This approach, however, has been questioned by some authors and is not widely implemented.<sup>364</sup> Another approach employs a reflex rule to analyze samples in duplicate whenever the result is below or above some predefined value (e.g., cTnI concentration <0.04 or >5.00 µg/L). If a reflex measurement exceeds the limits of acceptance (>20%), an aliquot is recentrifuged and the sample is reanalyzed.<sup>357</sup>

Unfortunately, many instruments cannot consistently detect and appropriately respond to samples of questionable quality due to residual and latent fibrin strands. Until the quality of blood collection systems and instrument performance is improved, laboratory professionals must stringently monitor reported results and implement corrective strategies to minimize the risk of such preanalytical errors.

#### Interference Caused by Sample Carryover

Sample carryover in automated analyzers occurs as a result of inefficient probe and cuvette washing and the subsequent inability of an instrument to successfully remove any remnants of the sample or reagent. Due to improper washing, a certain amount of reagent or analyte can be transferred (carried) by the measuring system from one assay reaction to a subsequent reaction, thereby erroneously affecting test results. Those instruments that do not use disposable probes are more susceptible to carryover problems with some highly sensitive assays. Carryover is, of course, not unique to immunoassays and may occur in all assay types. Nevertheless, the effect of carryover is more pronounced in sensitive assays, such as highly sensitive immunoassays and in assays measuring analytes with a very wide range of concentrations.

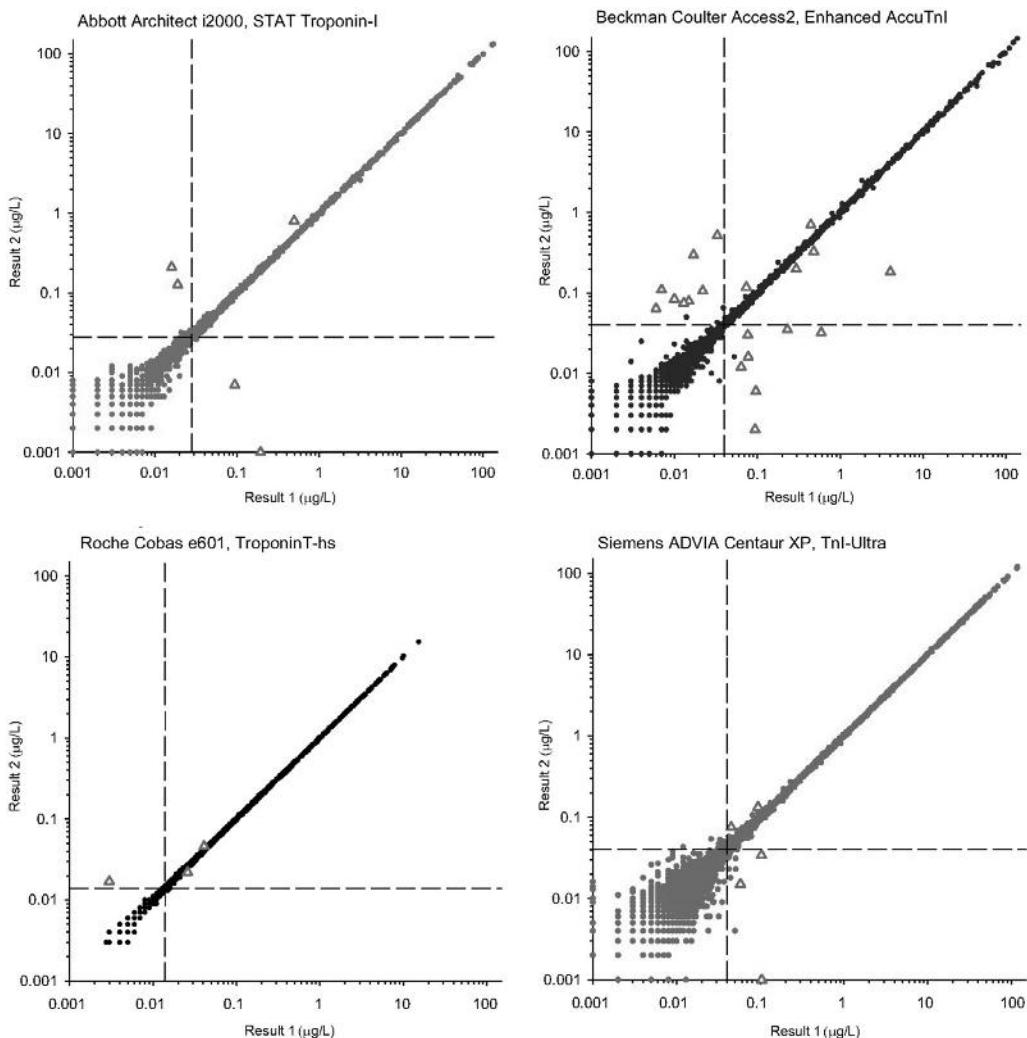
Sample carryover has been reported for the Enzymun-Test CEA assay on the ES-300 automated immunochemistry instrument (Boehringer-Mannheim, Germany).<sup>365</sup> Carryover was observed in samples being tested subsequent to those with extremely high CEA concentrations.

Though not confirmed by all participating laboratories in the study,<sup>366</sup> sample carryover as a cause of faulty results was also reported for cTnI on Beckman Access 2, UniCel DxI600, and UniCel DxI800 immunochemistry analyzers (Beckman Coulter, Inc.). Specimens with extremely high cTnI concentrations have been causing false increases that resulted in clinically significant changes in subsequent patients tested for cTnI.<sup>367,368</sup> To mitigate this problem, a reflex rule may be implemented to reanalyze cTnI in two subsequent specimens. Also, additional probe and cuvette washing may be performed after extremely high results that are associated with an increased risk of carryover.

Finally, it is highly recommended that laboratories perform testing for interferences due to sample carryover where evidence for the absence of an analyte is of clinical importance (i.e., at the limit of quantification), such as cardiac markers, tumor markers, and infectious disease (e.g., hepatitis) markers.

According to International Union of Pure and Applied Chemistry (IUPAC), sample carryover testing is performed by running one sample with high concentration of an analyte at least two times, followed by at least three runs of a sample with low concentration of that analyte.<sup>369</sup> If the instrument probe washing procedure is not done correctly, the results in the sample with low analyte concentration will be higher, and subsequent results will show a gradually decreasing pattern. The performance of the cuvette washing procedure is somewhat more difficult to assess because it may require multiple runs of a sample with high and low analyte concentrations (the exact number of runs depends on the number of cuvettes).<sup>173</sup>

According to CLSI, carryover testing for immunoassays is done by running four consecutive analyses of two samples at different concentrations (samples with extremely high analyte concentrations (A), followed by samples with very low



**FIGURE 5.13** Distribution of duplicate results for cardiac troponin measurement on four analyzers for a series of duplicate measurements in 2391 patient sera. Samples were analyzed with (1) Abbott Architect i2000SR analytical system with STAT Troponin-I reagent (Abbott Diagnostics); (2) Beckman Coulter Access2 analyzer with Enhanced AccuTnI reagent; (3) Roche Cobas e601 with TroponinT hs reagent (Roche Diagnostics); and (4) Siemens ADVIA Centaur XP with TnI-Ultra reagent (Siemens Healthcare Diagnostics). Dashed line marks the 99th percentile, as declared by the manufacturer. Red triangles represent outliers. (From Pretorius CJ, Dimeski G, O'Rourke PK, Marquart L, Tyack SA, Wilgen U, Ungerer JP. Outliers as a cause of false cardiac troponin results: Investigating the robustness of four contemporary assays. *Clin Chem* 2011;57:710–8, with permission by American Association for Clinical Chemistry.)

(B) concentrations for the same analyte).<sup>370</sup> The order of samples may be as follows:

A1, A2, A3, A4, B1, B2, B3, B4

According to the IUPAC protocol, carryover is expressed as the amount of analyte transferred from sample A2 to sample B1 (Fig. 5.14),

$$\text{Carryover, \%} = 100 \times (B_1 - B_3) / (A_2 - B_3)$$

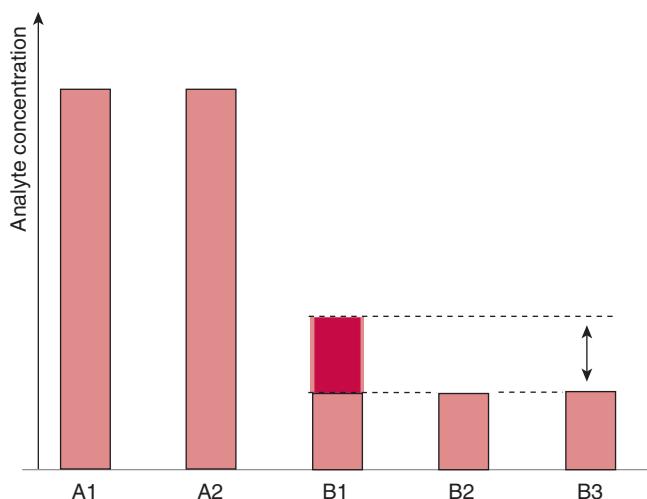
### Interference Caused by Antibiotics

Antibiotics are commonly added into reagents and buffer solutions to prevent microbial growth. Carryover from reagents containing gentamicin may cause spuriously high gentamicin results. Gentamicin is present in some diagnostic reagents

(glucose, urate, direct bilirubin, CK, ALT, AST, betahydroxybutyrate) as an antibacterial additive and is known to cause spuriously high gentamicin results on the Beckman Coulter AU480 analyzer (Beckman Coulter, Inc.) due to reagent carryover.<sup>371,372</sup> Therefore gentamicin measurement should be processed in a separate batch (i.e., before or after the measurement of any of the listed parameters) to prevent carryover.

### SPECIFIC CONSIDERATIONS

Besides the general features of preanalytical quality (i.e., patient and sample identification, controllable and noncontrollable variables), some specific aspects related to urine, saliva, and blood gas testing are of particular relevance to laboratory medicine. The following sections examine the preanalytical



**FIGURE 5.14** Carryover analysis. Samples A1 and A2 Have a high concentration of an analyte. Samples B1, B2, and B3 have a low concentration of the same analyte. The red rectangle depicts the amount of an analyte transferred from sample A2 to sample B1.

aspects of sample types other than serum and plasma, and other types of testing that deserve special preanalytical considerations. For additional discussion on body fluids, refer to Chapter 45.

### Urine

Urine composition is usually more variable and exhibits a broader reference interval compared to plasma.<sup>373</sup> Fluid intake, feeding, starvation, and muscular activity all influence the excreted amounts of urine constituents. Immobilization has also been shown to affect calcium excretion.<sup>374</sup> Calcium excretion increased 2.5-fold over 6 weeks of immobilization, reflecting bone degradation, and returned to normal when the patient was remobilized.<sup>29</sup>

Some of the most important preanalytical issues in the analysis of urine relate to patient preparation, choice of urine type, type of collection vessel, sample stability during transport and storage, sample homogeneity, and sample contamination.<sup>375,376</sup> These aspects are covered below.

### Types of Urine

Not every urine sample is suitable for every type of laboratory testing. Whereas some samples are well suited for hormone or drug testing, other types are appropriate for microbiological examinations. The possible types of urine samples and their intended uses are described in Table 5.12.

When collecting urine specimens and to prevent contamination with bacteria that normally reside on the skin, hygiene of the hands and genitalia is essential. Hands should be washed with soap, thoroughly rinsed with water, and dried. The positive predictive value of urinalysis for urinary tract infection is significantly higher and the contamination rate is much lower in urine specimens collected after proper hand and genital hygiene.<sup>377</sup> Thorough rinsing is also extremely important because baby soaps, for example, are known to cause significant interference with tetrahydrocannabinol (THC) immunoassays.<sup>378</sup> To prevent sample contamination by skin microorganisms, it is very important not to touch the inner surface of the cap and the container during collection.

**TABLE 5.12 Urine Specimens and Their Diagnostic Use**

Urine Type	Use
First morning urine	Qualitative urinalysis
Second morning urine	Proteins measurement, urinalysis
Timed urine (6–24 h)	Hormones and drugs measurement and detection
First void urine	Chlamydia
Midstream urine	Urinalysis, microbiological examination
Urine obtained through a catheter or sterile suprapubic aspiration	Exclusion and confirmation of urinary tract infection

Urine should be collected while holding the skin folds (labia) apart (females) or retracting the foreskin (uncircumcised men) during voiding.

First morning urine (also called “overnight” or “early morning” specimen) is collected immediately after arising from bed. In the cases of patients who suffer from insomnia or work the night shift, another 8-hour period can be used for first morning urine collection. It is important that the patient’s bladder is emptied immediately before going to sleep and that any amount of urine voided during the night is also collected and added to the first morning voided specimen.<sup>379</sup>

Timed urine specimens are collected at a specified time or in relation to some activity (before a meal, after a meal, before therapy, after exercise, etc.) during the 24-hour period. The exact time of the collection should always be indicated in the test report.

First void urine comprises the first portion of the urine (usually the first 15 to 50 mL) voided at any time of the day. It is collected after a patient has not urinated for at least 1 to 2 hours. The exact time depends on the sensitivity of the actual test method and is usually designated on the method insert sheet.

Midstream urine (also called a “clean catch specimen”) is a urine specimen collected during the middle of a urine flow after the urinary opening has been carefully cleaned. The first few drops of urine should go into the toilet. This prevents contamination with skin, vaginal, or urethral cells and bacteria. The midstream of the urine is collected, and once the container is filled, the rest of urine is voided into the toilet until the bladder is emptied.

Suprapubic aspiration and catheterization are procedures that are usually used in bacteriologic studies to obtain uncontaminated bladder urine. A catheter specimen is collected after inserting a catheter into the bladder through the urethra. A suprapubic specimen is collected by aspirating urine from the distended bladder through the abdominal wall. Both collection methods use sterile techniques and serve as an alternative to traditional methods of obtaining urine due to their high sensitivity and low risk of contamination.

### Patient Preparation and Sampling

European Urinalysis Guidelines provide recommendations on the proper patient preparation and sampling techniques to ensure the most accurate and reliable test results.<sup>373</sup> It is the responsibility of the laboratory to explain to a patient why a urine specimen needs to be collected and provide correct

information regarding optimal preparation and sampling procedure. Proper interpretation of laboratory reports is possible only if these conditions are fulfilled. Patients should be informed of all practical aspects of urine sampling and the possible effects of diet and fluid intake, diuresis, exercise, and other interferences.

The ideal container for any urine specimen is a wide-mouthed bottle of appropriate size. Containers should be clean, inert, leakproof, particle-free, and preferably made of a clear, disposable material that is inert with regard to urinary constituents.<sup>379</sup> The use of various “re-used” home items are discouraged, as they may lead to erroneous results due to the interaction of bottle material, detergent, or remnant bottle content with the urinary constituents.<sup>373</sup> If the urine is to be transported, the container should have a secure lid that will not leak during transport. If the urine is to be analyzed bacteriologically, the container must be sterile.

Nonspecific adsorption of various compounds (hormones, drugs, and proteins) to container surfaces in which the samples are collected, stored, or processed can cause analyte losses and affect the accuracy of quantification methods.<sup>380,381</sup> This is especially significant for protein analytes present in urine in low concentrations. For example, at low albumin concentrations, the adsorption of albumin to the surface of the urine container can cause a significant relative loss.<sup>382</sup> Analyte absorption to the vessel wall is avoided, or at least minimized, by the addition of some agents that either increase analyte solubility or minimize the interaction of analytes with the surface of a urine container.<sup>383</sup>

For pediatric and newborn patients, urine specimen collection bags with hypoallergenic skin adhesive should be used. First, the pubic and perineal areas should be cleaned with soap and water. Then the adhesive strip should be pressed all around the vagina or the bag fixed over the penis and the flaps pressed to the perineum.

### Transport and Storage of the Sample

Because some urine parameters have limited stability in unpreserved urine, temperature and time conditions during transport and storage of urine are of utmost importance for adequate sample quality. The stability of different particles in urine (Table 5.13) and test strip parameters (Table 5.14) is limited and decreases during prolonged storage and at higher temperatures.

Urine samples may be stored for up to 1 hour at room temperature and up to 4 hours if refrigerated without significant variation in the results of the physical, chemical, and morphologic analyses of particles.<sup>373,384</sup> With prolonged storage, urine particles lyse and bacteria grow. Bacterial growth causes the increase of urine pH. The stability of urine particles is lower in urine with alkaline pH (vegetarian diet), low relative density, and low osmolality (polyuria). Also, some preservatives (e.g., ethanol, polyethylene glycol, sodium fluoride, mercuric chloride, boric acid, formaldehyde- and formate-based solutions, etc.) may enhance the stability of urine particles.

As a general rule, if test strip analysis is performed within 2 to 4 hours from urine collection and urine has been kept at room temperature, preservatives for chemical constituents examined with test strips are not necessary. In unrefrigerated urine, bacterial growth may occur, leading to false-positive test strip results.

The addition of stabilizers inhibits the bacterial growth, metabolic processes, and degradation of urine analytes and

**TABLE 5.13 Stability of Urine Particles at Different Storage Conditions (Temperatures)**

Particle	-20 °C	4–8 °C	20–25 °C
Red blood cell	NA	1–4 h	1–24 h (>300 mOsmol/kg)
White blood cell	NA	1–4 h	1 h (pH >7.5) to 24 h (pH <6.5)
Acanthocytes	NA	2 days	1 day (>300 mOsmol/kg)
Casts	Not allowed	NA	2 days
Bacteria	NA	24 h	1–2 h
Epithelial cells	NA	NA	3 h

NA, Data not available.

From Delanghe J, Speeckaert M. Preanalytical requirements of urinalysis. *Biochem Med* 2014;24:89–104, with permission by Croatian Society of Medical Biochemistry and Laboratory Medicine.

**TABLE 5.14 Effect of Urine Storage Temperature on Chemical Constituents Examined by Urine Test Strips**

Analyte	4–8 °C	20–25 °C
Red blood cells	1–3 h	4–8 h
White blood cells	1 day	1 day
Proteins	NA	>2 h (unstable at pH >7.5)
Glucose	2 h	<2 h
Nitrites	8 h	4 days

NA, Data not available.

From Delanghe J, Speeckaert M. Preanalytical requirements of urinalysis. *Biochem Med* 2014;24:89–104, with permission by Croatian Society of Medical Biochemistry and Laboratory Medicine.

**TABLE 5.15 Stability of Urine Particles in the Most Commonly Used Urinary Preservatives**

Particle	Borate + Formiate + Sorbitol	10 mL/L Formaline + 0.15 mol/L NaCl	80 mL/L Ethanol + 20 g/L PEG
Red blood cell	Good	Not good	Not good
White blood cell	Very good	Very good	Very good
Casts	Good	Not good	Not good
Epithelial cells	Very good	Good	Not good
Bacteria	Very good	Very good	Good

From Delanghe J, Speeckaert M. Preanalytical requirements of urinalysis. *Biochem Med* 2014;24:89–104, with permission by Croatian Society of Medical Biochemistry and Laboratory Medicine.

particles. While the addition of some preservatives and the use of commercially available preservative tubes may be beneficial if sample transport time is more than 2 hours, it must be kept in mind that these preservatives can affect some chemical properties of the urine and particle integrity.<sup>385</sup> It is therefore important to select appropriate preservatives, taking into account potential effects on the analyte of interest. Table 5.15 provides data on the influence of the most

**TABLE 5.16 The Influence of Preservatives on the Stability of Some Chemical Constituents Examined by Urine Test Strips**

Analyte	Boric Acid	Formaldehyde	Hg Salts	Chloral Hexidine
Red blood cells	Stabilization	No stabilization	Stabilization	Limited stabilization
White blood cells	No stabilization	No stabilization	Limited stabilization	Limited stabilization
Proteins	No stabilization	Stabilization	Stabilization	Stabilization
Glucose	No stabilization	No stabilization	No stabilization	No stabilization
pH	No stabilization	No stabilization	No stabilization	No stabilization
Bacteria	Stabilization	Stabilization	Stabilization	Stabilization

From Delanghe J, Speeckaert, M. Preanalytical requirements of urinalysis. *Biochem Med* 2014;24:89–104 with permission by Croatian Society of Medical Biochemistry and Laboratory Medicine.

common preservatives on the stability of urine particles. Data about the influence of preservatives on the stability of some chemical constituents examined by urine test strips are provided in Table 5.16.

### POINTS TO REMEMBER

#### Urine

- The contamination rate is much lower in urine specimens collected after proper hand and genital hygiene.
- Laboratories should provide instructions to patients about reasons for urine collection, how to prepare for urine sampling (e.g., effects of diet and fluid intake, diuresis, exercise, and other interferents), and how to properly collect urine.
- The stability of different analytes in urine is limited and decreases during prolonged storage and at higher temperatures.
- Preservatives like ethanol, polyethylene glycol, sodium fluoride, mercuric chloride, boric acid, and formaldehyde- and formate-based solutions may enhance the stability of urine particles.
- The addition of urine stabilizers inhibits the bacterial growth, metabolic processes, and degradation of urine analytes and particles.

#### Saliva

Saliva is produced by the major salivary glands (parotid, submandibular, and sublingual glands) and by oral secretion of the mucus by hundreds of minor salivary glands. The major function of saliva is to provide oral protection by lubrication, digestion, and immune response.<sup>386</sup> Saliva is an attractive alternative to blood because it is collected noninvasively. As a diagnostic sample, saliva offers many advantages and some challenges (Box 5.1).

Today, saliva is increasingly used for assessing the genetic susceptibility to various conditions and for testing numerous analytes, as shown in Table 5.17.<sup>387,388</sup>

#### Types of Saliva Samples

The easiest way to collect saliva is to collect whole oral fluid (whole saliva). Whole saliva is representative of the oral milieu. However, depending on the intended aim of the sample collection and analyte to be tested, some other sample types are also possible. Various sampling techniques and devices

### BOX 5.1 Advantages and Challenges of Using Saliva as a Diagnostic Specimen

#### Advantages

- Rapid and easy collection by minimally trained individuals
- Sampling can be done by patients, at home, or outside hospital
- Multiple sample collection is possible
- Procedure is safe and painless for the patient
- Convenient in children, psychiatric patients, and stress research
- Availability
- Low cost associated with sampling (skilled staff not required)
- Convenient method for population screening programs
- Low risk of infections associated with sampling

#### Challenges

- Low analyte concentration
- Some analytes may be affected by circadian cycle
- Questionable recovery for some analytes
- Risk of contamination during collection
- Difficult sampling and low patient compliance (in small children).

are available. Whereas collection of whole oral fluid is an easy procedure, other sampling methods are more complicated, require trained staff, and are not commonly used.<sup>389</sup> See Chapter 4.

Whole saliva may be collected as both unstimulated and stimulated samples. Stimulation of saliva is obtained by oral movements (yawning, chewing gum) or by the use of a cotton ball soaked with citric acid. It must be noted that stimulated and unstimulated saliva are of different origins (produced by different salivary glands) and compositions (concentration of some analytes may vary) and are thus not equally suitable for all assays.<sup>390</sup> Unstimulated saliva is mostly produced by the submandibular glands, while stimulated saliva predominantly originates from the parotid glands.<sup>391</sup> Moreover, saliva stimulated by citric acid has a much lower pH (~3.0), and this may affect antigen antibody binding and interfere with immunoassays.<sup>392</sup>

Several commercially available devices may be used for the collection of saliva. Unstimulated whole saliva can be collected by passive drooling into a plastic vial or spitting into a collector vial. Passive drooling is considered by many to be the gold standard collection method for many analytes because it enables collection of a representative portion of saliva

**TABLE 5.17 Suitability of Different Specimen Types for Downstream Analyses**

	<b>Whole Blood</b>	<b>Plasma</b>	<b>Serum</b>	<b>Buffy Coat</b>	<b>ACP</b>	<b>DBS</b>	<b>Urine</b>	<b>Saliva</b>
Chemistry		✓	✓			✓	✓	✓
Hematology	✓					✓	✓	
Coagulation			✓					
Glucose	✓	✓				✓	✓	✓
Hemoglobin A <sub>1c</sub>	✓					✓		
Hormones		✓	✓			✓	✓	✓
Inflammation		✓	✓			✓	✓	✓
Cytokines			✓			✓	✓	✓
Vitamins			✓			✓		
Live cells				✓		✓		
Proteomics		✓	✓			✓	✓	✓
Metabolomics		✓	✓			✓	✓	✓
Genomics/germline DNA	✓			✓	✓	✓	✓	✓
ccfDNA			✓				✓	✓
Transcriptomics/mRNA	✓			✓		✓	✓	✓
miRNA (circulating)		✓	✓				✓	✓

ACP, All-cell pellet; ccfDNA, circulating cell-free DNA; DBS, dried blood spot; miRNA, microRNA.

From Ellervik C, Vaught J. Preanalytical variables affecting the integrity of human biospecimens in biobanking. *Clin Chem* 2015;61:914–34, with permission by American Association for Clinical Chemistry.

in the oral cavity.<sup>393</sup> Moreover, passive drooling is preferred over spitting because saliva collected by spitting is more likely to be contaminated with bacteria.<sup>394</sup>

Stimulated saliva can be obtained by adsorbing saliva with cotton balls, cotton swabs, or filter paper. Cotton is not an ideal collection material because of its unpleasant texture and because it may induce some variations in salivary immunoassays (some analytes are difficult to elute from cotton). To address this problem, some synthetic materials were made available, like inert polymers, polystyrene, rayon, and polyester.<sup>389</sup> Saliva is collected by placing the cotton ball in the patient's mouth. Placement of the ball or swab may vary depending on the analyte of interest. The ball is gently chewed for a couple of minutes and then placed in the vial. Saliva is obtained from the ball by centrifugation or expressed by a syringe. It must be emphasized that the composition of saliva collected by the use of various adsorbent devices may differ from the whole saliva because adsorbent devices mostly collect localized saliva. It is therefore important to be well aware of the collection method, its characteristics, and its performance.

Saliva collection in the elderly may be challenging, mostly due to xerostomia (dry mouth), which is quite common in the geriatric population, and low compliance. Challenges characteristic to infants and small children include the following<sup>392</sup>:

- Irregular cycles of sleep and periods of being awake
- Frequent napping (small babies often fall asleep, even during collection)
- Residue from liquids (e.g., milk in babies, juices in children) in the patient's mouth
- Food residue in an infant's mouth
- Anxiety about strangers (in children)
- Low compliance

### Patient Preparation for Saliva Sampling

The laboratory is responsible for providing information to the patient about the purpose of saliva collection, as well as

how to prepare for collection and how to collect saliva. Gloves should be worn during saliva collection to avoid sample contamination. Here are some important measures to minimize errors and ensure a high-quality saliva sample:

- If not otherwise decided by the requesting physician, saliva should be collected in the morning, preferably in the fasting state (the exact time of collection is important because some analytes may have diurnal variations).
- The patient should wash his or her face with water and soap and rinse it thoroughly to avoid contamination with facial creams and lotions.
- The patient should not consume alcohol 12 hours before the collection.
- The patient should not brush or floss his or her teeth at least 30 minutes before the collection.
- The patient should not have any dental work done 2 days before the collection (dental bleeding may affect test results).
- The patient should not ingest any food or drink (except water) within 30 minutes before the collection.
- The patient should not chew gum for at least 30 minutes before collection.
- Before the collection, the patient should rinse his or her mouth with water to remove any food particles. To avoid sample dilution with water, the sample should be collected 10 to 15 minutes after rinsing the mouth.

Any sample that is visibly contaminated with blood or food remnants should be rejected for analysis, and sample recollection should be requested.

### Storage of Saliva

After taking the sample, the saliva should be properly stored to prevent bacterial growth and protein degradation and to maintain sample integrity. Storage recommendation depends on the intended use of the saliva and duration of storage (Table 5.18). Some additives may also be used to inhibit the protease activity and retard bacterial growth (sodium azide).<sup>387</sup>

**TABLE 5.18 Recommendations for Different Storage Conditions for Saliva Specimens**

Storage Condition	Storage Time
Room temperature	30–90 min
+4 °C	3–6 h
–80 °C	Several months

From Chiappin S, Antonelli G, Gatti R, et al. Saliva specimen: a new laboratory tool for diagnostic and basic investigation. *Clin Chim Acta* 2007;383:30–40.

### POINTS TO REMEMBER

#### Saliva

- Saliva is increasingly used for assessing the genetic susceptibility to various conditions and for testing numerous analytes.
- The advantages of saliva as a diagnostic specimen are that it is inexpensive, convenient, rapid, easy, safe, and painless.
- The easiest way to collect saliva is to collect whole oral fluid (whole saliva), which is representative of the oral milieu.
- Whole saliva may be collected as either an unstimulated or a stimulated sample.
- Stimulation of saliva is obtained by oral movements (yawning, chewing gum) or by using a cotton ball soaked with citric acid.
- Stimulated and unstimulated saliva are of different origin and composition.

### Arterial Blood Gas Testing

Arterial blood gas testing requires attention for several reasons<sup>395</sup>:

- Blood gas testing is commonly requested in patients with a critical, life-threatening condition or who are experiencing some unexpected deterioration. Such patients may have a serious metabolic (acute complications of diabetes mellitus, drug intoxication) or respiratory disorder (respiratory failure, sepsis, or multiorgan failure) and need immediate medical intervention.
- Arterial blood sampling is an invasive procedure associated with a risk of complications such as bruising, bleeding, infections, and arterial thrombosis.
- Arterial blood samples have very limited stability. Due to the low biological variability of many blood gas parameters, allowable total error is quite low, and even small differences in serial measurements can be clinically meaningful.

International standards are available and may serve as a good resource for standardization of preanalytical steps respective to laboratory testing for blood gases, pH analysis, and ionized calcium.<sup>396–398</sup> See Chapters 37 and 50 for further discussion.

### Patient Condition

To ensure that test results reflect the actual condition of the patient, blood sampling should be done when a patient is in a stable, resting state. Furthermore, the exact time of the blood collection should always be recorded and reported with a test result. Any deviation from the steady state should be noted as a comment and accompany the test result in

order to allow proper interpretation of the results and patient management.

Relevant patient condition determinants (at the time of blood collection) include the following:

- Patient status (resting, exercising, crying (children), anxious)
- Ventilatory setting (spontaneous breathing or assisted mechanical ventilation)
- Mode of oxygen delivery (fraction of inspired oxygen ( $\text{FiO}_2$ ) through nasal cannula or Venturi mask)
- Respiratory rate (hyperventilation, hypoventilation)
- Body temperature

If the patient's condition is changing, a sufficient time should be allowed for the patient to stabilize.<sup>399,400</sup> For example, crying leads to a rapid decrease of oxygen saturation.<sup>401</sup> It has been shown that even a short walk or mild exercise may lead to a significant decrease in oxygen saturation in patients who are suffering from chronic obstructive pulmonary disease.<sup>402</sup> Hypoventilation and increasing body temperature are associated with an increase in ionized calcium and  $\text{pCO}_2$  and a decrease in pH.<sup>403</sup> Thus if patient temperature deviates from normal body temperature, information about that should accompany the test report to allow proper interpretation of results. Although blood gas instruments offer temperature-corrected values, their use is not recommended because currently data are not available to quantify the balance between oxygen delivery and oxygen demand at temperatures other than 37 °C.<sup>404</sup> If the temperature-adjusted results are reported anyway, it is absolutely mandatory that the report be clearly labeled as such and that the uncorrected values (at 37 °C) are also made available on the test report.<sup>405</sup>

If there has been any change in the ventilatory setting or mode of oxygen delivery, the patient should be left in a resting state to stabilize. For patients without pulmonary disease, a period of 3 to 5 minutes is usually enough to stabilize. However, in patients with pulmonary disease, this period is significantly longer. According to the CLSI C46-A2 standard for blood gas and pH analysis and related measurements, adequate time for most patients to reach a stable state following ventilatory changes is 20 to 30 minutes.<sup>397</sup>

### Sample Type

In healthy individuals, oxygen content in arterial blood is higher than in venous blood. The composition of arterial blood is constant throughout the body, whereas the composition of venous blood largely depends on the time of blood sampling, local and global circulatory conditions, and metabolic activity of the organ or tissue from which it carries blood to the heart. The major difference between arterial and venous blood is in their oxygen content. However, other parameters ( $\text{pCO}_2$ , pH) may also vary. The differences are more pronounced in conditions associated with compromised local or global circulation. Although venous blood is the specimen of choice for most routine laboratory tests, it is not the appropriate sample choice for the assessment of oxygen content in the blood. Arterial blood collected under anaerobic conditions is therefore the only acceptable sample type for an accurate evaluation of the gas exchange function of the lungs ( $\text{pO}_2$  and  $\text{pCO}_2$ ).<sup>405</sup>

If arterial blood is not available (e.g., neonates, small children, patients with burns) and during medical transport and prehospital critical care, a capillary sample is an acceptable

alternative. Capillary blood is obtained by puncturing the dermis layer of the skin and collecting it from the capillary beds running through the subcutaneous layer of the skin. Capillary blood is therefore a mixture of unknown proportions of the blood from the smallest veins (venules) and arteries (arterioles), the capillaries, and surrounding interstitial and intracellular fluids. Due to large differences in oxygen content between arterial and capillary blood, the results obtained from a capillary sample should be interpreted with extra caution.

Whereas capillary blood, if sampled properly, may accurately reflect arterial pCO<sub>2</sub> and pH over a wide range of values, unfortunately, it may never serve as an adequate substitute for arterial blood for accurate pO<sub>2</sub> measurement.<sup>406,407</sup> The difference in oxygen content between arterial and capillary blood is even more pronounced in hypotensive patients and in patients with an increase of arterial pO<sub>2</sub>.<sup>408,409</sup> Moreover, capillary blood sampling is not recommended in patients with circulatory shock or with poorly perfused (cyanotic), infected, inflamed, swollen, or edematous tissues. Capillary blood should be collected using an arterialization technique by warming the skin to 40 to 45 °C with a warm towel or by using a vasodilating cream containing, for example, methyl nicotinate or capsaicin. Arterialization increases the blood flow through the capillary beds and thus the proportion of arterial blood relative to venous blood in the capillary sample. An earlobe is a better sampling site than a fingertip because the blood sampled from an arterialized earlobe better reflects arterial blood values. Arterialized earlobe capillary blood gas sampling is widely used across primary and secondary health care settings, especially in patients requiring frequent monitoring of blood gas parameters.<sup>410</sup> To circumvent difficulties in capillary blood collection from an earlobe and minimize room air contamination, some special devices have been designed and are currently being evaluated.<sup>411</sup> These devices have been primarily designed to improve medical emergency management of patients in some extreme environments (e.g., space, high altitudes), but they could also become routinely used in the near future.

### Anticoagulants

The recommended anticoagulant for arterial blood gas and ionized calcium testing is lyophilized balanced Li-heparin. According to the CLSI C46-A2 standard on blood gas and pH analysis and related measurements, the final heparin concentration in the sample should be 20 IU/mL blood. Because the pH of heparin is 7.0 (slightly acidic) and its pO<sub>2</sub> and pCO<sub>2</sub> values are near room air values, the excess of heparin in the sample can alter sample pH, pO<sub>2</sub>, and pCO<sub>2</sub>.<sup>412</sup>

**Why use balanced heparin?** Heparin is traditionally used to prevent blood sample coagulation. However, heparin is negatively charged and binds various cations (e.g., Ca<sup>++</sup>, Na<sup>+</sup>, K<sup>+</sup>) in a dose-dependent manner. This may cause underestimation of electrolyte concentration. To prevent such direct binding of heparin and cations from the sample, balanced heparin was introduced. The binding sites of balanced heparin are presaturated with calcium.

**Why use lyophilized heparin?** Most commercially available dedicated syringes for arterial blood sampling contain spray-dried balanced heparin as an anticoagulant. Syringes with liquid heparin are also available. Whereas liquid heparin

enables better sample mixing, it may introduce sample dilution in cases of incomplete draw. Using ordinary syringes (without heparin) and flushing them before use are strongly discouraged. Flushing the syringe with liquid heparin causes sample contamination with heparin and sample dilution, resulting in significant differences among blood gas parameters (Fig. 5.15).<sup>413</sup>

### Sample Contamination

Sample contamination may substantially affect blood sample quality and cause significant bias. Arterial blood gas samples are most commonly contaminated with liquid heparin (discussed above), venous blood, or air bubbles. Contamination of an arterial sample with venous blood occurs if a vein is accidentally punctured. This may happen if the needle is not correctly positioned during arterial blood sampling. In an arterial blood sample contaminated with venous blood, pO<sub>2</sub> and sO<sub>2</sub> may be falsely decreased, while pCO<sub>2</sub> may be falsely increased.

When making an arterial puncture, the needle should be inserted at a 30 to 45-degree angle. Moreover, it is also recommended that short-beveled needles be used because they are much easier to position inside the artery. Self-filling dedicated syringes are also highly recommended. These syringes fill more quickly and easily when a needle is puncturing an artery instead of a vein as a result of the difference in blood pressure between the vein and the artery.<sup>414</sup>

The aspiration of air during arterial blood sampling or bubble formation can result in significant changes in the concentration of some blood gas parameters ( $\uparrow$ pH,  $\uparrow$ pO<sub>2</sub>,  $\uparrow$ sO<sub>2</sub>,  $\downarrow$ pCO<sub>2</sub>). The exchange between the air bubble and the arterial blood sample is rapid. It starts immediately and becomes significant after only 1 to 2 minutes. The exchange rate does not depend on the size of the bubble.<sup>415</sup> The longer the delay between blood sampling and sample analysis, the greater the effect of the contamination with atmospheric air and deviation from the true patient values. It should be noted that even a bubble as small as 1% of the total sample volume may cause significant changes in the oxygen content of the specimen.

Contamination with air bubbles can be prevented by visual inspection of the specimen immediately after blood sampling. If air bubbles are present in the sample, they should be expelled as soon as possible and certainly prior to the sample mixing. If there is a visible froth in the sample, such samples should not be analyzed because froth may contain a significant amount of atmospheric air.

The degree of contamination also depends on sample agitation during transport. The increment in pO<sub>2</sub> in the presence of air bubbles in the sample is even more pronounced if the samples are transported by pneumatic tubes due to the exaggerated oxygen movement between the blood sample and ambient air caused by sample turbulences in the pneumatic tube.<sup>416-418</sup> Pneumatic tubes are thus not recommended for the transport of arterial blood samples.

The use of blood gas syringes with a vented mechanism is also recommended to avoid sample contamination with air. Once such a syringe has been filled up to the dedicated volume, the vent allows the air to be pushed out from the syringe. After the air has been pushed out, the vent is closed, preventing the subsequent contamination of the sample with atmospheric air.

### Hemolysis

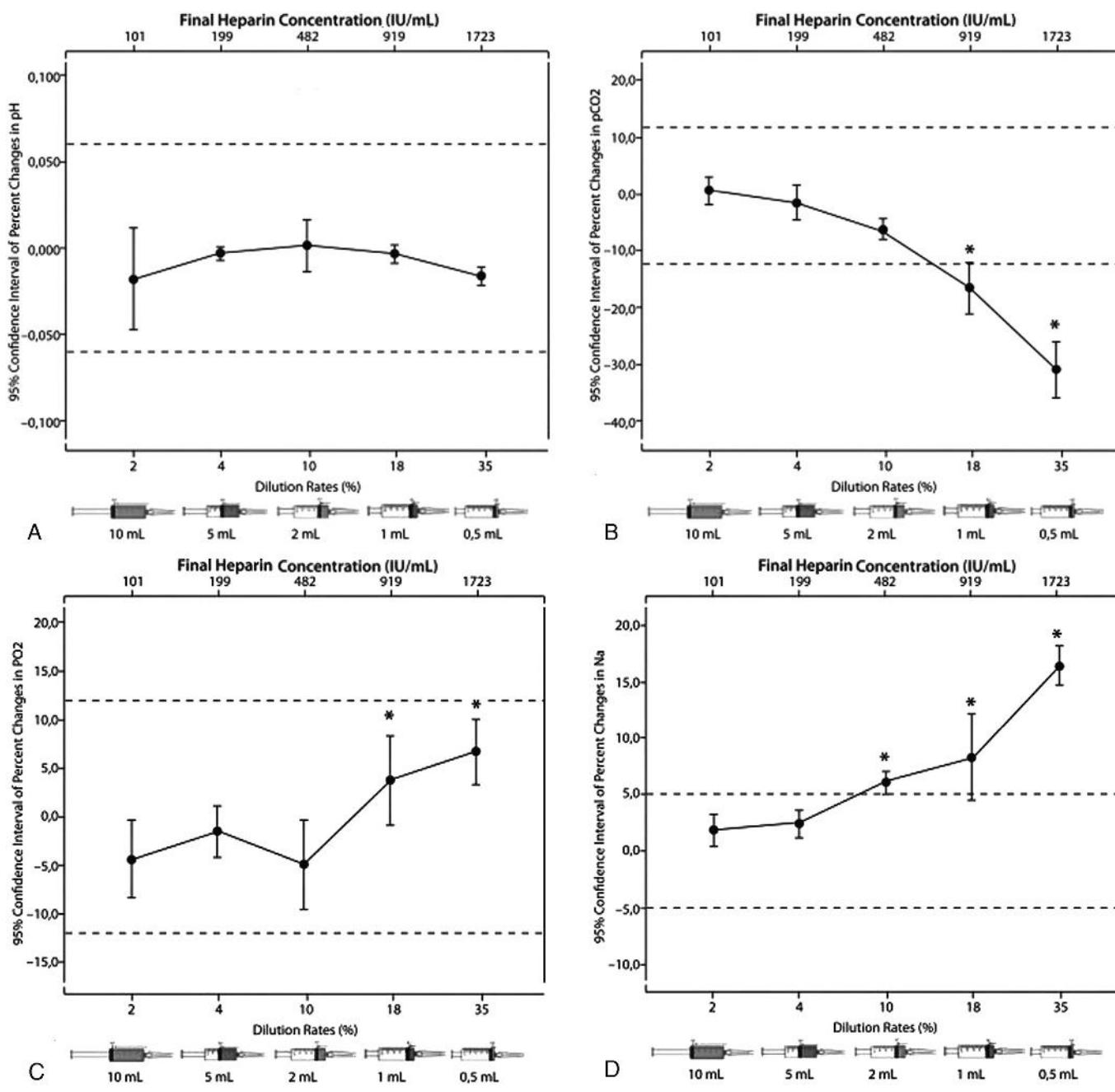
Sample hemolysis is another big concern related to blood gas testing. Although sample hemolysis is difficult (or almost impossible) to assess in arterial blood samples, it has been demonstrated that a significant proportion (up to 4%) of arterial blood samples are hemolyzed.<sup>419,420</sup> Hemolysis leads to a significant decrease in  $pO_2$  and an increase in  $pCO_2$ .<sup>421</sup> Electrolyte concentrations (potassium, calcium) are also dramatically affected by hemolysis. Hemolysis can occur during sampling and due to inadequate transport and storage conditions.

Because sample hemolysis cannot be detected in arterial blood gas samples, all necessary precautions should be taken

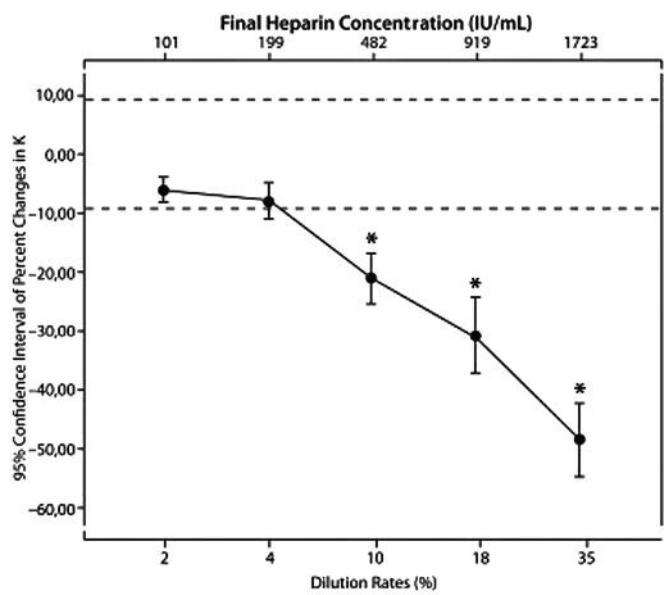
to minimize the risk of sample hemolysis. The following conditions should be avoided to minimize the risk of hemolysis:

- Vigorous mixing. Sample mixing should be done gently.
- Any source of sample turbulence
- Pneumatic tube systems
- High force during sample aspiration
- Cooling the sample directly on ice cubes. If the sample needs to be cooled, an ice slurry should be used instead.

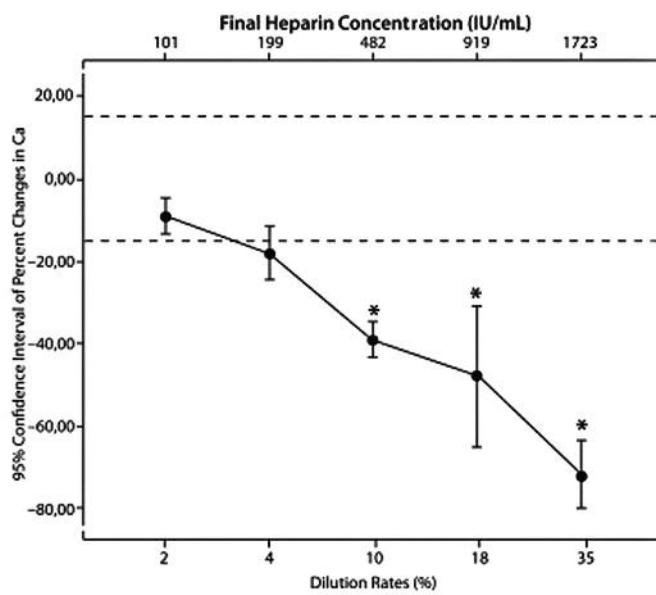
If capillary blood is collected, excessive pressure ("milking") should be avoided. Sample milking leads to significant hemolysis and contamination of the sample with surrounding tissue fluid (see Chapter 4). If milking is applied, many parameters in the sample will substantially deviate from the



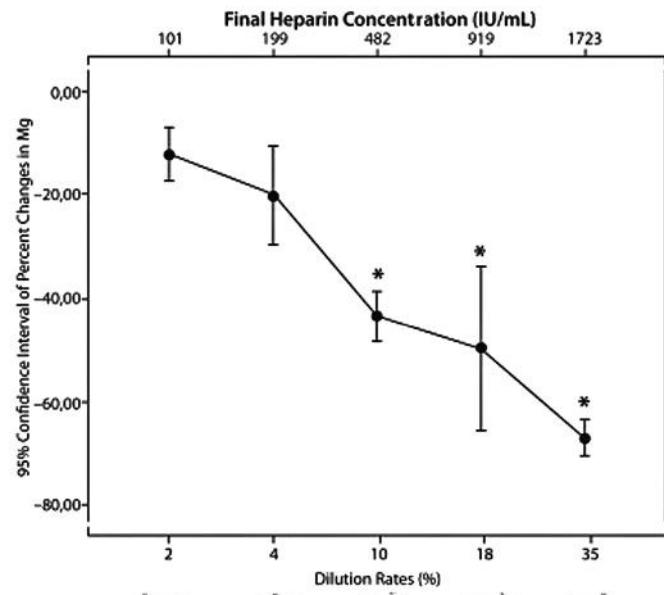
**FIGURE 5.15** The changes in pH (A) and the percent changes in  $pCO_2$  (B),  $pO_2$  (C),  $Na^+$  (D),  $K^+$



E 10 mL 5 mL 2 mL 1 mL 0,5 mL



F 10 mL 5 mL 2 mL 1 mL 0,5 mL



G 10 mL 5 mL 2 mL 1 mL 0,5 mL

true values (Fig. 5.16). Possible difficulties during capillary blood sampling should always be recorded and reported with the test results to enable proper interpretation of test results by the clinician.

### Sample Mixing

Blood samples must be properly mixed to prevent clot formation, promote heparin dissolution, and ensure that blood cells are uniformly suspended in the sample. Blood samples should be mixed immediately after the sampling but only after visible air bubbles have been expelled. Mixing should be done gently to avoid hemolysis. Samples can be mixed manually and automatically. Manual sample mixing is done by gently inverting the syringe several times and rolling it between the palms (Fig. 5.17). If manual mixing

**FIGURE 5.15 cont'd** (E), iCa+2 (F), and iMg+2 (G) at different dilutions and final heparin concentrations. The horizontal dotted line indicates the acceptable total analytical error (TEa) limits according to RiliBAK. \* $P < .05$  for differences from the true values (full sampling done to the full volume without dilution). (From Kume T, Sisman AR, Solak A, Tuglu B, Çirkoglu B, Çoker C. The effects of different syringe volume, needle size and sample volume on blood gas analysis in syringes washed with heparin. *Biochem Med* 2012;22:189–201, with permission from Croatian Society of Medical Biochemistry and Laboratory Medicine.)

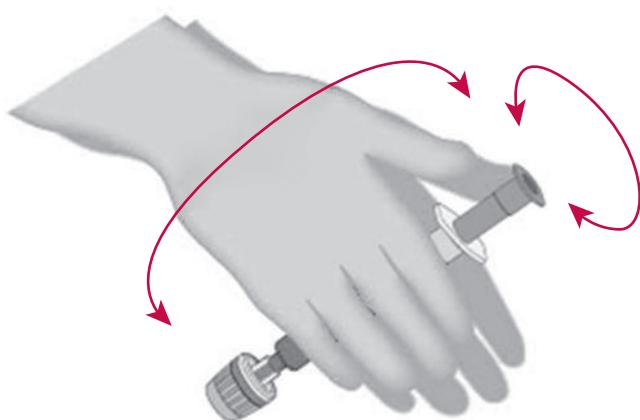
is not performed properly, the sample is unsuitable for analysis. The automatic arterial sample mixing is done with the use of a small metal ball located in the syringe barrel. The ball in the sample is moved through the sample by the force of an external magnet.

A clotted sample will cause analyzer malfunction. Moreover, if a clotted sample is analyzed, potassium concentration will be increased due to the efflux of the potassium from the platelets during blood clotting.

If the analysis is not done immediately and the sample needs to be transported to another location, the sample must be mixed again immediately prior to analysis. This is necessary to obtain a homogeneous sample and to ensure accurate test results. Mixing time depends on the time span between the sample collection and analysis. A shorter mixing time

Without milking		Milking applied	
ELECTROLYTES		ELECTROLYTES	
Na <sup>+</sup>	140.1	Na <sup>+</sup>	137.1
K <sup>+</sup>	3.76	K <sup>+</sup>	4.12
Ca <sup>++</sup>	0.97 ↓	Ca <sup>++</sup>	0.70 ↓
Ca <sup>++</sup> (7.4)	0.99	Ca <sup>++</sup> (7.4)	0.71
Cl <sup>-</sup>	104	Cl <sup>-</sup>	101

**FIGURE 5.16** Example of the effect of excessive repetitive pressure (the so-called sample “milking”) on sample hemolysis and sample contamination with tissue fluid. These two samples were obtained from the same patient in the resting state within 2 minutes. Sample without milking was obtained after an arterilization with a warm towel, while the other sample was obtained by the use of excessive finger pressure. (Laboratory data from the Clinical Institute of Chemistry, Clinical Hospital Center “Sestre milosrdnice,” Zagreb, Croatia. Data used with patient consent.)



**FIGURE 5.17** Manual sample mixing is done by gently inverting the syringe several times and rolling it between the palms. If manual sample mixing is not done properly, the sample is unsuitable for analysis.

(<1 minute) is acceptable if the time span from sampling to analysis is no longer than a couple of minutes. If longer delays occur between sampling and analysis, longer mixing intervals are required. The longer the time span, the longer the mixing time required. In samples that have been left to stand for 20 to 30 minutes, the homogeneity of the samples can be achieved by continuous mixing for at least 2 minutes.<sup>422</sup>

### Sample Transport

Arterial blood samples should be transported by hand and at room temperature. As already mentioned, vigorous movement during sample transport should be avoided. Time is a critical variable in blood gas testing. It is important to avoid delays and to analyze the sample as soon as possible. Prolonged storage prior to analysis introduces significant bias due to cell metabolism and oxygen exchange between the sample and the atmosphere. Moreover, prolonged storage may also cause spurious results due to blood sedimentation. To avoid that, samples should be visually inspected and properly mixed to homogenize the blood inside the syringe.

As a general rule, samples drawn in plastic syringes should be analyzed immediately. If analysis is delayed, the samples should be stored in glass syringes.<sup>423</sup> Plastic syringes should not be cooled because plastic molecules contract when cooled to 0 to 4 °C. Contraction of plastic molecules creates pores in the syringe wall through which oxygen easily diffuses. Because carbon dioxide is a much larger molecule than oxygen, it cannot diffuse through the syringe wall (Fig. 5.18). The deviation in an inappropriately cooled plastic syringe is therefore greatest for oxygen and oxygen-related parameters.

According to the CLSI C46-A2 standard (blood gas and pH analysis and related measurements), samples should be transported by hand in a plastic syringe at room temperature and analyzed within 30 minutes of collection. In cases when expected delivery time is longer than 30 minutes, glass syringes should be used and the sample should be transported on ice to reduce the rate of metabolism and exchange of gases between the sample and the ambient air.<sup>397</sup> For more information on blood gas measurement and interpretation of results, refer to Chapter 37.

### Hemostasis Testing

Many preanalytical variables may affect the results of routine coagulation testing, including general requirements regarding adequate test selection and ordering, patient preparation, patient identification, as well as all of the well-known potential sources of variation regarding sampling technique (fasting state, length of the venous stasis, order of draw, sampling from a catheter, sample handling (centrifugation), transport, and storage prior to analysis).<sup>424–431</sup> Some specific considerations related to hemostasis testing are explained further in the text.

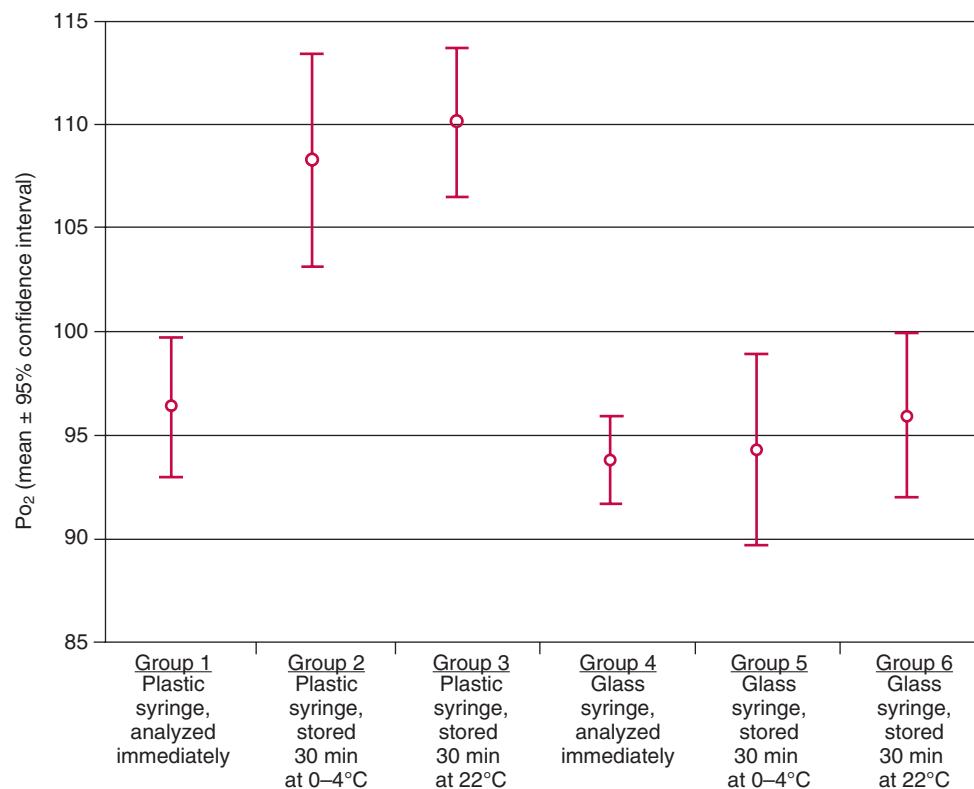
One of the major preanalytical issues in hemostasis testing is clotted samples. Samples that contain visible clots should not be accepted for analysis and must be rejected. To prevent clot formation, some precautionary measures should be taken during blood sampling, handling, and transport; the following errors should be avoided:

- Inappropriate anticoagulant
- Too slow blood flow (during blood sampling)
- Transferring to citrated tube after sample collection with a syringe
- Underfilled tubes
- Prolonged use of a tourniquet
- Considerable manipulation of the vein by the needle
- Incomplete mixing

Clot formation induces the activation of platelets and clotting factors and allows the release of granules from the platelets; these factors may lead to false results in coagulation assays. It is very important to keep in mind that even small clots that are invisible to the human eye may significantly impact coagulation assays. Longer venous stasis should be avoided because it results in hemoconcentration, activation of fibrinolysis, and other changes, such as an increase in fibrinogen and factors VII, VIII, and XII.<sup>432</sup> Clotted samples must be rejected.

### Choice of Anticoagulant

Historically, tubes made of glass have been used to collect blood samples. Since the coagulation cascade can be activated by contact with glass surface, these tubes must be siliconized to prevent glass-induced coagulation activation.<sup>433</sup> In the past



**FIGURE 5.18** Plastic syringes should not be cooled because plastic molecules contract when cooled to 0 to 4 °C. Contraction of plastic molecules creates pores in the syringe wall, through which oxygen easily diffuses. If samples must be cooled, glass syringes should be used to avoid the exchange of oxygen through the syringe wall. This graph shows the differences in  $\text{PO}_2$  values in arterial blood samples stored at different temperatures for different periods of time. (From Knowles TP, Mullin RA, Hunter JA, Douce FH. Effects of syringe material, sample storage time, and temperature on blood gases and oxygen saturation in arterialized human blood samples. *Respir Care* 2006;51:732–36, with permission by Dallas, Texas: Daedalus Enterprises for the American Association for Respiratory Therapy.)

decade blood collection tubes made of a variety of plastic materials have started to replace glass tubes, thus providing improved safety due to their increased shock resistance and tolerance to higher centrifugation speed; evidence showed equality in performance between the two types of tubes.<sup>434</sup> However, such plastic tubes should contain polypropylene as a nonactivating material.<sup>435,436</sup> Samples for hemostasis testing should be anticoagulated with 3.2% sodium citrate, although 3.8% may also be acceptable. It is important that the same concentration of sodium citrate be used within the laboratory because clotting times may be longer in 3.8% than in 3.2% sodium citrate.<sup>435</sup> As 3.8% sodium citrate binds more assay-added calcium than 3.2%, it has been reported that PT and aPTT may be overestimated in 3.8% sodium citrate, whereas fibrinogen is underestimated compared to values obtained in 3.2% citrated samples.<sup>437,438</sup> Additionally, it is important to emphasize that only 3.2% trisodium citrate is used for thromboplastin international sensitivity index (ISI) assignment that is also recommended by Scientific and Standardization Committee (SSC) of the International Society on Thrombosis and Hemostasis, so the anticoagulant of choice should preferably be 3.2% sodium citrate.

Other anticoagulants (e.g., EDTA or heparin) are not acceptable for hemostasis testing because they can lead to erroneous results and can cause clinically significant errors. For

some analyses, especially platelet function assays, buffered citrate solution or other anticoagulants, such as lepirudin or synthetic inhibitors of thrombin and factor Xa, are used.<sup>439–441</sup>

### Blood/Anticoagulant Ratio

The required blood to anticoagulant ratio is 9:1; it is therefore very important that tubes for coagulation assays be filled to the mark noted on the tube.<sup>435</sup> As filling volume decreases, clotting time in seconds tend to increase. This effect is more pronounced in tubes with 3.8% sodium citrate, compared to 3.2%. Acceptable deviation is maximum 10% of the total volume; filling the tube over 10% (overfilling) or less than 10% (underfilling) of the designated volume is strictly discouraged because this can introduce bias into test results. Although there is a great deal of evidence in the literature that demonstrates that unbiased results can be observed in 3.2% citrate tubes filled as low as 60%, especially in samples with low hematocrit value,<sup>442–444</sup> the general rule of thumb is to reject citrate anticoagulant tube filled below 90%.<sup>435,445</sup> Laboratories should have preanalytical procedures in place for the rejection of over- or underfilled tubes.

### Order of Draw

In the past, a discard tube was recommended whenever the coagulation tube was the first tube.<sup>446</sup> However, more recent

evidence suggests that a “discard” tube may not be necessary<sup>447–451</sup>; CLSI stopped that recommendation, except in selected circumstances.

When the coagulation tube is collected as the first or the only tube:

- and a straight needle is used for blood collection, no discard tube is needed
- and a winged blood collection set (butterfly devices) is used, a discard tube must be collected to prevent underfilling<sup>435,452</sup>

If a safe venipuncture is not possible, blood specimens have to be obtained through a VAD. If intravenous catheter systems are used for blood sampling, it is recommended to flush the central catheter with saline and to discard the first 5 mL of blood or the catheter dead space volume corresponding to six times the line volume prior to tube collection for coagulation testing.<sup>435,436</sup> If blood is obtained from a normal saline lock, two dead space volumes of the catheter and extension set should be discarded.<sup>435</sup> If the catheter might be contaminated with heparin due to heparin infusion or flushing with heparin-containing fluid, the option of heparin neutralization of the sample should be considered.<sup>435</sup>

The order of draw should be followed without exceptions during every blood collection.

### Mixing

Mixing of samples is extremely important for adequate coagulation testing. Samples must be promptly mixed to avoid in vitro clot formation. Tubes should be mixed by gently inverting the tubes (at 180 degrees) several times. For proper mixing, instructions from the tube manufacturer should be followed. Vigorous mixing and shaking of the tubes are discouraged because this may lead to sample hemolysis; the activation of platelets and coagulation factors, resulting in false shortening of clotting times; and even possibly a false increase in clotting factor activity.<sup>425,427,430</sup>

### Transport

Following collection, citrated samples should ideally be transported to the laboratory immediately and at room temperature but no later than within 1 hour from blood draw.<sup>424</sup> Transport of samples by pneumatic tubes is still under debate. While some claim that pneumatic tube transport is acceptable for the transport of samples for routine hematology and coagulation parameters,<sup>428,453</sup> others argue that samples transported by pneumatic tube are not suitable for platelet aggregation assays.<sup>454</sup> It is therefore recommended that each institution verifies the acceptability of its tube transport systems by conducting a comparison study.

### Centrifugation

Whole blood coagulation assays should be performed within 4 hours after blood sampling. For platelet function assays, samples should rest (at room temperature) for 30 minutes before analysis. Centrifugation is normally performed at 1500 g at room temperature for 10 to 15 minutes to obtain platelet-poor plasma.<sup>435</sup> Although there are many studies that aim to optimize centrifugation of coagulation samples at higher speeds with shorter centrifugation times,<sup>455–457</sup> these shortened protocols are not recommended because they may induce hemolysis and activation of platelets. Moreover, centrifuge breaks should also be avoided to prevent remixing of samples.<sup>427,458</sup>

The preparation of PRP for platelet function assays requires special care, and centrifugation speeds and duration need to be carefully optimized to ensure optimal results. Generally, centrifugation is performed at 200 to 250 g for 10 minutes without application of a rotor brake.<sup>459,460</sup>

### Stability and Storage

Blood samples for coagulation testing must be kept at room temperature (20 to 25 °C) until analysis. Storage at lower temperatures or on ice is discouraged because cold temperature activates some coagulation factors. For example, in citrated whole blood stored in an ice bath or refrigerated (2 to 8 °C), the activation of platelets, activation of factor VII, and significant time-dependent loss of both FVIII and von Willebrand factor (VWF) will occur.<sup>424</sup> PT and aPTT should generally be performed using fresh plasma within 4 hours after blood sampling and stored at room temperature.<sup>461</sup> If the centrifuged plasma is left to sit on the blood cells, this may result in shortening of PT and prolongation of aPTT.<sup>462</sup> Stability data for screening coagulation test according to CLSI *Collection, Transport, and Processing of Blood Specimens for Testing Plasma-Based Coagulation Assays and Molecular Hemostasis Assays* H21-A5,<sup>435</sup> and CLSI *Quantitative d-dimer for the Exclusion of Venous Thromboembolic Disease* H59-A<sup>463</sup> guidelines are listed in Table 5.19.

Stability data for all coagulation factors can be found in the literature.<sup>435,464</sup> Although sample stability for coagulation assays has been extensively investigated by ample studies, and results are discrepant from aforementioned guidelines,<sup>465–468</sup> the samples should be routinely processed according to current CLSI H21-A5 guidelines.<sup>435</sup> If possible, all coagulation analyses should be performed using fresh material; freezing of samples should be an exception.<sup>469</sup> Previously frozen samples should be rapidly thawed in a 37 °C water bath for 5 to 10 minutes or until completely thawed. The samples should not be allowed to sit in the water bath for an extended period of time due to some thermolabile factors (FVIII, FV). Once thawed, samples should be thoroughly and adequately mixed prior to testing.<sup>470,471</sup>

For more information on coagulation and platelet function tests and interpretation of results, refer to Chapters 80 and 81.

### Hematology

CBC is the most prevalent laboratory test, even across different continents, and regardless of the country's income.<sup>472</sup> However, even with the high prevalence of CBC determinations, reported frequency of preanalytical errors in the hematology laboratory is low, a mere ~1%.<sup>137,473–477</sup> Errors in patient identification and specimen labelling are pan-disciplinary and not related specifically to hematology testing.<sup>478</sup> Of the preanalytical errors related to samples for hematology testing, some of the most commonly reported problems are related to the clotting of samples and samples with an inappropriate blood-to-anticoagulant ratio (Table 5.20).

Some of the most common variables leading to an unsatisfactory sample for hematology testing are listed further in the text.

### Choice of Anticoagulant

Being efficacious for preserving cellular morphology, EDTA is the anticoagulant of choice for hematology testing. EDTA

**TABLE 5.19 Stability data for screening coagulation tests and d-dimer<sup>435,463</sup>**

Assay	STORED AS WHOLE BLOOD			PROCESSED AND PLASMA ALIQUOTED				
	Room Temp	Refrigerated	Frozen	Room Temp	Refrigerated	Frozen	-20 °C	Frozen ≥ -70 °C
PT	24 h	Unacceptable	Unacceptable	24 h	Unacceptable	2 weeks	12 months	
aPTT	4 h	Unknown	Unknown	4 h	4 h	2 weeks	12 months	
d-dimer	NA	NA	NA	24 h	24 h	24 months	24 months	

aPTT, Activated partial thromboplastin time; PT, prothrombin time.

Adapted from Clinical and Laboratory Standards Institute (CLSI). Collection, Transport, and Processing of Blood Specimens for Testing Plasma-Based Coagulation Assays and Molecular Hemostasis Assays; Approved Guideline—Fifth Edition. CLSI document H21-A5. Clinical and Laboratory Standards Institute, Wayne, Pennsylvania, USA, 2008; Clinical and Laboratory Standards Institute (CLSI). Quantitative D-dimer for the Exclusion of Venous Thromboembolic Disease; Approved Guideline. CLSI document H59-A. WaYNE, PA: Clinical and Laboratory Standards Institute; 2011.

**TABLE 5.20 Comparison of Prevalence of Reported Preanalytical Errors in Hematology Laboratory**

	Lippi, 2007 <sup>473</sup>	Simundic, 2010 <sup>137</sup>	Upreti, 2013 <sup>474</sup>	Narang, 2016 <sup>475</sup>	Arul, 2018 <sup>476</sup>	Narula, 2019 <sup>477</sup>
Wrong/missing identification	3.0%	27.1%	36.0%	2.6%	4.5%	23.4%
Inappropriate container	9.4%	0.4%	16.5%	5.3%	11.7%	2.3%
Inappropriate volume	7.0%	14.2%	19.7%	18.2%	45.8%	67.4%
Clotted samples	76.8%	31.5%	13.4%	73.9%	27.1%	2.3%
Contamination	3.8%	NA	4.3%	NA	4.5%	NA
Hemolysis	NA	23.7%	10.1%	NA	6.4%	NA
Samples lost/not found	NA	7.8%	NA	NA	NA	2.3%
Samples damaged in transport	NA	3.8%	NA	NA	NA	2.3%

Percentages are calculated as number of observed samples within category/total number of unsuitable samples.

Highest observed prevalence among preanalytical errors in hematology laboratory is marked in *light red*.

was chosen for hematologic tests when aniline-derived dyes were proposed for preparing blood smear from peripheral venous blood.<sup>479</sup> Anticoagulant function of EDTA is exerted through its potential to chelate calcium. Because EDTA as a free acid is not water soluble, it comes as disodium, dipotassium, and tripotassium salt. Potassium EDTA salts are more soluble than sodium salts. EDTA salts cause osmotically induced cell shrinkage and swelling to a different degree. Also, pH of EDTA increases with the number of ions bound to EDTA. Whereas the pH of EDTA in a free acid form is 2.5, tripotassium EDTA ( $K_3$ EDTA) has a pH of 7.5. In dipotassium and disodium EDTA, cell swelling is counteracted by cell shrinkage (due to the lower pH in dipotassium salts). Because cell shrinkage is less apparent, dipotassium EDTA ( $K_2$ EDTA) salts are superior to  $K_3$ EDTA. Also, mean cell corpuscular volume (MCV) based on the minihematocrit values in disodium and  $K_2$ EDTA samples provide acceptable results, as opposed to  $K_3$ EDTA samples.<sup>480</sup> Additionally, EDTA allows optimal dying with May-Grünwald Giemsa stain and optimal extended stabilization of blood cells and particles.<sup>479</sup> For these reasons, due to its higher solubility, lower osmotic effect, and best overall performance, the International Council for Standardization in Haematology (ICSH) recommends  $K_2$ EDTA salt as the anticoagulant of choice for hematology testing.<sup>481</sup> However, it is important to emphasize that these recommendations were based exclusively on the liquid form of  $K_3$ EDTA, and the conclusions drawn based on those type of tubes. Since the release of the recommendations, various studies have been published comparing  $K_2$  and  $K_3$ EDTA in

the automated CBC analysis, using glass and plastic tubes, and the results only partially agree with the recommendation.<sup>482</sup> Van Cott and colleagues concluded that the differences between results obtained with  $K_3$ EDTA glass tubes versus  $K_2$ EDTA plastic tubes are minimal and unlikely to be of any clinical significance.<sup>483</sup> With the release of spray-dried  $K_3$ EDTA tubes, some authors demonstrated the equivalence of spray-dried  $K_2$ EDTA, spray-dried  $K_3$ EDTA, and liquid  $K_3$ EDTA blood collection tubes for routine donor center or transfusion service testing.<sup>484</sup> It was confirmed that new  $K_3$ EDTA spray-dried tubes do not represent a clinically relevant new source of error in the clinical laboratory for several routine hematologic laboratory parameters.<sup>485</sup> Even the manufacturers themselves (Greiner Bio-One) have challenged the ICSH recommendation and demonstrated substantially equivalent performance of spray-dried  $K_3$ EDTA tubes to spray-dried  $K_2$ EDTA tubes.<sup>486</sup> Nevertheless, up to now, the recommended anticoagulant for hematology testing remains  $K_2$ EDTA. However, laboratory managers should be aware that the use of  $K_2$ EDTA vacuum tubes from different manufacturers may represent a clinically relevant source of variation for some CBC parameters.<sup>487</sup>

### Blood-to-Anticoagulant Ratio

According to CLSI GP39-A6 standard, EDTA tubes have a fixed fill volume that gives the optimum concentration of anticoagulant and should be filled to  $\pm 10\%$  of the stated draw volume.<sup>488</sup> In underfilled EDTA tubes, cell count and hematocrit might be falsely decreased due to the excess

EDTA, which leads to cell shrinkage. On the other side, tube overfilling may lead to clot formation and platelet clumping, because sample mixing becomes difficult due to the small head space.

These recommendations are based on some outdated studies, some of which are based on collection tubes with liquid anticoagulant.<sup>489–491</sup> Today, the predominant tubes for hematologic testing are spray-dried tubes, and evidence is emerging that shows that even underfilled spray-dried EDTA tubes might be acceptable for automated CBC counting.<sup>492–494</sup> Nevertheless, until some new studies prove differently tube overfilling is strongly discouraged.

### Contamination

Contamination with infusion fluids may be a cause of spurious anemia.<sup>478</sup> An unexplained decrease in CBC parameters may be due to IV contamination. A sudden shift of 4 to 5 fL in MCV, a parameter which should not fluctuate more than 1 to 2 fL within an individual, in the absence of a recent transfusion, is another reliable indicator of intravenous (IV) fluid contamination.<sup>495</sup> The fluid inside the RBC is hypertonic compared to the IV fluid which causes a shift of fluid into the cells, thus increasing their volume. Delta checks, the process of flagging differences in specific analytes between consecutive analyses, are one way to detect such problems. Delta checks are most efficient when done on parameters with low short-term biologic variability (e.g., within 24 hours), such as MCV and mean cell hemoglobin concentration (MCHC).<sup>496</sup>

Whenever possible, specimens should be collected from the arm opposite the line to avoid contamination. Specimens should not be collected distal to a catheter because fluids tend to pool in the periphery of the limb. Collection of samples proximal to a catheter will be diluted by the infusion fluid. When vascular access is limited, a specimen may need to be collected from the line. This decision should only be made after weighing the risk of specimen contamination versus the risk of phlebotomy from another site. Before drawing a specimen from the line, the infusion fluid should be completely stopped for several minutes and an amount of blood equal to three or more times the dead space of the catheter should be discarded.

### Tube Mixing

Blood should be mixed immediately after the specimen is drawn to allow proper mixing of the additive with blood and to prevent sample clotting. Poor sample mixing may lead to sample clotting. Clotted specimens are among the most common reasons for sample rejection in automated cell counting and coagulation.<sup>475,478,497</sup> While some authors have argued that mixing may be avoided, because blood is mixed with anticoagulant spontaneously, during the blood draw,<sup>498</sup> tube mixing is still recommended as an essential step which ensures the quality of hematology samples.

Tubes should also be mixed immediately before the analysis to achieve sample homogeneity.

Adequate mixing is achieved by gently inverting the tube at 180° and back to the upright position. For optimal results, the number of full rotations should correspond to manufacturers' instructions.<sup>499</sup> Although it has been suggested that vigorous mixing does not promote laboratory variability,<sup>500</sup> general advice is that vigorous mixing should be avoided.

### Sample Stability

Sample stability is a crucial aspect for the quality of results in the hematologic laboratory. Although the official recommendations of the ICSH evolved with time,<sup>501,502</sup> as a general rule, the EDTA anticoagulated blood should be stored at room temperature and analyzed within 3 hours of collection. It should be noted that analyte stability may differ depending on the parameters that are being measured, instrument type, and transport and storage conditions.<sup>503–506</sup> Therefore on some occasions, a shorter storage time is necessary to ensure accurate and reliable results, whereas some parameters show excellent stability even over much longer time intervals. Some parameters are very stable (hemoglobin and RBC), while others appear less stable (reticulocytes, MCV, and hematocrit). The stability of hematologic parameters is improved if samples are kept at 4 °C. Medium to long-term storage of whole blood samples at high temperatures (37 °C) should absolutely be avoided, both during transportation and within the laboratory.<sup>504</sup> Special attention should be paid to sample stability for preparation of peripheral blood smears. Morphologic changes in anticoagulated blood begin within 30 minutes in neutrophils and other granulocytes and consist of swelling, nuclear lobe structure, loss of cytoplasmic granulation, and vacuolization in anticoagulated blood.<sup>501</sup> These changes are attenuated with the time and conditions of blood storage. Some most relevant changes observed for WBC after prolonged storage at room temperature are abnormal chromatin clumping, abnormal band forms (so called “pseudo-bands”), loss of cytoplasmic margin definition, neutrophil and/or basophil degranulation, smudge cells, and so on.<sup>507</sup>

Although some of the changes are delayed in samples stored at 4°C, they are not eliminated, and for this reason, it is important to make smears as soon as possible.<sup>502</sup>

The ICSH data on the stability of some hematology parameters are summarized in Table 5.21.

**TABLE 5.21 International Council for Standardization in Haematology Data on the Stability of Some Hematology Parameters<sup>502</sup>**

Parameter	Storage Conditions	
	Room Temperature (18–25 °C)	4 °C
Hemoglobin	NA	72 h
RBC count	NA	72 h
MCV (Hct)	1–4 h	6–12 h
Reticulocyte count	NA	24–72 h
Platelets	NA	24–72 h
WBC count	6 h	24–72 h
Automated differential count	6 h	24–72 h
Peripheral blood smear	<3 h	8 h

*Hct*, Hematocrit; *MCV*, mean corpuscular volume; *RBC*, red blood cell; *WBC*, white blood cell.

Adapted from Zini G. International Council for Standardization in Haematology (ICSH): stability of complete blood count parameters with storage: toward defined specifications for different diagnostic applications. *Int J Lab Hematol* 2014;36:111–3.

As already emphasized above, the stability may vary depending on the instrument used. It is therefore the responsibility of the individual laboratory to verify the stability of the hematologic parameters on their instruments.

### Antibodies

Antibodies may affect the cell count of erythrocytes, leukocytes, and platelets. The following antibodies are known to interfere with hematologic analytes:

- EDTA-dependent antibodies with thrombocyte and leukocyte specificity
- Cold agglutinins (erythrocyte specific)
- Cryoglobulins

**EDTA pseudothrombocytopenia.** EDTA may in some individuals cause pseudothrombocytopenia—that is, platelet clumping or platelet satellitism (platelets adhering to neutrophils) and subsequently inaccurate platelet results.<sup>508,509</sup> Although EDTA-dependent pseudothrombocytopenia is a rare phenomenon (around 0.1% in the general population),<sup>510,511</sup> the reliable and timely identification is essential since it can lead to inappropriate clinical decisions.<sup>512</sup> The phenomenon was described more than 50 years ago and has been observed in healthy and diseased individuals, unrelated to gender, age, age of onset, disease, hemostasis alterations, or ingestion of specific drugs.<sup>513,514</sup> The hypothesized mechanism in pseudothrombocytopenia involves IgM autoantibodies directed against platelet IIb/IIIa fibrinogen receptors. EDTA induces steric conformation on negatively charged phospholipids and membrane receptors which then react with autoantibodies that triggers platelet activation through enhanced expression of GMP140, Gp55, and thrombospondin, activation of the tyrosine kinase pathway, and finally platelet agglutination and clumping in vitro (Fig. 5.19).<sup>512</sup>

This is further supported by the fact that platelets from patients with Glanzmann thrombasthenia (in which platelets have either defective or low concentrations of glycoprotein

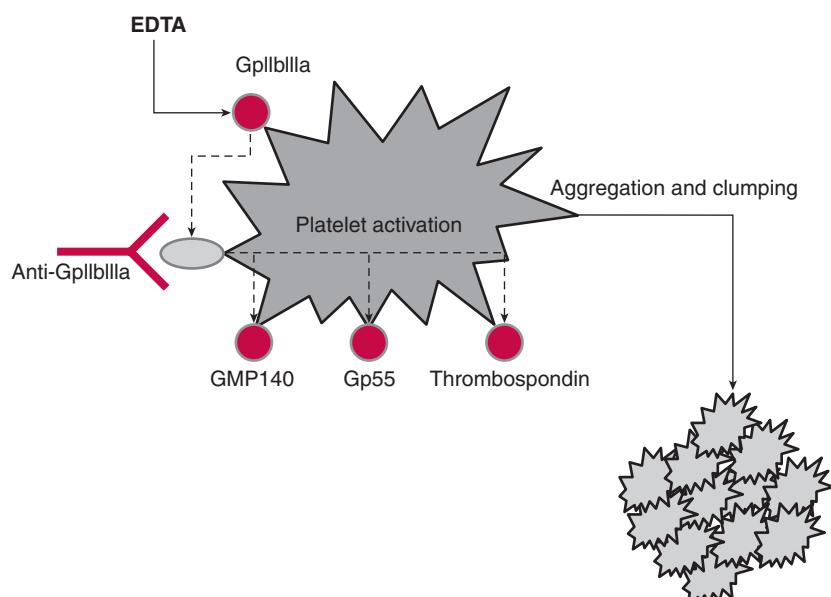
IIb/IIIa) do not react to autoantibodies from pseudothrombocytopenic patients.<sup>515</sup> Transplacental transmission from mother to child during pregnancy has also been described.<sup>516,517</sup> Major criteria for establishing a diagnosis of EDTA-dependent pseudothrombocytopenia are as follows<sup>512</sup>:

1. Platelet count typically  $<100 \times 10^9/L$
2. Onset in only EDTA-anticoagulated sample kept at room temperature
3. Time-dependent fall of the platelet count in EDTA specimens
4. Presence of platelet aggregates and clumps in EDTA-anticoagulated samples
5. Lack of clinical signs or symptoms of platelet disorder

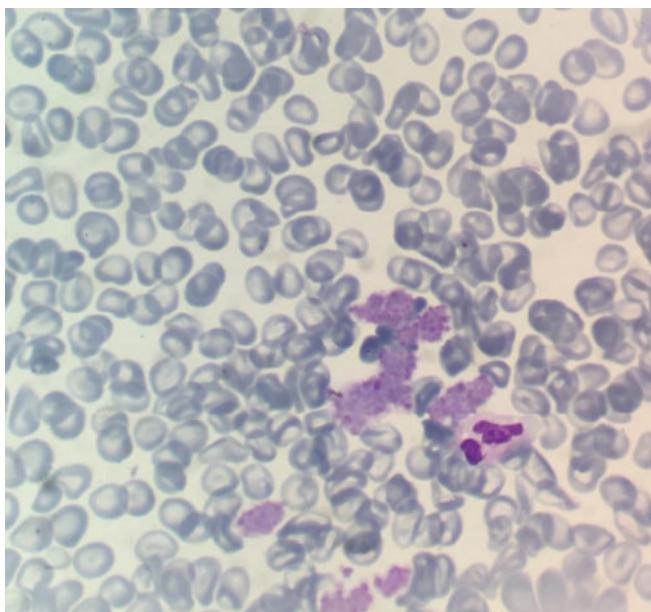
Because most cell counters are not able to identify this preanalytical problem, platelets are thus counted as WBC, resulting in spurious leukocytosis and false thrombocytopenia.<sup>518</sup> Although there were some efforts in screening for EDTA-dependent deviations in platelet counts by exploring platelet distribution histograms on hematology analyzers,<sup>519</sup> visual evaluation of blood smears is regarded as the gold standard for detection of EDTA-pseudothrombocytopenia, showing a typical pattern of platelet aggregates (Fig. 5.20).<sup>520,521</sup>

Although promising in many aspects of morphologic assessment of the blood smear, digital morphology analyzers have shown insufficient sensitivity for platelet clump detection; a manual microscopic review is recommended.<sup>522</sup>

Several approaches have been described to resolve this in vitro phenomenon. Platelet aggregation was not observed in samples anticoagulated with mixtures of EDTA and aminoglycosides.<sup>523</sup> Magnesium sulfate, previously used as anticoagulant for estimating manual platelet count, has also been proven promising as an alternative anticoagulant for platelet counts estimation in EDTA-pseudothrombocytopenia,<sup>524</sup> and some novel methods to dissociate platelet clumps based on the pathophysiologic mechanism were described.<sup>525</sup> Nevertheless, the most suitable and practical approach for most



**FIGURE 5.19** Pathogenesis of EDTA-dependent pseudothrombocytopenia.<sup>512</sup> EDTA from the anticoagulant enables IgM autoantibodies binding against platelet IIb/IIIa fibrinogen receptors. This triggers platelet activation and, finally, platelet agglutination and clumping in vitro.



**FIGURE 5.20** EDTA induced platelet aggregates in the peripheral blood smear. (Figure provided by the Department of Medical Laboratory Diagnostics, Clinical Hospital Sveti Duh, Zagreb, Croatia. Photo taken by Vanja Radisic Biljak.)

clinical laboratories will be recollection of specimens using tubes with other anticoagulant (sodium citrate, CPT,  $\text{CaCl}_2$ /heparin), and immediate processing of those blood samples.

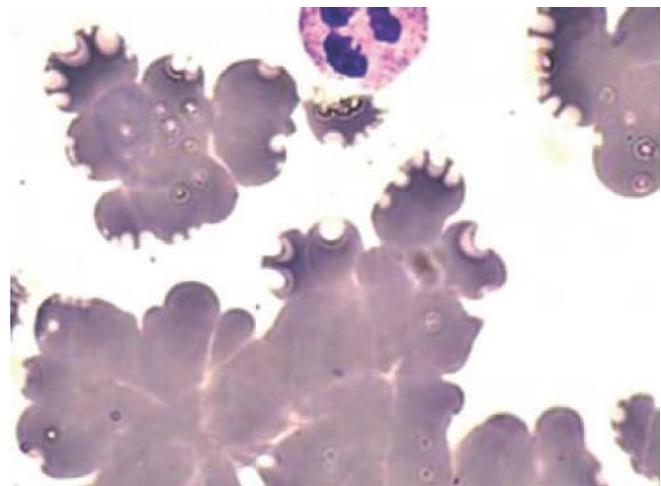
**Cold agglutinins.** Cold agglutinins are antibodies that are specific for erythrocyte surface carbohydrate antigens, which bind to the erythrocyte surface at temperatures of 0 to 4 °C. Binding of agglutinins causes agglutination of erythrocytes, induces complement activation and hemolysis, and impairs peripheral circulation.<sup>526</sup> Cold agglutinins may be monoclonal or polyclonal. Monoclonal agglutinins are found in patients with idiopathic forms of cold agglutinin disease or lymphoproliferative disorders, whereas polyclonal antibodies are often found in patients recovering from some infectious diseases.<sup>527</sup> Some rare cases of cold agglutinins toward platelets have also been described, causing pseudothrombocytopenia independent of EDTA.<sup>528</sup>

Cold agglutinins, if undetected, may cause diagnostic confusion and lead to subsequent extensive diagnostic workup and incorrect and unnecessary therapy, risking patient safety and increasing health care costs. It is therefore very important to recognize cold agglutinins promptly. Cold agglutinins should be suspected if the following anomalies are observed<sup>529,530</sup>:

- RBC counts too low, even at normal hemoglobin concentration, with a false decrease in hematocrit
- Falsely increased Red Cell Distribution Width (RDW) values
- Grossly enhanced MCV values due to measurements of erythrocyte clumps
- Falsely increased MCH and MCHC values without any obvious explanation.

Since hemoglobin is measured after RBC are lysed, it is generally not affected. The blood smear will show agglutination of erythrocytes (Fig. 5.21).<sup>531</sup>

For adequate analysis of samples in which cold agglutinins are suspected, it is essential to warm up the EDTA blood



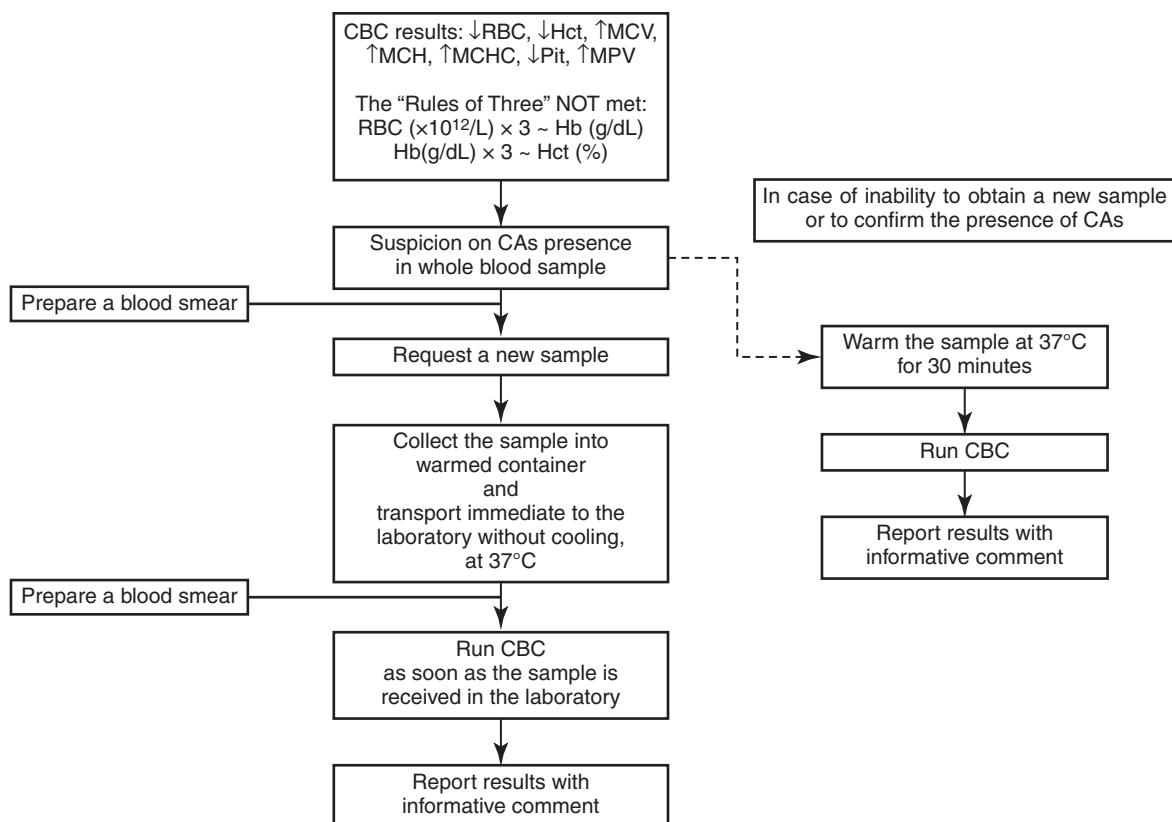
**FIGURE 5.21** Cold agglutinins in peripheral blood smear with typical clusters of erythrocytes, May-Grünwald—Giemsa stain, 1000x. Cold agglutinins bind to the erythrocyte surface antigen at a temperature optimum of 0 to 4 °C, which causes agglutination of erythrocytes. Red blood cells cluster together in an irregular mass giving the “bunch of grapes” appearance, unlike rouleaux formation which resembles a “stack of coins” agglutinates.<sup>531</sup>

sample at 37 °C and analyze immediately afterward.<sup>529,530,532</sup> This anomaly will appear again if a sample is kept at 4 °C and analyzed while it is cold. Samples in which cold agglutination is suspected should be examined by microscopic observation of a blood smear. In such cases, both the sample and the glass slide should be prewarmed to avoid agglutination.<sup>533</sup> A schematic overview of a proposed laboratory procedure to prevent preanalytical errors caused by cold agglutinins is presented in Fig. 5.22.<sup>534</sup>

**Cryoglobulins.** Cryoglobulins are immunoglobulins that precipitate in vitro at cold temperature and dissolve at 37 °C.<sup>535</sup> Cryoglobulins are often associated with infections, autoimmune disorders, and malignancies, and they can cause organ damage through immune-mediated mechanisms and vascular damage due to increased viscosity of the blood.<sup>536</sup> Precipitation of cryoglobulins depends on the immunoglobulin class to which they belong. Also, precipitation is absent at pH less than 5.0 or greater than 8.0.

In samples kept at room temperature, cryoglobulins tend to form globular or cylindric precipitates that are then counted by automated hematologic analyzers as cells, thus affecting hematologic laboratory tests and leading to false leukocytosis (pseudoleukocytosis) or false thrombocytosis (pseudothrombocytosis), while RBC values are generally unaffected in cryoglobulinemias. The degree of pseudoleukocytosis and pseudothrombocytosis depends on the time of exposure, temperature, cryoglobulin concentration, and the interaction of cryoglobulins with other plasma proteins.<sup>537</sup> Recognition of the hematologic abnormalities may be the first clue leading to the diagnosis of cryoglobulinemia.<sup>538,539</sup> The following indices may point to the presence of cryoglobulins<sup>538</sup>:

1. Abnormalities in automated WBC and/or platelet counts
2. Visualization of the cryoglobulins as blue sediments in differential count samples
3. In a sample warmed up to 37 °C, significantly lower cell counts



**FIGURE 5.22** Laboratory algorithm for investigating suspected cold agglutinins in whole blood samples. CAs, Cold agglutinins; CBC, complete blood count; Hct, hematocrit; MCH, mean corpuscular hemoglobin; MCV, mean corpuscular volume; MCHC, mean corpuscular hemoglobin concentration; MPV, mean platelet volume; Plt, platelet; RDW, red cell distribution width; WBC, white blood cell count (From Topic A, Milevoj Kopcinovic L, Bronic A, et al. Effect of cold agglutinins on red blood cell parameters in a trauma patient: a case report. *Biochem Med* 2018;28:031001, with permission by Croatian Society of Medical Biochemistry and Laboratory Medicine.)

Diagnosis of cryoglobulinemia is further confirmed by biochemical analysis of the serum, followed by isolation, purification, and immunochemical analysis of cryoglobulins. As in the case of cold agglutinins, for adequate analysis of samples in which cryoglobulins are suspected, preincubation of blood samples at 37 °C for 10 to 15 minutes often yields reliable CBC results, especially if analyzed immediately afterward.<sup>540</sup>

## MANAGEMENT OF THE QUALITY OF THE PREANALYTICAL PHASE

The ultimate goal of preanalytical quality management is not to improve the quality of the sample per se but to improve patient outcome.<sup>205</sup> Preanalytical quality management achieves its primary goal only if (1) the importance of preanalytical processes in the TTP is fully understood; (2) all sources of preanalytical variability and their effects are known; (3) patient-centered and evidence-based guidelines are available; (4) compliance with guidelines can be ensured; and (5) quality is continuously monitored and improved.<sup>205</sup>

Unfortunately, the preanalytical part of the TTP is not fully understood by all involved (laboratory staff, nurses, clinicians, and patients). Patients are usually not aware of the importance of proper preanalytical procedures and how improper sample collection could affect the results of requested tests.<sup>541,542</sup> Education is the key to the improvement of the

level of understanding of the importance of the preanalytical phase.<sup>543–545</sup> However, the effects of educational interventions are usually short-lived, and education should therefore be a continuous quality improvement activity.

### Outcome-based Preanalytical Studies

Although many potential sources of variability and how they affect the quality of samples and test results are well recognized, there is little evidence demonstrating the effect of preanalytical variation on patient outcomes and how particular errors may affect health care organization and expenditure. Most studies so far have been descriptive and reported failures of processes without linking those to patient harm. Quality improvement should focus on reducing patient harm rather than just eliminating process defects and waste.<sup>546</sup> Original studies need to focus more on patient-relevant outcomes—for example, how some preanalytical errors, such as improper sampling, delayed transport, or hemolyzed or clotted samples, may lead to patient discomfort, additional diagnostic workup, increased LOS in the hospital, increased costs, disease prevalence, and so on.<sup>547–550</sup> Error reduction strategies should focus on those most critical errors that have the greatest potential to impact patient outcomes.

### Quality Indicators

Preanalytical quality should continuously be monitored and improved. To measure the degree of improvement, quality

indicators (QI) should be used (see Chapter 3). QI are measurable, objective, quantitative measures of key system elements that show to what extent a laboratory meets the needs and expectations of the customers.<sup>551</sup> To allow consistent and comparable use across settings over time, a unique definition is needed for QIs.<sup>552–554</sup> While different professional groups have proposed some interesting programs on the use of QI in the TTP, until very recently there was no consensus on the definition, measurement methodology, and reporting practices for QI. Recently, a harmonized model of QI has been established by an expert panel during a consensus conference in Padua in October 2013.<sup>555</sup> The proposed QI model is based on a patient-centered approach, and the essential prerequisites taken into account were the following:

- Importance and applicability to a wide range of clinical laboratories worldwide

- Scientific soundness (focused on some most important areas in laboratory medicine)
- Definition of evidence-based thresholds for acceptable performance
- Timeliness and possible utilization as a measure of laboratory improvement

The model proposes 22 high-priority and 6 lower-priority preanalytical QIs (Table 5.22).

QIs enable the measurement of the quality of care and services, with the aim of assisting in quality improvement efforts. Collecting data on QI per se does not automatically mean quality improvement. Laboratories should strive for a system of continuous preanalytical quality improvement based on the “plan-do-check-act” cycle and using corrective and preventive actions with subsequent system redesign. Only then can patient outcomes be improved, preanalytical errors be reduced, and waste be eliminated.

**TABLE 5.22 Proposed Preanalytical (Priority 1) Quality Indicators Based on a Harmonized Consensus Model**

Quality Indicator	Reporting Systems
Misidentification errors	Samples suspected to be from wrong patients Percentage of “Number of misidentified requests/Total number of requests” Percentage of “Number of misidentified samples/Total number of samples” Percentage of “Number of samples with fewer than two identifiers initially supplied/Total number of samples” Percentage of “Number of unlabeled samples/Total number of samples”
Test transcription errors	Percentage of “Number of outpatient requests with erroneous data entry (test name)/Total number of outpatient requests” Percentage of “Number of outpatient requests with erroneous data entry (missed test)/Total number of outpatient requests” Percentage of “Number of outpatient requests with erroneous data entry (added test)/Total number of outpatient requests” Percentage of “Number of inpatient requests with erroneous data entry (test name)/Total number of inpatient requests” Percentage of “Number of inpatient requests with erroneous data entry (missed test)/Total number of inpatient requests” Percentage of “Number of inpatient requests with erroneous data entry (added test)/Total number of inpatient requests”
Incorrect sample type	Percentage of “Number of samples with wrong or inappropriate type (i.e., whole blood instead of plasma)/Total number of samples”
Incorrect fill level	Percentage of “Number of samples collected in wrong containers/Total number of samples” Percentage of “Number of samples with insufficient sample volume/Total number of samples” Percentage of “Number of samples with inappropriate sample-anticoagulant volume ratio/Total number of samples with anticoagulant”
Unsuitable samples for transportation and storage problems	Percentage of “Number of samples not received/Total number of samples” Percentage of “Number of samples not properly stored before analysis/Total number of samples” Percentage of “Number of samples damaged during transportation/Total number of samples” Percentage of “Number of samples transported at inappropriate temperature/Total number of samples” Percentage of “Number of samples with excessive transportation time/Total number of samples”
Contaminated samples	Percentage of “Number of contaminated samples rejected/Total number of samples”
Samples hemolyzed	Percentage of “Number of samples with free hemoglobin >0.5 g/L/Total number of samples (clinical chemistry)” <sup>a</sup>
Samples clotted	Percentage of “Number of samples clotted/Total number of samples with an anticoagulant”

<sup>a</sup>Clinical chemistry: all samples that are analyzed on the chemistry analyzer that is used for detection of HIL indices. If laboratories are detecting hemolysis visually, they count all samples with visible hemolysis (clinical chemistry). A color chart should be provided for this purpose.

From Plebani M, Astion ML, Barth JH, Chen W, de Oliveira Galoro CA, Escuer MI, et al. Harmonization of quality indicators in laboratory medicine: a preliminary consensus. *Clin Chem Lab Med* 2014;52:951–8, with permission by Walter de Gruyter.

## POINTS TO REMEMBER

### Management of the Quality of the Preanalytical Phase

- Quality improvement should focus on how to improve patient outcomes and reduce patient harm rather than to eliminate process defects and waste.
- Original studies need to focus more on outcomes and provide evidence for the effect of preanalytical errors (e.g., improper sampling, delayed transport, hemolyzed or clotted sample) on patient discomfort or harm, additional diagnostic workup, increased length of stay, increased costs, disease prevalence, and so on.
- Knowing the errors with the greatest potential to impact patient outcomes helps to prioritize error-reduction strategies and focus on the most critical errors.

## SUGGESTED READING

9. Guder WG. History of the preanalytical phase: a personal view. *Biochem Med* 2014;24:25–30.
24. Simundic AM, Cornes M, Grankvist K, et al. Survey of national guidelines, education and training on phlebotomy in 28 European countries: an original report by the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) working group on the preanalytical phase (WG-PA). *Clin Chem Lab Med* 2013;51:1585–93.
25. Simundic AM, Bölenius K, Cadamuro J, et al. Joint EFLM-COLABIOCLI recommendation for venous blood sampling. *Clin Chem Lab Med* 2018;56(12):2015–38. doi:[10.1515/cclm-2018-0602](https://doi.org/10.1515/cclm-2018-0602).
103. Sanchis-Gomar F, Lippi G. Physical activity—an important preanalytical variable. *Biochem Med* 2014;24:68–79.
117. Simundic AM, Cornes M, Grankvist K, et al. Standardization of collection requirements for fasting samples: for the Working Group on Preanalytical Phase (WG-PA) of the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM). *Clin Chim Acta* 2014;432:33–7.
133. Bowen RA, Remaley AT. Interferences from blood collection tube components on clinical chemistry assays. *Biochem Med* 2014;24:31–44.
135. Simundic AM, Baird G, Cadamuro J, Costelloe SJ, Lippi G. Managing hemolyzed samples in clinical laboratories. *Crit Rev Clin Lab Sci* 2020;57(1):1–21. doi:[10.1080/10408363.2019.1664391](https://doi.org/10.1080/10408363.2019.1664391).
142. Vecellio E, Li L, Mackay M, et al. A benchmark study of the frequency and variability of haemolysis reporting across pathology laboratories—The implications for quality use of pathology and safe and effective patient care. Report to Royal College of Pathologists Australasia. Australian Institute of Health Innovation, Macquarie University, Sydney; July 2015.
166. Nikolac N. Lipemia: causes, interference mechanisms, detection and management. *Biochem Med* 2014;24:57–67.
169. Tate J, Ward G. Interferences in immunoassay. *Clin Biochem Rev* 2004;25:105–20.
173. Dimeski G. Interference testing. *Clin Biochem Rev* 2008;29(Suppl 1):S43–8.
203. Nikolac N, Simundic AM, Miksa M, et al. Heterogeneity of manufacturers' declarations for lipemia interference—An urgent call for standardization. *Clin Chim Acta* 2013;426:33–40.
323. Li D, Ferguson A, Cervinski MA, et al. AACC guidance document on biotin interference in laboratory tests. *J Appl Lab Med* 2020;5(3):575–87.
375. Delanghe J, Speeckaert M. Preanalytical requirements of urinalysis. *Biochem Med (Zagreb)* 2014;24:89–104.
388. Ellervik C, Vaught J. Preanalytical variables affecting the integrity of human biospecimens in biobanking. *Clin Chem* 2015;61:914–34.
395. Baird G. Preanalytical considerations in blood gas analysis. *Biochem Med* 2013;23:19–27.
424. Bonar R, Favaloro EJ, Adcock DM. Quality in coagulation and haemostasis testing. *Biochem Med* 2010;20:184–99.
431. Gosselin RC, Marlar RA. Preanalytical variables in coagulation testing: Setting the stage for accurate results. *Semin Thromb Hemost* 2019;45:433–48.
551. Simundic AM, Topic E. Quality indicators. *Biochem Med* 2008;18:311–19.
554. Plebani M, Sciacovelli L, Aita A, et al. Harmonization of pre-analytical quality indicators. *Biochem Med* 2014;24:105–13.

## REFERENCES

1. Kohn LT, Corrigan JM, Donaldson MS, Institute of Medicine. To err is human: building a safer health system. Washington, DC: National Academy Press; 2000.
2. James J. A new, evidence-based estimate of patient harms associated with hospital care. *J Patient Saf* 2013;9:122–8.
3. Available from: [http://www.who.int/features/factfiles/patient\\_safety/en/](http://www.who.int/features/factfiles/patient_safety/en/).
4. Available from: [http://ec.europa.eu/health/patient\\_safety/policy/index\\_en.htm](http://ec.europa.eu/health/patient_safety/policy/index_en.htm).
5. Graber ML, Franklin N, Gordon R. Diagnostic error in internal medicine. *Arch Intern Med* 2005;165:1493–9.
6. Beastall GH. Adding value to laboratory medicine: a professional responsibility. *Clin Chem Lab Med* 2013;51:221–7.
7. Hallworth MJ. The “70% claim”: what is the evidence base? *Ann Clin Biochem* 2011;48:487–8.
8. Schiff GD, Hasan O, Kim S, et al. Diagnostic error in medicine: analysis of 583 physician-reported errors. *Arch Intern Med* 2009;169:1881–7.
9. Guder WG. History of the preanalytical phase: a personal view. *Biochem Med* 2014;24:25–30.
10. Simundic AM, Lippi G. Preanalytical phase—A continuous challenge for laboratory professionals. *Biochem Med* 2012;22:145–9.
11. Guder WG. Einfluss von Probennahme, Probentransport und Probenverwahrung auf klinisch chemische Untersuchungen. *Ärztl Lab* 1976;22:69–75.
12. Statland BE, Winkel P. Physiological variation of the concentrations values of selected analytes determined in healthy young adults. In: Elevitch FR, editor. Proceedings of the 1976 Aspen conference on Analytical Goals in Clinical Chemistry College of American Pathologists Chicago 1977.
13. Statland BE, Winkel P. Effects of preanalytical factors on the intraindividual variation of analytes in the blood of healthy subjects: consideration of preparation of the subject and time of venipuncture. *Crit Rev Clin Lab Sci* 1977;8:105–44.
14. Keller H, Guder WG, Hansert E, et al. Biological influence factors and interference factors in clinical chemistry: general considerations. *J Clin Chem Clin Biochem* 1985;23:3–6.
15. Lundberg GD. Critical (panic) value notification: an established laboratory practice policy (parameter). *JAMA* 1990; 263:709.
16. Godolphin W, Bodtker K, Uyeno D, et al. Automated blood sampling handling in the clinical laboratory. *Clin Chem* 1990;36:1551–5.
17. Büttner J. Unspecificity and interference in analytical systems: concepts and theoretical aspects. *Klin Chem Mitt* 1991;22:3–12.
18. Guder WG, Wahlefeld AW. Specimens and samples in clinical chemistry. In: Bergmeyer HU, editor. Methods of enzymatic analysis, vol. 2. 3rd ed. Weinheim: Verlag Chemie; 1983. p. 2–20.
19. Guder WG, Narayanan S. Pre-examination procedures in laboratory diagnostics. Berlin: Walter de Gruyter; 2015.
20. Plebani M, Carraro P. Mistakes in a stat laboratory: types and frequency. *Clin Chem* 1997;43:1348–51.
21. International Organization for Standardization (ISO). ISO 15189:2012: Medical laboratories: Particular requirements for quality and competence. Geneva, Switzerland: ISO; 2012.
22. Schiettecatte J, Anckaert E, Smitz J. Chapter 3: Interferences in immunoassays. In: Chiu NHL, Christopoulos TK, editors. Advances in immunoassay technology. InTechOpen; 2012.
23. Jones AM, Honour JW. Unusual results from immunoassays and the role of the clinical endocrinologist. *Clin Endocrinol (Oxf)* 2006;64:234–44.
24. Simundic AM, Cornes M, Grankvist K, et al. Survey of national guidelines, education and training on phlebotomy in 28 European countries: an original report by the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) working group on the preanalytical phase (WG-PA). *Clin Chem Lab Med* 2013;51:1585–93.
25. Simundic AM, Bölenius K, Cadamuro J, et al. Joint EFLM-COLABIOCLI recommendation for venous blood sampling. *Clin Chem Lab Med* 2018;56(12):2015–38. doi:10.1515/cclm-2018-0602.
26. Young DS. Effects of preanalytical variables on clinical laboratory tests. 3rd ed. Washington DC: AACC Press; 2007.
27. Nguyen LT, Buse JD, Baskin L, Sadzadeh SMH, Naugler C. Influence of diurnal variation and fasting on serum iron concentrations in a community-based population. *Clin Biochem* 2017;50(18):1237–42. doi:10.1016/j.clinbiochem.2017.09.018.
28. Harrop IS, Ashwell K, Hapton MR. Circannual and within individual variation of thyroid function tests in normal subjects. *Ann Clin Biochem* 1985;22:371–5.
29. Nordin BEC, Wilkinson R, Marshall DH, et al. Calcium absorption in the elderly. *Calcif Tissue Res* 1976;21:442–51.
30. Hagemann P. Qualität im Arztlabor, Optimierung der Präanalytik. Berlin-Heidelberg: Springer; 1994.
31. Keller H. Klinisch chemische Labordiagnostik für die Praxis. 2nd ed. Stuttgart: Thieme; 1991.
32. Das SS, Chaudharry R, Khetan D, et al. Calcium and magnesium levels during automated plateletpheresis in normal donors. *Transfus Med* 2005;15:233–6.
33. Das SS, Chaudharry R, Verma SK, et al. Pre- and post-donation haematological values in healthy donors undergoing plateletpheresis with five different systems. *Blood Transfus* 2009;7:188–92.
34. Dimeski G, Badrick T, St John A. Ion selective electrodes (ISEs) and interferences – A review. *Clin Chim Acta* 2010;411:309–17.
35. Lippi G, Daves M, Mattiuzzi C. Interference of medical contrast media on laboratory testing. *Biochem Med (Zagreb)* 2014; 24:80–8.
36. Barr ML, Chiu HK, Li N, et al. Thyroid dysfunction in children exposed to iodinated contrast media. *J Clin Endocrinol Metab* 2016;101:2366–70.
37. Al Dieri R, Béguin S, Hemker HC. The ionic contrast medium ioxaglate interferes with thrombin-mediated feedback activation of factor V, factor VIII and platelets. *J Thromb Haemost* 2003;1:269–74.
38. Laskey WK, Gellman J. Inflammatory markers increase following exposure to radiographic contrast media. *Acta Radiol* 2003;44:498–503.
39. Vlasveld LT, van 't Wout J, Castel A. False elevation of chromogranin A due to proton pump inhibitors. *Neth J Med* 2011;69:207.
40. Butt MU, Jabri A, Elayi SC. Azithromycin-induced thrombocytopenia: a rare etiology of drug-induced immune thrombocytopenia. *Case Rep Med* 2019;2019:6109831.
41. Thomas K, Eisele J, Rodriguez-Leal FR, et al. Acute effects of alemtuzumab infusion in patients with active relapsing-remitting MS. *Neurol Neuroimmunol Neuroinflamm* 2016;3:e228.
42. Kieboom BCT, Zietse R, Ikram MA, et al. Thiazide but not loop diuretics is associated with hypomagnesaemia in the general population. *Pharmacopidemiol Drug Saf* 2018;27:1166–73.

43. Delanaye P, Mariat C, Cavalier E, et al. Trimethoprim, creatinine and creatinine-based equations. *Nephron Clin Pract* 2011;119:c187–93.
44. Dasgupta A, Bernard DW. Herbal remedies: effects on clinical laboratory tests. *Arch Pathol Lab Med* 2006;130:521–8.
45. Singh A, Zhao K. Herb–drug interactions of commonly used Chinese medicinal herbs. *Int Rev Neurobiol* 2017;135: 197–232.
46. Leite PM, Martins MAP, Castilho RO. Review on mechanisms and interactions in concomitant use of herbs and warfarin therapy. *Biomed Pharmacother* 2016;83:14–21.
47. Christensen CM, Morris RS, Kapsandoy SC, et al. Patient needs and preferences for herb-drug-disease interaction alerts: a structured interview study. *BMC Complement Altern Med* 2017;17:272.
48. Escher M, Desmeules J, Giostra E, et al. Hepatitis associated with Kava, a herbal remedy. *BMJ* 2001;322:139.
49. Li XZ, Ramzan I. Role of ethanol in Kava hepatotoxicity. *Phytother Res* 2010;24:475–80.
50. Favreau JT, Ryu ML, Braunstein G, et al. Severe hepatotoxicity associated with the dietary supplement. *Ann Intern Med* 2002;136:590–5.
51. Jorge OA, Jorge AD. Hepatotoxicity associated with the ingestion of Centella asiatica. *Rev Esp Enferm Dig* 2005;97:115–24.
52. Arum SM, He X, Braverman LE. Excess iodine from an unexpected source. *N Engl J Med* 2009;360:424–6.
53. Müssig K, Thamer C, Bares R, et al. Iodine-induced thyrotoxicosis after ingestion of kelp-containing tea. *J Gen Intern Med* 2006;21:C11–4.
54. Preiss DJ, Godber IM, Lamb EJ, et al. The influence of a cooked-meat meal on estimated glomerular filtration rate. *Ann Clin Biochem* 2007;44(Pt 1):35–42.
55. Young DS. Preanalytical variables and biological variation. In: Carl A. Burtis, Edward R. Ashwood and David E. Bruns (eds): *Tietz Textbook of clinical chemistry and molecular diagnosis*, 5th ed. St Louis: Elsevier, 2012.
56. Steinmetz J, Panek E, Sourieau F, et al. Influence of food intake on biological parameters. In: Siest G, editor. *Reference values in human chemistry*. Basel: Karger; 1973. p. 195–200.
57. Buttar HS, Li T, Ravi N. Prevention of cardiovascular diseases: role of exercise, dietary interventions, obesity and smoking cessation. *Exp Clin Cardiol* 2005;10:229–49.
58. Kleiner RE, Hutchins AM, Johnston CS, et al. Effects of an 8-week high-protein or high-carbohydrate diet in adults with hyperinsulinemia. *MedGenMed* 2006;8:39.
59. Irwin MI, Staton AJ. Dietary wheat starch and sucrose: effect on levels of five enzymes in blood serum of young adults. *Am J Clin Nutr* 1969;22:701–9.
60. Fuhrman MP, Charney P, Mueller CM. Hepatic proteins and nutrition assessment. *J Am Diet Assoc* 2004;104:1258–64.
61. Olusi SO, McFarlane H, Osunkoya BO, et al. Specific protein assays in protein calorie malnutrition. *Clin Chim Acta* 1975;62:107–16.
62. Rifai N, Merill JR, Holly RG. Postprandial effect of a high-fat meal on plasma lipid, lipoprotein, cholesterol and apolipoprotein measurements. *Ann Clin Biochem* 1990;27:489–93.
63. Evans K, Laker MF. Intraindividual factors affecting lipid, lipoprotein and apolipoprotein measurement: a review. *Ann Clin Biochem* 1995;32:261–80.
64. Guder WG, Narayanan S, Wisser H, et al. Diagnostic samples: from the patient to the laboratory. 4th ed. Weinheim: Wiley-Blackwell; 2009.
65. Lima-Oliveira G, Salvagno GL, Lippi G, et al. Influence of a regular, standardized meal on clinical chemistry analytes. *Ann Lab Med* 2012;32(4):250–6. doi:10.3343/alm.2012.32.4.250.
66. Kackov S, Simundic AM, Nikolac N, et al. The effect of high-calorie meal consumption on oxidative stress and endothelial dysfunction in healthy male adults. *Physiol Res* 2013;62:643–52.
67. Bajaña W, Aranda E, Arredondo ME, et al. Impact of an Andean breakfast on biochemistry and immunochemistry laboratory tests: an evaluation on behalf COLABIOCLI WG-PRE-LATAM. *Biochem Med (Zagreb)* 2019;29(2):020702. doi:10.11613/BM.2019.020702.
68. van Oostrom AJHHM, Sijmonsma TP, Rabelink TJ. Postprandial leukocyte increase in healthy subjects. *Metabolism* 2003; 52:199–202.
69. Lippi G, Lima-Oliveira G, Salvagno GL, et al. Influence of a light meal on routine haematological tests. *Blood Transfus* 2010;8:94–9.
70. Kościelniak BK, Charchut A, Wójcik M, et al. Impact of fasting on complete blood count assayed in capillary blood samples. *Lab Med* 2017;48:357–61.
71. Onbas,i K, Efe B, Celer Ö. Postprandial phase fluctuations can trigger the coagulation cascade. *Int J Clin Exp Med* 2016;9:5891–901.
72. Arredondo ME, Aranda E, Astorga R, et al. Breakfast can affect routine hematology and coagulation laboratory testing: an evaluation on behalf of COLABIOCLI WG-PRE-LATAM. *TH Open* 2019;3:e367–76.
73. Miller GJ. Postprandial lipaemia and haemostatic factors. *Atherosclerosis* 1998;141(Suppl 1):S47–51.
74. Silveira A. Posprandial triglycerides and blood coagulation. *Exp Clin Endocrinol Diabetes* 2001;109:S527–32.
75. Lima-Oliveira G, Salvagno GL, Lippi G, et al. Could light meal jeopardize laboratory coagulation tests? *Biochem Med (Zagreb)* 2014;24:343–9.
76. Robertson D, Fröhlich JC, Carr RK, et al. Effects of caffeine on plasma renin activity, catecholamines and blood pressure. *N Engl J Med* 1978;298:181–6.
77. Bergman EA, Massey LK, Wise KJ, et al. Effects of dietary caffeine on renal handling of minerals in adult women. *Life Sci* 1990;47:557–64.
78. Farias-Pereira R, Park CS, Park Y. Mechanisms of action of coffee bioactive components on lipid metabolism. *Food Sci Biotechnol* 2019;28(5):1287–96.
79. Superko HR, Bortz Jr W, Williams PT, et al. Caffeinated and decaffeinated coffee effects on plasma lipoprotein cholesterol, apolipoproteins, and lipase activity: a controlled, randomized trial. *Am J Clin Nutr* 1991;54:599–605.
80. Boekema PJ, Samsom M, van Berge Henegouwen GP, et al. Coffee and gastrointestinal function: facts and fiction. *Scand J Gastroenterol Suppl* 1999;230:35–9.
81. Grunst J, Dietze G, Wicklmayr M, et al. Einfluss von Ethanol auf den Purinkatabolismus der menschlichen Leber. *Verh Dtsch Ges Inn Med* 1973;79:914–7.
82. Statland BE, Winkel P. Effects of preanalytical sources of variation. In: Gräsbeck R, Alström T, editors. *Reference values in laboratory medicine*. Chichester: Wiley; 1981. p. 127–37.
83. Freer DE, Statland BE. The effects of ethanol (0.75 g/kg body weight) on the activities of selected enzymes in sera of healthy adults: 1. Intermediate-term effects. *Clin Chem* 1977;23:830–4.
84. Bortolotti F, Sorio D, Bertaso A, Tagliaro F. Analytical and diagnostic aspects of carbohydrate deficient transferrin (CDT): a critical review over years 2007–2017. *J Pharm Biomed Anal* 2018;147:2–12. doi:10.1016/j.jpba.2017.09.006.

85. Andresen-Streichert H, Müller A, Glahn A, Skopp G, Sterneck M. Alcohol biomarkers in clinical and forensic contexts. *Dtsch Arztbl Int* 2018;115(18):309–15.
86. Stibler H. Carbohydrate-deficient transferrin in serum: a new marker of potentially harmful alcohol consumption reviewed. *Clin Chem* 1991;37:2029–37.
87. Rico H. Alcohol and bone disease. *Alcohol Alcohol* 1990;25: 345–52.
88. Bode JC. Die internistischen Folgeerkrankungen des Alkoholismus. In: Kisker KP, Lauter H, Meyer JE, et al., editors. *Psychiatrie der Gegenwart: Abhängigkeit und Sucht*. 3rd ed. Berlin: Springer; 1987. p. 206–41.
89. Grassi G, Seravalle G, Calhoun DA, et al. Cigarette smoking and the adrenergic nervous system. *Clin Exp Hypertens A* 1992;14:251–60.
90. Frati AC, Iniestra F, Ariza CR. Acute effect of cigarette smoking on glucose tolerance and other cardiovascular risk factors. *Diabetes Care* 1996;19:112–18.
91. Ainsworth MA, Hogan DL, Koss MA, et al. Cigarette smoking inhibits acid-stimulated duodenal mucosal bicarbonate secretion. *Ann Intern Med* 1993;119:882–6.
92. Lee HD, Lee HS, Lee JS, et al. Do cigarette smoking and obesity affect semen abnormality in idiopathic infertile males? *World J Mens Health* 2014;32:105–9.
93. Wannamethee SG, Lowe GD, Shaper AG, et al. Associations between cigarette smoking, pipe/cigar smoking, and smoking cessation, and haemostatic and inflammatory markers for cardiovascular disease. *Eur Heart J* 2005;26:1765–73.
94. Ahlgren C, Pottgiesser T, Robinson N, Sottas PE, Ruecker G, Schumacher YO. Are 10 min of seating enough to guarantee stable haemoglobin and haematocrit readings for the athlete's biological passport? *Int J Lab Hematol* 2010;32:506–11.
95. Astolfi T, Schumacher YO, Crettaz von Roten F, Saugy M, Faiss R. Does body position before and during blood sampling influence the athlete biological passport variables? *Int J Lab Hematol* 2020;42:61–7.
96. Lima-Oliveira G, Guidi GC, Salvagno GL, et al. Patient posture for blood collection by venipuncture: recall for standardization after 28 years. *Rev Bras Hematol Hemoter* 2017;39:127–32.
97. Lippi G, Salvagno GL, Lima-Oliveira G, et al. Influence on posture on routine hemostasis testing. *Blood Coagul Fibrinolysis* 2015;26:716–9.
98. Renoe BW, McDonald JM, Ladenson JH. Influence of posture on free calcium and related compounds. *Clin Chem* 1979;25: 1766–9.
99. Kuipers H, Brouwer T, Dubravcic-Simunjak S, et al. Hemoglobin and hematocrit values after saline infusion and tourniquet. *Int J Sports Med* 2005;26:405–8.
100. Lippi G, Salvagno GL, Montagnana M, Franchini M, Guidi GC. Venous stasis and routine hematologic testing. *Clin Lab Haematol* 2006;28:332–7.
101. Lippi G, Salvagno GL, Montagnana M, Guidi GC. Short-term venous stasis influences routine coagulation testing. *Blood Coagul Fibrinolysis* 2005;16:453–9.
102. Sanchis-Gomar F, Lippi G. Physical activity—An important preanalytical variable. *Biochem Med* 2014;24:68–79.
103. Lombardi G, Ricci C, Banfi G. Effect of winter swimming on haematological parameters. *Biochem Med* 2011;21:71–8.
104. Don BR, Sebastian A, Cheitlin M, et al. Pseudohyperkalemia caused by fist clenching during phlebotomy. *N Engl J Med* 1990;322:1290–2.
105. Lamprecht M, Moussali H, Ledinski G, et al. Effects of a single bout of walking exercise on blood coagulation parameters in obese women. *J Appl Physiol* 2013;115: 57–63.
106. Naesh O, Hindberg I, Trap-Jensen J, et al. Post-exercise platelet activation-aggregation and release in relation to dynamic exercise. *Clin Physiol* 1990;10:221–30.
107. Corsetti R, Lombardi G, Barassi A, et al. Cardiac indexes, cardiac damage biomarkers and energy expenditure in professional cyclists during the Giro d'Italia 3-weeks stage race. *Biochem Med* 2012;22:237–46.
108. Kratz A, Lewandrowski KB, Siegel AJ, et al. Effect of marathon running on hematologic and biochemical laboratory parameters, including cardiac markers. *Am J Clin Pathol* 2002;118:856–63.
109. Tjora S, Gjestland H, Mordal S, et al. Troponin rise in healthy subjects during exercise test. *Int J Cardiol* 2011;151:375–6.
110. Lippi G, Salvagno GL, Danese E, Tarperi C, Guidi GC, Schena F. Variation of red blood cell distribution width and mean platelet volumen after moderate endurance exercise. *Adv Hematol* 2014;2014:192173.
111. Romagnoli M, Alis R, Aloe R, et al. Influence of training and a maximal exercise test in analytical variability of muscular, hepatic, and cardiovascular biochemical variables. *Scand J Clin Lab Invest* 2014;74:192–8.
112. Lippi G, Banfi G, Montagnana M, Salvagno GL, Schena F, Guidi GC. Acute variation of leucocytes counts following a half-marathon run. *Int J Lab Hematol* 2010;32:117–21.
113. Williams ME, Gervino EV, Rosa RM, et al. Catecholamine modulation of rapid potassium shifts during exercise. *N Engl J Med* 1985;312:823–7.
114. Linton RAF, Lim M, Wolff CB, et al. Arterial plasma potassium measured continuously during exercise in man. *Clin Sci* 1984; 67:427–31.
115. Garrett WE, Donald T. Exercise and sport science. Philadelphia: Lippincott Williams & Wilkins; 2000.
116. Kanaley JA, Weltman JY, Pieper KS, et al. Cortisol and growth hormone responses to exercise at different times of day. *J Clin Endocrinol Metab* 2001;86:2881–9.
117. Simundic AM, Cornes M, Grankvist K, et al. Standardization of collection requirements for fasting samples: for the Working Group on Preanalytical Phase (WG-PA) of the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM). *Clin Chim Acta* 2014;432:33–7.
118. Heil W, Ehrhardt V. Reference ranges for adults and children: Preanalytical considerations. 8th ed. Mannheim: Roche Diagnostics; 2012.
119. Harris EK, Wong ET, Shaw Jr ST. Statistical data for separate reference intervals: race and gender groups in creatine kinase. *Clin Chem* 1991;37:1580–2.
120. McKenzie SB, Landis-Piwowar K, Williams JL. Clinical laboratory hematology. 4th edition. New Jersey: Pearson Prentice Hall; 2020.
121. Karayalcin G, Rosner F, Sawitsky A. Pseudo-neutropenia in American negroes. *Lancet* 1972;1:387.
122. Lim E, Miyamura J, Chen JJ. Racial/Ethnic-specific reference intervals for common laboratory tests: a comparison among Asians, Blacks, Hispanics, and White. *Hawaii J Public Health* 2015;74:302–10.
123. Lim EM, Cembrowski G, Cembrowski M, Clarke G. Race-specific WBC and neutrophil count reference intervals. *Int J Lab Hematol* 2010;32:590–7.

124. Thomas L. Clinical laboratory diagnostics. Frankfurt/Main: TH book; 1998.
125. Abbassi-Ghanavati M, Greer LG, Cunningham FG. Pregnancy and laboratory studies: a reference table for clinicians. *Obstet Gynecol* 2009;114:1326–31.
126. Available from: <http://www.caliperdatabase.com/>.
127. Colantonio DA, Kyriakopoulou L, Chan MK, 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* 2012;58: 854–68.
128. Adeli K, Raizman JE, Chen Y, 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* 2015;61:1075–86.
129. Snoeck HW. Aging of the hematopoietic system. *Curr Opin Hematol* 2013;20:355–61.
130. Fischer PE, DeLoughery TG, Schreiber MA. Hematologic changes with aging. In: Yelon JA, Luchette FA, editors. Geriatric trauma and critical care. New York: Springer Science+Business Media; 2014.
131. Chung SS, Park CY. Aging, hematopoiesis, and the myelodysplastic syndromes. *Blood Adv* 2017;1:2572–8.
132. Sandberg S, Fraser CG, Horvath AR, et al. Defining analytical performance specifications: consensus statement from the 1st Strategic Conference of the European Federation of Clinical Chemistry and Laboratory Medicine. *Clin Chem Lab Med* 2015;53:833–5.
133. Bowen RA, Remaley AT. Interferences from blood collection tube components on clinical chemistry assays. *Biochem Med* 2014;24:31–44.
134. Bowen RA, Hortin GL, Csako G, et al. Impact of blood collection devices on clinical chemistry assays. *Clin Biochem* 2010;43:4–25.
135. Simundic AM, Baird G, Cadamuro J, Costelloe SJ, Lippi G. Managing hemolyzed samples in clinical laboratories. *Crit Rev Clin Lab Sci* 2020;57(1):1–21. doi:10.1080/10408363.2019.1664391.
136. Guder WG. Haemolysis as an influence and interference factor in clinical chemistry. *J Clin Chem Clin Biochem* 1986;24:125–6.
137. Simundic AM, Nikolac N, Vukasovic I, Vrkic N. The prevalence of preanalytical errors in a Croatian ISO 15189 accredited laboratory. *Clin Chem Lab Med* 2010;48:1009–14.
138. Ong ME, Chan YH, Lim CS. Reducing blood sample hemolysis at a tertiary hospital emergency department. *Am J Med* 2009;122:1054.e1–6.
139. Jones BA, Calam RR, Howanitz PJ. Chemistry specimen acceptability: a College of American Pathologists Q-Probes study of 453 laboratories. *Arch Pathol Lab Med* 1997;121:19–26.
140. Lippi G, Salvagno GL, Favaloro EJ, et al. Survey on the prevalence of hemolytic specimens in an academic hospital according to collection facility: opportunities for quality improvement. *Clin Chem Lab Med* 2009;47:616–8.
141. Burns ER, Yoshikawa N. Hemolysis in serum samples drawn by emergency department personnel versus laboratory phlebotomists. *Lab Med* 2002;33:378–80.
142. Vecellio E, Li L, Mackay M, et al. A benchmark study of the frequency and variability of haemolysis reporting across pathology laboratories—The implications for quality use of pathology and safe and effective patient care. Report to Royal College of Pathologists Australasia. Australian Institute of Health Innovation, Macquarie University, Sydney; July 2015.
143. Cook SL, Bruns DE. Persistent hemolysis in a patient with pancreatitis. *Clin Chem* 2012;58:974–7.
144. Carraro P, Servidio G, Plebani M. Hemolyzed specimens: a reason for rejection or a clinical challenge? *Clin Chem* 2000;46:306–7.
145. Cooper CE, Schaer DJ, Buehler PW, et al. Haptoglobin binding stabilizes hemoglobin ferryl iron and the globin radical on tyrosine β145. *Antioxid Redox Signal* 2013;18: 2264–73.
146. Barbaryan A, Iyinagoro C, Nwankwo N, et al. Ibuprofen-induced hemolytic anemia. *Case Rep Hematol* 2013;2013: 142865.
147. Prueksaritanond S, Barbaryan A, Mirrakhimov AE, et al. A puzzle of hemolytic anemia, iron and vitamin B12 deficiencies in a 52-year-old male. *Case Rep Hematol* 2013; 2013:708489.
148. Fairbanks VF, Ziesmer SC, O'Brien PC. Methods for measuring plasma hemoglobin in micromolar concentration compared. *Clin Chem* 1992;38:132–40.
149. Walker HK, Hall WD, Hurst JW. Clinical methods: the history, physical, and laboratory examinations. 3rd ed. Emory University School of Medicine, Atlanta, Georgia. Boston: Butterworths; 1990.
150. Lippi G, Cervellin G, Favaloro EJ, et al. In vitro and in vivo hemolysis: an unresolved dispute in laboratory medicine. Berlin: Walter de Gruyter; 2012.
151. Elin RJ, Hosseini JM. Magnesium content of mononuclear blood cells. *Clin Chem* 1985;31:377–80.
152. Bugg NC, Jones JA. Hypophosphataemia: pathophysiology, effects and management on the intensive care unit. *Anaesthesia* 1998;53:895–902.
153. Koseoglu M, Hur A, Atay A, et al. Effects of hemolysis interferences on routine biochemistry parameters. *Biochem Med* 2011;21:79–85.
154. Nijsten MWN, Dofferhoff ASM. Pseudohyperkalemia and platelet counts. *N Engl J Med* 1991;325:1107.
155. Szasz G, Gerhardt W, Gruber W. Creatine kinase in serum: 3. Further study of adenylate kinase inhibitors. *Clin Chem* 1977; 23:1888–92.
156. Matsuura S, Igarashi M, Tanizawa Y, et al. Human adenylate kinase deficiency associated with hemolytic anemia: a single base substitution affecting solubility and catalytic activity of the cytosolic adenylate kinase. *J Biol Chem* 1989;264: 10148–55.
157. Szasz G, Gerhardt W, Gruber W, et al. Creatine kinase in serum: 2. Interference of adenylate kinase with the assay. *Clin Chem* 1976;22:1806–11.
158. van der Woerd-de Lange JA, Guder WG, Schleicher E, et al. Studies on the interference by haemoglobin in the determination of bilirubin. *J Clin Chem Clin Biochem* 1983;21: 437–43.
159. Sodi R, Darn SM, Davison AS, et al. Mechanism of interference by haemolysis in the cardiac troponin T immunoassay. *Ann Clin Biochem* 2006;43:49–56.
160. Snyder JA, Rogers MW, King MS, et al. The impact of hemolysis on Ortho-Clinical Diagnostic's ECi and Roche's Elecsys immunoassay systems. *Clin Chim Acta* 2004;348:181–7.
161. Bellomo G, Sulias MG, Mairate E, et al. Hemolysis is a major cause of variability in insulin measurement during oral glucose tolerance test in children. *Clin Lab* 2012;58:67–74.

162. Bais R. The effect of sample hemolysis on cardiac troponin I and T assays. *Clin Chem* 2010;56:1357–9.
163. Florkowski C, Wallace J, Walmsley T, et al. The effect of hemolysis on current troponin assays—a confounding preanalytical variable? *Clin Chem* 2010;56:1195–7.
164. Dasgupta A, Wells A, Biddle DA. Negative interference of bilirubin and hemoglobin in the MEIA troponin I assay but not in the MEIA CK-MB assay. *J Clin Lab Anal* 2001;15:76–80.
165. Kroll MH, McCudden CR. Endogenous interferences in clinical laboratory tests. *Icteric, Lipemic and Turbid Samples*. Berlin: Walter De Gruyter; 2012.
166. Nikolac N. Lipemia: causes, interference mechanisms, detection and management. *Biochem Med* 2014;24:57–67.
167. Bossuyt X, Schiettekatte G, Bogaerts A, et al. Serum protein electrophoresis by CZE 2000 clinical capillary electrophoresis system. *Clin Chem* 1998;44:749–59.
168. Bossuyt X. Separation of serum proteins by automated capillary zone electrophoresis. *Clin Chem Lab Med* 2003;41:762–72.
169. Tate J, Ward G. Interferences in immunoassay. *Clin Biochem Rev* 2004;25:105–20.
170. Dimeski G, Mollee P, Carter A. Increased lipid concentration is associated with increased hemolysis. *Clin Chem* 2005;51:2425.
171. Clinical and Laboratory Standards Institute (CLSI). C56-A: Hemolysis, icterus, and lipemia/turbidity indices as indicators of interference in clinical laboratory analysis—Approved guideline. Wayne PA, USA; 2012.
172. Dimeski G, Jones BW. Lipaemic samples: effective process for lipid reduction using high speed centrifugation compared with ultracentrifugation. *Biochem Med* 2011;21:86–94.
173. Dimeski G. Interference testing. *Clin Biochem Rev* 2008;29(Suppl 1):S43–8.
174. Ferraz TP, Fiúza MC, Dos Santos ML, et al. Comparison of six methods for the extraction of lipids from serum in terms of effectiveness and protein preservation. *J Biochem Biophys Methods* 2004;58:187–93.
175. Saracevic A, Nikolac N, Simundic AM. The evaluation and comparison of consecutive high speed centrifugation and LipoClear reagent for lipemia removal. *Clin Biochem* 2014;47:309–14.
176. Bartlett D. Intravenous lipids: antidotal therapy for drug overdose and toxic effects of local anesthetics. *Crit Care Nurse* 2014;34:62–6.
177. Christian MR, Pallasch EM, Wahl M, et al. Lipid rescue 911: are poison centers recommending intravenous fat emulsion therapy for severe poisoning? *J Med Toxicol* 2013;9:231–4.
178. American College of Medical Toxicology. ACMT position statement: interim guidance for the use of lipid resuscitation therapy. *J Med Toxicol* 2011;7:81–2.
179. Bucklin MH, Gorodetsky RM, Wiegand TJ. Prolonged lipemia and pancreatitis due to extended infusion of lipid emulsion in bupropion overdose. *Clin Toxicol (Phila)* 2013;51:896–8.
180. Grunbaum AM, Gilfix BM, Gosselin S, et al. Analytical interferences resulting from intravenous lipid emulsion. *Clin Toxicol (Phila)* 2012;50:812–7.
181. Agarwal S, Vargas G, Nordstrom C, et al. Effect of interference from hemolysis, icterus and lipemia on routine pediatric clinical chemistry assays. *Clin Chim Acta* 2015;438:241–5.
182. Steen G, Vermeer HJ, Naus AJ, et al. Multicenter evaluation of the interference of hemoglobin, bilirubin and lipids on Synchro LX-20 assays. *Clin Chem Lab Med* 2006;44:413–9.
183. Ji JZ, Meng QH. Evaluation of the interference of hemoglobin, bilirubin, and lipids on Roche Cobas 6000 assays. *Clin Chim Acta* 2011;412:1550–3.
184. Cobbaert CM, Baadenhuijsen H, Weykamp CW. Prime time for enzymatic creatinine methods in pediatrics. *Clin Chem* 2009;55:549–58.
185. Hermida FJ, Lorenzo MJ, Pérez A, et al. Comparison between ADVIA chemistry systems enzymatic creatinine\_2 method and ADVIA chemistry systems creatinine method (kinetic Jaffe method) for determining creatinine. *Scand J Clin Lab Invest* 2014;74:629–36.
186. Dimeski G, McWhinney B, Jones B, et al. Extent of bilirubin interference with Beckman creatinine methods. *Ann Clin Biochem* 2008;45(Pt 1):91–2.
187. Owen LJ, Keevil BG. Does bilirubin cause interference in Roche creatinine methods? *Clin Chem* 2007;53:370–1.
188. Yamauchi Y, Yamanouchi I. Initial response of serum bilirubin levels to phototherapy. *Biol Neonate* 1991;60:314–9.
189. Simundic AM, Topic E, Nikolac N, et al. Hemolysis detection and management of hemolysed specimens. *Biochem Med* 2010;20:154–9.
190. Glick MR, Ryder KW, Glick SJ, et al. Unreliable visual estimation of the incidence and amount of turbidity, hemolysis, and icterus in serum from hospitalized patients. *Clin Chem* 1989;35:837–9.
191. Simundic AM, Nikolac N, Ivankovic V, et al. Comparison of visual vs. automated detection of lipemic, icteric and hemolyzed specimens: can we rely on a human eye? *Clin Chem Lab Med* 2009;47:1361–5.
192. Hawkins R. Discrepancy between visual and spectrophotometric assessment of sample haemolysis. *Ann Clin Biochem* 2002;39:521–2.
193. Jeffery J, Sharma A, Ayling RM. Detection of haemolysis and reporting of potassium results in samples from neonates. *Ann Clin Biochem* 2009;46:222–5.
194. Lippi G, Daves M, Mattiuzzi C. Interference of medical contrast media on laboratory testing. *Biochem Med* 2014;24:80–8.
195. Aslan B, Stemp J, Catomeris P, et al. Clinical chemistry sample interferences reporting patterns in Ontario laboratories. *Clin Chem* 2012;58:A23–4.
196. Simundic AM, Bilic-Zulle L, Nikolac N, et al. The quality of the extra-analytical phase of laboratory practice in some developing European countries and Mexico—A multicentric study. *Clin Chem Lab Med* 2011;49:215–28.
197. Bilic-Zulle L, Simundic AM, Supak Smolcic V, et al. Self-reported routines and procedures for the extra-analytical phase of laboratory practice in Croatia—Cross-sectional survey study. *Biochem Med* 2010;20:64–74.
198. Lippi G, Avanzini P, Campioli D, et al. Systematical assessment of serum indices does not impair efficiency of clinical chemistry testing: a multicenter study. *Clin Biochem* 2013;46:1281–4.
199. Dolci A, Panteghini M. Harmonization of automated hemolysis index assessment and use: is it possible? *Clin Chim Acta* 2014;432:38–43.
200. Lippi G, Luca Salvagno G, Blancaert N, et al. Multicenter evaluation of the hemolysis index in automated clinical chemistry systems. *Clin Chem Lab Med* 2009;47:934–9.
201. CLSI EP7-A2. *Interference testing in clinical chemistry* approved guideline, second edition, (ISBN 1-56238-584-4). Clinical and Laboratory Standards Institute (CLSI), Wayne, PA.

202. Fernandez P, Llopis MA, Perich C, et al. Harmonization in hemolysis detection and prevention: a working group of the Catalonian Health Institute (ICS) experience. *Clin Chem Lab Med* 2014;52:1557–68.
203. Nikolac N, Simundic AM, Miksa M, et al. Heterogeneity of manufacturers' declarations for lipemia interference—An urgent call for standardization. *Clin Chim Acta* 2013;426: 33–40.
204. Kristensen GB, Aakre KM, Kristoffersen AH, et al. How to conduct external quality assessment schemes for the pre-analytical phase? *Biochem Med (Zagreb)* 2014;24:114–22.
205. Lippi G, Banfi G, Church S, et al. Preanalytical quality improvement. In pursuit of harmony, on behalf of European Federation for Clinical Chemistry and Laboratory Medicine (EFLM) Working Group for Preanalytical Phase (WG-PRE). *Clin Chem Lab Med* 2015;53:357–70.
206. Thomas MA, Davies G, Jones S, et al. A pre-analytical EQA scheme for sample integrity: a WEQAS study to monitor the effectiveness of serum indices. *Clin Chem Lab Med* 2015;53(Suppl 4):eA1–89.
207. Lippi G, Cadamuro J, von Meyer A, et al. Local quality assurance of serum or plasma (HIL) indices. *Clin Biochem* 2018;54:112–8. doi:10.1016/j.clinbiochem.2018.02.018.
208. Darby D, Broomhead C. Interference with serum indices measurement, but not chemical analysis, on the Roche Modular by Patent Blue V. *Ann Clin Biochem* 2008;45: 289–92.
209. McTaggart MP, Cannon LP, Kearney EM. Effect of Patent Blue V dye on sample interference indices on the Abbott Architect. *Ann Clin Biochem* 2012;49:510–11.
210. Thompson JF, Agarwala SS, Smithers BM, et al. Phase 2 study of intralesional PV-10 in refractory metastatic melanoma. *Ann Surg Oncol* 2015;22:2135–42.
211. Dimeski G, Jones B, Ungerer JP. Interference from rose bengal with total bilirubin measurement. *Clin Chem* 2009;55: 1040–1.
212. Monk C, Wallage M, Wassell J, et al. A monoclonal protein identified by an anomalous lipaemia index. *Ann Clin Biochem* 2009;46(Pt 3):250–2.
213. Fliser E, Jerkovic K, Vidovic T, et al. Investigation of unusual high serum indices for lipemia in clear serum samples on Siemens analysers dimension. *Biochem Med* 2012;22: 352–641.
214. Lippi G, Banfi G, Buttarello M, et al. Recommendations for detection and management of unsuitable samples in clinical laboratories. *Clin Chem Lab Med* 2007;45:728–36.
215. Ismail A, Shingler W, Seneviratne J, et al. In vitro and in vivo haemolysis and potassium measurement. *Brit Med J* 2005; 330:949.
216. Sheppard CA, Allen RC, Austin GE, et al. Paraprotein interference in automated chemistry analyzers. *Clin Chem* 2005;51:1077–8.
217. Barbier A, Vuillaume I, Baras A, et al. Interferences by a monoclonal IgM in biochemical analyses: detection and recommendations. [Article in French] *Ann Biol Clin (Paris)* 2007;65:411–5.
218. Dimeski G, Carter A. Rare IgM interference with Roche/Hitachi Modular glucose and gamma-glutamyltransferase methods in heparin samples. *Clin Chem* 2005;51:2202–4.
219. John R, Oleesky D, Issa B, et al. Pseudohypercalcaemia in two patients with IgM paraproteinaemia. *Ann Clin Biochem* 1997;34:694–6.
220. Chakraborty S, Sen S, Gupta D, et al. Spurious hyperphosphatemia in a case of multiple myeloma. *Indian J Clin Biochem* 2014;29:250–2.
221. Cohen AM, Magazanik A, van-der Lijn E, et al. Pseudohyperphosphataemia incidence in an automatic analyzer. *Eur J Clin Chem Clin Biochem* 1994;32:559–61.
222. Bakker A, Boxma H, Christen P. Influence of monoclonal immunoglobulins in three different methods for inorganic phosphorus. *Ann Clin Biochem* 1990;27:227–31.
223. Bowles SA, Tait RC, Jefferson SG, et al. Characteristics of monoclonal immunoglobulins that interfere with serum inorganic phosphate measurement. *Ann Clin Biochem* 1994;31:249–54.
224. Mandry JM, Posner MR, Tucci JR, et al. Hyperphosphatemia in multiple myeloma due to a phosphate binding immunoglobulin. *Cancer* 1991;68:1092–4.
225. Bakker A. Influence of monoclonal immunoglobulins in direct determinations of iron in serum. *Clin Chem* 1991;37: 690–4.
226. Smogorzewska A, Flood JG, Long WA, et al. Paraprotein interference in automated chemistry analyzer. *Clin Chem* 2004;50:1691–3.
227. Pantanowitz L, Horowitz GL, Upalakalin JN, et al. Artifactual hyperbilirubinemia due to paraprotein interference. *Arch Pathol Lab Med* 2003;127:55–9.
228. Nauti A, Barassi A, Merlini G, et al. Paraprotein interference in an assay of conjugated bilirubin. *Clin Chem* 2005;51: 1076–7.
229. Yang Y, Howanitz PJ, Howanitz JH, et al. Paraproteins are a common cause of interferences with automated chemistry methods. *Arch Pathol Lab Med* 2008;132:217–23.
230. Berth M, Delanghe J. Protein precipitation as a possible important pitfall in the clinical chemistry analysis of blood samples containing monoclonal immunoglobulins: 2 case reports and a review of the literature. *Acta Clin Belg* 2004;59:263–73.
231. Datta P, Graham GA, Schoen I. Interference by IgG paraproteins in the Jaffe method for creatinine determination. *Am J Clin Pathol* 1986;85:463–8.
232. Smith JD, Nobiletti J, Freed M, et al. Interference with the Astra 8 and Synchron CX3 assays of urea nitrogen in serum by a high-M(r) inhibitor in a patient with multiple myeloma. *Clin Chem* 1992;38:598–9.
233. Yu A, Pira U. False increase in serum C-reactive protein caused by monoclonal IgM-lambda: a case report. *Clin Chem Lab Med* 2001;39:983–7.
234. Yamada K, Yagihashi A, Ishii S, et al. Interference with nephelometric assay of C-reactive protein and antistreptolysin-O by monoclonal IgM-k from a myeloma patient. *Clin Chem* 1997;43:2435–7.
235. Schnebelen A, Sweat K, Marshall J, et al. Alleviation of IgM monoclonal protein interference in nephelometric assays by sample treatment with reducing agent in a chaotropic salt solution. *Am J Clin Pathol* 2012;137:26–8.
236. Covinsky M, Laterza O, Pfeifer PD, et al. An IgM-λ antibody to *Escherichia coli* produces false positive results in multiple immunometric assays. *Clin Chem* 2000;46:1157–61.
237. Imperiali M, Jelmini P, Ferraro B, et al. Interference in thyroid-stimulating hormone determination. *Eur J Clin Invest* 2010;40:756–8.
238. Dimeski G, Bassett K, Brown N. Paraprotein interference with turbidimetric gentamicin assay. *Biochem Med* 2015;25:117–24.

239. Gunther M, Saxinger L, Gray M, et al. Two suspected cases of immunoglobulin-mediated interference causing falsely low vancomycin concentrations with the Beckman PETINIA method. *Ann Pharmacother* 2013;47:e19.
240. LeGatt DF, Blakney GB, Higgins TN, et al. The effect of paraproteins and rheumatoid factor on four commercial immunoassays for vancomycin: implications for laboratorians and other health care professionals. *Ther Drug Monit* 2012;34:306–11.
241. Simons SA, Molinelli AR, Sobhani K, et al. Two cases with unusual vancomycin measurements. *Clin Chem* 2009;55: 578–80.
242. Brauchli YB, Scholer A, Schwietert M, et al. Undetectable phenytoin serum levels by an automated particle-enhanced turbidimetric inhibition immunoassay in a patient with monoclonal IgM lambda. *Clin Chim Acta* 2008;389:174–6.
243. Dutta AK. A curious case of hyperbilirubinemia. *Indian J Clin Biochem* 2012;27:200–1.
244. Bakker AJ, Mücke M. Gammopathy interference in clinical chemistry assays: mechanisms, detection and prevention. *Clin Chem Lab Med* 2007;45:1240–3.
245. Dimeski G, Morgan TJ, Presneill JJ, et al. Disagreement between ion selective electrode direct and indirect sodium measurements: estimation of the problem in a tertiary referral hospital. *J Crit Care* 2012;27:326.e9–16.
246. Lyon AW, Baskin LB. Pseudohyponatremia in a myeloma patient: direct electrode potentiometry is a method worth its salt. *LabMed* 2003;34:357–60.
247. CLSI EP07-ED3:2018 Interference testing in clinical chemistry guideline, 3rd edition, (ISBN 1-56238-847-9). Clinical and Laboratory Standards Institute (CLSI), Wayne, PA.
248. AACC Effects on clinical laboratory tests; Drugs, disease, herbs & natural products, Wiley. Available from <http://clinfx.wiley.com/aaccweb/aacc/login>. Accessed January 2020.
249. Dimeski G. Interference testing. *Clin Biochem Rev* 2008;29: S43–8.
250. Yao H, Rayburn ER, Shi Q, et al. FDA-approved drugs that interfere with laboratory tests: a systematic search of U.S. drug labels. *Crit Rev Clin Lab Sci* 2017;54:1–17.
251. Sonntag O. Analytical interferences and analytical quality. *Clin Chim Acta* 2009;404:37–40.
252. van Balveren JA, Verboeket-van de Venne WPHG, Erdem-Eraslan L, et al. Impact of interactions between drugs and laboratory test results on diagnostic test interpretation - a systematic review. *Clin Chem Lab Med* 2018;56:2004–9.
253. Kailajärvi M, Takala T, Grönroos P, et al. Reminders of drug effects on laboratory test results. *Clin Chem* 2000;46:1395–1400.
254. Green AJ, Halloran SP, Mould GP, et al. Interference by newer cephalosporins in current methods for measuring creatinine. *Clin Chem* 1990;36:2139–40.
255. Kaburaki J, Yamada M, Kamikawara M, et al. In vivo and in vitro positive interference by cefpirome in measurement of serum creatinine by the Jaffe method. *Keio J Med* 1999;48:93–6.
256. Luna-Záizar H, Virgen-Montelongo M, Cortez-Álvarez CR, et al. In vitro interference by acetaminophen, aspirin, and metamizole in serum measurements of glucose, urea, and creatinine. *Clin Biochem* 2015;48:538–41.
257. Fliser E, Modrič E, Klavž J, et al. Influence of metamizole on serum creatinine measurements (case report). 2nd EFLM-BD European conference on preanalytical phase. *Biochem Med (Zagreb)* 2013;23:A14.
258. Fliser E, Krajnc K. Determination of metamizole interference on measurements of selected analytes in accordance with CLSI standard EP7-A2. 3rd EFLM-BD European Conference on Preanalytical Phase. *Clin Chem Lab Med* 2015;53:eA19–20.
259. Garay RP, Chiavaroli C, Hannaert P. Therapeutic efficacy and mechanism of action of ethamsylate, a long-standing hemostatic agent. *Am J Ther* 2006;13:236–47.
260. Wiewiorka O, Dastych M, Čermáková Z. Strong negative interference of ethamsylate (Dicynone) in serum creatinine quantification via enzymatic assay using Trinder reaction. *Scand J Clin Lab Invest* 2013;73:449–51.
261. Dastych M, Wiewiorka O, Benovská M. Ethamsylate (Dicynone) interference in determination of serum creatinine, uric acid, triglycerides, and cholesterol in assays involving the Trinder reaction—In vivo and in vitro. *Clin Lab* 2014;60:1373–6.
262. Saenger AK, Lockwood C, Snozek CL, et al. Catecholamine interference in enzymatic creatinine assays. *Clin Chem* 2009;55:1732–6.
263. Choy KW, Wijeratne N, Doery JC. Eltrombopag: liver toxicity, kidney injury or assay interference? *Pathology* 2016;48:754–6.
264. Baird GS. Ionized calcium. *Clin Chim Acta* 2011;412:696–701.
265. Hubæk I, Abrahams AC, de Bruin M, et al. Falsely decreased ionized calcium results due to analytical interference by teriflunomide, the active metabolite of leflunomide (Arava). *Clin Chem Lab Med* 2012;50:755–6.
266. Gruber M, Nehring C, Creutzzenberg M, et al. Perchlorate (Irenat) may falsely lower measured ionised calcium. *Clin Chem Lab Med* 2011;49:1019–24.
267. Schiemsky T, Brandt I. Bicarbonate interference on cobas 6000 c501 chloride ion-selective electrodes. *Clin Chem Lab Med* 2018;56:e214–5.
268. Feyen BF, Coenen D, Jorens PG, et al. Falsely elevated sodium levels during thiopental treatment in the ICU: technical interference on a laboratory device with important clinical relevance. *Neurocrit Care* 2013;18:64–9.
269. Halbmayer WM, Weigel G, Quehenberger P, et al. Interference of the new oral anticoagulant dabigatran with frequently used coagulation tests. *Clin Chem Lab Med* 2012;50:1601–5.
270. Amanatullah DF, Lopez MJ, Gosselin RC, et al. Artificial elevation of prothrombin time by telavancin. *Clin Orthop Relat Res* 2013;471:332–5.
271. Vidali M, Bianchi V, Bagnati M, et al. False negativity to carbohydrate-deficient transferrin and drugs: a clinical case. *Biochem Med* 2014;24:175–9.
272. Saitman A, Park HD, Fitzgerald RL. False-positive interferences of common urine drug screen immunoassays: a review. *J Anal Toxicol* 2014;38:387–96.
273. Reidy L, Walls C, Steele BW. Crossreactivity of bupropion metabolite with enzyme-linked immunosorbent assays designed to detect amphetamine in urine. *Ther Drug Monit* 2011;33:366–8.
274. Kingery JM, Radke JB, Maakestad J, et al. Data on hydroxy-chloroquine interference with urine laboratory testing. *Data Brief* 2019;27:104781.
275. Reisfield GM, Chronister CW, Goldberger BA, et al. Unexpected urine drug testing results in a hospice patient on high-dose morphine therapy. *Clin Chem* 2009;55:1765–8.
276. da Silva AS, Falkenberg M. Analytical interference of quinolone antibiotics and quinine derived drugs on urinary protein determined by reagent strips and the pyrogallol red-molybdate protein assay. *Clin Biochem* 2011;44:1000–4.

277. Scotti da Silva-Colombelli A, Falkenberg M. Analytical interferences of drugs in the chemical examination of urinary protein. *Clin Biochem* 2007;40(13-14):1074–6.
278. Joyce C, Melvin A, Costeloe S, et al. Case report of phantom phaeochromocytoma. 5th EFLM Conference on Preanalytical Phase Zagreb. *Clin Chem Lab Med* 2019;57(4):eA9–10.
279. Danese E, Salvagno GL, Guzzo A, et al. Urinary free cortisol assessment by liquid chromatography tandem mass spectrometry: a case study of ion suppression due to unacquainted administration of piperacillin. *Biochem Med (Zagreb)* 2017;27:031001.
280. Zhang Z, Hu W, Li L, et al. Therapeutic monoclonal antibodies and clinical laboratory tests: when, why, and what is expected? *J Clin Lab Anal* 2018;32(3):e22307.
281. Genzen JR, Kawaguchi KR, Furman RR. Detection of a monoclonal antibody therapy (ofatumumab) by serum protein and immunofixation electrophoresis. *Br J Haematol* 2011;155:123–5.
282. McCudden CR, Voorhees VR, Hainsworth SA, et al. Interference of monoclonal antibody therapies with serum protein electrophoresis tests. *Clin Chem* 2010;56:1897–1904.
283. Chen PP, Tormey CA, Eisenbarth SC, et al. False-positive light chain clonal restriction by flow cytometry in patients treated with alemtuzumab: potential pitfalls for the misdiagnosis of B-cell neoplasms. *Am J Clin Pathol* 2019;151:154–63.
284. Velliquette RW, Aeschlimann J, Kirkegaard J, et al. Monoclonal anti-CD47 interference in red cell and platelet testing. *Transfusion* 2019;59:730–7.
285. Murata K, McCash SI, Carroll B, et al. Treatment of multiple myeloma with monoclonal antibodies and the dilemma of false positive M-spikes in peripheral blood. *Clin Biochem* 2018;51:66–71.
286. Jialal I, Pahwa R. Quantification of daratumumab in the serum protein electrophoresis. *Clin Chem Lab Med* 2017;55:e27–8.
287. McCudden C, Axel AE, Slaets D, et al. Response to: interference of daratumumab on the serum protein electrophoresis. *Clin Chem Lab Med* 2017;55:e29–30.
288. Deneys V, Thiry C, Frelik A, et al. Daratumumab: Therapeutic asset, biological trap! *Transfus Clin Biol* 2018;25:2–7.
289. Thoren KL, Pianko MJ, Maakaroun Y, et al. Distinguishing drug from disease by use of the Hydrashift 2/4 daratumumab assay. *J Appl Lab Med* 2019;3:857–63.
290. van de Donk NW, Otten HG, El Haddad O, et al. Interference of daratumumab in monitoring multiple myeloma patients using serum immunofixation electrophoresis can be abrogated using the daratumumab IFE reflex assay (DIRA). *Clin Chem Lab Med* 2016;54:1105–9.
291. Liu L, Shurin MR, Wheeler SE. A novel approach to remove interference of therapeutic monoclonal antibody with serum protein electrophoresis. *Clin Biochem* 2020;75:40–7.  
doi:10.1016/j.clinbiochem.2019.10.011.
292. Hamilton RG. Accuracy of US Food and Drug Administration-cleared IgE antibody assays in the presence of anti-IgE (omalizumab). *J Allergy Clin Immunol* 2006;117:759–66.
293. Young DS. Effects on clinical laboratory tests: drugs, disease, herbs and natural products. Available from: <http://eu.wiley.com/WileyCDA/WileyTitle/productCd-1118477979.html>.
294. Narayanan S, Young DS. Effects of herbs and natural products on clinical laboratory tests. Washington: AACC-Press; 2007.
295. Kaleta EJ, Jaffe AS, Baumann NA, et al. A case of floating gel. *Clin Chem* 2012;58:1604–5.
296. Daves M, Lippi G, Cosio G, et al. An unusual case of a primary blood collection tube with floating separator gel. *J Clin Lab Anal* 2012;26:246–7.
297. Spiritus T, Zaman Z, Desmet W. Iodinated contrast media interfere with gel barrier formation in plasma and serum separator tubes. *Clin Chem* 2003;49:1187–9.
298. Lin CT, Lee HC, Voon WC, et al. Positive interference from contrast media in cardiac troponin I immunoassays. *Kaohsiung J Med Sci* 2006;22:107–13.
299. Proctor KA, Rao LV, Roberts WL. Gadolinium magnetic resonance contrast agents produce analytic interference in multiple serum assays. *Am J Clin Pathol* 2004;121:282–92.
300. de Cordova CM, Nogara MS, de Cordova MM. Interference on the laboratory measurement of bilirubin: the effect of in vitro interactions. *Clin Chim Acta* 2009;407:77–9.
301. Löwe A, Balzer T, Hirt U. Interference of gadolinium-containing contrast-enhancing agents with colorimetric calcium laboratory testing. *Invest Radiol* 2005;40:521–5.
302. Normann PT, Frøysa A, Svaland M. Interference of gadodiamide injection (OMNISCAN) on the colorimetric determination of serum calcium. *Scand J Clin Lab Invest* 1995;55:421–6.
303. Lin J, Idee JM, Port M, et al. Interference of magnetic resonance imaging contrast agents with the serum calcium measurement technique using colorimetric reagents. *J Pharm Biomed Anal* 1999;21:931–43.
304. Prince MR, Erel HE, Lent RW, et al. Gadodiamide administration causes spurious hypocalcemia. *Radiology* 2003;227:639–46.
305. Ryan JB, Grant S, Walmsley T, et al. Falsely elevated plasma selenium due to gadolinium contrast interference: a novel solution to a preanalytical problem. *Ann Clin Biochem* 2014;51:714–6.
306. Walter A, Nelms S, Harrington CF, et al. Interference of gadolinium on the measurement of selenium in human serum by inductively coupled plasma-quadrupole mass spectrometry. *Ann Clin Biochem* 2011;48:176–7.
307. Harrington CF, Walter A, Nelms S, et al. Removal of the gadolinium interference from the measurement of selenium in human serum by use of collision cell quadrupole inductively coupled plasma mass spectrometry (Q-ICP-MS). *Ann Clin Biochem* 2014;51:386–91.
308. Samson E, Allouche S. Iohexol interference in the  $\alpha$ 2-globulin fraction of the serum protein capillary electrophoresis. *Clin Chem Lab Med* 2015;53:e337–8.
309. Quirós C, Cillero AI, Bretaña A, et al. In vivo interference of Ioversol in serum and urine capillary elecrtoptophoresis: an optimized protocol for sample collection. *Clin Chem Lab Med* 2018;56:e53–5.
310. McCudden CR, Jacobs JFM, Keren D, et al. Recognition and management of common, rare, and novel serum protein electrophoresis and immunofixation interferences. *Clin Biochem* 2018;51:72–9.
311. Simundic AM, Filipi P, Vrataric A, et al. Patient's knowledge and awareness about the effect of the over-the-counter (OTC) drugs and dietary supplements on laboratory test results: a survey in 18 European countries. *Clin Chem Lab Med* 2018;57(2):183–94. doi:10.1515/cclm-2018-0579.
312. Dasgupta A, Bernard DW. Herbal remedies: effects on clinical laboratory tests. *Arch Pathol Lab Med* 2006;130:521–8.
313. Bhuiyan MB, Fant ME, Dasgupta A. Study on mechanism of action of Chinese medicine Chan Su: dose-dependent biphasic production of nitric oxide in trophoblastic BeWo cells. *Clin Chim Acta* 2003;330:179–84.

314. Dasgupta A, Biddle D, Wells A, et al. Positive and negative interference of Chinese medicine Chan Su in serum digoxin measurement: elimination of interference by using a monoclonal chemiluminescent digoxin assay or monitoring free digoxin concentration. *Am J Clin Pathol* 2000;114:174–9.
315. Dasgupta A, Wu S, Actor J, et al. Effect of Asian and Siberian ginseng on serum digoxin measurement by five digoxin immunoassays. Significant variation in digoxin-like immunoreactivity among commercial ginsengs. *Am J Clin Pathol* 2003;119:298–303.
316. Dasgupta A, Tso G, Wells A. Effect of Asian ginseng, Siberian ginseng, and Indian ayurvedic medicine Ashwagandha on serum digoxin measurement by Digoxin III, a new digoxin immunoassay. *J Clin Lab Anal* 2008;22:295–301.
317. Baugher BW, Berman M, Dierksen JE, et al. Digoxin immunoassays on the ARCHITECT i2000SR and ARCHITECT c8000 analyzers are free from interferences of Asian, Siberian, and American ginseng. *J Clin Lab Anal* 2015;29:1–4.
318. Graham-Brown RA, Bourke JF, Bumphrey G. Chinese herbal remedies may contain steroids. *BMJ* 1994;308:473.
319. Saper RB, Kales SN, Paquin J, et al. Heavy metal content of Ayurvedic herbal medicine products. *JAMA* 2004;292:2868–73.
320. Tourbah A, Lebrun-Frenay C, Edan G, et al. MD1003 (high-dose biotin) for the treatment of progressive multiple sclerosis: a randomised, double-blind, placebo-controlled study. *Mult Scler* 2016;22:1719–31.
321. Luong JHT, Vashist SK. Chemistry of Biotin-Streptavidin and the growing concern of an emerging biotin interference in clinical immunoassays. *ACS Omega* 2019;5(1):10–18. doi:10.1021/acsomega.9b03013.
322. Piketty ML, Polak M, Flechtnner L, et al. False biochemical diagnosis of hyperthyroidism in streptavidin-biotin-based immunoassays: the problem of biotin intake and related interferences. *Clin Chem Lab Med* 2017;55:780–8.
323. Li D, Ferguson A, Cervinski MA, et al. AACC guidance document on biotin interference in laboratory tests. *J Appl Lab Med* 2020;5(3):575–87.
324. Samarasinghe S, Meah F, Singh V, et al. Biotin interference with routine clinical immunoassays: understand the causes and mitigate the risks. *Endocr Pract* 2017;23:989–98.
325. Bowen R, Benavides R, Colon-Franco JM, et al. Best practices in mitigating the risk of biotin interference with laboratory testing. *Clin Biochem* 2019;74:1–11.
326. Dasgupta A, Welsh KJ, Hwang SA, et al. Bidirectional (negative/positive) interference of oleandrins and oleander extract on a relatively new Lichi digoxin assay using Vista 1500 analyzer. *J Clin Lab Anal* 2014;28:16–20.
327. Fink SL, Robey TE, Tarabar AF, et al. Rapid detection of convallatoxin using five digoxin immunoassays. *Clin Toxicol (Phila)* 2014;52:659–63.
328. Killorn E, Lim RK, Rieder M. Apparent elevated creatinine after ingestion of nitromethane: interference with the Jaffe reaction. *Ther Drug Monit* 2011;33:1–2.
329. Spielvogel RM, Haddad M. A 5-year-old asymptomatic boy with an elevated serum creatinine level. *J Pediatr* 2012;161:1179.
330. Padmanabhan P, Spiller HA, Ross MP, et al. Is elevated creatinine a reliable marker for methanol toxicity in nitromethane-containing model fuel ingestions in children? *Clin Toxicol (Phila)* 2011;49:45–7.
331. Ngo AS, Rowley F, Olson KR. Case files of the California poison control system, San Francisco division: blue thunder ingestion: methanol, nitromethane, and elevated creatinine. *J Med Toxicol* 2010;6:67–71.
332. Booth C, Naidoo D, Rosenberg A, et al. Elevated creatinine after ingestion of model aviation fuel: interference with the Jaffe reaction by nitromethane. *J Paediatr Child Health* 1999;35:503–4.
333. Cao D, Maynard S, Mitchell AM, et al. Point of care testing provides an accurate measurement of creatinine, anion gap, and osmolal gap in ex-vivo whole blood samples with nitromethane. *Clin Toxicol (Phila)* 2014;52:611–7.
334. Wood DM, Andreyev J, Raja K, et al. Factitiously elevated blood chromium. *Clin Toxicol (Phila)* 2010;48:388–9.
335. Chowdry FR, Rodman H, Bleicher SJ. Glycerol-like contamination of commercial blood sampling tubes. *J Lipid Res* 1971;12:116.
336. Bowen RAR, Chan Y, Cohen J, et al. Effect of blood collection tubes on total triiodothyronine and other laboratory assay. *Clin Chem* 2005;51:424–33.
337. Drake SK, Bowen RAR, Remaley AT, et al. Potential interferences from blood collection tubes in mass spectrometric analyses of serum polypeptides. *Clin Chem* 2004;50:2398–401.
338. Sampson M, Ruddel M, Albright S, et al. Positive interference in lithium determinations from clot activator in collection container. *Clin Chem* 1997;43:675–9.
339. Wang C, Shiraishi S, Leung A, et al. Validation of a testosterone and dihydrotestosterone liquid chromatography tandem mass spectrometry assay: interference and comparison with established methods. *Steroids* 2008;73:1345–52.
340. La'ulu SL, Straseski JA, Schmidt RL, Genzen JR. Thrombin-mediated degradation of parathyroid hormone in serum tubes. *Clin Chim Acta* 2014;437:191–6. doi:10.1016/j.cca.2014.07.030.
341. Shi RZ, van Rossum HH, Bowen RA. Serum testosterone quantitation by liquid chromatography-tandem mass spectrometry: interference from blood collection tubes. *Clin Biochem* 2012;45:1706–9.
342. Schouwers S, Brandt I, Willemse J, et al. Influence of separator gel in Sarstedt S-Monovette serum tubes on various therapeutic drugs, hormones, and proteins. *Clin Chim Acta* 2012;413:100–4.
343. Bowen RA, Chan Y, Ruddel ME, et al. Immunoassay interference by a commonly used blood collection tube additive, the organosilicone surfactant silwet L-720. *Clin Chem* 2005;51:1874–82.
344. Bush VJ, Janu MR, Bathur F, et al. Comparison of BD Vacutainer SST Plus Tubes with BD SST II Plus Tubes for common analytes. *Clin Chim Acta* 2001;306:139–43.
345. Plebani M, Banfi G, Bernardini S, et al. Serum or plasma? An old question looking for new answers. *Clin Chem Lab Med* 2020;58(2):178–87. doi:10.1515/cclm-2019-0719.
346. Wu AHB, Apple FS, Gibler WB, et al. National Academy of Clinical Biochemistry Standards of Laboratory Practice: recommendations for the use of cardiac markers in coronary artery disease. *Clin Chem* 1999;45:1104–21.
347. Stiegler H, Fischer Y, Vazquez-Jimenez JF, et al. Lower cardiac troponin T and I results in heparin-plasma than in serum. *Clin Chem* 2000;46:1338–44.
348. Gerhardt W, Nordin G, Herbert AK, et al. Troponin T and I assays show decreased concentrations in heparin plasma compared with serum: lower recoveries in early than in late phase of myocardial injury. *Clin Chem* 2000;46:817–21.

349. Hermsen D, Hermsen D, Apple F, et al. Results from a multicenter evaluation of the 4th generation Elecsys TnT assay. *Clin Lab* 2007;53:1–9.
350. Tate JR. Troponin revisited 2008: assay performance. *Clin Chem Lab Med* 2008;46:1489–500.
351. English E, McFarlane I, Taylor KP, et al. The effect of potassium EDTA on the stability of parathyroid hormone in whole blood. *Ann Clin Biochem* 2007;44:297–9.
352. Evans MJ, Livesey JH, Ellis MJ, et al. Effect of anticoagulants and storage temperatures on stability of plasma and serum hormones. *Clin Biochem* 2001;34:107–12.
353. Glendenning P, Musk AA, Taranto M, et al. Preanalytical factors in the measurement of intact parathyroid hormone with the DPC IMMULITE assay. *Clin Chem* 2002;48:566–7.
354. CLSI GP41-Ed7 Collection of diagnostic venous blood specimens. 7th ed. Wayne, USA: CLSI; 2017.
355. Kocjancic M, Cargonja J, Delic-Knezevic A. Evaluation of the BD Vacutainer RST blood collection tube for routine chemistry analytes: clinical significance of differences and stability study. *Biochem Med* 2014;24:368–75.
356. Ng WY, Yeo CP. Thrombin-accelerated quick clotting serum tubes: an evaluation with 22 common biochemical analytes. *Adv Hematol* 2013;2013:769479.
357. Dimeski G. Evidence on the cause of false positive troponin I results with the Beckman AccuTnI method. *Clin Chem Lab Med* 2011;49:1079–80.
358. Dimeski G, Coogan M, Jones B, et al. Is the new Beckman AccuTnI+3 assay capable of producing false-positive troponin I results? *Clin Chem Lab Med* 2015;53:e101–3.
359. Dimeski G, Masci PP, Trabi M, et al. Evaluation of the Becton-Dickinson rapid serum tube: does it provide a suitable alternative to lithium heparin plasma tubes? *Clin Chem Lab Med* 2010;48:651–7.
360. Nosanchuk JS, Combs B, Abbott G. False increases of troponin I attributable to incomplete separation of serum. *Clin Chem* 1999;45:714.
361. Er TK, Tsai LY, Jong YJ, et al. Falsely elevated troponin I attributed to inadequate centrifugation using the Access immunoassay analyzer. *Clin Chem Lab Med* 2006;44:908–9.
362. Pretorius CJ, Dimeski G, O'Rourke PK, et al. Outliers as a cause of false cardiac troponin results: investigating the robustness of 4 contemporary assays. *Clin Chem* 2011;57:710–8.
363. Hejl CG, Astier HT, Ramirez JM. Prevention of preanalytical false-positive increases of cardiac troponin I on the Unicel DxI 800 analyzer. *Clin Chem Lab Med* 2008;46:1789–90.
364. Pfäfflin A. Doubt on prevention of false-positive results of cardiac troponin I by recentrifugation. *Clin Chem Lab Med* 2009;47:892–3.
365. Devine PL. High dose hook effect and sample carryover in carcinoembryonic antigen assay performed on the Boehringer-Mannheim ES-300 automated immunoassay system. *Eur J Clin Chem Clin Biochem* 1996;34:573–4.
366. Lippi G, Avanzini P, Musa R, et al. Carryover does not affect results of Beckman Coulter highly sensitive-AccuTnI assay on Access 2. *Clin Chem Lab Med* 2013;51:e141–3.
367. Gould MJ, Wilgen U, Pretorius CJ, et al. Probing indiscretions: contamination of cardiac troponin reagent by very high patient samples causes false-positive results. *Ann Clin Biochem* 2012;49:395–8.
368. Dimeski G, Jones B, Brown N. Carryover can be a cause of false-positive results with the Beckman AccuTnI assay. *Clin Chem Lab Med* 2012;50:1135–6.
369. Haeckel R. IUPAC Proposals for the description and measurement of carry-over effects in clinical chemistry. *Pure Appl Chem* 1991;63:301–6.
370. Clinical Laboratory Standards Institute (CLSI). EP 10-A3 Preliminary evaluation of quantitative clinical laboratory measurement procedures. Approved guideline, 3rd ed. Wayne, USA: CLSI; 2014.
371. Dimeski G, Johnston J, Bassett K, et al. Cuvette carryover with the gentamicin assay on the Beckman AU480 analyser. *Clin Chem Lab Med* 2015;53:e293–5.
372. Lima-Oliveira G, Salvagno GL, Danese E, et al. Contamination of lithium heparin blood by K2-ethylenediaminetetraacetic acid (EDTA): an experimental evaluation. *Biochem Med* 2014;24:359–67.
373. Kouri T, Fogazzi G, Gant V, et al. European urinalysis guidelines. *Scand J Clin Lab Invest* 2000;60:1–96.
374. Deitrick JE, Whedon GD, Shorr E. Effects of immobilization upon various metabolic and physiological functions in normal men. *Am J Med* 1948;4:3–36.
375. Delanghe J, Speeckaert M. Preanalytical requirements of urinalysis. *Biochem Med (Zagreb)* 2014;24:89–104.
376. Garza D. Urine collection and preservation. In: Ross DL, Neely AE, editors. *Textbook of urinalysis and body fluids*. Norwalk, CT: Appleton-Century-Crofts; 1983. p. 57–66.
377. Vaillancourt S, McGillivray D, Zhang X, et al. To clean or not to clean: effect on contamination rates in midstream urine collections in toilet-trained children. *Pediatrics* 2007;119: e1288–93.
378. Cotten SW, Duncan DL, Burch EA, et al. Unexpected interference of baby wash products with a cannabinoid (THC) immunoassay. *Clin Biochem* 2012;45:605–9.
379. CLSI GP16-A3 Urinalysis: Approved guideline. 3rd ed. GP16A3. Wayne, PA: Clinical and Laboratory Standards Institute; 2009.
380. Miller WG, Bruns DE, Hortin GL, et al. Current issues in measurement and reporting of urinary albumin excretion. *Clin Chem* 2009;55:24–38.
381. Robinson MK, Caudill SP, Koch DD, et al. Albumin adsorption onto surfaces of urine collection and analysis containers. *Clin Chim Acta* 2014;431:40–5.
382. Hara F, Shiba K. Nonspecific binding of urinary albumin on preservation tube. *Jpn J Clin Chem* 2003;32:28–9.
383. Silvester S, Zang F. Overcoming non-specific adsorption issues for AZD9164 in human urine samples: consideration of bioanalytical and metabolite identification procedures. *J Chromatogr B Anal Technol Biomed Life Sci* 2012;893–4: 134–43.
384. Manoni F, Valverde S, Caleffi A, et al. Stability of common analytes and urine particles stored at room temperature before automated analysis. *RIMeL/Ital J Lab Med* 2008;4:192–8.
385. Eisinger SW, Schwartz M, Dam L, et al. Evaluation of the BD Vacutainer Plus Urine C&S preservative tubes compared with nonpreservative urine samples stored at 4°C and room temperature. *Am J Clin Pathol* 2013;140:306–13.
386. Holsinger FC, Bui DT. Anatomy, function, and evaluation of the salivary glands. In: Myers EN, Ferris RL, editors. *Salivary gland disorders*. Berlin Heidelberg: Springer; 2007.
387. Nunes LA, Mussavira S, Bindhu OS. Clinical and diagnostic utility of saliva as a non-invasive diagnostic fluid: a systematic review. *Biochem Med (Zagreb)* 2015;25:177–92.
388. Ellervik C, Vaught J. Preanalytical variables affecting the integrity of human biospecimens in biobanking. *Clin Chem* 2015;61:914–34.

389. Chiappin S, Antonelli G, Gatti R, et al. Saliva specimen: a new laboratory tool for diagnostic and basic investigation. *Clin Chim Acta* 2007;383:30–40.
390. Ameringer S, Munro C, Elswick Jr RK. Assessing agreement between salivary alpha amylase levels collected by passive drool and eluted filter paper in adolescents with cancer. *Oncol Nurs Forum* 2012;39:E317–23.
391. Mohamed R, Campbell JL, Cooper-White J, et al. The impact of saliva collection and processing methods on CRP, IgE, and myoglobin immunoassays. *Clin Transl Med* 2012; 1:19.
392. Granger DA, Kivilighan KT, Fortunato C, et al. Integration of salivary biomarkers into developmental and behaviourally oriented research: problems and solutions for collecting specimens. *Physiol Behav* 2007;92:583–90.
393. Munro CL, Grap MJ, Jablonski R, et al. Oral health measurement in nursing research: state of the science. *Biol Res Nurs* 2006;8:35–42.
394. Nurkka A, Obiero J, Kaythy H, et al. Effects of sample collection and storage methods on antipneumococcal immunoglobulin A in saliva. *Clin Diagn Lab Immunol* 2003;10:357–61.
395. Baird G. Preanalytical considerations in blood gas analysis. *Biochem Med* 2013;23:19–27.
396. CLSI C31-A2—Ionized calcium determinations: Precollection variables, specimen choice, collection, and handling; approved guideline—second edition. Wayne, PA: Clinical and Laboratory Standards Institute; 2001.
397. CLSI C46-A2—Blood gas and pH analysis and related measurements; approved guideline. Wayne, PA: Clinical and Laboratory Standards Institute; 2010.
398. CLSI H11-A4—Procedures for the collection of arterial blood specimens. Wayne, PA: Clinical and Laboratory Standards Institute; 2004.
399. Weinreich UM, Thomsen LP, Hansen A, et al. Time to steady state after changes in FIO(2) in patients with COPD. *COPD* 2013;10:405–10.
400. Dong SH, Liu HM, Song GW, et al. Arterialized capillary blood gases and acid-base studies in normal individuals from 29 days to 24 years of age. *Am J Dis Child* 1985;139: 1019–22.
401. Lim K, Wheeler KI, Gale TJ, et al. Oxygen saturation targeting in preterm infants receiving continuous positive airway pressure. *J Pediatr* 2014;164:730–6.
402. Roberts MM, Cho JG, Sandoz JS, et al. Oxygen desaturation and adverse events during 6-min walk testing in patients with COPD. *Respirology* 2015;20:419–25.
403. Groenendaal F, De Vooght KM, van Bel F. Blood gas values during hypothermia in asphyxiated term neonates. *Pediatrics* 2009;123:170–2.
404. Shapiro BA. Temperature correction of blood gas values. *Respir Care Clin N Am* 1995;1:69–76.
405. Davis MD, Walsh BK, Sittig SE, et al. AARC clinical practice guideline: blood gas analysis and hemoximetry: 2013. *Respir Care* 2013;58:1694–703.
406. Zavorsky GS, Cao J, Mayo NE, et al. Arterial versus capillary blood gases: a meta-analysis. *Respir Physiol Neurobiol* 2007;155:268–79.
407. Higgins C. Capillary-blood gases: to arterialize or not. *MLO Med Lab Obs* 2008;40:42–7.
408. Yildizdas D, Yapicioglu H, Yilmaz HL, et al. Correlation of simultaneously obtained capillary, venous, and arterial blood gases of patients in a paediatric intensive care unit. *Arch Dis Child* 2004;89:176–80.
409. Vaquer S, Masip J, Gili G, et al. Earlobe arterialized capillary blood gas analysis in the intensive care unit: a pilot study. *Ann Intensive Care* 2014;4:11.
410. Young J. Arterialised earlobe capillary blood gases in the COPD population. *Br J Nurs* 2014;23:838–42.
411. Vaquer S, Masip J, Gili G, et al. Operational evaluation of the earlobe arterialized blood collector in critically ill patients. *Extrem Physiol Med* 2015;4:5.
412. Trulock EP. Chapter 49. Arterial Blood Gases. In: Walker HK, Hall WD, Hurst JW, editors. *Clinical methods: The history, physical, and laboratory examinations*. 3rd ed. Boston: Butterworths; 1990.
413. Küme T, Şışman AR, Solak A, et al. The effects of different syringe volume, needle size and sample volume on blood gas analysis in syringes washed with heparin. *Biochem Med* 2012;22:189–201.
414. Bender JJ, Allison JR, Goehring JJ, et al. Arterial sampler filling time during arterial and venous punctures, and its relationship with mean arterial pressure in human subjects. *Respir Care* 2012;57:1945–8.
415. Biswas CK, Ramos JM, Agroyannis B, et al. Blood gas analysis: effect of air bubbles in syringe and delay in estimation. *Br Med J (Clin Res Ed)* 1982;284:923–7.
416. Lu JY, Kao JT, Chien TI, et al. Effects of air bubbles and tube transportation on blood oxygen tension in arterial blood gas analysis. *J Formos Med Assoc* 2003;102:246–9.
417. Victor Peter J, Patole S, Fleming JJ, et al. Agreement between paired blood gas values in samples transported either by a pneumatic system or by human courier. *Clin Chem Lab Med* 2011;49:1303–9.
418. Astles JR, Lubarsky D, Loun B, et al. Pneumatic transport exacerbates interference of room air contamination in blood gas samples. *Arch Pathol Lab Med* 1996;120:642–7.
419. Salvagno GL, Lippi G, Gelati M, et al. Hemolysis, lipaemia and icterus in specimens for arterial blood gas analysis. *Clin Biochem* 2012;45:372–3.
420. Lippi G, Ippolito L, Fontana R. Prevalence of hemolytic specimens referred for arterial blood gas analysis. *Clin Chem Lab Med* 2011;49:931–2.
421. Lippi G, Fontana R, Avanzini P, et al. Influence of spurious hemolysis on blood gas analysis. *Clin Chem Lab Med* 2013; 51:1651–4.
422. Grenache DG, Parker C. Integrated and automatic mixing of whole blood: an evaluation of a novel blood gas analyzer. *Clin Chim Acta* 2007;375:153–7.
423. Knowles TP, Mullin RA, Hunter JA, et al. Effects of syringe material, sample storage time, and temperature on blood gases and oxygen saturation in arterialized human blood samples. *Respir Care* 2006;51:732–6.
424. Bonar R, Favaloro EJ, Adcock DM. Quality in coagulation and haemostasis testing. *Biochem Med* 2010;20:184–99.
425. Lippi G, Salvagno GL, Montagnana M, Lima-Oliveira G, Guidi GC, Favaloro EJ. Quality standards for sample collection in coagulation testing. *Semin Thromb Hemost* 2012;38:565–75.
426. Adcock Funk D, Lippi G, Favaloro EJ. Quality standards for sample processing, transportation, and storage in hemostasis testing. *Semin Thromb Hemost* 2012;38:576–85.
427. Favaloro EJ, Adcock Funk DM, Lippi G. Pre-analytical variables in coagulation testing associated with diagnostic errors in hemostasis. *LabMed* 2012;43:1–10.

428. Koçak FE, Yöntem M, Yücel O, et al. The effects of transport by pneumatic tube system on blood cell count, erythrocyte sedimentation and coagulation tests. *Biochem Med* 2013;23: 206–10.
429. Lima-Oliveira G, Salvagno GL, Lippi G, et al. Could light meal jeopardize laboratory coagulation tests? *Biochem Med* 2014;24:343–9.
430. Magnette A, Chatelain M, Chatelain B, Ten Cate H, Mullier F. Pre-analytical issues in the haemostasis laboratory: guidance for the clinical laboratories. *Thromb J* 2016;14:49.
431. Gosselin RC, Marlar RA. Preanalytical variables in coagulation testing: setting the stage for accurate results. *Semin Thromb Hemost* 2019;45:433–48.
432. Lippi G, Salvagno GL, Montagnana M, et al. Short-term venous stasis influences routine coagulation testing. *Blood Coagul Fibrinolysis* 2005;16:453–8.
433. Kratz A, Stanganelli N, Van Cott EM. A comparison of glass and plastic blood collection tubes for routine and specialized coagulation assays. *Arch Pathol Lab Med* 2006;130:39–44.
434. Toulon P, Aillaud MF, Arnoux D, Boissier E, Borg JY, Gourmel C. Multicenter evaluation of a bilayer polymer blood collection tube for coagulation testing: effect on routine hemostasis test results and on plasma levels of coagulation activation markers. *Blood Coagul Fibrinolysis* 2006;17:625–31.
435. Clinical and Laboratory Standards Institute (CLSI). Collection, transport, and processing of blood specimens for testing plasma-based coagulation assays and molecular hemostasis assays; approved guideline – Fifth Edition. CLSI document H21-A5. Clinical and Laboratory Standards Institute, Wayne, Pennsylvania, USA, 2008.
436. Bronić A, Coen Herak D, Margetić S, Milić M. Croatian Society of Medical Biochemistry and Laboratory Medicine: National recommendations for blood collection, processing, performance and reporting of results for coagulation screening assays prothrombin time, activated partial thromboplastin time, thrombin time, fibrinogen and D-dimer. *Biochem Med* 2019;29:020503.
437. Adcock DM, Kressin DC, Marlar RA. Effect of 3.2% vs 3.8% sodium citrate concentration on routine coagulation testing. *Am J Clin Pathol* 1997;107:105–10.
438. Ratzinger F, Lang M, Schmetterer KG, Haslacher H, Perkmann T, Quehenberger P. The effect of 3.2% and 3.8% sodium citrate on specialized coagulation tests. *Arch Pathol Lab Med* 2018;142:992–7.
439. Loreth RM, Klose G. Comparison of two different blood sample tubes for platelet function analysis with the multiplate system. *Transfus Med Hemother* 2010;37:289–92.
440. Hellstern P, Sturzebecher U, Wuchold B, et al. Preservation of in vitro function of platelets stored in the presence of a synthetic dual inhibitor of factor Xa and thrombin. *J Thromb Haemost* 2007;5:2119–26.
441. Favaloro EJ, Lippi G. The new oral anticoagulants and the future of haemostasis laboratory testing. *Biochem Med* 2012; 22:329–41.
442. Adcock DM, Kressin DC, Marlar RA. Minimum specimen volume requirements for routine coagulation testing: dependence on citrate concentration. *Am J Clin Pathol* 1998;109:595–99.
443. Ver Elst K, Vermeiren S, Schouwers S, Callebaut V, Thomson W, Weekx S. Validation of the minimal citrate tube fill volume for routine coagulation tests on ACL TOP 500 CTS®. *Int J Lab Hematol* 2013;35:614–9.
444. Haas T, Spielmann N, Cushing M. Impact of incorrect filling of citrate blood sampling tubes on thromboelastometry. *Scand J Clin Lab Invest* 2015;75:717–9.
445. Lippi G, Salvagno GL, Radišić Biljak V, et al. Filling accuracy and imprecision of commercial evacuated sodium citrate coagulation tubes. *Scand J Clin Lab Invest* 2019;79: 276–9.
446. Harrison P, Mackie I, Mumford A, et al. Guidelines for the laboratory investigation of heritable disorders of platelet function. *Br J Haematol* 2011;155:30–44.
447. Rajmakers MT, Menting CH, Vader HL, et al. Collection of blood specimens by venipuncture for plasma-based coagulation assays: necessity of a discard tube. *Am J Clin Pathol* 2010;133:331–5.
448. Smock KJ, Crist RA, Hansen SJ, Rodgers GM, Lehman CM. Discard tubes are not necessary when drawing samples for specialized coagulation testing. *Blood Coagul Fibrinolysis* 2010;21:279–82.
449. Favaloro EJ, Lippi G. Discard tubes are sometimes necessary when drawing samples for hemostasis. *Am J Clin Pathol* 2010;134:849–53.
450. Lippi G, Funk DMA, Favaloro EJ. Discard tube for coagulation testing: the debate continues. *Blood Coagul Fibrinolysis* 2012; 23:572–3.
451. Masih M, Kakkar N. Routine coagulation testing: do we need a discard tube? *Indian J Hematol Blood Transfus* 2014;30:347–50.
452. Simundic AM, Bölenius K, Cadamuro J, et al. Joint EFLM-COLABIOCLI recommendation for venous blood sampling. *Clin Chem Lab Med* 2018;56:2015–38.
453. Wallin O, Söderberg J, Grankvist K, et al. Preanalytical effects of pneumatic tube transport on routine haematology, coagulation parameters, platelet function and global coagulation. *Clin Chem Lab Med* 2008;46:1443–9.
454. Hübner U, Böckel-Frohnhofer N, Hummel B, et al. The effect of a pneumatic tube transport system on platelet aggregation using optical aggregometry and the PFA-100. *Clin Lab* 2010;56:59–64.
455. Lippi G, Salvagno GL, Montagnana M, Manzato F, Guidi GC. Influence of the centrifuge time on primary plasma tubes on routine coagulation testing. *Blood Coagul Fibrinolysis* 2007; 18:525–8.
456. Suchsland J, Friedrich N, Grotevendt A, et al. Optimizing centrifugation of coagulation samples in laboratory automation. *Clin Chem Lab Med* 2014;52:1187–91.
457. Boissier E, Sevin-Allouet M, Le Thuaut A, et al. A 2-min at 4500 g rather than a 15-min at 2200 g centrifugation does not impact the reliability of 10 critical coagulation assays. *Clin Chem Lab Med* 2017;55:e118–21.
458. Daves M, Giacomuzzi K, Tagnin E, et al. Influence of centrifuge brake on residual platelet count and routine test in citrated plasma. *Blood Coagul Fibrinolysis* 2014;25:292–5.
459. Lippi G, Salvagno GL, Montagnana M, et al. Influence of the centrifuge time of primary plasma tubes on routine coagulation testing. *Blood Coagul Fibrinolysis* 2007;18: 525–8.
460. Femia EA, Pugliano M, Podda G, et al. Comparison of different procedures to prepare platelet-rich plasma for studies of platelet aggregation by light transmission aggregometry. *Platelets* 2012;23:7–10.
461. Zürcher M, Sulzer I, Barizzi G, et al. Stability of coagulation assays performed in plasma from citrated whole blood

- transported at ambient temperature. *Thromb Haemost* 2008;99:416–26.
462. Lippi G, Franchini M, Montagnana M, et al. Quality and reliability of routine coagulation testing: can we trust that sample? *Blood Coagul Fibrinolysis* 2006;17:513–9.
463. Clinical and Laboratory Standards Institute (CLSI). Quantitative D-dimer for the Exclusion of Venous Thromboembolic Disease; Approved Guideline. CLSI document H59-A. WaYNE, pa: Clinical and Laboratory Standards Institute;2011.
464. Guder WG, Fiedler GM, da Fonseca-Wollheim F, et al. Quality of diagnostic samples. 4th ed. Oxford: BD Diagnostics Preanalytical Systems; 2015.
465. Zhao Y, Feng G, Zhang J, Gong R, Cai C, Feng L. Effects of preanalytical frozen storage time and temperature on screening coagulation tests and factors VIII and IX activity. *Sci Rep* 2017;7:12179.
466. Rimac V, Coen Herak D. Is it acceptable to use coagulation plasma samples stored at room temperature and 4°C for 24 hours for additional prothrombin time, activated partial thromboplastin time, fibrinogen, antithrombin and D-dimer testing? *Int J Lab Hematol* 2017;39:475–81.
467. van Geest-Daalderop JH, Mulder AB, Boonman-de Winter LJ, Hoekstra MM, van der Besselaar AM. Preanalytical variables and off-site blood collection: influences on the results of the prothrombin time/international normalized ratio test and implications for monitoring of oral anticoagulant therapy. *Clin Chem* 2005;51(3):561–8.
468. Feng L, Zhao Y, Zhao H, Shao Z. Effects of storage time and temperature on coagulation tests and factors in fresh plasma. *Sci Rep* 2014;4:3868.
469. Alesci S, Borggrefe M, Dempfle CE. Effect of freezing method and storage at -20°C and -70°C on prothrombin time, aPTT and plasma fibrinogen levels. *Thromb Res* 2009;124:121–6.
470. Foshat M, Bates S, Russo W, et al. Effect of freezing plasma at -20°C for 2 weeks on prothrombin time, activated partial thromboplastin time, dilute Russell viper venom time, activated protein C resistance, and D-dimer levels. *Clin Appl Thromb Hemost* 2015;21:41–7.
471. Gosselin RC, Dwyre DW. Determining the effect of freezing on coagulation testing: comparison of results between fresh and once frozen-thawed plasma. *Blood Coagul Fibrinolysis* 2015;26:69–74.
472. Horton S, Fleming KA, Kuti M, et al. The top 25 laboratory tests by volume and revenue in five different countries. *Am J Clin Pathol* 2019;151:446–51.
473. Lippi G, Bassi A, Solero GP, et al. Prevalence and type of preanalytical errors on inpatient samples referred for complete blood count. *Clin Lab* 2007;53:555–6.
474. Upreti S, Upreti S, Bansal R, et al. Types and frequency of preanalytical errors in hematology lab. *J Clin Diagn Res* 2013;7:2191–3.
475. Narang V, Kaur H, Kaur Selhi P, et al. Preanalytical errors in hematology laboratory – an avoidable incompetence. *Iran J Pathol* 2016;11:151–4.
476. Arul P, Pushparaj M, Pandian K, et al. Prevalence and types of preanalytical error in hematology laboratory of a tertiary care hospital in South India. *J Lab Physicians* 2018;10:237–40.
477. Narula A, Yadav SK, Jahan A, et al. Pre-analytical error in a hematology laboratory: an avoidable cause of compromised quality in reporting. *Clin Chem Lab Med* 2019;57:e262–4.
478. De la Salle B. Pre- and postanalytical error in haematology. *Int J Lab Hematol* 2019;41(Suppl 1):170–6.
479. Banfi G, Salvagno GL, Lippi G. The role of ethylenediamine tetraacetic acid (EDTA) as in vitro anticoagulant for diagnostic purposes. *Clin Chem Lab Med* 2007;45:565–76.
480. Goosens W, van Duppen V, Verwilghen RL. K2- or K3-EDTA: the anticoagulant of choice in routine haematology? *Clin Lab Haematol* 1991;13:291–5.
481. International Council for Standardization in haematology (ICSH). Recommendations of the ICSH for ethylene diamine tetraacetic acid anticoagulation of blood for blood cell counting and sizing: expert panel on cytometry. *Am J Clin Pathol* 1993;100:371–2.
482. Buttarello M. Quality specifications in haematology: the automated blood cell count. *Clin Chim Acta* 2004;346:45–54.
483. Van Cott EM, Lewandrowski KB, Patel S, et al. Comparison of glass K3EDTA versus plastic K2EDTA blood-drawing tubes for complete blood counts, reticulocyte counts, and white blood cell differentials. *Lab Hematol* 2003;9:10–4.
484. Leathem S, Zantek ND, Kemper M, Korte L, Langeberg A, Sandler SG. Equivalence of spray-dried K2EDTA, spray-dried K3EDTA, and liquid K3EDTA anticoagulated blood samples for routine blood center or transfusion service testing. *ImmunoHematology* 2003;19:117–21.
485. Lima-Oliveira G, Lippi G, Salvagno GL, et al. K3EDTA vacuum tubes validation for routine hematological testing. *ISRN Hematology* 2012;Article ID 875357, 5pages.
486. Available from: [https://www.gbo.com/fileadmin/user\\_upload/Downloads/White\\_Papers/WP\\_Evaluation\\_K2\\_K3\\_comparison\\_hematology\\_WE01\\_Rev00.pdf](https://www.gbo.com/fileadmin/user_upload/Downloads/White_Papers/WP_Evaluation_K2_K3_comparison_hematology_WE01_Rev00.pdf). Accessed January 11, 2020.
487. Lima-Oliveira G, Lippi G, Salvagno GL, et al. Brand of dipotassium EDTA vacuum tubes as a new source of preanalyticl variability in routine hematology testing. *Brit J Biomed Sci* 2013;70:6–9.
488. Clinical and Laboratory Standards Institute (CLSI). Tubes and additives for venous and capillary blood specimen collection; approved standard—Sixth Edition. CLSI document GP39-A6 (ISBN 1-56238-740-5). Clinical and Laboratory Standards Institute, 950 West Valley Road, Suite 2500, Wayne, Pennsylvania 19087 USA, 2010.
489. Lampasso JA. Error in hematocrit value produced by excessive ethylenediamine-tetraacetic acid. *Am J Clin Pathol* 1965;44:109–110.
490. Lewis SM, Stoddart CTH. Effects of anticoagulants and containers (glass and plastic) on the blood count. *Lab Practice* 1977;20:787–92.
491. Sacker LS. Specimen collection. In: Lewis SM, Coster JF, editors. *Quality Control in Haematology*. New York: Academic Press; 1975;224–7.
492. Xu M, Robbe VA, Jack RM, et al. Under-filled blood collection tubes containig K<sub>2</sub>EDTA as anticoagulant are acceptable for automated complete blood counts, white blood cell differential, and reticulocyte count. *Int J Lab Hematol* 2010;32:491–7.
493. Radišić Biljak V, Božić ević S, Krhac' M, et al. Impact of under-filled blood collection tubes containing K2EDTA and K3EDTA as anticoagulants on automated complete blood count (CBC) testing. *Clin Chem Lab Med* 2016;54:e323–6.
494. Pewarchuk W, Vanderbroom J, Blajchman MA. Pseudopolycythemia, pseudothrombocytopenia, and pseudoleukopenia due to overfilling of blood collection vacuum tubes. *Arch Pathol Lab Med* 1992;116:90–2.

495. Available from: <http://www.clinlabnavigator.com/guidelines-for-detecting-iv-contamination-of-blood-samples.html>. Accessed January 28, 2020.
496. CAP Today. Available from [http://www.captodayonline.com/Archives/q\\_and\\_a/1206qa.html](http://www.captodayonline.com/Archives/q_and_a/1206qa.html). Accessed January 28, 2020.
497. Ye Y, Wang W, Zhao H, et al. Haematology specimen acceptability: a national survey in Chinese laboratories. *Biochem Med* 2018;28:420–9.
498. Lippi G, Salvagno GL, Montagnana M, et al. Evaluation of different mixing procedures for K2EDTA primary samples on hematological testing. *Labmedicine* 2007;38:723–5.
499. Simundic AM, Bölenius K, Cadamuro J. Joint EFLM-COLABIOCLI recommendation for venous blood sampling. *Clin Chem Lab Med* 2018;56:2015–38.
500. Lima-Oliveira G, Lippi G, Salvagno GL, et al. Effects of vigorous mixing of blood vacuum tubes on laboratory test results. *Clin Biochem* 2013;46:250–4.
501. Tatsumi N, Miwa S, Lewis SM. International Society of hematology, and the International Council for Standardization in Haematology. Specimen collection, storage, and transmission to the laboratory for hematological tests. *Int J Hematol* 2002; 75:261–8.
502. Zini G. International Council for Standardization in Haematology (ICSH): stability of complete blood count parameters with storage: toward defined specifications for different diagnostic applications. *Int J Lab Hematol* 2014;36:111–3.
503. Imeri F, Herklotz R, Risch L, et al. Stability of hematological analytes depends on the hematology analyser used: a stability study with Bayer Advia 120, Beckman Coulter LH 750 and Sysmex XE 2100. *Clin Chim Acta* 2008;397:68–71.
504. Daves M, Zagler EM, Cemin R, et al. Sample stability for complete blood cell count using the Sysmex XN haematological analyser. *Blood Transfus* 2015;13:576–82.
505. Pintér E, László K, Schüszler I, et al. The stability of quantitative blood count parameters using the ADVIA 2120i hematology analyzer. *Pract Lab Med* 2016;4:16–21.
506. Lippi G, Salvagno GL, Solero GP, et al. Stability of blood cell counts, hematologic parameters and reticulocytes indexes on the Advia A120 hematologic analyzer. *J Lab Clin Med* 2005;146:333–40.
507. Vives-Corrons JL, Briggs C, Simon-Lopez R, et al. Effect of EDTA-anticoagulated whole blood storage on cell morphology examination. A need for standardization. *Int J Lab Hematol* 2014;36:222–6.
508. Kopcinovic LM, Pavic M. Platelet satellitism in a trauma patient. *Biochem Med* 2012;22:130–4.
509. Vidranski V, Laskaj R, Sikiric D, et al. Platelet satellitism in infectious disease? *Biochem Med* 2015;25:285–94.
510. Vicari A, Banfi G, Bonini PA. EDTA-dependent pseudothrombocytopenia: a 12-month epidemiological study. *Scand J Clin Lab Invest* 1988;48:537–42.
511. Bizzaro N. EDTA-dependent pseudothrombocytopenia: a clinical and epidemiological study of 112 cases, with 10-year follow up. *Am J Hematol* 1995;50:103–9.
512. Lippi G, Plebani M. EDTA-dependent pseudothrombocytopenia: further insights and recommendations for prevention of a clinically threatening artifact. *Clin Chem Lab Med* 2012;50:1281–5.
513. Gowland E, Kay HEM, Spillman JC, et al. Agglutination of platelets by a serum factor in the presence of EDTA. *J Clin Path* 1969;22:460–4.
514. Mant MJ, Doery JCG, Gauldie J, et al. Pseudothrombocytopenia due to platelet aggregation and degranulation in blood collected in EDTA. *Scand J Haematol* 1975;15:161–70.
515. Patel KJ, Hughes CG, Parapia LA. Pseudoleucocytosis and pseudothrombocytosis due to cryoglobulinemia. *J Clin Pathol* 1987;40:120–1.
516. Ohno N, Kobayashi M, Hayakawa S, et al. Transient pseudothrombocytopenia in a neonate: transmission of a maternal EDTA-dependent anticoagulant. *Platelets* 2012;23:399–400.
517. Korterink JJ, Boersma B, Schoolr M, et al. Pseudothrombocytopenia in a neonate due to mother? *Eur J Pediatr* 2013;172: 987–9.
518. Xiao Y, Xu Y. Concomitant spuriously elevated white blood cell count, a previously underestimated phenomenon in EDTA-dependent pseudothrombocytopenia. *Platelets* 2015; 26:627–31.
519. Bartels PCM, Schoorl M, Lombarts AJPF. Screening for EDTA-dependent deviations in platelet counts and abnormalities in platelet distribution histograms in pseudothrombocytopenia. *Scand J Clin Lab Invest* 1997;57:629–36.
520. Akinci S, Hacibekiroglu T, Guney T, et al. Evaluation of pseudothrombocytopenia causes. *Basic Research Journal of Medicine and Clinical Sciences* 2014;3:24–7.
521. Nagler M, Keller P, Siegrist D, et al. A case of EDTA-dependent pseudothrombocytopenia: simple recognition of an underdiagnosed and misleading phenomenon. *BMC Clin Pathol* 2014;14:19.
522. Kratz A, Lee S, Zini G, et al. Digital morphology analyzers in hematology: ICSH review and recommendations. *Int J Lab Hematol* 2019;41:437–47.
523. Ahn HL, Jo YI, Choi YS, et al. EDTA-dependent pseudothrombocytopenia confirmed by supplementation of kanamycin: a case report. *Korean J Intern Med* 2002;17:65–8.
524. Choccalingam C, Radha RKN, Snigdha N. Estimation of platelet counts and other hematological parameters in pseudothrombocytopenia using alternative anticoagulant: magnesium sulfate. *Clin Med Insights Blood Disord* 2017; 10:1–6.
525. Chae H, Kim M, Lim J, et al. Novel method to dissociate platelet clumps in EDTA-dependent pseudothrombocytopenia based on the pathophysiological mechanism. *Clin Chem Lab Med* 2012;50:1387–91.
526. Berentsen S, Beiske K, Tjønnfjord GE. Primary chronic cold agglutinin disease: an update on pathogenesis, clinical features and therapy. *Hematology* 2007;12:361–70.
527. Berentsen S, Tjønnfjord GE. Diagnosis and treatment of cold agglutinin mediated autoimmune hemolytic anemia. *Blood Rev* 2012;26:107–15.
528. Kumar TB, Bhardwaj N. Platelet cold agglutinins and thrombocytopenia: a diagnostic dilemma in the intensive care unit. *J Anaesthesiol Clin Pharmacol* 2014;30:89–90.
529. Glassy EF. Color Atlas of Hematology. An illustrated field guide based on proficiency testing. Illinois: College of American Pathologists; 2018.
530. McKenzie SB, Landis-Piwowar K, Williams JL. Clinical Laboratory Hematology. 4 th edition. New Jersey: Pearson Prentice Hall; 2020.
531. Ercan S, Caliskan M, Koptur E. 70-year old female patient with mismatch between hematocrit and hemoglobin values: the effects of cold agglutinin on complete blood count. *Biochem Med* 2014;24:391–5.

532. Rim JH, Chang MH, Oh J, et al. Effects of cold agglutinin on the accuracy of complete blood count results and optimal sample pretreatment protocols for eliminating such effects. *Ann Lab Med* 2018;38:371–4.
533. La Gioia A. Eliminating or minimizing the effects of cold agglutinins on the accuracy of complete blood count results. *Ann Lab Med* 2019;39:499–500.
534. Topic A, Milevoj Kopcinovic L, Bronic A, et al. Effect of cold agglutinins on red blood cell parameters in a trauma patient: a case report. *Biochem Med* 2018;28:031001.
535. Kolopp-Sarda MN, Miossec P. Cryoglobulins: an update on detection, mechanisms and clinical contribution. *Autoimmun Rev* 2018;17:457–64.
536. Ramos-Casals M, Stone JH, Cid MC, et al. The cryoglobulinaemias. *Lancet* 2012;379:348–60.
537. Abela M, McArdle B, Qureshi M. Pseudoleucocytosis due to cryoglobulinaemia. *J Clin Pathol* 1980;33:796.
538. Fohlen-Walter A, Jacob C, Lecompte T, et al. Laboratory identification of cryoglobulinemia from automated blood cell counts, fresh blood samples, and blood films. *Am J Clin Pathol* 2002;117:606–14.
539. Keuren JFW, Raijmakers MTM, Oosterhuis WP, et al. Drastic effects of cryoglobulin on blood cell counts, erythrocyte morphology and M-protein analysis. *Scand J Clin Lab Invest* 2010;70:462–4.
540. Gulati G, Song J, Dulau Florea A, et al. Purpose and criteria for blood smear scan, blood smear examination, and blood smear review. *Ann Lab Med* 2013;33:1–7.
541. Miler M, Simundic AM. Low level of adherence to instructions for 24-hour urine collection among hospital outpatients. *Biochem Med* 2013;23:316–20.
542. Kackov S, Simundic AM, Gatti-Drnic A. Are patients well informed about the fasting requirements for laboratory blood testing? *Biochem Med* 2013;23:326–31.
543. Bölenius K, Lindkvist M, Brulin C, et al. Impact of a large-scale educational intervention program on venous blood specimen collection practices. *BMC Health Serv Res* 2013;13:463.
544. Lima-Oliveira G, Lippi G, Salvagno GL, et al. Impact of the phlebotomy training based on CLSI H03-a6—Procedures for the collection of diagnostic blood specimens by venipuncture. *Biochem Med* 2012;22:342–51.
545. Salinas M, López-Garrigós M, Flores E, et al. Education and communication are the key for the successful management of vitamin D test requesting. *Biochem Med* 2015;25(2):237–41.
546. Epner PL, Gans JE, Gruber ML. When diagnostic testing leads to harm: a new outcomes-based approach for laboratory medicine. *BMJ Qual Saf* 2013;22(Suppl 2):ii6–10.
547. Simundic AM, Nikolac N, Miler M, et al. Efficiency of test report delivery to the requesting physician in an outpatient setting: an observational study. *Clin Chem Lab Med* 2009;47:1063–6.
548. Juricic G, Kopcinovic LM, Saracevic A, et al. Liquid citrate acidification introduces significant glucose bias and leads to misclassification of patients with diabetes. *Clin Chem Lab Med* 2016;54(2):363–71. doi:10.1515/cclm-2015-0358.
549. Maranda B, Cousineau J, Allard P, et al. False positives in plasma ammonia measurement and their clinical impact in a pediatric population. *Clin Biochem* 2007;40:531–5.
550. Jacobs P, Costello J, Beckles M. Cost of haemolysis. *Ann Clin Biochem* 2012;49:412.
551. Simundic AM, Topic E. Quality indicators. *Biochem Med* 2008;18:311–9.
552. Sciacovelli L, O’Kane M, Skaik YA, et al. Quality indicators in laboratory medicine: from theory to practice. Preliminary data from the IFCC Working Group Project “Laboratory Errors and Patient Safety.” *Clin Chem Lab Med* 2011;49:835–44.
553. Plebani M, Chiozza ML, Sciacovelli L. Towards harmonization of quality indicators in laboratory medicine. *Clin Chem Lab Med* 2013;51:187–95.
554. Plebani M, Sciacovelli L, Aita A, et al. Harmonization of pre-analytical quality indicators. *Biochem Med* 2014;24:105–13.
555. Plebani M, Astion ML, Barth JH, et al. Harmonization of quality indicators in laboratory medicine. A preliminary consensus. *Clin Chem Lab Med* 2014;52:951–8.

## MULTIPLE CHOICE QUESTIONS

1. Which of the following statements best describes the way through which interference factors may be reduced or eliminated?
  - a. Standardizing the preanalytical conditions
  - b. Providing proper instructions to the patients on how to prepare for blood sampling
  - c. Selecting a more specific method
  - d. Selecting the appropriate sampling procedure
  - e. Maintaining the analytical variability (method CV<sub>A</sub>) to a minimum
2. Spectrophotometric interference of hemolysis occurs due to the ability of hemoglobin to absorb light at which wavelengths?
  - a. 400, 500, and 600 nm
  - b. 550 and 570 nm
  - c. 540 and 600 nm
  - d. 415 and 540 nm
  - e. 415, 540, and 570 nm
3. The recommended sample for the accurate measurement of potassium is:
  - a. Plasma
  - b. Serum
  - c. Whole blood
  - d. Capillary blood
  - e. Arterial blood
4. Which of the following statements is true for lipid testing and testing for lipid-soluble drugs and hormones?
  - a. Testing should always be done in a delipidated sample
  - b. Testing should always be done on the native sample before delipidation
  - c. Delipidation does not affect the concentration of lipids and lipid-soluble drugs
  - d. The most suitable delipidation method is ultracentrifugation
  - e. The most suitable delipidation method is lipid removal using the lipid-clearing agents
5. Which of these sources of interferences is endogenous?
  - a. Prescribed medications
  - b. Supportive medical therapy like parenteral emulsions, contrast media agents, or infusion solutions
  - c. Dietary supplements
  - d. Substances occurring in the blood through accidental exposure and poisoning
  - e. In vivo hemolysis
6. Which of these mechanisms is inherent to paraprotein interference?
  - a. Binding of the paraprotein to the blood tube walls
  - b. Precipitation of the paraprotein and its binding to the assay components
  - c. Paraprotein interference due to an increase in the aqueous phase volume in the sample
  - d. Paraprotein interference due to a change in the sample ionic strength
  - e. Proteolytic cleavage of the binding sites
7. According to the International Council for Standardization in Hematology, the anticoagulant of choice for hematology testing is:
  - a. Dipotassium EDTA
  - b. Tripotassium EDTA
  - c. Disodium EDTA
  - d. 3.8% sodium citrate
  - e. 3.2% sodium citrate
8. In which cases is the discard tube necessary?
  - a. In patients with difficult vein access
  - b. In patients with high blood pressure
  - c. Whenever coagulation tube is collected as the first or the only tube
  - d. When coagulation tube is collected as the first or the only tube and a straight needle is used for blood collection
  - e. When coagulation tube is collected as the first or the only tube and a winged blood collection set (butterfly devices) is used
9. Which of the below stated effects occur as a consequence of the ingestion of cooked meat?
  - a. A decrease of up to 20% of plasma creatinine concentration
  - b. Sample hemolysis
  - c. Metabolic acidosis
  - d. An increase of up to 20% of plasma creatinine concentration
  - e. An increase of AST and ALT
10. Which of the below stated effects occur as a consequence of the change in body posture from the supine to the upright position?
  - a. The decrease in the concentration of molecules with large molecular weight
  - b. The increase in plasma volume
  - c. The increase in the concentration of molecules with large molecular weight
  - d. An increase of up to 30% of plasma sodium concentration
  - e. Sample hemolysis

# Quality Control of the Analytical Examination Process

*W. Greg Miller and Sverre Sandberg<sup>a</sup>*

## ABSTRACT

Quality control (QC), also called internal QC, monitors a measurement procedure to verify that results for patient samples meet performance specifications appropriate for patient care or that an error condition exists that must be corrected. QC samples are measured at intervals along with patient samples. Recovery of the expected target values for the QC samples allows the laboratory to verify that a measurement procedure is working correctly and the results for patient samples can be reported. The QC plan specifies the number of controls, the frequency they are to be measured, and the rules to determine if the QC results are consistent with expected measurement procedure performance. External QC, also called external quality assessment (EQA) or proficiency testing (PT), is an assessment process in which control samples are received from an independent external organization and the expected values are not known by the laboratory. The results for the EQA/PT samples are compared with target values assigned to the samples to verify that a laboratory's measurement procedures conform to expected performance. EQA/PT schemes that use commutable samples assess trueness of patient sample results when a reference measurement procedure is used for target value assignment, or harmonization among results when no reference measurement value is available.

## Background

The purpose of a clinical laboratory test is to provide information on the pathophysiologic condition of an individual

patient to assist with diagnosis, to guide or monitor therapy, or to assess risk for developing a disease or for progression of a disease. QC, also called internal QC, monitors a measurement procedure to verify that it meets performance specifications appropriate for patient care or that an error condition exists that must be corrected.

## Content

Internal QC ensures that measurement procedures meet specifications at the time patient testing occurs. QC samples are measured at intervals along with patient samples. Recovery of the expected target values for the QC samples allows the laboratory to verify that a measurement procedure is working correctly and the results for patient samples can be reported. The QC plan specifies the number of controls, the frequency they are to be measured, and the rules to determine if the QC results are consistent with expected measurement procedure performance. External QC, also called *external quality assessment (EQA)* or *proficiency testing (PT)*, is an assessment process in which control samples are received from an independent external organization and the expected values are not known by the laboratory. The results for the EQA/PT samples are compared with target values assigned to the samples to verify that a laboratory's measurement procedures conform to expected performance. In addition to internal and external QC, the results from patient sample testing (e.g., medians of patient results) can be used to assess and monitor the performance of measurement procedures.

## INTRODUCTION

The purpose of a clinical laboratory test is to provide information on the pathophysiologic condition of an individual patient to assist with diagnosis, to guide or monitor therapy, or to assess risk for developing a disease or for progression of a disease. Quality control (QC) monitors a measurement procedure to verify that it meets performance specifications

appropriate for patient care or that an error condition exists that must be corrected. QC includes both internal and external components.

Internal QC includes control procedures applied within a laboratory to assess the performance of an analytical examination procedure. The most common approach is to substitute surrogate QC samples that are intended to simulate clinical samples from patients. The QC samples are measured at intervals along with patient samples. Recovery of the expected target values for the QC samples allows the laboratory to verify that a measurement procedure is working correctly and the results for patient samples are reliable enough to be

<sup>a</sup>Some of this material was previously published in McPherson RA and Pincus MR, editors. *Henry's clinical diagnosis and management by laboratory methods*. 24th ed. St. Louis: Elsevier; 2020.

reported. Note the term internal QC is distinct from control processes and fluids that are “built-in” to a measurement technology or to the reagent cartridges or strips used by a measurement procedure. The performance of a measurement procedure can, for some measurands, also be monitored using the consistency of results from patient samples as part of the internal or external QC process.

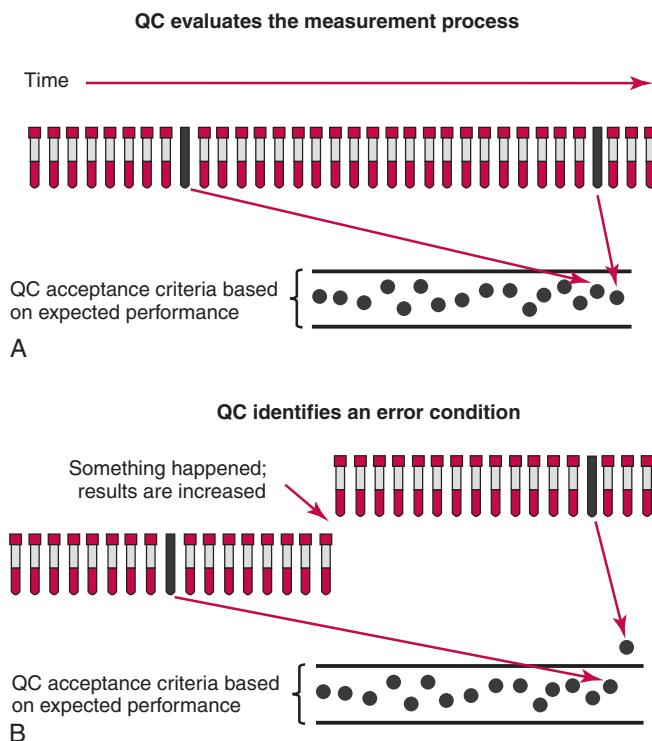
External QC, also called external quality assessment (EQA) or proficiency testing (PT), is a monitoring process in which surrogate samples are received from an independent external organization and the expected values are not known by the laboratory. The results for the EQA/PT samples are sent to the provider and compared with results from other laboratories to examine if a laboratory’s measurement procedures conform to expected performance. When commutable EQA samples are used (see later section on EQA) the performance can be assessed for agreement with a true value assigned by a reference procedure. Some EQA providers also provide follow up of erroneous results with advice and site visits to individual participants. In these organizations, the acronym EQA stands for external quality *assurance*.

As illustrated in Fig. 6.1, internal QC evaluates a measurement procedure by periodically measuring a QC sample for which the expected result is known in advance. If the result

for a QC material is within acceptable limits of the expected value (see Fig. 6.1A), the measurement procedure is verified to be stable, which means that it is performing as expected, and results for patient samples can be reported with high probability that they are suitable for clinical use. If a QC result is not within acceptable limits (see Fig. 6.1B), the measurement procedure is not performing correctly, there is a high probability that results for patient samples are not suitable for clinical use, and corrective action is necessary. Note that QC acceptance criteria may be designed to provide a warning of, for example, calibration drift that can be corrected before the error becomes large enough to adversely affect patient results. If corrective action is indicated, patient sample measurements will need to be repeated when the measurement procedure has been restored to its stable performance condition. If erroneous results have already been reported before an error condition is identified, a corrected report must be issued.

Measurement procedures fall into one of two general categories from a QC plan perspective. One type of procedure is a “batch” measurement process in which the results for patient samples and QC samples are completed before the results are reported. For batch measurement procedures, results are not reported if an error condition is identified by the QC sample measurements. The other type of procedure, which is becoming more common, is a “continuous” measurement process in which patient sample results are reported during the interval between QC sample measurements and continue to be reported after a QC measurement event with no intervention made to the measuring system. For continuous measurement procedures, there is a possibility that erroneous results have already been reported if an error condition is identified by the next QC sample measurement(s). In either category, a random measurement error that affects only one or a few patient results, called a nonpersistent error, may not be identified by the results for the QC samples. QC procedures only identify persistent error conditions at the point in time when a QC sample is actually measured. Consequently, the choice of criteria to evaluate QC results and the frequency that QC results are measured become important QC plan design considerations.

The design of a QC plan must consider the analytical performance capability of a measurement procedure and the risk of harm to a patient that might occur if an erroneous laboratory test result is used for a clinical care decision. An erroneous laboratory test result is a hazardous condition that may or may not cause harm to a patient depending on how the laboratory test is used for patient monitoring and treatment, the magnitude of error, and what action or inaction is taken by a clinical care provider based on the erroneous result. The following sections in this chapter explain how to establish a QC plan for monitoring a measurement procedure based on information about a measurement procedure’s analytical performance, the analytical performance required to meet medical care requirements, and the risk of harm from an erroneous result. However, establishing the analytical performance specifications to meet medical requirements and evaluating the probability of harm from an erroneous result are challenging because the link between analytical performance and the outcome for the patient can be difficult to establish.<sup>1</sup>



**FIGURE 6.1** Quality control (QC) process for a measurement procedure. (A) QC samples (black) are periodically measured in place of patient samples (red/grey) to determine if the results for QC samples are within expected performance limits for a measurement procedure. (B) If an error occurs in the measuring system, such as a shift to higher results, a QC result can identify that a measurement error condition occurred at some point in time since the last acceptable QC result was measured.

## POINTS TO REMEMBER

### Quality Control

- The primary role of internal QC is to ensure release of correct patient results in real time.
- The primary role of EQA/PT (external QC) is to compare performance between laboratories and, when possible with commutable samples and true values, to determine that a laboratory's measurement procedures conform to expected performance.

## MEASUREMENT PROCEDURE PERFORMANCE AS A PREREQUISITE FOR A QUALITY CONTROL PLAN

### Calibration Traceability to a Reference System

Chapter 7 describes that calibration of clinical laboratory measurement procedures should, whenever possible, be traceable to a higher order reference measurement procedure (RMP) or certified reference material.<sup>2-4</sup> Such calibration ensures that results for patient samples are equivalent within medically acceptable limits irrespective of the measurement procedure or laboratory making the measurements. Calibration is provided by the in vitro diagnostic (IVD) manufacturer for commercially available measurement procedures. In the case of a laboratory developed test, the clinical laboratory produces the measurement procedure and is responsible for its calibration hierarchy including traceability to a reference system when available.

Internal QC is not intended to verify that a measurement procedure is calibrated to a higher order reference system. Rather, QC is intended to verify that the performance, for example, the bias and imprecision, of a measurement procedure remains within acceptable limits during use. A clinical laboratory may wish to verify that a measurement procedure's calibration conforms to an IVD manufacturer's claim for traceability to the reference system used for a given measurand. Some measurement procedure manufacturers provide materials specifically intended for this purpose. Such materials may be provided as measurement procedure-specific QC materials that typically have matrix characteristics and target values that are intended only for use with the specific measurement procedures claimed in the instructions for use and cannot be used with any other manufacturer's measurement procedure.

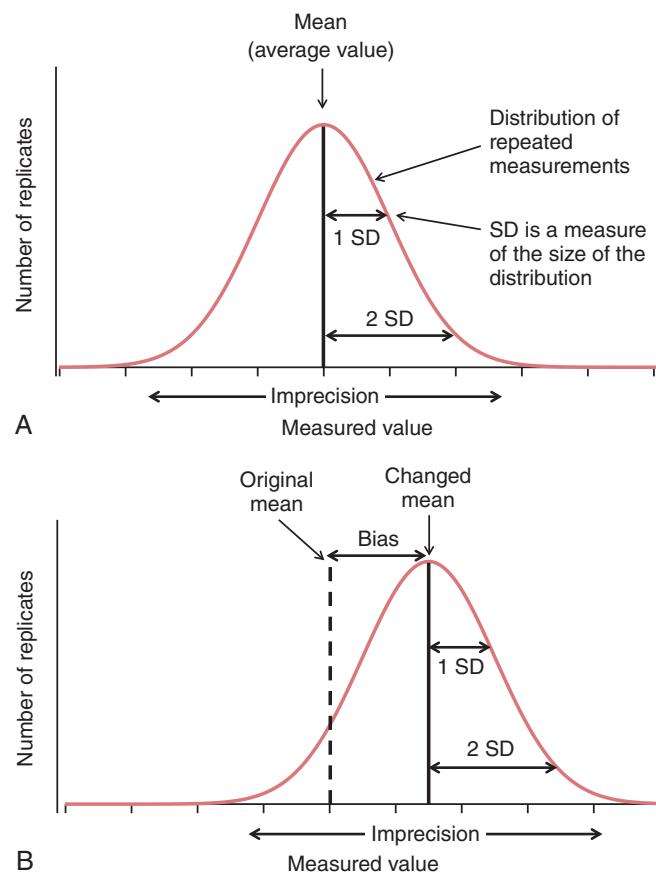
A clinical laboratory has limited resources to verify the calibration traceability of a commercially available or laboratory developed measurement procedure. National and international certified reference materials are available for some measurands. As described in Chapter 7, certified reference materials can be used to verify calibration when those certified reference materials are commutable with clinical samples for use with a specific measurement procedure. The certificate or published validation of a certified reference material should be reviewed for commutability documentation. A laboratory can split clinical samples with a laboratory that offers an RMP to verify calibration. In most cases, a clinical laboratory is dependent on the IVD manufacturer for metrologic traceability of calibration of measurement procedures.

Third-party QC materials intended for statistical process control (i.e., those provided by a manufacturer other than the routine measurement procedure's manufacturer) are not

suitable to verify calibration traceability. These materials are not validated for commutability with clinical samples for different routine measurement procedures, and they do not have target values that are traceable to higher-order RMPs. Such QC materials are designed to be used as QC samples, with target values and standard deviation (SD) values assigned as described later in this chapter. When third-party QC materials are used in an interlaboratory comparison program with measurement procedure-specific peer group mean values, these values can be used to confirm that a laboratory is using a specific measurement procedure in conformance with other users of the same measurement procedure when the results are not influenced by different reagent lots (see External Quality Assessment or Proficiency Testing section).

### Analytical Bias and Imprecision

**Fig. 6.2** illustrates the meaning of bias and imprecision that must be known to develop a QC plan for a measurement procedure. In **Fig. 6.2A**, the horizontal axis represents the numeric value for an individual result, and the vertical axis represents the number of repeated measurements with the same value made on aliquots of a QC material. The red line shows the dispersion of results for repeated measurements



**FIGURE 6.2** (A) Distribution of results showing the mean value and distribution of results (standard deviation, SD) for repeated measurements of a quality control sample. (B) Bias when a change in calibration has occurred. (From Miller WG. Quality control. In: McPherson RA, Pincus MR. *Henry's clinical diagnosis and management by laboratory methods*. 24th ed. Philadelphia: Elsevier; 2020.)

of aliquots of the same QC material, which is the random imprecision of the measurement. Assuming that the dispersion of results follows a Gaussian (normal) distribution, it is described by the SD. The SD is a measure of expected imprecision in a measurement procedure when it is performing within specifications. Note that results near the mean (average value) occur more frequently than results farther away from the mean. An interval of  $\pm 1$  SD includes 68% of the measured values,  $\pm 2$  SDs includes 95% of the values. A result that is more than 2 SDs from the mean is expected to occur 5% of the time (100%–95%) with 2.5% of results in a positive and 2.5% of results in a negative direction from the mean value. Correct calibration of a measurement procedure eliminates systematic bias (within uncertainty limits), so the mean of repeated measurements of a QC sample becomes the expected or target value for that QC sample when the measurement procedure is performing within specifications.

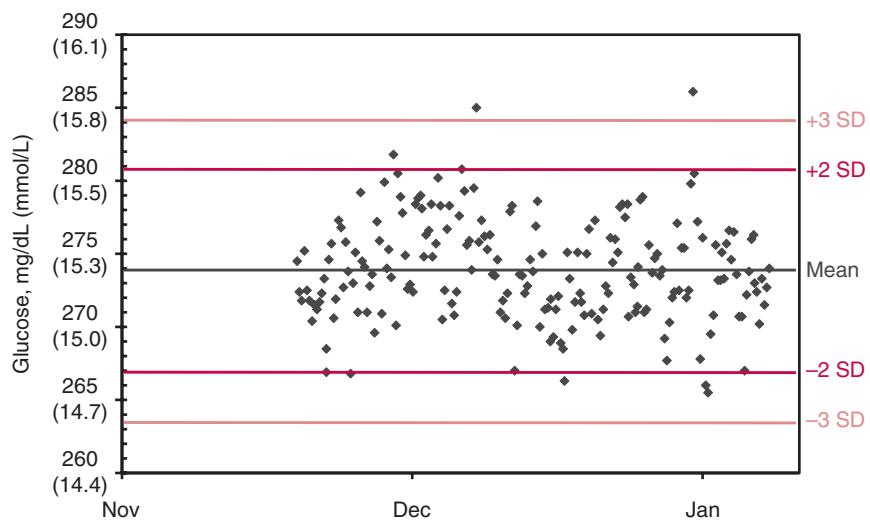
**Fig. 6.2B**, illustrates that if the calibration changes for any reason, a systematic bias is introduced into the results. The bias is the difference between the observed mean and the expected value for a QC material (for more discussion on bias, refer to Chapter 2). Note that the imprecision is the same as before the bias occurred because it is unlikely, although not impossible, that a change in imprecision would occur at the same time as a bias shift. The primary purpose of measuring QC samples is to statistically evaluate the measurement procedure performance to verify that it continues to perform within the specifications consistent with its acceptable expected stable condition or to identify that a change in performance occurred that needs to be corrected. Acceptance criteria for QC results, discussed in a later section, are based on the probability for an individual QC result to be different from the variability in results expected when the measurement procedure is performing within specifications.

The term *accuracy* is used for closeness of agreement of an individual result and a true value and is the combination of bias and imprecision that occurred for that specific measurement

(refer to Chapter 2 for more discussion on accuracy). The bias for an individual patient sample includes any systematic bias in the measurement procedure and the influence of any interfering substances that could be present in that sample. An individual QC sample is only influenced by systematic bias and imprecision of the measurement procedure. Statistical QC does not evaluate possible interfering substances that may affect results for an individual patient sample. The influence of interfering substances needs to be examined during the evaluation that a measurement procedure is suitable for use (refer to Chapter 5 for additional discussion on interference). However, the imprecision observed for QC results provides a measure of the variability expected for an individual patient result caused by the inherent imprecision of a measurement procedure and is usually independent of interfering substances that typically affect the bias for an individual patient result.

The term *trueness* is inversely related to a bias that may be present in a measurement procedure. Trueness is an attribute of a measurement procedure that reflects how correctly its calibration is traceable to a reference system.

**Fig. 6.3** shows a Levey-Jennings<sup>5</sup> plot that was an adaptation for clinical laboratory measurements of the Shewhart<sup>6</sup> plot developed for statistical process control in manufacturing. The Levey-Jennings plot is the most common presentation for evaluating QC results. This format shows each QC result sequentially over time and allows a quick visual assessment of performance. Assuming the measurement procedure is performing in a stable condition consistent with its specifications, the mean value represents the target (or expected) value for the QC result, and the SD lines represent the expected imprecision. Assuming a Gaussian (normal) distribution of imprecision, the results should be distributed uniformly around the mean with results observed more frequently closer to the mean than near the extremes of the distribution. Note that a few results in **Fig. 6.3** are greater than 2 SDs, and two results slightly exceed 3 SDs, which is expected for a Gaussian distribution of imprecision. For a



**FIGURE 6.3** Levey-Jennings plot of quality control (QC) results ( $n = 199$ ) for a single lot of QC material used for a 49-day period. SD, Standard deviation. (From Miller WG. Quality control. In: McPherson RA, Pincus MR. *Henry's clinical diagnosis and management by laboratory methods*. 24th ed. Philadelphia: Elsevier; 2020.)

large number of repeated measurements, the number of results expected within the SD intervals is as follows:

- $\pm 1$  SD = 68.3% of observations
- $\pm 2$  SD = 95.4% of observations
- $\pm 3$  SD = 99.7% of observations

Interpretation of an individual QC result is based on its probability to be part of the expected distribution of results for the measurement procedure when the procedure is performing correctly. A later section provides details regarding interpretive rules for evaluation of QC results. Note that evaluation of individual QC results may be performed by computer algorithms without visual examination of a Levey-Jennings chart. However, the underlying logic of such algorithms is illustrated by the Levey-Jennings chart example.

### Performance of a Measurement Procedure for Its Intended Medical Use

It is necessary to determine how the performance of a measurement procedure relates to the intended medical use for interpreting results in order to determine the frequency to measure QC samples and the criteria to use to evaluate the QC results. The sigma metric is commonly used to assess how well a measurement procedure performs relative to the analytical performance specifications that ideally should be based on the intended medical use of the results. Sigma is the Greek letter used to denote SD. The sigma scale expresses the variability of a measurement process in SDs in relation to the variability that is acceptable because it will not cause an error in diagnosis or treatment of a patient.

For laboratory measurements, the sigma metric is calculated as:

$$\text{Sigma} = \frac{(\text{TE}_a - \text{bias})}{\text{SD}}$$

where  $\text{TE}_a$  is the total error allowed based on analytical performance specifications that ideally should be related to the intended medical use for interpreting results, and bias and SD refer to performance characteristics of the measurement procedure. The SD is estimated from the QC data as previously described. It is critically important that the estimate of SD be made using QC data that represent all or most components of variability that occur over an extended time period (see the section called Establishing the Quality Control Target Value and Standard Deviation That Represent a Stable Measurement Operating Condition). The bias is difficult for a laboratory to estimate because it is difficult to evaluate if a particular measurement procedure has a bias compared with a reliable estimate of a true value such as based on an RMP. For internal QC, a laboratory is usually interested to determine if a bias has occurred compared with the condition established by calibration of a measurement procedure. Such a bias represents a QC result that is sufficiently different from its target value that corrective action is needed. Consequently, the bias is usually assumed to be zero for calculating sigma.

However, a bias term may be needed in situations when there are two or more different measurement procedures used for the same measurand and those different measurement procedures have a bias between them that cannot be removed, or when changes in lots of reagents or calibrators introduce shifts in bias that cannot otherwise be corrected. Note that it is preferable to adjust the calibration of different measurement procedures or different lots of reagents or

calibrators to provide equivalent results, but this solution may not be applicable for some technologies. In such cases, this relative bias can be estimated based on comparison of results for patient samples following a procedure such as described in Clinical and Laboratory Standards Institute (CLSI) document EP9.<sup>7</sup> That bias should be considered in determining the sigma metric and in establishing a QC plan for such measurement procedures.

$\text{TE}_a$  represents the measurement procedure performance required to enable suitable medical decisions based on a test result. A test result may be used for different medical decisions in different disease conditions. In a main lab setting where samples from different medical practices are measured, the most stringent decision parameter should be used as the basis for the  $\text{TE}_a$ . In a setting where the samples are used for one specific clinical situation, for example, in a point-of-care (POC) setting, the medical requirements of the setting can be used as the basis for the  $\text{TE}_a$ .  $\text{TE}_a$  can be estimated using three models.<sup>8</sup> The preferred model (model 1) to set a performance specification is to base it on an outcome study (i.e., investigating the impact of analytical performance of the measurement procedure on the clinical outcome). Outcome studies can be direct assessment of clinical outcome for a group of patients or “indirect” outcome when the consequences of analytical performance on, for example, clinical classifications or decisions and thereby on the probability of patient outcomes can be investigated. These probabilities can be discussed with clinical experts who then can recommend a performance specification.<sup>8</sup>

Indirect outcome studies are often used to set  $\text{TE}_a$  in laboratory practice guidelines. For example, the National Cholesterol Education Program recommends that total cholesterol be measured with a  $\text{TE}_a$  of 9% or less,<sup>9</sup> and the National Kidney Disease Education Program recommends that creatinine be measured with a  $\text{TE}_a$  of less than 7 to 10% in the concentration interval 1 to 1.5 mg/dL (76.3 to 114.4 mmol/L).<sup>10</sup> The limitation of this model is that it is only useful when the links between the measurand, clinical decision-making, and clinical outcomes are strong, which is the case for a minority of measurands.

Another model (model 2) tries to minimize the ratio of the “analytical” noise to the “biologic signal” with an assumption that a small ratio will identify measurement procedure performance that relates to the medical requirements. The biological variation is composed of within and between subject variation. Performance specifications for imprecision, bias, and  $\text{TE}_a$  are based on a fraction of the within and between individual biologic variations of the measurand.<sup>11,12</sup> Tables of optimal, desirable, and minimal  $\text{TE}_a$  based on biologic variation are available and may provide useful information.<sup>13,14</sup> However, biologic variation-based estimates of  $\text{TE}_a$  should be used with caution because the estimates of biologic variation in many cases are based on limited data, and the experimental designs of the estimates and the process to select the estimates to list in the tables have been challenged.<sup>15–17</sup> Estimates of biologic variation typically vary among different investigations.<sup>18–20</sup> The newly established EFLM database on biological variation<sup>14</sup> evaluates published reports on biological variation using a critical appraisal checklist<sup>21</sup> and calculates point estimates with confidence intervals for each measurand after a meta-analysis of eligible reports. In addition, the way the  $\text{TE}_a$  is calculated is flawed because the calculation

combines maximum allowable imprecision with maximum allowable bias (both based on a fraction of biologic variation) that has no theoretical basis and leads to overestimation of the  $TE_a$ .<sup>15</sup> Another limitation is that the biologic variability has typically been derived from data for nondiseased individuals and may be different for pathologic conditions. Additional examples and discussion of biologic variation are provided in Chapter 8.

Model 3 bases the performance specifications on the “state of the art.” The advantage of this model is that data are readily available from QC and EQA/PT information. The disadvantage is that there may be no relationship between what is technically achievable and what is needed to make a medical decision for diagnosis or treatment of a patient. It is generally agreed that preference should be given to model 1 whenever such information is available or to model 2 as a starting point to estimate  $TE_a$ .<sup>8,22</sup> A laboratory director should consult with clinical care providers to agree on an appropriate  $TE_a$  for the patient population served. An extended presentation of analytical performance specifications is given in Chapter 8.

Because sigma assumes a Gaussian or normal distribution for repeated measurements, the probability of a defect (i.e., an erroneous laboratory result) can be predicted as shown in Table 6.1. The sigma metric represents the probability that a given number of erroneous results that could cause risk of harm to a patient are expected to occur when the test measurement procedure is performing to its specifications. The phrase “six sigma” refers to a condition when the variability in the measurement process is sufficiently smaller than the medical requirement that erroneous results are very uncommon. A “four-sigma” measurement procedure would be less robust and have a higher probability that erroneous results could be produced but still at a fairly low frequency. A “two-sigma” measurement procedure would produce enough erroneous results even though it met its performance specifications that it would not be very reliable for patient care.

**TABLE 6.1 Probability of Acceptable or Erroneous Results Based on the Sigma Scale<sup>a</sup>**

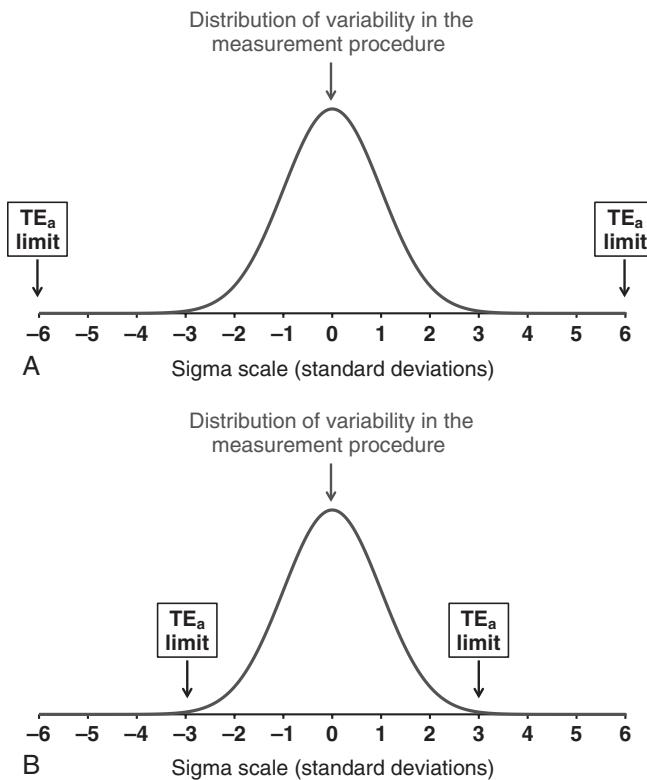
Sigma Value	Percent of Results Within Specification	Percent of Results With an Error (Defect)	Errors (Defects) per Million Opportunities
1	68	32	317,311
2	95.5	4.5	45,500
3	99.7	0.3	2700
4	99.994	0.006	63
5	99.99994	0.00006	0.6
6	99.999998	0.0000002	0.002

<sup>a</sup>The values in this table are based on a Gaussian statistical distribution and do not include the “1.5 sigma shift” frequently introduced to recognize that many manufacturing processes have been observed to have a long-term drift approximately  $\pm 1.5$  SD when operating in a stable condition. The 1.5 sigma shift is not used for QC rules design.

From Miller WG. Quality control. In: McPherson RA, Pincus MR. *Henry's clinical diagnosis and management by laboratory methods*. 24th ed. Philadelphia: Elsevier; 2020.

Fig. 6.4 shows how the sigma metric describes the performance of a laboratory test relative to the  $TE_a$ . Parts A and B show that a measurement procedure with the same analytical performance characteristics can have different sigma metrics depending on how the imprecision relates to the  $TE_a$ . Fig. 6.4A shows a “six-sigma” measurement procedure that has the  $TE_a$  limits 6 SDs away from the center point of the distribution of variability in measurements when the measurement procedure is performing to its analytical specifications. In the “six-sigma” situation, a small amount of bias or increased imprecision will have little influence on the number of erroneous results produced, and less stringent QC will be appropriate because the risk of producing an erroneous result even with some loss of performance is very low.

Fig. 6.4B shows a “three-sigma” measurement procedure that has the  $TE_a$  limits 3 SDs away from the center point of the expected distribution of variability in measurements when the measurement procedure is performing to its analytical specifications. In the “three-sigma” situation, a small amount of bias or increased imprecision will cause the number of erroneous results to increase substantially, and more stringent QC is needed to identify when such an error condition occurs so that corrective action can be initiated. Note that no amount of QC will improve the performance of a marginal measurement procedure. However, more frequent QC and more stringent acceptance criteria will allow the



**FIGURE 6.4** Measurement procedure performance relative to the sigma scale to describe how well performance meets medical requirements expressed as the allowable total error ( $TE_a$ ). (A) A “six-sigma” measurement procedure. (B) A “three-sigma” measurement procedure. (From Miller WG. Quality control. In: McPherson RA, Pincus MR. *Henry's clinical diagnosis and management by laboratory methods*. 24th ed. Philadelphia: Elsevier; 2020.)

laboratory to more quickly identify when small changes in performance occur so they can be corrected to minimize the risk of harm to a patient from erroneous results being acted on to make medical care decisions. It is important to emphasize that the sigma calculations are dependent on what  $TE_a$  is chosen. As discussed earlier, an “objective”  $TE_a$  is often difficult to establish, and good data to set a  $TE_a$  are often lacking.

### POINTS TO REMEMBER

#### Measurement Procedure Performance as a Prerequisite for a Quality Control Plan

- The performance characteristics of a measurement procedure when it is performing in a stable in-control condition must be understood.
- The allowable total error for a measurement procedure must be established based on analytical performance specifications that ideally should be based on the intended medical use of a laboratory result in patient care decisions.
- The sigma metric represents the probability that a given number of erroneous results that could cause risk of harm to a patient are expected to occur when the test measurement procedure is performing to its specifications.

## DEVELOPING A QUALITY CONTROL PLAN AND IMPLEMENTING QUALITY CONTROL PROCEDURES

### Selection of Quality Control Materials

Generally, two different concentrations are necessary for adequate statistical QC. For quantitative measurement procedures, QC materials should be selected to provide measurand concentrations that monitor the analytical measuring interval of the measurement procedure. In practice, laboratories are frequently limited by concentrations available in commercial QC products. When possible, it is important to confirm that measurement procedure performance is stable near the limits of its analytical measuring interval because defects may affect these concentrations before others. Many quantitative measurement procedures have a linear response over the analytical measuring interval, and it is reasonable to assume that their performance over the interval is acceptable if the results near the interval limits are acceptable. In the case of nonlinear analytical response, it may be necessary to use additional controls at intermediate concentrations. Concentrations of control materials close to clinical decision values (e.g., glucose, therapeutic drugs, thyroid-stimulating hormone, prostate-specific antigen, hemoglobin A<sub>1c</sub> [HbA<sub>1c</sub>], troponin) may also be appropriate for additional QC monitoring. In many cases, the imprecision near the limit of quantitation may be relatively large, in which case the concentration should be chosen to provide adequate SD for practical evaluation of QC results. For procedures with extraction or other pretreatment steps, controls must be used that are subject to the same pretreatment steps.

This chapter primarily focuses on QC procedures for quantitative measurement procedures. However, the principles can be adapted to most qualitative procedures with allowances for the lack of numeric results. For measurement

procedures based on qualitative interpretation of quantitative measurements (e.g., drugs of abuse, human chorionic gonadotropin, hepatitis markers), the same principles of QC assessment can be applied to the numeric results even if they are only expressed as instrument signal values. For qualitative results, the negative and positive controls should be selected to have concentrations relatively near the clinical decision threshold to adequately control for discrimination between negative and positive. For qualitative procedures with graded responses (e.g., dipstick urinalysis), negative, positive, and graded response controls are required. For qualitative tests based on other properties (e.g., electrophoretic procedures, stain adequacy, immunofluorescence, organism identification), it is necessary to ensure that the QC procedure will appropriately evaluate that the measurement procedure correctly discriminates normal from pathologic conditions.

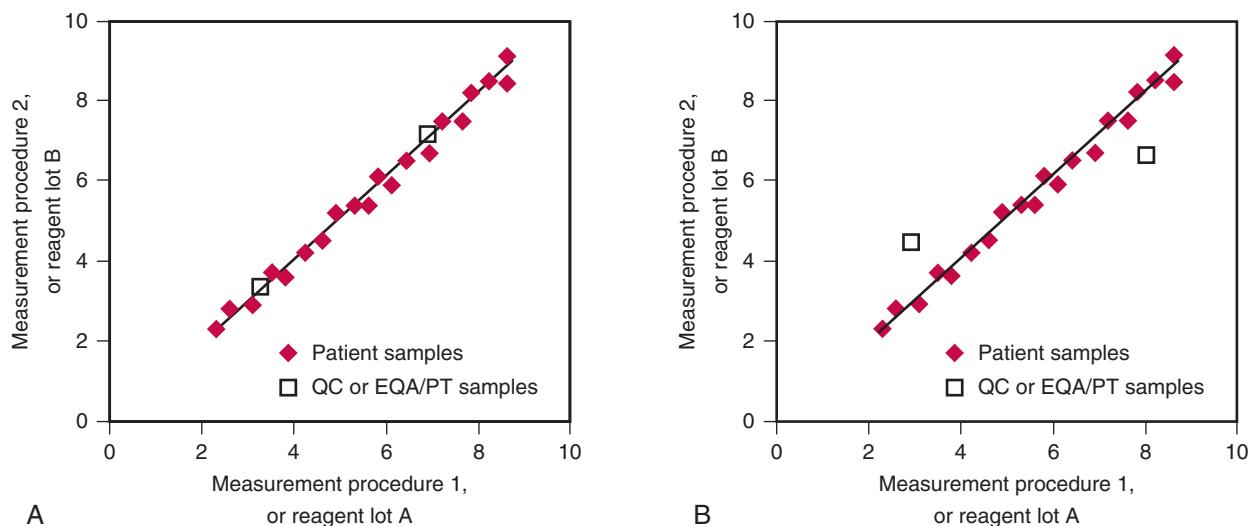
The QC samples selected must be manufactured to provide stable materials that can be used for an extended time period, preferably 1 or more years for stable measurands. Use of a single lot for an extended period allows reliable interpretive criteria to be established that will permit efficient identification of a measurement procedure problem, avoid false alerts caused by poorly defined expected ranges for the QC results, and minimize limitations in interpreting values after reagent and calibrator lot changes.

### Limitations of Quality Control Materials

Limitations are inherent in most commercially available QC materials. One limitation is that the QC material is frequently noncommutable with patient samples. Commutability is a property of a reference or control material that refers to how well that material mimics patient samples in measurement procedures. A commutable QC material is one that reacts in a measurement procedure to give a result that would closely agree with that expected for a patient sample containing the same amount of measurand. Fig. 6.5A shows that results from patient samples and from commutable QC (or EQA/PT) samples have the same relationship between two measurement procedures or between two reagent lots used with the same measurement procedure. Fig. 6.5B shows that noncommutable samples have a different relationship than observed for patient samples.

QC, as well as EQA/PT materials, are typically noncommutable with patient samples because the serum or other biologic fluid matrix is usually altered from that of a patient sample.<sup>23–27</sup> The matrix alteration is due to processing of the biologic fluid during product manufacturing; use of partially purified human and nonhuman additives to achieve desired concentrations of the measurands; and various stabilization processes that alter proteins, cells, and other components. The impact of the matrix alteration on measurement of a measurand is not predictable and is frequently different for different lots of QC material, for different lots of reagent within a given measurement procedure, and for different measurement procedures.<sup>28,29</sup> Because of the noncommutability limitation, special procedures are required (discussed in later sections) when changing lots of reagent or comparing QC results among two or more measurement procedures.

A second limitation of QC materials is deterioration of the measurand during storage. Measurand stability during unopened storage is generally excellent, but slow deterioration eventually limits the shelf life of a product and can introduce



**FIGURE 6.5** Illustration of commutable and noncommutable samples. (A) Quality control (*QC*) or external quality assessment/proficiency testing (*EQA/PT*) samples (open white squares) are commutable when they have the same relationship between two measurement procedures, or two reagent lots, as observed for patient samples (red diamonds). (B) Noncommutable QC or EQA/PT samples have a different relationship than observed for patient samples.

a gradual drift into QC data that may require correction over the life of a lot of QC material. Measurand stability after reconstitution, thawing, or vial opening can be an important source of variability in QC results and can vary substantially among measurands in the same vial. Variables to be controlled are the time spent at room temperature and the time spent uncapped with the potential for evaporation. An expiration time after opening is provided by the QC material manufacturer but may need to be established by a laboratory for each QC material under the conditions of use in that laboratory and may be different for different measurands in the same QC product. For QC materials reconstituted by adding a diluent, vial-to-vial variability can be minimized by standardizing the pipetting procedure (e.g., using the same pipet or filling device, preferably an automated device, and having the same person prepare the controls) whenever practical.

Another limitation of QC materials is that measurand concentrations in multiconstituent control materials may not be at levels optimal for all measurement procedures. This limitation may be caused by solubility considerations or potential interactions between different constituents, particularly at higher concentrations. It may be necessary to use supplementary QC materials to adequately monitor the measuring interval and clinically important decision concentrations.

### Frequency to Measure Quality Control Samples

Determining the frequency to measure QC samples should use a risk assessment approach. The frequency to measure QC samples is a function of several parameters:

- Analytical stability of the measurement procedure
- Risk of harm to a patient from clinical action being taken before a significant error is detected at the next scheduled QC event
- Number of patient results produced in a period of time when an error condition could exist but is not yet detected
- Scheduled events such as recalibration or maintenance that may alter the current performance condition of the measurement procedure

- Training and competency of the test operator, particularly for manual or semiautomated measurement procedures

### Analytical Stability of the Measurement Procedure

The stability of the measurement procedure is a fundamental determinant of how frequently a QC sample needs to be measured. The more stable the measurement procedure, the less frequently a QC evaluation needs to be performed. Note, however, that all of the considerations in the preceding list must be evaluated together to determine a suitable frequency to perform QC. Some measurement procedures have been designed with sophisticated built-in control procedures to mitigate the risk that an erroneous result may be produced. Built-in control procedures may include calibrators and QC materials integrated with reagent packaging, and sensors that monitor electronic components and the measurement process with algorithms that prevent a result from being produced if any monitored conditions fail to meet criteria. Examples of built-in controls are frequently found in POC instruments. These measurement systems may be sufficiently stable and self-monitored to justify reduced frequency of traditional surrogate QC sample testing. However, there is little information in the literature that has examined the optimal frequency or control rules to be used in these cases.

### Risk of Harm to a Patient and Number of Patients Who May Be at Risk

The risk of clinical action being taken before a significant measurement error is detected is an important consideration for more frequent QC measurements than one based strictly on analytical stability of the measurement procedure or on regulatory minimum requirements. More frequent QC measurements are appropriate to avoid the situation of discovering a measurement procedure defect many hours after a physician has made a clinical treatment or nontreatment decision based on an erroneous result. For example, QC sampling performed on a 24-hour cycle might be performed at 9 A.M. If the next QC results indicate a measurement procedure

problem, the erroneous condition could have started at any time during the previous 24 hours. If the problem had occurred at 3 P.M. the previous day, erroneous results could have been reported for 18 hours, likely putting a large number of patients at risk of an inappropriate medical decision. Parvin<sup>30</sup> reported an assessment of the frequency of QC testing and the number of potentially incorrect patient results that could be reported before errors of different magnitudes were detected.

The medical risk of harm to a patient from erroneous results must be considered and the frequency of QC testing established to reduce the risk to an acceptable level. From a practical perspective, the cost of a medical error, or simply the cost of repeating questionable patient samples since the last acceptable QC results, could be more expensive than a more frequent QC measurement schedule that would detect an error condition in a more timely manner.

The CLSI has published guideline EP23 addressing risk-based QC procedures.<sup>31</sup> The document provides guidance to clinical laboratories on how to develop a QC plan based on evaluation of risk of harm to a patient and assessment of the effectiveness of risk mitigation procedures, including QC, used with a measurement procedure. Information about measurement procedure performance is obtained from the manufacturer, from laboratory validation and QC data, and from other literature sources that is combined with the clinical requirements of the local health care setting. In general terms, the laboratory director is responsible for ensuring that a result has a high probability of being correct at the time it is reported for clinical use. To make this judgment, the laboratory director needs to understand the risks that can cause a measurement technology to perform incorrectly, needs to evaluate the effectiveness of built-in control processes to mitigate those risks, and needs to ensure that adequate control procedures are in place to confirm that a result is correct at the time it is reported. A combination of built-in and QC samples-based monitoring procedures can be used to ensure that all risks have been appropriately mitigated or monitored at a frequency commensurate with the risk of malfunction and the risk of harm to a patient if an incorrect result was reported.

### Measurement of Quality Control Samples Based on Scheduled Events

Laboratory operations typically implement two types of testing schedules. Batch processes measure a group of clinical samples as a batch and testing for all samples is completed before reporting the results. For batch processes, QC samples can be measured concurrent with the patient samples, for example, multiple well reaction plates, or at the beginning and end of a sequential series of measurements to ensure all results have a high probability of being correct before reporting.

Continuous measurement processes are common in higher-volume settings, for example, automated chemistry or hematology, when patient samples are measured and results reported continuously as they are received in the laboratory with QC sample measurements made periodically during the course of the process. When using a continuous measurement system, scheduled events such as recalibration or maintenance are performed to ensure the measurement conditions continue to meet specifications and to correct for any calibration drift or component deterioration that may have

occurred. Measuring QC samples before such scheduled events allows the laboratory to verify that no significant errors in results have occurred since the last time QC samples were measured.<sup>32</sup> If QC samples are not measured before such scheduled events, a laboratory will not know if erroneous results for patient samples were reported since the time of measuring the last QC samples and initiation of the scheduled event.

The laboratory director uses a risk assessment approach to determine when QC samples should be measured before a scheduled event. For example, daily cleaning procedures are intended to maintain the measuring system in good working condition and may not require QC assessment before the event. Whereas, a maintenance event that replaces components will produce an altered measuring system and QC assessment before the event is the only way to determine that no erroneous results were reported since the last acceptable QC results before the event. QC samples should be measured after scheduled events to verify that the operations were performed correctly, and that measurement procedure performance meets specifications before restarting to measure patient samples.

If a malfunction occurs during a continuous measurement process such that the measuring system becomes non-operable, that condition must be treated as a QC failure with follow up to repeat and confirm the acceptability of already reported patient sample results. This follow up action is required because the malfunction prevents measuring QC samples to determine if erroneous results were reported since the time of the last acceptable QC assessment.

### Establishing the Quality Control Target Value and Standard Deviation That Represent a Stable Measurement Operating Condition

QC target values and acceptable performance limits are established to optimize the probability of detecting a measurement defect that is large enough to have an impact on clinical care decisions while minimizing the frequency of “false alerts” caused by statistical limitations of the criteria used to evaluate QC results. The measurement system must be correctly calibrated and operating within acceptable performance specifications before the statistical parameters to establish QC interpretive rules can be established. Some sources of measurement variability that are expected to occur during typical operation of a measurement procedure are listed in Table 6.2. Measurement variability includes sources with short time interval frequencies, many of which can be described by Gaussian error distributions, and intermittent and longer time interval sources, which can cause cyclic fluctuations over several days or weeks, gradual drift over weeks or months, and intermittent abrupt small shifts in results. The SD used in QC interpretive rules needs to adequately represent all sources of variability in results that are expected to occur over time when the measurement procedure is performing to specifications.

### Quality Control Material Target Value

A QC material must have a reliable target value that represents the condition when systematic bias is as small as possible. This condition requires the measurement procedure to be calibrated correctly. For practical reasons, the sources of

**TABLE 6.2 Common Sources of Measurement Variability**

Source	Time Interval for Fluctuation	Likely Statistical Distribution
Pipet volume	Short	Gaussian
Pipet seal deterioration	Long	Drift
Instrument temperature control	Short or long	Gaussian or other
Electronic noise in the measuring system	Short	Gaussian
Calibration cycles	Short to long	Gaussian or periodic drift/shift
Reagent deterioration in storage	Long	Drift
Reagent deterioration after opening	Intermediate	Cyclic, periodic drift or shift
Calibrator deterioration in storage	Long	Drift
Calibrator deterioration after opening	Intermediate	Cyclic, periodic drift or shift
Control material deterioration in storage	Long	Drift
Control material deterioration after opening	Intermediate	Cyclic, periodic drift or shift
Environmental temperature and humidity	Variable	Variable
Reagent lot changes <sup>a</sup>	Long	Periodic shift
Calibrator lot changes	Long	Periodic shift
Instrument maintenance	Variable	Cyclic or periodic shift
Deterioration of instrument components	Variable	Cyclic, periodic drift or shift

<sup>a</sup>Note that reagent lot changes can have an artifactual influence on quality control values and require special handling as discussed in the section Verifying Quality Control Evaluation Parameters After a Reagent Lot Change.

From Miller WG. Quality control. In: McPherson RA, Pincus MR. *Henry's clinical diagnosis and management by laboratory methods*. 24th ed. Philadelphia: Elsevier; 2020.

variability in Table 6.2 are not represented in the time available to establish a target value. Consequently, the target value has uncertainty and needs to be refined over time to reflect the expected variability in measurement conditions within the performance specifications for a measurement procedure.

The generally accepted minimum protocol for target value assignment is to use the mean value from at least 10 measurements of the QC material on 10 different days.<sup>32</sup> The 10 or more measurements can be performed over a longer time interval to include other important sources of variability in the estimate of target value. When applicable, more than one calibration should be represented in the 10 measured values to include this source of variability in the target value. If a QC material will be used for longer than 1 day, a single vial should be stored correctly and measured on as many days as the material is planned to be used to allow variability caused by opened storage to be represented in the target value. If a 10-day protocol is not possible (e.g., if an emergency replacement of a lot of QC material is necessary), a provisional target value can be established with fewer data but should be updated when additional replicate results are available.

Because all sources of variability cannot be captured in 10 measurements, it is recommended to update the target value after more data have been acquired during use of the QC material. Target values may also need to be updated after reagent lot changes or other alterations in measurement conditions as described in a later section called Verifying Quality Control Evaluation Parameters After a Reagent Lot Change.

### Quality Control Material Standard Deviation

SD is the conventional way to express measurement procedure variability and assumes the QC data can be described by a Gaussian (normal) distribution even though non-Gaussian components of variability influence the QC results. The statistical QC packages in instrument and laboratory computer systems are designed with the assumption that the SD for

a Gaussian distribution is used for the QC rules criteria. Because there are non-Gaussian components to measurement variability over time, it is very important that they be represented in the data used to estimate an SD used to make conclusions regarding the acceptability of an individual QC value. The SD must be as realistic as possible to represent the variability expected for a measurement procedure when its performance meets its specifications.

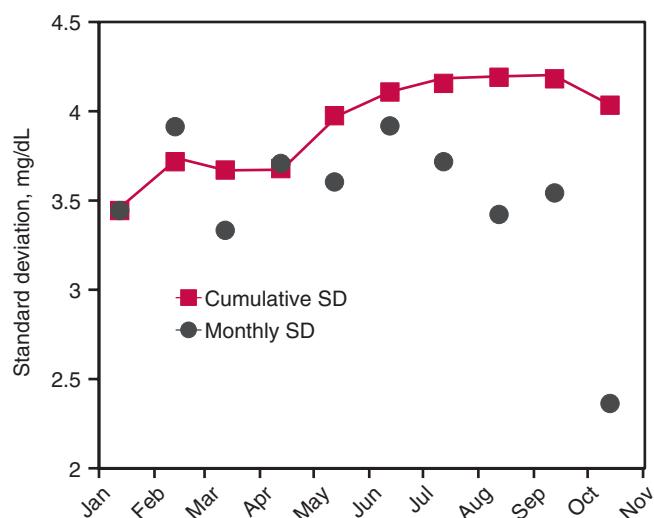
When a measurement procedure has been established in a laboratory and a new lot of QC material is being introduced, the target value for the new lot of QC material is used along with the well-established long-term SD from the previous lot. This practice is appropriate because in most cases, measurement imprecision is a property of the measurement procedure and equipment used and is unlikely to change with a different lot of QC material.<sup>32</sup> The SD from a small number of replicate measurements made on a new lot of QC material should not be used because it will not include most sources of variability, will underestimate the SD, and thus cause meaningless false positive QC rule failures. When a new QC material from a different manufacturer is introduced, it is possible for the observed SD to be different than the historic value because of matrix differences, and the SD should be monitored and adjusted as experience with the new material is accumulated. If target values for the old and new lots are substantially different, a different SD may be needed. Assuming the coefficient of variation (CV) is approximately constant over the difference in concentration between target values for the old and new QC lots, the SD can be calculated by applying the existing CV to the new target value and then converting to the corresponding SD value. Because it is possible for the SD to be influenced by a lot of QC material, adjustment to the SD may be necessary as additional experience with the new lot is accumulated.

If a new measurement procedure is introduced for which no historical performance information is available, the SD for stable performance can be established using QC data obtained during the measurement procedure validation. A

minimum of 20 results on different days is recommended for the initial estimation of the SD.<sup>32</sup> Not all of the sources of variability in Table 6.2 will be included in the initial estimate of SD and this SD will likely be smaller than an estimate based on longer-term data that includes most sources of variability. It is desirable to have some of the events that contribute to measurement variability, in particular calibration and maintenance, included during the time interval over which the SD is estimated. Note that reagent lot changes should not be included in the estimate of SD because QC results are frequently artificially influenced by different reagent lots (see the section Verifying Quality Control Evaluation Parameters After a Reagent Lot Change). The CLSI document EP05 provides guidance on establishing the SD for a measurement procedure.<sup>33</sup> Note that EP05 does not include longer-term sources of variability, so the long-term SD is underestimated by this protocol.

When a new measurement procedure replaces an existing procedure, the SD for the existing procedure can in many cases be used to inform the initial estimate of the SD for the new measurement procedure. With the assumption that the SD for the existing measurement procedure was appropriate to ensure the results were suitable for use in medical decisions, that SD can be used as the basis for QC decisions for the new measurement procedure. This approach is suitable when the initial estimates of SD for the new procedure are smaller than the SD in use for the old procedure. This approach may allow an initial estimate of SD that is consistent with the intended use of the results for medical decisions until sufficient QC results have been obtained for the new measurement procedure to estimate a new SD that includes sources of variability that are not possible to include in the initial estimate of SD determined during validation of the new measurement procedure.

The initial estimate of SD will likely not include contributions from all expected sources of variability and will need to be updated when additional QC data are available. An estimate of SD can vary with measurement conditions over different time intervals with more robust values obtained for longer time intervals that include most sources of variability.<sup>34</sup> An SD that represents stable measurement performance can usually be estimated from the cumulative SD over a 6- to 12-month period, ideally for a single lot of QC material and reagent. Fig. 6.6 illustrates the fluctuation in SD that occurred when calculated for monthly intervals compared with the relatively stable value observed for the cumulative SD after a period of 6 months. Note that the cumulative SD is not the average of the monthly values but is the SD determined from all individual results obtained over a time interval since the lot of QC material was first used. Different sources of long-term variability occur at different times during the use of a measurement procedure. The monthly SD does not adequately reflect the longer-term components of variability. Consequently, the cumulative SD is typically larger than the monthly values because it includes more sources of variability (see Table 6.2) and better represents the actual variability of the measurement procedure. If the SD expected during normal stable operation is underestimated, the acceptable range for QC results will be too small, and the false-alert rate will be unacceptably high. If SD for the stable condition is overestimated, the acceptable range will be too large, and a significant measurement error might go undetected.



**FIGURE 6.6** Cumulative standard deviation (SD) versus single monthly values calculated from the data in Fig. 6.9. (From Miller WG. Quality control. In: McPherson RA, Pincus MR. *Henry's clinical diagnosis and management by laboratory methods*. 24th ed. Philadelphia: Elsevier; 2020.)

Data can be combined from more than one lot of QC material or for more than one lot of reagent using a statistical pooling approach to obtain a long-term estimate of SD that represents most sources of variability in a measurement procedure. For example, this approach may be needed when the stability of a QC product is limited, and a single lot is not available for an extended time interval or when reagent stability is limited, and a new reagent lot must be used frequently. See the later section called Verifying Quality Control Evaluation Parameters After a Reagent Lot Change that explains why QC results can be artificially influenced by reagent lot changes. Pooling of data to obtain an SD requires the SD for each stable interval of use to be determined separately; then the SD for each stable interval can be combined using the following formula where  $n$  is the number of QC results in a given time interval from 1 to  $i$ .

$$SD_{\text{pooled}} = \sqrt{\frac{(n_1 - 1) \times SD_1^2 + (n_2 - 1) \times SD_2^2 + \dots + (n_i - 1) \times SD_i^2}{n_1 + n_2 + \dots + n_i - i}}$$

It is important to include all valid QC results in the calculation of SD to ensure that the SD correctly represents expected measurement procedure variability. A valid QC result is one that was, or would have been in the case of preliminary value assignment, used to verify acceptable measurement procedure performance and supported reporting patient results. Only QC results that were, or would have been, responsible for not reporting patient results should be excluded from summary calculations.

### Quality Control Materials With Preassigned Values

Some QC materials are provided by the measurement procedure manufacturer with preassigned target values and acceptable ranges intended to confirm that the measurement procedure meets the manufacturer's specifications. Such assigned values may be used to verify the manufacturer's

specifications. However, it is recommended that both the target value and the SD should be reevaluated and assigned by the laboratory after adequate replicate results have been obtained because the QC interpretive rules used in a single laboratory should reflect performance for the measurement procedure in that laboratory. The acceptability limits (product insert ranges) suggested by a manufacturer typically account for sources of variability, such as among instruments, among reagent lots, and among calibrator lots, that will be greater than the variability expected in an individual laboratory. Use of product insert ranges that are too large will reduce a laboratory's ability to detect an erroneous measurement condition. It is often a problem for POC devices that the manufacturer's preset limits must be used that can reduce the possibilities to detect errors. A laboratory can institute additional QC testing, but such options may be limited for some technologies.

QC materials with assigned target values and SDs are also available from third-party manufacturers (i.e., manufacturers not affiliated with the measurement procedure manufacturer) and typically have values that are applicable to specifically stated measurement procedures to accommodate the influence of noncommutability. Caution should be used with target values and SDs assigned by third-party QC material providers because the target values may have been assigned by a small number of measurements and using reagent and calibrator lots that are no longer available. The SD values will not be suitable for use in QC acceptance rules because they do not reflect the measurement conditions in an individual laboratory. Of particular concern is, for example, when a QC material manufacturer assigned SD is larger than that observed in an individual laboratory, the acceptable limits for QC results will be too large, and an erroneous measurement condition may not be identified appropriately.

Some QC material providers offer an interlaboratory comparison program to which participants send QC results for aggregation with those of other laboratories. Such interlaboratory summary data are similar to those from EQA/PT programs and can be useful to laboratories to compare their target values to those from a group of laboratories using the same measurement procedure and lot of QC material (see the section called External Quality Assessment or Proficiency Testing). The within group SD values from aggregated QC data are not suitable for use in an individual laboratory because the values do not reflect the measurement conditions in an individual laboratory. The SD from aggregated QC data is likely to be larger than the SD in an individual laboratory causing the acceptable limits for QC results to be too large, and an erroneous measurement condition to not be identified appropriately.

### Establishing Rules to Evaluate Quality Control Results

The acceptable range and rules for interpretation of QC results are based on the probability of detecting a significant analytical error condition with an acceptably small false-alert rate. The desired process control performance characteristics must be established for each measurand before the appropriate QC rules can be selected.

The conventional way to express QC interpretive rules is by using an abbreviation nomenclature popularized among clinical laboratories by Westgard<sup>35</sup> and summarized in Table 6.3. Note that fractional standard deviation intervals (SDIs) can be used as in the  $2_{2.5S}$  and  $8_{1.5S}$  examples and that combinations of numbers of controls and limits can be used as appropriate for QC interpretive rules. Statistical procedures, such as cumulative sum (CUSUM), moving average, or exponentially weighted moving average (EWMA), are preferred to

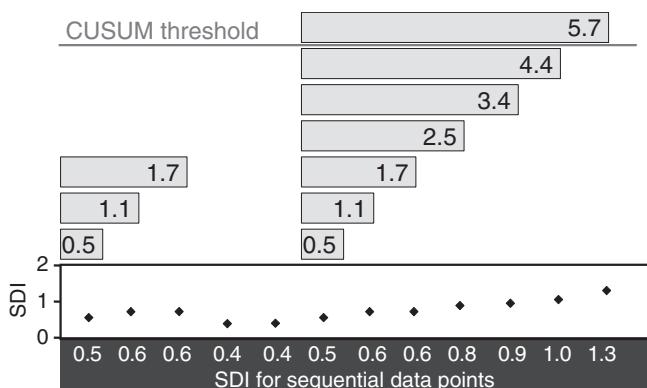
**TABLE 6.3 Abbreviation Nomenclature for Quality Control Evaluation Rules**

Rule	Meaning	Detects
$1_{2S}$	One observation exceeds 2 SDs from the target value. The $1_{2S}$ rule has a large false-alert rate and is not recommended except for low sigma measurement procedures.	Bias or imprecision (use with caution)
$1_{3S}$	One observation exceeds 3 SDs from the target value.	Bias or large imprecision
$2_{2.5S}$ ( $2_{2.5S}$ )	Two sequential observations, or observations for two QC samples measured at approximately the same time, exceed 2 SDs (or 2.5 SDs) from the target value in the same direction.	Bias
2of3 $2_{2S}$	Two observations for three QC samples measured at approximately the same time exceed 2 SDs from the target value in the same direction. Note that this type of rule is used when three QC materials are used for a measurement procedure.	Bias
$R_{4S}$	Range between observations for two QC samples measured at approximately the same time, or for two sequential observations of the same QC sample, exceeds 4 SDs.	Imprecision
$10_x$ or $10_m$	Ten sequential observations for the same QC sample are on the same side of the target value ( $x$ or mean). The $10_x$ rule is not recommended because it has a large false-alert rate.	Bias trend (not recommended)
$8_{1.5S}$ ( $8_{1.5S}$ )	Eight sequential observations for the same QC sample exceed 1 SD (or 1.5 SD) in the same direction from the target value.	Bias trend
CUSUM	Cumulative sum of SDI for a specified number of previous results.	Bias trend
MA	Moving average for a specified number of previous results.	Bias trend
EWMA	Exponentially weighted moving average with newer results having more influence (weight).	Bias trend

CUSUM, cumulative sum; EWMA, exponentially weighted moving average; MA, moving average, QC, quality control; SD, standard deviation; SDI, standard deviation interval.

Modified from Miller WG. Quality control. In: McPherson RA, Pincus MR. *Henry's clinical diagnosis and management by laboratory methods*.

24th ed. Philadelphia: Elsevier; 2020.



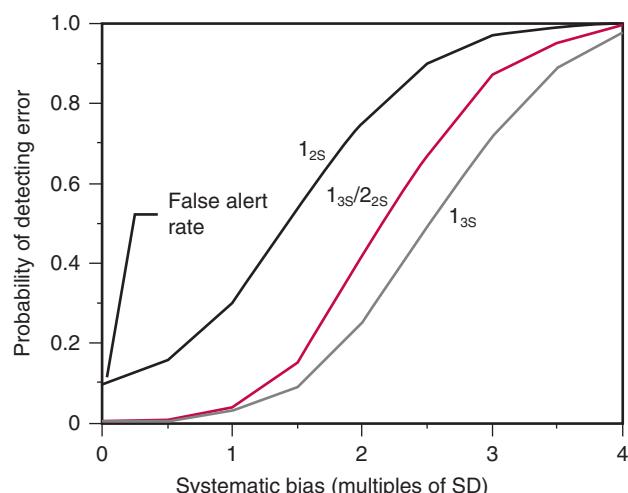
**FIGURE 6.7** Cumulative sum (CUSUM) process to identify trends in sequential results. The standard deviation interval (SDI) for a result vs. its target value to initiate a CUSUM is 0.45, and the threshold for an alert is 5.0.

monitor for bias trends.<sup>36</sup> One of these more advanced trend detection procedures is recommended if supported by an available computer system because they are more powerful for detecting trends than approaches based on a number of sequential observations exceeding a specified SD interval from the target value.

CUSUM expresses the difference between a QC result and its target value as an SDI, or z-score. For example, if the target value is 25.3 and the SD is 1.4, a QC result of 27.5 would have an SDI of  $(27.5 - 25.3)/1.4 = 1.6$ . Fig. 6.7 illustrates the CUSUM of SDI values for the most recent QC result and previous results for the same QC material since the last CUSUM reset. A minimum value for the SDI is used to initiate the cumulative summation to prevent relatively small increments from giving false alerts. If a QC value does not exceed the minimum SDI, the CUSUM is reset to zero. When the CUSUM exceeds a threshold value, an alert is given. The CUSUM alert may occur before a trend in bias causes the result for an individual QC value to be recognized as exceeding its QC evaluation rules. The threshold value for the CUSUM and the minimum value for the SDI to initiate the summation are set to provide a high probability to identify a trend in bias that may represent a defect in the measurement procedure that needs to be investigated. The threshold can be set to provide a warning that may not require immediate corrective action but rather an alert to a potential problem.

Moving average or EWMA operates similarly to CUSUM by taking the average of the most recent QC result and a specified number of previous results. For EWMA, a function in the calculation decreases the “weights” of each result in an exponential manner such that recent results contribute a greater proportion, and older results contribute a smaller proportion to the “average” value. The moving average or EWMA value represents a bias trend in the QC results, and a threshold is set that represents a defect in the measurement procedure that needs to be investigated. As for CUSUM, the threshold can be set to provide a warning that may not require immediate corrective action but rather an alert to a potential problem.

Power function graphs have been used to express the probability that a QC interpretive rule will detect an analytical error of a given magnitude.<sup>37</sup> Software to calculate power function graphs assumes Gaussian (normal) error distributions. Consequently, because there are influences on QC



**FIGURE 6.8** Power function graphs for the ability of different quality control interpretive rules to detect systematic error using two controls. Systematic error is expressed as number of standard deviations (SDs) from the target value. (Modified with permission from Westgard JO, Groth T. Power functions for statistical control rules. *Clin Chem* 1979;25:863–9.)

results that are non-Gaussian, the conclusions about QC rule performance from power function graphs are most useful as general guidance for selecting rules to interpret QC data. Other literature reports have addressed rule selection criteria using various statistical models and assumptions regarding distribution of errors.<sup>38–41</sup>

Power function graphs are useful to indicate relationships and relative effectiveness among different QC rules. Fig. 6.8 shows an example power function graph that plots the probability to detect a measurement error (y-axis), which is the probability that a result will exceed a particular interpretive rule, versus a systematic bias of known magnitude in a result (x-axis) with a fixed random imprecision of 1 SD when there are two QC samples being measured. The three lines in Fig. 6.8 represent the probabilities of different interpretive rules to detect biases of various magnitudes. For example, for the 1<sub>2s</sub> rule, a result with a bias of 1 SD (x-axis) has a 0.35 (35%) probability (y-axis) of violating the rule (i.e., of having a result >2 SDs from the target value). Note that this figure shows only bias as SD on the x-axis, and a result with 1 SD bias will also have an imprecision component that may cause the 1<sub>2s</sub> rule to be exceeded. A 1<sub>2s</sub> interpretive rule has approximately 35% probability to detect a systematic error that is 1 SD in magnitude and approximately 80% probability to detect a systematic error that is 2 SD in magnitude. Similar graphs can be created for other interpretive rules for both bias and imprecision error conditions.

Note in Fig. 6.8 that none of these interpretive rules has a 100% probability to detect a systematic bias until the error becomes relatively large. The 1<sub>2s</sub> rule has a good probability of detecting errors (e.g., almost 90% probability of detecting a 2.5-SD bias) but has a high false-alert rate as indicated by the y-intercept that indicates that because of imprecision, the probability of indicating an error condition for zero bias is approximately 10%. Because of this high false-alert rate, it is generally not recommended to use a 1<sub>2s</sub> rule unless the measurement procedure has marginal performance (i.e., is a “low sigma” measurement procedure) and the laboratory desires

to identify small biases that could cause inappropriate risk for a patient care decision. The  $1_{3S}$  rule has a low false-alert rate but a lower probability to detect an error (e.g., a 50% probability to detect a 2.5-SD bias).

Combining two or more rules and applying them simultaneously as multi-rule criteria is recommended to improve the efficiency of QC interpretive rules. For example in Fig. 6.8, the  $1_{3S}/2_{2S}$  multi-rule identifies an error condition if one control exceeds  $\pm 3$  SD from the target value or if two controls exceed  $\pm 2$  SDs in the same direction from the target value. The  $1_{3S}/2_{2S}$  multi-rule has a low false-alert rate similar to the  $1_{3S}$  rule but improved probability to detect an error (e.g., a 65% probability to detect a 2.5-SD bias and a 90% probability to detect a 3.2-SD bias). In this multi-rule example, the  $1_{3S}$  component is sensitive to imprecision or large bias, the  $2_{2S}$  component is sensitive to bias.

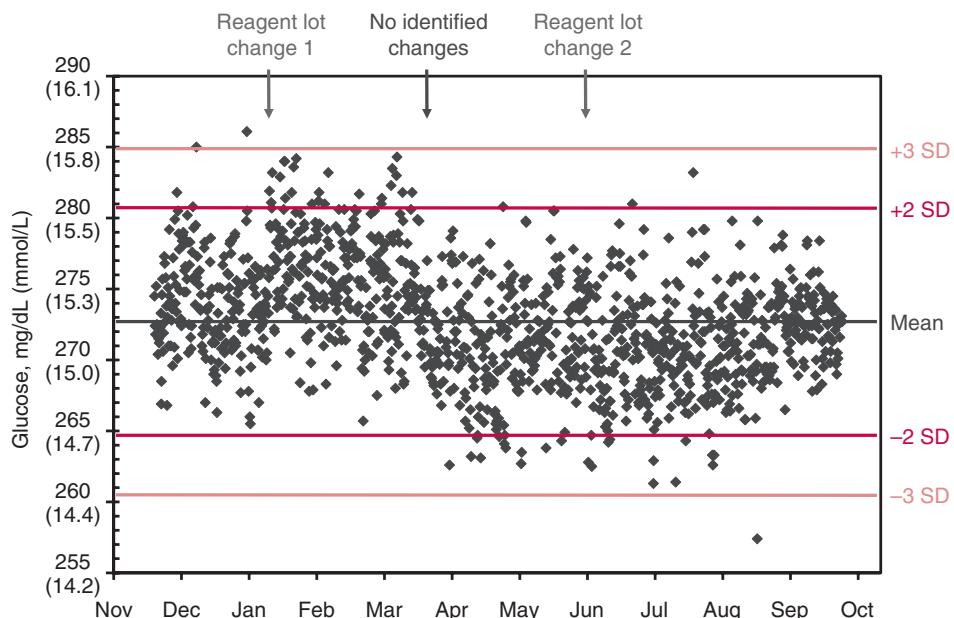
A challenge in selecting interpretive rules for evaluating QC results is that the different longer-term sources of variation listed in Table 6.2 occur at different times when using a measurement procedure. These types of variability are not adequately described by Gaussian models for rules selection. At certain periods of time, the short-term SD will be noticeably smaller than the long-term cumulative value (see Fig. 6.6). One must avoid concluding, based on a short-term estimate of SD, that the SD used for evaluation is too large because, over time, the cumulative value will be more consistent with measurement procedure performance as periodic sources of variability are encountered. Using an estimate of SD that is inappropriately small will lead to increased frequency of false alerts.

Fig. 6.9 shows how non-Gaussian error sources influenced results for a single lot of QC material used over a 10-month period for an automated glucose measurement procedure. The stability and performance over the 10 months were considered acceptable for clinical use. Data for the first 49 days

are the same as in Fig. 6.3 and represent the initial experience with this lot of QC material. 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 that was too small to initiate a change in target value. The second reagent lot change had no effect on QC results. Between March and April, a transition to lower values occurred that did not correspond to any maintenance, reagent lot change, or calibration events. Throughout the 10-month period, intervals of several weeks in duration occurred when the imprecision was better or worse than at other time periods (also see Fig. 6.6 calculated from the same data).

In practice, empirical judgment is frequently used to establish acceptance criteria (rules) to evaluate QC results based on data acquired over a long enough time to adequately estimate the expected variability when a measurement procedure is working correctly. QC rules should not be selected based only on Gaussian models of imprecision because the rules will not correctly accommodate all the types of variability observed for many analytical measuring systems.

Table 6.4 gives an example of an empirically developed multi-rule based on the data in Fig. 6.9. An empirical approach can be used by obtaining a set of QC data that represents a time interval expected to include most sources of variability. Using those QC data, the false-alert rate for a rule can be determined, and bias errors of different magnitudes can be added to estimate the ability of a rule, or a combination of rules, to identify that error. This multi-rule had 0.6% false alerts when applied to the data in Fig. 6.9 using the mean from the November to January (see Fig. 6.3) period as the target value and the cumulative SD for the 10-month period to represent overall imprecision. If a  $2_{2S}$  rule was used instead of a  $2_{2.5S}$  rule, the false-alert rate would increase by 1.2%, but the rule would detect slightly smaller biases. An  $8_{1.5S}$  rule was used to provide detection of bias



**FIGURE 6.9** Levey-Jennings plot of quality control (QC) results ( $n = 1232$ ) for a single lot of QC material used over a 10-month period. SD, Standard deviation. (Adapted with permission from Miller WG, Nichols JH. Quality control. In: Clarke WA, editor. *Contemporary practice in clinical chemistry*. 2nd ed. Washington, DC: AACC Press; 2010.)

**TABLE 6.4 Empirical Multi-Rule for the Quality Control Data Presented in Fig. 6.11**

Multi-Rule Components	Type of Variability Detected
$1_{3S}$	Imprecision or bias
$2_{2.5S}$	Bias
$R_{4S}$	Imprecision
$8_{1.5S}$	Bias trend

QC, Quality control.

From Miller WG. Quality control. In: McPherson RA, Pincus MR. *Henry's clinical diagnosis and management by laboratory methods*. 24th ed. Philadelphia: Elsevier; 2020.

trends because it had a 0% false-alert rate (compared with 0.5% for an  $8_{1S}$  rule) and was adequate to detect a developing trend before it became clinically important because the CV was small, 1.5%, at the concentration of the QC material. If a 10% TE<sub>a</sub> is considered acceptable for glucose at this concentration, then the sigma metric for this measurement procedure is 6.7, suggesting that it has a very low error rate, and these QC rules with a low false-alert rate will be suitable to alert the laboratory to an error condition large enough to affect patient care decisions. Such control rules should allow the laboratory to detect errors before they are of a magnitude that will affect clinical actions. At other clinical concentration ranges or for other analytes, a different set of QC evaluations rules may be more appropriate. A  $10_x$  rule was not used because it would have increased the false-alert rate by 10.6%. A  $10_x$  rule or other rule that counts the number of sequential QC results on one side of the target value is not recommended because this condition does not indicate a problem with clinical interpretation of patient results when the magnitude of the difference from the target value is small. Counting the number of sequential results that exceed a larger SD from the target value, such as  $8_{1.5S}$  in this example, is more likely to represent a measurement condition that might need investigation.

The balance between false alerts and the probability of detecting an error is improved when multiple rules are used in combination. When establishing rules to interpret QC results, it is important to remember that statistical process control can only verify at a point in time that a measurement procedure is producing results that conform to the expected variability when the procedure is performing in a stable operating condition. It is important to remember that periodic measurement of QC samples does not identify random events (e.g., a temporary clot in a sample pipet, a random reagent pipetting error) that do not persist until the next QC sample is measured. QC rules are chosen to detect changes in calibration and changes in imprecision that persist until the next QC measurement and are significant enough to require correction before patient results are reported.

### Poorly Performing Measurement Procedures

In the process of reviewing statistical parameters for QC data, a measurement procedure's performance may be identified as marginal or inadequate to meet medical requirements. Determining the sigma metric is helpful to make this assessment because the sigma value gives a prediction of the number of erroneous results expected. If the measurement procedure performance cannot be improved and a better measurement procedure is not available, the laboratory can either discontinue

the test if the performance is inadequate or apply more stringent QC practices to identify small deviations from the expected performance. More stringent QC practices include selecting rules that will give an alert at smaller error conditions, using additional rules in a multi-rule set, measuring QC more frequently, using more than two QC samples, and not releasing patient results until QC assessment is complete for the time interval during which patient samples were measured. Lower sigma measurement procedures usually require more frequent measurement of QC samples, more stringent criteria for accepting QC results, as well as applying patient-based QC monitoring when possible.

More stringent QC rules will not improve measurement procedure performance but will identify smaller changes in measurement procedure performance that could affect patient care decisions based on the results. More stringent QC rules will have more false alerts, but this is an unavoidable cost when lower sigma measurement procedures are used. Because the analytical requirements are not easy to establish and are themselves somewhat uncertain, one should regularly reassess the requirements to see if they remain appropriate or should be updated, and should reconsider that the QC rules are appropriate to identify an error condition. In addition, the measurement procedure limitations should be communicated to patient care providers.

### Specifying the Quality Control Plan

The preceding subsections describe the considerations for each component in a QC plan. The laboratory director is responsible for considering the components, making judgments regarding the considerations, and approving the final plan for each analyte measured in a laboratory. A plan for internal QC using surrogate samples specifies the following components:

- The number of controls to be measured and the approximate concentrations of analytes in those controls
- The target values for each control
- The SD to be used in the QC rules
- The rules for evaluating the QC results
- The frequency to test the QC samples

Overall, the choice of the parameters in the QC plan depends on the performance characteristics of the measurement procedure, the number of potentially erroneous patient results that could occur before the error condition is identified, and the risk of harm to a patient if potentially erroneous results were used in medical care decisions. Yundt-Pacheco and Parvin described a methodology using a Gaussian statistical model to compute the expected number of unreliable patient results produced based on an out-of-control condition and the performance characteristics of a measurement procedure.<sup>41</sup>

### Considerations for Point-of-Care Testing

Internal QC of POC instruments offers extra challenges compared with those addressed at the central laboratory. The main reasons for this are that POC instruments are often operated by persons without laboratory training; they often use methodologic principles that are different from those in the central laboratory; they often have "built-in" controls; and the number of measurements is rather small, making the use of QC samples in the traditional way expensive. Because the instruments are used by nonlaboratory personnel who also should run the QC program, these people have to be

convinced that measuring QC samples is useful and will detect errors important for patient safety. Unfortunately, the evidence for measuring QC samples is scarce, probably because there is little agreement on how to implement QC for POC instruments and how to handle the QC alerts. The ISO 22870 document<sup>42</sup> states that if an institution wants to be accredited to this standard, internal and external QC of POC instruments should be done, but it is not stated how this should be done. Recommendations from different countries generally state that it is “mandatory” to measure QC samples, but they are usually vague concerning how it should be done, from analyzing two levels of QC materials a day to one level of QC material every sixth month or as “recommended by the manufacturer” or “recommended to use control material independent from the manufacturer” and use specifications as defined by the local laboratory.<sup>43–45</sup>

It is not surprising that there are no uniform or concrete recommendations because POC instruments use different methodologies and technologies, and they are used at different locations, from wards at a hospital to remote areas. Before establishing an internal QC program for POC instruments, at least three issues should be taken into consideration: (1) the type of POC instrument and what “built-in” control processes it has, (2) the location of the instrument, and (3) the operator of the instrument. Broadly, the current instruments have been divided into three categories: (1) instruments that are similar to the wet chemistry instruments used in the central laboratory, (2) cartridge-based instruments, and (3) strip-based instruments,<sup>46</sup> although there is a significant “overlap” between the two last categories. The instruments using wet chemistry and similar technologies to those used in the central laboratory should follow the principles for internal QC as outlined in this chapter albeit taking into consideration the number of measurements per day. In cartridge-based and some strip-based instruments, the manufacturer often has placed the technology in the cartridge or strip together with QC materials, and in some cases, QC rules are built in so that patient results cannot be reported unless the QC is “satisfactory.”<sup>47</sup> The instrument is then merely an electronic reader that often has incorporated an “electronic quality control” where the electronics of the measurement procedure are verified. The electronic instrument checks do not verify the reagents in the cartridges or strips, and unless each cartridge has internal QC materials, the reagent cartridges or strips should be checked at delivery and then at intervals (e.g., with the arrival of a new shipment or lot or at a suitable interval such as monthly or based on the number of patient measurements in a time interval).

Not all POC instruments include enhanced QC features. In these cases, one must rely on daily liquid surrogate QC performed by the operator.<sup>48</sup> The limitation of using the liquid QC sample in this situation is that it only checks if one disposable cartridge or strip meets the performance specifications. This limitation requires an assumption that all devices in a lot were manufactured uniformly and will perform equivalently. Some tests, such as dipstick urinalysis, require liquid QC because there are no built-in control mechanisms. Others, such as urine pregnancy tests, have a built-in positive control band to ensure that the device is functioning properly, but this may not be an adequate substitute for a traditional QC sample that can assess suitable recovery of concentrations. Some POC protocols run patients’ samples on a POC instrument and send the same samples to a central laboratory

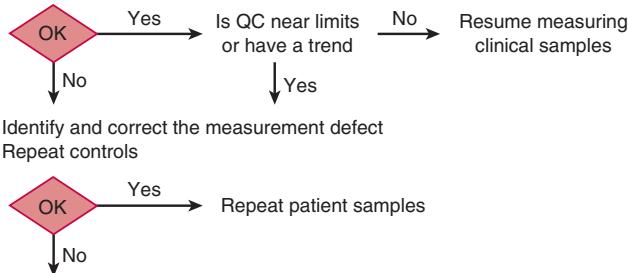
as a form of internal QC. It has been shown for international normalized ratio that this periodic split patients’ samples procedure cannot be recommended because it had a lower probability of error detection and a higher rate of false alarms compared with using commercial lyophilized QC material.<sup>48</sup> The operator of the instrument is also important. In patient self-testing, it is difficult to implement internal QC procedures other than what is built into the instruments or strips or cartridges.

How internal QC should be performed and supervised also depends on the location of the POC instruments. In a hospital, it is now possible with real-time bidirectional connectivity between the POC devices and the central laboratory to transfer both patient and QC results and to set lockout parameters for conformance to a QC protocol. As technology advances, the general trend is for more sophisticated POC devices with built-in control systems to be incorporated to minimize or prevent the possibility for an incorrect result.<sup>49</sup>

### Corrective Action When a Quality Control Result Indicates a Measurement Problem

A QC alert occurs when a QC result fails an evaluation rule, which indicates that an analytical problem may exist. A QC alert means there is a high probability that the measurement procedure is producing results that have the potential to be unreliable for patient care. Some types of QC rules, for example, trend rules, can be established as warnings intended to initiate an investigation but not require immediate discontinuing reporting of results. Other types of QC rules are intended to indicate error conditions for which discontinuing reporting patients’ results is indicated. Fig. 6.10 presents a generalized troubleshooting sequence. Remeasuring the same QC sample is not recommended because, with properly designed control rules, it is more likely that a measurement procedure problem exists than the QC result was a statistical outlier or gave a false alert. However, QC materials can deteriorate after opening because of improper handling and storage or because of unstable measurands. Thus repeating the measurement on a new vial of the QC material is a useful step to determine if the alert was caused by deteriorated QC material rather than by a measurement procedure problem. In this situation, if the result for the new QC sample is acceptable, testing of patient samples can resume. One caution when the repeat QC result is near acceptability limits is to consider

1. Stop reporting results
2. Measure a new container of control



**FIGURE 6.10** Generalized troubleshooting sequence showing the initial steps after an unacceptable quality control (QC) result. The details of troubleshooting the defect may be different for different rules violations or if more than one unacceptable QC result was obtained.

whether the repeat and original results are essentially the same. In this situation, the probability is fairly high that a measurement problem exists, and this possibility should be investigated. In addition, current and preceding QC results should be examined for a trend in bias that indicates a measurement issue that needs to be corrected. These precautions in evaluating repeat results for a new QC sample can be challenging or impossible for automated evaluation by computer systems, thus requiring the laboratory technologist to be vigilant in reviewing results.

When repeat testing of a new QC sample does not resolve the alert situation, the instrument and reagents should be inspected for component deterioration, empty reagent containers, mechanical problems, and so on. In many cases, recalibration is indicated.

When the problem is identified and corrected, QC samples should be measured to verify the correction, and all patient samples since the time of the last acceptable QC results, or the time when the error condition occurred, should be measured again. It may be difficult to establish the time when an error condition occurred but repeating selected patient results may be considered. For example, every few samples can be repeated back to the time of the last acceptable QC results. The repeated results are then compared with acceptable criteria for repeated results agreement (see next paragraph) to identify a point in time when the error condition occurred. When selecting the samples to repeat, it is important to ensure that a substantial representation of the potentially erroneous samples is repeated and that samples at a concentration consistent with that of the unacceptable QC are represented. Alternatively, groups of 10 patient samples can be repeated, again ensuring that samples at a concentration consistent with that of the unacceptable QC are represented, until all repeat results in at least two sequential groups are within acceptable criteria for repeated results. When the point at which the error condition was likely to have occurred is identified, all patient samples must be repeated from that point until the unacceptable QC result was obtained. Any assessment of the point at which an error condition occurred by repeating selected patient samples has a risk to incorrectly identify that point and laboratories must repeat enough patient samples to have confidence in the assessment.

The laboratory director must establish acceptable criteria to determine if the repeat results agree adequately to permit reporting of original results without issuing a corrected report. Otherwise, corrected results must be reported. As an example, Table 6.5 lists empirical criteria used in one author's laboratory for this purpose. The criteria for acceptability of repeated tests are based on measurement procedure performance characteristics and the intended medical use of the results. The considerations for determining the TE<sub>a</sub> described in the section Performance of a Measurement Procedure for its Intended Medical Use and in Chapter 8 related to the reference change value may be helpful to set acceptance criteria for repeated patient sample results.

In some cases, residual sample volume may not be adequate for repeat testing (quantity not sufficient [QNS]). In these situations, no results can be reported unless it is documented that the impact of the measurement procedure defect on the original results was small enough to have minimal effect on clinical interpretation. A protocol to evaluate the clinical impact of the measurement defect involves repeating

**TABLE 6.5 Example for Selected Chemistry Analytes of Empirical Criteria for Patient Test Result Agreement Between Repeated Assays and for Agreement Among Results for a Single Patient Sample Measured on Multiple Instruments**

Analyte	Acceptance Criteria (Difference Between Results)
Albumin	0.4 g/dL (4.0 g/L)
ALP	10 U/L or 10% <sup>a</sup>
ALT	10 U/L or 10% <sup>a</sup>
Amylase	15 U/L or 10% <sup>a</sup>
AST	10 U/L or 10% <sup>a</sup>
Bilirubin, total	0.3 mg/dL (5 µmol/L) or 10% <sup>a</sup>
Calcium, total	0.5 mg/dL (0.125 mmol/L)
Chloride	4 mmol/L
Cholesterol	5%
CK	10 U/L or 10% <sup>a</sup>
CO <sub>2</sub>	4 mmol/L
Creatinine	0.2 mg/dL (18 µmol/L) or 10% <sup>a</sup>
GGT	10 U/L or 10% <sup>a</sup>
Glucose	6 mg/dL (0.33 mmol/L) or 5% <sup>a</sup>
Iron	10 µg/dL (1.8 µmol/L) or 10% <sup>a</sup>
Lactate	0.32 mmol/L
LDH	10 U/L or 10% <sup>a</sup>
Lipase	10 U/L or 10% <sup>a</sup>
Magnesium	0.3 mg/dL (0.1 mmol/L)
Phosphorus	0.4 mg/dL (0.13 mmol/L)
Potassium	0.3 mmol/L
Protein, total	0.4 g/dL (4.0 g/L)
Sodium	4 mmol/L
Triglycerides	10%
Urea nitrogen (BUN)	3 mg/dL (1.1 mmol/L urea) or 10% <sup>a</sup>
Uric acid	0.4 mg/dL (24 µmol/L)

<sup>a</sup>Whichever is greater.

ALP, Alkaline phosphatase; ALT, alanine aminotransferase; AST, aspartate aminotransferase; BUN, blood urea nitrogen; CK, creatine kinase; CO<sub>2</sub>, carbon dioxide; GGT, γ-glutamyl transferase; LDH, lactate dehydrogenase.

From Miller WG. Quality control. In: McPherson RA, Pincus MR. *Henry's clinical diagnosis and management by laboratory methods*. 24th ed. Philadelphia: Elsevier; 2020.

those samples that have adequate volume. The repeated samples must represent the concentration range of the QNS samples and the time span since the previous acceptable QC results and should include a substantial proportion of the total samples originally assayed while the measurement procedure was in the unacceptable condition. If the repeat results for this subgroup are within established "acceptable" criteria for repeat testing of patient samples, the original results for the QNS samples can be reported. Otherwise, the original results for the QNS samples are considered erroneous; no results can be reported, and any original results already reported need to be corrected to a "no result" condition.

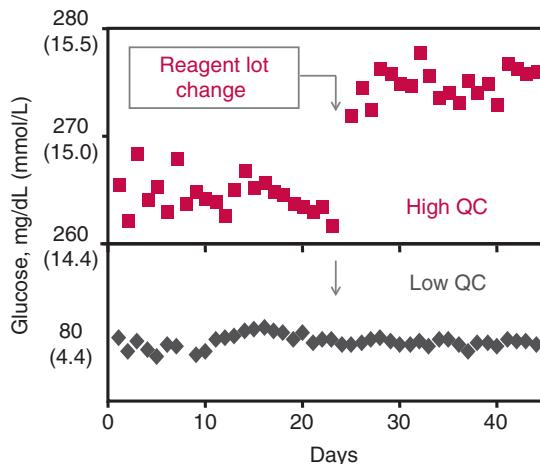
Alternatively, for QNS samples, when there was not an outright malfunction but rather a drift or shift in calibration, it may be possible to estimate the magnitude of a bias error from the results for other patient samples that were repeated, apply that bias as a correction to the QNS samples, and report

the corrected result with a comment regarding the increased uncertainty in the values. When there are inadequate data to estimate a correction factor from repeated results for other patients, the laboratory may consider reporting the original result for a QNS sample with a comment regarding its uncertainty. This approach can be useful especially in cases when it might be difficult or take a long time to obtain a new sample, or in cases when the clinical question can be answered by whether the result is very high or very low (e.g., hypo- or hyperthyroidism). The laboratory director should be consulted for guidance in reporting results for QNS samples that may have greater uncertainty than the usual quality performance for a laboratory.

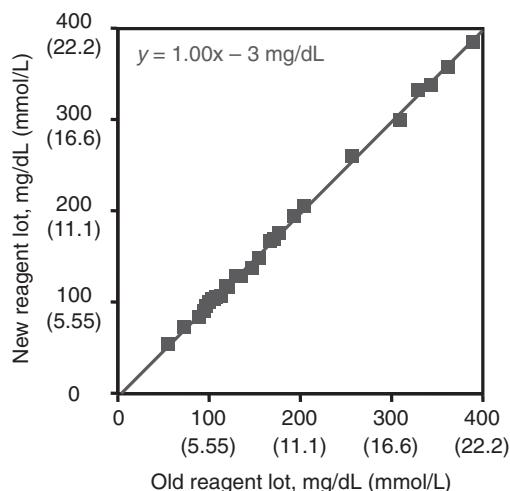
### Verifying Quality Control Evaluation Parameters

#### After a Reagent Lot Change

Changing reagent lots can have an unexpected impact on QC results. Careful reagent lot crossover evaluation of QC target values is necessary. Because the matrix-related interaction between a QC material and a reagent can change with a different reagent lot, QC results may not be a reliable indicator of a measurement procedure's performance for patient samples after a reagent lot change. In a large study of 661 reagent lot changes for eight QC materials measured for 82 analytes using seven different instrument platforms, 41% of 1483 QC material-reagent lot combinations had significant differences in QC values between the reagent lots that were not observed for patient samples.<sup>28</sup> In the example in Fig. 6.11, QC values for the high-concentration control shifted after the change to a new lot of reagents, but there was no change in results for the low control. A comparison of results for a panel of patient samples measured using the new and old reagent lots, as shown in Fig. 6.12, verified that patient results were the same when either lot of reagents was used. Patient results spanning the measuring interval had nearly identical values, as indicated by the slope of 1.00 and the small intercept of  $-3 \text{ mg/dL}$  ( $0.17 \text{ mmol/L}$ ). Consequently, the change in values for the high-concentration QC material was due to a difference in matrix-related bias between the QC material and each of the reagent lots.



**FIGURE 6.11** Levey-Jennings plot showing impact of a reagent lot change on matrix bias with quality control (QC) samples. (Modified with permission from Miller WG, Nichols JH. Quality control. In: Clarke WA, editor. *Contemporary practice in clinical chemistry*. 2nd ed. Washington, DC: AACC Press; 2010.)

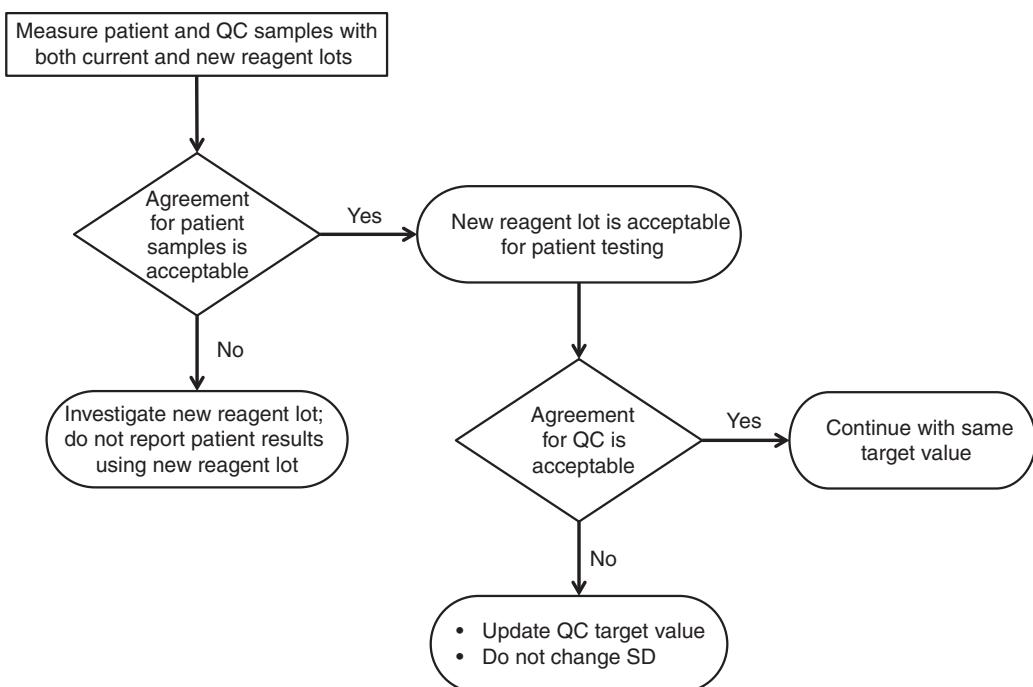


**FIGURE 6.12** Deming regression analysis of results from a patient sample comparison between the same old and new lots of reagent shown in Fig. 6.11 for quality control samples. (From Miller WG. Quality control. In: McPherson RA, Pincus MR. *Henry's Clinical diagnosis and management by laboratory methods*. 24th ed. Philadelphia: Elsevier; 2020.)

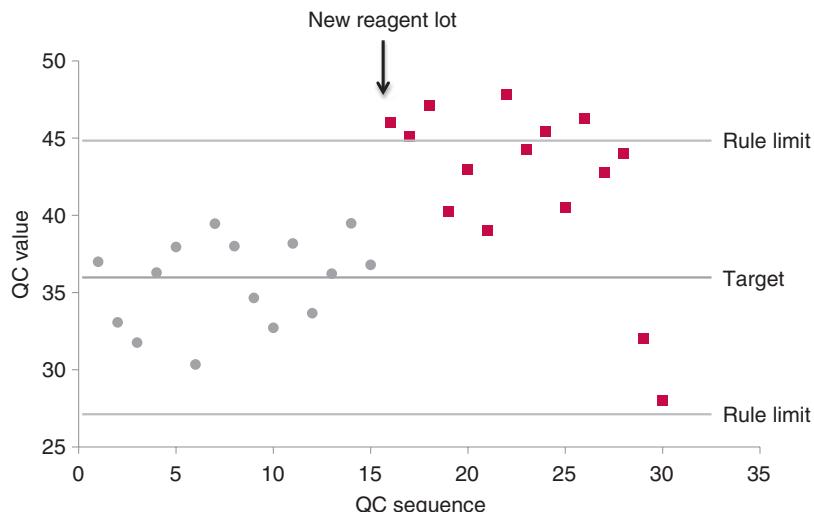
Clinical patient samples must be used to verify the consistency of results between old and new lots of reagents because of the unpredictability of a matrix-related bias being present for QC materials. Fig. 6.13 presents a protocol to verify or adjust QC material target values after a reagent lot change. A group of patient samples and the QC samples are measured using both the current (old) and new reagent lots. The first step is to verify that results for a group of patient samples measured with the new reagent lot are consistent with results from the current (old) lot. The patient sample results, not the QC results, provide the basis for verifying that the new reagent lot is acceptable for use. If a problem is identified, the calibration of the new reagent lot must be investigated and corrected, or the new reagent lot may be defective and should not be used. When evaluating the patient results, keep in mind that the calibration of the old reagent lot may have drifted and should be verified before concluding that the new reagent lot is not giving acceptable results for the patient samples.

The number of patient samples to use for verifying the performance of a new reagent lot will depend on the measuring interval, the imprecision of a measurement procedure, and the concentrations at which clinical decisions are made. CLSI document EP26<sup>50</sup> recommends a minimum of three patient samples and more patient samples depending on the number of important clinical decision concentrations and the imprecision of a measurement procedure. This CLSI guideline includes a statistical analysis to determine if a difference in patient results is less than a critical difference that would represent risk for an inappropriate patient care decision based on a particular laboratory test result. An alternate approach is to select 10 or more patient samples that span the measuring interval and use orthogonal regression analysis or a difference plot to evaluate average performance over the interval of concentrations represented by the patient samples.

There are no well-established clinical acceptance criteria for agreement between results; consequently, the laboratory must establish acceptance criteria consistent with the



**FIGURE 6.13** Process for assessment of potential matrix impact on quality control (QC) samples after a reagent lot change. *SD*, Standard deviation. (From Miller WG. Quality control. In: McPherson RA, Pincus MR. *Henry's clinical diagnosis and management by laboratory methods*. 24th ed. Philadelphia: Elsevier; 2020.)



**FIGURE 6.14** Illustration of the influence on the failure rate for a quality control (QC) rule when failing to adjust the target value for a matrix-related shift. QC results before a new reagent lot are shown as gray circles and after the new reagent lot as red squares. *SD*, Standard deviation. (From Miller WG. Quality control. In: McPherson RA, Pincus MR. *Henry's clinical diagnosis and management by laboratory methods*. 24th ed. Philadelphia: Elsevier; 2020.)

relatively small number of samples used, the analytical performance characteristics of a measurement procedure, and the intended medical use for interpreting results. As an example, empirical acceptance criteria used in one author's laboratory for assessment of individual results are provided in Table 6.5.

When the results for patients' samples are acceptable, the second step in Fig. 6.13 evaluates results for each QC material to determine if its target value is correct for use with the new

lot of reagent(s). If the target value has changed, it must be adjusted to correct for the change in matrix-related bias between old and new lots of reagent(s). This adjustment keeps the expected variability centered around the QC target value so that QC interpretive rules will remain valid. Failure to make a target value adjustment will introduce an artifactual bias in subsequent QC results, causing both an increased false-alert rate and a decreased ability to detect some error conditions. These effects are illustrated in Fig. 6.14 where the

shift in target value would cause some of the results shown by red squares to exceed the old upper QC rule limit when in reality there is no defect in patient results because the QC results are artificially increased because of the matrix-related bias with the new reagent lot. Similarly, the increased magnitude of the gap to the old lower QC rule limit will permit a low bias condition, as shown by the square red points at sequence number 29 and 30, to be undetected until the low bias becomes larger than the QC rule is intended to detect.

The SD used to evaluate QC results will not typically change when a new lot of reagent(s) is put into use. The SD represents expected variability when the measurement procedure is stable and is performing according to specifications. In most cases, the variability of a measurement procedure will be the same with any lot of reagent(s). However, occasional exceptions may occur; for example, if the new reagent lot is a reformulation, it may be necessary to adjust the SD after additional numbers of QC results are accumulated with the new reagent lot. A reagent lot verification is typically performed on a single day and will likely provide only a few QC results from which to evaluate if the target value has changed. Consequently, it is necessary to carefully monitor QC results as more data are acquired using the new reagent lot and, if needed, to further refine the new target value.

Note that a reagent lot induced matrix-related change in the numeric values for the QC results will cause an artificial increase in the cumulative SD, thus making it larger than the inherent measurement variability and not suitable for use in QC rules. For this reason, it is recommended to use the cumulative SD from a single reagent lot or the pooled SD from more than one reagent lot (see subsection called Quality Control Material Standard Deviation) when determining the SD to use for interpreting QC rules.

Experience in clinical laboratories has shown that there are changes, other than reagent lot changes, in measurement procedures that can also affect the QC values but not the results for patient samples. Such changes could be caused by instrument component replacement or other causes. In theory, there should be an assignable cause for such effects, but such a cause is not always identifiable. In practice, any condition that affects QC results but does not affect patient results is treated in the same manner as described for reagent lot changes.<sup>32</sup> The important QC principle is that if the results for patient samples are consistent between the two conditions, then the target value for the QC sample should be adjusted, if necessary, to reflect its value under the new condition. Failure to adjust the QC target value will cause inappropriate acceptability criteria to be used for evaluating the QC results and incorrect QC rules evaluation with increased false alerts and some potential error conditions missed (see Fig. 6.14).

### Verifying Measurement Procedure Performance After Use of a New Lot of Calibrator

When a new lot of calibrator is used, with no change in reagents, there is no change in matrix interaction between the QC material and the reagents. In this situation, QC results provide a reliable indication of calibration status with the new lot of calibrator. If the QC results indicate a bias after use of a new lot of calibrator, the calibration has changed and needs to be corrected to ensure consistent results for patient samples.

Some measurement procedures are packaged as kits that include reagents, calibrators, and QC materials. In this case, QC results could fail to identify a calibration shift when a new kit lot is used, and it is necessary to measure patient samples with the old and new kit lots to verify consistency of patient results. When possible, it is recommended to use QC materials that are independent of the kit lot and to avoid changing lots of QC material at the same time as changing lots of reagent or calibrators. Measuring patient samples always provides a reliable approach to verify the consistency of results after changes in lots of reagents or calibrators and changes in other measurement conditions.

### Review of Quality Control Data and the Effectiveness of the Quality Control Plan

The immediate use of QC data is to determine if the results for patient samples can be reported for use in clinical care decisions as described in the preceding sections. In addition, QC data must be reviewed by laboratory management on a regular schedule. Typical review schedules are weekly by senior technologists or supervisors and at least monthly by the laboratory director. However, the laboratory director should promptly review items such as reagent or calibrator lot change validations, changes in QC target values associated with reagent lot or other changes, EQA/PT results review, and other occurrences that may affect quality of the laboratory results.

The weekly review process should determine that correct follow up of any QC alerts was conducted, that all patient samples that may have had erroneous results were repeated, that any corrected reports were issued, and that the process was properly documented in QC records. The monthly review should include any issues identified by the weekly review process and examination of the Levey-Jennings chart or other tool to identify trends or changes in assay performance that may need to be addressed before they have effects on clinical care decisions. Note that automated systems to assist in the review of QC data are acceptable, and individual Levey-Jennings charts do not need to be examined every month. A report that compares the mean and SD for QC results over a defined time interval, such as 1 month, to the expected values consistent with stable performance can be useful to focus the review on measurement procedures that may need attention. For example, the report might identify a QC mean value that is more than a specified amount, such as 1 SD, from its target value, an SD that exceeds its expected value, or the number of individual results that exceed 2 or 3 SDs from the target value. QC values that are identified as needing further examination can then be followed up with review of a Levey-Jennings chart or other records of measurement procedure performance such as maintenance, calibration, and reagent change. The monthly review should also include any patient data-based QC procedures described in the following sections, as well as notation of any adjustments made to QC parameters during the month.

The QC review process serves three major functions, which are to (1) verify that the measurement procedures are stable and meeting their performance specifications, (2) identify measurement procedures that may need intervention to address performance issues, and (3) make adjustments as needed to the QC plan based on review of relevant quality indicators. Quality indicators and implementation of

**TABLE 6.6 Examples of Quality Indicators for the Examination Process**

Quality Indicator	Interpretation
Frequency of QC alerts	Compare with the frequency expected for the measurement procedure sigma metric and QC rules used. A higher frequency may indicate an issue with the measurement procedure or inappropriate QC rules. A lower frequency may indicate inappropriate QC rules.
Frequency of recalibration based on QC alerts	May indicate that recalibration should be performed more frequently.
Number of reagent changes due to QC alerts	May indicate that reagents are not stable and smaller quantities should be used or perhaps more frequent recalibration should be performed to compensate for reagent changes.
Number of times controls were repeated without confirming a measurement error	May indicate that the QC rules allow too high frequency of false alerts or the QC material is not stable after opening, is stored incorrectly, or other QC handling issue.
Frequency of unscheduled maintenance due to QC alerts	May indicate that maintenance is needed on a more frequent schedule.
Number of patient samples repeated based on QC alerts	May indicate that QC should be performed more frequently to minimize the risk of an erroneous result causing harm to a patient.
Frequency that patient samples are repeated based on QC alerts	May indicate inadequate calibration or maintenance schedules, that QC rules are inappropriate, or the QC sample target value or SD is not a correct reflection of measurement procedure performance.
Number of patient results corrected	May indicate an unstable measurement procedure and that QC should be performed more frequently to minimize the risk of an erroneous result causing harm to a patient.
Number of EQA/PT unacceptable results	May indicate that a measurement procedure is not calibrated correctly or some part of the measurement is not being performed correctly.

EQA, External quality assessment; PT, proficiency testing; QC, quality control; SD, standard deviation.

the laboratory quality management program are described in Chapter 3. **Table 6.6** lists some useful quality indicators related to the examination process and its QC plan. The quality indicators should be reviewed at regular intervals as part of the overall quality management program and can also be reviewed at suitable intervals during regularly scheduled QC review meetings to determine if changes in the QC plan may be needed.

### POINTS TO REMEMBER

#### Internal Quality Control/Statistical Process Control

- QC samples are measured along with patient samples.
- The target value and SD expected for a QC sample are established by the laboratory.
- Results from QC samples are evaluated using interpretive rules that are established after considering the probability of detecting errors that represent a risk of harm to a patient and the probability of false alerts.
- The QC plan is designed to confirm acceptable performance of a measurement procedure and to identify error conditions that may cause risk of harm to a patient.

### USING PATIENT DATA IN QUALITY CONTROL PROCEDURES

In addition to using patient samples to verify consistency of patient results when changing lots of reagent or calibrators for a measurement procedure discussed previously, patient data are used in other QC applications. A delta check process compares current with previous results for a patient to identify inconsistencies that may represent a pre-examination or measurement error. Comparison of patient results among

different measurement procedures used in a healthcare system for the same measurand is used to ensure that calibration of the different measurement procedures produces consistent results. There is increasing interest in using patient results to monitor the performance of a measurement procedure in real time as a supplement to the surrogate QC sample approach. Each of these patient data-based techniques is described in more detail in the next sections.

#### 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 against a previous result for the same measurand. This comparison is called a "delta check." If there is a difference between the two results exceeding the delta check value, this difference can be caused either by a pre-examination error, a laboratory error, or a change in the patient's physiologic condition. Delta check values can be developed in three ways. The first approach is to set delta check values based empirically on experience and then adjust them with time so as not to generate too many false delta check failures. The second involves collecting large numbers of consecutive pairs of patient results for the same measurand from a population that is representative of the patient population to which the delta check values will be applied. The population of delta values, which are actually empirically obtained reference change values (see Chapter 8), are plotted in a frequency distribution histogram. Delta check values are then identified to flag a certain percentage, for example, 5% or 1%, of the population of observed delta values. The third approach is to calculate the reference change value based on analytical and within-subject biologic variation. Reference change values can then be used to flag reports to alert users to those serial results in an individual where, for example, there is less than 1% probability that the change can be explained by the combined

**TABLE 6.7 Example Delta Check Criteria Intended to Identify Samples That Are Potentially Mislabeled or Contaminated With Intravenous Fluids Because of Incorrect Collection**

Test	Delta Criteria
Sodium	5% change within previous 48 h
Urea nitrogen	60% change within previous 48 h
Creatinine	50% change within previous 48 h
Calcium	25% change within previous 48 h
Osmolality	5% change within previous 48 h

From Miller WG. Quality control. In: McPherson RA, Pincus MR. *Henry's clinical diagnosis and management by laboratory methods*. 24th ed. Philadelphia: Elsevier; 2020.

influence of pre-examination error, examination (or analytical) error, and biologic variation.

The most common use of delta checks is to identify mislabeled samples and samples altered by dilution with intravenous fluid. The difference between results that causes a delta check alert must be sufficiently large to avoid excessive numbers of false alerts yet adequate to allow identification of samples that may be compromised and require follow-up investigation. Table 6.7 shows, as an example, the delta check parameters for automated chemistry used in one author's laboratory designed to identify compromised patient samples. A relatively small number of measurands are sufficient to identify potentially compromised specimens. The delta criteria were based on assessment of the delta values for consecutive pairs of results from the same patient that had differences larger than the 99 percentile of all the differences found, that is, an alert rate less than 1%.

Delta checks can detect analytical measurement errors; however, the threshold values necessary to identify analytical errors are usually fairly small compared with physiologic changes and may cause a large number of false alerts that reduce the efficiency of laboratory workflow. A well-designed statistical QC plan is more effective to detect analytical errors. However, delta checks might be useful to identify an interfering substance (e.g., from a drug) that may appear in a patient's sample. Kazmierczak<sup>51</sup> has reviewed and presented recommendations for using delta check and other patient data-based QC procedures. CLSI has published guideline EP33 for using delta checks in the clinical laboratory.<sup>52</sup>

### Verify Consistency of Results Between More Than One Instrument or Measurement Procedure

Another common use of patient results as part of the QC process is to verify the consistency of patient results when a measurand is measured using more than one instrument or measurement procedure within the same health delivery system. Verification of consistent results is necessary even when identical measurement procedures from the same manufacturer are used. Good laboratory practice requires that multiple instruments or measurement procedures for the same measurand be calibrated to produce the same results for patient samples whenever possible. It may be necessary to modify the calibration settings of one measurement procedure so that results for patient samples will be equivalent to

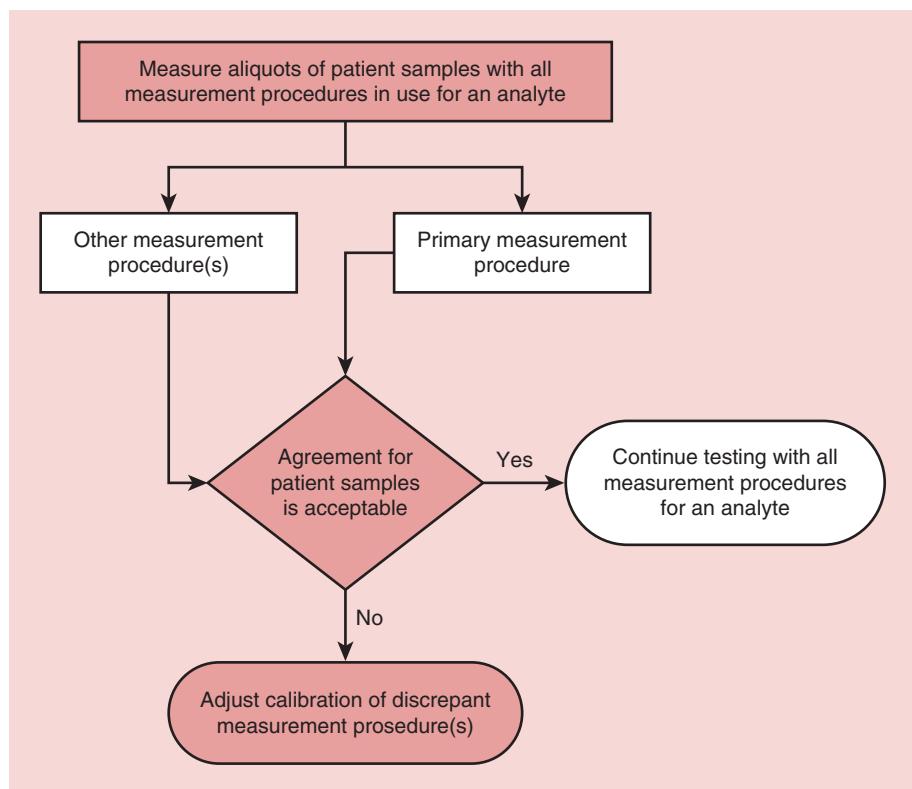
those for another measurement procedure. This strategy allows a common reference interval to be used, provides continuity in results among different laboratory testing locations, and avoids clinical confusion regarding interpretation of laboratory results.

As illustrated in Fig. 6.15, consistency of patient results is verified by measuring aliquots of patient samples using each of two or more measurement procedures to evaluate and, if necessary, adjust the calibration or use a postmeasurement correction function as needed to achieve agreement in results for patient samples. Such an analysis design is called a "round robin." One procedure may be chosen to represent the primary measurement procedure (or designated comparison measurement procedure) to which others will be adjusted to achieve equivalent results. The primary measurement procedure should be chosen based on quality and reliability of results with consideration of its calibration traceability to a national or international reference system, its performance stability, its analytical selectivity for the analyte, and its susceptibility to interfering substances. An alternate approach is to evaluate each measurement procedure for agreement with the mean of all measurement procedures and to adjust the calibration of any measurement procedures as necessary to produce equivalent results among the group.

There are no well-established guidelines regarding the number of samples to use for a round-robin exchange. The laboratory needs to establish the frequency of evaluation and the number of samples based on the stability of the measurement procedures, the frequency of reagent and calibrator lot changes, and the intended medical use of results in the health delivery system. Common practices include a round-robin exchange of one or more individual patient samples or a pool prepared from several samples on a weekly basis for high-volume measurement procedures or on a monthly or quarterly basis for lower-volume or very stable measurement procedures. For frequent comparisons with one or two samples, concentrations should be chosen to evaluate the measuring interval over a period of several examinations. For less frequent comparisons, a larger number of patient samples is recommended to cover the measuring interval. When establishing interpretation criteria, the laboratory needs to consider the limited statistical power for the number of results available. CLSI document EP31 provides a statistical approach suitable for using one to five samples in a comparison.<sup>53</sup>

Table 6.5 provides, as an example, empirical criteria used in one author's laboratory for evaluation of agreement among results for a single patient sample assayed weekly among multiple analyzers. These criteria were established based on the expected imprecision of the measurement procedures used and the influence of discrepant results on medical decisions. To allow for the limitations of a single measurement of a single sample in a comparison, a result outside a criterion is typically not acted on unless the magnitude of a difference is much larger than the criterion or the situation persists for 2 or more weeks.

Patient samples are recommended to verify agreement between multiple measurement procedures or instruments of the same type even when from the same manufacturer. Results for QC materials should not be used for the purpose of verifying consistency of results for patient samples measured using different measurement procedures or instruments. As discussed in an earlier section, QC materials are not validated



**FIGURE 6.15** Process used to evaluate agreement between measurement procedures and to adjust calibration, if necessary, to achieve equivalent results from different measurement procedures. (From Miller WG. Quality control. In: McPherson RA, Pincus MR. *Henry's clinical diagnosis and management by laboratory methods*. 24th ed. Philadelphia: Elsevier; 2020.)

to be commutable with patient samples between different measurement procedures. Even when more than one measurement procedure from the same manufacturer is used, differences may be seen in the measured values for QC materials between different reagent lots and different instruments.<sup>34</sup> In principle, if more than one of the same model of an instrument with the same reagent lots is used, all should have the same results for the same lot of QC material. In practice, differences in measurement details or maintenance condition between different instruments frequently cause small differences in QC results. The acceptance criteria for a QC result can be set to allow for such differences. However, more reliable conclusions will be drawn when patient samples are used to evaluate agreement among different instruments.

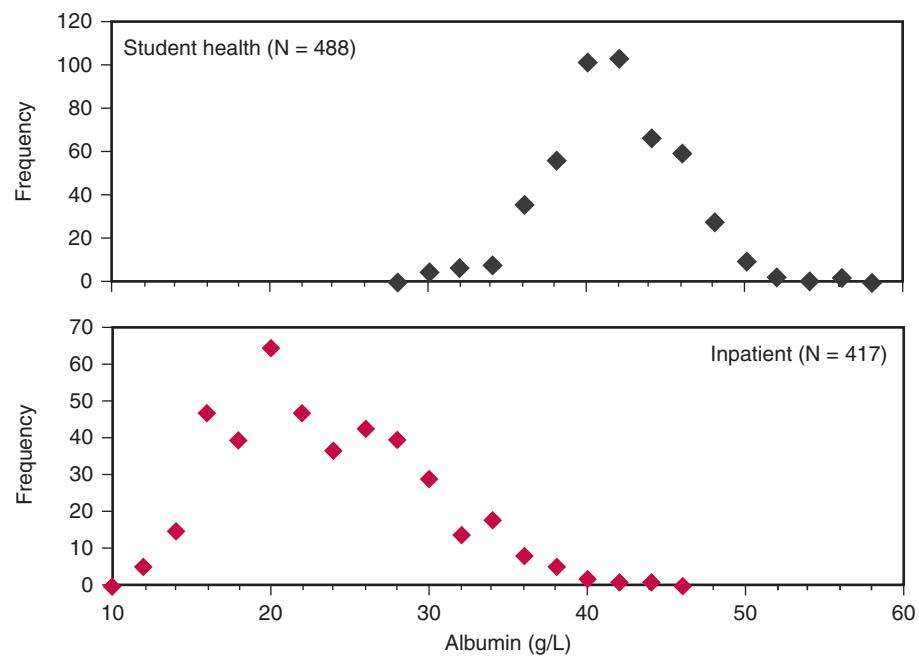
### Using Patient Data for Statistical Quality Control

Patient results can be used in a statistical QC process to monitor measurement procedure performance. For a sufficiently large number of results, the mean (or preferably, median) value may be sufficiently stable to be used as an indicator of measurement procedure consistency over time. This approach can be used on a periodic basis by extracting data for a specified time period (e.g., 1 day, week, month depending on the number of results available), calculating the mean and SD for the distribution of results, and comparing one time period versus another to determine whether any changes have occurred. This type of periodic evaluation can identify changes in calibration stability or in overall imprecision for a

measurement procedure. The mean and SD can also be compared for consistency among two or more measurement procedures for the same measurand.

An important limitation for using patient data to evaluate consistency within a single measurement procedure or between different measurement procedures for the same measurand is the physiologic homogeneity of results. Fig. 6.16 shows an example of the potential impact of a non-homogeneous sample of patients on distribution of serum albumin results for hospital general medicine inpatients compared with a student health outpatient clinic. The histograms are very different because the two patient groups differ in severity of disease and in recumbent versus supine position for blood collection, which influences vascular water volume and the concentration of albumin. Using patient data for process control for albumin measurement, similar to other measurands, requires partitioning of the patient population into a homogeneous subgroup.

Automated approaches to use the mean (or median) for groups of sequential patient results as a continuous process control parameter have been described.<sup>51,54–60</sup> These approaches are called “average of normals” (a misnomer because not all results are from normal individuals) or “moving average” techniques and are suitable for use in higher-volume measurement procedures in chemistry and hematology. In general, these approaches evaluate sequential patient results over time intervals from minutes to hours or days. The time interval that can be confidently used depends on the number of results in the time interval and the relative homogeneity of



**FIGURE 6.16** Histograms for distribution of sequential patient results for albumin from a student health outpatient clinic and a hospital general medicine inpatient unit. (From Miller WG. Quality control. In: McPherson RA, Pincus MR. *Henry's clinical diagnosis and management by laboratory methods*. 24th ed. Philadelphia: Elsevier; 2020.)

the clinical conditions represented in the patients' results. For some measurands, patients may need to be partitioned to obtain subgroups whose results are expected to be homogeneous. Considerations for partitioning include age, gender, ethnicity, and disease conditions. It is easier to obtain homogenous groups when monitoring one laboratory than when comparing several laboratories. Some approaches have arbitrarily ignored abnormal results in an attempt to restrict results to more normal health conditions. Removing abnormal results must be used with caution because excessive deletion will create an artificial subset of results that may not reflect a measurement procedure's calibration condition. The median of a group of patient results is sensitive to the overall distribution of results but is minimally influenced by extreme values and is recommended as the most robust estimate for tracking measurement procedure stability over time.

There is no consensus regarding the number of sequential patient results to include in a group for which the mean or median is calculated. An empirical approach based on extracting patient data from the laboratory computer system and simulating different group sizes in a spreadsheet can identify group sizes that have sufficiently small variation over time to provide useful information for tracking consistency of a measurement procedure. The same data can be used to assess the influence of partitioning considerations and to determine the statistical parameters to use for interpreting the data described in the next paragraph.

The mean or median for groups of patient results is tracked over sequential time intervals to monitor measurement procedure performance. The mean or median for groups of patient results can be treated as a QC sample value. A target value for the average mean or median and an SD for the distribution of mean or median values is determined, and these parameters are used to establish acceptance rules

similar to those used to interpret an individual surrogate QC sample result. Process control using patient data is primarily useful to identify bias and is less useful to identify changes in precision because of the inherent differences in measurand values among a group of patient results. Statistical procedures such as moving average, EWMA, or CUSUM are used to monitor trends in calibration status based on patient data.

Moving averages are frequently applied in hematology blood count measurement procedures based originally on the Bull's Algorithm approach.<sup>61</sup> Other than this application, patient results-based QC monitoring procedures have not been widely adopted because of lack of consensus guidelines for their use and lack of computer support from instrument and laboratory information system (LIS) suppliers. When computer support is available, patient data-based statistical monitoring is a useful addition to conventional QC measurements especially for analytes such as calcium, sodium, and others with low sigma metric performance capability. The advantage is that aggregated patient results give information in real time, the results are free from any matrix effects (i.e., the patients' samples are by definition "commutable"), and the results also include the pre-examination (preanalytical) variation. The disadvantage is that to obtain reliable results, the patient populations need to be stable<sup>59,60</sup> or be partitioned into stable subgroups as shown in the albumin example.

## EXTERNAL QUALITY ASSESSMENT OR PROFICIENCY TESTING

EQA/PT is a program used to evaluate measurement procedure performance by comparing a laboratory's results with those of other laboratories for the same set of samples.<sup>62</sup> EQA/PT providers circulate a set of samples among a group

**TABLE 6.8 Evaluation Capabilities of External Quality Assessment or Proficiency Testing Related to Scheme Design**

Category	Commutable	SAMPLE CHARACTERISTICS				EVALUATION CAPABILITY					
						Accuracy					
						Individual Laboratory					
						Relative to Participant Results					
						Reproducibility					
		Value Assigned With RMP or CRM		Replicate Samples in Survey	Absolute vs. RMP or CRM	Peer Group	Individual Laboratory Intralaboratory CV	Measurement Procedure Interlaboratory CV	Absolute vs. RMP or CRM	Relative to Participant Results	
1	Yes	Yes	Yes		X	X	X	X	X	X	
2	Yes	Yes	No		X	X		X	X	X	
3	Yes	No	Yes			X	X	X	X	X	
4	Yes	No	No			X	X		X		X
5	No	No	Yes				X	X	X		
6	No	No	No				X		X		

<sup>a</sup>Standardization when patient results are equivalent between measurement procedures and calibration is traceable to the Système International using a reference measurement procedure, harmonization when patient results are equivalent between measurement procedures, and calibration traceability is not based on a reference measurement procedure.

CRM, Certified reference material; CV, coefficient of variation; RMP, reference measurement procedure.

Reproduced with permission from Miller WG, Jones GRD, Horowitz GL, Weykamp C. Proficiency testing/external quality assessment: current challenges and future directions. *Clin Chem* 2011;57:1670–80.

of laboratories. Each laboratory measures the EQA/PT samples as if they were patient samples and reports the results to the EQA/PT provider for evaluation. The EQA/PT provider assigns a target value for the EQA/PT samples and determines if the results for an individual laboratory are in close enough agreement with the target value to be consistent with acceptable measurement procedure performance.

Ideally, an EQA/PT program should circulate commutable materials that are measured in replicates by the participating laboratories to provide participants with results that inform them if their measurement procedure has a bias from a true value assigned using an RMP or RM, or with results from all other measurement procedures. Unfortunately, commutable materials are often not used, and in some cases, circulation of unsuitable EQA/PT materials can cause harm by misclassifying measurement procedure performance.

EQA/PT is not available for some measurands because a particular measurement procedure may be new to the clinical laboratory, is not commonly performed, or because measurand stability makes it difficult to include in an EQA/PT material. In these situations, the laboratory should use an alternate approach to periodically verify acceptable performance of the measurement procedure. CLSI guideline QMS24 provides approaches for verifying measurement procedure performance when formal EQA/PT is not available.<sup>63</sup>

Before enrolling in an EQA/PT program, the laboratory should consider the quality of the program. The following questions have to be addressed: (1) Is the EQA/PT material commutable with patient samples? (2) How many replicates are measured? (3) How is the target value established? (4) What is the number of participants in the scheme and in a particular method group? (5) How is the measurement

method group established? And (6) How are the performance specifications set? This information is necessary to be able to interpret the feedback report from the organizer in a sensible way. Types of EQA/PT schemes are summarized in Table 6.8, with the most desirable type of program listed first and schemes that provide less information at the bottom.

### External Quality Assessment or Proficiency Testing Programs That Use Commutable Samples

EQA/PT programs that use commutable samples are preferred whenever available.<sup>62</sup> Refer to Fig. 6.5 for an explanation of commutability. Commutable samples are typically prepared by using an individual donor's specimen or by pooling patient samples with minimal processing or additives to avoid any alteration of the sample matrix. To achieve samples with abnormal values for measurands, donors can be identified with known pathologic conditions, or blood, plasma, serum, or urine units from a general donor population can be prescreened for a selected measurand. Supplementing donor samples or pools with purified analytes may be acceptable in some cases, but assessment of commutability should be performed to confirm that supplementation did not inappropriately alter the matrix.<sup>64</sup>

When commutable EQA/PT samples can be prepared, the results reflect what would be expected if individual patient samples were sent to each of the different laboratories. Thus agreement among different laboratories and measurement procedures (harmonization) can be correctly evaluated. The agreement between an individual laboratory result measured in singlicate and a reference measurement result gives an assessment of accuracy, the agreement between an individual laboratory result measured in replicate and a reference measurement result

gives an assessment of trueness, and the agreement between a measurement procedure group mean value and the reference measurement result gives an assessment of trueness and calibration traceability for the measurement procedure group. The latter information is of particular interest to the producers of measurement procedures and can be used as part of a surveillance program for the metrologic traceability scheme.<sup>65</sup>

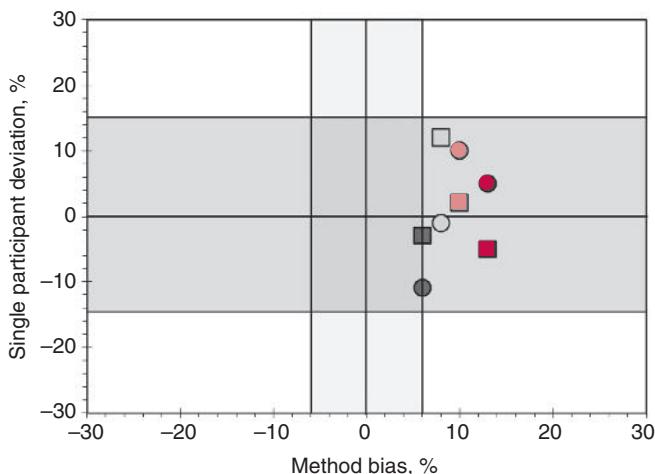
For example, EQA/PT programs for HbA<sub>1c</sub> in many countries use pooled, freshly collected whole blood from both normal and diabetic donors. The target values for the pooled blood are assigned by RMPs for HbA<sub>1c</sub>. In these surveys, the accuracy of individual laboratory results and the trueness of measurement procedure group means versus the RMP values can be evaluated because the EQA/PT samples are commutable with clinical patient samples. In these cases, the performance of different measurement procedures has been used to monitor and improve the calibration traceability processes used by the measurement procedure manufacturers for the benefit of improved patient care decisions regarding diabetes.<sup>66,67</sup>

Sufficient volumes of commutable materials have been challenging to prepare for use in large EQA/PT programs. An alternative is to combine commutable and noncommutable samples in the same EQA/PT event. An example from such a survey is shown in Fig. 6.17. The measurement procedure bias was evaluated based on results from a smaller group of the participating laboratories that measured the commutable samples, and individual participant's performance was evaluated based on agreement within a measurement procedure peer group using the noncommutable materials.<sup>68</sup> Use of commutable materials adds substantial value to the information obtained from EQA/PT survey results and is recommended whenever possible.<sup>62</sup> Refer to Chapter 7 for information on procedures to validate the commutability of QC, EQA/PT, and reference materials.<sup>64,69–73</sup>

### External Quality Assessment or Proficiency Testing Programs That Use Noncommutable Samples

Table 6.8 includes EQA/PT programs that use noncommutable samples. The materials commonly used for EQA/PT samples are derived from blood, urine, or other body fluids but are altered in the process to manufacture EQA/PT samples such that the matrix is modified and the samples frequently do not have the same measurement characteristics as observed for unaltered clinical patient samples.<sup>24–27,62,73</sup> In addition, some EQA/PT samples (e.g., cerebrospinal fluid or blood gas) are prepared as synthetic materials that are not derived from patient fluids. Consequently, many EQA/PT samples, as for QC samples, are noncommutable with authentic patient samples. The results for a noncommutable EQA/PT sample will have a different relationship in their numeric values between different measurement procedures and sometimes for different reagent lots within a measurement procedure than would be observed for patient samples.

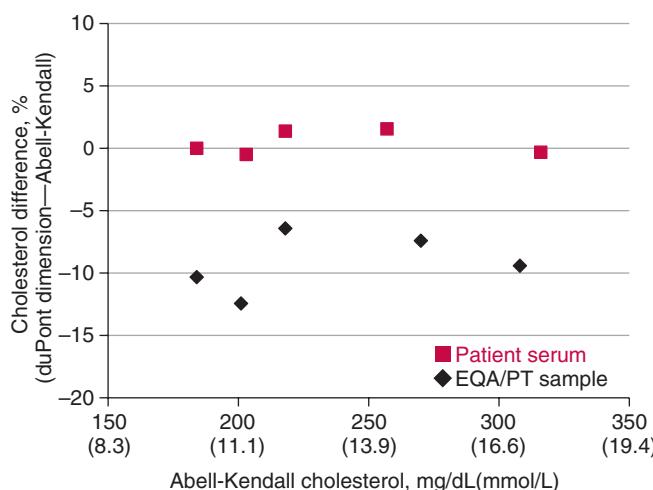
Because EQA/PT samples are frequently noncommutable with patient samples, it is a common practice for EQA/PT providers to organize results into “peer groups” of measurement procedures that represent similar technology expected to have the same result for a noncommutable EQA/PT sample. The mean or median value of the peer group results is the target value. Because the peer group mean value may be influenced by a matrix-related bias, that value can only be used to evaluate laboratories using the same or very similar



**FIGURE 6.17** Example of part of an international normalized ratio external quality control report from the Norwegian Organisation for Quality Improvement of Laboratory Examinations (Noklus). The bias of each measurement procedure from the conventional true value was obtained using patient samples for each measurement procedure (x-axis). A single participant was evaluated against the peer group target value for a given measurement procedure (y-axis). *Vertical lines* represent acceptable bias for measurement procedure performance, and *horizontal lines* represent acceptable deviation from a peer group target value. The *circle* and *square* represent two samples in one survey with each color representing a different survey (four surveys). The results show that, whereas the measurement procedures had a bias of about 10%, the participant's deviations from the peer group target value were within the performance specifications. This graph indicates that the participants perform the measurement procedures correctly, but the measurement procedures had an unacceptable positive bias. (Modified with permission from Stavelin A, Petersen PH, Sølvik UØ, Sandberg S. External quality assessment of point-of-care methods: model for combined assessment of method bias and single-participant performance by the use of native patient samples and noncommutable control materials. *Clin Chem* 2013;59:363–71.)

measurement procedures and cannot be used to evaluate if results from different measurement procedures agree with each other. Peer groups may be formed arbitrarily based on the apparent agreement among results for different measurement procedures, but there is no scientific basis for this practice, and unexpected changes may occur, such as when new reagent lots or formulations are introduced by a manufacturer. The measurement procedures included in an arbitrarily formed peer group may not be similar. In this situation, one of the measurement procedures in the “peer group” may dominate the number of results and inappropriately influence the target value if set as the mean/median of all results. An alternative is to set the target value not as the mean/median of all results, but as a mean/median of the means/medians determined for each different measurement procedure in the peer group.

However, even within a peer group using the same measurement procedure, differences can occur because of different reagent lots used in different laboratories because the matrix of the EQA/PT material can influence the results from different reagent lots even if the patient samples give similar results.<sup>29</sup> Therefore in some cases, reagent lots should be registered, and even reagent lot-specific target values may need



**FIGURE 6.18** Example of noncommutable results between proficiency testing samples and pooled patient serum samples for a specific measurement procedure. *EQA*, External quality assessment; *PT*, proficiency testing. (Data replotted from Naito HK, Kwak YS, Hartfiel JL, Park JK, Travers EM, Myers GL, 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* 1993;117:345–51.)

to be assigned.<sup>74,75</sup> When target values are set for noncommutable materials, it is important how the target values are calculated, how outlier results are treated, and what uncertainty is associated with the target value. When target values are assigned from relatively small numbers of results in a peer group, the target value should be given with an uncertainty and the criteria for acceptable performance should be “extended” to high and low values that include the uncertainty.

Fig. 6.18 illustrates the effects of noncommutable materials on interpretation of EQA/PT results and demonstrates why “peer group” evaluation is used. In this older but still valid example, pooled patient sera and EQA/PT samples were measured by the DuPont Dimension Analyzer and by the Abell-Kendall RMP for cholesterol.<sup>76</sup> The Abell-Kendall measurement procedure was shown to be unaffected by matrix-induced changes in EQA/PT samples.<sup>77</sup> The patient samples showed excellent agreement between the two measurement procedures (average bias, 0.2%). However, the EQA/PT samples had a large negative bias (average –9.5%) between measurement procedures, caused by a matrix-related bias with the DuPont measurement procedure that was not present with the RMP.<sup>78</sup>

In this example, the routine measurement procedure was correctly calibrated and produced results for patient samples that were traceable to the RMP. However, EQA/PT results gave an incorrect impression of the measurement procedure’s calibration relationship to the RMP. If the routine measurement procedure’s calibration had been erroneously adjusted on the basis of EQA/PT results, the results for patient samples would then be incorrect. EQA/PT results were useful for evaluating the performance of all laboratories using the DuPont measurement procedure because the matrix-related bias was uniform within this peer group. Consequently, if an individual laboratory’s results agreed with those of the peer group, the individual laboratory could conclude that the measurement procedure was performing in conformance

with the manufacturer’s specifications. In general, an individual laboratory depends on the manufacturer to correctly calibrate a clinical laboratory measurement procedure to be traceable to the reference system for a measurand.

QC material manufacturers may provide a data analysis service that compares results from different laboratories using the same lot of QC material by calculating group statistics for performance evaluation. This type of interlaboratory QC data analysis provides similar information to that from EQA/PT programs that use noncommutable samples. Interlaboratory QC data comparison allows a laboratory to verify that it is producing QC results that are consistent with those of other laboratories using the same measurement procedure and lot of QC material. This information can be helpful for troubleshooting measurement procedure issues and for assessing performance of a new measurement procedure being introduced to a laboratory.

### External Quality Assessment or Proficiency Testing Programs for Measurements on a Nominal or Ordinal Scale

Many constituents in laboratory medicine can be measured on a nominal scale (all types of classification without any quantitative value, e.g., identity of a bacteria as *Escherichia coli*, Group A Strep, or *Klebsiella*; a virus as SARS-COV-2 or different mutations) or an ordinal scale (all types of graded response, e.g., urine strips with values 0, 1, 2, 3, 4 for increasing amounts of analyte present). Often, measurements performed on an ordinal scale are measurements that can also be performed on a ratio or interval (numeric) scale. The quantities are often measured on an ordinal scale because a more rapid result can be obtained and because such tests can be performed by nonprofessional users (e.g., in a physician’s office or by lay people using a POC device). When setting up an EQA/PT program for such measurement procedures, it is important to note that all aspects regarding commutability of the QC samples and thereby establishing target values will be the same as for measurement procedures on the interval or ratio scale (quantitative measurements).

EQA/PT programs addressing identifications of species or mutations often circulate multiple samples where different mutations of species should be identified and the participants are classified according to the percentage of correct identifications.<sup>79,80</sup> Results from measurement procedures using the ordinal scale can be dichotomous (often called qualitative tests) or multinary with more steps (often called semi-quantitative tests) in which each category can be considered as a dichotomous test. It is possible to evaluate the results from these tests using a rankit ordinal model.<sup>81,82</sup> The performance characteristics of the measurement procedure should be described from the manufacturer giving a detection limit for dichotomous tests and, for example, the concentrations below which 5% of the samples should be negative, the concentration at which 50% of the samples should be positive, and the concentration above which 95% of the samples should be positive ( $C_5$ ,  $C_{50}$ ,  $C_{95}$ , respectively) when related to a ratio scale. Performance specifications for such measurement procedures should use the same models as described earlier<sup>8</sup> and can, for example, relate to the percentage of results that should be positive or negative above or below a certain concentration.<sup>82</sup> These performance specifications

are, however, easier to apply for method evaluation than for single participant evaluation in an EQA/PT scheme because numerous samples with different concentrations are necessary.

In an EQA/PT assessment, it is useful to circulate samples with concentrations that are expected to give “positive” or “negative” results and samples with an intermediate concentration that can have both positive and negative results. In the feedback report, participants will typically be evaluated with respect to the positive or negative samples because failure to obtain the expected results will be evaluated as “poor” performance. Samples with intermediate concentrations may be included to assess the robustness of threshold values by reporting to the participants how many obtained positive results and how many obtained negative results. However, intermediate concentrations are typically not graded because the results are expected to be mixed between “positive” and “negative.” Such information is useful to assess and to monitor the performance of the measurement procedure. For example, a study using EQA/PT results showed that six of eight POC measurement procedures for human chorionic gonadotropin gave 3 to 11% false-positive results.<sup>82</sup> Using a commutable EQA/PT material, different measurement procedures can be compared and monitored over time, and it is possible to identify opportunities when the threshold discrimination needs to be improved among different measurement procedures.<sup>83,84</sup>

Some EQA/PT programs, often for rare diseases, are examining the whole testing procedure (e.g., the correct measurands to request for a certain diagnostic problem, the appropriate sample collection and transportation, the performance of the analytical measurement procedures, the adequacy of the diagnosis, and the report provided to the clinicians).<sup>79,85,86</sup> These programs are often run on an international level.

### Reporting External Quality Assessment or Proficiency Testing Results When One Measurement Procedure Is Adjusted to Agree With Another Measurement Procedure

It is good laboratory practice to adjust the calibration of different measurement procedures for the same measurand used within a large hospital system that can have several satellite laboratories or a collection of several hospitals with the same management structure so that the results for patient samples are consistent, irrespective of which measurement procedure is used. Such harmonization of results is important for uniform use of reference intervals and decision thresholds within a hospital or clinic system.

It is important to report EQA/PT results such that they can be properly evaluated against a true target value or the peer group target value that reflects the calibration established by the measurement procedure manufacturer. When a laboratory has applied a calibration correction, it is important to inform the EQA provider what result is provided. Usually the individual EQA/PT result should be reported to the EQA/PT provider after removing any calibration adjustments so that the reported result is consistent with the manufacturer’s nonadjusted calibration. The most convenient way to remove a calibration adjustment is to first measure the EQA/PT samples with the calibration adjustment applied to

the measurement procedure, as would be the usual measurement process for patient samples. After the measurement, the EQA/PT results should be adjusted “in reverse” by mathematically removing the calibration adjustment factors, and the results should be reported to the EQA/PT provider with any adjustment factors removed. One should not recalibrate the instrument with a new set of calibrators for the purpose of measuring the EQA/PT samples because this practice would violate regulations requiring the EQA/PT material to be measured in the same manner as patient samples.

For example, a laboratory has performed a patient sample comparison between measurement procedure A used in the main laboratory and measurement procedure B used in a satellite laboratory. Measurement procedure B consistently gave 10% higher results (i.e., a slope of 1.10 and a negligible intercept were observed for a regression analysis). Measurement procedure B was adjusted to agree with measurement procedure A by putting the adjustment factor  $1/1.10 = 0.9091$  in the measurement procedure B instrument to automatically multiply each measured result by 0.9091 to lower the reported result to be equivalent to a value that would have been reported by method A. When EQA/PT results from measurement procedure B are reported, it is necessary to remove the 0.9091 factor to allow the reported result to be compared with the peer group mean of results from all laboratories using measurement procedure B. Removing the 0.9091 factor is accomplished by multiplying the reported EQA/PT result from measurement procedure B by the factor  $1/0.9091 = 1.100$  to increase its numeric value by 10% to the nonadjusted value that was actually measured according to the manufacturer’s defined calibration procedure for measurement procedure B. This process allows the EQA/PT result measured by measurement procedure B to be appropriately evaluated in comparison with its peer group mean, which will reflect the manufacturer’s established calibration. This process permits the EQA/PT sample to be measured in the same manner as patient samples and the numeric result reported to the EQA/PT provider to reflect the actual measured result using the manufacturer’s calibration settings.

### Interpretation of External Quality Assessment or Proficiency Testing Results

Many countries have regulations requiring EQA/PT and specifying the evaluation criteria for acceptable performance. When criteria are set by regulations, an EQA/PT provider is required to use them. When criteria are not set by regulations, the EQA/PT provider sets evaluation criteria on the basis of clinically acceptable performance, biologic variation, or the analytical capability of the measurement procedures in use. EQA/PT evaluation criteria are usually designed to evaluate the accuracy of a single measurement. In some cases, measurements are made several times, and it is possible to separately assess the bias and the imprecision. The acceptability limits for EQA/PT include bias and imprecision components considered acceptable for a measurand plus other error components that are unique to EQA/PT samples such as between-laboratory variation in calibration; variable matrix-related bias with different lots of reagent within a peer group; uncertainty in the target value; stability variability in the EQA/PT material, both in storage and shipping, and after reconstitution or opening in the laboratory; and homogeneity of the EQA/PT material vials. Consequently, the acceptability

limits for EQA/PT samples are frequently larger than what might be expected for clinically acceptable total error with patient samples.

The acceptability limits are partly dependent on the quality of the EQA/PT material (e.g., its commutability, stability, homogeneity, and methods of reconstitution). A commutable EQA/PT material often has a target value assigned by an RMP or by value transfer from suitable measurement procedures calibrated using commutable certified reference materials. A commutable EQA/PT material should in principle have the same results for all measurement procedures and lots of reagent and measurement procedure calibrator. In this case, the variation in the results will reflect the different measurement procedures, reagent lots, and calibrator lots in use from all of the manufacturers represented in the survey. For example, a multi-country European EQA/PT program that used commutable samples documented the performance of 17 measurands among six commercially available measurement procedures.<sup>87</sup>

With a noncommutable EQA/PT material, the target value is set by the mean or median of the peer group that should include only very similar measurement procedures, and the acceptance criteria could be stricter because the variability only includes variation within the same measurement procedure. However, in some cases, a noncommutable EQA/PT material is not even commutable among reagent lots for the same measurement procedure, and in theory, each reagent lot could have its own target value with even stricter acceptability limits.<sup>28,29,74,75</sup>

**Fig. 6.19** is an example of a typical evaluation report sent to a participating laboratory when noncommutable samples were used. Each reported result is compared with the mean result for the peer group using the same measurement procedure. The report also includes the SD for the distribution of results in the peer group, the number of laboratories in the peer group, and the SDI, which expresses the reported result as the number of SDs it is from the mean value (SDI = [Result – Mean]/SD). The limits of acceptability are shown. Acceptability

#### A External Quality Assessment (Proficiency Testing) Participant Report

Shipment date: 1 May 2015

Evaluation date: 12 June 2015

Analyte Units Method	Specimen	Reported Result	Limits of Acceptability					
			Mean	SD	Labs (n)	SDI	Lower	Upper
Calcium mg/dL	1	9.6	9.92	0.23	587	-1.4	8.9	11.0
	2	8.8	8.86	0.26	592	-0.2	7.8	9.9
Arsenazo dye	3	7.5	7.65	0.23	587	-0.7	6.6	8.7
Manufacturer A	4	8.2	8.43	0.23	590	-1.0	7.4	9.5
	5	10.8	10.87	0.25	589	-0.3	9.8	11.9
Iron µg/dL	1	190	192.5	7.0	397	-0.4	154	232
	2	65	65.0	3.4	394	0.0	51	78
Pyridylazo dye	3	74	69.2	3.2	395	+1.5	55	83
Manufacturer A	4	124	107.9	4.6	395	+3.5	86	130
	5	277	260.9	8.8	396	+1.8	208	314

#### B External Quality Assessment (Proficiency Testing) Participant Report

Shipment date: 1 May 2015

Evaluation date: 12 June 2015

Analyte Units Method	Specimen	Reported Result	Limits of Acceptability					
			Mean	SD	Labs (n)	SDI	Lower	Upper
Calcium mmol/L	1	2.40	2.48	0.06	587	-1.4	2.22	2.74
	2	2.20	2.21	0.06	592	-0.2	1.95	2.47
Arsenazo dye	3	1.87	1.91	0.06	587	-0.7	1.65	2.17
Manufacturer A	4	2.05	2.10	0.06	590	-1.0	1.85	2.37
	5	2.69	2.71	0.06	589	-0.3	2.45	2.97
Iron µmol/L	1	34.0	34.5	1.3	397	-0.4	27.6	41.5
	2	11.6	11.6	0.6	394	0.0	9.1	14.0
Pyridylazo dye	3	13.3	12.4	0.6	395	+1.5	9.8	14.9
Manufacturer A	4	22.2	19.3	0.8	395	+3.5	15.4	23.3
	5	49.6	46.7	1.6	396	+1.8	37.2	56.2

**FIGURE 6.19** Example of an external proficiency testing evaluation report sent to a participating laboratory. (A) conventional units; (B) SI units. The red value represents a large SDI. SD, Standard deviation;

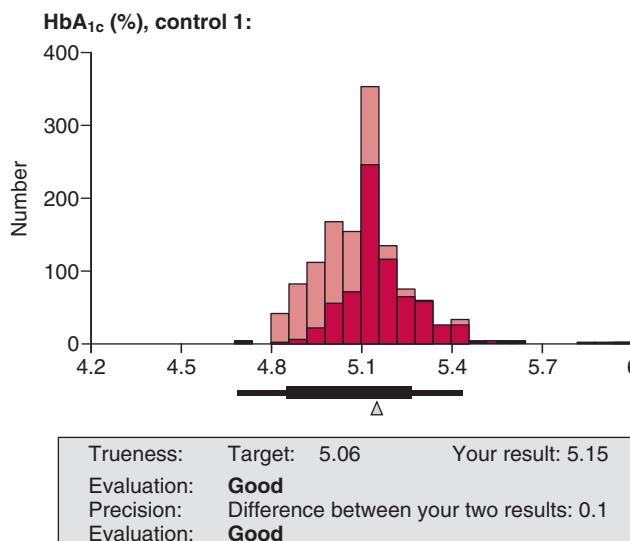
SDI, standard deviation interval. (A, from Miller WG. Quality control. In: McPherson RA, Pincus MR.

Henry's clinical diagnosis and management by laboratory methods. 24th ed. Philadelphia: Elsevier; 2020.)

criteria may be a number of SDs from the mean value, a fixed percent from the mean value, or a fixed concentration from the mean value. For example, in Fig. 6.19, calcium acceptability criteria are  $\pm 1$  mg/dL (0.25 mmol/L) from the mean value, and iron criteria are  $\pm 20\%$  from the mean value.

Peer group evaluation allows a laboratory to verify that its EQA/PT results are consistent with those of other laboratories using the same measurement procedure and by extension that its results for patient results are in agreement with those of other laboratories in the peer group. Consequently, the laboratory can conclude that it is using a commercially available measurement procedure according to the manufacturer's specifications.<sup>62</sup> In Fig. 6.19, the calcium results are in close agreement with the peer group mean (SDI ranges from  $-0.2$  to  $-1.4$ ). However, the iron results show greater variability, with one result  $+3.5$  SDI. Although this iron result is within the acceptability criteria, it is recommended to investigate the measurement procedure because a  $+3.5$  SDI is more likely to be different from, than to be in agreement with, the peer group.

Fig. 6.20 shows a typical evaluation report sent to a primary care office for POC measurement of HbA<sub>1c</sub> for one of two EQA/PT samples. In this situation, the EQA/PT provider is communicating directly with the user of the measurement procedure, the clinician or the coworker in the general practice office, and the feedback must be easy to understand for nonlaboratory professionals. The EQA/PT result is evaluated



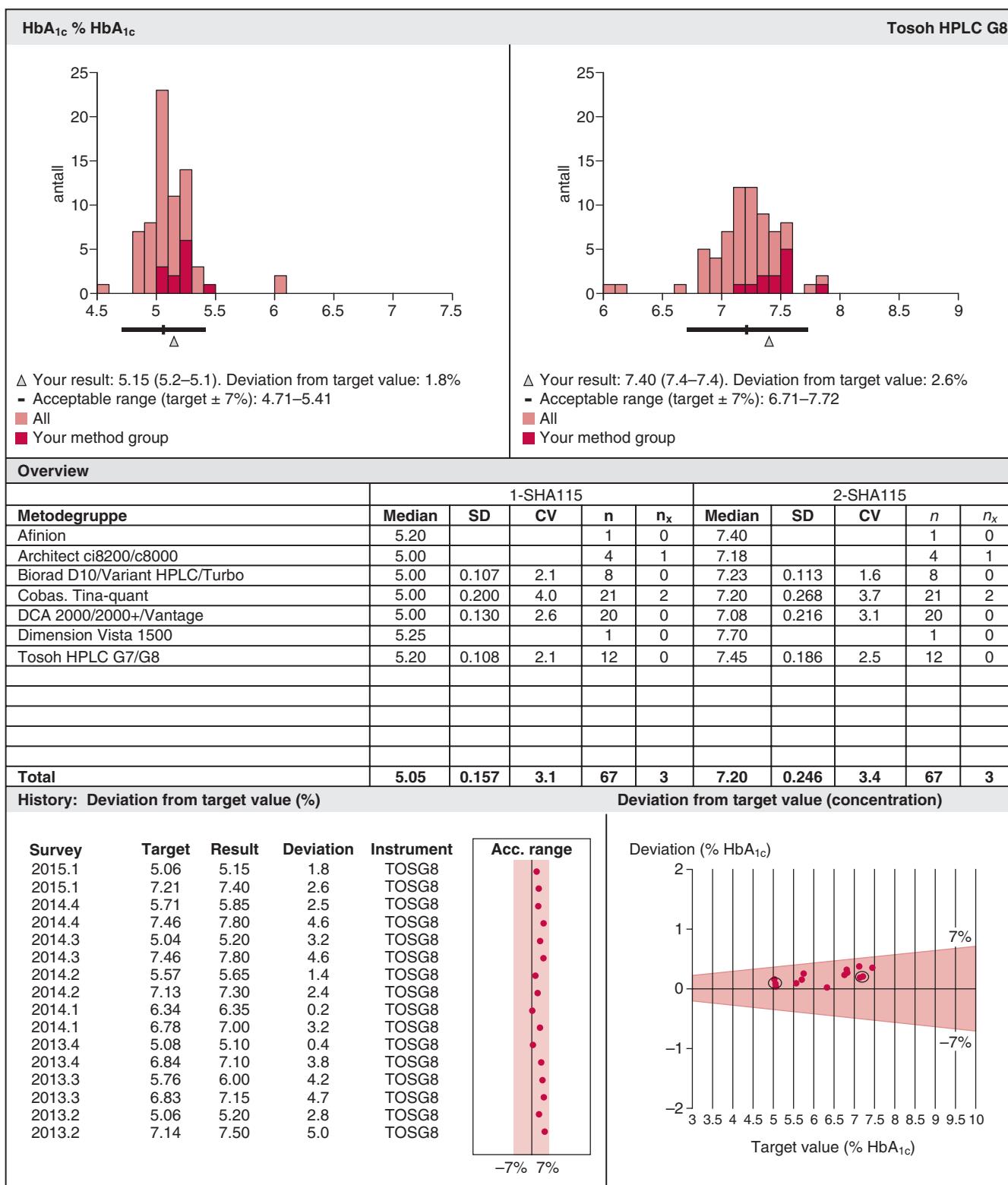
**FIGURE 6.20** Example of part of a feedback report to hemoglobin A<sub>1c</sub> (HbA<sub>1c</sub>) point-of-care (POC) users in a survey for general practitioners' offices and nursing homes. Commutable EQA/PT material was measured in duplicate. The participant is informed about the bias (mean of the two results) compared with a reference measurement procedure target (x-axis) and "precision" as the difference between the two results. The histogram represents the distribution of results among all participants (pink) and for the participant's method group (red). The thick black line represents the interval for "good" results, and the thin black line represents the interval for "acceptable" results. Results outside these limits are characterized as "poor." The triangle points to the result of the participant. (Modified with permission from the Norwegian Organisation for Quality Improvement of Laboratory Examinations (Noklus), the external quality assessment provider in Norway.)

as "good," "acceptable," or "poor." The lot numbers of the reagent are always registered so that the participant, in case of an aberrant result, can get information if the result was due to the measurement procedure used, the reagent lot used, or the performance of the user. In this case, commutable samples were used and in addition to participant reports, the manufacturers were informed of aberrant reagent lots. In cases when noncommutable samples are used, comments to results for aberrant reagent lots typically include a sentence that the EQA/PT result may not necessarily reflect results for patient samples. In all cases, the participants are encouraged to contact the organizers to sort out problems.

Fig. 6.21 shows a similar report from the same HbA<sub>1c</sub> survey provided to hospital laboratories. In addition to the figures about the distribution of results, information is provided on how different measurement procedures performed and a historical overview of performance on consecutive EQA/PT samples and performance related to the concentration of the sample. The EQA/PT material used for the HbA<sub>1c</sub> is pooled fresh patient blood (commutable), and the target value was set by an RMP and is therefore the same for all measurement procedures. Each sample was measured in duplicate (as requested by the EQA/PT provider), and the mean of the duplicates was used to estimate bias versus the RMP. In the present example, the performance was within the acceptability limits but with a generally high bias during the whole period. Because this observation was true for all the participants using this measurement procedure, the EQA/PT organizer discussed the results with the manufacturer to solve the problem. Until the problem was solved (the manufacturer had to make a new calibrator), the participants were advised by the EQA/PT provider to use a correction factor when reporting their results for patient samples.<sup>88</sup>

If an unacceptable EQA/PT result is identified, the measurement procedure must be investigated for possible causes and the necessary corrective action taken. Even when an EQA/PT result is within acceptability criteria, it is a good laboratory practice to investigate results that are more than approximately 2.5 SDI from the peer group mean. When the SDI is 2.5, there is only a 0.6% probability that the result will be within the expected distribution for the peer group; consequently, the probability is reasonable that a measurement procedure problem may need to be corrected. In addition, EQA/PT results that have been near the failure limit for more than one EQA/PT event, even if the results have met the EQA/PT acceptance criteria, should initiate a review for systematic problems with the measurement procedure. These practices support identification of potential problems before they progress to more serious situations. When results are investigated, a limitation of SD-based grading criteria should be considered. A peer group with very precise measurement procedures may have a very small SD, and even if a result is outside an SD limit, the finding may be inconsequential regarding the intended use of results for medical decisions.

Common causes for EQA/PT failure are listed in Box 6.1. Incorrect handling and reporting are unique to EQA/PT events and may not reflect the process used in the laboratory for patient samples. Nonetheless, these situations reflect the attention to detail, which is a necessary attribute for quality laboratory testing. Occasionally, the EQA/PT material may have a defect that causes it to perform inappropriately for all or a subgroup of measurement procedures or reagent lots. In



**FIGURE 6.21** Example of a part of a feedback report to hemoglobin A<sub>1c</sub> (HbA<sub>1c</sub>) users in hospital laboratories. Same survey and same materials as presented in Fig. 6.20. The histogram represents the distribution of results among all participants (pink) and for the participant's method group (red). Only limits for "acceptable" results are given (black lines in figures). Information about performance of measurement procedures is given in addition to a historical overview of percentage deviation from target values dependent on time and concentration of HbA<sub>1c</sub>. Shaded area represents the limits of acceptable performance. CV, Coefficient of variation; HPLC, high pressure liquid chromatography; SD, standard deviation. (Modified with permission from the Norwegian Organization for Quality Improvement of Laboratory Examinations (Noklus), the external quality assessment provider in Norway.)

**BOX 6.1 Classification of Potential Problems Identified When Investigating Unacceptable External Quality Assessment or Proficiency Testing Results<sup>a</sup>**

<b>1. Clerical errors</b>	Equipment component malfunction (e.g., light source, membrane, fluidics, detector)
Incorrectly transcribed EQA/PT result from the instrument read-out to the report form	Incorrect instrument conditions (e.g., water quality, surrounding temperature)
The EQA/PT sample was mislabeled in the laboratory	Instrument maintenance not performed appropriately
Incorrect instrument or measurement procedure was reported on the results submission form	
Incorrect units were reported	<b>4. Technical problems caused by personnel errors</b>
Decimal point was misplaced	Did not operate equipment correctly or did not conform to measurement procedure SOP
<b>2. Measurement procedure problems</b>	Incorrect storage, preparation, or handling of reagents or calibrators
Inadequate standard operating procedure (SOP)	Delay causing evaporation or deterioration of the EQA/PT sample
Problem with manufacture or preparation of reagents or calibrators (e.g., unstable)	Failure to follow recommended instrument function checks or maintenance
Lot-to-lot variation in reagents or calibrators	Pipetting or dilution error
Incorrect value assignment of calibrators	Calculation error
Measurement procedure lacks adequate specificity for the measurand	Misinterpretation of test result
Measurement procedure lacks adequate sensitivity to measure the concentration	
Carry-over from a previous sample	<b>5. A problem with the EQA/PT material such as:</b>
Inadequate QC procedures used	Incorrect storage, preparation, or handling of EQA/PT materials
<b>3. Equipment problems</b>	Differences between EQA/PT samples and patient samples (e.g., matrix, additives, stabilizers)
Obstruction of instrument tubing or orifice by clot	Sample deteriorated in transit or during laboratory storage
Misalignment of instrument probes	Sample had weak or borderline reaction
Incorrect instrument data processing functions	Sample contained interfering factors (which may be measurement procedure specific)
Incorrect instrument setting	Sample was not homogeneous among vials
Automatic pipettor not calibrated to acceptable precision and accuracy	

<sup>a</sup>This classification scheme assists in developing an appropriate corrective action plan.

EQA, External quality assessment; PT, proficiency testing; QC, quality control.

From Miller WG, Jones GRD, Horowitz GL, Weykamp C. Proficiency testing/external quality assessment: current challenges and future directions. *Clin Chem* 2011;57:1670–80.

this case, the EQA/PT provider should recognize the problem and not grade participants for that sample. Because the influence of reagent lots on noncommutability related bias is well documented,<sup>28,29,74,75</sup> EQA/PT programs are recommended to register reagent lots as part of the reporting process so this limitation can be more appropriately addressed in the scoring and investigating schemes.<sup>74,75</sup>

EQA/PT results are usually received several weeks after the date of testing. Consequently, investigation of unacceptable results requiring review of QC, reagent lots, calibration frequency and lots, and maintenance records for the date of the test and the preceding several weeks or months is necessary. It is common practice to save any remaining EQA/PT samples for use to investigate an unacceptable result. Care must be taken to store the residual EQA/PT samples to preserve the stability of the measurands. In some cases, the measurand will not be stable during storage. In addition, degradation that could affect any remeasured value may have occurred during storage, for example freeze-thaw, or possibly before storage while the materials were still being tested in the laboratory for the EQA/PT event. It may be possible to obtain additional vials from the EQA/PT provider. If a review of records suggests a stable operating condition, and a review of the EQA/PT material handling and documentation does not identify a cause for the erroneous EQA/PT result, one can

conclude that the EQA/PT failure was a random event. Investigative steps, data reviews, conclusions, and all corrective actions must be documented in a written report to address the unacceptable EQA/PT results and reviewed by the laboratory director. Some EQA/PT programs provide the participants with flow charts or checklists to be used to identify the reason for the EQA/PT failure.

### Interpreting External Quality Assessment or Proficiency Testing Summary Reports

EQA/PT providers typically provide a summary report, which includes the mean and SD for all peer groups represented by the participants' results (see Figs. 6.19 to 6.21). When commutable materials were used in the surveys, the trueness compared with an RMP or the harmonization among different measurement procedures can be assessed.

When summary reports are for surveys with noncommutable materials, assessment of mean results among different peer groups or to an RMP is not possible. The EQA/PT results from noncommutable samples are not reliable to infer agreement or lack of agreement for patient results among different measurement procedures for the same measurand. In this case, the peer group mean and SD are useful for evaluating the uniformity of results among laboratories using the same measurement procedure, and to evaluate the consistency

of an individual measurement procedure's performance over time intervals from one EQA/PT event to the next (trend monitoring). A limitation using EQA/PT results for trend monitoring is that differences in matrix-related bias in different sample materials can be different within the same peer group over time.

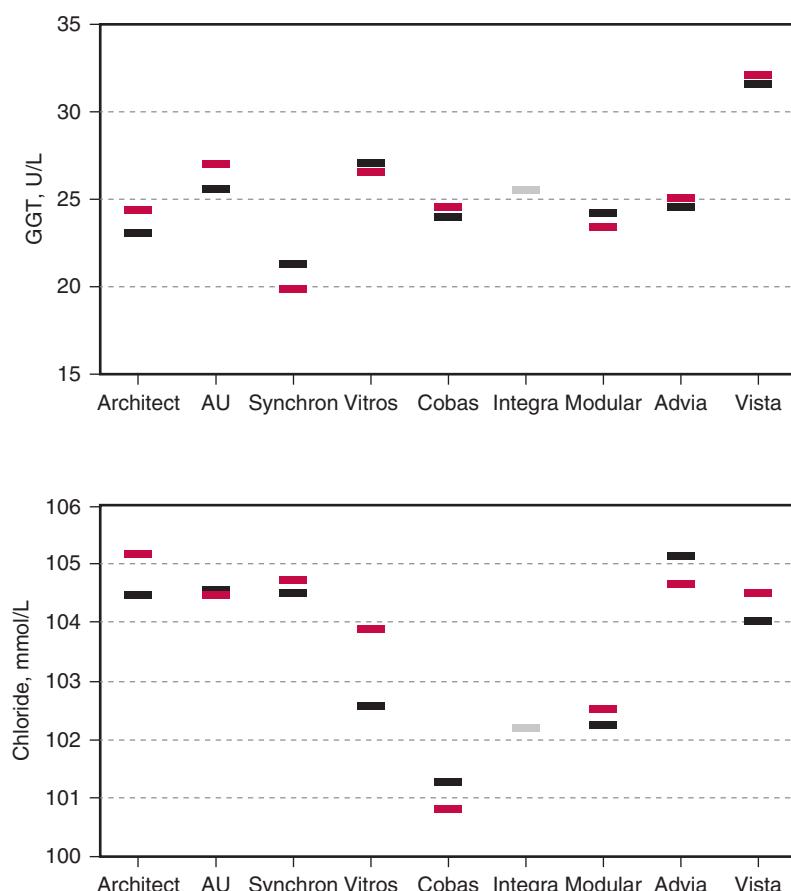
Summary information also allows evaluation of the imprecision of various measurement procedure groups, within the limitations of EQA/PT material and reagent lot matrix-bias differences. The number of users in each measurement procedure group reveals which measurement procedures are commonly used.

### Using Patient Medians for External Quality Assessment

One of the main objectives of EQA/PT is to compare results between different measurement procedures and, if possible, to compare them with a true value obtained by an RMP or reference material. This information is important both for the laboratory and also for the IVD manufacturers. However, a comparison between measurement procedures is only possible by using commutable samples. Commutable samples can be difficult to prepare for numerous measurands since they are often more expensive, unstable, and not possible to obtain for many measurands.

Comparison of results between measurement procedures can be done by comparing medians of patient results between measurement procedures performed in different laboratories. It is then possible in real time to monitor the effect of harmonization and standardization efforts. One prerequisite for using patient medians for this purpose is that the patient medians originate from similar patient populations. Patient medians also include pre-examination (preanalytical) factors in addition to measurement biases. Commutable EQA/PT samples can be circulated to the same laboratories to validate that differences in patient medians between measurement procedures reflect "real" measurement procedure differences if these differences can be reproduced with the commutable samples.

An example of such a comparison is shown in Fig. 6.22 where results of patient medians reflect the same differences as observed for fresh frozen patient samples.<sup>89</sup> When this relationship is established, daily or weekly monitoring of the measurement procedures can be performed using the patient medians. A laboratory can then monitor in a "continuous" manner its own measurement procedure in comparison with others to determine: (i) results comparability, (ii) measurement procedure stability including lot-to-lot reagent and calibrator changes, and (iii) individual laboratory bias, for example, caused by changes in preanalytical conditions. This



**FIGURE 6.22** Comparison of medians of patient results (black bars) from 100 laboratories with a total of 182 devices with the results from peer group medians (red bars) of 20 fresh frozen serum samples analyzed by 20 laboratories per peer group for chloride and gamma-glutamyl transferase (GGT).

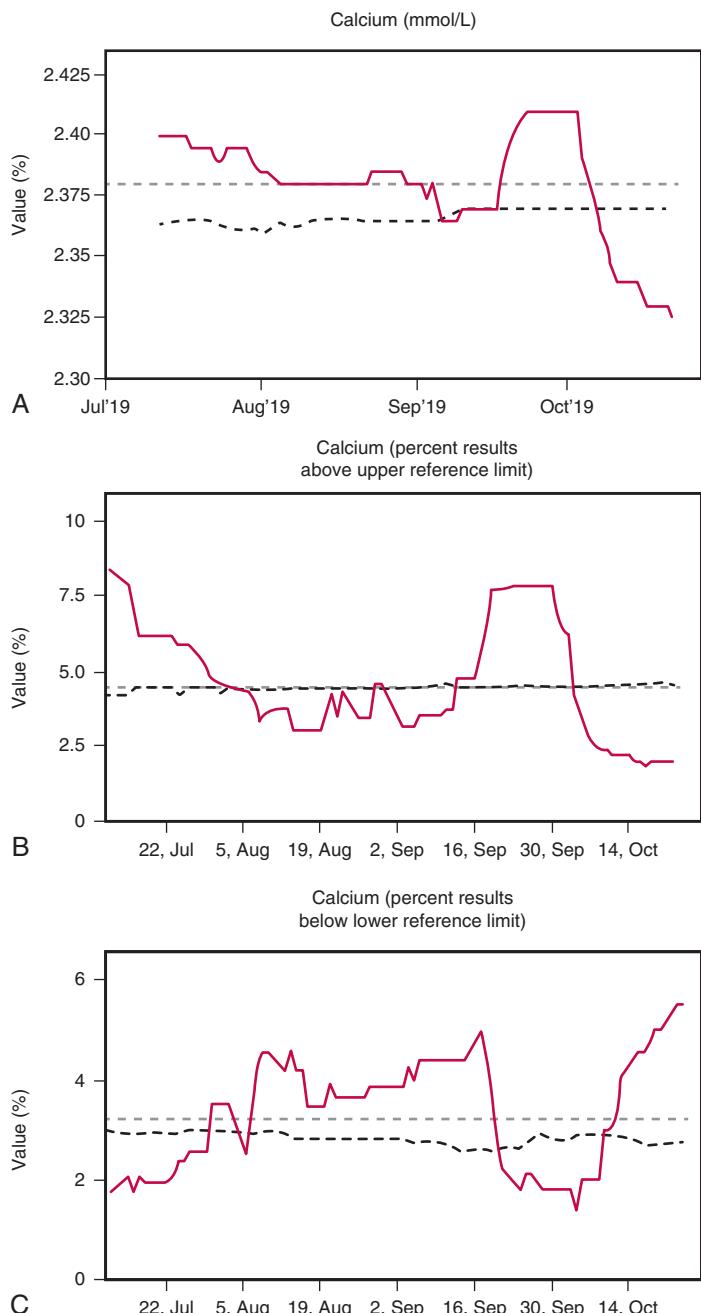
(From De Grande LAC, Goossens K, Van Uytfanghe K, Stöckl D, Thienpont LM. The empower project—a new way of assessing and monitoring test comparability and stability. *Clin Chem Lab Med* 2015;53:1197–204.)

near-continuous monitoring allows remediation shortly after an event takes place.

It is also possible to calculate the percentage of patient results that is outside a reference interval or a specific clinical cut-off interval. Thus the consequences for patients of any changes in measurement performance can be easily shown. For example, Fig. 6.23 shows moving daily patient medians for the previous 16 days for serum calcium and the percent of

median results that are above or below the reference interval limits. The moving median could be determined from other time periods, for example, 5 days or 8 days as appropriate for a given measurand. The laboratory can, in addition to being compared with all laboratories, choose to be compared with their own instrument group.

Development of more sophisticated information management systems makes it easier for laboratories to obtain



**FIGURE 6.23** Moving daily patient median (average medians of the last 16 days) of (A) the concentration of calcium (mmol/L); (B) the percentage of calcium results outside the upper reference limit; and (C) the percentage of results below the lower reference limit. The red line represents the results from one laboratory, the grey broken line represents the long-term median of the laboratory, and the black broken line represents the results from the moving median (16 days) from all the laboratories participating in the program ( $\approx 150$  laboratories). (Reprinted with permission from the Norwegian Organization for Quality Improvement of Laboratory Examinations (Noklus), the external quality assessment provider in Norway.)

patient medians and percentage of results outside defined cut-offs, and to aggregate the data as an EQA process. Preferably, the calculations are done in the laboratory by an automatic function either in the LIS or via a middleware or homemade solution and then transferred, preferably on a daily basis, or batch-wise (frequency as convenient). Some LIS providers and IVD manufacturers offer their customers a solution for these patient data-based transfers.<sup>90</sup>

It is likely that such systems based on patient data will be a valuable supplement to traditional EQA programs.

### Responsibility of the External Quality Assessment or Proficiency Testing Provider

The EQA/PT provider is responsible for producing programs that fulfill the goal of evaluating a measurement procedure's performance in a single laboratory to that of other laboratories, or to a true value when possible.<sup>91</sup> EQA/PT providers should strive to use commutable materials whenever possible.<sup>62</sup> The frequency of distribution and the number of EQA/PT samples must address the need of the laboratory and conform to applicable regulatory requirements.

The EQA/PT provider should have the knowledge to advise the participants when they have questions regarding their EQA/PT results. Some EQA/PT providers organize "user meetings" to address and evaluate the results of the different schemes, facilitate discussions on topics of common interest, provide a "national overview" of the performance of measurement procedures and laboratories, and develop national "expert" groups within different topics.<sup>92,93</sup>

The EQA/PT providers should communicate directly with manufacturers concerning findings related to their measurement procedures. Furthermore, they should, especially when commutable samples are used, perform postmarketing surveillance and report any deficiency that could affect patient safety to the appropriate regulatory body.<sup>94</sup>

### POINTS TO REMEMBER

#### External Quality Assessment or Proficiency Testing

- An independent external organization circulates samples with unknown target values.
- The quality of the EQA/PT sample is critical for interpretation of the results.
- When commutable, "patient-like," samples are used, a laboratory can compare its own results with results from all other measurement procedures and often with a true value from an RMP.
- When noncommutable samples are used, a laboratory can only compare its own results with results from participants using a similar measurement procedure.

### SELECTED REFERENCES

1. Horvath AR, Bossuyt PMM, Sandberg S, et al. Setting analytical performance specifications based on outcome studies—is it possible? *Clin Chem Lab Med* 2015;53:841–8.
2. Sandberg S, Fraser CG, Horvath AR, et al. Defining analytical performance specifications: consensus statement from the 1st Strategic Conference of the European Federation of Clinical Chemistry and Laboratory Medicine. *Clin Chem Lab Med* 2015;53:833–5.
3. European Federation of Clinical Chemistry and Laboratory Medicine on-line biological variation database. <<https://biologicalvariation.eu>> [accessed 2020.07.04].
4. Røraas T, Petersen PH, Sandberg S. Confidence intervals and power calculations for within-person biological variation: effect of analytical imprecision, number of replicates, number of samples, and number of individuals. *Clin Chem* 2012;58:1306–13.
5. Aarsand AK, Røraas T, Sandberg S. Biological variation—reliable data is essential. *Clin Chem Lab Med* 2015; 53:153–4.
6. Aarsand AK, Røraas TR, Fernandez-Calle P, et al. The Biological Variation Data Critical Appraisal Checklist: A Standard for Evaluating Studies on Biological Variation. *Clin Chem* 2018;64:501–14.
7. Panteghini M, Sandberg S. Defining analytical performance specifications 15 years after the Stockholm conference. *Clin Chem Lab Med* 2015;53:829–32.
8. Miller WG, Erek A, Cunningham TD, et al. Commutability limitations influence quality control results with different reagent lots. *Clin Chem* 2011;57:76–83.
9. Parvin CA. Assessing the impact of the frequency of quality control testing on the quality of reported patient results. *Clin Chem* 2008;54:2049–54.
10. CLSI. Laboratory quality control based on risk management; approved guideline EP23-A. Wayne, PA: Clinical and Laboratory Standards Institute; 2011.
11. CLSI. Statistical quality control for quantitative measurement procedures: principles and definitions; approved guideline C24-A4. Wayne, PA: Clinical and Laboratory Standards Institute; 2016.
12. CLSI. User evaluation of between-reagent lot variation; approved guideline EP26-A. Wayne, PA: Clinical and Laboratory Standards Institute; 2013 (under revision at the time of publication).
13. CLSI. Verification of comparability of patient results within one healthcare system; approved guideline EP31-A-IR. Wayne, PA: Clinical and Laboratory Standards Institute; 2012.
14. Ng D, Poliyo FA, Cervinski MA. Optimization of a Moving Averages Program Using a Simulated Annealing Algorithm: The Goal is to Monitor the Process Not the Patients. *Clin Chem* 2016;62:1361–71.
15. Bennett ST. Continuous Improvement in Continuous Quality Control. *Clin Chem* 2016;62:1299–1301.
16. Badrick T, Bietenbeck A, Cervinski MA, et al. Patient-Based Real-Time Quality Control: Review and Recommendations. *Clin Chem* 2019;65:962–71.
17. Miller WG, Jones GRD, Horowitz GL, et al. Proficiency testing/external quality assessment: current challenges and future directions. *Clin Chem* 2011;57:1670–80.
18. Stavelin A, Riksheim BO, Christensen NG, et al. The Importance of reagent lot registration in external quality assurance/proficiency testing schemes. *Clin Chem* 2016; 62:708–15.
19. Miller WG. Time to pay attention to reagent and calibrator lots for proficiency testing. *Clin Chem* 2016; 62:666–7.
20. De Grande LAC, Goossens K, Van Uytfanghe K, et al. The Empower project - a new way of assessing and monitoring test comparability and stability. *Clin Chem Lab Med* 2015; 53:1197–204.

## REFERENCES

1. Horvath AR, Bossuyt PMM, Sandberg S, et al. Setting analytical performance specifications based on outcome studies—is it possible? *Clin Chem Lab Med* 2015;53:841–8.
2. ISO 17511:2020. In vitro diagnostic medical devices—requirements for establishing metrological traceability of values assigned to calibrators, trueness control materials and human samples. ISO, Geneva, Switzerland; 2020.
3. Vesper HW, Thienpont LM. Traceability in laboratory medicine. *Clin Chem* 2009;55:1067–75.
4. Miller WG, Tate JR, Barth JH, et al. Harmonization: the sample, the measurement and the report. *Ann Lab Med* 2014;34:187–97.
5. Levey S, Jennings ER. The use of control charts in the clinical laboratory. *Am J Clin Pathol* 1950;20:1059–66.
6. Shewhart WA. Economic control of quality of manufactured product. New York: Van Nostrand; 1931.
7. CLSI. Measurement procedure comparison and bias estimation using patient samples, 3<sup>rd</sup> Edition, corrected; approved guideline EP09c. Wayne, PA: Clinical and Laboratory Standards Institute; 2018.
8. Sandberg S, Fraser CG, Horvath AR, et al. Defining analytical performance specifications: consensus statement from the 1st Strategic Conference of the European Federation of Clinical Chemistry and Laboratory Medicine. *Clin Chem Lab Med* 2015;53:833–5.
9. National Institutes of Health. Recommendations for improving cholesterol measurement: a report from the Laboratory Standardization Panel of the National Cholesterol Education Program. Bethesda, MD: National Institutes of Health; 1990. NIH Publication No: 90-2964.
10. Myers GL, Miller WG, Coresh J, et al. Recommendations for improving serum creatinine measurement: a report from the Laboratory Working Group of the National Kidney Disease Education Program. *Clin Chem* 2006;52:5–18.
11. Fraser CG. Biological variation: from principles to practice. American Association for Clinical Chemistry, ISBN 1-890883-49-2, AACC Press, 2001.
12. Lund F, Petersen PH, Fraser CG, et al. Calculation of limits for significant bidirectional changes in two or more serial results of a biomarker based on a computer simulation model. *Ann Clin Biochem* 2015;52:434–40.
13. Perich C, Minchinela J, Ricos C, et al. Biological variation database: structure and criteria used for generation and update. *Clin Chem Lab Med* 2015;53:399–405.
14. European Federation of Clinical Chemistry and Laboratory Medicine on-line biological variation database. <<https://biologicalvariation.eu>> [accessed 2020.07.04].
15. Oosterhuis WP. Gross overestimation of total allowable error based on biological variation. *Clin Chem* 2011;57:1334–6.
16. Røraas T, Petersen PH, Sandberg S. Confidence intervals and power calculations for within-person biological variation: effect of analytical imprecision, number of replicates, number of samples, and number of individuals. *Clin Chem* 2012;58:1306–13.
17. Aarsand AK, Røraas T, Sandberg S. Biological variation—reliable data is essential. *Clin Chem Lab Med* 2015;53:153–4.
18. Carobene A, Braga F, Røraas T, et al. A systematic review of data on biological variation for alanine aminotransferase, aspartate aminotransferase and g-glutamyl transferase. *Clin Chem Lab Med* 2013;51:1997–2007.
19. Marco JD-G, Fernandez-Calle P, Minchinela J, et al. Biological variation data for lipid cardiovascular risk assessment biomarkers. A systematic review applying the biological variation data critical appraisal checklist (BIVAC). *Clin Chim Acta* 2019;495:467–75.
20. Coskun F, Braga F, Carobene A, et al. Systematic review and meta-analysis of within-subject and between-subject biological variation estimates of 20 haematological parameters. *Clin Chem Lab Med* 2019;58:25–32.
21. Aarsand AK, Røraas TR, Fernandez-Calle P, et al. The Biological Variation Data Critical Appraisal Checklist: A Standard for Evaluating Studies on Biological Variation. *Clin Chem* 2018;64:501–14.
22. Panteghini M, Sandberg S. Defining analytical performance specifications 15 years after the Stockholm conference. *Clin Chem Lab Med* 2015;53:829–32.
23. Franzini C, Ceriotti F. Impact of reference materials on accuracy in clinical chemistry. *Clin Biochem* 1998;31:449–57.
24. Thienpont LM, Stockl D, Friedecky B, et al. Trueness verification in European external quality assessment schemes: time to care about the quality of the samples. *Scand J Clin Lab Invest* 2003;63:195–201.
25. Miller WG. Specimen materials, target values and commutability for external quality assessment (proficiency testing) schemes. *Clin Chim Acta* 2003;327:25–37.
26. Miller WG, Myers GL, Ashwood ER, et al. Creatinine measurement: state of the art in accuracy and inter-laboratory harmonization. *Arch Pathol Lab Med* 2005;129:297–304.
27. Miller WG, Myers GL, Ashwood ER, et al. State of the art in trueness and inter-laboratory harmonization for 10 analytes in general clinical chemistry. *Arch Pathol Lab Med* 2008;132:838–46.
28. Miller WG, Erek A, Cunningham TD, et al. Commutability limitations influence quality control results with different reagent lots. *Clin Chem* 2011;57:76–83.
29. Kristensen GBB, Christensen NG, Thue G, et al. Between-lot variation in external quality assessment of glucose: clinical importance and effect on participant performance evaluation. *Clin Chem* 2005;51:1632–6.
30. Parvin CA. Assessing the impact of the frequency of quality control testing on the quality of reported patient results. *Clin Chem* 2008;54:2049–54.
31. CLSI. Laboratory quality control based on risk management; approved guideline EP23-A. Wayne, PA: Clinical and Laboratory Standards Institute; 2011.
32. CLSI. Statistical quality control for quantitative measurement procedures: principles and definitions; approved guideline C24-A4. Wayne, PA: Clinical and Laboratory Standards Institute; 2016.
33. CLSI. Evaluation of precision of quantitative measurement procedures; approved guideline EP05-A3. Wayne, PA: Clinical and Laboratory Standards Institute; 2014.
34. Ellis AD, Gross AR, Budd JR, Miller WG. Influence of reagent lots and multiple measuring systems on estimating the coefficient of variation from quality control data; implications for uncertainty estimation and interpretation of QC results. *Clin Chem Lab Med* 2020. DOI: <https://doi.org/10.1515/cclm-2020-0320> | Published online: 28 Apr 2020.
35. Westgard JO, Barry PL, Hunt MR. A multi-rule Shewhart chart for quality control in clinical chemistry. *Clin Chem* 1981;27:493–501.
36. Ryan TP. Statistical measurement procedures for quality control. New York: John Wiley & Sons; 1989.

37. Westgard JO, Groth T. Power functions for statistical control rules. *Clin Chem* 1979;25:863–9.
38. Parvin CA. Quality-control (QC) performance measures and the QC planning process. *Clin Chem* 1997;43:602–7.
39. Parvin CA, Gronowski AM. Effect of analytical run length on quality-control (QC) performance and the QC planning process. *Clin Chem* 1997;43:2149–54.
40. Linnet K. Choosing quality-control systems to detect maximum clinically allowable errors. *Clin Chem* 1989;35:284–8.
41. Yundt-Pacheco J, Parvin CA. Validating the performance of QC procedures. *Clin Lab Med* 2013;33:75–88.
42. ISO 22870:2016. Point-of-care testing (POCT)—requirements for quality and competence. ISO, Geneva, Switzerland, 2016.
43. Farrant I. *Review policies, procedures and guidelines for point-of-care testing. Report from RCPA Quality Assurance Programs*, 2012;1–63. <<http://www.aacb.asn.au/documents/item/635>> [accessed 01.25.20].
44. C. P. Price, I. Smith, and A. Van den Bruel, “Improving the quality of point-of-care testing,” *Fam Pract*, vol. 35, no. 4, pp. 358–364. 2018
45. R. Huddy, M. Ni, S. Misra, S. Mavroveli, J. Barlow, and G. B. Hanna, “Development of the Point-of-Care Key Evidence Tool (POCKET): a checklist for multi-dimensional evidence generation in point-of-care tests,” *Clin Chem Lab Med* 2019;57:845–55.
46. Martin CL. Quality control issues in point of care testing. *Clin Biochem Rev* 2008;29(Suppl. 1):S79–82.
47. Lewandrowski K, Gregory K, Macmillan D. Assuring quality in point-of-care testing: evolution of technologies, informatics, and program management. *Arch Pathol Lab Med* 2011;135:1405–14.
48. Stavelin A, Petersen PH, Solvik U, et al. Internal quality control of prothrombin time in primary care: comparing the use of patient split samples with lyophilised control materials. *Thromb Haemost* 2009;102:593–600.
49. Jani IV, Peter TF. How point-of-care testing could drive innovation in global health. *N Engl J Med* 2013;368:2319–24.
50. CLSI. User evaluation of between-reagent lot variation; approved guideline EP26-A. Wayne, PA: Clinical and Laboratory Standards Institute; 2013 (under revision at the time of publication).
51. Kazmierczak SC. Laboratory quality control: using patient data to assess analytical performance. *Clin Chem Lab Med* 2003;41:617–27.
52. CLSI. Delta checks; approved guideline EP33-A. Wayne, PA: Clinical and Laboratory Standards Institute; 2016.
53. CLSI. Verification of comparability of patient results within one healthcare system; approved guideline EP31-A-IR. Wayne, PA: Clinical and Laboratory Standards Institute; 2012.
54. Westgard JO, Smith FA, Mountain PJ, et al. Design and assessment of average of normals (AON) patient data algorithms to maximize run lengths for automatic process control. *Clin Chem* 1996;42:1683–8.
55. Smith FA, Kroft SH. Exponentially adjusted moving mean procedure for quality control. An optimized patient sample control procedure. *Am J Clin Pathol* 1996;105:44–51.
56. Cembrowski GS, Chandler EP, Westgard JO. Assessment of “average of normals” quality control procedures and guidelines for implementation. *Am J Clin Pathol* 1984;81:492–9.
57. Ye JJ, Ingels SC, Parvin CA. Performance evaluation and planning for patient-based quality control procedures. *Am J Clin Pathol* 2000;113:240–8.
58. Ng D, Poliyo FA, Cervinski MA. Optimization of a Moving Averages Program Using a Simulated Annealing Algorithm: The Goal is to Monitor the Process Not the Patients. *Clin Chem* 2016;62:1361–71.
59. Bennett ST. Continuous Improvement in Continuous Quality Control. *Clin Chem* 2016;62:1299–1301.
60. Badrick T, Bietenbeck A, Cervinski MA, et al. Patient-Based Real-Time Quality Control: Review and Recommendations. *Clin Chem* 2019;65:962–71.
61. Lunetzky ES, Cembrowski GS. Performance characteristics of Bull’s multirule algorithm for the quality control of multichannel hematology analyzers. *Am J Clin Pathol* 1987;88:634–8.
62. Miller WG, Jones GRD, Horowitz GL, et al. Proficiency testing/external quality assessment: current challenges and future directions. *Clin Chem* 2011;57:1670–80.
63. CLSI. Using Proficiency Testing and Alternative Assessment to Improve Medical Laboratory Quality, 3rd Edition, QMS24. Wayne, PA: Clinical and Laboratory Standards Institute; 2016.
64. Miller WG, Schimmel H, Rej R, et al. IFCC Working Group Recommendations for Assessing Commutability Part 1: General Experimental Design. *Clin Chem* 2018;64:447–54.
65. Miller WG, Myers GL, Gantzer ML, et al. Roadmap for harmonization of clinical laboratory measurement procedures. *Clin Chem* 2011;57:1108–17.
66. Little RR, Rohlfing CL, Sacks DB. Status of hemoglobin A1c measurement and goals for improvement: from chaos to order for improving diabetes care. *Clin Chem* 2011;57:205–14.
67. Solvik UØ, Røraas T, Christensen NG, et al. Diagnosing diabetes mellitus: performance of hemoglobin A1c point-of-care instruments in general practice offices. *Clin Chem* 2013;59:1790–801.
68. Stavelin A, Petersen PH, Solvik UØ, et al. External quality assessment of point-of-care methods: model for combined assessment of method bias and single-participant performance by the use of native patient samples and noncommutable control materials. *Clin Chem* 2013;59:363–71.
69. CLSI. Characterization and qualification of commutable reference materials for laboratory medicine; approved guideline EP30-A. Wayne, PA: Clinical and Laboratory Standards Institute; 2010.
70. CLSI. Evaluation of commutability of processed samples; approved guideline EP14-A3. Wayne, PA: Clinical and Laboratory Standards Institute; 2014.
71. Nilsson G, Budd JR, Greenberg N, et al. IFCC Working Group Recommendations for Assessing Commutability Part 2: Using the Difference in Bias Between a Reference Material and Clinical Samples. *Clin Chem* 2018;64:455–64.
72. Budd JR, Weykamp C, Rej R, et al. IFCC Working Group Recommendations for Assessing Commutability Part 3: Using the Calibration Effectiveness of a Reference Material. *Clin Chem* 2018;64:465–74.
73. Ross JW, Miller WG, Myers GL, et al. The accuracy of laboratory measurements in clinical chemistry. A study of 11 routine chemistry analytes in the College of American Pathologists chemistry survey with fresh frozen serum, definitive measurement procedures, and reference measurement procedures. *Arch Pathol Lab Med* 1998;122:587–608.
74. Stavelin A, Riksheim BO, Christensen NG, et al. The Importance of reagent lot registration in external quality assurance/proficiency testing schemes. *Clin Chem* 2016;62:708–15.
75. Miller WG. Time to pay attention to reagent and calibrator lots for proficiency testing. *Clin Chem* 2016;62:666–7.

76. Naito HK, Kwak YS, Hartfiel JL, 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* 1993;117:345–51.
77. Ellerbe P, Myers GL, Cooper GR, et al. Comparison of results for cholesterol in human serum obtained by the reference measurement procedure and by the definitive measurement procedure of the National Reference System for cholesterol. *Clin Chem* 1990;36:370–5.
78. Kroll MH, Chesler R. Effect of serum lyophilization on the rate constants of enzymatic measurement procedures for measuring cholesterol. *Clin Chem* 1990;36:534–7.
79. Seneca S, Morris MA, Patton S, et al. Experience and outcome of 3 years of a European EQA scheme for genetic testing of the spinocerebellar ataxias. *Eur J Hum Genet* 2008;16:913–20.
80. Kalman LV, Lubin IM, Barker S, et al. Current landscape and new paradigms of proficiency testing and external quality assessment for molecular genetics. *Arch Pathol Lab Med* 2013;137:983–8.
81. Petersen PH, Christensen NG, Sandberg S, et al. How to deal with semi-quantitative tests? Application of an ordinal scale model to measurements of urine glucose. *Scand J Clin Lab Invest* 2009;69:662–72.
82. Petersen HP, Christensen GN, Sandberg S, et al. How to deal with dichotomous tests? Application of a rankit ordinal scale model with examples from the Nordic ordinal scale project on screening tests. *Scand J Clin Lab Invest* 2008;68:298–311.
83. Petersen PH, Sandberg S, Fraser CG, et al. A model for setting analytical quality specifications and design of control for measurements on the ordinal scale. *Clin Chem Lab Med* 2000;38:545–51.
84. Nordin G. Before defining performance criteria we must agree on what a “qualitative test procedure” is. *Clin Chem Lab Med* 2015;53:939–41.
85. Seneca S, Morris MA, Patton S, et al. Experience and outcome of 3 years of a European EQA scheme for genetic testing of the spinocerebellar ataxias. *Eur J Hum Genet* 2008;16:913–20.
86. Aarsand AK, Villanger JOH, Stole E, et al. European specialist porphyria laboratories: diagnostic strategies, analytical quality, clinical interpretation, and reporting as assessed by an external quality assurance program. *Clin Chem* 2011;57:1514–23.
87. Weykamp C, Secchiero S, Plebani M, et al. Analytical performance of 17 general chemistry analytes across countries and across manufacturers in the INPUTS project of EQA organizers in Italy, the Netherlands, Portugal, United Kingdom and Spain. *Clin Chem Lab Med* 2017;55:203–11.
88. Carlsen S, Thue G, Cooper JG, et al. Benchmarking by HbA1c in a national diabetes quality register - does measurement bias matter? *Clin Chem Lab Med* 2015;53:1433–9.
89. De Grande LAC, Goossens K, Van Uytfanghe K, et al. The Empower project - a new way of assessing and monitoring test comparability and stability. *Clin Chem Lab Med* 2015;53:1197–204.
90. [https://www.noklus.no/media/qzrl1thn/26\\_introduction-to-percentiler-and-flagger-programs.pdf](https://www.noklus.no/media/qzrl1thn/26_introduction-to-percentiler-and-flagger-programs.pdf) [accessed 2020.07.07].
91. ISO/IEC 17043:2010. Conformity assessment—general requirements for proficiency testing. ISO, Geneva, Switzerland; 2010.
92. Stavelin A, Meijer P, Kitchen D, et al. External quality assessment of point-of-care international normalized ratio (INR) testing in Europe. *Clin Chem Lab Med* 2012;50:81–8.
93. United Kingdom National External Quality Assessment Service. (UK NEQAS) Participant Meetings. <<https://ukneqas.org.uk/events/>> [accessed 2020.07.04].
94. World Health Organization. *Guidance for post-market surveillance of in vitro diagnostics, version 5*, 19 January 2015. <[http://www.who.int/diagnostics\\_laboratory/postmarket/150210\\_pms\\_ivds\\_guidance.pdf?ua](http://www.who.int/diagnostics_laboratory/postmarket/150210_pms_ivds_guidance.pdf?ua)> [accessed 2020.07.04].

## MULTIPLE CHOICE QUESTIONS

1. Why should internal QC be performed?
  - a. To be sure that the QC material meets specifications
  - b. To examine if the control material is commutable (behaves like patient samples)
  - c. To have a high probability that correct patient results are released
  - d. To be able to pass the accreditation inspection
  - e. To examine if my measurement procedure gives results similar to other laboratories
2. How should I interpret results when an EQA/PT program uses noncommutable materials?
  - a. When my result is within the performance specifications, I can be confident that I have no bias compared with a true value.
  - b. When my result is different than the all method mean, I have to recalibrate my measurement procedure.
  - c. When my result is close to the target value for the peers in my measurement procedure group, I can be confident my laboratory is performing as well as my peers.
  - d. I should compare my results with results from other measurement procedures to be confident my laboratory is not biased.
  - e. I should compare my results with the average of all measurement procedure groups.
3. What is the advantage of using commutable QC materials?
  - a. They are similar to patient samples and should therefore not be used for QC.
  - b. They are suitable for internal QC but not for EQA and proficiency testing.
  - c. They should be avoided because they are contagious.
  - d. Their results provide information on the accuracy for patient samples if the target value is set by an RMP.
  - e. Their target values are always assigned by RMPs.
4. How can a reagent lot change affect a QC target value?
  - a. QC target values are not affected by reagent lot changes.
  - b. The QC target should be used to verify acceptability of a new reagent lot.
  - c. The non-commutability bias can be different, so the QC target value may need to be adjusted.
  - d. The non-commutability bias may be different, so the SD may need to be adjusted.
  - e. The reagent lot should be rejected if the QC target value is not recovered.
5. How is the SD estimated for an internal QC material?
  - a. From measurements made during the target value assignment of a QC material
  - b. From the long-term SD that includes most types of variability expected to influence the measurement procedure
  - c. From the instructions for use provided by the measurement procedure manufacturer
  - d. From reports of interlaboratory summaries
  - e. From the range of acceptable results provided by the QC material manufacturer
6. How frequently should QC samples be measured?
  - a. Based on the stability of a measurement procedure and the risk of harm from a potentially erroneous result being reported
  - b. Based on the stability of a measurement procedure
  - c. Based on the magnitude of the SD used for evaluating QC results
  - d. Every 24 hours
  - e. Whenever a result is suspicious
7. How should a target value be established for a QC sample?
  - a. As the mean of the first 25 results
  - b. As the mean of the first 10 results
  - c. As the mean from interlaboratory comparison data
  - d. As the mean of at least 10 results and updated when more results are available
  - e. As the standard error of the mean of 10 results
8. What is the first thing to do when a QC result fails an evaluation rule?
  - a. Repeat the QC sample
  - b. Recalibrate then repeat the QC sample
  - c. Check if the previous QC result was acceptable or not
  - d. Check if the result from a different QC sample measured at the same time was acceptable or not
  - e. Stop reporting results for patient samples
9. What are the key attributes of a rule used to interpret QC results?
  - a. The rule should identify a QC result that has a 95% probability of being incorrect.
  - b. The rule should identify a bias that exceeds the manufacturer's specification for the measurement procedure.
  - c. The rule should identify either a bias or an imprecision that exceeds the manufacturer's specification for the measurement procedure.
  - d. The rule should identify an error condition that is large enough to increase the risk of an erroneous medical decision based on results for the test.
  - e. The rule should identify when a measurement procedure is at risk to produce an erroneous result.
10. The moving average of patient sample results is useful in which of the following situations?
  - a. When there are a large number of results generated in a short time interval
  - b. When a physiologically homogeneous population of patients can be identified
  - c. When the cost of QC samples is very expensive
  - d. When the measurement procedure is very stable after long time intervals
  - e. When results can vary within a patient over relatively short time intervals.

# Standardization and Harmonization of Analytical Examination Results\*

*W. Greg Miller*

## ABSTRACT

### Background

The purpose of a clinical laboratory test is to provide information on the pathophysiologic condition of an individual patient to assist with diagnosis, to guide or monitor therapy, or to assess risk for a disease. Results for the same measurand must be equivalent when measured using different measurement procedures (MPs) to avoid medical errors when using clinical decision values to interpret those results. Standardization or harmonization of results is accomplished by metrological traceability of calibration to the same reference system and by MPs having adequate selectivity for the measurand being measured.

### Content

The purpose of a clinical laboratory test is to provide information on the pathophysiologic condition of an individual patient to assist with diagnosis, to guide or monitor therapy, or to assess risk for a disease. Results for the same measurand must be equivalent when measured using different MPs to avoid medical errors when using clinical decision values to interpret those results. Standardization or harmonization of results is accomplished by metrological traceability of calibration to the same reference system and by MPs having adequate selectivity for the measurand being measured.

The International Organization for Standardization (ISO) has published a series of standards that describe requirements for metrological traceability of results for patients' samples to higher-order references, including a harmonization protocol, for reference materials (RMs) and reference MPs used in metrological traceability, and for calibration laboratories that offer reference measurement services. The Joint Committee for Traceability in Laboratory Medicine reviews and approves RMs and reference MPs that conform to the ISO standards.

In vitro diagnostic manufacturers of end-user MPs used in clinical laboratories, including clinical laboratories that develop test procedures, establish metrological traceability to the available higher-order reference system for a measurand. When metrological traceability is successful, results for patients' samples agree among different MPs. Important limitations in achieving harmonized results are that higher-order references are available for only a little over 100 measurands and some matrix-based RMs in use are not commutable with patients' samples, which causes disagreement among results from different MPs. External quality assessment or proficiency testing using commutable samples is an important procedure to monitor the success of harmonization and provide feedback to identify measurands that need better harmonization of results.

\*The full version of this chapter is available electronically on [ExpertConsult.com](http://ExpertConsult.com).

## INTRODUCTION

The purpose of a clinical laboratory test is to provide information on the pathophysiologic condition of an individual patient to assist with diagnosis, to guide or monitor therapy, or to assess risk for a disease. Results from different measurement procedures (MPs) for the same measurand must be equivalent within a total allowable error consistent with an acceptable risk of harm from decisions based on a test result. Equivalent results are essential when using clinical guideline decision values to make medical decisions for patient care based on those results.<sup>1</sup> For example, Almond and colleagues reported that results for parathyroid hormone (PTH) differed approximately fourfold among five clinical laboratory MPs.<sup>2</sup> Using guidelines from the UK Renal Association for treating hypophosphatemia in kidney disease, these differences in PTH results altered drug treatment decisions for one-half of the patients in the investigation. In principle, using reference intervals rather than decision values to interpret such results could improve outcomes. However, for these PTH MPs, the reference intervals were very similar despite a fourfold difference in results suggesting that determining adequate reference intervals is challenging.

Equivalent results among different MPs can be achieved by having the calibration of all clinical laboratory MPs traceable to the same higher-order reference system, and having all MPs measure the same measurand without influence from other molecules present in a patient's sample.

## Regulations

Although standardization and harmonization of test results have been important goals in laboratory medicine for many decades, there have been few regulations requiring calibration of clinical laboratory MPs to be traceable to higher-order references. One of the first regulations was a directive passed by the European Commission in 1998 that by 2003 all in vitro diagnostic (IVD) devices sold in the European Union were required whenever possible to have calibration traceable to a higher-order reference system.<sup>3</sup> The directive was replaced by a regulation in 2017, effective 2022, with essentially the same requirements but adding a formal review and approval process.<sup>4</sup> 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. Other countries are introducing regulations requiring calibration hierarchies that are metrologically traceable to higher-order reference systems. In addition, many countries have regulations and a regulatory review agency that approve all medical devices including clinical laboratory MPs as safe and effective before they can be sold. Unfortunately, not all regulatory agencies require metrological traceability to approved reference systems when available. Consequently, harmonization of regulatory approval procedures in different countries would contribute to improved harmonization of results among different MPs for the same measurand.

## Terminology

The term *analyte* refers to the name of the substance being analyzed or measured. The term "measurand" refers to the

quantity intended to be measured or the *quantity* subject to measurement in a stated matrix where *quantity* means the property of the molecular substance being measured. For example, in the phrase "mass of creatinine in blood serum," the analyte is creatinine, the measurand is creatinine in blood serum, and the quantity being measured is the mass of creatinine. If we measured the mass of creatinine in urine, the analyte and quantity are the same, but the measurand is different because the sample matrix is now urine rather than blood serum. The measurand can be difficult to specify for complex molecules such as proteins. For example, when measuring the "mass of albumin in urine" using a mass spectrometry MP after trypsin digestion, the analyte is albumin, the measurand defined as the quantity intended to be measured is albumin in urine, but the quantity actually measured is a specific trypsin amino acid fragment presumably derived only from the intact albumin in the urine. In clinical laboratory medicine, we frequently use the terms *analyte* and *measurand* colloquially and the reader or listener must infer what is the correct measurand from the context of the usage.

Various terms are used to describe a *measurement procedure* in different contexts. An MP is a written description of a measurement process including reagents, calibrators, equipment, software, procedure for calibration, calibration hierarchy, etc., and how measurements are made on specified samples using these items. A *measuring system* is the physical embodiment of an MP that is used by a laboratory to make measurements on a physical sample. A *measuring system* is also called an IVD *medical device* developed by an IVD manufacturer for use by a clinical laboratory. Note that an IVD *measuring system* may be produced by a commercial manufacturer or by a clinical laboratory for its own use when it is also called a laboratory developed test. In this chapter, the abbreviation MP is used to refer to a written description of a MP or to a *measuring system* used to make measurements on samples. The intended meaning is clear from the context.

The term *method* is typically used to describe the technology or measurement principle used in an MP; examples include ion-selective electrode, kinetic spectrophotometry, mass spectrometry. The term *method* is sometimes used to mean a specific MP or physical *measuring system* based on that MP used in a laboratory. The term *assay* is used similarly to *method* to mean either a type of measurement principle or a specific MP or physical *measuring system* used in a laboratory. The terms *assay* and *method* are vague and not used in this chapter.

The terms *standardization* and *harmonization* (or *standardized* and *harmonized*) are frequently used interchangeably to refer to achieving equivalent results, within clinically acceptable limits, for patients' samples measured using different clinical laboratory MPs. The term *standardization* has traditionally been used when calibration is metrologically traceable to a certified reference material (CRM) and/or a reference measurement procedure (RMP) as described later. The ISO standard 17511:2020 uses the term *harmonization* in the context of an international harmonization protocol to achieve equivalent results based on a consensus approach for metrological traceability to harmonization reference materials (RMs) as described later.

## POINTS TO REMEMBER

### Standardization and Harmonization

- Equivalent results from different measurement procedures are necessary to interpret laboratory results using clinical practice guidelines.
- Equivalent results are achieved by calibration of all measurement procedures traceable to the same higher-order reference system and by all measurement procedures having suitable selectivity for the measurand.

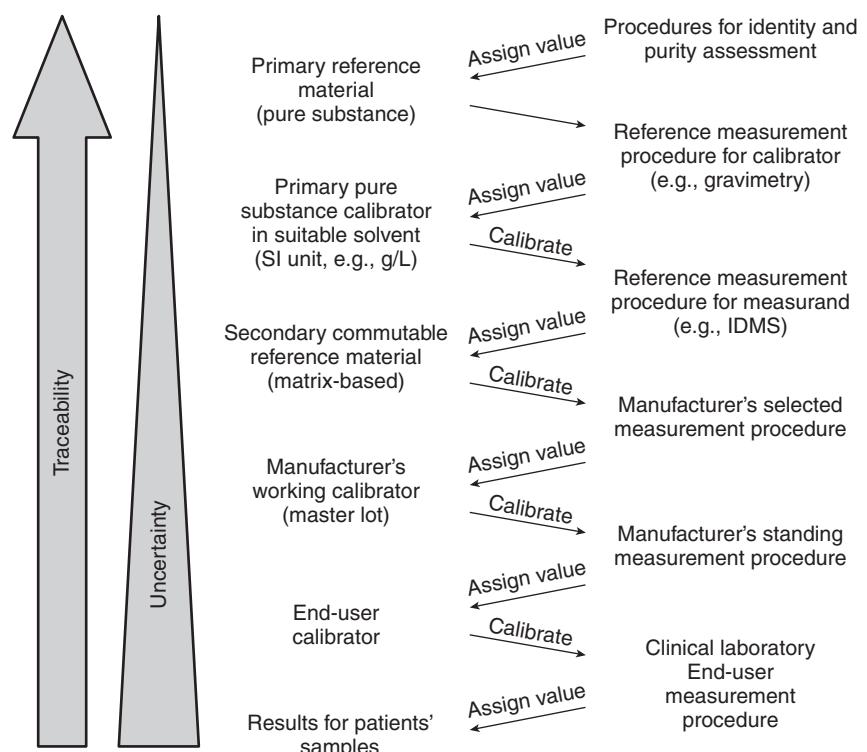
## METROLOGICAL TRACEABILITY TO A REFERENCE SYSTEM

Whenever possible, calibration of clinical laboratory end-user MPs should be metrologically traceable to a higher-order CRM and/or RMP.<sup>1-7</sup> Metrology is the science of making measurements. Metrological traceability (Fig. 7.1) means establishing calibration of a clinical laboratory end-user MP by an unbroken chain of metrological traceability steps to the best available CRM or RMP, called higher-order reference system components. Fig. 7.1 is based on the ISO standard 17511:2020<sup>5</sup> that specifies requirements for establishing metrological traceability of values assigned to calibrators, trueness control materials, and human samples and is referred to as a complete reference system because all components are available. Achieving standardized results using a complete

metrological traceability system is desirable because the pure substance CRM and the RMP can be reproduced in different locations and times in the future. Consequently, the reference system provides a stable and reproducible calibration hierarchy for use as needed by end-user MP producers.

A complete reference system provides traceability of the results for patient samples from an end-user MP used in a clinical laboratory to the Système Internationale (SI) unit based on a series of calibrations that link the end-user calibrator to a higher-order pure substance CRM or RMP for measurands defined by the RMP. A CRM is an RM accompanied by a certificate issued by an authoritative body that provides property values, such as mass fraction or concentration, with associated uncertainties and traceabilities. Authoritative bodies are typically national metrology institutes or designated institutes with expertise to develop and produce CRMs and RMPs. Examples of metrology institutes include the National Institute for Standards and Technology in the United States and the Joint Research Center in the European Union. A list of national metrology institutes is available at the listing of members of the Consultative Committee for Amount of Substance: Metrology in Chemistry and Biology.<sup>8</sup>

The highest order pure substance calibrator in the traceability chain is prepared from a well-characterized pure substance primary CRM using a primary RMP such as gravimetry to establish the SI unit for the analyte in a well-defined solution that is suitable for use as a calibrator for the next RMP in the hierarchy, for example one based on isotope dilution mass



**FIGURE 7.1** Metrological traceability of calibration of end-user measurement procedures based on the International Organization for Standardization standard 17511:2020. *SI*, Système Internationale; *IDMS*, isotope dilution mass spectrometry. (Adapted from ISO 17511:2020. *In vitro diagnostic medical devices—Requirements for establishing metrological traceability of values assigned to calibrators, trueness control materials and human samples*. 2nd ed. Geneva, Switzerland: International Organization for Standardization; 2020.)

spectrometry. An RMP for the measurand is one with high selectivity for the measurand, acceptable imprecision, and minimal influence from sample matrix components. ISO 17511:2020 also describes metrological traceability when the highest component is the RMP itself that defines the measurand and for cases such as enzyme activity or coagulation factors. In such cases, the measurand is specified by the measurement conditions.

The RMP for the measurand is then used to assign values for a measurand to a secondary commutable RM, frequently a CRM, that typically has a matrix similar or identical to that of clinical patients' samples. A measurand is the substance, called a quantity in metrology, intended to be measured and includes the analyte name with the sample type and specification of the molecular substance intended to be measured.<sup>9</sup> When a suitable matrix-based RM is not available, a panel of clinical patients' samples can be used as RMs at this point in the metrological traceability chain.

A special category called an international conventional RMP or RM is recognized in ISO 17511:2020. An international conventional RMP is one that gives values not metrologically traceable to SI but which by international agreement are used as reference values for a defined measurand. An international conventional RM is referred to as an international conventional calibrator or calibration material whose value is not metrologically traceable to the SI but is assigned by international agreement. An international conventional RMP or RM can be used in the positions of the RMP for the measurand or the secondary commutable RM in a metrological traceability chain.

The secondary commutable RM is then typically used as a calibrator for a manufacturer's internal selected MP that is used to assign values to the manufacturer's working calibrator, frequently called a *master lot of calibrator*, that is used in the manufacturing process to assign values to the end-user calibrators that are used to calibrate end-user MPs used by clinical laboratories. The series of metrological traceability steps are referred to as the calibration hierarchy for the end-user MP.

Note that the manufacturer's selected and standing MPs may be different but are frequently the same and used to fulfill two purposes in the calibration hierarchy. A selected MP is used to transfer values from the secondary commutable RM to the manufacturer's master lot of working calibrator. The standing MP transfers values from the working calibrator, typically used for many years, to many lots of end-user calibrator distributed to clinical laboratories. Depending on manufacturing considerations, the step using the working calibrator and standing MP could be eliminated from the calibration hierarchy. The manufacturer's selected, standing, and end-user MPs may also use the same measurement principle and measuring system in which case the measuring systems used in the selected and standing positions will be operated with different protocols for calibration and replication to reduce the uncertainty of the result. The same metrological traceability steps are applicable when laboratories develop MPs for their own use in which case the laboratory is the manufacturer and the steps assigned to manufacturers are the responsibility of the laboratory that develops an end-user MP.

### Uncertainty

The values assigned at each step in the metrological traceability sequence have an uncertainty starting with the mass

fraction of the analyte in the primary pure substance RM. The uncertainty of a value assigned at each step adds cumulatively to that of preceding steps to have a final combined uncertainty of the results for patients' samples measured in clinical laboratories. The standard uncertainty at each step is usually determined as the SD for each value assignment procedure. The combined standard uncertainty is calculated as follows:

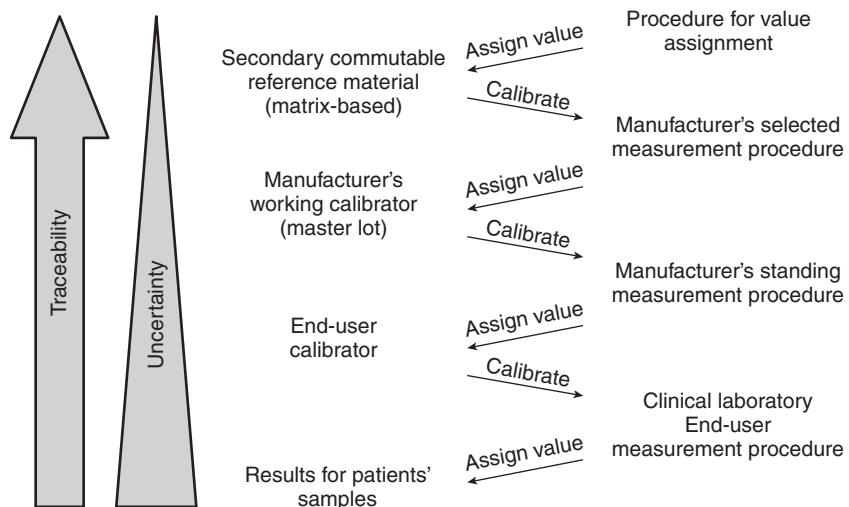
$$U_{\text{combined}} = \sqrt{u_1^2 + u_2^2 + \dots + u_i^2} \quad (7.1)$$

where lower case  $u$  is the uncertainty at a step and  $i$  denotes a discrete metrological traceability step. The expanded uncertainty, upper case  $U$ , is frequently reported as  $U = u \times 2$  where 2 is the coverage factor,  $k$ , so the uncertainty interval has approximately 95% probability to include the true measured or assigned value. The performance specifications for MPs in the calibration hierarchy are defined to contribute small enough uncertainties to ensure the final combined uncertainty for patient sample results is acceptable for making medical decisions based on those results. Approaches to determine uncertainty of a result are described in Chapter 2, and approaches to determine the allowable total measurement error for end-user MPs are described in Chapter 8. Braga and Panteghini have described an approach to allocate fractions of the allowable uncertainty in a patient's sample result to the different steps in the metrological traceability chain.<sup>10</sup>

### Metrological Traceability When Pure Substance Certified Reference Materials or Reference Measurement Procedures Are Not Available

Unfortunately, for technical reasons, pure substance RMs and RMPs are not available for a large number of the tests offered by clinical laboratories (see section on JCTLM). The ISO 17511:2020 standard describes cases when various reference system components are not available for a measurand. When there is no pure substance RM or RMP for value assignment of matrix-based RMs, metrological traceability can end with a matrix-based secondary commutable RM as shown in Fig. 7.2. In this situation, a matrix-based RM is produced and a value assigned for a measurand. Since there is no RMP, value assignment is done using alternative analytical approaches such as the all-MP mean for MPs that meet defined performance specifications, for example, parallel responses over the measuring interval, selectivity for the measurand and precision that are suitable for value assignment.

For example, ERM-DA470k/IFCC from the Joint Research Center of the European Commission has values assigned for 13 proteins in human serum. The values were assigned for 12 measurands by transfer from the predecessor CRM, ERM-DA470, using results from 8 clinical laboratory end-user immunoassay MPs operated by 22 different laboratories (not all measurements were made by all laboratories).<sup>11</sup> The value transfer protocols included procedures to assess quality and comparability of the results among the different MPs used. Overall, this approach can be considered a consensus value assignment based on comparisons of results for the new and old CRM from carefully qualified end-user MPs. One of the proteins, beta-2-microglobulin, was added gravimetrically as a purified protein with mass determined by amino acid analysis and confirmed by dry mass determination that established traceability to the SI for this measurand.<sup>12</sup> The approach to value assignment for ERM-DA470k/IFCC provided continuity



**FIGURE 7.2** Metrological traceability of calibration of end-user measurement procedures when the calibration hierarchy ends at a secondary commutable reference material. (Adapted from the International Organization for Standardization standard 21151:2020.)

with the preceding CRM that itself was value assigned by a combination of analytical procedures including value transfer from its predecessor CRM and procedures that established traceability to the SI for some measurands (see reference 12 for additional information).

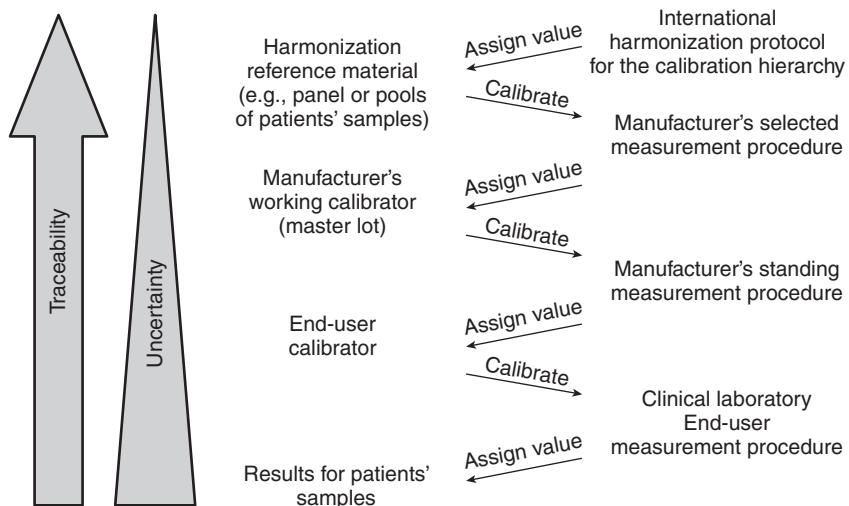
### Metrological Traceability When a Matrix-Based Reference Material Is Prepared by Dilution of Another Certified Reference Material

Pure substance CRMs are always diluted for use into a suitable matrix for the MP in the metrological traceability position for which they will be used as calibrators. Some matrix-based CRMs have high concentrations and are diluted into a suitable biological matrix to prepare one or more calibrators for the manufacturer's selected or standing MP. Values are assigned based on the dilution ratios and

dilutions are usually performed gravimetrically to minimize the uncertainty of the values assigned. The choice of matrix for dilutions is critical because it will affect the commutability of the diluted CRM that must be verified for each dilution (see section on commutability).

### Metrological Traceability When There Is No Primary or Matrix-Based Certified Reference Material or Reference Measurement Procedure: Harmonization Protocol

The ISO 17511:2020 standard describes metrological traceability using an international harmonization protocol when there is no primary or matrix-based secondary CRM or RMP available for a measurand. Fig. 7.3 shows an international harmonization protocol as the highest order component in a metrological traceability chain. ISO published a new



**FIGURE 7.3** Metrological traceability of calibration of end-user measurement procedures when the calibration hierarchy ends at an international harmonization protocol for the calibration hierarchy. (Adapted from the International Organization for Standardization standard 21151:2020.)

**BOX 7.1 Critical Steps in a Harmonization Protocol Described in ISO 21151:2020**

1. How to prepare harmonization RMs. Harmonization RMs will typically be a panel of human samples, or pools of human samples, for which details regarding clinical characteristics of sample donors, restrictions on known interfering substances, concentrations, preparation, and storage conditions are specified. Validation of the commutability of a representative subset of harmonization RMs is typically required unless the harmonization RMs are individual patients' samples measured using the same pre-examination conditions used for clinical samples.
2. How to value assign the harmonization RMs. Value assignment can be a consensus approach similar to that described in the preceding section for matrix-based CRMs when no pure substance CRM or RMP is available. The performance characteristics of the MPs used for value assignment are specified in the protocol.
3. How IVD manufacturers of end-user MPs use the harmonization RMs in their calibration hierarchies. Each IVD manufacturer develops a bias correction based on results for the harmonization RMs and applies the correction as a step in their calibration hierarchy to achieve results for patients' samples that are equivalent to results measured with other MPs.
4. How to sustain the harmonization of results over extended time intervals. For example, a second reserve panel of harmonization RMs could be prepared at the same time as the original set, stored under stable conditions, and used to verify that harmonization was achieved in step 3 or to verify that harmonization was sustained over time. The protocol also specifies how replacement harmonization RMs can be prepared that will be consistent with the original materials. Consequently, the specifications in steps 1 and 2 must be thoroughly documented to enable the harmonization protocol to be sustained. Reserve and replacement harmonization RMs can be used to bring new end-user MPs into the harmonization scheme.
5. How to monitor the harmonization condition over time. External quality assessment using commutable samples as described in Chapter 6 is one approach. Monitoring harmonization using consistency of results from patients' samples in different clinical laboratories using different end-user MPs is another possibility also described in Chapter 6. The harmonization protocol must be sufficiently detailed that it can be repeated when there is evidence that harmonization has deteriorated.

CRM, Certified reference material; IVD, in vitro diagnostic; MP, measurement procedure; RM, reference material; RMP, reference measurement procedure.

standard 21151:2020 that specifies requirements for international harmonization protocols that establish metrological traceability of values assigned to calibrators and human samples with a worked example in its Annex.<sup>13</sup> Box 7.1 shows the critical steps in a harmonization protocol applied to the calibration hierarchies of end-user MPs to achieve equivalent results for patients' samples measured using any of the end-user MPs. The scientific basis for applying a harmonization protocol to achieve equivalent results among different clinical laboratory end-user MPs has evolved in the literature.<sup>14–20</sup> Fig. 7.4 from the ISO 21151:2020 standard shows how the sequence of steps are implemented for a harmonization protocol.

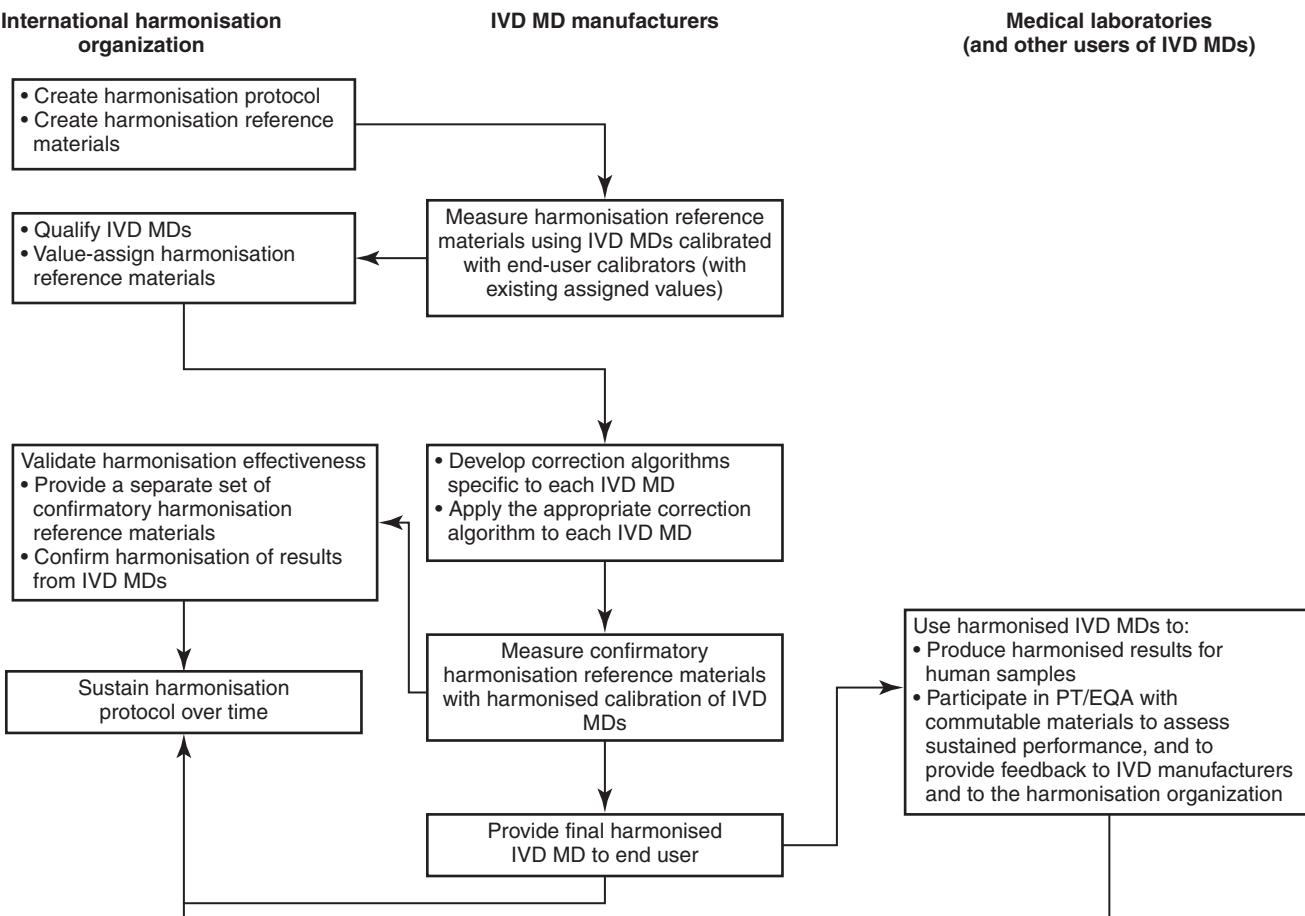
### Metrological Traceability When the End-User Measurement Procedure Manufacturer Provides a Reference Material

The lowest level of metrological traceability described in ISO 17511:2020 is when none of the higher-order calibration hierarchies described above exist. In these situations, the calibration hierarchy ends with the manufacturer's or laboratory's working calibrator. Unfortunately, each producer of an end-user MP selects what they feel is a suitable working calibrator but there is no coordination regarding materials or analytical properties of those materials used as calibrators. Although metrological traceability can be described, results for patients' samples frequently differ depending on the end-user MP used or the laboratory performing the measurements. End-user MPs should be selected that use the highest-order calibration hierarchy available for a measurand.

### POINTS TO REMEMBER

#### *Metrological Traceability to a Reference System*

- Metrological traceability of calibration of an end-user measurement procedure to a higher-order reference system enables results from different measurement procedures to have equivalent results within limits suitable for making medical decisions.
- Pure substance primary certified reference materials are used to prepare calibrators for a reference measurement procedure for the measurand.
- The reference measurement procedure for the measurand assigns values to matrix-based secondary commutable certified or other reference materials that are used to calibrate in vitro diagnostic manufacturers' procedures to value assign the end-user calibrator used by clinical laboratories.
- When a pure substance primary certified reference material or a reference measurement procedure for the measurand is not available, secondary commutable reference materials can be prepared and a value assigned using a consensus approach.
- A secondary commutable reference material must be commutable with patients' samples to be used as a calibrator; otherwise results for patients' samples will not agree among different end-user measurement procedures.
- When a secondary commutable reference material and a reference measurement procedure for the measurand is not available, an international harmonization protocol can be used to achieve harmonized results among different measurement procedures.



**FIGURE 7.4** Flowchart for steps in a harmonization protocol. *IVD MD*, In vitro diagnostic medical device; *PT/EQA*, proficiency testing/external quality assessment. (Used with permission from ISO 21151:2020.)

## THE 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 International Federation for Clinical Chemistry and Laboratory Medicine (IFCC), and the International Laboratory Accreditation Cooperation.<sup>21</sup> The JCTLM maintains an online database that lists RMs, RMPs, and reference laboratories that offer RMP services that conform to the ISO standards for these reference system components listed in Box 7.2. JCTLM uses a structured review process and examines evidence that reference system components meet the requirements in the ISO standards. IVD manufacturers and clinical laboratories use the JCTLM database to identify higher-order reference systems for use in their calibration hierarchies for end-user MPs. As of 2019, JCTLM listed pure substance CRMs for 92 analytes, matrix-based CRMs for 107 measurands, RMPs for 86 analytes (103 measurands considering the analytes in different clinical sample matrices), and 41 measurands with RMP services offered by calibration laboratories.

### BOX 7.2 ISO Standards Used by JCTLM to List Conforming Reference System Components

- ISO 15193:2009. In vitro diagnostic medical devices—measurement of quantities in samples of biological origin—requirements for content and presentation of reference measurement procedures. ISO, Geneva, Switzerland, 2009.
- ISO 15194:2009. In vitro diagnostic medical devices—measurement of quantities in samples of biological origin—requirements for certified reference materials and the content of supporting documentation. ISO, Geneva, Switzerland, 2009.
- ISO 15195:2018. Laboratory medicine—requirements for the competence of calibration laboratories using reference measurement procedures. ISO, Geneva, Switzerland, 2018.
- ISO 21151:2020. In vitro diagnostic medical devices—requirements for international harmonisation protocols establishing metrological traceability of values assigned to calibrators and human samples. International Organization for Standardization, Geneva, Switzerland, 2020.

Many clinical laboratory tests do not have higher-order reference system components available. Reasons for the limited number of RMs and RMPs include technical complexity and high cost to develop CRMs and RMPs. Introduction of ISO 21151:2020 offers an alternative approach for harmonization of many important measurands. In laboratory medicine, we must remember that although a CRM or RMP is desirable, a more pragmatic approach such as a harmonization protocol will improve the usefulness of laboratory tests and reduce the risk from incorrect interpretation of test results. The International Consortium for Harmonization of Clinical Laboratory Results maintains a table of measurands with information regarding harmonization status and priority for harmonization of measurands that have different results from different MPs.<sup>22</sup>

One important limitation of some of the CRMs listed by JCTLM is they were not examined for commutability with patients' samples. As described in the next section, commutability is an essential property of a matrix-based RM to be suitable for use as a calibrator in a calibration hierarchy. Following publication of the requirements for CRMs in ISO 15194:2009, 2nd edition, the JCTLM required commutability validation for newer matrix-based CRMs to be listed but as of this writing has not removed matrix-based CRMs for which commutability has not been validated.

RMs are also provided by organizations that do not submit for review and listing by JCTLM. For example, the World Health Organization (WHO) operates separately from the JCTLM and the ISO standards, and offers over 400 RMs approved by their Expert Committee on Biological Standardization. Most of the WHO RMs are intended for qualitative testing for infectious diseases, but almost 100 RMs are intended for quantitative immunoassays for endocrine hormones and proteins. Since almost none of the WHO RMs have been validated to be commutable with patients' samples, they are not entirely suitable for use but are used because there

is nothing else available and IVD manufacturers are required to use what is available. One challenge in laboratory medicine is to ensure the reference system components used by IVD manufacturers and clinical laboratories are suitable for use.

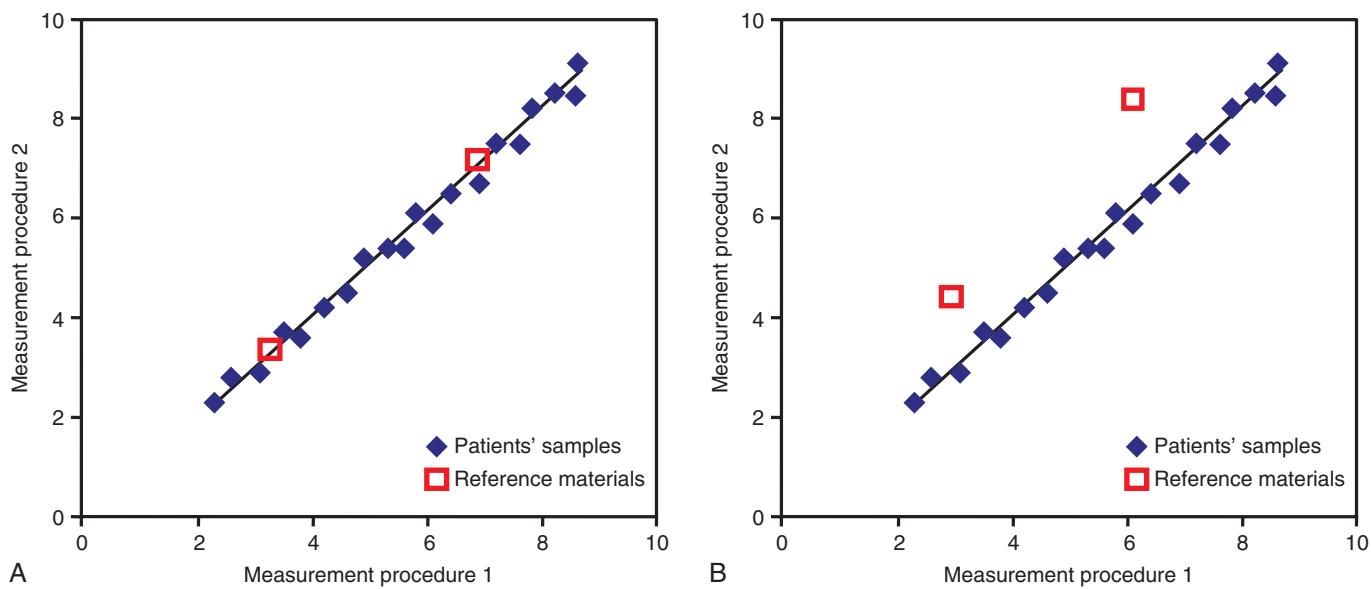
### POINTS TO REMEMBER

#### *The Joint Committee for Traceability in Laboratory Medicine*

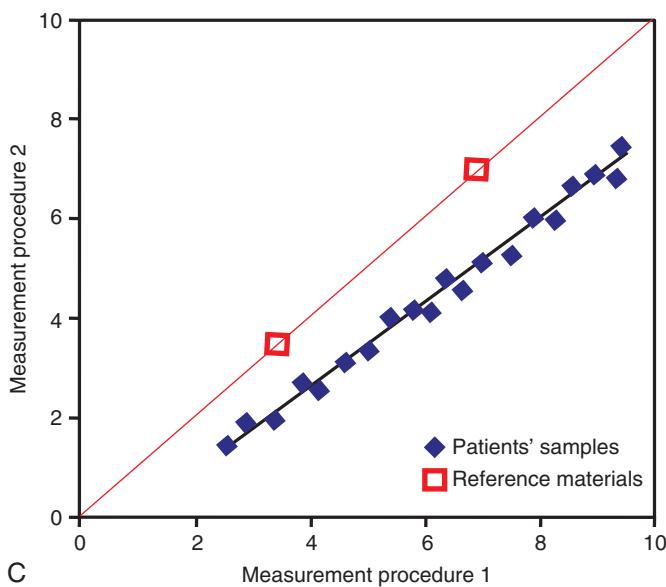
- The Joint Committee for Traceability in Laboratory Medicine examines and approves certified reference materials, reference measurement procedures, and calibration laboratories that offer reference measurement services for conformance to the requirements in the applicable International Organization for Standardization (ISO) standards for each reference system component.
- The Joint Committee for Traceability in Laboratory Medicine is expected to add international harmonization protocols to its review and listing process based on ISO 21151:2020.

### COMMUTABILITY OF A REFERENCE MATERIAL WITH PATIENTS' SAMPLES

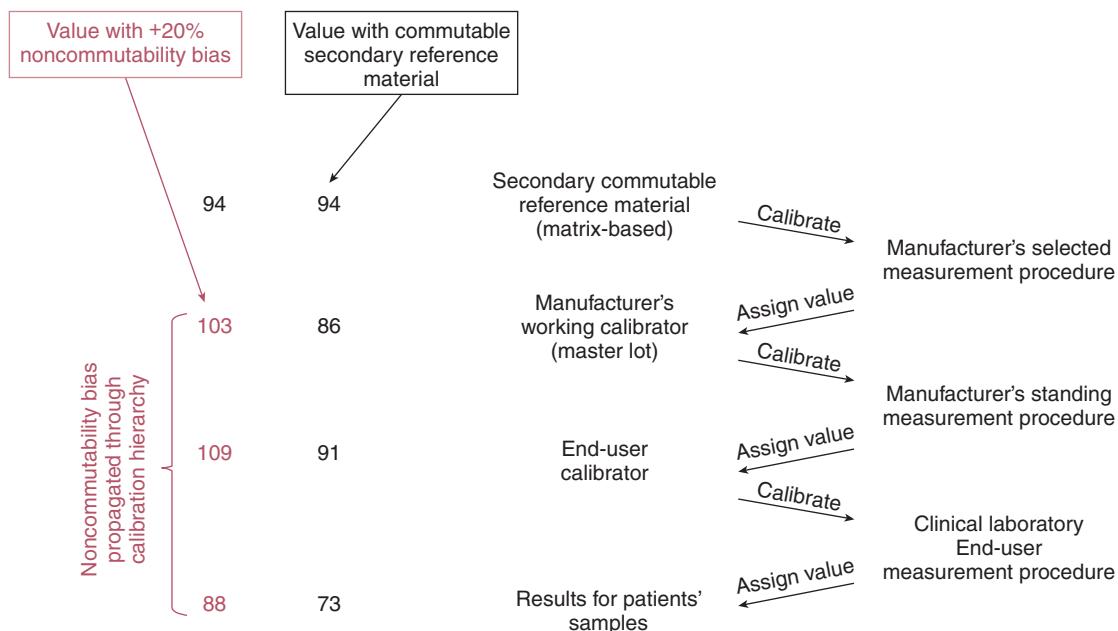
Matrix-based RMs intended for use as calibrators in the calibration hierarchies of end-user MPs must have the property called *commutability*.<sup>7,23,24</sup> As illustrated in Fig. 7.5A, a commutable RM has a numeric relationship between two (or more) MPs that closely agrees with the relationship observed for a panel of patients' samples. Consequently, a commutable RM (that may be used as a calibrator) reacts in an MP to give a numeric result that would be in close agreement to that observed for a patient's sample with the same amount of measurand. Commutability is a challenge for secondary



**FIGURE 7.5** Illustration of commutable and noncommutable reference materials. A, Commutable reference materials (red squares) have the same relationship between two measurement procedures as observed for patient samples (blue diamonds). B, Noncommutable reference materials have a different relationship than observed for patient samples.



**FIGURE 7.5—Cont'd C,** If the noncommutable reference materials were used as calibrators, there would be apparent agreement for the reference materials but the results for patients' samples would be offset by the magnitude of the noncommutability bias and would not agree between the two measurement procedures.



**FIGURE 7.6** Propagation of noncommutability bias for a secondary reference material in the calibration hierarchy of an end-user measurement procedure. The value 94 (arbitrary units) is assigned to the secondary reference material by its producer. The *black values* represent the condition when the secondary reference material is commutable with patients' samples for use with the manufacturer's selected measurement procedure and the end-user measurement procedure. The *red values* represent the condition when the secondary reference material has a +20% noncommutability bias when used as a calibrator for the manufacturer's selected measurement procedure.

matrix-based RMs because their matrix may be modified from that of patients' specimens during preparation. Differences in matrix-related bias between a RM and patients' samples causes a noncommutable relationship between the RM and the patients' samples as illustrated in Fig. 7.5B. Fig. 7.5C shows that if noncommutable RMs from Fig. 7.5B were used as calibrators, the RMs would have an apparent agreement between the two MPs, but the patients' samples

results would not agree. Fig. 7.6 shows how the noncommutability bias (also called matrix-related bias or matrix-bias) of a noncommutable RM used as a calibrator is propagated through the calibration hierarchy and cause results for patients' samples to be biased.

A matrix-based secondary commutable RM is typically not used as a calibrator for end-user MPs because sufficient amounts of such RMs are not available and would be cost

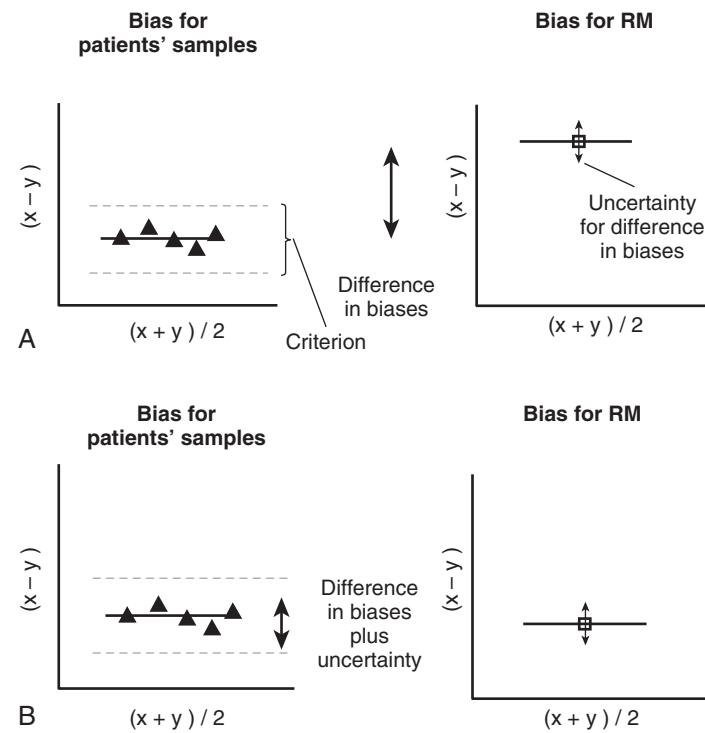
prohibitive. Rather, IVD manufacturers produce end-user calibrators with values assigned according to the calibration hierarchies shown in Figs. 7.1 and 7.2. Commutability of matrix-based secondary commutable RMs is assessed using end-user MPs and the manufacturer is responsible for the production of working and end-user calibrators with values that successfully achieve metrological traceability from the patients' results to the value assigned to the matrix-based secondary commutable RM. The manufacturer's selected, standing and end-user MPs may utilize the same or different method principles. When the same, the selected and standing MPs are usually operated with more stringent calibration and replication protocols to achieve smaller uncertainties for the results at those steps in the calibration hierarchy. The property of commutability is a function of the reagent formulation and measurement conditions and is not influenced by calibration or replication designs. Consequently, the commutability assessment for the end-user MP will be applicable to the selected and standing MPs when all are the same MP. If the selected MP utilizes a different method principle, for example, HPLC versus direct spectrophotometry, then the manufacturer needs to determine the commutability of the matrix-based secondary commutable RM with patients' samples for the selected MP versus the end-user MP.

An IVD manufacturer's working calibrator and end-user calibrator do not need to be commutable with patients' samples. An IVD manufacturer assigns a value to these MP-specific calibrators that compensates for any noncommutability bias (matrix-related bias) that may be present so that results for patients' samples are metrologically traceable to the secondary commutable RM in the calibration hierarchy.<sup>24</sup> Consequently, end-user calibrators typically have matrix characteristics and target values that are intended only for use with a specific end-user MP. Such an MP-specific calibrator cannot be used with a different MP because it does not compensate for a different noncommutability bias with that different MP and results for patients' samples will be incorrect.

### Validating Commutability of a Reference Material

Commutability of a RM can be validated using several approaches. All approaches are based on showing that results for a RM are equivalent to results for patients' samples when measured by different MPs. Historically, CLSI EP14 and EP30 used linear regression on results for patients' samples measured using two MPs with the statistical 95% prediction interval used as a criterion for an equivalent relationship between results for a RM and the panel of patients' samples.<sup>25,26</sup> Key limitations of the CLSI approaches are different magnitudes of prediction intervals among different pairs of two MPs that have no relationship to a criterion based on medical requirements for MP performance, and the uncertainty of the difference between results for RM and patients' samples is not considered.<sup>24</sup>

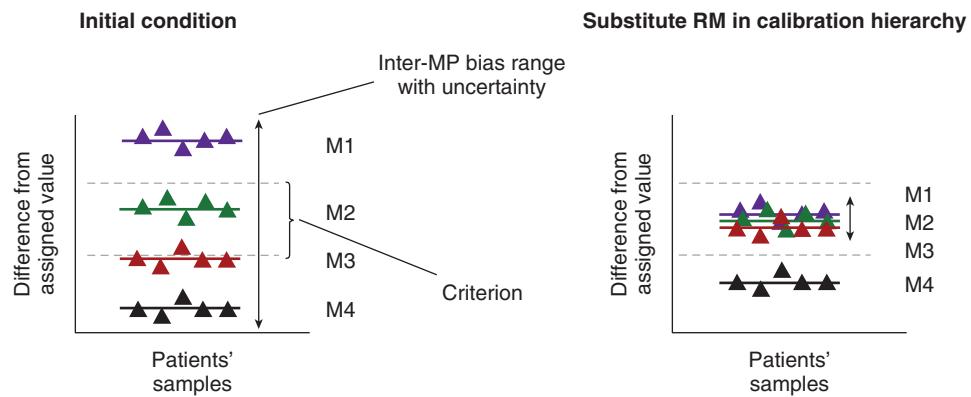
An IFCC working group published improved procedures for validating commutability of matrix-based RMs.<sup>24,27,28</sup> The IFCC recommendations included qualifying performance characteristics of MPs as suitable for inclusion in a commutability assessment, and specifying characteristics of the individual patients' samples as suitable for use in a commutability assessment.<sup>24</sup> Fig. 7.7 shows the principle of the



**FIGURE 7.7** Difference in bias procedure for commutability assessment. A, The difference between the bias in results for patients' samples measured using two measurement procedures (black triangles) and the bias for a reference material (RM, red square) is compared to a criterion for commutability (dashed lines). Panel A shows a noncommutable RM because the difference in bias plus its uncertainty (blue vertical arrow) exceeds the criterion. B, A commutable RM has a difference in bias plus its uncertainty that is within the criterion. The symbols x and y represent results from each of two different measurement procedures.

difference in bias procedure for commutability assessment.<sup>27</sup> The bias between two MPs is shown as a difference plot for the patients' samples and for the matrix-based RM. The difference in the two biases is compared to a fixed criterion based on the uncertainty required for the RM in the calibration hierarchy of the end-user MP that is derived from how the test results are used in medical decisions. In Fig. 7.7A, the difference in bias plus its uncertainty are outside the criterion so the RM is considered noncommutable. Fig. 7.7B shows a commutable RM with difference in bias plus its uncertainty within the criterion. The difference in bias is evaluated for all combinations of two MPs in the commutability assessment to determine if the matrix-based RM is suitable for use with a sufficiently large fraction of MPs to be practical. All combinations of end-user MPs do not need to be examined when an RMP for the measurand is available and the results from each end-user MP are compared to results from the RMP. Nilsson and colleagues<sup>27</sup> includes a detailed worked example for the difference in bias approach for commutability assessment.

Fig. 7.8 shows the principle of the calibration effectiveness procedure for commutability assessment.<sup>28</sup> A panel of patients' samples is measured with each MP using its current calibration. The inter-MP bias range with its uncertainty is calculated and compared to a criterion determined as described in the

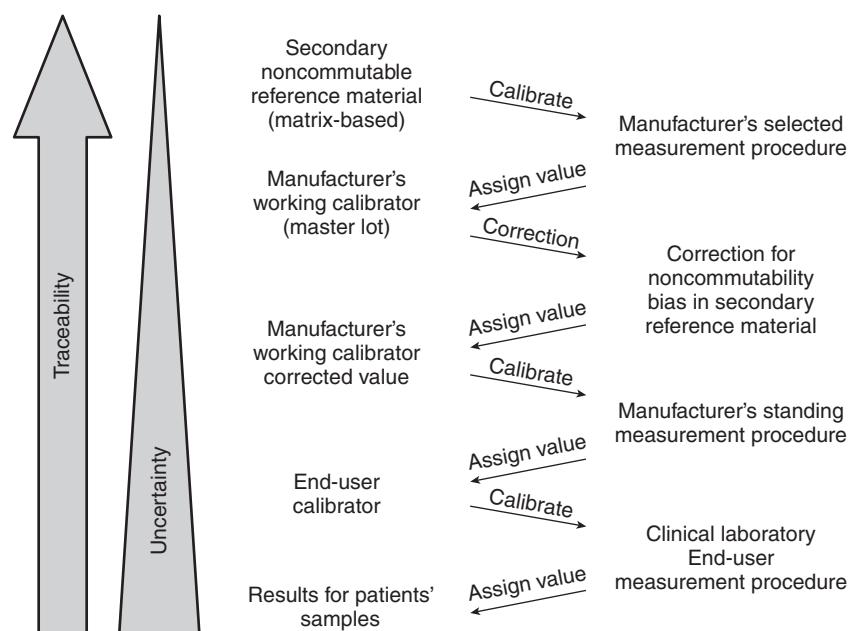


**FIGURE 7.8** Calibration effectiveness procedure for commutability assessment. The difference between results for patients' samples measured by each measurement procedure in the assessment and the value assigned to each patient's sample is plotted for each patients' sample (*triangles*). The median differences (M, *horizontal lines*) are used to calculate an inter-measurement procedure (MP) bias range plus its uncertainty. The reference material (RM) is substituted in the calibration hierarchy for each MP and the inter-MP bias recalculated from the results for patients' samples. The RM is commutable for use with those MPs (1, 2, and 3) for which the inter-MP bias plus its uncertainty is within the criterion (*dashed horizontal lines*). The RM is noncommutable for use with MP4.

preceding difference in bias approach. The matrix-based RM is then substituted for the current calibrator used in the calibration hierarchies for each MP and the patients' samples re-measured. The inter-MP bias range is recalculated after using the RM as a calibrator. The RM is commutable for use with each MP for which, when included in the calculation, the inter-MP bias range is within the criterion. This approach is only suitable when the matrix-based RM can be used in the calibration hierarchies of the end-user MPs, so cannot be used for assessment of commutability of external quality assessment/proficiency testing materials or trueness controls. Budd and colleagues<sup>28</sup> includes a detailed worked example for the calibration effectiveness approach for commutability assessment.

### Correcting for Noncommutability Bias

When a matrix-based CRM is commutable for use with a large fraction of IVD MPs in use but is noncommutable for use with one or a few IVD MPs, a correction for the bias caused by noncommutability can be added to the metrological traceability chain.<sup>29</sup> Fig. 7.9 shows that, for a noncommutable CRM, the magnitude of the difference in bias, called the noncommutability bias, can be used as a correction step in the metrological traceability chain for the end-user MP for which the CRM was noncommutable for use. In the original commutability assessment, the difference in bias was determined for all combinations of MPs for which the matrix-based CRM is intended to be used as a calibrator in their calibration hierarchies. Because the uncertainty of the noncommutability



**FIGURE 7.9** A correction for noncommutability bias of the secondary reference material is added to the calibration hierarchy to allow the results for patients' samples to be metrologically traceable to the value assigned to the secondary reference material.

bias used as a correction will be added to the uncertainties from the other steps in the metrological traceability chain, in most cases an uncertainty smaller than that from the original commutability assessment will be needed to ensure the combined uncertainty of the patients' results will be within the performance specifications. Consequently, a new experiment is typically needed with a larger number of patients' samples and more replication to ensure a small uncertainty in the difference in bias to be used as a correction factor in the calibration hierarchy. See Miller and colleagues<sup>29</sup> for details of experimental designs with worked examples to determine a correction for noncommutability bias in a CRM when used with a particular end-user MP. Adding the correction step in the calibration hierarchy of the end-user MP provides metrological traceability to the value assigned to the CRM.

### POINTS TO REMEMBER

#### **Commutability of a Reference Material With Patients' Samples**

- Commutability is a property of a reference material characterized by a numeric relationship between two (or more) measurement procedures that closely agrees, within a medical use defined criterion, with the relationship observed for a panel of patients' samples.
- A commutable certified reference material reacts in a measurement procedure to give a numeric result that would be in close agreement to that observed for a patient's sample with the same amount of measurand.
- Matrix-based secondary reference materials must be commutable to be used as calibrators in the calibration hierarchy for an end-user measurement procedure.
- If a matrix-based secondary calibrator is noncommutable, the noncommutability bias will be propagated in the metrological traceability chain such that results for patients' samples will not agree with results from other end-user measurement procedures.
- A correction for noncommutability bias can be added to the metrological traceability chain to allow correct metrological traceability of results for patients' samples to the value assigned to the matrix-based secondary reference material.

### VERIFYING METROLOGICAL TRACEABILITY OF AN END-USER MEASUREMENT PROCEDURE TO A HIGHER-ORDER REFERENCE SYSTEM

A clinical laboratory may wish to verify that an MP's calibration conforms to an IVD manufacturer's claim for traceability to the reference system used for a given measurand. A clinical laboratory has limited resources to verify the calibration traceability of a commercially available or laboratory developed end-user MP. National and international CRMs are available for some measurands. Not all matrix-based CRMs listed by JCTLM or available from other providers have been validated for commutability. Some matrix-based CRMs and most pure substance CRMs are not intended for use with clinical laboratory end-user MPs. Rather, they are intended for use with higher-order RMPs that have better selectivity for the measurand. Very few pure substance CRMs include diluents in their instructions for use that are suitable

for use to prepare commutable matrix-based RMs by a laboratory.

A CRM's certificate of analysis should be reviewed for commutability documentation. If a CRM is commutable with patients' samples for use with a given end-user MP, it can be measured as if a patient's sample and the result compared to the assigned value to verify the traceability of calibration to the reference system. The criterion for agreement between the value measured in a laboratory and the assigned value should consider the combined uncertainty of both values (see [Equation 7.1](#)). Using a noncommutable CRM or other RM as a calibrator will cause the routine MPs to be miscalibrated and produce erroneous patient results.<sup>7,23,29–31</sup> Similarly, measuring a noncommutable CRM or other RM as if it were a patient sample to verify metrological traceability will give incorrect information regarding the traceability of an end-user MP. If a CRM's or other RM's commutability status is unknown, it must be assumed not to be commutable with patient samples, and conclusions regarding metrological traceability of an end-user MP may be erroneous when based on such an RM.

Some MP manufacturers provide materials, called trueness controls, specifically intended to verify metrological traceability of their specific end-user MP. Such materials may be provided as MP-specific quality control or calibration verification materials. As for MP-specific calibrators, such MP-specific quality control materials typically have matrix characteristics and target values that are intended only for use with the specific MP claimed in the instructions for use and cannot be used with any other manufacturer's MP. Such MP-specific materials to verify metrological traceability may have target values that are specific for stated reagent lots, or they may have values certified by the IVD manufacturer to be suitable for all reagent lots.

A laboratory can verify the calibration traceability of an end-user MP by sending a set of patients' samples to a laboratory that offers an RMP for that measurand. However, there are few reference laboratories that offer JCTLM listed RMPs and the cost may be prohibitive. Note that the term "reference laboratory" is often used colloquially to mean a referral laboratory that offers a wide menu of testing frequently including specialized testing. Such referral laboratories do not generally offer RMPs that meet the metrological traceability requirements described in the ISO standards. However, some referral laboratories or other clinical laboratories may offer well-characterized MPs with calibration hierarchies suitable for use as high-quality comparison MPs. When a suitable RMP or comparison MP is available, results for a set of patient samples can be compared to results from the clinical laboratory's MP to determine if the measured values for the measurand are the same, within the uncertainties of both values, as the values assigned by the RMP or comparison MP.

### **THE ROLE OF EXTERNAL QUALITY ASSESSMENT OR PROFICIENCY TESTING IN SURVEILLANCE FOR STANDARDIZATION/HARMONIZATION**

Monitoring the status of harmonization of results among different MPs and different laboratories is essential to sustain technical procedures to achieve standardized or harmonized results. Results from external quality assessment or proficiency testing using commutable samples are useful to laboratories to monitor their performance and to IVD manufacturers to know when their calibration hierarchies need to be

realigned with higher-order references. The criteria for agreement are typically set by the program provider and the uncertainty should be reviewed for suitability. Chapter 6 describes how external quality assessment or proficiency testing programs are operated. Newer approaches for surveillance based on monitoring medians of results from patients' samples from different laboratories using different MPs are also described in Chapter 6.

### POINTS TO REMEMBER

#### **Verifying Metrological Traceability and External Quality Assessment (Proficiency Testing)**

- A laboratory can verify the calibration traceability of an end-user measurement procedure by measuring commutable external quality assessment samples and determining if the measured values for the measurand are the same, within the uncertainties of both values, as the values assigned to the sample by a reference measurement procedure or as the mean value of participants' results from suitably qualified end-user measurement procedures.
- A laboratory can verify the calibration traceability of an end-user measurement procedure by measuring a commutable secondary reference material as a patient's sample and determining if the measured value for the measurand is the same, within the uncertainties of both values, as the value assigned to the commutable secondary referenced material.
- When a suitable reference or comparison measurement procedure is available, a clinical laboratory can determine if the results for a set of patients' samples from the end-user measurement procedure agree, within the uncertainties of both values, with the results from the reference or comparison measurement procedure.

### HARMONIZATION OF OTHER PATH OF WORKFLOW AREAS IN LABORATORY MEDICINE

Although not addressed in this chapter, **Box 7.3** lists other areas in laboratory medicine where harmonization is needed. Additional information on activities that address these areas is available in the literature.<sup>7,32–34</sup> Harmonization of practices in all of these areas is important for reducing risk of harm to patients caused by errors in the steps in what has been called the “brain-to-brain” loop<sup>35</sup> or path of workflow for laboratory services.

#### **BOX 7.3 Areas in Addition to Measurement Results Where Standardization/Harmonization Is Important**

1. Test ordering nomenclature
2. Patient identification
3. Specimen collection
4. Specimen transportation
5. Specimen storage before examination
6. Specimen preparation for examination
7. Result reporting units
8. Result reference intervals or decision limits
9. Result interpretive information

### SELECTED REFERENCES

1. Beastall GH, Brouwer N, Quiroga S, Myers GL. Traceability in laboratory medicine: a global driver for accurate results for patient care. *Clin Chem Lab Med* 2017;55:1100–8.
5. ISO 17511:2020 (2nd edition). In vitro diagnostic medical devices—Requirements for establishing metrological traceability of values assigned to calibrators, trueness control materials and human samples. International Organization for Standardization, Geneva, Switzerland, 2020.
6. Vesper HW, Thienpont LM. Traceability in laboratory medicine. *Clin Chem* 2009;55:1067–75.
7. Miller WG, Tate JR, Barth JH, et al. Harmonization: the sample, the measurement and the report. *Ann Lab Med* 2014;34:187–97.
10. Braga F, Panteghini M. Defining permissible limits for the combined uncertainty budget in the implementation of metrological traceability. *Clin Biochem* 2018;57:7–11.
13. ISO 21151:2020. In vitro diagnostic medical devices—Requirements for international harmonisation protocols establishing metrological traceability of values assigned to calibrators and human samples. International Organization for Standardization, Geneva, Switzerland, 2020.
15. Miller WG, Myers GL, Gantzer ML, et al. Roadmap for harmonization of clinical laboratory measurement procedures. *Clin. Chem* 2011;57:1108–1117.
21. Joint Committee for Traceability in Laboratory Medicine. <http://www.bipm.org/jctlm/>. Accessed 12 January 2020.
22. International Consortium for Harmonization of Clinical Laboratory Results. <http://www.harmonization.net/>. Accessed 12 January 2020.
23. Miller WG, Myers GL. Commutability still matters. *Clin Chem* 2013;59:1291–3.
24. Miller WG, Schimmel H, Rej R, et al. IFCC Working Group Recommendations for Assessing Commutability Part 1: General Experimental Design. *Clin Chem* 2018;64:447–54.
27. Nilsson G, Budd JR, Greenberg N, et al. IFCC Working Group Recommendations for Assessing Commutability Part 2: Using the Difference in Bias Between a Reference Material and Clinical Samples. *Clin Chem* 2018;64:455–64.
28. Budd JR, Weykamp C, Rej R, et al. IFCC Working Group Recommendations for Assessing Commutability Part 3: Using the Calibration Effectiveness of a Reference Material. *Clin Chem* 2018;64:465–74.
29. Miller WG, Budd J, Greenberg N, et al. IFCC working group recommendations for correction of bias caused by non-commutability of a certified reference material used in the calibration hierarchy of an end-user measurement procedure. *Clin Chem* 2020;66:769–78.
31. Vesper HW, Miller WG, Myers GL. Reference materials and commutability. *Clin Biochem Rev* 2007;28:139–47.
32. Tate JR, Johnson R, Barth J, Panteghini M. Harmonization of laboratory testing—Current achievements and future strategies. *Clin Chim Acta* 2014;432:4–7
34. Tate JR, Graziani MS, Plebani M. Special Issue: Harmonization in Laboratory Medicine: the Request, the Sample, the Measurement, and the Report – an Update. Part 2. *Clin Chem Lab Med* 2019; 57:1–143.
35. Plebani M, Laposata M, Lundberg GD. The brain-to-brain loop concept for laboratory testing 40 years after its introduction. *Am J Clin Pathol* 2011;136:829–33.

## REFERENCES

1. Beastall GH, Brouwer N, Quiroga S, Myers GL. Traceability in laboratory medicine: a global driver for accurate results for patient care. *Clin Chem Lab Med* 2017;55:1100–8.
2. Almond A, Ellis AR, Walker SW. Current parathyroid hormone immunoassays do not adequately meet the needs of patients with chronic kidney disease. *Ann Clin Biochem* 2012;49:63–7.
3. 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 1998 (Dec 7);L331:1–37.
4. 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 2017 (May 5);L117:176–332.
5. ISO 17511:2020 (2nd edition). In vitro diagnostic medical devices—Requirements for establishing metrological traceability of values assigned to calibrators, trueness control materials and human samples. International Organization for Standardization, Geneva, Switzerland, 2020.
6. Vesper HW, Thienpont LM. Traceability in laboratory medicine. *Clin Chem* 2009;55:1067–75.
7. Miller WG, Tate JR, Barth JH, et al. Harmonization: the sample, the measurement and the report. *Ann Lab Med* 2014; 34:187–97.
8. Consultative Committee for Amount of Substance: Metrology in Chemistry and Biology. <https://www.bipm.org/en/committees/cc/ccqm/members-cc.html>; accessed 09 July 2020.
9. International vocabulary of metrology—basic and general concepts and associated terms (VIM). 3rd Edition. Joint Committee for Guides on Metrology, JCGM 200:2012.
10. Braga F, Panteghini M. Defining permissible limits for the combined uncertainty budget in the implementation of metrological traceability. *Clin Biochem* 2018;57:7–11.
11. European Commission, Joint Research Center, Certification of proteins in the human serum Certified Reference Material ERM-DA470k/IFCC. <https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/certification-proteins-human-serum-certified-reference-material-erm-da470kifcc> Accessed 09 July 2020.
12. European Commission, Joint Research Center, The certification of the mass concentration of beta-2-microglobulin in human serum: ERM-DA470k/IFCC. <https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/certification-mass-concentration-beta-2-microglobulin-human-serum-erm-da470kifcc>. Accessed 09 July 2020.
13. ISO 21151:2020. In vitro diagnostic medical devices— Requirements for international harmonisation protocols establishing metrological traceability of values assigned to calibrators and human samples. International Organization for Standardization, Geneva, Switzerland, 2020.
14. Thienpont LM, Van Houcke SK. Traceability to a common standard for protein measurements by immunoassay for in-vitro diagnostic purposes. *Clin Chim Acta* 2010;411(23 – 24):2058–61.
15. Miller WG, Myers GL, Gantzer ML, et al. Roadmap for harmonization of clinical laboratory measurement procedures. *Clin. Chem* 2011;57:1108–1117.
16. Van Houcke SK, Van Aeist S, Van Uytfanghe K, et al. 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* 2013;51:e103–5.
17. Van Uytfanghe K, De Grande LA, Thienpont LM. A “Step-Up” approach for harmonization. *Clin Chim Acta* 2014;432:62–7.
18. Stöckl D, Van Uytfanghe K, Van Aeist S, et al. A statistical basis for harmonization of thyroid stimulating hormone assays using a robust factor analysis model. *Clin Chem. Lab Med* 2014;52: 965–72.
19. Thienpont LM, Van Uytfanghe K, De Grande LA, et al. Harmonization of serum thyroid-stimulating hormone measurements paves the way for the adoption of a more uniform reference interval. *Clin Chem* 2017;63:1248–60.
20. Miller WG. Harmonization: its time has come. *Clin Chem* 2017;63:1184–6.
21. Joint Committee for Traceability in Laboratory Medicine. <http://www.bipm.org/jctlm/>. Accessed 09 July 2020.
22. International Consortium for Harmonization of Clinical Laboratory Results. <http://www.harmonization.net/>. Accessed 09 July 2020.
23. Miller WG, Myers GL. Commutability still matters. *Clin Chem* 2013;59:1291–3.
24. Miller WG, Schimmel H, Rej R, et al. IFCC Working Group Recommendations for Assessing Commutability Part 1: General Experimental Design. *Clin Chem* 2018;64:447–54.
25. Evaluation of Commutability of Processed Samples; Approved Guideline – Third Edition. CLSI document EP14-A3, Clinical and Laboratory Standards Institute, Wayne, PA, 2014.
26. Characterization and qualification of commutable reference materials for laboratory medicine; approved guideline. CLSI document EP30-A, Clinical and Laboratory Standards Institute, Wayne, PA, 2010.
27. Nilsson G, Budd JR, Greenberg N, et al. IFCC Working Group Recommendations for Assessing Commutability Part 2: Using the Difference in Bias Between a Reference Material and Clinical Samples. *Clin Chem* 2018;64:455–64.
28. Budd JR, Weykamp C, Rej R, et al. IFCC Working Group Recommendations for Assessing Commutability Part 3: Using the Calibration Effectiveness of a Reference Material. *Clin Chem* 2018;64:465–74.
29. Miller WG, Budd J, Greenberg N, et al. IFCC working group recommendations for correction of bias caused by non-commutability of a certified reference material used in the calibration hierarchy of an end-user measurement procedure. *Clin Chem* 2020;66:769–78.
30. Franzini C, Ceriotti F. Impact of reference materials on accuracy in clinical chemistry. *Clin Biochem* 1998;31:449–57.
31. Vesper HW, Miller WG, Myers GL. Reference materials and commutability. *Clin Biochem Rev* 2007;28:139–47.
32. Tate JR, Johnson R, Barth J, Panteghini M. Harmonization of laboratory testing—Current achievements and future strategies. *Clin Chim Acta* 2014;432:4–7.
33. Plebani M. Harmonization in laboratory medicine: Requests, samples, measurements and reports. *Crit Rev Clin Lab Sci* 2016;53:184–96.
34. Tate JR, Graziani MS, Plebani M. Special Issue: Harmonization in Laboratory Medicine: the Request, the Sample, the Measurement, and the Report – an Update. Part 2. *Clin Chem Lab Med* 2019;57:1–143.
35. Plebani M, Laposata M, Lundberg GD. The brain-to-brain loop concept for laboratory testing 40 years after its introduction. *Am J Clin Pathol* 2011;136:829–33.

## MULTIPLE CHOICE QUESTIONS

1. What is metrological traceability?
  - a. An unbroken chain of calibrations from a patient's sample result to a higher-order reference material or reference measurement procedure.
  - b. A process used by a manufacturer to assign values to its working calibrator.
  - c. The relationship between a measurement procedure and its calibration curve.
  - d. A property of a reference material that means it can be used as a calibrator for an end-user measurement procedure.
  - e. A series of comparisons among different measurement procedures intended to make results for patients' samples agree.
2. What is a certified reference material?
  - a. A reference material provided by an in vitro diagnostic manufacturer with assigned values and uncertainties described in the instructions for use (package insert).
  - b. A reference material of high purity that a manufacturer certifies is suitable for use with one or more measurement procedures.
  - c. A reference material accompanied by a certificate issued by an authoritative body that provides property values with associated uncertainties and traceabilities.
  - d. A reference material that is used by at least 80% of clinical laboratories.
  - e. A reference material that comes with a package insert that gives the concentration of a measurand.
3. Where is a pure substance primary certified reference material usually used in metrological traceability?
  - a. To prepare a calibrator in a suitable matrix for a reference measurement procedure for the measurand.
  - b. To prepare a matrix-based secondary commutable reference material for use as a calibrator for a manufacturer's selected measurement procedure.
  - c. To prepare a matrix-based working calibrator for calibration of a manufacturer's standing measurement procedure.
  - d. To prepare end-user calibrators for clinical laboratory end-user measurement procedures.
  - e. To prepare a calibrator in a suitable matrix for use by a clinical laboratory to calibrate an end-user measurement procedure.
4. What is a secondary commutable reference material?
  - a. A reference material that is used to calibrate an end-user measurement procedure because it is metrologically traceable to a higher-order reference material.
  - b. A reference material that is considered secondary because its matrix is similar to that of patients' samples.
  - c. A reference material that has a secondary purpose to verify calibration of a manufacturer's selected measurement procedure.
  - d. A reference material that has a matrix similar to that of clinical patients' samples and gives a measurement response equivalent to that of a patient's sample containing the same amount of a measurand.
  - e. A reference material whose assigned value is established by consensus of qualified end-user measurement procedures.

5. What is commutability?
  - a. A property of a reference material that means the results for that reference material measured using different measurement procedures have the same relationship between measurement procedures that is observed for a panel of patients' samples.
  - b. A property of a reference material that allows that material to be substituted for calibrators already used in the calibration hierarchies of end-user measurement procedures.
  - c. A property of end-user measurement procedures that means the results for patients' samples are equivalent when measured using any of the measurement procedures.
  - d. A condition when results for patients' samples are harmonized among most end-user measurement procedures.
  - e. A condition when the results for patients' samples measured using most end-user measurement procedures agree well enough that there is no risk of harm to a patient when interpreting a result using clinical practice guidelines.
6. What happens when a noncommutable secondary reference material is used in the calibration hierarchies of end-user measurement procedures?
  - a. Reference intervals will have to be reestablished for each end-user measurement procedure.
  - b. A common reference interval will be applicable for all end-user measurement procedures.
  - c. Results for patients' samples will agree better among different end-user measurement procedures than before the reference material was used.
  - d. Results for patients' samples will remain essentially unchanged among different end-user measurement procedures.
  - e. Results for patients' samples will not agree among different end-user measurement procedures.
7. How is a value assigned to a secondary commutable reference material when there is no higher-order reference measurement procedure for the measurand?
  - a. The mean value for all current end-user measurement procedures is used.
  - b. The trimmed mean value for all current end-user measurement procedures is used.
  - c. The mean value for the 10 most frequently used end-user measurement procedures is used.
  - d. The mean value for end-user measurement procedures that meet defined performance specifications is used.
  - e. The mean value for end-user measurement procedures that use a well-defined calibrator is used.
8. When should an international harmonization protocol be used?
  - a. When all available reference measurement procedures are influenced by interfering substances.
  - b. When there are no higher-order reference system components available.
  - c. When a secondary commutable reference material is available but it is noncommutable for use with 20% of the end-user measurement procedures.
  - d. When a clinical laboratory cannot afford to purchase a secondary commutable reference material.

- e. When an IVD manufacturer determines that a secondary commutable reference material is actually noncommutable for use in the calibration hierarchy of its end-user measurement procedure.
- 9. What are three critical steps in implementing an international harmonization protocol?
  - a. Informing in vitro diagnostic manufacturers, preparing harmonization reference materials, informing clinical laboratories to use the harmonization reference materials to calibrate their end-user measurement procedures.
  - b. Preparing harmonization reference materials, assigning values to the harmonization reference materials, developing corrections to the calibration hierarchies of each end-user measurement procedure.
  - c. Preparing harmonization reference materials, assigning values to the harmonization reference materials, informing clinical laboratories to use the harmonization reference materials to calibrate their end-user measurement procedures.
  - d. Preparing harmonization reference materials, assigning values to the harmonization reference materials, informing in vitro diagnostic manufacturers that the harmonization reference materials are available.
  - e. Preparing harmonization reference materials, assigning values to the harmonization reference materials, informing external quality assessment or proficiency testing organizations to use the harmonization reference materials in their programs.
- 10. What is the role of the Joint Committee for Traceability in Laboratory Medicine?
  - a. To accredit in vitro diagnostic manufacturers who implement metrological traceability.
  - b. To develop educational programs for in vitro diagnostic manufacturers on how to implement metrological traceability.
  - c. To review and list higher-order reference system components that meet International Organization for Standardization standards for these components.
  - d. To list higher-order reference system components that are available.
  - e. To develop standards for implementing metrological traceability.

# Biological Variation and Analytical Performance Specifications\*

*Sverre Sandberg, Thomas Røraas, and Aasne K. Aarsand*

## ABSTRACT

### Background

There are many sources of variation in numerical results generated by examinations performed in laboratory medicine. Some measurands have biological variations over the span of life and others have predictable cyclical or seasonal variations. Most measurands in an individual display random variation around homeostatic set points and this is termed within-subject biological variation. The homeostatic set points vary between individuals and the variation between the set points of different individuals is termed between-subject biological variation. An understanding of these sources of variation is required to enable appropriate application of clinical laboratory measurements.

### Content

In this chapter, we explain that numerical estimates of analytical, within-subject, and between-subject biological

variation are usually generated by prospective studies; series of specimens from a cohort of individuals are examined, followed by statistical analysis to identify and quantify the different types of variation. Furthermore, sources of evidence-based data on biological variation and tools for the appraisal of the quality of biological variation studies are presented. The chapter also provides an overview of what applications biological variation data have in laboratory medicine such as the “index of individuality” and “reference change value” where the latter is used to determine whether changes in serial results from an individual can be explained by analytical and within-subject biological variation only. Additionally, models for setting analytical performance specifications, for imprecision, bias, total error, and measurement uncertainty, which can be created using estimates of within-subject and between-subject variation, are presented.

\*The full version of this chapter is available electronically on ExpertConsult.com.

## INTRODUCTION

There are many causes of variation that contribute to the uncertainty of any result generated in laboratory medicine. Biological variation is one of the most important sources and should be taken into account in any interpretation made. This chapter is based on a chapter in the previous edition of the Tietz textbook.<sup>1</sup>

There are various types of biological variation. The concentration or activity of some measurands changes over the span of life, some slowly and some more quickly, particularly at times of rapid physiologic development, such as the neonatal period, childhood, puberty, menopause, adults, and advanced age. The concentration or activity of measurands can also differ between men and women. This variation is taken care of by the creation of age- and/or sex-stratified (partitioned) reference intervals. A number of measurands have predictable cyclical rhythms in their concentrations. These can be daily (e.g., iron), monthly (e.g., pituitary gonadotrophins in females), or seasonal (e.g., vitamin D) in nature. Knowledge of the expected values throughout the cycles mentioned above is vital for clinical interpretation, and specimen collection should occur at appropriate times. An absence of rhythm may indicate disease. These types of biological variations are described in detail in Chapter 9.

In 1960, Schneider<sup>2</sup> defined the distribution of values observed in a group as caused by “the effect of a large number of undefined forces acting randomly to displace the values of the individual members of the group away from the true group value.” He described three factors which contribute to the overall variation in a dataset, where one result from each individual has been included.

- Factors which make for true differences between individuals (interindividual/between-subject)
- Factors which make for true differences from time to time in each single individual (intraindividual/within-subject)
- Factors which make for true differences from measurement to measurement of each sample that may be measured (analytical).

Under the assumption of constant and homogenous within-subject variation, as well as that the variation of laboratory error is the same for all measurements, he divided the observed variance (squared standard deviation [SD])  $SD_{\text{observed}}^2$  using the formula:

$$SD_{\text{observed}}^2 = SD_{\text{between observed}}^2 + SD_{\text{within-subject}}^2 + SD_{\text{analytical}}^2$$

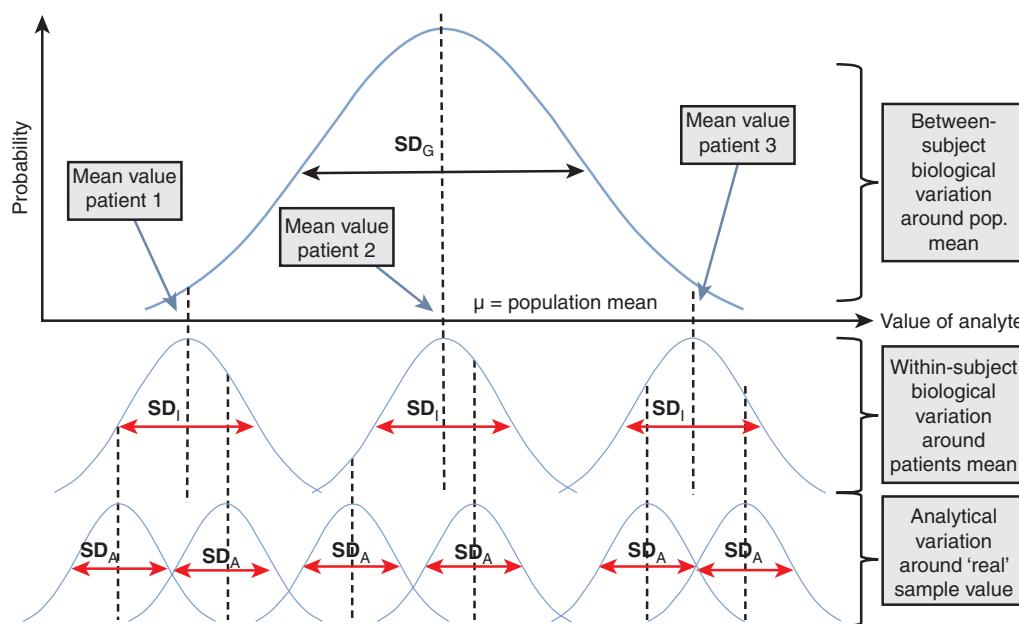
The concept was further elaborated on in a series of four articles in clinical chemistry titled “Biological and analytic components of variation in long-term studies of serum constituents in normal subjects.”<sup>3–6</sup> An overview of the analytical, within-subject, and between-subject variation is displayed in Fig. 8.1.

As an example, four specimens were taken from four individuals at daily intervals, and serum sodium activity was examined (reference interval: 135 to 147 mmol/L). The results are provided in Table 8.1. It is evident that the results for each individual vary from day to day; which is ascribed to three sources of variation: preanalytical, analytical, and within-subject biological variations. The mean value is termed the homeostatic set point. In addition, each individual has a different average serum sodium concentration; the variation among the homeostatic set points of individuals is the between-subject variation, whereas the average variation within each individual is the within-subject variation. Generation and subsequent application of numerical data on the components of biological variation are crucial facets in laboratory medicine, and both of these are described in detail in this chapter.

## Terminology

The terms and symbols used throughout this chapter are:

- SD: standard deviation
- CV: coefficient of variation
- $SD_I/SD_{\text{CV}}$ : within-subject biological variation (variation within a single individual estimated as a pooled variation from a [homogenous] group of individuals)



**FIGURE 8.1** The relationship between the between-subject ( $SD_G$ ), within-subject ( $SD_I$ ), and analytical ( $SD_A$ ) variation. (Adapted from Røraas T, Petersen PH, Sandberg S. Confidence intervals and power calculations for within-person biological variation: effect of analytical imprecision, number of replicates, number of samples, and number of individuals. *Clin Chem* 2012;58:1306–1313.)

**TABLE 8.1 Serum Sodium Activity in Four Specimens Collected at Daily Intervals From Each of a Cohort of Four Individuals**

Sodium	Day 1	Day 2	Day 3	Day 4
Individual 1	137	139	136	138
Individual 2	144	146	145	144
Individual 3	141	143	142	140
Individual 4	139	138	141	140

Values are measured in millimoles per liter.

- $SD_G/CV_G$ : between-subject biological variation (variation between the homeostatic set points of a group of individuals)
- $SD_A/CV_A$ : analytical variation (analytical examination variation).

## GENERATION OF DATA ON COMPONENTS OF BIOLOGICAL VARIATION

Production of data on biological variation is quite similar to derivation of population-based reference intervals (see Chapter 9) with the exception being that, instead of one specimen being taken from a large number of reference individuals, at least two specimens are needed. Biological variation studies are usually undertaken as prospective experimental studies that include a higher number of specimens taken from a smaller cohort of reference individuals where estimates thereafter are derived from a traditional statistical approach such as, for example analysis of variance, ANOVA or similar analyses (traditional approach), or by more recently published approaches such as Bayesian statistics. Additionally, there is a renewed interest in basing estimates on a lower number of specimens, using a larger cohort (big data), such as was the case in a recently published study where measurements available from patient cohorts from hospital data were used.<sup>7</sup>

### Prospective Experimental Studies

#### Design of Studies

In general terms, this approach recommends that numerical estimates of  $CV_A$  and both  $CV_I$  and  $CV_G$  components of biological variation should be generated using the following experimental approach:

- Select a group of reference individuals.
- Take a set of specimens from each of the individuals at regular time intervals while minimizing all sources of pre-analytical variation in preparation of the subjects for specimen collection.
- Transport specimens in a standardized way and store aliquots under controlled conditions until ready for analysis.
- Undertake the analyses in duplicate while minimizing analytical sources of variation.

This design has been widely used and is very suitable for those measurands that have a low  $CV_I$  and strict homeostatic control. For example, a typical study design includes 10 specimens collected on a weekly basis from 20 individuals recruited from a smaller cohort of reference individuals. In such a setting, the aim is to have the same number of specimens from each individual (i.e., a balanced design), but in reality, the dataset in the end is usually unbalanced with an unequal number of specimens from each individual, as not all participants will be able to participate in every sampling,

and some data points may be considered as outliers. Fraser and Harris<sup>8</sup> stated that “the components of variation can be obtained from a relatively small number of specimens collected from a small group of subjects over a reasonably short period of time”; however, solid evidence to support this statement was lacking until recently.<sup>9</sup> Basing the design on a small group of subjects can have some weaknesses, especially when we want to evaluate whether subgroups have different within-subject biological variation estimates. If the initial study population consisted of 20 subjects, consisting equally of males and females, the estimates of the subgroups for sex will only consist of 10 subjects. If we further want to evaluate age groups or ethnicity, the subgroups will be even smaller. If there is clinical reason to assess subgroups, we need to design the study with this in mind. It is a general concern that design may often be based on simplicity, for example, on how many individuals are easily recruited from the local staff at the hospital, and this may lead to the derived biological variation estimates not being representative for the general population at all.

Other types of laboratory data are often accompanied by confidence intervals (CIs); unfortunately, this has rarely been included in the most recently published reports that provide estimates of the components of biological variation. CIs are essential for appraising the results of and comparing results between studies. The determination of CIs for different balanced designs for a two-level nested variance analysis model with varying analytical imprecision has been examined in detail.<sup>9</sup> Data sets based on this model were simulated to calculate the power of different study designs for estimation of  $CV_I$ . It was found that the reliability of an estimate for  $CV_I$  and the power are greatly influenced by the study design and by the ratio between  $CV_A$  and  $CV_I$ . The study provided data where it was indicated what the effects were of increasing the number of included individuals and the number of replicates at different levels of imprecision.<sup>9</sup>

Some measurands for which biological variation estimates are sought may be unstable and examinations must therefore be performed soon after the collection of specimens (e.g., for some hematologic measurands, such as mean cell volume, number of erythrocytes and leukocytes per volume). In this case, to obtain the necessary statistically unconfounded estimate of  $CV_I$ , the  $CV_A$  is estimated by analyzing all specimens taken at each sampling point, in duplicate. However, this only represents the within-run CV. Thus in addition, quality control materials have to be analyzed between each run to ascertain that variation due to systematic deviations in the examination procedure between each examination is excluded. Using this strategy, the quality control material should preferentially be commutable, meaning that the variations of the analyses of the individual samples and the quality control materials are comparable. Furthermore, it must be assured that the concentrations or activities of the quality control materials are similar to those of the samples from the subjects studied, because  $CV_A$  often varies with concentration or activity.

### Data Analysis—Traditional Approach

The most frequently used method for sample collection and data analysis is detailed by Fraser and Harris,<sup>8</sup> where duplicate analyses are performed on samples from a cohort of individuals. However, before any components of variation are estimated with this type of method, it is important to (1) verify that the individuals are in a steady state; (2) exclude

outliers in the data set; and (3) assess whether the individuals have a homogeneously distributed within-subject CV<sub>I</sub>.

**Steady state.** The calculation of biological variation data assumes that the individuals assessed are in a “steady state,” that is, the homeostatic set points do not change during the duration of the study. If the population displays a trend over the sampling period, data should be transformed to a “steady state,” for example, by correcting for trends using regression analysis or using other methods such as multiple of medians (MoM) and its natural logarithm.<sup>10</sup>

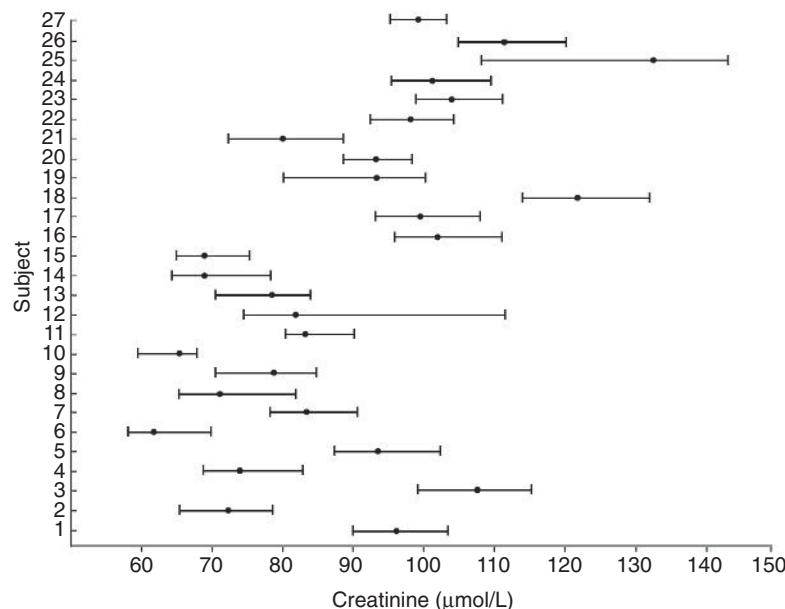
**Data transformation/normal distribution.** If one wishes to estimate the CI for the components of biological variation, most methods assume that data is normally distributed. As most biological data are naturally logarithmically distributed, the examination and calculations must be performed on the logarithms of the observations. This both helps in extracting the CVs and ensuring that the data distribution is closer to normal.<sup>11</sup> It is important to specify that the normality relates to the model effects; that is, both the analytical variation around the true sample value and the individuals’ variation around their homeostatic set point are normally distributed. It does not relate to the total pooled data. Pooling the standardized residuals for each level, namely, residuals from replicates (difference between replicates and mean of replicates from each sample), residuals from samples (difference between mean of samples and mean for the individual), and residuals from subjects (differences between individual means and total mean), can be used to assess normality.

Example: Assume observations of 2.25, 2.50, and 2.75 are from individual A and 3.50, 4.00, and 4.50 from individual B. Standardized residuals are generated by first dividing by the mean for each individual. Individual A will have an average of 2.50, so standardized observations will be 0.90, 1.00, and 1.10, and corresponding standardized residuals -0.10, 0.00, and 0.10. Individual B will have an average of 4.00, standardized observations 0.875, 1.00, and 1.125, and corresponding residuals -0.125, 0.00, and 0.125. The pooled standardized residuals will be -0.10, 0.00, 0.10, -0.125, 0.00, and 0.125. The standardized residuals can then be examined using

Kolmogorov-Smirnov or Anderson-Darling or other techniques for the assessment of normality.

If a log-transformation is applied, it is important to transform the estimated SDs back to CVs afterward. An alternative approach is using the CV-ANOVA as described by Røraas and colleagues.<sup>12</sup> Using this approach, data are transformed by dividing each subject’s measurement values by that subject’s mean value, so a distribution of values around 1 is obtained, and the CV<sub>A</sub> and CV<sub>I</sub> can be derived directly, as all subjects have a mean of 1. However, this approach does not lend itself to estimation of CV<sub>G</sub>.

**Outliers.** The assessment of outliers is important, because such aberrant values will lead to erroneous estimates of the components of biological variation if applying the method of Fraser and Harris, or a similar approach. It is important that this assessment is done using the same measure of variability that is estimated; that is, if CVs are estimated, which is usually the case, all the calculations should be performed using CVs. This can be achieved by normalizing data, for example, through log-transformation as described above. After any transformation, outlier assessments are performed at three levels: (1) between duplicates or replicates, (2) between samples within an individual, and (3) between individuals. For levels 1 and 2, typically Cochran’s test is used on the CV<sup>2</sup>, but applied to SD<sup>2</sup> if we work on log-transformed data. Failure to remove outliers in the replicates can result in a falsely high CV<sub>A</sub> and an erroneous CV<sub>I</sub> estimate, while failure to remove outliers from results from each of the individuals can result in a falsely increased CV<sub>I</sub>. Finally, outliers among the mean values of the individuals (level 3) are assessed. A simple strategy to perform this process is to use Reed’s criterion, where the difference between any mean value and the next highest or lowest value in the series should be less than one-third of the absolute range of all values. Another useful approach for the assessment of outliers between individuals is a simple graphical approach in which the mean values and the range of all these values are plotted for each individual (on the y-axis) against concentration or activity (on the x-axis). An example is provided in Fig. 8.2 and discussed later in this chapter. Failure to exclude outliers of the mean



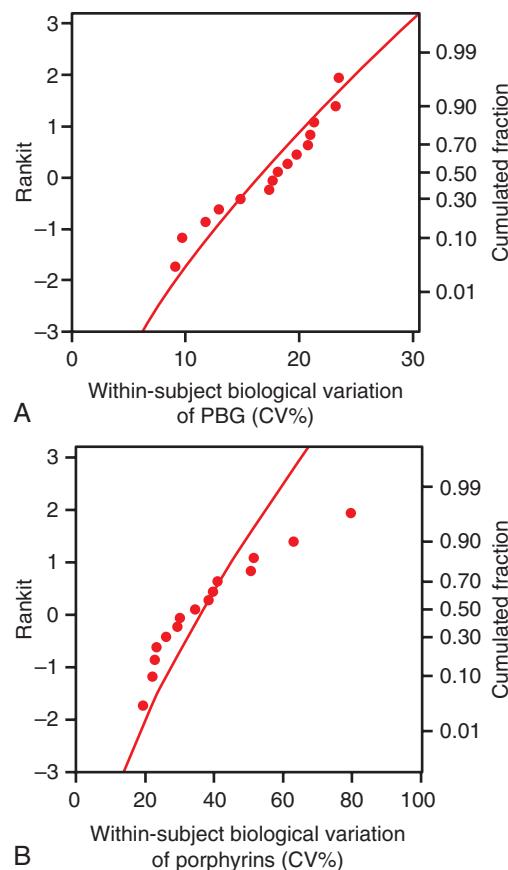
**FIGURE 8.2** Means and extreme values for serum creatinine in 27 older adults. Note: 100 µmol/L = 1.13 mg/dL. (From Fraser CG. Inherent biological variation and reference values. *Clin Chem Lab Med* 2004;42:758–764.)

values of the different individuals will result in a falsely large  $CV_G$ , and because the overall mean value will be different; this may also affect  $CV_A$  and  $CV_I$  depending on the transformation chosen. The number of outliers and concentrations that have been found and the number of datapoints used to derive the components of biological variation should always be reported.

**Homogeneity.** Applications of within-subject biological variation data, particularly for estimation of reference change values (RCVs) (which will be dealt with later), depend on the within-subject variation data, for example, the variances being homogeneously distributed. If the data are derived from a study cohort with heterogeneously distributed within-subject estimates, the results may not be representative of the population, except for an “average” individual. To remedy this, stratified analysis by subgroups should be applied if possible, or an alternative approach for deriving biological variation estimates, such as Bayesian statistics, should be used. Thus although estimates of  $CV_I$  and RCVs can be calculated, these are not generalizable to the overall population. It is therefore always necessary to check that the variances in specimens drawn from a population are “homogeneous” by definition, and consequently, that the ranked cumulative distributions of these variances are distributed around the true variance of the population according to  $\chi^2/\text{df}$  ( $\chi^2$ ) distribution for degrees of freedom (df) according to the individual sample sizes. In contrast, when a series of different variances has a dispersion around the pooled variance different from a  $\chi^2/\text{df}$  distribution, they are considered to be heterogeneous. Ideally if we have normality, a balanced design and a low  $CV_A$  compared to  $CV_I$ , then homogeneity can be illustrated by plotting the cumulated ranked fractions of within-subject variations as a function of the within-subject variation estimates on a Rankit scale (Fig. 8.3). If homogeneous, this curve will fit to the theoretical of the square root of the pooled variance multiplied by  $\chi^2/\text{df}$ . Variance homogeneity can also be tested further by Bartlett’s test. Although Cochrane’s test of the variances of the mean values is primarily an outlier test, it also can be used as a test for homogeneity.

After testing for homogeneity and outliers, it is important to indicate how many individuals and results have to be removed to obtain homogeneity of the data used to estimate the  $CV_I$ . This again provides an indication of the representative nature of the data and underscores its suitability for wide application. It is also recommended to check if excluded individuals share common traits that can explain their heterogeneity, which can be valuable information to the applicability of the estimated  $CV_I$ .

**Calculation of analytical, within-subject, and between-subject variation.** There are several estimators available: ANOVA, maximum likelihood (ML), restricted maximum likelihood (REML), minimum variance quadratic unbiased estimator (MIVQUE), and weighted analysis of means (WAM). For balanced designs, the choice of estimator is not important; however, for unbalanced designs, the estimators can yield different results and should be chosen carefully. These estimators are available in most statistical software packages using general linear models or generalized linear models. In historical publications, many have not used formal ANOVA but instead have simply used subtracted variances. The assumption is that, because preanalytical sources of variation have been minimized and can be considered negligible, the total CV ( $CV_T$ ) of a set of results from each cohort of individuals includes  $CV_A$ ,  $CV_I$ , and

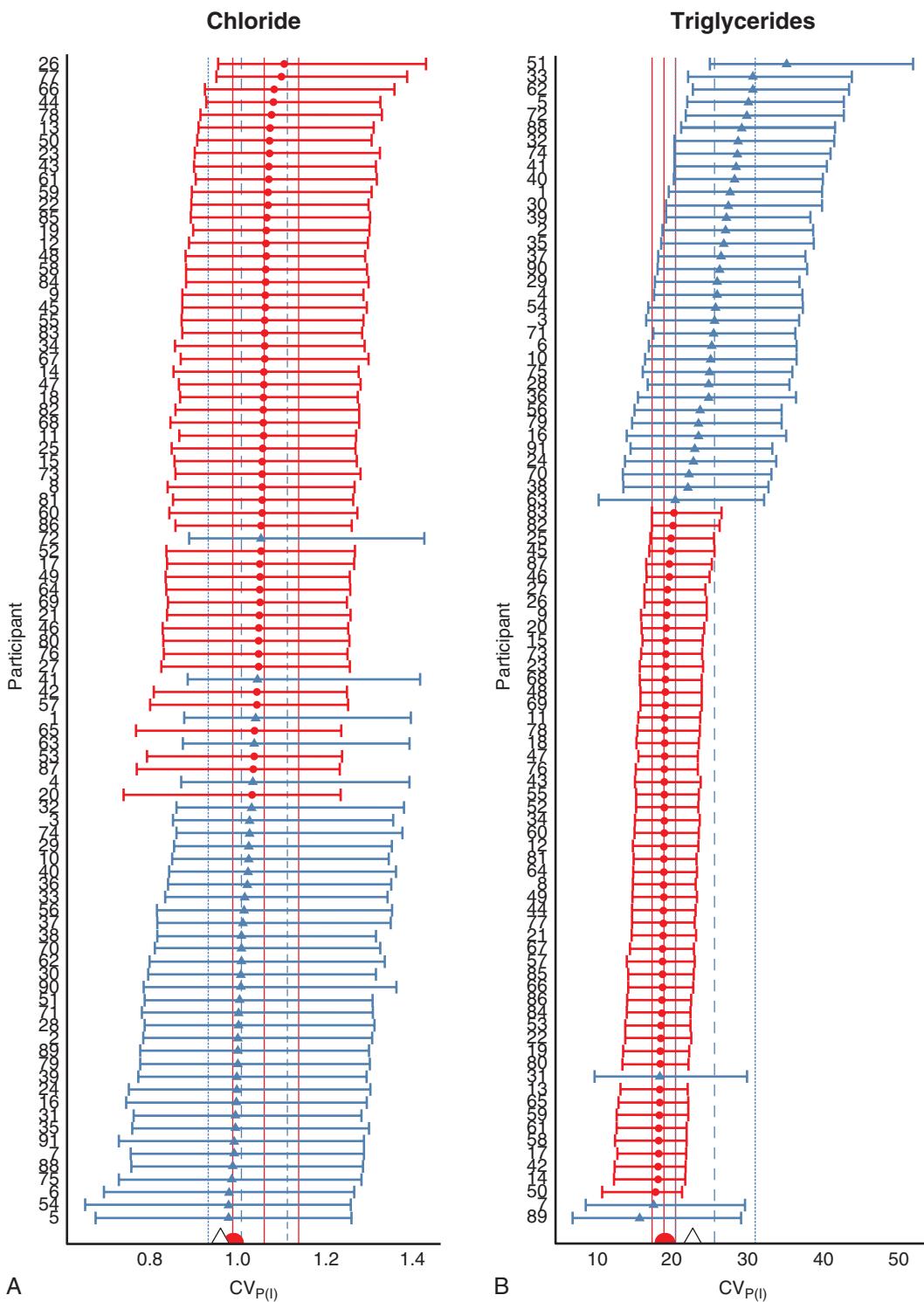


**FIGURE 8.3** Variance homogeneity plots. Rankit plots show the accumulated fractions as function of within-subject biological coefficient of variations (CVs). The filled circles represent individual CVs for healthy individuals; the line indicates the expected distribution of measured CV for “true” CV values of 17.3% for porphobilinogen (PBG) (A) and 38.1% for porphyrins (B). The Cochrane test gives values of 0.11 for PBG and 0.25 for porphyrins, and a value indicating heterogeneity is greater than 0.17. By this follows that it is not recommendable to derive common RCVs for porphyrins. (From Aarsand AK, Petersen PH, Sandberg S. 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* 2006;5:650–656.)

$CV_G$ . Then, because  $CV_T = [(CV_A)^2 + (CV_I)^2 + (CV_G)^2]^{1/2}$ , the components can be calculated by simple subtraction. However, this approach will require calculations including degrees of freedom and therefore a formal ANOVA is a simpler approach for correct calculations and can take into account unbalanced designs and further give the ability to calculate CIs. Additionally, in many of these studies,  $CV_A$  has been estimated based on quality control materials, the possibility of outliers has not been assessed, and a normal distribution and homogeneity of data are assumed; as a consequence, estimates of the components of biological variation are likely to be less precise.

### Bayesian Analysis

Applying Bayesian statistical models with different degrees of robustness using adaptive Student- $t$  distributions instead of the normal distributions to derive estimates of biological variation has been explored by Røraas and colleagues.<sup>13</sup> This Bayesian approach allows for the data to be heterogeneously distributed with different  $CV_I$  estimates for each individual (Fig. 8.4). This



**FIGURE 8.4** The estimated personal  $CV_I$ ,  $CV_{P(l)}$ , with 95% Crl (credible interval), women (red circles) and men (blue triangles) for chloride (A) and triglycerides (B) with use of a Bayesian model on the raw data from the EuBIVAS.<sup>32</sup> The participants are sorted by their  $CV_{P(l)}$ . The CV-ANOVA-derived EuBIVAS results for the different subgroups are shown on the x axis by corresponding symbols (red circle = women; white triangle = men) for both analytes. The vertical lines show the 20%, 50%, and 80% percentiles for the predicted distributions for  $CV_{P(l)}$ . The model uses an adaptive Student *t* likelihood for the distribution for samples and replicates. (From Røraas T, Sandberg S, Aarsand AK, Støve B. A Bayesian approach to biological variation analysis. *Clin Chem* 2019;65:995–1005).

has an advantage over the traditional approach, such as the nested ANOVA, and also reduces the need for laborious statistical tests as described in detail above. Furthermore, the possible subjectivity in data trimming and exclusion of outliers to achieve homogeneity and normality is avoided. A Bayesian approach can also illustrate the degree of heterogeneity, and the ability to crudely estimate personal within-subject CVs can be used to explore relevant sub-groups or correlation between  $CV_1$  and age or homeostatic set points (see Fig. 8.4). Applying the Bayesian approach on a raw log transformed data set was shown to give results comparable to a traditional approach with outlier assessments and removal when the data set was homogeneous.<sup>13</sup> However, different estimates were derived when the data showed a high degree of heterogeneity and where the traditional approach required exclusion of many data points to achieve homogeneity of the data. This indicates that for heterogeneous, “noisy” data, a Bayesian approach may provide more generalizable and reliable results, as it is not affected by these issues in the same way as the more traditional approaches. The flexibility of the Bayesian approach further allows for the analytical imprecision to be based on precision profiles and not simply as a  $CV_A$  estimate and to include a trend in the model. Furthermore, by incorporating reliable prior information (priors) into the model, based on knowledge derived from, for example, previous studies of similar analytes or another study population, precise estimates are possible even when based on small data sets.

### Retrospective Studies Using “Big Data”

#### Design

Results from patient cohorts from hospital data such as pathology laboratory databases may be used to derive biological variation estimates. This may be particularly relevant if the measurand in question is not present in matrices from apparently healthy subjects (such as unusual proteins found in myeloma) or if it would be unacceptable or unethical to collect specimens from individuals in whom the measurand is most interesting, such as children. In some of these settings, specimens could be collected from patients in different states of health and in a stable phase.<sup>10,14</sup> Assessment of data collected for diagnostic or monitoring purposes may also give us the ability to assess sub-groups, the effect of time between sampling, or other factors, without the efforts of prospectively collecting large data sets. For the majority of tests requested for a patient, the levels of the analytes are not impacted by the pathology request and may represent values obtained for the healthy population. An approach on how to use big data in the form of pathology databases for estimating within-subject biological variation and RCVs equivalent to the indirect method for the reference interval has been published.<sup>7</sup> This approach uses cohorts where at least two samples are collected for routine clinical purposes from a large number of individuals.

#### Data Analysis

The proposed method for big data from patient cohorts is using the Bhattachayra analysis<sup>15</sup> to extract the underlying distribution for the ratio between measurements, where the second measurement is divided by the first measurement for each patient. If more than two measurements are available for a patient, each sequential pair give a separate ratio to the data pool. The  $CV_{ratio}$  is calculated from the mean and SD

extracted from the underlying distribution of the ratios using the Bhattachayra analysis. The  $CV_{total}$  describing the total variability of the individual measurement results, is approximated by  $CV_{ratio}/\sqrt{2}$ . Finally, a  $CV_1$  is approximated by subtracting the appropriate  $CV_A$  from  $CV_{total}$ . This approach produced estimates of  $CV_1$  that, for most of the 26 measurands studied, showed rather good agreement with published data based on traditional approaches.<sup>7</sup> However, it must be kept in mind that the method is only applicable for normally distributed processes while the traditional approach is also valid for processes that are not normally distributed. Furthermore, the method does not provide measures of uncertainty or CI for the estimates.

### POINTS TO REMEMBER

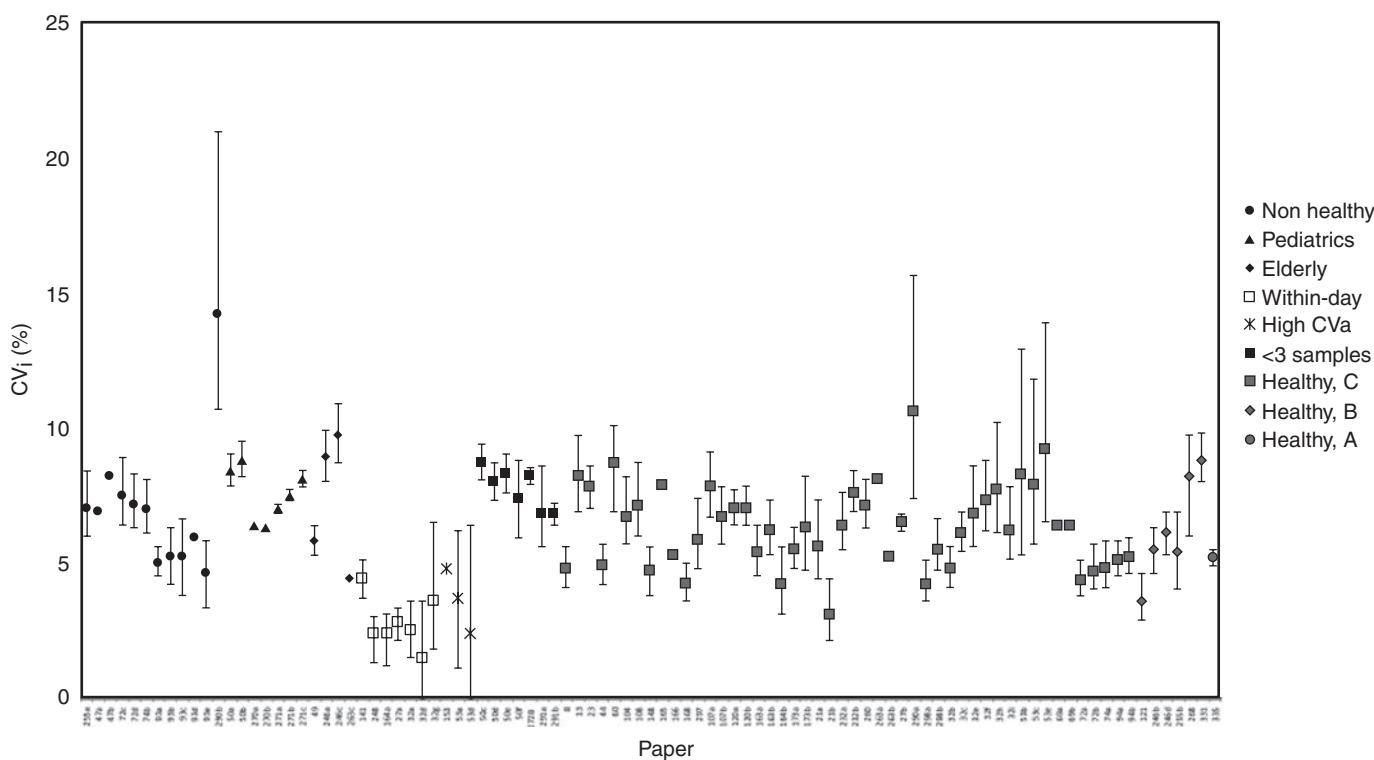
#### *Study Design and Data Treatment*

- The design of a study on biological variation should be carefully planned for its intended use.
- The optimal study design involves the use of a prospective cohort study.
- To be able to assess biological variation in subgroups, where a traditional approach may not be suited, a retrospective study using (big) data from the hospital laboratory system can be considered.
- Data analysis can be performed by using different types of ANOVA, Bayesian statistics, or in the case of big data, Bhattachayra analysis.

### QUALITY OF AVAILABLE BIOLOGICAL VARIATION DATA

#### *Historical Databases and Developments*

The applications of biological variation data for diagnosis and monitoring of disease and setting analytical performance specifications (APS) in the laboratory requires that biological variation data used for these approaches must be robust and of high quality. Furthermore, the data must be relevant to the laboratory in which they will be used, both in terms of results being transferable to the examination method in use and relevant for the populations the laboratory serves. This implies that biological variation data are reference data, and those who use published biological variation data for diagnosis and monitoring must appraise the data similarly to the approach that is usually applied when adopting population-based reference intervals from previously published studies. A review of available biological variation data in literature has shown that independent studies performed in healthy volunteers have delivered estimates that vary substantially for the same measurand (Fig. 8.5). The reason for this variation is probably multi-factorial, but important factors are likely the lack of harmonization in study design, differences in analytical principle (measuring a different measurand), and different statistical approaches applied to deliver the biological variation estimates. A considerable number of biological variation studies, with varying quality, have been published over the last 40 years. A compilation of these historical studies presenting biological variation data for a range of measurands together with APS for bias, imprecision, and total error has been available. This was the work of the Analytical Quality



**FIGURE 8.5** CV<sub>i</sub> and 95% CI estimates for total cholesterol. The different symbols indicate that the estimates have been derived in the following study populations/settings: (1) healthy adult study subjects, Biological Variation Data Critical Appraisal Checklist (BIVAC)-compliant studies; (2) elderly and pediatric subjects, BIVAC-compliant studies; (3) nonhealthy study subjects; (4) within-day studies; (5) studies with CV<sub>A</sub> estimates higher than desirable APS based on the data derived from the historical biological variation database; and (6) studies with less than three samples per subject. The numbers on x-axis refer to different biological variation studies and with letters indicating subset of data from (subgroups) derived from the same publication. (From Diaz-Garzon J, Fernandez-Calle P, Minchinela J, et al. Biological variation data for lipid cardiovascular risk assessment biomarkers. A systematic review applying the biological variation data critical appraisal checklist [BIVAC]. *Clin Chim Acta* 2019;495:467–475.)

Commission of the Spanish Society of Laboratory Medicine (SEQC<sup>ML</sup>), first presented at the Stockholm Conference in 1999.<sup>16</sup> A scoring system was applied to decide which studies were to be included, and this resulted in biological variation estimates for 358 analytes being documented, with 2-yearly updates until 2014. However, over time questions were raised regarding the robustness of estimates presented in this database, including the method for selecting studies to be included, as well as the methodology applied for reporting the common estimates. Furthermore, for some measurands estimates were based on single or a limited number of studies or derived from studies applying obsolete methods. Following the First Strategic Conference of the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) in November 2014<sup>17</sup> where APS were defined, the EFLM established a Task and Finish Group for the Biological Variation Database (TFG-BVD), with the objective to appraise the quality of biological variation data that are publicly available. The groups' terms of reference were to develop a critical appraisal check list for evaluation of biological variation studies, to use this to assess the existing literature on biological variation, and to extract essential information from those studies and to summarize the results. First, the Task and Finish Group, in collaboration with the EFLM Working Group on Biological Variation, developed the Biological Variation

Data Critical Appraisal Checklist (BIVAC),<sup>18</sup> a tool to appraise the quality of biological variation studies. Second, the groups developed a meta-analysis approach to pool estimates from BIVAC compliant studies with a similar study design to provide global biological variation estimates. Finally, the two groups designed a new quality-based biological variation database, the EFLM Biological Variation Database, which is currently populating biological variation data for a large number of measurands.<sup>19</sup>

### The Biological Variation Data Critical Appraisal Checklist

The BIVAC<sup>18</sup> has been designed as a tool to assess the methodological quality of biological variation publications by verifying whether all essential elements that may impact upon veracity and utility of the associated biological variation estimates are present. It focuses mainly on the effect of study design, the measurement procedure, and statistical handling of data on CV<sub>i</sub> estimates. It was developed based on the structure of the Biological Variation Data Reporting Checklist<sup>20</sup> developed in 2015 by the EFLM Working Group on Biological Variation, in which key elements were identified by six main minimum data set domains.

The BIVAC consists of 14 quality items, which can be awarded scores A, B, C, or D. An overview of the 14 items is

**TABLE 8.2 The Quality Items of the Biological Variation Data Critical Appraisal Checklist (BIVAC) With Achievable Scores**

Quality Item Number	Quality Item	Achievable Scores		
1	Scale	A	B	—
2	Subjects	A	B	C D
3	Samples	A	B	C D
4	Measurand	A	B	C D
5	Preanalytical procedures	A	B	C —
6	Estimates of analytical variation	A	B	C —
7	Steady state	A	B	C —
8	Outliers	A	B	C —
9	Normally distributed data	A	B	— —
10	Variance homogeneity	A	—	C —
11	Statistical method	A	B	C —
12	Confidence limits	A	—	C —
13	Number of included results	A	B	C —
14	Concentrations	A	B	— —

Adapted from Aarsand AK, Røraas T, Fernandez-Calle P, et al. The biological variation data critical appraisal checklist: a standard for evaluating studies on biological variation. *Clin Chem* 2018;64:501–514.

provided in Table 8.2. The overall BIVAC grade of a publication is based on the individual scores for each of the quality items. The grade A is achieved if the publication is fully compliant with the BIVAC; that is a score of A is awarded for all individual quality items. The grade B is awarded if the lowest score for any quality items is a B, and similarly, grade C or D if the lowest score for any quality item is a C or D, respectively. The BIVAC quality items relating to subjects (quality item 2), samples (quality item 3), and the measurand (quality item 4) are considered critical for the reliability and applicability of the biological variation estimates and these quality items therefore can be awarded a D (see Table 8.2). The BIVAC recommends that biological variation estimates derived from studies that receive one or more D scores are not applied in clinical practice. When applying biological variation data for diagnosis and monitoring, it is a prerequisite that the population from which the data has been derived is adequately characterized to ascertain transferability of results. Additionally, to compare with other studies, to deliver CIs and to generate global biological variation estimates by meta-analysis, details on the study population, samples, sample material, and timing of samples are essential. The measurement procedure must also be described in sufficient detail (quality item 4) to clarify that historical publications have assessed the same measurand as contemporary methods in use. Quality items 5, 6, and 7 refer to preanalytical procedures, estimates of  $CV_A$ , and the demonstration of steady state. These quality items must be fulfilled in order to obtain high-quality BV estimates, and can be awarded scores from A to C; but are not as essential as to elicit a D score. The BIVAC has five quality items relating to the statistical approach for data analysis, for example, quality items 8 (outliers), 9 (normally distributed data), 10 (variance homogeneity), 11 (statistical method), and 12 (confidence limits). These quality items are essential for delivery of robust biological variation estimates and to highlight their applicability and can be awarded different alternatives A, B; A and C; and A, B, and C scores (see Table 8.2). As for the quality items 1 (scale), 13 (number of included results), and 14 (concentrations), these

do not reflect the reliability of the biological variation estimates in themselves. However, these elements refer to properties of the studies which are necessary for the interpretation and application of the associated biological variation data. Over the past few years, systematic reviews and critical appraisal of biological variation studies by the use of the BIVAC have been carried out, for lipids,<sup>21</sup> enzymes,<sup>22</sup> diabetes-related measurands,<sup>23</sup> and hematologic parameters.<sup>24</sup> It has been concluded from these studies that the BIVAC quality items for outlier analysis and variance homogeneity testing are for most studies awarded a C, which implies that appropriate analysis of these elements is absent in the majority of historical publications. Though the studies in question were performed according to existing standards at the time of publication, it is evident that many publications did not adequately address essential elements, such as statistical analysis, which affected the applicability of the associated biological variation estimates derived from these studies.

The BIVAC not only allows for appraisal of already published studies, but also, in combination with the Biological Variation Data Reporting Checklist,<sup>20</sup> provides a framework that may help those planning and performing future studies. Furthermore, appraisal by the BIVAC will enable authors, reviewers, and journal editors to review studies which are fit for purpose and deliver appropriate estimates of  $CV_1$  and  $CV_G$ , accompanied by key metadata. The BIVAC defines a standard for the reporting of studies on biological variation akin to the well-known standard for reporting of studies on diagnostic accuracy (STARD),<sup>25</sup> and this will hopefully imply that in future, only studies accompanied by a complete checklist will be considered acceptable by reviewers and editors.

### Critical Review and Meta-analysis to Deliver Global Biological Variation Estimates

For frequently requested measurands, biological variation estimates obtained from many different studies are available. A meta-analysis approach for pooling published estimates to deliver global estimates has been developed.<sup>18</sup> As a first step,

biological variation publications are identified by systematic literature searches. Relevant publications are thereafter appraised by the BIVAC. As expected, the majority of biological variation studies have been performed in healthy adult volunteers, often laboratory or hospital workers, with a smaller number of studies performed in groups with different population characteristics such as age and different states of well-being (see Fig. 8.5). Nevertheless, even when studies are performed in healthy adults, widely varying estimates are reported for the same measurand. This is exemplified by total cholesterol, for which one of the highest number of publications is available. A systematic review published in 2019 identified 57 different publications, producing biological variation estimates for 95 different population subgroups, which are defined by age, sex, and health status.<sup>21</sup> The CV<sub>I</sub> estimates reported in these studies showed large variation (see Fig. 8.5). Of the 57 publications, 10 were awarded a BIVAC grade D, due to quality items 2 or 4, and 42 studies were graded C, 4 B, and 1 A. This review indicated that biological variation estimates had been published for a number of settings such as different sampling intervals (within-day, daily/weekly/monthly), different age groups (pediatric, adult, elderly), and states of well-being, but even so, few high-quality studies were identified. This indicates that there is a lack of evidence for drawing conclusions regarding differences in biological variation estimates between different states of well-being or age groups (see Fig. 8.5).

Some may argue that most of the data on the components of biological variation are inappropriate for wide use in laboratory medicine because they have generally been derived from studies on healthy individuals, and not from diseased patients who are the source of most requests for examinations.<sup>26</sup> Previous research has indicated that estimates of CV<sub>I</sub> are generally independent of the state of health, except when the measurand is one that is pathologically changed, such as tumor markers in patients with cancers.<sup>27</sup> If reviewing previously published studies for HbA<sub>1c</sub>, where biological variation has been assessed in different studies for healthy individuals and for diabetes patients,<sup>23</sup> one might draw the conclusion that diabetic patients have a higher CV<sub>I</sub> than healthy individuals. However, one study performed in parallel, where both diabetes mellitus patients and healthy individuals were investigated under the same conditions, found that CV<sub>I</sub> estimates were similar in the two groups for HbA<sub>1c</sub>. However, for glucose, the diabetes patients exhibited higher biological variation.<sup>28</sup>

Generally, very few studies fulfill all the BIVAC criteria and subsequently receive a BIVAC grade A. The European Biological Variation Study (EuBIVAS) is a large-scale BIVAC compliant biological variation study designed by the EFLM Working Group on Biological Variation.<sup>29</sup> In the EuBIVAS, six European clinical laboratories in five different countries followed a strict, detailed protocol for the recruitment of subjects and for the preanalytical phase. The study population included 91 healthy volunteers (38 males and 53 females, aged 21 to 69 years), where fasting blood samples were determined for 10 consecutive weeks and samples were frozen at  $-80^{\circ}\text{C}$  before shipment to the coordinating center. All samples for a specific participant were analyzed in the same run for each measurand, applying contemporary methods and strict quality control.<sup>29</sup> The resultant data sets underwent rigorous scrutiny and appropriate statistical analysis to deliver

biological variation estimates accompanied by CIs. So far, the EuBIVAS has produced updated biological variation estimates for various measurands<sup>30–34</sup> and additional studies are underway.

### The European Federation of Clinical Chemistry and Laboratory Medicine Biological Variation Database

The availability of robust and relevant biological variation estimates may be considered an essential requirement for effective application of clinical laboratory results. This emphasizes the importance of evidence-based estimates, with relevant metadata, being readily available for interested users. The results from the systematic reviews and appraisal processes, undertaken by the Task Group on the Biological Variation Database and the Working Group on Biological Variation in the EFLM, are freely available for users worldwide via the new Biological Variation Database, published on the EFLM website.<sup>19</sup> Here, CV<sub>I</sub> and CV<sub>G</sub> estimates and the score for each individual BIVAC quality item and the overall BIVAC grade are displayed for all appraised studies. In addition, a biological variation minimum data set is extracted and entered into the database, which encompasses approximately 30 descriptive items related to study duration, subjects' age, sex, and health status, preanalytical considerations, sample types, sampling time and interval, number of samples, information on examination methods, etc. As of April 2021, 530 publications were referenced in the database and 2573 records of biological variation data sets had been published for 232 different measurands, covering areas such as lipids, enzymes, diabetes-related measurands, hormones, hematology parameters, tumor markers, etc. Review, critical appraisal, and data entry for the remaining measurands are underway. Global estimates derived from meta-analysis have been published for more than 100 measurands. In addition to the global estimates and the detailed data for all the appraised studies, the database also offers automatic calculations of RCVs and APS. The EFLM Biological Variation Database aims to provide updated evidence-based global estimates which are easily accessible to users and that may contribute to a harmonized approach so that where appropriate, the same global biological variation estimates and APS may be applied in laboratories worldwide.

### POINTS TO REMEMBER

#### *Quality of Biological Variation Data*

- A considerable number of studies on biological variation have been published over the last 50 years, reporting varying biological variation estimates for the same measurand.
- The Biological Variation Data Critical Appraisal List is a standard designed to appraise the quality of biological variation. It consists of 14 quality items, which can be awarded a score of A, B, C, or D, which is indicative of increasing noncompliance.
- The European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) Biological Variation Database available on the EFLM website presents global CV<sub>I</sub> and CV<sub>G</sub> estimates derived by meta-analysis accompanied by automatically calculated reference change values and analytical performance specifications, as well as detailed metadata for users worldwide.

## INTERPRETATION AND USE OF DATA

### Index of Individuality

The results of most examinations in laboratory medicine are compared to conventional population-based reference intervals or sometimes fixed clinical decision-making limits. This is mandatory when previous results are unavailable, as is often the case in clinical settings of diagnosis, for case findings, and screening. However, reference intervals represent the values found in a fraction (usually 0.95) of the reference population rather than the values found in a single individual, although the value of “personal reference intervals” has recently been emphasized.<sup>35,36</sup> The ramification of biological variation on the use of reference intervals is determined by the individuality of the measurand, an aspect which has been explained in detail.<sup>37</sup> An example thereof is reproduced in part here.

Fig. 8.2 shows a graph, which as stated earlier, should be prepared by all who are generating data on the components of biological variation, to assess the visual presence of outlier observations. This shows the means and extreme values for a cohort of 27 older adults for serum creatinine concentration. Subjects 1 to 13 were women and subjects 14 to 27 were men. The conventional reference intervals for creatinine in individuals older than 55 years of age generated in the laboratory were 60 to 98 µmol/L (0.68 to 1.10 mg/dL; reference intervals may vary depending on methodology) for women and 66 to 128 µmol/L (0.75 to 1.45 mg/dL) for men. As documented previously,<sup>38</sup> analysis of the data in Fig. 8.2 allows for the following conclusions to be drawn:

- No individual has observed values that span the entire reference interval, and the range of values from each individual occupies only a small part of the dispersion of the reference interval.
- Most individuals have all observed values within the reference interval.
- The means of the observed values of most individuals lie within the reference interval, but they differ from one another.
- A few individuals have observed values that span the lower reference limit, and these individuals have values that change from usual to unusual over time.
- A few individuals have observed values that span the upper reference limit, and these individuals also have values that change from usual to unusual.

It is clear that the  $CV_I$  of creatinine (the variation around the homeostatic set points) is smaller than the  $CV_G$  (i.e., the difference between the homeostatic set points). In numerical terms,  $CV_I$  was 4.3% and  $CV_G$  was 18.3%. It is evident that  $CV_I$  is smaller than  $CV_G$ , and in such situations, the measurand is said to have marked individuality. This characteristic can be expressed mathematically as an index of individuality (II) and is best calculated as the ratio of analytical plus within-subject biological variation estimate to the between-subject biological variation estimate, mathematically:  $II = (CV_A^2 + CV_I^2)^{1/2} / CV_G$ . However, it is now widely accepted that II simply be calculated as  $CV_I / CV_G$ . This is satisfactory if  $CV_A$  is less than  $0.5 \times CV_I$ , as is often the case with modern analytical technology and methodology, since  $CV_A$  will then contribute limited analytical noise to the numerator ( $CV_A^2 + CV_I^2$ )<sup>1/2</sup>. Most commonly examined measurands, like creatinine here, have a low II, implying that they have marked individuality.

This example provides a biological explanation for the fact that serum creatinine concentration, compared to conventional reference intervals, does not have high sensitivity for the detection of mild renal impairment in an individual patient, and provides a reason for adopting estimated glomerular filtration rate, which is calculated using formulas that take age, sex, and ethnicity into account. This example also provides a sound rationale for the well-established fact that most measurands examined in laboratory medicine are not very useful in population screening or in case reports if reference limits are used as action limits.

### Consequences for Population-based Reference Intervals

The consequences of individuality were first postulated by Harris<sup>39,40</sup> who showed that, when  $CV_I / CV_G$  is high (the criterion usually applied is  $CV_I / CV_G > 1.4$ ), the distribution of values from any single individual will cover virtually the entire dispersion of the reference interval derived from values found in reference individuals. In contrast, if  $CV_I / CV_G$  is low (especially when  $CV_I / CV_G < 0.6$ ), the dispersion of values for any individual will span only a small part of the conventional population-based reference interval.

The ramifications of this individuality on the interpretation of the results of examinations are profound. When II is low, individuals may have values that are abnormal for them, but these will often still lie within the reference interval. As a consequence, these results would not be flagged by laboratories as deserving of further attention, because although they are abnormal for that individual, they are within the reference interval. Moreover, the users of the laboratory results would be highly unlikely to pay attention to such values. Thus taking one specimen from an individual and comparing the result of an examination with the population-based reference interval will not, as shown previously for creatinine, be an effective way of detecting the small changes often seen in early pathologic processes. However, when only one sample of an individual is examined, the II will have no influence on the percentage of false positives and true positives detected, irrespective of whether the upper reference limit or a selected clinical decision-making value is used. However, if a “confirmatory” measurement is performed, the II is of importance. For quantities with a very low II, which is the usual situation in laboratory medicine, a new result of measurement is likely to be close to the first and only provide limited new information. For quantities with a high II, a repeat measurement will decrease the number of true positives and false positives. In a low prevalence situation (e.g., in screening and case findings), in which it is important to prevent healthy individuals being incorrectly labelled, a positive result will “confirm” the first. In a relatively high prevalence situation (e.g., in diagnosis) in which the number of false positives is low, and it is important to discover most of the diseased patients, the measurement need not be repeated. Thus the only clear reason for a “confirmation” measurement is in a low prevalence situation when the II is high; this is rare in laboratory medicine.<sup>41,42</sup>

If II is defined as  $CV_I / CV_G$ , this ratio must be increased to make conventional population-based reference intervals of higher clinical usefulness, especially in diagnosis, case finding, and screening, when no previous results on an individual are available. This can easily be achieved as  $CV_G$  can be made smaller by stratifying (or partitioning) the data.<sup>41</sup> An example

**TABLE 8.3 Within-subject ( $CV_I$ ) and Between-subject ( $CV_G$ ) Biological Variation estimates of Urine Creatinine and Indices of Individuality (II)**

Group	$CV_I$ (%)	$CV_G$ (%)	II
Men ( $n = 7$ )	11.0	6.0	1.83
Women ( $n = 8$ )	15.7	11.0	1.42
Total	13.0	28.2	0.46

is shown in *Table 8.3* where the total cohort, II, is 0.46, and therefore the reference intervals will be of low usefulness, especially for monitoring individuals. However, where men and women are viewed separately, the II are 1.83 and 1.42, respectively, and reference intervals will be more useful. Stratification according to gender has vastly increased the usefulness of conventional population-based reference intervals. As most measurands have low II, stratification must be considered when reference intervals are being developed (see Chapter 9). If  $CV_I$  is stratified, which can happen,<sup>34</sup> this will, in contrast to a stratification of  $CV_G$ , result in a smaller II. Personalized reference intervals will be most useful for measurands with a low II.

### Reference Change Values and Differences in Serial Results

The results of examinations in laboratory medicine are used for many purposes; mostly for monitoring, either of acute disease (short term) or for chronic disease. Monitoring, by definition, means assessment of results over time. As most measurands have marked individuality and low II, conventional population-based reference intervals have disadvantages as aids to interpretation of serial results in an individual.

Harris and Yasaka<sup>43</sup> introduced the concept of the RCV, which is also sometimes called the “critical difference.” The generation and application of RCVs have been reviewed in depth.<sup>11,45</sup>

The result of one examination will have a dispersion =  $Z \times \sqrt{SD_A^2 + SD_1^2}$ , where Z is the Z-score equal to the number of SDs appropriate for the probability desired. The result of a second examination will have the same dispersion, and so the total dispersion of two results will be  $\sqrt{2} \times Z \times \sqrt{SD_A^2 + SD_1^2}$ . Thus for the change between two results in the same individual not to be explained by analytical and biological variation only, this inherent difference due to  $CV_A$  and  $CV_I$  must be exceeded, and this is the RCV. However, as we usually use CVs to describe biological variation, a transformation is needed in order to apply the formula, as detailed below.

### Reference Change Values When Coefficient of Variations and Not Standard Deviations Are Used

The original RCV formula, termed the  $RCV_{SD}$ , is only applicable to SDs, for example, changes in units. When used for changes in percent, for example as estimates given as CVs (which is a ratio), we need, in principle, to use a transformation. The reason for this is that the sum of normally distributed variables is normally distributed while the ratio between normally distributed variables is not. The relative difference between measurements M1 and M2 is defined as  $(M2 - M1)/M1 = M2/M1 - 1$ , and if M1 and M2 are normally distributed,  $M2/M1$  is not and the original RCV formula will not

provide the correct differences, which in reality are skewed. A sufficient approximation for  $RCV_{CV}$  when using a relevant  $CV_I$  estimate from a publication or a database and the laboratory's corresponding  $CV_{A,lab}$  can be estimated using the formulas below:

$$SD_A^2 = \log e(CV_A^2 + 1)$$

$$SD_1^2 = \log e(CV_1^2 + 1)$$

$$RCV_{\%} = 100\% \times \left( \exp \left[ \pm Z_a \times \sqrt{2} \times \sqrt{SD_A^2 + SD_1^2} \right] - 1 \right)$$

Typically,  $Z_a$  is 1.96 or 1.64 depending on the chosen level of probability.

Example:

$$CV_1 = 5.1\%; CV_A = 1.4\%;$$

$$SD_1^2 = \log e(0.051^2 + 1) = 0.00259762$$

$$SD_A^2 = \log e(0.014^2 + 1) = 0.00019598$$

$$RCV_{CV} = 100\%$$

$$\times \left( \exp \left( \pm 1.64 \times \sqrt{2} \times \sqrt{0.00019598 + 0.00259762} \right) - 1 \right)$$

$$= (-11.6\% + 13.1\%)$$

If, however, using estimates given as CVs in the  $RCV_{SD}$ , one would obtain the following result:

$$RCV_{\%} = 100\% \times 1.64 \times \sqrt{2} \times \sqrt{0.051 \times 0.051 + 0.014 \times 0.014}$$

$$= \pm 12.5\%$$

Using the log transformation is therefore most important when the included  $CV_I$  or  $CV_A$  estimate is high.

### Considerations When Calculating and Applying Reference Change Value

It is assumed that preanalytical sources of variation are considered constant and negligible, and in clinical and laboratory practice, this means having well-documented standard operating procedures for patient preparation and specimen collection, transport, and handling before examination, and also good training of healthcare staff performing these tasks. Moreover, it is important to realize that changes in the bias of the examination between the collections of the serial specimens can also add to the RCV; if these can be quantitated, as a difference, due to bias in percentage terms,  $\Delta B$ , the formula becomes  $RCV_{SD} = \Delta B + \sqrt{2} \times Z \times \sqrt{SD_A^2 + SD_1^2}$ . However, in a single laboratory in practice, the main source of  $\Delta B$  over time is due to recalibration, and this random bias is usually an integral component of the longer term  $CV_A$  as estimated from replicate examinations of internal quality control materials. It should be emphasized that the  $CV_A/SD_A$  used in these calculations should be comparable to the length of time between which the two samples are drawn. If this is done, it can be assumed that the bias of the examination does not change during the period of the two examinations, and the simpler formula applies.

It is often assumed that a Z-score of 1.96 for  $P < .05$  (and sometimes also 2.58 for  $P < .01$ ) are appropriate. It is almost ubiquitously stated in studies on biological variation that the  $RCV_{SD}$  is calculated as  $2.77 \times \sqrt{SD_A^2 + SD_1^2}$ . This is incorrect for a number of reasons.

First, these Z-scores are termed bidirectional (or two-tailed or two-sided), and this infers that the difference between the two serial results can be either an increase or a decrease. However, in most clinical situations, the decision making is the assessment of a significant fall (decline, decrease or reduction, for example, HbA<sub>1c</sub> after treatment for diabetes mellitus or in blood glucose after adjustment of insulin dosage), or a significant rise (for example, an increase in serum creatinine to assess acute kidney injury or serum troponin after acute chest pain). Thus unidirectional (one-tailed or one-sided) Z-scores must be used in most clinical situations to facilitate correct interpretation; these are 1.65 for  $P < .05$  and 2.33 for  $P < .01$ . Correct definitions of the clinical decision-making context and the major differences between the term “change” and “rise or fall,” and their synonyms, are required for the correct calculation of appropriate RCVs.

Second, clinical decision making is not always done at  $P < .05$ , which is the most commonly used probability in analysis of research data. The semantics used are crucial to understanding the probability that is appropriate. It will always be dependent on the consequences of the actions taken; the advantage of an action must always be weighted against the consequences/risk of no action. RCVs should be used in a spectrum of postanalytical processes, including provision of graphs and tables of change versus probability,  $\Delta$ -checking, and flagging of significant changes at different levels of probability on electronic and paper reports of results of examinations.<sup>46</sup>

In addition, the RCV for each laboratory will always depend on the CV<sub>A</sub> in that laboratory. Furthermore, generation of data on CV<sub>I</sub> is not easy, and it is important that the CV<sub>I</sub> estimate used is obtained from a population with a homogenous CV<sub>I</sub> distribution and that it is similar to the population for whom the RCV is being created. This means that when RCV is interpreted for diagnostic purposes, exceeding the RCV limits means that the change is unlikely explained by analytical and within-subject biological variation in a healthy population. When RCV is interpreted for monitoring purposes, we assume that the within-subject biological variation is similar for healthy and diseased persons and therefore changes can be explained in the same way. This will usually be the case, but in some cases the biological variation for diseased persons will be different from healthy as discussed above. In addition, the time interval used for obtaining both the CV<sub>A</sub> and the CV<sub>I</sub> must be comparable to the one used in practice. As an example, if HbA<sub>1c</sub> is controlled every 6 months in a diabetic patient, the CV<sub>A</sub> should be taken from the internal quality control during this period and should reflect the relevant concentration. An overview of the publications on CV<sub>I</sub> can be found in the EFLM database<sup>19</sup>; thus laboratories can calculate relevant RCVs by using CV<sub>A</sub> derived from their own internal quality control programs, using data close to clinical decision-making concentrations or activities, along with the CV<sub>I</sub> estimates from publications cited in the most up-to-date database to create a RCV to use for a variety of purposes.

Traditional RCVs only address how likely it is that a certain change can be explained by analytical and within-subject biological variation, but not the probability that a change in the patient’s disease state has occurred. A tool for better understanding and interpretation of measured differences in monitoring has therefore been suggested. The concepts of sensitivity, specificity, likelihood ratios, and odds used for

diagnostic test evaluations were applied to monitoring by substituting measured concentrations with measured differences.<sup>47</sup> It is suggested that this idea expanded the earlier concept of RCV by making it possible to have an estimate of the post-test odds for a certain difference to occur. Consequently, the likelihood ratio for change increases with a larger measured difference, and when used together with pretest odds or pretest probability, the post-test odds and post-test probability, which are related to the clinical situation, can be calculated. An example of this is shown in Fig. 8.6.

It has been proposed<sup>44,45,48</sup> that the probability of significance of any difference seen in clinical practice between two results can be readily calculated using a simple rearrangement of the RCV equation making the Z-score (and thus the probability), the unknown, namely:

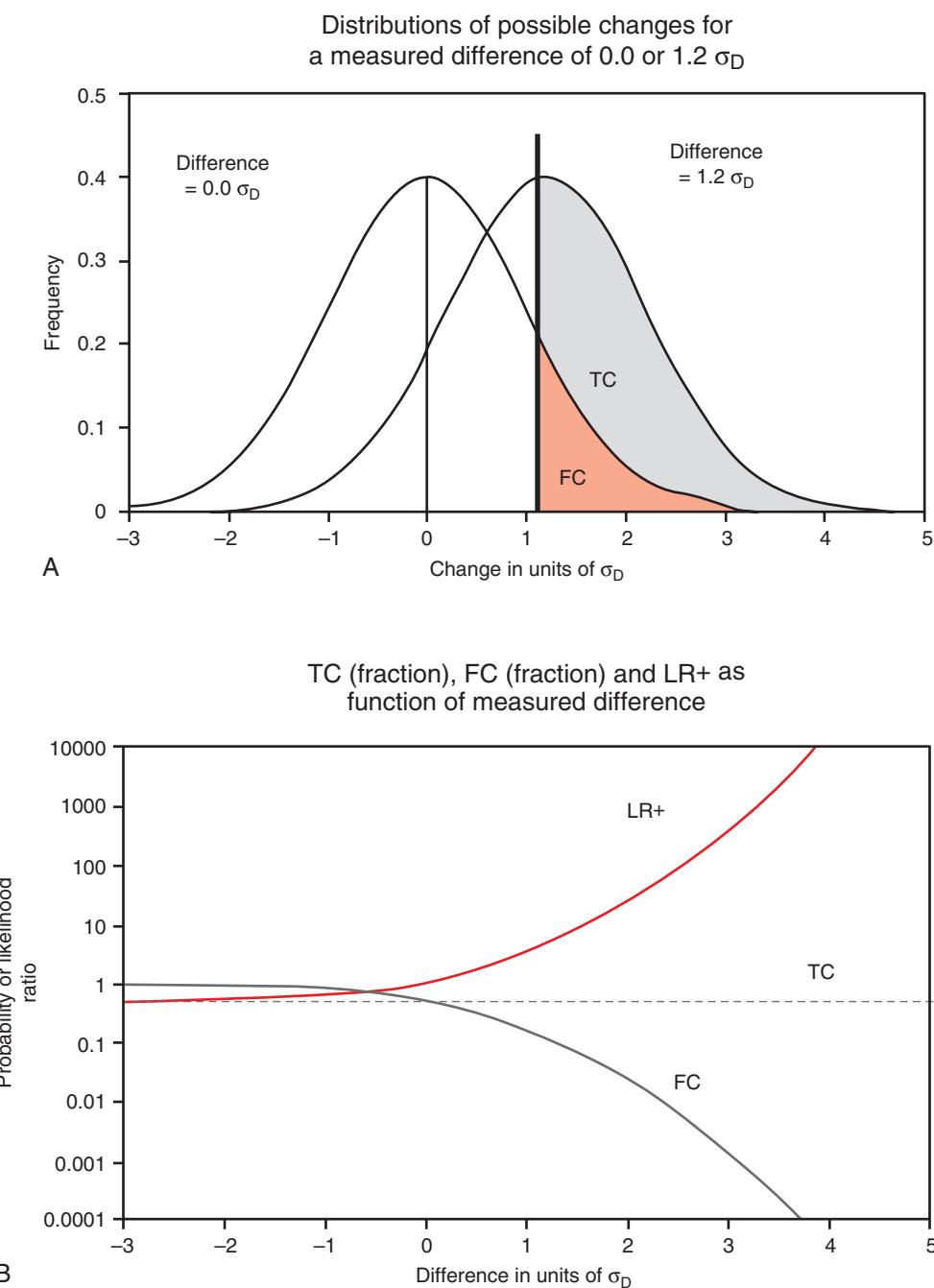
$$Z = \frac{\text{Difference}}{\left( \sqrt{2} + \sqrt{\left( \sqrt{CV_A^2} + CV_I^2 \right)} \right)}.$$

Thus it calculates the probability of significance of your observed change and lets the person interpreting the result decide, in that clinical context, what is good enough a percentage certainty for them. It also means they may not ignore a true change which has not yet reached the 95% significance threshold. This does not seem to have found its place in practice, but there is clearly a scope for adoption of this technique to enhance interpretation of serial results. Exactly as for RCVs, this application would have important consequences for CV<sub>A</sub>: the smaller the CV<sub>A</sub>, the smaller the RCV will be for any probability, and the significance of any difference seen will be of higher probability.

### Reference Change Values for More Than Two Serial Examinations

Frequently, in practice, more than two results of examinations are available for individuals over time. Using the traditional RCVs described previously, it is only possible to calculate the significance of changes between each of the two consecutive examinations. Thus an “RCV method” including all available serial results might be useful for interpretation of significant differences over time.

Mathematical methods have been developed to assess serial results from an individual, which is sometimes termed methods for time series analysis.<sup>39</sup> One model is called the “homeostatic model.” The model assumes that the measurand varies randomly around a homeostatic set point. After collection of results from a small number of examinations, the mean and SD are calculated. The next result should fall within the range calculated from the mean and the SD. Thereafter, new data from the same individual is used going forward to refine the estimates of the mean and SD for that individual to interpret whether the next new result has changed significantly from previous data. In contrast, the “random walk” model assumes that the measurand behaves randomly, and there is no homeostatic set point. The dispersion of each result from the previous result is calculated. However, instead of calculating a mean for the individual and recalculating this mean as further data are gathered, it is assumed that the most recent result is the starting point from which to assess whether change has occurred. These models have not been widely applied.



**FIGURE 8.6** (A) Illustrations of distributions of possible changes for a “measured” difference of  $0.0 = \sigma_D$  and a “measured” difference of  $1.2 = \sigma_D$ , respectively, and where  $\sigma_D = \sqrt{2} \times \sigma_{\text{within-subject}}$ . The areas describing the true positive change, TC, and false positive change, FC, are located to the right of the “measured” difference. The likelihood ratio, LR+, is  $\text{TC}/\text{FC} = \text{TP fraction}/\text{FP fraction} = 50/11.5 = 4.3$ . (B) Assuming increasing “measured” difference from  $-3 \times \sigma_D$ s to  $-5 \times \sigma_D$ s, the TC fractions = 0.5, and the FS fraction decreasing from 1 to very low values are shown (for  $2 \times \sigma_D = \text{RCV}$ , the FC fraction is 0.0227). Further, the increasing likelihood ratio LR+ function is shown as function of “measured” difference. (From Petersen PH, Sandberg S, Iglesias N, et al. “Likelihood-ratio” and “odds” applied to monitoring of patients as a supplement to “reference change value” [RCV]. *Clin Chem Lab Med* 2008;46:157–164.)

Recently, studies on both unidirectional and bidirectional changes in serial results, that used computer simulations, have been performed.<sup>49,50</sup> Factors used to multiply the first result from an individual were calculated to create the limits for constant cumulated significant differences. The factors

were shown to become a simple function of the number of results and the total CV. The first result is multiplied by the appropriate factor for an increase or a decrease, which gives the limits for a significant difference. It remains to be seen if such apparently “simple techniques” become used in practice.

## Number of Samples Needed

Often in clinical practice, only one sample is taken. Examination result variation can be reduced by multiple sampling (or multiple examinations), and the variation decreased by the square root of the number of replicates. To estimate the number of specimens needed to determine the homeostatic set point within a certain percentage error with a stated probability, a simple rearrangement of the usual standard error of the mean formula is used, namely:  $n = \left[ Z \times (CV_A^2 + CV_I^2)^{1/2} / D \right]^2$ , where  $Z$  is the  $Z$ -score appropriate for the probability, and  $D$  is the desired percentage closeness to the homeostatic set point. It is important to note that taking multiple specimens and undertaking replicate analyses does affect the overall variability of the individual examination result; the dispersion (expressed as 1 CV) can be calculated as: dispersion =  $Z \times \left[ Z \times (CV_A^2/nA \times CV_I^2/nS) \right]^{1/2}$ , where  $Z$  is the number of SDs appropriate to the probability selected,  $nA$  is the number of replicate examinations, and  $nS$  is the number of specimens. The relative magnitudes of  $CV_A$  and  $CV_I$  are important in deciding if a lower dispersion is required, whether it is better to undertake replicate examinations on one specimen or singleton examinations on multiple specimens. Further examples and the detailed reasons why knowledge of numerical data on the components of biological variation is of crucial importance have been provided in a review article.<sup>51</sup>

## Selecting the Best Specimen to Collect, and Choosing the Best Examination

It may be possible to report the results of examinations in different ways. For example, measurands in urine, such as creatinine, can be reported as concentration or output per day, and many are reported as a ratio with creatinine concentration. Moreover, for some measurands, it is possible to collect different samples for the same clinical purpose (e.g., early morning or random or timed urine specimens for low concentration albumin and protein examinations). In certain clinical situations, examinations that might be considered to have a somewhat similar purpose are available, such as serum creatinine and cystatin-C, or blood HbA<sub>1c</sub> and serum fructosamine. Knowledge of the components of biological variation can assist in making decisions about selecting the best specimen to collect and choosing the best test.<sup>52</sup>

To undertake such comparisons, the influences of biological variation should be considered. The ideal measurand would have low  $CV_I$  so that a single examination will give a good measure of the true value for that individual. Moreover, this would allow easy monitoring over time and detection of significant differences, because the RCV would be low, provided that the  $CV_A$  was also low. In addition, the ideal measurand would have no heterogeneity of  $CV_I$  among individuals and across studies and would not be dependent on age and gender and other possible confounding factors so that the simple general formulas given in this chapter would be applicable for all.

## Reporting Results

Laboratory results are usually reported and compared to a reference interval. As mentioned earlier, most measurands have a low index of individuality and therefore population-based reference intervals are often of limited value. Since many individuals already have a previous measurement of the same measurand, it is possible to provide information on the RCV for that specific measurand. This can be done by

implementing an auto-verification procedure based on RCV<sup>6</sup> (see Chapter 6 in this book) and it can be reported to the physicians by using an alert next to the result, indicating that the difference from the previous result is higher or lower than what can be expected from analytical and biological variation alone, for example, indicating that it exceeds the RCV.<sup>7</sup>

## Method Development and Evaluation

The introduction of new tests is an ongoing task for the IVD industry and thereafter for most medical laboratories. Some years ago, Zweig and Robertson suggested that the introduction of a new procedure should be similar to the structured evolution of a new drug through phase trials and should follow a linear model.<sup>53</sup> In practice, however, it is more common to follow a dynamic cyclic model moving back and forth between different components. These components which are essential in the clinical pathway are: analytical performance, clinical performance, clinical effectiveness, and cost effectiveness.<sup>54</sup> First of all, it is important to define clinical goals and how the intended application of the biomarker in the clinical pathway should drive each component of test evaluation. It is important to emphasize the interaction of the different components, and that clinical effectiveness data should be fed back to refine analytical and clinical performances to achieve improved outcomes. Desirable clinical performance criteria for a test is the ability to predict a diagnosis or clinically significant event. To be able to do that it is, among other factors, important that the test has a high signal-to-noise ratio, for example, a single test result is likely to detect a clinically significant change and to differentiate it from what can be explained by biological and analytical variation.

Therefore there is a need to generate and apply data on biological variation early on in any examination.<sup>38,54</sup> Moreover, the data are also necessary for objective analysis of the often somewhat subjective guidelines from professional bodies. The latter which provide recommendations on interpretation of the numerical results of examinations and on examination performance specifications; however, these recommendations often do not take into account analytical or biological variation.<sup>51</sup>

## POINTS TO REMEMBER

### *Interpretation and Use of Data*

- A low index of individuality ( $CV/CV_G$ ) diminishes the usefulness of population-based reference intervals.
- The reference change value (RCV) gives information about how likely it is that the difference between two serial results can be explained by analytical and biological variation.
- Estimates of biological variability are important when developing and evaluating new biomarkers.

## ANALYTICAL PERFORMANCE SPECIFICATIONS

APS are in principle the numerical standards of examination performance required to facilitate optimum patient care. There are many strategies available to set these specifications, and these have been described over time as this facet

of laboratory medicine has evolved.<sup>55</sup> A concise historical perspective has been published.<sup>56</sup>

Following pioneering studies done in the United States<sup>3</sup> on the definition of the components of biological variation, a College of American Pathologists conference held in 1976 supported the concept that specifications should be best based on biology.<sup>57</sup> In the conclusion from the conference it was stated:

"For group screening, in which an individual is to be selected from a population, a specification for imprecision ( $CV_A$ ) is defined as:

$$CV_A = 0.5 \times \sqrt{CV_I^2 + CV_G^2}$$

and "For individual single and multipoint testing, in which an individual is evaluated on the basis of discrimination values:  $CV_A = 0.5 \times CV_I$ ."

The Stockholm Conference held in 1999 on "Strategies to set global analytical quality specifications in laboratory medicine" advocated the ubiquitous application of a hierarchical structure of approaches, based on a model proposed by Fraser and Petersen.<sup>58</sup> The hierarchy had five levels, namely: (1) evaluation of the effect of analytical performance on clinical outcomes in specific clinical settings; (2) evaluation of the effect of analytical performance on clinical decisions in general, using (a) data based on components of biological variation or (b) analysis of clinicians' opinions; (3) published professional recommendations from (a) national and international expert bodies or (b) expert local groups or individuals; (4) performance goals set by (a) regulatory bodies or (b) organizers of external quality assessment (EQA) schemes; and (5) goals based on the current state of the art as (a) demonstrated by data from EQA or proficiency testing (PT) scheme, or (b) found in current publications on methodology.<sup>59</sup>

Because laboratory medicine has evolved considerably over the last few years, the EFLM organized the First Strategic Conference in 2014 in Milan on how to define APS. The presentations are available on the EFLM website.<sup>60</sup> The consensus statement and formal papers emanating from the speakers are included in a Special Issue of Clinical Chemistry and Laboratory Medicine,<sup>17</sup> and provide an invaluable resource on all aspects of setting analytical performance standards, including the generation and application of data on biological variation. In the consensus statement from Milan, the hierarchy was simplified and represented by three different models to set APS<sup>61</sup>:

Model 1. Based on the effect of analytical performance on clinical outcomes

This can, in principle, be established using different types of studies:

- Direct outcome studies—investigating the impact of analytical performance of the test on clinical outcomes;
- Indirect outcome studies—investigating the impact of analytical performance of the test on clinical classifications or decisions and thereby on the probability of patient outcomes, such as by simulation or decision analysis.

The advantage of this approach is that it addresses the influence of analytical performance on clinical outcomes that are relevant to patients and society. The primary disadvantage is that it is only useful for examinations where the links between the test, clinical decision making, and clinical outcomes are straightforward and strong. Furthermore, analytical specifications derived from direct or indirect outcome studies will often be influenced by the current measurement quality and results may vary according to the actual test method used, the investigated population, and healthcare settings.

Model 2. Based on components of biological variation of the measurand

This attempts to minimize the ratio of "analytical noise" to the "biological signal." The advantage is that it can be applied to most measurands for which population-based or subject-specific biological variation data can be established. There are limitations to this approach, including the need to carefully assess the relevance and validity of the biological variation data published and the rigor of study design, whose aspects that are dealt with elsewhere in this chapter.

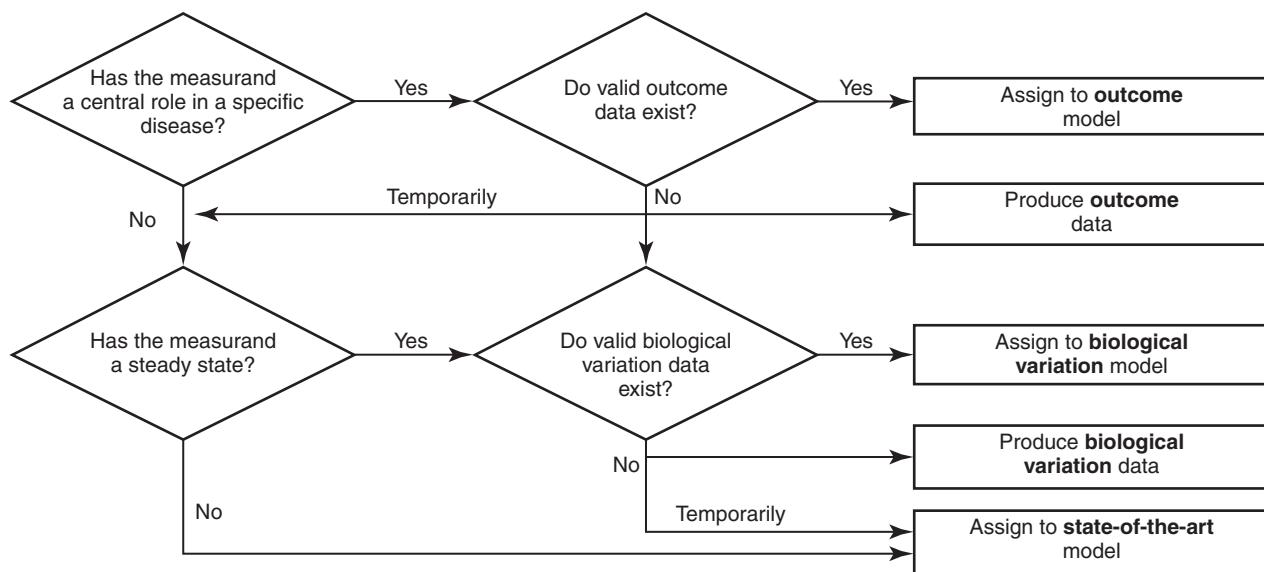
Model 3. Based on state of the art

This relates to the highest level of analytical performance technically achievable. The advantage of this model is that state-of-the-art performance data are readily available. The disadvantage is that there may be no relationship between what is technically achievable and what is needed to minimize the ratio of "analytical noise" to the biological signal or needed to obtain an improved clinical outcome. This approach is sometimes used by external quality assurance (EQA)/PT providers at least for some of their analytes.<sup>62</sup> An overview of the different models and their advantages and disadvantages is provided in Table 8.4.

The three models use different principles, and some models will be better suited for certain measurands than for others. An overview of principles for how to allocate measurands to

**TABLE 8.4 Advantages and Disadvantages of the Three Models Used to Set Analytical Performance Specifications**

Model Based on	Study/Principle	Advantage	Disadvantage
Clinical outcomes	Outcome studies (type 1a and type 1b)	Address the needs of patients and society	Difficult to perform studies. Possible for only a limited number of measurands
Biological variation	Studies on biological variation Analytical noise vs. biological signal	Can be applied to most measurands	Reliable data can be difficult to obtain
State of the art	Empirical data	Easy to obtain data	Does not relate to what is medically needed (outcome) or to noise/signal minimalization



**FIGURE 8.7** Workflow for assignment of a measurand to a defined analytical performance specification model starting from the upper left: “Has the measurand a central role in a specific disease.” (From Ceriotti F, Fernandez-Calle P, Klee GG, et al. Criteria for assigning laboratory measurands to models for analytical performance specifications defined in the 1st EFLM Strategic Conference. *Clin Chem Lab Med* 2017;55:189–194.)

different models in an article by Ceriotti and colleagues<sup>63</sup> and is summarized in Fig. 8.7.

Some measurands could have different performance specifications defined when the test has multiple intended clinical applications. For example, performance specifications could be defined for blood glucose in a critical care setting by simulation of the impact of the test on probable patient outcomes (model 1b), for self-monitoring of blood glucose in type 1 diabetes or diagnosis of gestational diabetes by clinical outcome studies (model 1a) or by a more general approach based on biological variation (model 2).

Although application of model 1 is difficult and probably will be limited to a few measurands, it is in principle the “best” model. A review of the different ways to estimate APS based on outcome studies using method 1b identified 82 studies, most of which evaluated the impact of uncertainty (imprecision, bias, biological variation) on diagnostic/clinical accuracy.<sup>62</sup> Surprisingly, however, whereas most of the studies simulated the impact of bias and imprecision, only 22% took biological variation into account. A common analytical framework underpinning the various methods was identified, consisting of three key steps: (a) calculation of “true” test values; (b) calculation of “measured” test values incorporating uncertainty; and (c) calculation of the “impact” of discrepancies between (a) and (b) on specified outcomes.

As it is the users of the results of examinations in laboratory medicine that make use of the data provided, it might be believed that they should be able to inform about the analytical quality required to facilitate decision making, model 1b. One way to do this is with clinical vignettes, whereby the clinician provides information about the change in serial results that is required to make a clinical decision. Some early studies were carried out, but these were not ideal, as it was assumed that changes were all due to analytical imprecision, and within-subject biological variation was not considered.<sup>65,66</sup>

The principle behind the studies should be that a critical difference, CD (corresponding to the RCV), in serial results is due to pre-examination sources of variation ( $CV_p$ ),  $CV_A$ ,  $CV_b$ , and changes in bias ( $\Delta B$ ):

$$CD = \sqrt{2} \times Z \times \sqrt{CV_p^2 + CV_b^2 + CV_A^2 + \Delta B}$$

Now, let  $CV_p$  and  $\Delta B$  be zero and rearrange the equation:

$$CV_A = \sqrt{\frac{CD^2}{2 \times Z^2} - CV_b^2}$$

Studies have applied this model for a variety of measurands including involving patients undertaking self-testing.<sup>67</sup> However, the derived APS depends on the clinical setting, how the question is phrased, how close the result was to a decision limit or reference limit, and probably on current examination performance. Moreover, there was large inter-clinician variation in responses. This is also a type 1b approach, but again it is hampered by the fact that the clinicians are “used” to using a test with a certain analytical performance and in their own specific clinical situations.

### Analytical Performance Specifications Based on Biological Variation

It has been considered that generally applicable APS for the laboratory in which most measurands are used for multiple purposes will be best based on biological variation data. Updated and critically appraised biological variation data can now be found on the database hosted by EFLM.<sup>19</sup> A number of EQA schemes base their acceptable examination performance on a combination of the criteria from the Milan conference, especially a combination of state of the art and biological variation (models 2 and 3). An overview is given in a report from one of the groups established after the Milan conference.<sup>68</sup>

### Specifications for Analytical Imprecision

This approach became of even greater interest as the quantity of data on the components of biological variation increased. In most cases, the examination performance specification for  $CV_A$  is simply taken as  $CV_A = 0.5 \times CV_B$ , with the rationale that the noise (analytical imprecision) is then small (12%) compared to the within-subject biological variation when using this 0.5 factor.

### Specifications for Analytical Bias

Examination bias was not mentioned in the first documents on performance specifications, possibly because the laboratories all had their own reference intervals to which the results of their examinations were compared. For diagnosis, it was therefore proposed that the total analytical variation should be less than 0.5 of the total biological variation (including within- and between-subject variation).<sup>4</sup> However, as interest in harmonization of data across time and geography developed, it was realized that harmonization of reference intervals was important, and it was proposed that the APS for bias ( $B$ ), to allow use of harmonized reference intervals, was  $B < 0.25 \times (CV_I^2 + CV_G^2)^{1/2}$ .<sup>69</sup> It must be emphasized, however, that this was not intended as a general formula for a bias specification, but to be used for diagnosis when laboratories wanted to share common reference intervals. Other models combine analytical (state of the art) and biological variation.<sup>70,71</sup> What formulae to apply should be dependent on the situation in which it is to be used.

### Specifications for Total Error or Measurement Uncertainty

As the concepts of total laboratory quality management evolved, and the idea that random and systematic sources of variation (imprecision and bias) were both important, it became clear that total analytical error was important clinically.<sup>72,73</sup> It was proposed that a linear model of combining imprecision and bias could be used to set APS for total error (TEa) as a simple linear addition; for  $P < .05$ :  $TEa < 1.64 \times 0.5 CV_I + 0.25(CV_I^2 + CV_G^2)^{1/2}$  where the latter represents the assay bias and the  $1.64 \times 0.5 CV_I$  adds imprecision at the 95% probability level.<sup>74</sup> It should be emphasized that this model for combining bias and imprecision is only one of the models available; the advantages and disadvantages of these have been discussed in detail,<sup>73</sup> and the disadvantages of this particular model have been recently restated where it is emphasized that there is no theoretical basis for combining imprecision and bias in a linear way and that it results in overestimation of the allowable total error.<sup>75,76</sup> However, the linear error model is widely used in current practice in laboratory medicine because the formula is simple to use and to calculate, albeit sometimes with a different multiplier for the imprecision term, to deliver different levels of probability. Because the fundamental principles in the "Guide to the expression of uncertainty of measurement" (GUM) are that bias should be eliminated, if possible, and all sources of variation should be added linearly as variances, possible APS for measurement uncertainty (MU) is that  $MU < Z \times 0.5 \times CV_I$ .

The EFLM established a Task and Finish Group to address the difficulties with the total error concept and to compare it to measurement uncertainty.<sup>77</sup> In principle the

expanded uncertainty performance specification should combine the uncertainty established in the individual laboratory, as well as the uncertainty accumulated along all the steps of the metrological traceability chain. It is proposed that no more than 30% of the total uncertainty budget, established by appropriate APS, should be consumed by the uncertainty of reference measurement methods and approximately 20% of the total budget consumed by the manufacturer's calibration and value transfer protocol. The remaining 50% should be available for the commercial system imprecision (including the batch-to-batch variation of the reagents) and individual laboratory performance in order to fulfil the uncertainty goal.<sup>78</sup> The concept of uncertainty has yet to show its strength in laboratory medicine mainly due to difficulties in the calculations. However, ISO has recently released a document for a practical approach to calculate measurement uncertainty.<sup>79</sup>

### Minimum, Desirable, and Maximum Analytical Performance Specifications

A three-tier model for APS has been proposed, giving minimum, desirable, and maximum APS based on biological variation using  $0.25 CV_A$ ,  $0.5 CV_A$ , and  $0.75 CV_A$  for imprecision and  $0.125(CV_I^2 + CV_G^2)^{1/2}$ ,  $0.25(CV_I^2 + CV_G^2)^{1/2}$ , and  $0.325(CV_I^2 + CV_G^2)^{1/2}$  for bias, respectively.<sup>76,80</sup> However, there is no sound theoretical basis for this.

### POINTS TO REMEMBER

#### Analytical Performance Specifications

- Three different models can be used to establish analytical performance specifications: (1) to examine the impact of analytical performance on clinical outcomes; (2) to minimize analytical noise compared to biological variation; and (3) to use state of the art of examination methods.
- For each of the models, there are "submodels" to set concrete analytical performance specifications for imprecision, bias, total error, and uncertainty.
- The selection of the model and method to set analytical performance specifications should be decided by the specific measurand and its intended use.

### OVERALL CONCLUSIONS

Numerical estimates of within-subject and between-subject biological variation are in most cases best generated by prospectively examining a series of specimens taken from a cohort of individuals, followed by statistical analysis of the sources of variation. The proper design and performance of such studies is complex. However, databases of estimates are available that facilitate application in determining the individuality of a measurand and the usefulness of conventional population-based reference intervals, the statistical significance of changes in serial results from an individual, APS for imprecision, bias, total error, measurement uncertainty, and other characteristics. The biological variation estimates have many other uses. Recent recommendations should be followed in the generation and application of biological variation data.

## SELECTED REFERENCES

8. Fraser CG, Harris EK. Generation and application of data on biological variation in clinical chemistry. *Crit Rev Clin Lab Sci* 1989;27:409–37.
9. Røraas T, Petersen PH, Sandberg S. Confidence intervals and power calculations for within-person biological variation: effect of analytical imprecision, number of replicates, number of samples, and number of individuals. *Clin Chem* 2012;58: 1306–13.
12. Røraas T, Støve B, Petersen PH et al. Biological Variation: The Effect of Different Distributions on Estimated Within-Person Variation and Reference Change Values. *Clin Chem* 2016;62: 725–36.
13. Røraas T, Sandberg S, Aarsand AK et al. A Bayesian Approach to Biological Variation Analysis. *Clin Chem* 2019;65:995–1005.
16. Ricos C, Alvarez V, Cava F, et al. Current databases on biological variation: pros, cons and progress. *Scand J Clin Lab Invest* 1999;59:491–500.
17. Panteghini M, Sandberg S. 1th EFLM Strategic Conference “Defining Analytical Performance Goals - 15 years after the Stockholm Conference.” *Clin Chem Lab Med*. 2015;53: 829–958.
18. Aarsand AK, Raas TR, Fernandez-Calle P, et al. The Biological Variation Data Critical Appraisal Checklist: A Standard for Evaluating Studies on Biological Variation. *Clin Chem* 2018;64:501–514.
19. The EFLM Biological Variation Database: <https://biological-variation.eu/> [Accessed February 2020].
29. Carobene A, Strollo M, Jonker N, et al. Sample collections from healthy volunteers for biological variation estimates’ update: a new project undertaken by the Working Group on Biological Variation established by the European Federation of Clinical Chemistry and Laboratory Medicine. *Clin Chem Lab Med* 2016;54:1599–1608.
47. Petersen PH, Sandberg S, Iglesias N, et al. ‘Likelihood-ratio’ and ‘odds’ applied to monitoring of patients as a supplement to ‘reference change value’ (RCV). *Clin Chem Lab Med* 2008;46:157–164.
52. Fraser CG ed. *Biological Variation: From Principles to Practice*. Washington; AACC Press: 2001
54. Horvath AR, Lord SJ, Stjohn A, et al. From biomarkers to medical tests: the changing landscape of test evaluation. *Clin Chim Acta* 2014;427:49–57.
56. Fraser CG. The 1999 Stockholm Consensus Conference on quality specifications in laboratory medicine. *Clin Chem Lab Med* 2015;53:837–40.
59. Kenny D, Fraser CG, Petersen HP et al. Stockholm Consensus Statement1999. *Scand J Clin Lab* 1999;59:585.
60. 1th EFLM Strategic Conference; <https://www.eflm.eu/site/page/last/1054> (Accessed April 2021)
61. Sandberg S, Fraser CG, Horvath AR et al. Defining analytical performance specifications: Consensus Statement from the 1st Strategic Conference of the European Federation of Clinical Chemistry and Laboratory Medicine. *Clin Chem Lab Med* 2015;53:833–35.
63. Ceriotti F, Fernandez-Calle P, Klee GG, et al. Criteria for assigning laboratory measurands to models for analytical performance specifications defined in the 1st EFLM Strategic Conference. *Clin Chem Lab Med* 2017;55:189–94.
64. Smith AF, Shinkins B, Hall PS et al. Toward a Framework for Outcome-Based Analytical Performance Specifications: A Methodology Review of Indirect Methods for Evaluating the Impact of Measurement Uncertainty on Clinical Outcomes. *Clin Chem* 2019;65:1263–74
69. Gowans EM, Petersen PH, Blaabjerg O et al. Analytical goals for the acceptance of common reference intervals for laboratories throughout a geographical area. *Scand J Clin Lab Invest* 1988;48:757–64.
75. Oosterhuis WP. Gross overestimation of total allowable error based on biological variation. *Clin Chem*. 2011;57:1334–36.

## REFERENCES

1. Fraser CG, Sandberg S. Biological variation. In: Rifai N, Horvath A R, Wittwer CT, eds. 6th ed. Tietz textbook of Clinical Chemistry and Molecular Biology. 6th Ed. St. Louis: Elsevier 2018:157–70.
2. Schneider AJ. Some Thoughts on Normal, or Standard, Values in Clinical Medicine. *Pediatrics* 1960;26:973–84.
3. Williams GZ, Young DS, Stein MR et al. Biological and Analytic Components of Variation in Long-Term Studies of Serum Constituents in Normal Subjects. I Objectives, Subject Selection, Laboratory Procedures, and Estimation of Analytic Deviation. *Clin Chem* 1970;16:1016–21.
4. Harris EK, Kanofsky P, Shakarji G et al. Biological and Analytic Components of Variation in Long-Term Studies of Serum Constituents in Normal Subjects. II Estimating biological components of variation. *Clin Chem* 1970;16:1022–27.
5. Cotlove E, Harris EK, Williams GZ. Biological and Analytic Components of Variation in Long-Term Studies of Serum Constituents in Normal Subjects. III Physiological and Medical Implications *Clin Chem* 1970;16:1028–32.
6. Young DS, Harris EK, Cotlove E. Biological and Analytic Components of Variation in Long-Term Studies of Serum Constituents in Normal Subjects. IV Results of a Study Design to Eliminate Long-Term Analytic Deviations. *Clin Chem* 1971;17:403–10.
7. Jones GRD. Estimates of Within-Subject Biological Variation Derived from Pathology Databases: An Approach to Allow Assessment of the Effects of Age, Sex, Time Between Sample Collections, and Analyte Concentration on Reference Change Values. *Clin. Chem* 2019;65:579–88.
8. Fraser CG, Harris EK. Generation and application of data on biological variation in clinical chemistry. *Crit Rev Clin Lab Sci* 1989;27:409–37.
9. Røraas T, Petersen PH, Sandberg S. Confidence intervals and power calculations for within-person biological variation: effect of analytical imprecision, number of replicates, number of samples, and number of individuals. *Clin Chem* 2012;58:1306–13.
10. Kristoffersen A-H, Petersen PH, Sandberg S. A model for calculating the within-subject biological variation and likelihood ratios for analytes with a time-dependent change in concentrations; exemplified with the use of D-dimer in suspected venous thromboembolism in healthy pregnant women. *Ann Clinical Biochem* 2012;49:561–69.
11. Aarsand AK, Røraas T, Sandberg S. Biological variation - reliable data is essential. *Clin Chem Lab Med*, 2015;53:153–54.
12. Røraas T, Støve B, Petersen PH et al. Biological Variation: The Effect of Different Distributions on Estimated Within-Person Variation and Reference Change Values. *Clin Chem* 2016; 62:725–36.
13. Røraas T, Sandberg S, Aarsand AK et al. A Bayesian Approach to Biological Variation Analysis. *Clin Chem* 2019;65:995–1005.
14. Fraser CG. Making better use of differences in serial laboratory results. *Ann Clin Biochem* 2012;49:1–3.
15. Battacharya C. A simple method of resolution of a distribution into Gaussian components. *Biometrics* 1967;23:115–135.
16. Ricos C, Alvarez V, Cava F, et al. Current databases on biological variation: pros, cons and progress. *Scand J Clin Lab Invest* 1999;59:491–500.
17. Panteghini M, Sandberg S. 1th EFLM Strategic Conference “Defining Analytical Performance Goals - 15 years after the Stockholm Conference.” *Clin Chem Lab Med*. 2015;53:829–958.
18. Aarsand AK, Raas TR, Fernandez-Calle P, et al. The Biological Variation Data Critical Appraisal Checklist: A Standard for Evaluating Studies on Biological Variation. *Clin Chem* 2018; 64:501–514.
19. The EFLM Biological Variation Database: <https://biological-variation.eu/> [Accessed February 2020].
20. Bartlett WA, Braga F, Carobene A, et al. A checklist for critical appraisal of studies of biological variation. *Clin Chem Lab Med* 2015;53:879–85.
21. Diaz-Garcon Marco J, Fernandez-Calle P, Minchinela J, et al. Biological variation data for lipid cardiovascular risk assessment biomarkers. A systematic review applying the biological variation data critical appraisal checklist (BIVAC). *Clin Chim Acta* 2019;495:467–75.
22. Carobene A, Braga F, Røraas T et al. A systematic review of data on biological variation for alanine aminotransferase, aspartate aminotransferase and gamma-glutamyl transferase. *Clin Chem Lab Med* 2013;51:1997–2007.
23. González-Lao E, Corte Z, Simón M, et al. Systematic review of the biological variation data for diabetes related analytes. *Clin Chim Acta*. 2019;488:61–7.
24. Coskun A, Braga F, Carobene A, et al. Systematic review and meta-analysis of within-subject and between-subject biological variation estimates of 20 haematological parameters. *Clin Chem Lab Med*. 2020;58:25–32.
25. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: An Updated List of Essential Items for Reporting Diagnostic Accuracy Studies. *Clin Chem*. 2015;61:1446–52.
26. Lawson N. Is variation in biological variation a problem? *Ann Clin Biochem* 2007;44:319–20.
27. Ricos C, Iglesias N, García-Lario JV, et al. Within-subject biological variation in disease: collated data and clinical consequences. *Ann Clin Biochem* 2007;44:343–52.
28. Carlsen S, Petersen PH, Skeie S, et al. Within-subject biological variation of glucose and HbA(1c) in healthy persons and in type 1 diabetes patients. *Clin Chem Lab Med* 2011;49:1501–7.
29. Carobene A, Strollo M, Jonker N, et al. Sample collections from healthy volunteers for biological variation estimates' update: a new project undertaken by the Working Group on Biological Variation established by the European Federation of Clinical Chemistry and Laboratory Medicine. *Clin Chem Lab Med* 2016;54:1599–1608.
30. Carobene A, Aarsand AK, Guerra E, et al. European Biological Variation Study (EuBIVAS): Within- and Between-Subject Biological Variation Data for 15 Frequently Measured Proteins. *Clin Chem* 2019;65:1031–41
31. Carobene A, Guerra E, Locatelli M, et al. Biological variation estimates for prostate specific antigen from the European Biological Variation Study; consequences for diagnosis and monitoring of prostate cancer. *Clin Chim Acta* 2018;486:185–91.
32. Aarsand AK, Díaz-Garzón J, Fernandez-Calle P, et al. The EuBIVAS: Within- and Between-Subject Biological Variation Data for Electrolytes, Lipids, Urea, Uric Acid, Total Protein, Total Bilirubin, Direct Bilirubin, and Glucose. *Clin Chem* 2018;64:1380–93.
33. Carobene A, Røraas T, Sølvik UØ, et al. Biological Variation Estimates Obtained from 91 Healthy Study Participants for 9 Enzymes in Serum. *Clin Chem* 2017;63:1141–50.
34. Carobene A, Marino I, Coskun A, et al. The EuBIVAS Project: Within-and Between-Subject Biological Variation Data for Serum Creatinine Using Enzymatic and Alkaline Picrate

- Methods and Implications for Monitoring. *Clin Chem* 2017; 63:1527–36
35. Zaninetti C, Biino G, Noris P et al. Personalized reference intervals for platelet count reduce the number of subjects with unexplained thrombocytopenia. *Haematologica*. 2015;100:E338–40.
36. Coşkun A, Sandberg S, Unsal I et al. Personalized reference intervals in laboratory medicine: A new model based on within-subject biological variation. *Clin Chem*. 2021;67:374–84.
37. Fraser CG. Inherent biological variation and reference values. *Clin Chem Lab Med* 2004;42:758–64.
38. Fraser CG. Data on biological variation: essential prerequisites for introducing new procedures? *Clin Chem* 1994;40:1671–3.
39. Harris EK. Some theory of reference values. II. Comparison of some statistical models of intraindividual variation in blood constituents. *Clin Chem* 1976;22:1343–50.
40. Harris EK. Statistical aspects of reference values in clinical pathology. *Prog Clin Pathol* 1981;8:45–66.
41. Petersen PH, Fraser CG, Sandberg S et al. The index of individuality is often a misinterpreted quantity characteristic. *Clin Chem Lab Med* 1999;37:655–61.
42. Petersen PH, Sandberg S, Fraser CG et al. Influence of index of individuality on false positives in repeated sampling from healthy individuals. *Clin Chem Lab Med* 2001;39:160–165.
43. Harris EK, Yasaka T. On the calculation of a “reference change” for comparing two consecutive measurements. *Clin Chem* 1983;29:25–30.
44. Deleted in review.
45. Fraser CG. Improved monitoring of differences in serial laboratory results. *Clin Chem*. 2011;57:1635–37.
46. Plebani M, Sciacovelli L, Bernardi D et al. What information on measurement uncertainty should be communicated to clinicians, and how? *Clin Biochem* 2018;57:18–22.
47. Petersen PH, Sandberg S, Iglesias N, et al. ‘Likelihood-ratio’ and ‘odds’ applied to monitoring of patients as a supplement to ‘reference change value’ (RCV). *Clin Chem Lab Med* 2008;46:157–164.
48. Fraser CG. Reference change values. *Clin Chem Lab Med* 2011;50:807–812.
49. Lund F, Petersen PH, Fraser CG et al. Calculation of limits for significant bidirectional changes in two or more serial results of a biomarker based on a computer simulation model. *Ann Biol Clin* 2015;52:434–440.
50. Lund F, Petersen PH, Fraser CG et al. Calculation of limits for significant unidirectional changes in two or more serial results of a biomarker based on a computer simulation model. *Ann Biol Clin* 2014;52:237–244.
51. Fraser CG. Test result variation and the quality of evidence-based clinical guidelines. *Clin Chim Acta* 2004;346:19–24.
52. Fraser CG ed. Biological Variation: From Principles to Practice. Washington; AACPress: 2001
53. Zweig MH, Robertson EA. Why we need better test evaluations. *Clin Chem* 1982;28:1272–76.
54. Horvath AR, Lord SJ, Stjøn A, et al. From biomarkers to medical tests: the changing landscape of test evaluation. *Clin Chim Acta* 2014;427:49–57.
55. Fraser CG, Petersen PH. Desirable Standards for Laboratory Tests if They Are to Fulfill Medical Needs. *Clin Chem* 1993; 39:1447–55.
56. Fraser CG. The 1999 Stockholm Consensus Conference on quality specifications in laboratory medicine. *Clin Chem Lab Med* 2015;53:837–40.
57. Elevitch FR ed. Proceedings of the 1976 Aspen Conference on Analytical Goals in Clinical Chemistry. Illinois: College of American Pathologists;1977.
58. Fraser CG, Petersen PH. Analytical performance characteristics should be judged against objective quality specifications. *Clin Chem* 1999;45:321–323.
59. Kenny D, Fraser CG, Petersen HP et al. Stockholm Consensus Statement1999. *Scand J Clin Lab* 1999;59:585.
60. 1th EFLM Strategic Conference; <https://www.eflm.eu/site/page/last/1054> (Accessed April 2021)
61. Sandberg S, Fraser CG, Horvath AR et al. Defining analytical performance specifications: Consensus Statement from the 1st Strategic Conference of the European Federation of Clinical Chemistry and Laboratory Medicine. *Clin Chem Lab Med* 2015;53:833–35.
62. Jones GRD. Analytical performance specifications for EQA schemes - need for harmonisation. *Clin Chem Lab Med*. 2015;53:919–24.
63. Ceriotti F, Fernandez-Calle P, Klee GG, et al. Criteria for assigning laboratory measurands to models for analytical performance specifications defined in the 1st EFLM Strategic Conference. *Clin Chem Lab Med* 2017;55:189–94.
64. Smith AF, Shinkins B, Hall PS et al. Toward a Framework for Outcome-Based Analytical Performance Specifications: A Methodology Review of Indirect Methods for Evaluating the Impact of Measurement Uncertainty on Clinical Outcomes. *Clin Chem* 2019;65:1263–74
65. Skendzel LP, Barnett RN, Platt R. Medically useful criteria for analytic performance of laboratory tests. *Am J Clin Pathol* 1985;83:200–5.
66. Elion-Gerritzen WE. Analytic precision in clinical chemistry and medical decisions. *Am J Clin Pathol* 1980; 73:183–195.
67. Thue G, Sandberg S. Analytical performance specifications based on how clinicians use laboratory tests. Experiences from a post-analytical external quality assessment programme. *Clin Chem Lab Med* 2015;53:857–62.
68. Jones GRD, Albareda S, Kesseler D, et al. Analytical performance specifications for external quality assessment - definitions and descriptions *Clin Chem Lab Med* 2017;55:949–55. doi:[10.1515/cclm-2017-0151](https://doi.org/10.1515/cclm-2017-0151)
69. Gowans EM, Petersen PH, Blaabjerg O et al. Analytical goals for the acceptance of common reference intervals for laboratories throughout a geographical area. *Scand J Clin Lab Invest* 1988;48:757–64.
70. Haeckel R, Wosniok W. A new concept to derive permissible limits for analytical imprecision and bias considering diagnostic requirements and technical state-of-the-art. *Clin Chem Lab Med* 2011;49:623–35.
71. Oosterhuis WP, Sandberg S. Proposal for the modification of the conventional model for establishing performance specifications. *Clin Chem Lab Med* 2015;53:925–37.
72. Westgard JO, Hunt MR. Use and interpretation of common statistical tests in method-comparison studies. *Clin Chem* 1973;19:49–57.
73. Petersen HP, Fraser CG, Jørgensen LGM, et al. Combination of analytical quality specifications based on biological within- and between-subject variation. *Ann Biol Clin* 2002; 39:543–50.
74. Fraser CG, Petersen PH. Quality goals in external quality assessment are best based on biology. *Scand J Clin Lab Invest* 1993;53:8–9.

75. Oosterhuis WP. Gross overestimation of total allowable error based on biological variation. *Clin Chem*. 2011;57:1334–36.
76. Adams O, Cooper G, Fraser C, et al. Collective opinion paper on findings of the 2011 convocation of experts on laboratory quality. *Clin Chem Lab Med* 2012;50:1547–58.
77. Oosterhuis WP, Bayat H, Armbruster D, et al. The use of error and uncertainty methods in the medical laboratory. *Clin Chem Lab Med* 2018;56:45–11
78. Braga F, Infusino I, Panteghini M. Performance criteria for combined uncertainty budget in the implementation of metrological traceability. *Clin Chem Lab Med* 2015;53:905–12.
79. ISO. Medical laboratories — Practical guidance for the estimation of measurement uncertainty. ISO. July 2019:1–80.
80. Fraser CG, Petersen PH, Libeer JC et al. Proposals for setting generally applicable quality goals solely based on biology. *Ann Clin Biochem* 1997;34:8–12.

## MULTIPLE CHOICE QUESTIONS

1. Which of the following usually best describes the variation in the concentration or activity of most measurands in laboratory medicine?
  - a. Circadian rhythms
  - b. Monthly cycles
  - c. Systematic trends
  - d. Random variation
  - e. Seasonal fluctuations
2. Which of the following describes the most appropriate way to find numerical data on the components of biological variation?
  - a. Determine these in your own laboratory
  - b. Use a search engine on the Internet
  - c. Use databases on specific websites
  - d. Ask a colleague
  - e. E-mail a query to an internet forum
3. Analytical performance specifications (Select all that apply):
  - a. Can be calculated using measurement uncertainty.
  - b. Can be estimated based on biological variation data.
  - c. Should never be estimated using state of the art of the measurement method.
  - d. Is not important for setting quality control rules.
  - e. A measurand can have different analytical performance specifications based on its intended use.
4. Which of the following is irrelevant to the creation of reference change values?
  - a. Within-subject biological variation
  - b. Examination bias
  - c. Examination imprecision
  - d. Pre-examination sources of variation
  - e. Between-subject biological variation
5. Which of the following represents the individuality of most measurands in laboratory medicine?
  - a. The index of individuality is high.
  - b. The index of individuality is low.
  - c. The measurand has low individuality.
  - d. The within-subject variation is larger than the between-subject variation.
  - e. The analytical imprecision is lower than the within-subject variation.
6. What are the consequences for the use of the reference change value (RCV) if the applied within-subject variation estimates are derived from a heterogeneously distributed data set?
  - a. No consequences.
  - b. The within-subject biological variation will be too high.
  - c. The RCV can only be used on specific individuals.
  - d. The RCV is not applicable for use in the general population.
  - e. The RCV indicates only analytical variation.
7. If the difference between two consecutive examination results in a patient is larger than the reference change value (RCV), what does this mean?
  - a. That there is a medical change in the condition of the patient
  - b. There is an error in the analytical system
  - c. That the difference is larger than what can be explained by analytical and biological variation
  - d. That the patient has a 95% probability of being seriously ill
  - e. That the patient belongs to another population
8. What are the consequences of not excluding outliers within samples of each individual when using this data to estimate biological variation?
  - a. The reference change value cannot be generalized.
  - b. You have to log transform the data.
  - c. The within-subject biological variation will be underestimated.
  - d. The within-subject biological variation will be overestimated.
  - e. You cannot use the data for meta-analysis.
9. The rationale to set analytical performance specifications based on biology is:
  - a. To minimize the analytical noise to the biological signal.
  - b. Because this is the best way to measure patient outcomes.
  - c. Because most differences between two serial results from a patient can be explained by biological variation.
  - d. Because otherwise it is not possible to estimate total error.
  - e. Because it is the easiest way.
10. Select all that apply of the following statements.
  - a. The Biological Variation Data Critical Appraisal Checklist (BIVAC) may be used as a tool to appraise the quality of biological variations studies.
  - b. The BIVAC primarily focuses on the effect of study design and statistical handling on estimates for between-subject variation.
  - c. An overall BIVAC grade D indicates that the publication is BIVAC compliant and that data are fit for purpose.
  - d. The BIVAC consists of 14 quality items.
  - e. The BIVAC can be used to calculate the within-subject variation.

# Establishment and Use of Reference Intervals

*Gary Horowitz and Graham Ross Dallas Jones*

## ABSTRACT

### Background

One of the most important elements of a laboratory test result is the reference interval, a set of values against which physicians compare their patients' test results, facilitating interpretation. It is extremely important, therefore that the laboratory community devotes sufficient resources to ensure the reference limits they provide are well-founded. Most frequently, these reference limits represent values for healthy, adult patients, but other sets of values can be provided (such as values for pregnancy or for children). Sometimes, clinical decision limits are provided in place of conventional reference limits (such as for treating patients with diabetes or for diagnosing acute coronary syndromes).

### Content

In this chapter, we describe the techniques for properly establishing reference intervals, including selection of appropriate

reference individuals, implementation of preanalytical standardization, considerations for eliminating outliers, partitioning the reference group, and performance of statistical methods to calculate reference limits and their confidence intervals. In addition, since formal establishment of reference intervals may be beyond the capacity of many laboratories, we discuss alternative sources for reference limits (including manufacturers' package inserts, peer-reviewed literature, multicenter trials, historical laboratory data), along with techniques to verify the transferability of these data and common reference intervals. Consideration will be given to issues related to enhancing the display of patient test results with the appropriate reference limits. Even though most of the chapter is devoted to single tests (univariate) and population-based reference limits, we will also briefly describe the concept of subject-based and multivariate reference intervals. Lastly, we discuss techniques for ongoing verification of reference limits.

## CONCEPT OF REFERENCE LIMITS

### Interpretation by Comparison

Laboratory test results play a vital role in clinical medicine. Physicians use these results when screening for diseases in apparently healthy people, for confirming, excluding, or changing the probability of the diagnosis of specific diseases in patients with certain symptoms and signs, and for monitoring changes in a patient over time. To achieve these goals, interpretation is made by comparison with population reference limits, clinical decision limits, or previous results from the same patient. Population reference limits are generally derived from subjects without diseases, whereas clinical decision limits are generally based on clinical categories or outcomes of patients which can be separated on the basis of laboratory results. To facilitate these comparisons it is critical that laboratories provide not only the patient's test result but appropriate *reference limits* with which the patient's results can be compared. Ideally, for comparison with population limits, such reference limits should be available not only from healthy individuals but also from patients with relevant diseases, to assess whether a result is within the expected range for a clinical condition under consideration. Usually only health-associated reference limits are available in pathology

reports with expected values in diseases often estimated by doctors based on training and experience. Reference limits have been described as the most common decision support tool in laboratory medicine, and their inclusion on a pathology report is endorsed by the international clinical laboratory standard ISO 15189<sup>1</sup> and required by the College of American Pathologists (CAP).<sup>2</sup> A detailed history and commentary on the development of reference intervals is available.<sup>3</sup>

### Normal Values/Normal Ranges: Obsolete Terms

Historically, the term *normal values* was used to describe the laboratory data provided for purposes of comparison, and *normal ranges* as the expression of these on pathology reports. However, use of these terms often leads to confusion because the word "normal" has several different connotations.<sup>4</sup> For example, three medically important but very different meanings of "normal" are:

1. *Statistical sense*: Values can be described as "normal" if their observed distribution seems to follow closely the theoretical *normal distribution* of statistics—the Gaussian probability distribution. This use of "normal" has sometimes misled people to believe that the distribution of biological data is always symmetric and bell shaped, like the Gaussian distribution. However, on closer examination,

this usually is not correct. To exorcize the “ghost of Gauss,” Elveback and colleagues recommend not using the term *normal limits*.<sup>5</sup> For a similar reason, the term *normal distribution* should be avoided and replaced by the term *Gaussian distribution*.

2. *Epidemiologic sense:* Another meaning of “normal” is illustrated by the following statement: It is “normal” to find that the activity of gamma-glutamyltransferase (GGT) in serum is between 7 and 47 IU/L, whereas it is considered “abnormal” to have a serum GGT value outside these limits. Here a more exact statement would read as follows: Approximately 95% of the values obtained, when the activity of GGT in sera collected from individuals considered to be healthy is measured, are included in the interval 7 to 47 IU/L. The obsolete concept of *normal values* in part carried this meaning. Alternative terms for “normal” in this sense are *common, frequent, habitual, usual, and typical*.
3. *Clinical sense:* The term “normal” also is often used to indicate that values show the absence of certain diseases or the absence of risks for the development of diseases. In this sense, a *normal value* is considered a sign of health. Better descriptive terms for such values are *healthy, non-pathologic, and harmless*. As a corollary, when results are discussed with patients, it may be unhelpful to describe results outside reference limits as “abnormal” because this may be taken to indicate the presence of disease or ill health and therefore create unnecessary anxiety or concern. Because of confusion resulting from the different meanings of normal, the terms *normal values* and *normal ranges* are obsolete and should not be used.

To prevent the ambiguities inherent in the term *normal values*, the concept of *reference values*, from which the terms *reference intervals* and *reference limits* are derived, was introduced and implemented in the 1980s.<sup>6,7</sup> The term *reference* is appropriate because these values provide something to refer to when interpreting a result. This was an important event in establishing a scientific basis for clinical interpretation of laboratory data.<sup>8</sup> The term *reference range* is sometimes used in place of the term *reference interval* recommended by the International Federation of Clinical Chemistry and Laboratory Medicine (IFCC). This use is incorrect because the statistical term *range* denotes the difference (a single value!) between maximum and minimum values in a distribution.<sup>9</sup>

## Terminology

The IFCC recommends use of the term *reference individuals* and related terms such as *reference value, reference limit, reference interval, and observed values*.<sup>7,10–15</sup> The definitions and the presentation in the following sections of this chapter are in accordance with IFCC recommendations,<sup>10–15</sup> which have been adopted by the Clinical Laboratory Standards Institute (CLSI).<sup>16</sup>

*Reference individual:* An individual selected for comparison using defined criteria.

As mentioned previously, for the interpretation of values obtained from an individual under clinical investigation, appropriate comparison values are needed. To provide such values, suitable individuals must be selected. The characteristics of the individuals in each group chosen for comparison should be clearly defined. Their age and sex must be specified and whether they should be healthy or have a certain disease. The definition of a reference individual also covers cases in

which the individual under clinical investigation is his or her own reference, as discussed in a later section on subject-based reference values.

*Reference population:* The entire set of reference individuals.

*Reference value:* A value obtained by observation or measurement of a particular type of quantity on a reference individual.

If, for example, the activity of GGT is measured in sera collected from a group of reference individuals selected for comparison according to a sufficiently exact set of criteria, the GGT results are considered reference values.

*Reference distribution:* The distribution of the reference values.

*Reference limits:* The upper and lower bounds of the specified fraction of the reference distribution, typically the central 95% of the distribution.

*Reference interval:* The spread of values defined by the upper and lower reference limits.

*Observed value:* A value of a particular type of quantity obtained by observation or measurement and produced to make a medical decision. Observed values can be compared with reference values, reference distributions, reference limits, or reference intervals.

Or, rephrased: An observed value is the result obtained by analysis of a specimen collected from *an individual under clinical investigation*. The equivalent term used in the International Vocabulary of Metrology (VIM) is *measurement result*.<sup>17</sup>

The IFCC also defines other terms related to the concept of *reference values*: reference sample group, reference distribution, reference limit, and reference interval.<sup>10–15</sup> Some of these terms are introduced in later sections of this chapter.

## Clinical Decision Limits

The terms *reference limits* and *clinical decision limits* should not be confused.<sup>8,18</sup> *Reference limits* are descriptive of the distribution of results in the selected subset of reference individuals; they tell us something about the expected variation of values in the reference population. Comparison of new values with these limits conveys information about similarity to the given reference values. In contrast, *clinical decision limits* provide separation based on clinical categories or outcomes. The latter limits may be based on analysis of reference values from several groups of individuals (healthy persons and patients with relevant diseases) and are used for the purpose of differential diagnosis.<sup>18–20</sup> Alternatively, such values are established on the basis of outcome studies and are used as clinical guidelines for treatment. Examples of current decision limits include recommended concentrations for therapeutic drug levels (see Chapter 42), the National Cholesterol Education Program guidelines related to cholesterol,<sup>21</sup> the American Diabetes Association recommendations for diagnosis of diabetes with HbA<sub>1c</sub> or plasma glucose,<sup>22</sup> and the American Academy of Pediatrics guidelines on neonatal bilirubin.<sup>23</sup> A key factor with clinical decision limits is that each assumes that measurements of the involved analytes are accurate, with the metrological traceability similar to the method used in the clinical studies on which the clinical decision points were established (see Chapter 7).

In this context, it is critical to point out another difference between reference limits and clinical decision limits. For most analytes, a laboratory should establish (or verify) its

own reference limits. The processes to do this are described later in this chapter. But for analytes interpreted using clinical decision limits such as national or international laboratory guidelines, efforts that once would have been dedicated to establishing or verifying reference intervals should be redirected toward establishing accuracy (trueness). In the 2010 CLSI guidelines,<sup>16</sup> this point is given much-deserved emphasis. It does little good to establish one's own reference limits if physicians will (and should) use national guidelines or if the laboratory gives results which are biased compared with the results used to determine the clinical decision points. Methods to establish the accuracy of one's method are discussed in Chapter 7. It is also important for laboratories to communicate to clinicians the nature of reference limits provided with results, specifying whether these are population reference intervals or clinical decision limits, as well as any additional information required for appropriate use. In particular, information on populations with and without specified diseases allows for determination of important characteristics of diagnostic tests, including their sensitivity, specificity, predictive values, and likelihood ratios, all of which are discussed in detail in Chapter 2.

### Types of Reference Limits

In practice it is often necessary or convenient to give a short description associated with the term *reference limits*, such as *health-associated reference limits* (close to what was understood by the obsolete term *normal values*). With conditions such as obesity, which are prevalent in many populations and associated with poorer health outcomes, the definition of health-associated reference limits becomes more difficult, both to define (this is discussed in subsequent text with exclusions from the reference population) and to communicate to the end-user. Other examples of such qualifying words could be *hospital inpatient*, *pregnancy*, and *patients with well-controlled diabetes*. These short descriptions prevent the common misunderstanding that reference values are associated only with health.

### Subject-Based and Population-based Reference Values

*Subject-based* reference values are previous values from the same individual, obtained when he or she was in a known state of health. *Population-based* reference limits are those obtained from a group of well-defined reference individuals and are usually the types of values referred to when the term *reference limits* is used with no qualifying words. This chapter deals primarily with population-based values. It should be noted, however, that for some tests, intraindividual variation may be small relative to interindividual differences. The relationship of within- to between-individual variation is known as the index of individuality (see Chapter 8), and in cases in which this is low (e.g., creatinine,<sup>24</sup> immunoglobulins<sup>25</sup>), the use of population-based reference intervals may distract from clinically significant intraindividual changes, as noted later in this chapter. In this setting the concept of "reference change value" (RCV) can be seen as analogous to a population reference limit, as the RCV is defined using data from reference subjects from a reference population, and a statistical analysis used to determine significant changes.

It is also important to note that this chapter focuses on population-based *univariate reference limits* and quantities derived from them. For example, if separate reference limits

for calcium and parathyroid hormone (PTH) in plasma are used, two sets of univariate reference limits are produced. The term *multivariate reference limits* denotes that results of two or more analytes obtained from the same set of reference individuals are treated in combination. Plasma calcium and PTH values may be used, for example, to define a bivariate reference region, which would reflect the fact that, as calcium concentrations decrease, even within healthy reference limits, PTH levels rise. Thus a PTH level that is within health-associated univariate reference limits might not be within the health-associated bivariate reference limits. This subject is addressed briefly in a later section.

### Requirements for Valid Use of a Reference Interval

Certain conditions apply for a valid comparison between a patient's laboratory results and reference values<sup>26</sup>:

1. The reference individuals for each test should be clearly defined.
2. The patient examined should sufficiently resemble the reference individuals in all respects other than those under investigation.
3. The conditions under which the reference specimens were obtained and processed for analysis should be known and these conditions should be the same as for the patient specimen.
4. The measurand under examination in the patient and the reference individuals should be the same.
5. All laboratory results should be produced using adequately standardized methods under sufficient analytical quality control (see Chapters 6 and 7). The standardization should be sufficient that any bias or difference in precision or analytical specificity between the analytical system used for the patient sample and that used for the reference samples does not affect the interpretation.

To these general requirements one may add others that become necessary when more detailed and sophisticated approaches to decision making are applied.<sup>8</sup>

6. Stages in the pathogenesis of diseases that are the objectives for diagnosis should be demarcated beyond the separation between presence and absence of the disease. For example, although some overlap occurs, the clinical grades of congestive heart failure (CHF) are distinguished by progressive increases in levels of N-terminal pro-brain natriuretic peptide (NTproBNP).<sup>27</sup>
7. Clinical diagnostic sensitivity and specificity, prevalence, and clinical costs of misclassification should be known for all laboratory tests used. For example, in some instances, one might want to know whether a given NTproBNP value is "healthy," in which case one would want to use reference limits for age- and sex-matched individuals with no evidence of CHF. In contrast, when faced with a patient complaining of shortness of breath in the emergency room, one might want instead to know, not so much whether any degree of CHF is present, but whether the patient's CHF is sufficiently advanced to be the cause of the shortness of breath.<sup>28,29</sup>

### SELECTION OF REFERENCE INDIVIDUALS

A set of *selection criteria* determines which individuals should be included in the group of reference individuals.<sup>7,10–15</sup> Such selection criteria include statements describing the source

population and specifications of criteria for health or for the disease of interest.

Often, separate reference values for each sex and for different age groups,<sup>30</sup> as well as other criteria, are necessary. The overall group of reference individuals therefore may have to be divided into more homogeneous subgroups. For this purpose, specific rules for the division, called *stratification* or *partitioning criteria*, are needed.

It is important to distinguish between selection and partitioning criteria. First, selection criteria are applied to obtain a group of reference individuals. Thereafter, this group is divided into subgroups using partitioning criteria. Whether a specific criterion (e.g., sex) is a selection or a partitioning criterion depends on the purpose of the actual project. For example, sex is a selection criterion if reference values only from female subjects are necessary. Sex can also be a selection criteria where the data will be partitioned using this criterion to ensure sufficient numbers of each sex are collected.

### Concept of Health in Relation to Reference Values

There is an obvious requirement for health-associated reference values for quantities measured in the clinical laboratory. But the concept of health is problematic; as Grasbeck stated “Health is characterized by a minimum of subjective feelings and objective signs of disease, assessed in relation to the social situation of the subject and the purpose of the medical activity, and it is in the absolute sense an unattainable ideal state.”<sup>31</sup> Much confusion may arise if the selection criteria for health are not clearly stated for a specific project.

When reference values are produced, the following questions are asked: (1) Why are these values needed? (2) How are they going to be used? (3) To what extent does the intended purpose of the project determine how health is identified? For example when setting reference limits for cardiac-specific troponins, a “cardio-healthy” population is required that is in other ways similar to the patients who are likely to present with possible acute coronary syndrome (i.e., they should be of similar age and gender, and they may have hypertension or hyperlipidemia).<sup>32,33</sup>

### Strategies for Selection of Reference Individuals

Several methods have been suggested for the selection of reference individuals. Table 9.1 shows a variety of concepts that may be used to describe a sampling scheme. The concepts can be considered as pairs, each of which is mutually exclusive. For example, the sampling may be direct or indirect, and direct sampling may be a priori or a posteriori.

The merits and disadvantages of these strategies are described in the following sections. It is not possible to recommend one sampling scheme that is superior in all respects and applicable to all situations. One must choose the optimal approach for a given project and state clearly what has been done.

### Direct or Indirect Sampling?

Selection of reference intervals by *direct* sampling involves collection of specimens from selected members of the reference population for the purpose of establishing reference limits. *Indirect* sampling involves deriving reference limits from using results of samples collected for other purposes. Direct selection of reference individuals (see Table 9.1) concurs with the concept of reference values as recommended by

**TABLE 9.1 Strategies for Selection of Reference Individuals**

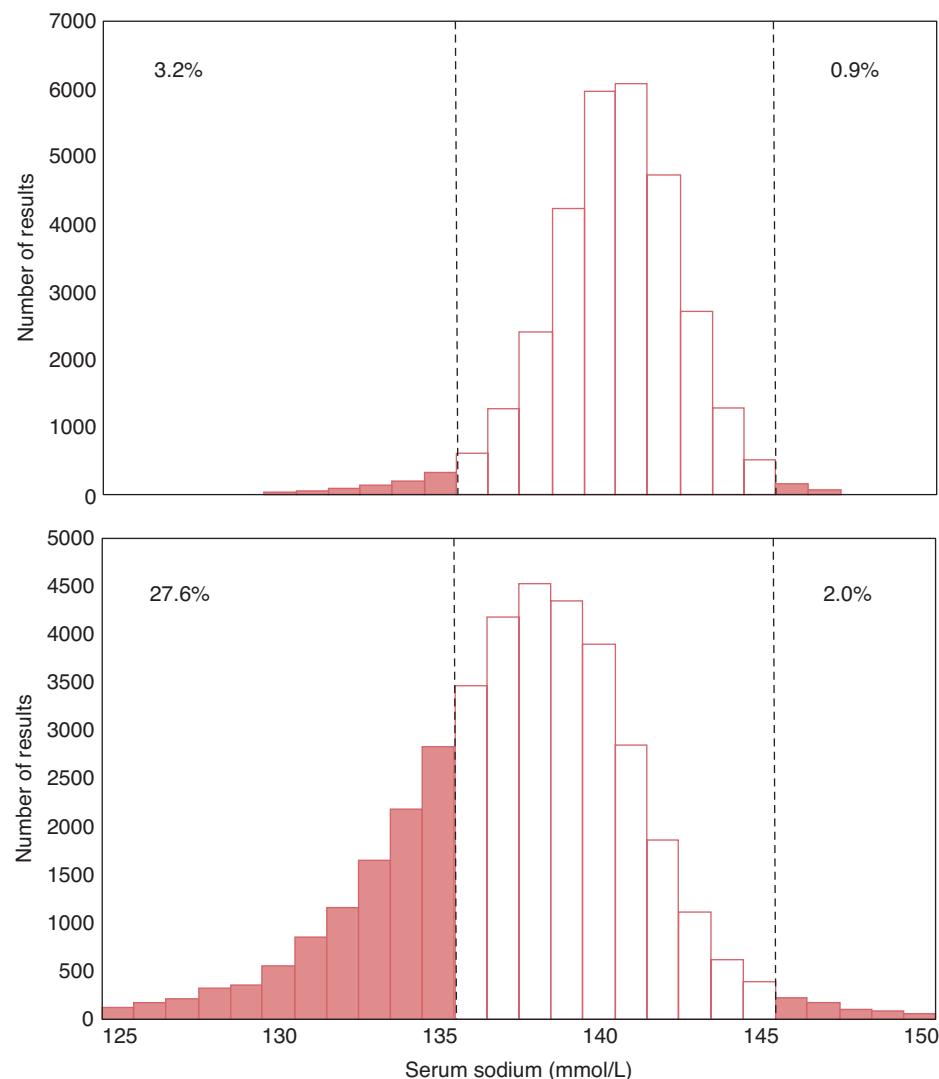
<b>Direct Versus Indirect</b>	
Direct	Individuals are selected from a parent population using defined criteria
Indirect	Individuals are not considered, but certain statistical methods are applied to analytical values in a laboratory database
<b>A Priori Versus A Posteriori</b>	
A Priori	Individuals are selected for specimen collection and analysis if they fulfill defined inclusion criteria
A Posteriori	Use of an already existing database containing both relevant clinical information and analytical results. Values of individuals meeting defined inclusion criteria are selected.
<b>Random Versus Nonrandom</b>	
Random	Process of selection giving each item (individual or test result) in the parent population an equal chance of being chosen
Nonrandom	Process of selection that does not ensure that each item in the parent population has an equal chance of being chosen

the IFCC,<sup>10–15</sup> and it is the basis for the presentation in this chapter. Its major disadvantages are the problems and costs of obtaining a representative group of reference individuals.

These practical problems have led to the search for simpler and less expensive approaches such as *indirect* methods.<sup>6,34</sup> Historically the indirect approach has been taken using results in a routine pathology database, often from laboratories serving a largely inpatient population. While these may be the only data available for some laboratories, the indirect approach may be applied to other data sources, such as samples collected for research, epidemiology, or “wellness testing,” where the expected prevalence of disease may be low. A key starting point with any indirect method is an understanding of the population from which the samples have been drawn, even if specific criteria have not been applied at the time of collection.

The indirect approach is based on the observation that many analysis results produced in the clinical laboratory seem to be “normal,” or at least unaffected by the reason for the sample collection. Two main concepts have been used to extract information about reference distributions from this type of data. The first is the use of statistical methods which allow identification of a distribution within the database which is then taken to represent the reference population. Note that for this approach no attempt is made to classify individual results as representing the reference population. The alternate method is to use additional clinical information to classify individual results and exclude those which are more likely to be from individuals with relevant disease or other factors which may affect the results. Typically both methods are applied in the development of reference intervals using the indirect approach.

An example of the results from a pathology database is shown in Fig. 9.1. As seen, the values of serum sodium



**FIGURE 9.1** Distribution of serum sodium concentrations obtained in a routine laboratory over 1 year. The top histogram (A) shows 31,183 sequential results from general practice sites and the lower histogram (B) shows 38,751 from hospital wards. The dark shaded areas and attached percentages show the fractions of the two populations outside the reference interval derived by direct methods in the same population (represented by the dashed vertical lines, 136 to 145 mmol/L) (From Koerbin G, Cavanaugh JA, Potter JM, Abhayaratna WP, West NP, Glasgow N, et al. "Aussie normals": an a priori study to develop clinical chemistry reference intervals in a healthy Australian population. *Pathology*. 2015;47:138–44).

concentrations from outpatients have a distribution with a preponderant central peak and a shape similar to a Gaussian distribution. The underlying assumption of the indirect method is that this peak is composed mainly of *normal values* or, more precisely, is derived from patients without the condition of interest or diseases that may affect the analyte under consideration. Advocates of the method therefore claim that it is possible to estimate a *reference interval* if the distribution of unaffected values from this distribution is extracted. Fig. 9.1 also shows serum sodium results from hospital inpatients showing both lower results on average, as well as an increased proportion of lower results (the data set is left skewed). This may be due to the presence of a significant proportion of the samples being derived from patients with a condition affecting the results, for example, in the case of

serum sodium, diuretic use, dehydration, and other fluid imbalances. It may also be due to systematic preanalytical differences, such as recumbence in inpatients compared with ambulatory outpatients. This shows the importance of selection of the data set to use for indirect analysis. Several mathematical methods have been used to extract a distribution for the derivation of reference limits from routine laboratory data.<sup>35–38</sup>

In short, the indirect method has at least two potential major deficiencies:

1. Estimates of the reference limits can depend heavily on the particular mathematical method used and on its underlying assumptions.
2. Estimates of the reference limits can be affected by the prevalence, nature, and severity of disease included in the laboratory database. This may be a particular problem

with databases containing only hospital inpatients. The use of ambulatory outpatients and general practice patients can reduce this variability considerably.

However, if appropriate exclusion criteria are applied, data derived by indirect sampling from pathology databases may be used for the establishment of reference values in a way that is fully concordant with IFCC recommendations.<sup>35–37,39,40</sup>

The requirement for this approach is that laboratory results should be *combined with other information* (i.e., to combine an *a posteriori* strategy with the indirect method). Laboratory results are to be used as reference values only if stated clinical criteria are fulfilled. The types of data which can be used include demographic information such as age, sex, source (e.g., inpatient, outpatient, specific clinics); patient sampling related information (e.g., by excluding multiple results from the same patient or limiting samples to those where only a single request for that test has been made in a specified time); information from other pathology results (e.g., using HbA<sub>1c</sub> or fasting glucose results to reduce the likelihood of overweight or obesity related effects); or from other clinical information available by linking with clinical databases. In practice the factors applied are analyte-specific and depend on what is available, and detailed understanding of the pathophysiology of the analyte being examined is required. For example results from inpatients should be excluded for analytes affected by recumbency (e.g., serum albumin or sodium) or intercurrent disease (e.g., C-reactive protein [CRP], other acute phase reactants). For tests used for both diagnosis and monitoring (e.g., serum creatinine, tumor markers), restricting analysis to patients with a single result may be preferred. This can also be seen as an acceptance by the treating doctor that further investigation was not warranted based on the results.

Reference values produced by indirect sampling techniques have a number of significant potential advantages over those based on direct sampling. With any indirect method, the preanalytical and analytical factors are exactly the same for the patient sample and the reference setting process, and also the reference population matches that of the patient. This can provide a more appropriate comparison group, as the role of clinical decision-making is to separate patients with the same clinical presentation on the basis of disease, rather than separating sick from healthy. For example, the need, in patients with chest pain, is to distinguish those having a myocardial infarction from those who are not.

The indirect approach can also be used in settings where collection of samples for reference interval studies may be particularly problematic such as extremes of age or during pregnancy. Additionally the numbers of samples which may be available for indirect techniques can be vastly greater than direct techniques, in the many tens or even hundreds of thousands and the costs are a fraction of those of direct studies. If direct studies are available for comparison, indirect studies will enable an assessment of whether there are differences in the local population, specimen collection techniques, or analytical methods. It is however important to note that the indirect approach is continuing to evolve and that if poorly performed, an indirect study can give misleading results.

### A Priori or A Posteriori Sampling?

When carefully performed, both *a priori* (before) and *a posteriori* (after) sampling (see Table 9.1) may result in reliable

reference values. The use of the *a priori* approach is limited to direct reference interval studies, but as discussed above, the *a posteriori* approach can be applied to direct and indirect studies. The choice is often a question of practicality. Both require the same set of successive steps, but the order of some of these operations differs depending on the mode of selection: *a priori* or *a posteriori*.<sup>6</sup>

The first step in the process of producing reference values for a laboratory test should always be the collection of quantitative information about sources of biological, preanalytical, and analytical variation for the analyte studied. In this setting, biological variation includes expected variation with time of day, with meals, with seasons, and with life stages. A search through relevant literature may yield the required information.<sup>41,42</sup> If relevant information cannot be found in the literature, pilot studies may be necessary before the selection of reference individuals is planned in detail. Serum sodium is an example of a biological analyte that is affected by only a few sources of biological variation. However, the list of factors may be rather long for other analytes, such as serum enzymes, proteins, and hormones.

It is important to distinguish between controllable and noncontrollable sources of biological variation. Some factors may be controlled by standardization of the procedure for preparation of reference individuals and specimen collection such as fasting status and time of day (see a later section of this chapter). Other factors, such as age and gender, may be relevant partitioning criteria. Remaining sources of variation should be considered when criteria for the selection of reference individuals are defined.

The *a priori* strategy is best suited for smaller studies and for analytes for which there are very specific confounding factors or for which the analytical process is very difficult or expensive. One such example is male sex hormone-related reference intervals.<sup>43</sup> Potential reference individuals from the parent population should be interviewed and examined clinically and by selected laboratory methods to decide whether they fulfill the defined inclusion criteria. If they do, specimens for analysis are collected by a standardized procedure (including the necessary preparation of individuals before the collection).

The *a posteriori* method is based on the availability of a large collection of data on medically examined individuals and measured quantities. Studies thoroughly planned by centers for health screening or preventive medicine may provide such data. It is important that data be collected by a strictly standardized and comprehensive protocol concerning (1) sampling from the parent population, (2) registration of demographic and clinical data on participating individuals, (3) preparation for and execution of specimen collection, and (4) handling and analysis of the specimens. If these requirements are met, values may be selected after application of the defined inclusion criteria to individuals found in the database. The selection of individuals from large pathology databases (see earlier discussion) is another example of the application of an *a posteriori* method. In this case, however, the quality of data may be lower than that in well-planned population studies.

A study performed in Kristianstad, Sweden,<sup>44</sup> highlights a practical problem often met when reference individuals are selected: the number of subjects fulfilling the inclusion criteria may be too small. In this study, only 17% of participants

were accepted into the study, according to the criteria used, leaving an insufficiently sized reference sample group and a risk of selection bias. The frequency of exclusion was higher among women and in older age groups, exacerbating the issues in these groups.

This problem has two possible solutions:

1. The exclusion criteria may be relaxed. As already discussed, the set of relevant sources of biological variation differs among different analytes. One may define a minimum set of exclusion criteria for a given laboratory test. In the Kristianstad study, the complete group of individuals could probably be used for establishment of reference values for serum sodium, and most of the individuals would be acceptable for the determination of reference values for several other analytes.<sup>44</sup>
2. Another design of the sampling procedure could reduce the practical problems and costs of obtaining a sufficiently large group of reference individuals. The Kristianstad study showed that 75% of excluded subjects could have been identified using only a simple questionnaire.<sup>44</sup> In the upper age group, this percentage was even higher. Therefore preliminary screening of a large number of individuals from the parent population, using a carefully designed questionnaire (i.e., of or related to the current or previous medical history of a patient), would result in a much smaller sample of individuals for examination clinically and by laboratory methods. If 3000 individuals had been prescreened in Kristianstad, and if only the individuals remaining in the reduced sample were subjected to a closer examination, a group of 240 reference individuals would have been obtained.

The two modifications of the protocol may also be combined.

### Random or Nonrandom Sampling?

Ideally, the group of reference individuals should be a random sample of all individuals fulfilling the inclusion criteria defined in the parent population. Statistical estimation of distribution parameters (and their confidence intervals) and statistical hypothesis testing require this assumption.

For several reasons, most collections of reference values are, in fact, obtained by a nonrandom process.<sup>45</sup> This means that all possible reference individuals in the entire population under study do not have an equal chance of being chosen for inclusion in the usually much smaller sample of individuals studied. A strictly random sampling scheme in most cases is impossible for practical reasons. It would imply the examination of and application of inclusion criteria to the entire population (thousands or millions of persons), and then the random selection of a subset of individuals from among those accepted. This approach has been used in selecting individuals at random to provide a cohort that is representative of the full population by several national organizations, such as the National Health and Nutrition Examination Survey (NHANES)<sup>46</sup> in the United States, the Canadian Health Measures Survey in Canada,<sup>47</sup> and the Australian Bureau of Statistics.<sup>48</sup>

Usually the situation is less satisfactory. The sampling process is highly affected by convenience and cost. For example, samples of reference individuals are commonly obtained by selecting (1) from blood donors, (2) from persons working in a nearby factory, (3) from hospital staff, or (4) from hospital databases, none of which represent a

random sampling of possible reference individuals in the general population.

The conclusions are obvious: (1) the best reference sample obtainable should be used with a balance between practical considerations and consideration of possible biases that may be introduced by the selection process, and (2) the data should be used and interpreted with due caution, with awareness of the possible bias introduced by the nonrandomness of the sample selection process. For example, lower iron stores may be expected in a sample of regular blood donors, and higher vitamin D concentrations may be expected in a sample drawn from outdoor workers. An additional effect of nonrandomness is an increased chance that results of different reference studies may produce different results even when the defined reference population is intended to be the same.

### Selection Criteria and Evaluation of Subjects

The selection of reference individuals consists essentially of applying defined criteria to a group of examined candidate persons.<sup>10–15</sup> The required characteristics of the reference values determine which criteria should be used in the selection process. Table 9.2 lists some important criteria to consider when production of health-associated reference values is the aim.

In practice, consideration of which *diseases* and *risk factors* to exclude is difficult (see the discussion on the concept of health earlier in this chapter). The answer lies in part in the intended purpose of establishing reference values; the project must be goal oriented.

Once a factor has been selected as an exclusion factor, a relevant and practical definition is required. For example, obesity is a common condition that is associated with a

**TABLE 9.2 Examples of Exclusion and Partitioning Criteria<sup>a</sup>**

Exclusion	Partitioning
Age	Age
Alcohol intake	Blood group
Blood donation (recent)	Circadian variation
Drug abuse	Ethnicity
Exercise intensity (recent)	Exercise intensity (recent)
Fasting vs. nonfasting	Fasting vs. nonfasting
Sex	Sex
Hospitalization (recent)	Menstrual cycle (by stage)
Hypertension	
Illness (recent)	
Lactation	
Obesity	Obesity
Occupation	Posture (when sampled)
Oral contraceptives	
Pregnancy	Pregnancy (by stage)
Prescription drugs	Prescription drugs
Recent transfusion	

<sup>a</sup>As indicated by the shaded boxes, some criteria may be considered as either exclusion criteria or partitioning criteria.

number of diseases; however, the definition of *obesity* is problematic. A definition might be based on a known assumed contribution to the risk of a development of specified disease. However, scientific data of this type are seldom available for the studied population. Another possibility for establishing obesity is to use upper limits based on weight measurements in different age, gender, and height groups of the general population (e.g., more than 20% above the national age-, sex-, and height-specific mean weight). For obesity, a common approach is to use definitions based on the body mass index (BMI),<sup>49</sup> although limiting subjects to the healthy range will exclude over 50% of some populations. Tables of optimum or ideal weights have been published by life insurance companies; they may be more appropriate for delineation of obesity. Similar problems relate to the definition of hypertension. And what if a potential reference individual is no longer obese as a result of bariatric surgery or is currently normotensive on drug therapy?

In addition, is it permissible to use exclusion criteria based on *laboratory measurements*? It has been argued that a circular process might happen when laboratory tests are used to assess the health of subjects who are subsequently used as healthy control subjects for laboratory tests. But actually there is no difference, in this context, between measuring height, weight, and blood pressure and performing selected laboratory tests, provided that these laboratory tests are neither those for which reference values are produced nor tests that are significantly correlated with them.<sup>31</sup>

The removal of reference results based on other laboratory results has been used in a process termed latent abnormal values exclusion (LAVE). In a multinational study it was shown that this process, using a standard group of exclusion tests and criteria, affected some analytes but had little effect on others.<sup>50</sup> As stated above, care should be taken that tests with correlated results are not used for this purpose. It is particularly difficult to define selection criteria when establishing reference values for older patients.<sup>51</sup> In older age groups, it is “normal” (i.e., common) to have minor or major diseases and to take therapeutic drugs. One solution is to collect values at one time and to use the values of survivors after a defined number of years.<sup>31,52</sup>

Usually the clinical evaluation of candidate individuals is based on (1) a detailed interview or questionnaire (i.e., the complete history recalled and recounted by a patient), (2) a physical examination, and (3) supplementary investigations. Questionnaires and examination forms tailored to the requirements of the actual project facilitate the evaluation and document the decisions made.

### Partitioning of the Reference Group

It may also be necessary to define *partitioning criteria* for the subclassification of the set of selected reference individuals into more homogeneous groups (see Table 9.2).<sup>10–15</sup> (The question of determining when stratification of the reference sample group is necessary and justified is discussed in later sections.) In practice, the number of partitioning criteria should usually be kept as small as possible to ensure sufficient sample sizes to derive valid estimates.

Age and sex are the most frequently used criteria for subgrouping, because several analytes vary notably among different age and sex groups.<sup>41,51,53</sup> Age may be categorized by equal intervals (e.g., by decades) or by intervals that are

narrower in the periods of life when greater variation is observed. In some cases, more appropriate intervals can be obtained from qualitative age groups, such as (1) postnatal, (2) infancy, (3) childhood, (4) prepubertal, (5) pubertal, (6) adult, (7) premenopausal, (8) menopausal, and (9) geriatric. Further subdivision may also be needed based on Tanner stage of puberty or based on phase of the menstrual cycle. Height and weight also have been used as criteria for categorizing children. The use of age and sex for partitioning has the advantage that reference limits derived from subpopulations on these criteria can be easily applied on pathology reports where these factors are usually known about the patient. In contrast, the application of limits based on other criteria requires knowledge not usually available to the laboratory.

## SPECIMEN COLLECTION

Several preanalytical factors can influence the values of measured biological quantities, such as the concentrations of components in a blood sample or the amount excreted in feces, urine, or sweat.<sup>54,55</sup> This topic is covered elsewhere (see Chapters 4 and 5). In this discussion, only aspects of special relevance to the generation of reliable reference values are highlighted.<sup>10–15,54</sup>

Standardization of the (1) preparation of individuals before specimen collection, (2) procedure of specimen collection itself, and (3) handling of the specimen before analysis may eliminate or minimize bias or variation from these factors. This reduces the “noise” that might otherwise conceal important biological “signals” of disease, risk, or treatment effect.

### Preanalytical Standardization

Preanalytical procedures used before routine analysis of patient specimens and when reference values are established should be as similar as possible. In general, it is much easier to standardize routines for studies of reference values than those used in the daily clinical setting, especially when specimens are collected in emergency or other unplanned situations. Thus two general approaches have been suggested:

1. Only such factors that may be relatively easily controlled in the clinical setting should be part of the standardization when reference values are produced.
2. The rules for preanalytical standardization when reference values are produced should also be used for the clinical situation. Such rules include food and beverage restrictions, exercise restrictions, time sitting (or lying down) prior to phlebotomy, and tourniquet time. It has been shown that it is possible to apply these rules rather closely in the clinical setting for both hospitalized and ambulatory patients.<sup>7</sup> The same philosophy forms the basis for recommendations concerning sample preparation preceding analysis.

However, either philosophy is concordant with the concept of reference values, provided that the conditions under which reference values are produced are clearly stated.

### Analyte-Specific Considerations

The types and magnitudes of preanalytical sources of variation clearly are not equal for different analytes (see Chapter 5).<sup>42</sup> In fact, some believe that only those factors that cause unwanted

variation in the biological quantities for which reference values are being generated should be considered. For example, body posture during specimen collection is highly relevant for the establishment of reference values for analytes that do not diffuse across blood vessel walls, such as albumin in serum or red cell count in blood, but posture is irrelevant for establishment of serum sodium values.<sup>42,55</sup>

Alternatively, several constituents are analyzed routinely in the same clinical specimen and therefore it would be impractical to devise special procedures for every single type of quantity.<sup>7</sup> Consequently, three standardized procedures for blood specimen collection by venipuncture have been recommended<sup>6,54</sup>: (1) collection in the morning from hospitalized patients, (2) collection in the morning from ambulatory patients, and (3) collection in the afternoon from ambulatory patients. Such schemes have to be modified depending on local conditions and necessities and on the intended use of the reference values produced. Published checklists<sup>7,10–15</sup> may be helpful in the design of a scheme.

A special problem is caused by drugs taken by individuals before specimen collection,<sup>56,57</sup> and it may be necessary to distinguish between indispensable and dispensable medications. If possible, dispensable medication should be avoided for at least 48 hours. The use of indispensable drugs, such as contraceptive pills or essential medication, may be a criterion for exclusion or partitioning if these affect the analyte of interest.

In emergency or other unplanned clinical situations, even a partial application of the standardized procedure for collection has been shown to be of great value.<sup>6</sup> When collections have been made under conditions other than those specified for a specific analyte, interpretation of results against reference limits requires awareness of the type and magnitude of variation that may be expected under those circumstances. For example, a serum cortisol collected in the evening cannot usually be compared with reference limits established for morning collections, the exception being that a high result is still of great clinical relevance because the upper limit for evening values is typically much lower than the upper limit for morning values.

### Necessity for Additional Information

The clinical situation is often different from a controlled research situation; for example, specimens have to be taken (1) during operations, (2) in emergency situations, and (3) when patients are unwilling or unable to follow instructions. Therefore the clinician may need additional information for interpretation of a patient's values in relation to reference values obtained under fairly standardized conditions.

An *empirical approach*<sup>7</sup> is to produce other sets of reference values, such as postprandial values, postexercise values, or postpartum values.<sup>6</sup> Such a method, however, is very expensive and does not cover all situations that could possibly arise. This approach is also limited by the variability in these events (i.e., for postprandial samples, the size of the meal, the types of food consumed, and the number of hours since the meal).

Another, more general solution to the problem is called the *predictive approach*.<sup>7</sup> Starting from a set of ordinary reference values and using quantitative information on the effects of various factors (e.g., intake of food, alcohol, and drugs; exercise; stress; posture; or time of day), expected reference

values that fit the actual clinical setting could be estimated.<sup>41,42</sup> An interesting example is provided by thyroid-stimulating hormone (TSH), where the effect of diurnal variation needs to be considered.<sup>58</sup>

More studies of such effects are needed, especially for the combined effect of two or more sources of variation. For example, is the combined effect of alcohol and contraceptive drugs on GGT activity in serum less than, equal to, or greater than the sum of their individual effects?

### ANALYTICAL PROCEDURES AND QUALITY CONTROL

Essential components of the required definition of a set of reference values are specifications concerning (1) the analysis method (including information on metrological traceability, equipment, reagents, calibrators, type of raw data, and calculation method; see Chapter 7), (2) quality control (see Chapter 6), and (3) reliability criteria.<sup>6,10–15</sup>

Specifications should be so carefully described that another investigator will be able to reproduce the study and that a user of reference values will be able to evaluate their comparability with values obtained by methods used for producing the patients' values in a routine laboratory. To ensure comparability between reference values and observed values a method with the same performance characteristics of traceability, reproducibility, and analytical specificity should be used.

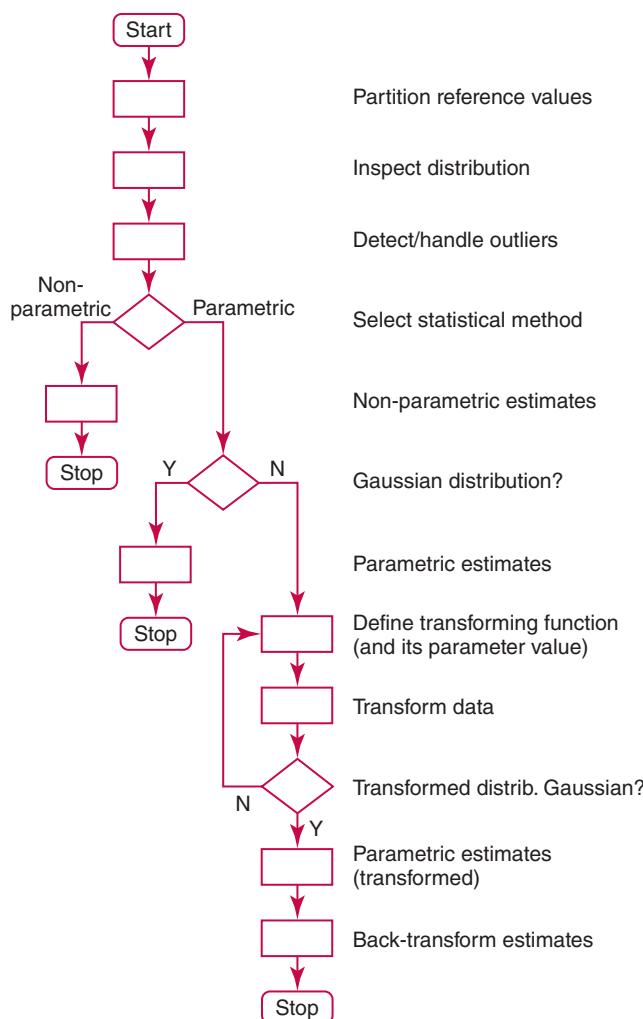
It is often claimed that analytical quality should be better when reference values rather than routine values are produced. This is certainly correct for trueness; all measures should be taken to select an appropriate reference standard (materials or methods) for traceability base and minimizing bias from that standard. The use of methods traceable to the Joint Committee for Traceability in Laboratory Medicine (JCTLM)-listed reference materials, methods, and services<sup>59</sup> (if available) increases the likelihood that reference limits are transferable among different laboratories. The question of imprecision is more difficult because it depends in part on the intended use of the reference values. Increases in analytical random variation result in widening of the reference interval.<sup>6,a</sup> For some special uses of reference values, the narrower reference interval obtained by a more precise analytical method may be appropriate. However, this usually is not true for routine clinical use of reference values. Interpretation is simplest if a patient's values and reference values are comparable with regard to analytical imprecision. For the same reason, it is advisable to analyze specimens from reference individuals in several runs to include between-run components of variation. A safe way to obtain comparability is to include these specimens in routine runs together with real patient specimens. Particular care must be taken with

<sup>a</sup>The width of a reference interval is a combination of the within-subject biological variation (coefficient of variation [ $CV_{I1}$ ]), the between-subject biological variation ( $CV_G$ ), the preanalytical variation ( $CV_{PA}$ ), and the analytical variation ( $CV_A$ ). Assuming for demonstration that these factors are all distributed in a Gaussian manner, the CV of the results produced in a reference interval study ( $CV_{RI}$ ) is  $\sqrt{CV_I^2 + CV_g^2 + CV_{pa}^2 + CV_A^2}$ . Thus a greater analytical imprecision leads to a wider reference interval.

analytical quality if multisite studies are performed with measurements performed at different locations. There is likely to be increased analytical variation due to between-instrument and between-laboratory factors, which must be kept small enough so that it does not adversely affect the results.

## STATISTICAL TREATMENT OF REFERENCE VALUES

This section deals with two main topics: the partitioning of reference values into more homogeneous classes, and the determination of reference limits and intervals.<sup>60</sup> The subject matter is presented in the order in which data are often treated. Fig. 9.2 gives an outline and refers to corresponding sections in the text. Before the presentation of methods, some statistical concepts used are briefly discussed (see also Chapter 2). A textbook by Harris and Boyd gives an excellent summary of the statistical bases of reference values in laboratory medicine.<sup>61</sup>



**FIGURE 9.2** The statistical treatment of reference values. The boxes in the flow chart refer to sections in the text. The order of the first three actions (partitioning, inspection, and detection and/or handling of outliers) may vary, depending on the distribution and the statistical methods applied. *N*, No; *Y*, yes.

## Basic Statistical Concepts

### Sample

The first step in the establishment of reference values is the selection of a group of reference individuals. In practice, it is not feasible to gather observations on all possible reference individuals of a certain category of the general population. Therefore a smaller group (sometimes called the reference *sample* group) is examined. This *subset* is chosen so that it is expected to give the desired information about the characteristics of the complete *set* of individuals (the reference *population*).<sup>10–15</sup>

The reference *population* is often considered to be *hypothetical* because its characteristics are not observed directly; neither the number (the set size) nor the properties of all of its individuals are known. An obvious requirement is that individuals in the subset are typical of those in the complete set. Statistical theory usually assumes that items in the subset are selected at *random* from among those in the set; otherwise, the subset may be biased. If items are not randomly selected, statistical techniques are still used, but only with due caution and with awareness of the possible bias introduced.

Two main types of inference may be made from values obtained from the subset (sample group) to the set (total reference population): estimating properties of the reference population (e.g., midpoint estimate and reference limits) and testing hypotheses related to the reference population (e.g., whether the distribution is Gaussian).

### Estimating Properties

In practice, properties of the set are estimated. A *reference limit* (a percentile) of a biological quantity, such as the activity of serum GGT, based on subset reference values, is an example of a *point estimate* (a single value). It is considered representative of the property that might have been found if all possible values in the set had been observed. If many randomly selected subsets from the same set are examined, several estimates with some variation around the “true” value of the set are obtained. Also, it is possible to produce an *interval estimate* bounded by limits within which the “true” value is located with a specified confidence: the *confidence interval*. The parameter is expressed as a percentage between 0 and 100%, indicating the degree on the scale between “never” and “always” that the point estimate lies within the interval estimate. A reference limit for serum GGT can thus be associated with a confidence interval showing its region of uncertainty (e.g., the 97.5th percentile for serum GGT is 47 IU/L, with 90% confidence limits of 39 to 50 IU/L).

### Testing Hypotheses

Hypotheses about the population distribution can be also tested. For example, one can state the *null hypothesis* that the distribution of values for serum GGT activities is Gaussian. If true, this will enable determination of the reference limits with relatively few points. If deviations of subset values from the Gaussian distribution are small, they can be ascribed to variation caused by chance alone. In that case, it is reasonable to use statistical methods based on the Gaussian distribution. However, the hypothesis must be rejected if it is unlikely that observed deviations from the Gaussian distribution are caused by chance alone. *Statistical tests* provide quantitative approaches to these types of decisions; the null hypothesis is rejected if the statistical test shows that the probability of the

hypothesis being true is less than a stated *significance level*. The *probability* ( $P$ ) is a number in the interval of 0 to 1, with higher values indicating a greater certainty. If a significance level of 0.05 is stated, the Gaussian hypothesis is tested for the distribution of serum GGT activities; it should be rejected if the probability obtained by the test is below this value (e.g., if  $P = .01$ , there is only a 1% chance that the distribution is Gaussian). Then the alternative hypothesis that the distribution is non-Gaussian is accepted. The *power* of a statistical test is the probability of rejection when the null hypothesis is false.

### Describing the Distribution

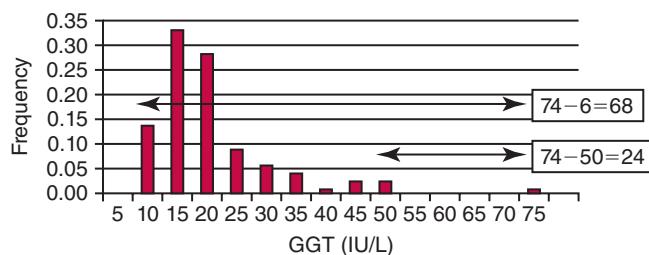
In the following sections, the term *reference distribution*<sup>10–15</sup> is used for the distribution of reference values ( $x$ ). For Gaussian distributions the two statistics *arithmetic mean* ( $\bar{x}$ ) and *standard deviation* ( $SD$ ) ( $s_x$ ) are measures of the location (based on a measure of the center of the distribution) and the dispersion of values in it, respectively. They are defined as follows:

$$\bar{x} = \frac{\sum x}{n}$$

$$s_x = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}} = \sqrt{\frac{\sum x^2 - (\sum x)^2/n}{n-1}}$$

where  $x$  represents each of the  $n$  reference values in the subset (or a subclass of it). The equations can be described in words to facilitate understanding. The arithmetic mean is the sum of all the values divided by the number of values. The SD is the square root of the result of the following calculation: the sum of the squares of all the differences of each value from the arithmetic mean divided by the number of samples minus one.

An observed distribution should be presented as a table or, preferably, as a graph (histogram) showing the number of observations in small intervals (Fig. 9.3). The number of observations in an interval divided by the total number of observations in the distribution (its size) is an estimator of the probability of finding a value in the corresponding interval of the hypothetical *probability distribution* of the population (assuming random sampling). By consecutive summing of all

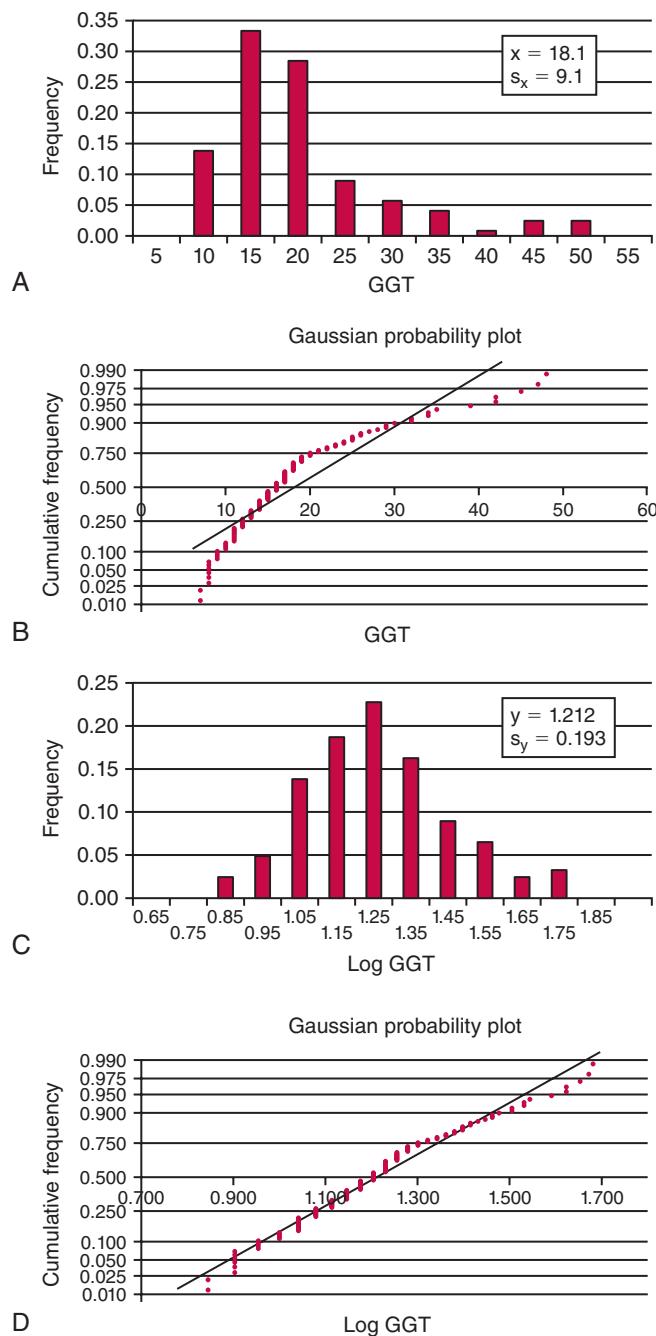


**FIGURE 9.3** Observed distribution of 124 gamma-glutamyltransferase (GGT) values in serum (IU/L). The *upper arrow* indicates the range of the observed values (highest – lowest, or  $74 - 6 = 68$ ); the *lower arrow* indicates the difference between the highest value and the next highest value ( $74 - 50 = 24$ ). Since the quotient ( $24/68 = 0.35$ ) exceeds 0.33, Dixon's range test indicates that the highest value is an outlier and is therefore omitted from further analyses.

these ratios, starting with the leftmost interval of the observed distribution, an estimate is obtained of the hypothetical *cumulative probability distribution*, shown as a normal probability plot in Fig. 9.4B.

### Reference Limits: Interpercentile Interval

As mentioned previously, reference values provide a basis for interpretation of laboratory data. In clinical practice, one



**FIGURE 9.4** Distribution of 123 remaining gamma-glutamyltransferase (GGT) values from reference subjects (A) is a histogram of the original, untransformed data, (B) shows the cumulative frequency of the data from (A) plotted on Gaussian probability paper, (C) is a histogram of the logarithmic transformed data, (D) shows the cumulative frequency of the data from (C) using a "normal probability plot."

usually compares a patient's result with the corresponding *reference interval*, which is bounded by a pair of *reference limits*.<sup>10–15</sup> This interval, which may be defined in different ways, is a useful condensation of the information carried by the total set of reference values.

This discussion will be confined to the *interpercentile interval*, which is (1) simple to estimate, (2) commonly used, and (3) recommended by the IFCC<sup>10–15</sup> and CLSI.<sup>16</sup> It is defined as an interval bounded by two percentiles of the reference distribution. A *percentile* denotes a value that divides the reference distribution such that specified percentages of its values have magnitudes less than or equal to the limiting value. For example, if 47 IU/L is the 97.5th percentile of serum GGT values, then 97.5% of the values are equal to or below this value.

It is an arbitrary but common convention to define the reference interval as the *central 95%-interval* bounded by the 2.5th and 97.5th percentiles.<sup>10–15</sup> Another size or an asymmetric location of the reference interval may be more appropriate in particular cases. For example, the 99th percentile has been recommended for cardiac troponins,<sup>62</sup> and the 80th percentile for lipoprotein(a).<sup>63</sup> To prevent ambiguity, the definition of the interval should always be stated. The estimation of percentiles presented in the following sections is based on the conventional central 95% interval, but the techniques are easily adapted to other locations of the limits.

The percentiles are point estimates of population parameters. Accordingly, they are unbiased estimates only if the subset of values was selected randomly from the population. But, as was discussed earlier, random sampling is often difficult to achieve. An interpercentile interval may always be used, however, as a summary or description of the *subset* reference distribution.

The precision of a percentile as an estimate of a population value depends on the size of the subset and the scatter of results around the percentile; it is less precise when few observations are reported or when data points are more widely scattered. If the assumption of random sampling is fulfilled, the *confidence interval* of the percentile (i.e., the limits within which the true percentile is located with a specified degree of confidence) can be determined. The 90% confidence interval of the 2.5th percentile (lower reference limit) for serum GGT values may, for example, be 6 to 8 IU/L, whereas the 90% confidence interval of the 97.5th percentile could be 39 to 50 IU/L. The upper limit confidence interval is wider because of a skewed distribution leading to more scattered data points.

### Methods Used to Determine Interpercentile Intervals

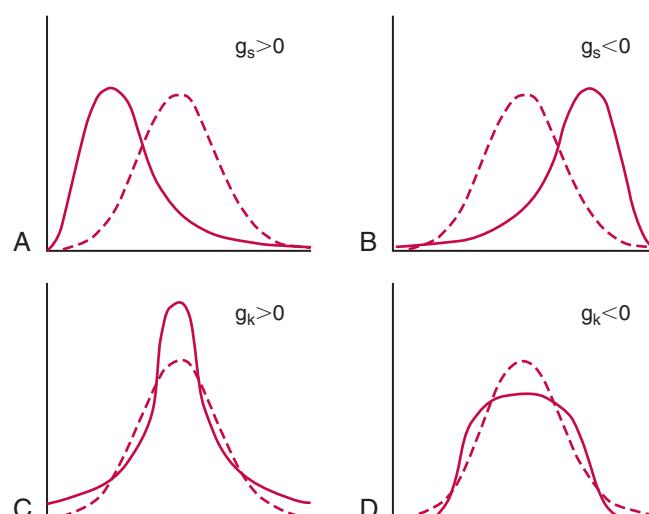
The interpercentile interval is typically determined based on one of two major method principles: parametric or nonparametric (Table 9.3).<sup>10–16</sup>

The *parametric method* has as its major advantage the need for fewer reference values to determine percentiles and their confidence intervals. It can be applied when the distribution can be described completely by a small number of population parameters. For example, in a Gaussian distribution, determination of the mean and SD allow for calculation of the 2.5th and 97.5th reference limits as the values located roughly two SDs below and above the mean. In fact, most of the parametric methods are based on the Gaussian distribution. If the reference distribution does not appear to be Gaussian, mathematical functions may be used to transform

**TABLE 9.3 Notable Differences Between Analysis Methods**

	Nonparametric	Parametric
Sample size (minimum number of reference individuals per partition)	120	40
Reference value distribution	No requirements Any distribution is acceptable	Gaussian distribution required
Ease of analysis	Straightforward No expertise required	Can be complicated Proof that distribution is Gaussian required Transformation of data may be required
Endorsements	IFCC, CLSI	

CLSI, Clinical Laboratory Standards Institute; IFCC, International Federation of Clinical Chemistry and Laboratory Medicine.



**FIGURE 9.5** Skewness and kurtosis. The two upper figures show asymmetric distributions (A, positive skewness; B, negative skewness). The two lower figures show distributions with non-Gaussian peakedness (C, positive kurtosis; D, negative kurtosis). The Gaussian distribution (dashed curve) is shown in all graphs for comparison. The values of the coefficients of skewness ( $g_s$ ) and kurtosis ( $g_k$ ) are also shown.

data to a distribution that approximates a Gaussian shape. Some positively skewed distributions (Fig. 9.5A) may, for example, be made symmetric by using logarithmic, Box-Cox, or other transformations of the data values.

In contrast, the *nonparametric method* has as its major advantage that it makes no assumptions concerning the type of distribution and does not use estimates of distribution parameters. Percentiles are determined simply by eliminating the required percentage of values in each tail of the subset reference distribution (typically 2.5%).

The simple nonparametric method for determination of percentiles is recommended by IFCC<sup>10–15</sup> and CLSI.<sup>16</sup> The

parametric method, which can be fairly complex, is seldom necessary, but it will be presented here because of its popularity and frequent misapplication.

Two other methods will be mentioned later in this chapter, but they are more complex and require the use of computer techniques (though these techniques are widely available in commercial software). It is worth emphasizing that, when results obtained using proper application of any of these methods are compared, it is usually found that estimates of the percentiles are very similar. Indeed if dissimilar results are obtained, investigation should be undertaken to consider possible causes. Detailed descriptions of nonparametric and parametric methods are given later in this chapter.

### Sample Size

For the parametric method, the theoretical lower limit of the sample size required for estimation of the  $100p$  and  $100(1 - p)$  percentiles is equal to  $1/p$ . Thus estimation of the 2.5th percentile requires at least  $1/0.025 = 40$  observations (per partition).

In contrast, for the nonparametric approach, a sample size of at least 120 reference values (per partition) has been recommended (the actual minimum is 119; however, this is commonly rounded up to 120); otherwise, one cannot determine confidence intervals for the reference limits.<sup>10–16</sup> It is important to note that 120 reference values allows for calculation of statistically valid 90% confidence limits; it does not necessarily provide the user with reference limits that would be considered clinically adequate. It is up to the scientists managing the study to determine whether the uncertainty of the estimates of the reference intervals meets the clinical need.

It should be noted that for any method (parametric or nonparametric), the precision of the percentile estimates improves as the number of observations increases. Additionally, different numbers of samples are required for different percentiles. For example, it requires 299 samples to determine the 90% confidence limit for the 99th percentile (vs. 120 for the 97.5th percentile). Also, the more highly skewed a distribution is, the larger is the number of reference values needed to obtain clinically reasonable confidence intervals at the tail end of the distribution.<sup>64</sup> The Nordic Reference Interval Project (NORIP)<sup>65</sup> provides a particularly good example of this phenomenon. The value for the 97.5th percentile for serum alanine aminotransferase (ALT) in males was 68, with 90% confidence limits of 63.4 to 73.6. Even though the study was based on 1080 subjects, the confidence interval represented more than 15% of the reference interval.

### Partitioning of Reference Values

The best order of the first three actions outlined in Fig. 9.2 (1—partitioning of reference values, 2—inspection of the distribution, and 3—detection/handling of outliers) may in some cases be different from that shown in the figure. For example, it might be more appropriate to detect outliers before testing for partitioning. No strict rules for the order of these actions can be given because it depends on data and the statistical methods applied. In addition, it can be argued that inspection of the data is important at each of the processes in the figure. With these cautions in mind, the presentation in this chapter follows Fig. 9.2.

The subset of reference individuals and corresponding reference values may be partitioned according to sex, age, and other characteristics (see Table 9.2). The process of partitioning

is also referred to as stratification, categorization, or subgrouping, and its results have been called partitions, strata, categories, classes, or subgroups. In this chapter, the terms *partitioning* (for the process) and *(sub)classes* (for its result) are used.

The aim of partitioning is to create more homogeneous subsets of data so as to provide a better basis for comparison of clinical laboratory results: *class-specific* reference intervals (e.g., age- and sex-specific reference intervals). An initial step is to graph the data against the relevant parameter. For example, plotting reference values against age will allow assessment as to whether partitioning is likely to be needed and, if so, which ages should be included in each class.

Various statistical criteria for partitioning have been suggested.<sup>61,66</sup> For example, an intuitive criterion states that partitioning is necessary if differences between classes are statistically significant (rejection of the “null” hypothesis of equal distributions). The distribution of reference values in the classes may show different locations (the mean values vary) or different intraclass variations (the SDs vary). These differences may be tested by statistical methods, which are not described here. The reader is referred to Chapter 2 and to standard textbooks of parametric<sup>67,68</sup> and nonparametric statistics.<sup>69</sup>

Differences in location or variation, however, may be statistically significant and still may be too small to justify replacing a single total reference interval with several class-specific intervals. In practice it is common to use a clinical assessment as to whether the proposed differences are likely to be important for patient care. For example, a difference within the analytical variation of the assay may be deemed sufficiently small to ignore. Alternatively, statistically nonsignificant differences can lead to situations in which the proportions of each subclass above the upper or below the lower reference limits (without partitioning) are much different from the desired 2.5% on each side. Harris and Boyd<sup>61</sup> therefore suggested criteria based on the ratio between subclass SDs, a normal deviate test of means, and calculation of critical decision values dependent on the sample size. Lahti and coworkers<sup>66,70</sup> suggested focusing directly on the proportions of each subgroup falling outside the combined population reference limits in order to determine whether partitioning is indicated. According to their approach, subgroup-specific reference limits are needed when more than 4.1% or less than 0.9% of any subgroup falls outside the combined reference limits. Advantages of their method are that it can be used with non-Gaussian and Gaussian distributions and that it takes into account differences in subgroup prevalences.

Partitioning requires large samples of reference values. If these are not used, subclass sizes may be too small for reliable estimates of reference intervals and there may be limited statistical power to identify true differences between classes.

To solve the subclass size problem, it has been suggested to estimate regression-based reference intervals. Instead of dividing, for example, the total material into several age classes, one may construct continuous age-dependent reference limits and their confidence regions. Simulation studies have shown that this method produces reliable estimates with smaller sample sizes.<sup>71,72</sup> When the intended purpose of the reference interval is to detect individual changes in biochemical status, subject-based reference values may be more appropriate than class-specific reference intervals for interpretation.<sup>61,73,74</sup>

In the following sections, a homogeneous reference distribution and either the complete distribution (if partitioning has been shown to be unnecessary) or a subclass distribution (after partitioning) are assumed.

### Inspection of the Distribution

It is always advisable to display the reference distribution graphically and to inspect it. A *histogram*, as shown in Fig. 9.3, is easily prepared and is the type of data display best suited for visual inspection. Examination of the histogram serves as a safeguard against misapplication or misinterpretation of statistical methods, and it may reveal valuable information about the data. Data should be evaluated for the following characteristics of the distribution:

1. Highly deviating values (*outliers*) may represent erroneous values.
2. *Bimodal* or *polymodal* distributions have more than one peak and may indicate that the distribution is nonhomogeneous because of mixing of two or more distributions. If so, the criteria used to select reference individuals should be reevaluated, or partitioning of the values according to age, sex, or other relevant factors should be attempted.
3. The shape of the distribution should be noticed. It may be asymmetrical, or it may be more or less peaked than the symmetrical and bell-shaped Gaussian distribution (see Fig. 9.5). The asymmetry most frequently observed with clinical chemistry data is positive *skewness* (see Fig. 9.5A). A symmetric distribution with positive *kurtosis* has a high and slim peak and a greater number of values in both tails than the Gaussian type of distribution (see Fig. 9.5C). Conversely, negative kurtosis indicates that the distribution has a broad and flat top with relatively few observations in the tails (see Fig. 9.5D). Asymmetry and non-Gaussian peakedness may be combined.

The visual inspection may also provide initial estimates of the location of reference limits that are useful as checks on the validity of computations. Assessment of the shape of the distribution can lead to a number of actions. It can guide the choice of approach (i.e., parametric or nonparametric) or the necessity for transformation before using a parametric approach. As noted earlier, a skewed distribution may be a true reflection of the population, but it, or a secondary population, may also raise a question about possible causes: for example, the effect of a covariate (e.g., age affecting part of the population), an analytical problem (e.g., bias during one analytical run), or a preanalytical problem (e.g., samples from one site handled differently from other sites). Visual inspection can also be a valuable tool to assess for data quality and a possible need for partitioning. Viewing the data with age, sex, or time of collection/analysis on the *x*-axis may highlight important changes with these parameters.

### Identification and Handling of Erroneous Values

An *erroneous value* may occur due to a gross deviation from the prescribed procedure for establishment of reference values.<sup>45</sup> Such values may deviate significantly from proper reference values (*outliers*) or may be hidden in the reference distribution. Only a strict experimental protocol, with adequate controls at each step, can eliminate the latter type of erroneous values.

An outlier has been defined as “the observation in a sample, so far separated in value from the remainder as to

suggest that it may be from a different population, or the result of an error in measurement.”<sup>75</sup> This definition has some particular utility because it focuses on the possible nature of the expected distribution as outlined in the subsequent text.

As stated previously, *visual inspection* of a histogram is useful in screening for identification of possible outliers. It is important to keep in mind, however, that values far out in the long tail of a skewed distribution may easily be misinterpreted as outliers. If the distribution is positively skewed, inspection of a histogram after logarithmic or some other transformation of the values may aid in the visual identification of outliers. In the end, though, statistical tests must be used to make a final determination.

Some outliers may be identified by statistical tests independent of visual inspection, but no single method is capable of detecting outliers in every situation that may occur. The number of techniques suggested or recommended is, for this reason, very large.<sup>61,76,77</sup> The two main problems encountered can be described as follows:

1. Many tests assume that the type of the true distribution is known before the test is used. Some of these specifically require that the distribution be Gaussian. However, biological distributions often are non-Gaussian, and their types are seldom known in advance. Furthermore, statistical tests of types of distribution are unreliable in the presence of outliers. This unreliability poses a difficult dilemma; some tests for outliers assume that the type of distribution is known, but tests for determining the type of distribution require that outliers be absent! As a consequence, it may be difficult to transform the distribution to Gaussian form before outliers are identified by statistical tests. Some tests are relatively insensitive to departures from a Gaussian distribution. This is the case with Dixon's *range test*, in which a value is identified as an extreme outlier if the difference between the two highest (or lowest) values in the distribution exceeds one third of the range of all values (see Fig. 9.3).<sup>10–15,61,78</sup>
2. Several tests for outliers assume that a data set contains only a single outlier. The limitation of these tests is obvious. Some tests may detect a specified number of outliers, or they may be run several times, discarding one outlier in each pass of data. The range test, however, usually fails in the presence of several outliers. It is possible to estimate the SD using data remaining after *trimming* of both tails of the distribution by a specified percentage of observations.<sup>61,79</sup> Outliers could be identified by this method as the values lying 3 or 4 SDs from the arithmetic mean. This method assumes, however, that the true distribution is Gaussian.

Horn and coworkers<sup>80</sup> published a novel method in two stages for outlier detection that seems to provide a promising solution to both of the problems just mentioned. With this method, one executes the following:

1. Mathematically transform the data to approximate a Gaussian distribution. Horn used the Box-Cox transformation,<sup>81</sup> but other transformations that correct for skewness (see later) probably would also work. As mentioned earlier, it is impossible to achieve exact symmetry by transformation in the presence of outliers, but this does not seem to be critical with Horn's method.
2. Identify (or eliminate) outliers using a criterion based on the central 50% of the distribution, thus reducing the

masking effect of several outliers. Compute the interquartile range (IQR) between the lower and upper quartiles of the distribution ( $Q_1$  and  $Q_3$ , respectively):  $IQR = Q_3 - Q_1$ . Then identify as outliers data lying outside the two fences

$$Q_1 - 1.5 \times IQR \text{ and } Q_3 + 1.5 \times IQR$$

Deviating values identified as possible outliers cannot always be discarded automatically. Values should be included or excluded on a rational basis. For example, records of the dubious values should be checked and errors corrected. In some cases, deviating values should be rejected because noncorrectable causes have been found, such as in previously unrecognized conditions that qualify individuals for exclusion from the group of reference individuals or analytical errors.

### Methods for Determining Reference Limits—Suitable for Direct Sampling

More details regarding four different approaches to determining reference limits are discussed in this section. In addition to the nonparametric and parametric methods described in general terms earlier, overviews of the bootstrap and robust methods are provided. In all four cases, it is important to remember that, at this stage, a homogeneous reference distribution, with outliers removed, is assumed.

#### Nonparametric Method

The nonparametric method is notable for at least three reasons: it is simple to perform, it does not require that the distribution is Gaussian, and it is recommended by both IFCC<sup>10–15</sup> and CLSI.<sup>16</sup> On the other hand, for estimates of the central 95 percentiles, it does require a minimum of 120 values (per partition). It consists essentially of eliminating a specified percentage of the values from each tail of the reference distribution. Very simple and reliable methods are based on *rank numbers*.<sup>10–15,78,82</sup> These methods also allow nonparametric estimation of the confidence intervals of the percentiles<sup>78</sup> and can easily be applied manually or with a spreadsheet program.

The rank-based method as recommended by the IFCC<sup>10–15</sup> and CLSI<sup>16</sup> requires the following steps:

1. First, the  $n$  reference values are sorted in ascending order of magnitude.
2. Next, the individual values are ranked. For example, the minimum value has rank number 1, the next value has rank number 2, and so on, until the maximum value, which has rank number  $n$ . Consecutive rank numbers should be given to two or more values that are equal (“ties”).
3. The rank numbers of the  $100p$  and  $100(1 - p)$  percentiles are computed as  $p(n + 1)$  and  $(1 - p)(n + 1)$ , respectively. Thus the limits of the conventional 95% reference interval have rank numbers equal to  $0.025(n + 1)$  and  $0.975(n + 1)$ ; in a data set of 120, these are the 3rd and 118th ranked results.
4. The percentiles are determined by finding the original reference values that correspond to the computed rank numbers, provided that the rank numbers are integers. Otherwise, one should interpolate between the two limiting values. Note that the final values selected should have the same number of significant figures as will be used to report clinical results.
5. Finally, the confidence interval of each percentile is determined by using the binomial distribution.<sup>78</sup> Table 9.4

**TABLE 9.4 Nonparametric Confidence Intervals of Reference Limits**

Sample Size	RANK NUMBERS		
	Lower 0.90 CI Limit	2.5th Percentile	Upper 0.90 CI Limit
119–132	1	4	7
133–160	1	4	8
161–187	1	5	9
188–189	2	5	9
190–200	2	5	10
201–219	2	6	10
219–240	2	7	10
240–248	2	7	11
249–249	2	7	12
250–279	3	7	12
280–307	3	8	13
308–309	4	8	13
310–320	4	8	14
321–340	4	9	14
341–360	4	9	15
361–363	4	10	15
364–372	5	10	15
373–400	5	9	16
401–403	5	11	16
404–417	5	11	17
418–435	6	11	17
436–440	6	11	18
441–468	6	12	18
469–470	6	12	19
471–481	6	13	19
471–500	7	13	19

The table shows the rank numbers of the 2.5th percentile together with the lower and upper limits of the 0.90 confidence interval for samples with 119 to 500 values. To obtain the corresponding rank numbers of the 97.5th percentile, subtract the rank numbers in the table from  $(n + 1)$ , where  $n$  is the sample size. Note that the 2.5th percentile values are the nearest number from the data set and may differ from results derived from statistical software packages that commonly derive the percentile values by interpolation between results from the data set when the rank value does not correspond to the exact percentile.

CI, Confidence interval.

provides data for the 0.90 confidence interval of the 2.5th and 97.5th percentiles. For the relevant sample size  $n$ , rank numbers for the lower and upper limits should be found for the 2.5th percentile; those same values are subtracted from  $(n + 1)$  to find the rank numbers for the 97.5th percentile. Several software packages are available to determine the percentiles and their uncertainties.<sup>83–85</sup>

Table 9.5 provides a detailed example of the nonparametric determination of 95% reference limits using the serum GGT reference values first shown in Fig. 9.3.

It is claimed that the nonparametric process is less affected by outliers than parametric methods, a statement that has some truth. For example, a single extreme outlier in a set of 120 results will affect the calculated mean and SD, but it will not affect the 2.5th and 97.5th percentiles. It will, however, affect the 90% confidence interval of the nonparametric method, as the extreme result is the boundary of the confidence

**TABLE 9.5 Nonparametric Determination of Reference Interval**

Calculation of Rank Numbers of Percentiles		
Lower:	0.025 (123 + 1) = 3.1 (i.e., Rank #3)	
Upper:	0.975 (123 + 1) = 120.9 (i.e., Rank #121)	
Original Values Corresponding to These Rank Numbers		
Lower limit (2.5th percentile):	7 IU/L	
Upper limit (97.5th percentile):	47 IU/L	
Rank Numbers and Values of the 0.90-Confidence Limits		
<b>Lower Reference Limits</b>		
Rank numbers (see Table 8.4):	#1 and #7	
Values:	6 and 8 IU/L	
<b>Upper Reference Limits</b>		
Rank numbers (see Table 8.4):	(123 + 1) – 7 = #117 and (123 + 1) – 1 = #123	
Values:	39 and 50 IU/L	
<b>Summary</b>		
Lower reference limit:	7 (6 to 8) IU/L	
Upper reference limit:	47 (39 to 50) IU/L	

GGT Value	Frequency	Rank Order
6	1	1
7	2	2, 3
8	6	4–9
9	4	10–13
10	4	14–17
11	9	18–26
12	7	27–33
13	7	34–40
14	9	41–49
15	9	50–58
16	8	59–66
17	11	67–77
18	8	78–85
19	5	86–90
20	3	91–93
21	2	94, 95
22	2	96, 97
23	2	98, 99
24	2	100, 101
25	3	101–104
26	2	105, 106
27	1	107
28	1	108
29	2	109, 110
30	1	111
32	2	112, 113
34	2	114, 115
35	1	116
39	1	117
42	2	118, 119
45	1	120
47	1	121
48	1	122
50	1	123

This table shows an example using the 123 serum gamma-glutamyltransferase (GGT) values displayed in Fig. 9.4A. See text for a description of the nonparametric method.

limit with these numbers. It can be easily seen, however, that a larger number of outliers will influence the percentile limits as well.

It should be emphasized that Table 9.4 is specific for the 2.5th and 97.5th percentiles and 90% confidence limits. It can be seen from the table that 0.90 CI are not available for fewer than 119 samples. For this reason, when fewer than 120 samples are included in reference interval studies provided by manufacturers, a narrower reference interval is sometimes provided [e.g., the central 90% (5th through 95th percentiles)]. As noted previously, a larger number of samples is required to generate 0.90 CI for wider reference intervals (e.g., the central 98%, or 1st through 99th percentiles); a smaller number is required for narrower reference limits (e.g., the central 90%, or 5th through 95th percentiles).

### Parametric Method

Although it can be more complicated, usually involving statistical software, the parametric method is advantageous (in comparison to the nonparametric method) in requiring fewer reference values to determine reference limits. The method is presented here under separate headings for testing the type of distribution, for transforming the data, and for estimating percentiles and their confidence intervals.

It should be noted that commonly used statistical computer program packages aid in the estimation of reference limits, but these packages may lack some of the techniques described in this chapter. Several programs have been

designed with clinical laboratories in mind and have specific functions to perform many of these processes, including CB-stat,<sup>83</sup> MedCalc,<sup>84</sup> and Analyse-it.<sup>85</sup> The availability of these and other specialized programs will change over time, but it can be most useful to select a program that meets the needs of the laboratory, gain skills in its correct use, and maintain use of the same program over time. Basic statistical analysis can also be performed in common spreadsheet programs (e.g., Microsoft Excel), but the more sophisticated features like confidence limits may require writing special functions into the spreadsheet.

**Testing fit to Gaussian distribution.** The parametric method for estimating percentiles assumes that the true distribution is Gaussian. This fact was frequently ignored in the past and caused Elveback<sup>5</sup> to warn against “the ghost of Gauss.” This assumption may result in seriously biased estimates of reference limits.<sup>86</sup> Simple signs that a distribution is highly unlikely to be Gaussian are skewed distribution on visual inspection of the distribution, a mean and median that are markedly different, and  $S_x$  above approximately 30% of the mean value. In any of these cases, formal assessment for Gaussian distribution is unnecessary. After elimination of the outlier from the GGT reference values in Fig. 9.3, the mean and SD of the remaining 123 serum GGT reference values are 18.1 and 9.1 (see Fig. 9.4A), from which the reference interval is calculated as  $\bar{x} \pm 1.96 \times S_x$ , or 0 to 36 IU/L (vs. the nonparametric values of 7 and 47 IU/L; Table 9.6). More highly positively skewed distributions may even result in negative values for the lower reference limit.

**TABLE 9.6 Summary of Gamma-Glutamyltransferase Reference Interval Determination by Three Methods**

Method	Midpoint	Lower Limit (CI)	Upper Limit (CI)	Values Below Lower Limit	Values Above Upper Limit
Nonparametric	16	7 (6 to 8)	47 (39–50)	1	2
Parametric—untransformed data	18	0 (−2 to 2)	36 (34–38)	0	7
Parametric—transformed data	16	7 (6 to 8)	40 (35–44)	1	6

The table summarizes the midpoint, central 95% and associated 90% confidence limits of the reference intervals generated by each of three methods for the same data set. The numbers of observed values deemed lower and higher than the corresponding interval for each method are given in the last two columns. Because the original data are positively skewed, note that the parametric techniques generate intervals that are biased low. Note, also that the parametric technique on untransformed data has a lower confidence interval, which is actually less than 0. CI, Confidence interval.

Therefore a critical phase in the parametric method is testing the goodness-of-fit of the reference distribution to a hypothetical Gaussian distribution. If the Gaussian hypothesis must be rejected at a specified significance level, one is left with two alternatives (see Fig. 9.2): either the nonparametric method can be used, or a mathematical transformation of data can be applied to approximate the Gaussian distribution. Only when the Gaussian hypothesis is not rejected by the test can one pass directly to parametric estimation of percentiles and their confidence intervals (see Fig. 9.2).

Formal goodness-of-fit tests have been reviewed by Mardia.<sup>87</sup> These tests can be broadly classified as (1) graphical procedures, (2) coefficient-based tests, and (3) tests that are based on shape differences between observed and theoretical distributions.

1. *The graphical procedure* consists of plotting the cumulative distribution on probability paper, which has a nonlinear vertical axis based on the Gaussian distribution (see Fig. 9.4B and D). The plot should be close to a straight line if the distribution is Gaussian.
2. *Coefficient-based tests* use statistical measures of skewness and kurtosis (see Fig. 9.5). Formulas for calculating these parameters are available elsewhere,<sup>10–15</sup> or they may be produced by statistical or reference interval software.<sup>83–85</sup> For Gaussian (and other symmetric distributions), the *coefficient of skewness* is zero; the sign of a nonzero coefficient indicates the type of skewness present in the data (see Fig. 9.5A and B). The *coefficient of kurtosis* is approximately zero for the Gaussian distribution. The sign of a nonzero coefficient indicates the type of kurtosis present in the data (see Fig. 9.5C and D). The statistical significance of these two coefficients may be found by referring to tables for testing skewness and kurtosis.<sup>68</sup>
3. Tests of *shape differences* that have been used to evaluate goodness-of-fit include the (1) Kolmogorov-Smirnov, (2) Cramer-von Mises, and (3) Anderson-Darling tests.<sup>10–15,88</sup> The Anderson-Darling test is recommended by the IFCC.<sup>10–15</sup>

**Transformation of data.** In the previous section, it was shown that  $\bar{x} \pm 1.96 \times S_x$  of the serum GGT data in Fig. 9.4A resulted in biased reference limits (too low values), as was to be expected with this positively skewed distribution. However, it is often possible to transform data mathematically to obtain a distribution of transformed values that approximates a Gaussian distribution. With these new values, the 2.5th and 97.5th percentiles are again localized at 1.96 SDs on both sides of the mean. The estimates may then be

transformed back to the original measurement scale by using the inverse mathematical function.

It is frequently observed that *logarithmically transformed* values,  $y = \log(x)$ , of a positively skewed distribution fit the Gaussian distribution rather closely. In other cases, *square roots* of the values,  $y = \sqrt{x}$ , result in a better approximation to the Gaussian distribution. In theory, any mathematical transformation of the data can be used. From a practical perspective, the family of Box-Cox transformations provides solutions in the vast majority of situations where the transformation parameter ( $\lambda$ ) can be selected to transform right-skewed distributions which may be more or less skewed than a logarithmic distribution.<sup>86</sup>

The following example uses the logarithmic transformation for convenience, but any other transformation can be used in the same way. The procedure is as follows:

1. Test the fit of the distribution of original data to the Gaussian distribution. If the distribution has approximately a Gaussian shape, the 2.5th and 97.5th percentiles are calculated directly as  $\bar{x} \pm 1.96 \times S_x$ . Otherwise, continue with the following steps.
2. Transform data by the logarithmic function  $y = \log(x)$  (or by another selected function), then test the fit to the Gaussian distribution. If the transformed distribution is significantly different from Gaussian shape, try another transformation or estimate the percentiles by the nonparametric method (see earlier in this chapter). Continue with the next step if the transformation resulted in a Gaussian distribution.
3. Compute the mean  $\bar{y}$  and the SD  $S_y$  of transformed data. Then estimate the 2.5th and 97.5th percentiles in the transformed data scale as  $\bar{y} \pm 1.96 \times S_y$ .
4. The final step is reconversion of these percentiles to the original data scale. The inverse function for the logarithmic transformation  $y = \log x$  is  $x = 10^y$ .
5. It is now possible to use the properties of the Gaussian distribution to estimate the reference limits and their confidence intervals. This method is presented in a later section.

**Example:** As noted earlier, the original GGT data reference distribution is not Gaussian but is, as with many biological distributions, skewed to the right (see Fig. 9.4A). However, by using the logarithm of the serum GGT values, a distribution very close to Gaussian shape (see Fig. 9.4C) is obtained. This observation is confirmed in Fig. 9.4B and D where the cumulative probabilities are shown graphed on Gaussian probability paper; the original data are not linear, but the transformed data form a reasonably good line.

### Parametric estimates of percentiles and their confidence intervals.

Once the distribution of reference data (original or transformed) is shown to be Gaussian, calculations of the  $100p$  and  $100(1 - p)$  percentiles and their 0.90 confidence intervals are straightforward:

As noted earlier, the  $100p$  and  $100(1 - p)$  percentiles are calculated as follows:

$$\text{mean} \pm c \times (\text{standard deviation})$$

where  $c$  is the  $(1 - p)$  standard Gaussian deviate, as can be found in statistical tables. For the 2.5th and 97.5th percentiles, the  $(1 - 0.025) = 0.975$  standard Gaussian deviate,  $c$ , has a value of 1.960.

The 0.9-confidence intervals of these percentiles are then determined as follows<sup>7,10–15</sup>:

$$\text{percentile} = \pm 2.81 \frac{s_y}{\sqrt{n}}$$

where  $s_y$  is the SD of the reference values (original or transformed) and  $n$  is the number of values.<sup>b</sup>

*Example:* The mean and SD of the transformed data in Fig. 9.4 are  $\bar{y} = 1.212$  and  $s_y = 0.193$ , respectively; that is, the mean value is 1.212 (corresponding to  $10^{1.212}$ , or 16 in the original scale). The transformed 2.5th percentile is then  $1.212 - (1.960 \times 0.193) = 0.835$ . On reconversion to the original data scale, a value of  $10^{0.835} = 6.84$  is obtained. The lower reference limit of serum GGT is thus 7 IU/L. Similarly, it is found that the upper reference limit is 39 IU/L. These values are in closer agreement with those found by the nonparametric method: 7 and 47 IU/L (see Tables 9.5 and 9.6).

The 0.90 confidence limits of the lower percentile are then

$$0.835 - 2.81 \left( 0.193 / \sqrt{123} \right) = 0.786 \quad 10^{0.786} = 6.1$$

$$0.835 + 2.81 \left( 0.193 / \sqrt{123} \right) = 0.884 \quad 10^{0.884} = 7.7$$

Thus the complete estimate of the 2.5th percentile (and its 0.90 confidence interval) is 7 (6 to 8) IU/L. The 97.5th percentile is, by the same method, found to be 39 (35 to 43) IU/L.

Table 9.6 summarizes data from the three methods used to determine reference intervals from GGT data. It can be seen that the application of parametric statistics to transformed data yields similar reference limits to those obtained by nonparametric methods. While nonparametric methods have the advantage of simpler mathematical processes and determining reference limits without assumptions on the underlying distribution, there can be some advantages to using parametric methods. If an underlying distribution can be defined, it can assist in assessing the likelihood of a result being a member of the reference population based on the parameters defining the population.

### Other Methods

As a brief introduction to the bootstrap and robust methods for determining reference limits, it is worth noting that they share two characteristics. First, neither of these methods makes assumptions about the underlying distribution; it need not be Gaussian. Second, both require the use of computer software because they involve numerous iterations and somewhat complicated calculations. These methods are appropriate for use in direct reference interval studies or indirect studies provided the likelihood of nonhealthy subjects can be made sufficiently small.

**Bootstrap method.** There are a number of variations on the “bootstrap” method, all of which can be used to generate reliable reference limits.<sup>61,82,89</sup> In principle, the technique is simple, but it involves many iterations (100 to 1000) and thus requires computers. As is the case with the nonparametric method, there is no requirement that the distribution be Gaussian. For reliable estimates of confidence intervals for the reference limits, it is recommended that there be a minimum of 100 reference values (per partition).<sup>82</sup> The following steps are involved in a typical bootstrap procedure:

1. First, random samples, each of size  $m$ , are selected, with replacement, from the original set of  $n$  reference values. One selects “with replacement” if each value randomly selected from the original set remains available, so that it may be selected again in the random selection of the next value. In other words, even if there is only one occurrence of a specific value in the original set of  $n$  values, it may appear more than once in one, or more, random samples of size  $m$ . The number of resamples should be high (500 is a reasonable number of iterations).
2. For each resample, the upper and lower reference limits (percentiles) are next estimated by the rank-based nonparametric procedure described previously. These estimates from each iteration are saved.
3. Upon completion of all iterations, the final lower reference limit is calculated as the mean of the estimates of the lower reference limit; similarly, the final upper reference limit is calculated as the mean of the estimates of the upper reference limit.
4. Finally, the 0.90 confidence interval of each reference limit is calculated from the distribution of the percentile estimates, that is, with 500 iterations, the confidence interval for the 2.5th percentile (the mean of ranks 13) would be the means of ranks 7 and of ranks 19.

The reader should note that the bootstrap version described here uses rank-based nonparametric percentile estimates. However, the bootstrap principle may be employed with any kind of estimation, parametric or nonparametric.

**Robust method.** The robust method has the form of the parametric method described earlier, but instead of using the mean and the SD of the sample, it uses robust measures of location and spread. For example, instead of using the mean, it uses the median: in a series of 10 values, if the highest value is doubled, the mean changes appreciably, but the median does not change at all. The process involves weighting the data to place more value on results near the middle of the distribution and less weight on more distant results. The rationale is that the scattered data are more likely to reflect results that are not members of the desired distribution. This resistance to the effect of outliers is the basis for the term *Robust Method*. This method has particular value with small sample sizes.

<sup>b</sup>This formula is a special case of a general formula that can be used for confidence intervals of other sizes or for other percentiles derived from Gaussian distributions. CI for percentiles are calculated as follows:  $\text{mean} \pm z_1 \times s_y \pm z_2 \times [s_y^2/n + (z_1^2 \times s_y^2)/(2 \times n)]^{0.5}$  where mean is the population mean,  $s_y$  is the population SD,  $n$  is the sample size,  $z_1$  is the probit value related to the selected percentile (=1.96 for 97.5th percentile), and  $z_2$  is the covering factor for the CI (=1.64 for 90%). ([http://www.statsdirect.com/help/Default.htm#parametric\\_methods/reference\\_range.htm](http://www.statsdirect.com/help/Default.htm#parametric_methods/reference_range.htm).)

Briefly, the steps involved are as follows:

1. Symmetry of the data is ensured, using transformations if necessary (e.g., Box-Cox transformation<sup>81</sup>).
2. Initial robust measures of location (median) and spread (median absolute deviation) are found.
3. Using a *biweight estimation* technique, in which more weight is given to observations closer to the center and progressively less to values farther from the center, new estimates of location and spread are found until successive results are satisfactorily close.
4. With final robust values of location and spread, the upper and lower limits are calculated, in a manner analogous to that described for the parametric technique.
5. Confidence intervals are then estimated using the bootstrapping technique described in the previous section.

Similar to the bootstrap method, this method does not require a Gaussian distribution. It is resistant to outliers and may be applied to very small numbers of observations. Details on the method are available.<sup>90</sup>

### Methods for Determining Reference Limits—Suitable for Indirect Sampling

The methods above are recommended for direct reference interval studies, but they are not suitable for indirect studies when appreciable numbers of results from diseased subjects is likely. To address this issue, a number of statistical approaches have been developed to minimize the effect of the presence of results not representing the reference population.<sup>37</sup>

### Graphical-Based Methods

In the precomputer era, a number of alternate methods for computing reference intervals were developed based on analysis of graphical display of reference values. The best known of these are the Hoffmann<sup>35</sup> and Bhattacharya<sup>36</sup> methods. Many reference interval studies have been performed using these tools. Both methods can be performed using basic computer spreadsheets or with third-party software. The methods are based on finding a Gaussian distribution in the midst of other data. With care and the right data set, both methods can provide useful results. Each method has some user-defined steps which can affect the accuracy, and therefore care and understanding of the processes is required to ensure accurate results are produced. The methods are more robust with larger numbers of values; while no minimum has been determined, inclusion of several thousand results is likely to lead to more accurate reference intervals. Approaches to minimize risk of such errors include the use of more than one statistical method to estimate the reference limits and, as mentioned previously, the visual inspection of the original data with the derived reference intervals. For both the Hoffmann and Bhattacharya techniques, a key factor is the use of transformations such as logarithms or one of the Box-Cox family. For analytes with a narrow, symmetrical between-person variation (e.g., serum sodium, calcium, albumin, and other analytes with a group coefficient of variation ( $CV_G$ ) less than about 15%) a good Gaussian fit is often possible on untransformed data. For analytes with a wider, skewed distribution, transformation is required. One approach may be to use a transformation (e.g., a value for lambda in a Box-Cox distribution) which has been shown to describe a healthy distribution from a direct study.<sup>50</sup> Applying a transformation

to give the best fit on the data may “normalize” pathology by bringing it into the central data peak. An example may be liver transaminases under the effect of fatty liver, where the “best fit” transformation may wrongly include subjects with this condition being included within the reference limits.<sup>50</sup> An important limitation of these methods is the lack of a confidence interval for the reference limits.

**Hoffmann Method.** The Hoffmann technique was developed in 1963 and is based on a display of the data as a cumulative distribution using normal probability paper (or electronic equivalent) and identifying the linear section which indicates a Gaussian distribution.<sup>35</sup> The assumption of this technique is that the majority of the data in the central region has a Gaussian distribution, and one of its key advantages is that distant outliers have no effect on the outcome. Key risks with the use of this procedure are closely overlapping distributions which may not be separated or the presence of a secondary population of significant size.<sup>91</sup> The decisions that must be made are whether any transformation of the original data is indicated and how much of the final data set to include as reflecting a Gaussian distribution. A revised version described as “computerized Hoffmann method” has been developed<sup>92,93</sup> although this does not use the normal probability data analysis and should be considered a separate method<sup>94</sup> with less satisfactory performance.<sup>95</sup>

**Bhattacharya Method.** The Bhattacharya method, like Hoffmann, is also a graphical method for identifying a Gaussian distribution in the midst of other data.<sup>36</sup> The procedure is able to separate overlapping distributions after graphical display.<sup>36,91</sup> Computer-based versions have been developed in Microsoft Excel and R programming language. The Bhattacharya method has been shown to be less influenced by data not included in the Gaussian distribution than is Hoffmann.<sup>91</sup> This method has been subject to review<sup>34,96,97</sup> and has been used in a number of published papers.<sup>98,99</sup> The variables that must be selected are the bin size and bin location to use in analyzing the data and, as with the Hoffmann technique, whether any transformation of the original data is indicated and how much of the final data set to include as reflecting a Gaussian distribution. An additional factor after transformation is variation in bin sizes of the data which can produce results which are difficult to model. The use of extra significant figures in the raw data can minimize this effect.

### Special Computerized Approaches

The key limitations to the graphical methods above are the quality of the separation of a deemed healthy subpopulation from other subpopulations that may be affected by disease, and the handling of data transformation. Computer-based solutions that address both of these issues have been developed.

**DGKL Reference Limit Estimator.** Developed by the German scientific society for Clinical Chemistry and Laboratory Medicine (DGKL) working group on reference intervals, a description and software is available on the DGKL website.<sup>100</sup> In this process a Gaussian distribution is automatically fitted to the majority of the data with an optimal transformation, and this fit is used to define the reference limits with the distribution of pathologic results also determined.<sup>38</sup> Additionally the software calculates possible clinical decision

limits as the intersection point of the nonpathologic and pathologic density curves (bimodal reference limit with the highest diagnostic efficiency) and checks for possible analytical trend during the time of data collection and considers automatic stratification according to sex and age.

**Mixed Likelihood Techniques** An alternate method is based on mixed likelihood techniques used in other fields.<sup>95,101</sup> This method separates likely healthy and diseased subpopulations, each of which may have Gaussian or skewed distributions and may overlap to varying degrees. There are a number of statistical tools available for this approach and superiority over Hoffmann and Bhattacharya has been claimed.<sup>95</sup>

## TRANSFERABILITY OF REFERENCE LIMITS FROM OTHER SOURCES

Determination of reliable reference values for each test in the laboratory's repertoire is a major task that is often far beyond the capabilities of the individual laboratory. This is especially important when ethical or practical considerations limit the number of available individuals (e.g., when establishing pediatric or cerebrospinal fluid [CSF] reference values). However, even in the absence of such considerations, most of the methods discussed in the previous sections require qualification of, and analysis of samples from, relatively large numbers of reference individuals.

Two issues are critical in considering adopting reference intervals derived from the other sources discussed in the following sections. First, the populations under consideration must be comparable (i.e., no major ethnic, social, or environmental differences should be noted between them that may be relevant to the analyte in question). If they are not, a separate reference interval study may well need to be done. Second, even if the populations are comparable, the analytical methods under consideration must be comparable. The optimal, but often unrealistic, situation assumes that analytical methods, including their calibration and quality assurance, are identical in the laboratories. The provision of methods from different manufacturers that are all traceable to equivalent higher-order reference materials and methods, such as are listed on the JCTLM database, facilitates the sharing of reference intervals.<sup>102</sup> In the absence of verified traceable methods, a pragmatic approach involves (1) standardization of analytical protocols, (2) common calibration, (3) design of a sufficiently efficient external quality control scheme, and (4) the use of mathematical transfer functions if results still are not directly comparable.<sup>103</sup> The parameters of transfer functions may be estimated from results obtained by analysis of a sufficient number of patient specimens spanning the relevant range of concentrations in all participating laboratories.<sup>16</sup> Provided both assays are linear, functions of the form  $y = \alpha \times x + \beta$  are generally appropriate, where the constant term  $\beta$  compensates for systematic shifts among methods, whereas the coefficient term  $\alpha$  adjusts for proportional differences. Care should be taken to ensure that errors do not occur due to the use of inappropriate statistical techniques. For example, simple linear regression can be affected by outlier values, variation in data dispersion at different analyte concentrations (heteroscedasticity), and a limited range of analyte concentrations. The use of Passing and Bablok<sup>104</sup> (or

weighted linear regression) provides a more robust estimate of the linear function (see Chapter 2). It should be noted that the mentioned transfer functions account only for analytical bias; however, adjustments for differences in imprecision may also be designed.

### Other Sources for Reference Limits

The reference limits used by clinical laboratories are often those provided by the manufacturers of diagnostic equipment. At the outset, it is important to note that, in some cases, manufacturers do not actually perform reference limit studies with their methods but instead cite other literature as the source for their reference limits (including earlier versions of this textbook). As stated earlier, some manufacturers' limits are central 90% rather than the usual central 95%. If, however, the manufacturer has indeed performed a good reference limit study using its method, and the laboratory uses the method exactly as prescribed by the manufacturer, it is reasonable to infer that the issue of method comparability has been addressed. Even after method comparability has been assured, though, it remains to be shown that the package insert data addresses the population comparability. Because supporting information is commonly not supplied in the manufacturers' Instructions for Use, it may be necessary to contact the manufacturer. The information required should include the age and sex distribution of the reference population, exclusion criteria, confidence intervals of the reference limits, and the statistical processes used. If relevant, additional information (e.g., ethnicity; lifestyle factors; body composition data, such as BMI or waist circumference) may be helpful. For example, the creatine kinase (CK) upper reference limits cited in a package insert based on a Caucasian population underestimated, by several-fold, the upper reference limits for blacks and Asians.<sup>105</sup>

A second source for reference limits is peer-reviewed publications. In this case, both method comparability and population comparability are at issue. Laboratories seeking to adopt these limits must proceed carefully, but it may well be considerably easier to address these issues than it would be to repeat the studies themselves. To return to the CK example just cited,<sup>105</sup> laboratories using the same method that was used in that publication, or other methods with demonstrated traceability to the same reference method (which is stated in the paper), could presumably adopt the reference limits determined in the study, whereas laboratories using other methods without this traceability would be aware of a potential problem but could not simply adopt those same reference limits for those ethnic groups. Another excellent example is the CALIPER initiative,<sup>53</sup> which established, using the methods described earlier, pediatric reference intervals, partitioned by age, sex, and ethnicity, for 40 different assays. As the authors emphasized, the published reference limits were specific to the methods they used. (In later publications<sup>106</sup> the authors provide reference limits for additional analytical methods based on transference studies alluded to in the previous section.) A third example, involving adult reference intervals, was structured along the same lines as the original CALIPER studies.<sup>107</sup>

A recent enhancement of this technique is reflected in the HAPPI Kids study in Australia, where a single set of collected

specimens was analyzed by multiple methods to generate reference intervals, thereby allowing adoption in a larger number of laboratories sooner.<sup>108</sup>

A third source for reference limits is multicenter trials. In contrast to the CALIPER initiative, in which all the analyses were done in a single laboratory, these studies seek to pool data from many laboratories spread over large regions and potentially among different countries. Although this decreases the number of reference individuals needed to be recruited from each laboratory, it also typically increases the number of analytical methods involved for each assay. Despite global efforts at harmonization and standardization, these multicenter trials have repeatedly shown that current methods do not always produce interchangeable results, and therefore methods to ensure method comparability are still needed. Nonetheless, once these issues have been addressed, multicenter trials are proving to be an excellent way to generate data for establishing reference limits.<sup>65,109</sup>

Another potential source of data to generate reference limits is a laboratory's own historical data. A laboratory's database is attractive on several counts: it encompasses the laboratory's own populations; it is generated with the laboratory's own methods and own preanalytical conditions; and it may include large quantities of data. The major limitation associated with this data is that it almost certainly includes data from unhealthy subjects and healthy subjects. However, as described previously, careful application of any of several indirect techniques has been used with data of this type to generate reliable reference limits.<sup>98,110</sup>

In summary, there are multiple different sources of information on reference intervals, including local formal reference interval studies, manufacturers' data, peer-reviewed publications, data mining techniques, and textbooks, including a compilation of reference interval data in various chapters and at the end of this textbook. Such information can be used as yet another source of data, with the same caveats noted at the beginning of this section and in the following section on verification. As is the case with any scientific process, it is good practice, when setting reference intervals, to access all available sources of information. This process allows for identification of discordant data sources and unexpected causes of variation, as well as for confirmation when different data sources provide the same information.

## POINTS TO REMEMBER

### **Sources for Reference Limits (Other Than Establishing One's Own)**

- Manufacturers' package inserts<sup>c</sup>
- Peer-reviewed literature<sup>c</sup>
- Multicenter trials<sup>c</sup>
- Analysis of laboratory's own historical data ("data mining")<sup>a</sup>

<sup>a</sup>Verification of the transferability of these values is *required*.

### **Verification of Transferability**

Whether a laboratory adopts reference values from (1) a manufacturer's package insert, (2) a peer-reviewed publication or

textbooks, (3) a multicenter trial, or (4) a data mining exercise, it is important that the laboratory verifies the appropriateness of those values for its own use.<sup>61</sup> This verification is the final check that the laboratory has implemented the analytical method correctly, and that the laboratory's own population is comparable with that used for the original reference value study.

Comparison of a locally produced, small subset of values with the large set produced elsewhere using traditional statistical tests often is not appropriate because the underlying statistical assumptions are not fulfilled and the sample sizes are unbalanced. Relatively sophisticated methods using non-parametric tests<sup>111</sup> or Monte Carlo sampling<sup>100,112</sup> have been described. In addition to providing its own recommendations for relatively sophisticated tests for verification, CLSI suggests a reasonably practical alternative that most laboratories should be able to adopt: with a sample size of 20 reference values, one verifies the appropriateness of a proposed reference interval so long as no more than two values are outside the proposed limits.<sup>16</sup> One obvious deficiency of this test is that it does not detect the situation in which the reference interval of the local group is narrower than that of the study group. Nonetheless, it does provide reasonable assurance that a proposed reference interval can be used. The use of a larger number of local samples can give greater assurance that the intervals are satisfactory and more chance to detect inappropriate intervals. If the data exists, local data mining techniques can also be used to verify, or at least assess, transferred reference intervals.<sup>113</sup> For example, the median or mode of an outpatient population should be close to the central point of the transferred reference interval, and the effect of the transferred interval on flagging rates, high and low, can be assessed.

## **PROCESS OF SELECTING REFERENCE INTERVALS**

Laboratories are responsible for, and usually take the lead in, providing reference limits for their test results. However, as indicated in the previous sections of this chapter, there are a large number of potentially complicated decisions involved in the process, and the final selection of reference limits will influence the decisions of many physicians about their patients. As a result, it is important for organizations to include individuals from outside the laboratory in the process of selecting reference limits for use. A multidisciplinary group should be involved in making decisions, such as whether to perform a reference interval study or to transfer intervals from another source; if the decision is made to do a local study, whether partitioning will be necessary (which will affect the number of subjects), and which preanalytical variables are relevant; and if limits will be transferred from another source, whether the methods and populations from that source are comparable. Some issues are more subjective: whether to set exclusion criteria in an effort to change sensitivity (e.g., exclude individuals who are obese or drink alcohol in order to increase sensitivity to detect the effects of these conditions); whether statistically significant differences in partitions are clinically significant; consideration of "rounding" the reference limits for ease of memory and to avoid an unwarranted impression of precision (e.g., whether the GGT upper limit calculated earlier could be rounded from 47 to 50 IU/L). As suggested earlier, even if a local reference

limit study is performed, it is appropriate to review all available data from the literature and from local data mining. When all data sources are concordant, there can be greater confidence in the limits. In contrast, when there is variability among the sources, it serves as a flag to assess possible reasons for the differences. There should be a written record of the people, decisions, and data used in selecting reference intervals, which will facilitate review and understanding in the future.

### Common Reference Intervals

As stated in the previous sections, there is a considerable amount of work involved in setting a reference interval. The usual concept of individual laboratories performing this work for all their tests is potentially overwhelming, even for large well-resourced laboratories. Additionally, when laboratories do determine their own, there are often variations in the final reference intervals that are not supported by analytical or population differences.<sup>114,115</sup> There is also an increasing need to reduce unnecessary variation among laboratories as patients move between health locations, data is combined in common databases, and patients view their results directly and wonder about small differences and the changes in classification that these may cause. One approach that is gaining in acceptance is to recommend common reference intervals for a country, region, or other grouping of laboratories. The aim is to gather an appropriately experienced group to assess all available data (as described in the previous section), including analytical and possibly population differences, and then to select reference intervals for participating laboratories. As well as the advantages of sharing the workload, better decisions may be made by involving a wider range of parties in the decisions. It is important to realize that common reference intervals are likely to be a little wider than single-site reference intervals, at least in part because of the greater analytical variation expected from a group of laboratories versus a single laboratory. For this reason, a single laboratory may observe that fewer than 5% of its patients are outside the traditional 95% limits. Approaches that have been used vary between sites, but work of this type has begun in Australia, Canada, and the Netherlands.<sup>102,110,116,117</sup>

## PRESENTATION OF AN OBSERVED VALUE IN RELATION TO REFERENCE VALUES

The purpose of reference limits is to allow a point of comparison for an observed value (patient's value). This comparison is similar to hypothesis testing, but it is seldom statistical testing in the strict sense. Ideally, the patient and the reference individuals should match (i.e., the hypothesis is stated that they were all picked from the same set [population]) in all aspects other than the presence of the medical condition for which the test has been requested. Often, however, this is not the case. Thus it is advisable to consider the reference values as the yardstick for a less formal assessment than hypothesis testing. It is this less formal assessment that typically directs the attention of the interpreting doctor to those results most likely to represent pathology. Indeed the results flagged as outside the reference intervals are commonly used as the starting point to begin interpretation of a report and to assess for possible major or urgent pathology.

The convenient presentation of an observed value in relation to reference values and clear flagging of results outside the intervals may be of great help for the busy clinician.<sup>6,10–15,26,118</sup> The most common presentation format for reference limits is the provision of the upper and lower reference limits on the same line of a report as the test name, test result (observed value), and units used for the result. Typically, when the result is outside the reference limits, the test result is accompanied by a "flag," which may be an asterisk, the letter H for "High" (i.e., above the upper reference limit), the letter L for "Low" (i.e., below the lower reference limit), or some other combinations of symbols. This format for reference limits has some significant strengths and weaknesses. The key strengths are the close proximity of the limits to the results, the ability to highlight values outside the reference limits, and the potential to tailor the displayed limits based on patient demographics typically available to the laboratory (such as age, sex, and fasting status). To the extent that additional factors about the patient are available (e.g., time since last drug dose or stage of pregnancy), even more specific limits can be supplied with the result. Limitations to this process include situations in which relevant factors are not available (e.g., stage of puberty or phase of the menstrual cycle). In these cases, a list of reference limits covering those factors can be included in the report, and the ordering clinician with knowledge of the patient can make the correct interpretation.

The preceding comments relate to reports in paper format and rendered electronic formats (e.g., portable document format [PDF]), situations in which the laboratory controls the way the data are presented. Recognizing that ease of reading pathology reports is a patient safety issue (a misread report can lead to a wrong medical decision; a report that is difficult to read will waste important clinical time), an Australian group has recommended strict formatting requirements for reports, including the placement and formatting of reference limits and flagging (e.g., columns of numerical results right-aligned, flagged with L or H to the right of results following a clear space, reference intervals in brackets to the right of results, and units of measure further to the right).<sup>119</sup>

A more difficult, and often under-appreciated, issue is the need to transfer reference limits along with patient results when they are communicated in other ways. For example, when individual patient values are transferred into third-party systems or common pathology databases, it is important to ensure that the laboratory-specific reference limits be as easily accessible as the patient values themselves. This issue is one driver toward developing common reference intervals.<sup>120</sup>

In addition, a very informative presentation of the observed value involves showing its location graphically relative to the reference limits. In the era of information overload, and review of pathology reports by patients and other non-medical personnel, a number of graphical tools are being developed to convey the information contained in a result and its relationship to reference limits.<sup>121</sup>

As stated earlier, results outside the reference limits may be flagged. A more detailed division of these results has been advocated to indicate how unusual the observed value is. For example, some laboratories highlight critical or extreme values with different flags such as C for critical, HH and LL for extremely high or low results respectively. In such cases, it is important for laboratories to educate their clinicians about

these flags because their use and meaning may vary considerably between laboratories and among tests. Similarly, subdivisions within the reference interval may also have clinical meaning and require communication to clinicians. For example, high-sensitivity C-reactive protein (hsCRP) results may be used not only for cardiovascular risk prediction but also as an indicator of acute inflammation.

Another method that can be used to express the observed value is by a *statistical distance measure*. All such distances are ratios of the following type:

$$\frac{\text{observed value} - \text{measure of location}}{\text{measure of dispersion}}$$

The *SD unit*, or normal equivalent deviate, is such a measure. It is calculated as the difference between the observed value and the mean of the reference values divided by their SD.<sup>122</sup> Several similar ratios under different names have been suggested and discussed.<sup>51,123–125</sup> This approach can be used to produce a graphical report for multiple analytes where the reference limits for different tests are aligned, and the observed values plotted as an SD unit against the reference limits. This allows a rapid assessment as to which results on a report are the “most abnormal.” One issue with these protocols is the need to develop methods for non-Gaussian distributions. Such methods include using transformed data<sup>126</sup> or using split data with different SDs for the upper and lower half of the reference population.<sup>127</sup>

A related concept is the use of *multiples of the median* (MoM), where observed values are expressed as a multiple of the median reference value. This is most commonly used in antenatal testing. By maintaining a current database comprising a large number of subjects unaffected by disease, this process seeks to reduce the effect of analytical bias by normalizing current reported values against current assay performance; the extent to which this is achieved has been questioned.<sup>128</sup>

Reporting the observed value as a *percentile* of the reference distribution provides a very accurate measure of the relation for results within the interval.<sup>26,129</sup> An observed serum GGT value of 48 IU/L may, for example, be reported as 48 IU/L (99th percentile). Results outside the reference interval will have very high or low percentiles and the accuracy of the assignment is likely to be limited. Alternatively, the probability of finding a value closer to the mean than the observed value, the *index of atypicality*, can be estimated.<sup>118,130</sup>

No matter how a reference interval is displayed, the clinician should always have access to as much information about the reference values as needed to use them appropriately. Reference values for all laboratory tests may be presented to clinicians in a booklet, web page, or other medium, together with information about (1) analytical methods, (2) their imprecision, and (3) descriptions of the reference values (e.g., whether they represent the central 90% rather than 95% of values) and any relevant limitations necessary for basic interpretation (e.g., diurnal variation). Graphical representations of the reference distributions can be informative (e.g., the use of a histogram [see Fig. 9.4A and C]), or a plot of the cumulative distribution [see Fig. 9.4B and D]), particularly with skewed distributions and tests with which the clinician is less familiar. If a reference limit is a clinical decision limit, it is important to make this distinction and to provide supporting documentation. The goal is to provide sufficient information to clinicians for them to make rational clinical judgments.

## ADDITIONAL TOPICS

### Multivariate, Population-Based Reference Regions

The topic of previous sections of this chapter has been univariate population-based reference values and quantities derived from them. However, such values do not fit the common clinical situation in which observed values of several different laboratory tests are available for interpretation and decision making. For example, the average number of individual clinical chemistry tests requested on each specimen received in the authors' laboratories is roughly 10; in many laboratories, this number is even larger. Two models are used for interpretation by comparison in this situation. Each observed value can be compared with the corresponding reference values or interval (i.e., a *multiple, univariate comparison is performed*), or the set of observed values can be considered as a single multivariate observation and can be interpreted as such by a *multivariate comparison*. In this section, the relative merits of these two approaches are discussed, and methods for the latter type of comparison are presented.

### Multivariate Concept

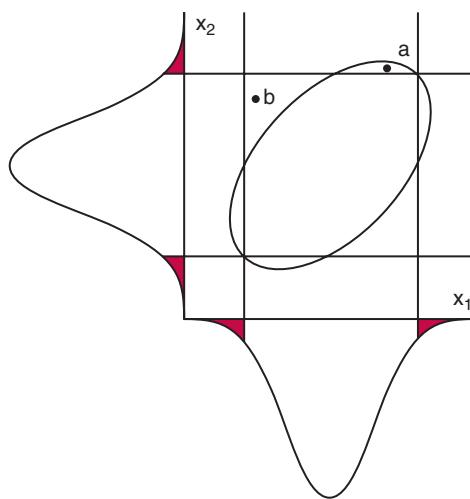
A univariate observation, such as a single laboratory result, may be represented graphically as a point on a line—the axis. Results obtained by two different laboratory tests performed on the same specimen (a bivariate observation) are then displayed as a point in a plane defined by two perpendicular axes. With three results, a trivariate observation and a point in a space are defined by three perpendicular axes, and so forth. The possibility of visualization of a multivariate observation is lost when there are more than three dimensions. Still, one can consider the multivariate observation as a point in a multidimensional hyperspace with as many mutually perpendicular axes as there are results of different tests. The prefix *hyper-* signifies, in this context, “more than three dimensions.” Such multivariate observations are also called *patterns* or *profiles*. A multivariate distribution thus is represented by a cluster of points on a plane, in a space, or in a hyperspace, depending on the dimensionality of the observation.<sup>76,126,131,132</sup> Several statistical methods are based on multivariate methods, some of which are straightforward extensions of well-known univariate methods.<sup>133</sup>

### Multiple, Univariate Reference Region

The univariate reference interval is bounded by two reference limits (lower and upper) on the result axis. Fig. 9.6 shows the univariate reference intervals for two laboratory tests: one depicted on the *x*-axis and the other on the *y*-axis.

Together, they describe a square in the plane of the two axes. Similarly, three or more univariate reference intervals define boxes or hyperboxes in the (hyper)space. By multiple, univariate comparison, it can be decided whether a multivariate observation point lies inside or outside this square, box, or hyperbox. However, this method has two very serious deficiencies<sup>134</sup>: an observation may lie outside the limits of the region without being unusual (see Fig. 9.6, point *a*), or it may be found on

the inside and still be an atypical observation (see Fig. 9.6, point *b*). If the central 95% interval is used, 5% of the values by definition are expected to be located in the two tails of the univariate reference distribution. However, more than 5% of the values would be located outside the square or (hyper)box created by several 95% intervals. To be exact,  $100(1 - 0.95^m)$



**FIGURE 9.6** Bivariate reference region (ellipse) compared with the region defined by the two univariate reference intervals (box).

percent of multivariate reference values would be excluded by the method of multiple, univariate comparison ( $m$  being the number of different tests, or the dimensionality). For example, provided the results are independent of each other, one would expect to find  $100(1 - 0.95^{10}) = 40\%$  of healthy subjects (members of the reference population) to have at least one result flagged as “abnormal” (i.e., 40% of healthy subjects will have false-positive results when 10 laboratory tests are performed). While this description is based on a multidimensional analysis, it is a model of standard laboratory practice. Additionally, an unusual combination (e.g., a low serum urea and high serum creatinine, each within its reference interval) would not be detected as abnormal, reflecting a limited sensitivity for abnormal patterns of results. This discouraging result has been verified in several multiphasic screening programs. Therefore a better method is needed.

### Multivariate Reference Region

It is possible to define a common multivariate reference region<sup>61,130–132,134,135</sup> on the basis of joint distribution of reference values for two or more laboratory tests. This multivariate region is not a right-angled area, or hyperbox, but is more like an ellipse in the plane (see Fig. 9.6) or an ellipsoid hyperbody in hyperspace. This region may be a straightforward extension of the univariate 95% interval to the multivariate situation; it may be set to enclose 95% of central multivariate reference data points. In this case, one would expect to find only 5% false positives.

The use of multivariate reference regions usually requires the assistance of a computer program, which takes a set of results obtained by several laboratory tests on the same clinical specimen and calculates an index. Interpretation of a multivariate observation in relation to reference values is then the task of comparing the index with a threshold value estimated from the reference values. Obviously, this is much simpler than comparing each result with its proper reference interval.

This index is essentially a distance measure and is known as *Mahalanobis' squared distance* ( $D^2$ ). It is analogous to the square of the SD for single reference values. It expresses the multivariate distance between the observation point and the common mean of the reference values, taking into account

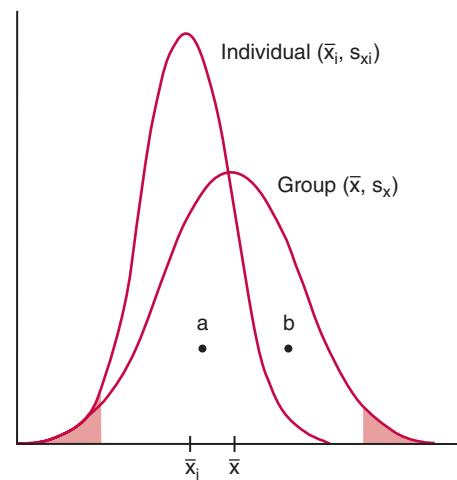
the dispersion and the correlation of the variables.<sup>61,130–132,134,135</sup> More interpretational guidance may be obtained from this distance by expressing it as a percentile analogous to the percentile presentation of univariate observed values.<sup>135</sup> Also, the index of atypicality has a multivariate counterpart.<sup>130,131</sup>

Although the theory of multivariate reference regions has been known for a while, surprisingly few applications of it have been reported in the literature. Some recent examples are two- and three-dimensional regions for thyroid function tests<sup>136</sup> and an outcome study in intensive care showing improved prediction with result pairs than with individual results.<sup>137</sup> An important report reviews the topic and presents the results of a very careful study on the multivariate 95% region for a 20-test chemistry profile.<sup>135</sup> Some of the most important findings can be summarized as follows:

1. Sixty-eight percent of subjects had at least one test result outside univariate reference intervals, which was close to what was theoretically expected:  $100(1 - 0.95^{20}) = 64\%$ .
2. By contrast, only 5% of patterns were outside the multivariate reference region (as expected).
3. Transformation to approximately Gaussian shape of the univariate distributions was necessary.
4. A test profile may be distinctly unusual in the multivariate sense even though each individual result is within its proper reference interval (e.g., see point *b* in Fig. 9.6).
5. The multivariate reference region could detect minor deviations of multiple analytes.
6. Conversely, it could also be insensitive to highly deviating results for a single analyte.
7. Sensitivity could be increased by defining multivariate reference regions for subsets of physiologically related tests.

### Subject-Based Reference Values

Fig. 9.7 depicts the inherent problem associated with population-based reference values. It shows two hypothetical reference distributions. One represents the common reference distribution based on single specimens obtained from a group of



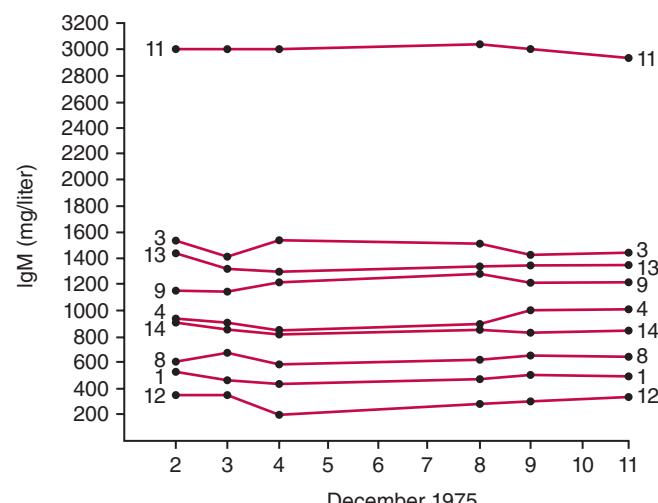
**FIGURE 9.7** Relationship between population- and subject-based reference distributions and reference intervals. The example is hypothetical, and the two distributions are, for simplicity, Gaussian. Note that both points *a* and *b* are within the population-based reference interval, but only point *a* would be “normal” for this particular subject. (Modified from Harris EK. Effects of intraindividual and interindividual variation on the appropriate use of normal ranges. *Clin Chem* 1974;20:1536.)

different reference individuals. It has a true (hypothetical) mean  $\bar{x}$  and a SD  $s_x$ . The other distribution is based on several specimens collected over time in a single individual, the  $i$ th individual. Its hypothetical mean is  $\bar{x}_i$  and its SD  $s_{xi}$ .

If an observed value is located outside the subject's 2.5th and 97.5th percentiles, the personal or *subject-based reference interval*, the cause may be a change in biochemical status, suggesting the presence of disease. Fig. 9.7 shows that such an observed value may still be within the population-based reference interval. The sensitivity of the latter interval to changes in a subject's biochemical status depends accordingly on the location of the individual's mean  $x_i$  relative to the common mean  $x$  and to the relative magnitudes of the corresponding SDs  $s_{xi}$  and  $s_x$ . A mean  $s_{xi}$  close to  $s_x$  and a small  $s_{xi}$  relative to  $s_x$  may conceal the individual's changes entirely within the population-based reference interval.

Harris<sup>73,74</sup> analyzed this topic and found that the ratio  $R$  of intraindividual (personal) variation over interindividual (among subjects) variation provides a criterion for the usefulness of the population-based reference interval. This is now known as the Index of Individuality (II). The population-based reference interval has less than the desired sensitivity to changes in biochemical status if II is  $\leq 0.6$ . The interval is a more trustworthy reference if II is greater than 1.4, at least for the individual whose SD  $\sigma_i$  is close to the average value. Published data<sup>73,138</sup> usually show that homeostatically tightly controlled quantities, such as serum electrolytes, have high ratio values. Population-based reference intervals of such analytes suffice for clinical use. In contrast, serum proteins and enzymes have very low ratios because they are not under the same degree of metabolic control. Here, subject-based reference intervals seem more appropriate, although limitations to the meaning of a low value for the II have been raised.<sup>139</sup>

Two specific examples mentioned earlier may help to clarify this concept further. Fig. 9.8 depicts immunoglobulin (Ig)M values from several healthy individuals over the course



**FIGURE 9.8** Serial immunoglobulin (Ig)M values over several days from reference individuals. Note that the intra-individual variability is very small compared to the inter-individual variability. (From Statland BE, Winkel P, Killingsworth LM. Factors contributing to intra-individual variation of serum constituents: physiological day-to-day variation in concentrations of 10 specific proteins in sera. *Clin Chem* 1976;22:1635–8.)

of several days. As illustrated, the intra-individual differences are small compared with interindividual differences. Even though the population-based reference interval might extend from 200 to 1600 mg/dL, it would be most unusual (abnormal) for any patient's IgM value to change by more than 200 mg/dL, even if the value remained within the population-based reference interval. Similarly, it is well known that any given patient's serum creatinine value is reasonably constant,<sup>24</sup> which is related both to glomerular filtration rate (GFR) and to lean muscle mass. If the latter is constant, then changes in GFR are inversely proportional to the serum creatinine (see Chapter 34). That is, even though a typical (population-based) reference interval for serum creatinine might extend from 62 to 106 µmol/L (0.7 to 1.2 mg/dL), a change from 65 to 105 µmol/L in a given patient would be distinctly abnormal, representing the loss of almost half of the GFR, a finding of great clinical importance.

Two solutions can be proposed to the problem of the clinical insensitivity of population-based reference intervals:

1. One can try to reduce variation in reference values by *partitioning* into more homogeneous subclasses, as was discussed in a previous section. However, increasing the index of individuality (II), for example, from 0.6 to 1.4 by partitioning requires that one can obtain the rather dramatic reduction of 37% in SD.<sup>74</sup> This is often difficult to attain in practice.
2. The other possibility is to use the patient's previous values, obtained when the patient was in a defined state of health, as the reference for any future value. Application of *subject-based reference values* becomes more feasible as health screening by laboratory tests and computer storage of results become available to large segments of the general population.

The current approach to adopting the second solution above is to use one or more previous results as the best estimate of the patient's previous state, and then assess whether a subsequent result is likely to be different from that baseline. There are many similarities in this approach with the descriptions of population reference intervals previously described in this chapter. The amount of change in analyte concentration which is considered significant, known as the RCV, although previously referred to as the *critical difference*, is commonly set at a 5% probability of the change being due to random effects. Just as this mirrors the 5% probability of a member of the reference population being outside standard reference limits, it is possible to set RCVs based on different probabilities. The RCV is a combination of the analytical variation, the within-subject biological variation ( $CV_i$ ), and the desired probability. For population reference intervals these also include the between-subject variation. It is recommended that values for  $CV_i$  are determined in formal studies with reference individuals under controlled conditions,<sup>140</sup> but methods have been developed to determine values for  $CV_i$  from laboratory databases,<sup>141</sup> these two approaches mirroring the direct and indirect approaches of population reference intervals. More information on biological variation, RCVs, and the relationship with population reference intervals is found in Chapter 8.

### When "Normal" May Appear to Be "Abnormal" and Vice Versa

Some, if not many, clinical laboratory test results are related to other clinical laboratory test results in a way that affects

their interpretation. For example, consider the relatively common situation of a patient with a low serum albumin. Because of the physiologic relationship between total calcium and albumin in serum (see Chapter 54), a total serum calcium concentration in the healthy reference interval might actually be pathologically high in this patient, and a total serum calcium concentration below the lower reference limit might be healthy. (For this reason, clinicians may choose to calculate an “adjusted calcium concentration” or to measure the “free calcium” concentration in these patients.) In these types of situations, it is important not to consider the normal (or abnormal) test results out of context. In most cases, laboratory reports, including reference intervals, do not take these situations into account.

As another example, consider a pregnant woman with typically high concentrations of serum binding proteins, including thyroxine binding globulin. She might well have what appears to be an “abnormally” high concentration of serum total thyroxine when compared with conventional reference limits (see Chapter 57), when in fact this is typical of a healthy pregnant state. Similarly, high concentrations of ceruloplasmin in pregnancy and other high estrogen states such as the oral contraceptive pill can produce high concentrations of serum copper, which does not indicate any change in copper metabolism. Other such examples in clinical medicine abound. Consider the prostate specific antigen (PSA) level in a postprostatectomy patient or the thyroglobulin level in a post-thyroidectomy patient. In these cases, it would be “healthy” to have abnormally low (undetectable) concentrations of these measurands, and it would be distinctly “unhealthy” to have levels in the healthy reference interval. Another scenario is urine sodium in investigation of acute kidney injury. If the kidney is “healthy” (i.e., with prerenal causes of acute kidney injury [AKI]), the urine sodium will be lower than the population reference interval, and a “healthy” result is consistent with intrinsic damage to the kidney. In all of these cases, the problem is that the traditional reference population (nonpregnant, healthy individuals) is not appropriate for the specific clinical setting.

Indeed, interpretations in much of the field of endocrinology are based not so much on “healthy” concentrations but rather on “appropriate” concentrations. For example, in a patient with a very high free T4 concentration, a thyroid stimulating hormone (TSH) within the traditional reference limits is used for differential diagnosis rather than likelihood of health. It should be undetectable in primary hyperthyroidism; otherwise, it may well represent secondary or tertiary hyperthyroidism (see Chapter 57). Similarly, in a patient with hypercalcemia, a PTH within the traditional healthy reference limits is distinctly abnormal.

In other words, even when reference limit studies are done well, one needs to remember there are dependencies that can render those reference limits misleading.

### Special Populations

As noted in the previous section, there are groups of patients for whom the typical populations used in establishing reference limits may not be appropriate. Such groups include, but are not limited to, children, pregnant women, and the elderly. Even within these groups there may be important subgroups (partitions): for example, children may need to be further divided by age, sex, ethnicity, and/or Tanner stage; pregnant

women, by trimester or week; elderly, by age, sex, and ethnicity. Over the past decade a number of studies have been conducted to establish reference intervals for these populations.

As noted earlier, the CALIPER initiative, a multicenter trial, recruited several thousand healthy pediatric subjects from across Canada and established reference limits, partitioned by age, sex, and ethnicity, for 40 measurands using one analytical system.<sup>53</sup> Its database was later extended by transference to several additional analytical systems and to an additional 29 measurands.<sup>106</sup> A recent study from Denmark established reference limits for 36 measurands based on 801 normal pregnancies in Caucasian women.<sup>142</sup> These limits might be extended to other analytical systems by transference, but additional pregnancies in non-Caucasian women will be required to determine whether the limits can be extended to other ethnic groups. Another study from Denmark established reference limits for 27 measurands based on 1016 70-year-old Caucasians.<sup>143</sup> Again, this involved a single set of analytical systems and a single ethnic group, so the authors were careful to point out that additional work will be required to determine to what extent their reference limits can be adopted by others.

In each of these cases, important differences from the traditional reference intervals were uncovered.

### Special Cases of Laboratory Test Result Interpretation

In the search for improved diagnostic performance, several individual test results may be combined to form an index or score with the aim of combining the discriminating power of the individual results. Examples include the “triple” and “quad” screens used to calculate risk for Down syndrome and trisomy 18 in early pregnancy (see Chapter 59), a variety of proprietary “liver fibrosis” screens<sup>144</sup> to estimate the likelihood of cirrhosis (see Chapter 51), and the prostate health index from Beckman Coulter.<sup>145</sup> Although a discussion of these methods is beyond the scope of this chapter, the basic concepts of reference populations, sampling, outlier exclusion, and so on remain vital to the process. Typically, the reference limit is a clinical decision point rather than a population reference interval limit. In these cases, the goal is not so much to determine whether the patient is healthy but to determine whether the likelihood of abnormality is high enough that additional testing is warranted (triple and quad screens) or to determine whether certain therapies are likely to be helpful without resorting to an invasive liver<sup>144</sup> or prostate biopsy.<sup>145</sup> More common examples are calculated tests such as the anion gap, osmolar gap, serum globulins, and calculated free testosterone. In each case there is a tendency for reference intervals to be derived from textbooks without taking the test methodology into account. Laboratories need to consider and promote accurate reference intervals for these tests with the same care as for individual tests.

### Ongoing Verification of Reference Limits

Whether a laboratory establishes values with its own reference limit studies or adopts reference limits after verification from another source, it is rare for any laboratory to assess those reference limits on a regular basis. Typically, the reference limits are not reevaluated until the laboratory implements new instrumentation, at which point it may be necessary to make changes. Even with changes in methods, though, there is no guarantee that a reassessment will occur. In an interesting

report from Australia,<sup>102</sup> studies showed that reference limits for common measurands differed significantly among laboratories, even when they used the same analytical systems; additional studies indicated that common reference intervals for these measurements could, and should, be implemented across all these laboratories.

Laboratories can, however, use “data mining” techniques discussed earlier to perform ongoing audits of reference limits and to investigate specific concerns when they arise. For example, with these techniques, a laboratory could determine the median of the distribution of bicarbonate values from its healthy outpatients, which should be extremely stable.<sup>113,146</sup> Similarly, it could determine what percentage of these values fall outside the reference limits. Assuming those limits were set to include the central 95% of healthy outpatient values, no more than 2.5% of the values should be below the lower limit or above the upper limit. If a change in median occurs, or if too many samples fall outside the limits, there may well be a problem. The laboratory could then investigate further by using the technique to see when the problem started, by reviewing whether any changes in methods were made, by investigating performance on proficiency testing, and even by repeating a short reference interval verification study with 20 individuals as described earlier. These data mining studies are relatively inexpensive, involving only retrieval and manipulation of data that already exists, but they provide great reassurance to the laboratory and its users of the accuracy of its test results and reference limits.

## CONCLUSION

---

In this chapter, we have emphasized the importance of reference limits, which provide physicians with values with which they can compare their patients’ results, thereby facilitating

interpretation. In most, but not all, cases, reference limits reflect values seen in healthy individuals.

When generating, verifying, or reporting reference limits, it is critical that the populations be well-defined and that the patients for whom the limits are used be comparable in terms of gender, age, and ethnicity where relevant, and any other measurable characteristics determined to be of importance. We have described in detail ways to ensure collection of high-quality data, methods to eliminate outliers when appropriate, and techniques to analyze the data (including nonparametric [preferred], parametric, graphical, computerized).

Recognizing that many laboratories do not have the resources to generate their own reference limits, we have described alternative sources (including manufacturers’ package inserts, peer-reviewed literature, multicenter trials, historical laboratory data) and techniques to verify the transferability of such data.

We have recommended a multidisciplinary approach to selecting and implementing reference limits, as well as support for the concept of common reference intervals amongst groups of laboratories in a region or country, and we have discussed a number of considerations related to the display of reference limits along with patient results.

We have also included discussion of several special topics, including multivariate reference limits, subject-based reference limits, reference limits for special populations, and ongoing verification of reference limits.

It is quite clear that laboratories face many issues and a great deal of effort in ensuring that their reference limits are valid, but we would argue that, in the absence of this effort, much of the data we generate would not be interpretable. The validity of our reference limits is at least as important as the accuracy and precision of our analytical techniques and should therefore warrant at least as much attention.

## SELECTED REFERENCES

3. Siest G, Henny J, Grasbeck R, et al. The theory of reference values: an unfinished symphony. *Clin Chem Lab Med* 2013;51:47-64.
10. Solberg HE. IFCC approved recommendation on the theory of reference values. Part 1. The concept of reference values. *J Clin Chem Clin Biochem* 1987;25:337-42.
16. Clinical And Laboratory Standards Institute: Defining, establishing, and verifying reference intervals in the clinical laboratory (EP28-A3c). Wayne, PA: Clinical and Laboratory Standards Institute; 2010.
20. Ozarda Y, Sikaris K, Streichert T, Macri J. Distinguishing reference intervals and clinical decision limits - A review by the IFCC Committee on Reference Intervals and Decision Limits. *Critical reviews in clinical laboratory sciences* 2018;55:420-31.
37. Jones GRD, Haeckel R, Loh TP, et al. Indirect methods for reference interval determination - review and recommendations. *Clin Chem Lab Med* 2018;57:20-9.
50. Ichihara K, Ozarda Y, Barth JH, et al. A global multicenter study on reference values: 1. Assessment of methods for derivation and comparison of reference intervals. *Clin Chim Acta* 2017;467:70-82.
53. Colantonio DA, Kyriakopoulou L, Chan MK, 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* 2012;58:854-68.
60. Ichihara K, Boyd JC. An appraisal of statistical procedures used in derivation of reference intervals. *Clin Chem Lab Med* 2010;48:1537-51.
61. Harris EK, Boyd JC. Statistical bases of reference values in laboratory medicine. New York: Marcel Dekker; 1995.
65. Rustad P, Felding P, Franzson L, et al. The Nordic Reference Interval Project 2000: recommended reference intervals for 25 common biochemical properties. *Scand J Clin Lab Invest* 2004;64:271-84.
86. Pavlov IY, Wilson AR, Delgado JC. Reference interval computation: which method (not) to choose? *Clin Chim Acta* 2012;413:1107-14.
90. Horn PS, Pesce AJ. Reference intervals: a user's guide. Washington, DC: AACC Press; 2005.
92. Katayev A, Fleming JK, Luo D, Fisher AH, Sharp TM. Reference intervals data mining: no longer a probability paper method. *Am J Clin Pathol* 2015;143:134-42.
95. Holmes DT, Buhr KA. Widespread Incorrect Implementation of the Hoffmann Method, the Correct Approach, and Modern Alternatives. *Am J Clin Pathol* 2019;151:328-36.
102. Tate JR, Sikaris KA, Jones GR, 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* 2014;35:213-35.
106. Karbasy K, Ariadne P, Gaglione S, Nieuwsteeg M, Adeli K. Advances in Pediatric Reference Intervals for Biochemical Markers: Establishment of the Caliper Database in Healthy Children and Adolescents. *J Med Biochem* 2015;34:23-30.
113. Jones GR. Validating common reference intervals in routine laboratories. *Clin Chim Acta* 2014;432:119-21.
117. Koerbin G, Sikaris K, Jones GRD, Flatman R, Tate JR. An update report on the harmonization of adult reference intervals in Australasia. *Clin Chem Lab Med* 2018;57:38-41.
119. Flatman R, Legg M, Jones GR, Graham P, Moore D, Tate J. Recommendations for reporting and flagging of reference limits on pathology reports. *Clin Biochem Rev* 2014;35:199-202.
121. O'Connor JD. Reducing post analytical error: perspectives on new formats for the blood sciences pathology report. *Clin Biochem Rev* 2015;36:7-20.

## REFERENCES

1. ISO15189 medical laboratories—requirements for quality and competence. 2015. (Accessed February 20, 2020, at [http://www.iso.org/iso/catalogue\\_detail?csnumber=56115](http://www.iso.org/iso/catalogue_detail?csnumber=56115).)
2. College of American Pathologists Lab General checklist item 41096. 2019.
3. Siest G, Henny J, Grasbeck R, et al. The theory of reference values: an unfinished symphony. *Clin Chem Lab Med* 2013; 51:47-64.
4. Murphy EA. The normal, and the perils of the sylleptic argument. *Perspect Biol Med* 1972;15:566-82.
5. Elveback LR, Guillier CL, Keating FR, Jr. Health, normality, and the ghost of Gauss. *JAMA* 1970;211:69-75.
6. Grasbeck R, Alstrom T. Reference values in laboratory medicine. Chichester, UK: John Wiley; 1981.
7. Solberg HE, Grasbeck R. Reference values. *Adv Clin Chem* 1989;27:1-79.
8. Sunderman FW, Jr. Current concepts of "normal values," "reference values," and "discrimination values," in clinical chemistry. *Clin Chem* 1975;21:1873-7.
9. Kendall MG, Buckland WR. A dictionary of statistical terms. 5th ed. London, UK: Longman; 1990.
10. Solberg HE. IFCC approved recommendation on the theory of reference values. Part 1. The concept of reference values. *J Clin Chem Clin Biochem* 1987;25:337-42.
11. PetitClerc C, Solberg HE. IFCC approved recommendation on the theory of reference values. Part 2. Selection of individuals for the production of reference values. *J Clin Chem Clin Biochem* 1987;25r:639-44.
12. Solberg HE, PetitClerc C. IFCC approved recommendation 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* 1988;26:593-8.
13. Solberg HE, Stamm D. IFCC approved recommendation on the theory of reference values. Part 4. Control of analytical variation in the production, transfer and application of reference values. *Eur J Clin Chem Clin Biochem* 1991;29:531-5.
14. Solberg HE. The theory of reference values Part 5. Statistical treatment of collected reference values. Determination of reference limits. *J Clin Chem Clin Biochem* 1983;21:749-60.
15. Dybkaer R. IFCC approved recommendation on the theory of reference values. Part 6. Presentation of observed values related to reference values. *J Clin Chem Clin Biochem* 1987;25:657-62.
16. Clinical And Laboratory Standards Institute: Defining, establishing, and verifying reference intervals in the clinical laboratory (EP28-A3c). Wayne, PA: Clinical and Laboratory Standards Institute; 2010.
17. Joint Committee for Guides in Metrology. JCGM 200: International vocabulary of metrology: basic and general concepts and associated terms. (Accessed February 20, 2020, at [http://www.bipm.org/utils/common/documents/jcgm/JCGM\\_200\\_2012.pdf](http://www.bipm.org/utils/common/documents/jcgm/JCGM_200_2012.pdf))
18. Pincus MR. Interpreting laboratory results: reference values and decision making. In: Henry JB, ed. *Clinical Diagnosis and management by laboratory methods*. Philadelphia, PA: WB Saunders; 1996:74-91.
19. Statland BE. Clinical decision levels for lab tests. Oradell, NJ: Medical Economics Books; 1987.
20. Ozarda Y, Sikaris K, Streichert T, Macri J. Distinguishing reference intervals and clinical decision limits - A review by the IFCC Committee on Reference Intervals and Decision Limits. *Critical reviews in clinical laboratory sciences* 2018;55:420-31.
21. Stone NJ, Robinson JG, Lichtenstein AH, 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* 2014;129:S1-45.
22. American Diabetes Association: Classification and diagnosis of diabetes. *Diabetes care* 2020;43(Supplement 1):S14-31.
23. Management of hyperbilirubinemia in the newborn infant 35 or more weeks of gestation. *Pediatrics* 2004;114:297-316.
24. Fraser CG. Biological variation: from principles to practice. Washington, DC: AACC Press; 2001.
25. Statland BE, Winkel P, Killingsworth LM. Factors contributing to intra-individual variation of serum constituents: Physiological day-to-day variation in concentrations of 10 specific proteins in sera of healthy subjects. *Clin Chem* 1976;22:1635-8.
26. Dybkaer R. Observed value related to reference values. In: Grasbeck R, Alstrom T, eds. *Reference values in laboratory medicine*. Chichester, UK: John Wiley; 1981:263-78.
27. Januzzi JL, Jr., Camargo CA, Anwaruddin S, et al. The N-terminal Pro-BNP investigation of dyspnea in the emergency department (PRIDE) study. *Am J Cardiol* 2005;95:948-54.
28. Lainchbury JG, Campbell E, Frampton CM, Yandle TG, Nicholls MG, Richards AM. Brain natriuretic peptide and n-terminal brain natriuretic peptide in the diagnosis of heart failure in patients with acute shortness of breath. *J Am Coll Cardiol* 2003;42:728-35.
29. Maisel AS, Krishnaswamy P, Nowak RM, et al. Rapid measurement of B-type natriuretic peptide in the emergency diagnosis of heart failure. *N Engl J Med* 2002;347:161-7.
30. Bjerner J. Age-dependent biochemical quantities: an approach for calculating reference intervals. *Scand J Clin Lab Invest* 2007;67:707-22.
31. Grasbeck R. Health as seen from the laboratory. In: Grasbeck R, Alstrom T, eds. *Reference values in laboratory medicine*. Chichester, UK: John Wiley; 1981:17-24.
32. Koerbin G, Tate JR, Hickman PE. Analytical characteristics of the Roche highly sensitive troponin T assay and its application to a cardio-healthy population. *Ann Clin Biochem* 2010;47:524-8.
33. Sandoval Y, Apple FS. The global need to define normality: the 99th percentile value of cardiac troponin. *Clin Chem* 2014;60:455-62.
34. Baadenhuijsen H, Smit JC. Indirect estimation of clinical chemical reference intervals from total hospital patient data: application of a modified Bhattacharya procedure. *J Clin Chem Clin Biochem* 1985;23:829-39.
35. Hoffmann RG. Statistics in the practice of medicine. *JAMA* 1963;185:864-73.
36. Bhattacharya CG. A simple method of resolution of a distribution into gaussian components. *Biometrics* 1967;23:115-35.
37. Jones GRD, Haeckel R, Loh TP, et al. Indirect methods for reference interval determination - review and recommendations. *Clin Chem Lab Med* 2018;57:20-9.
38. Arzideh F, Wosniok W, Gurr E, 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* 2007;45:1043-57.
39. Kouri T, Kairisto V, Virtanen A, et al. Reference intervals developed from data for hospitalized patients: computerized method based on combination of laboratory and diagnostic data. *Clin Chem* 1994;40:2209-15.

40. Solberg HE. Using a hospitalized population to establish reference intervals: pros and cons. *Clin Chem* 1994;40:2205-6.
41. Siest G, Henny J, Schiele F, Young DS. Interpretation of clinical laboratory tests: reference values and their biological variation. Foster City, CA: Biomedical Publications; 1985.
42. Ozcurumez MK, Haeckel R. Biological variables influencing the estimation of reference limits. *Scand J Clin Lab Invest* 2018;78:337-45.
43. Sikaris K, McLachlan RI, Kazlauskas R, de Kretser D, Holden CA, Handelsman DJ. Reproductive hormone reference intervals for healthy fertile young men: evaluation of automated platform assays. *J Clin Endocrinol Metab* 2005;90:5928-36.
44. Berg B, Nilsson JE, Solberg HE. Practical experience in the selection and preparation of reference individuals: empiric testing of the provisional Scandinavian recommendations. In: Grasbeck R, Alstrom T, eds. Reference values in laboratory medicine. Chichester, UK: John Wiley; 1981:55-64.
45. Harris EK. Statistical aspects of reference values in clinical pathology. In: Stefani M, Benson ES, eds. Progress in clinical pathology. New York: Grune & Stratton; 1981:45-66.
46. Centers for Disease Control and Prevention: National health and nutrition examination survey. 2020. (Accessed February 20, 2020, at [http://www.cdc.gov/nchs/nhanes/about\\_nhanes.htm](http://www.cdc.gov/nchs/nhanes/about_nhanes.htm).)
47. Adeli K, Higgins V, Nieuwesteeg M, 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* 2015;61:1049-62.
48. Australian Bureau of Statistics: Australian health survey. 2020. (Accessed February 20, 2020, at <http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/74D87E30B3539C53CA257BBB0014BB36?opendocument>.)
49. Centers for Disease Control and Prevention: About BMI for adults. 2020. (Accessed March 5, 2020, at [http://www.cdc.gov/healtyweight/assessing/bmi/adult\\_bmi](http://www.cdc.gov/healtyweight/assessing/bmi/adult_bmi).)
50. Ichihara K, Ozarda Y, Barth JH, et al. A global multicenter study on reference values: 1. Assessment of methods for derivation and comparison of reference intervals. *Clin Chim Acta* 2017;467:70-82.
51. Faulkner WR, Meites S. Geriatric clinical chemistry: reference values. Washington, DC: AACC Press; 1997.
52. Nilsson SE, Evrin PE, Tryding N, Berg S, McClearn G, Johansson B. Biochemical values in persons older than 82 years of age: report from a population-based study of twins. *Scand J Clin Lab Invest* 2003;63:1-13.
53. Colantonio DA, Kyriakopoulou L, Chan MK, 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* 2012;58:854-68.
54. Alstrom T, Grasbeck R, Lindblad B, Solberg HE, Winkel P, Viinikka L. Establishing reference values from adults: recommendation on procedures for the preparation of individuals, collection of blood, and handling and storage of specimens. Committee on Reference Values of the Scandinavian Society for Clinical Chemistry. *Scand J Clin Lab Invest* 1993;53:649-52.
55. Felding P, Tryding N, Hyltoft Petersen P, Horder M. Effects of posture on concentrations of blood constituents in healthy adults: practical application of blood specimen collection procedures recommended by the Scandinavian Committee on Reference Values. *Scand J Clin Lab Invest* 1980;40:615-21.
56. Tryding N, Tufvesson C, Sonntag O. Drug effects in clinical chemistry. Stockholm: Apoteksbolaget; 1996.
57. Young DS. Effects of drugs on laboratory tests. Washington, DC: AACC Press; 2000.
58. Roelfsema F, Veldhuis JD. Thyrotropin secretion patterns in health and disease. *Endocr Rev* 2013;34:619-57.
59. Armbruster D, Miller RR. The Joint Committee for Traceability in Laboratory Medicine (JCTLM): a global approach to promote the standardisation of clinical laboratory test results. *Clin Biochem Rev* 2007;28:105-13.
60. Ichihara K, Boyd JC. An appraisal of statistical procedures used in derivation of reference intervals. *Clin Chem Lab Med* 2010; 48:1537-51.
61. Harris EK, Boyd JC. Statistical bases of reference values in laboratory medicine. New York: Marcel Dekker; 1995.
62. Thygesen K, Alpert JS, Jaffe AS, et al. Third universal definition of myocardial infarction. *Circulation* 2012;126:2020-35.
63. Nordestgaard BG, Chapman MJ, Ray K, et al. Lipoprotein(a) as a cardiovascular risk factor: current status. *Eur Heart J* 2010; 31:2844-53.
64. Linnet K. Two-stage transformation systems for normalization of reference distributions evaluated. *Clin Chem* 1987;33:381-6.
65. Rustad P, Felding P, Franzson L, et al. The Nordic Reference Interval Project 2000: recommended reference intervals for 25 common biochemical properties. *Scand J Clin Lab Invest* 2004; 64:271-84.
66. Lahti A, Petersen PH, Boyd JC, Rustad P, Laake P, Solberg HE. Partitioning of nongaussian-distributed biochemical reference data into subgroups. *Clin Chem* 2004;50:891-900.
67. Sachs L. Applied statistics: a handbook of techniques. New York: Springer-Verlag; 1982.
68. Snedecor GW, Cochran WG. Statistical methods. 8th ed. Ames, IA: Iowa State University Press; 1989.
69. Bradley JV. Distribution-free statistics. Englewood Cliffs, NJ: Prentice-Hall; 1968.
70. Lahti A, Hyltoft Petersen P, Boyd JC, Fraser CG, Jorgensen N. Objective criteria for partitioning Gaussian-distributed reference values into subgroups. *Clin Chem* 2002;48:338-52.
71. Virtanen A, Kairisto V, Uusipaikka E. Regression-based reference limits: determination of sufficient sample size. *Clin Chem* 1998;44:2353-8.
72. Zierk J, Arzideh F, Haeckel R, Rascher W, Rauh M, Metzler M. Indirect determination of pediatric blood count reference intervals. *Clin Chem Lab Med* 2013;51:863-72.
73. Harris EK. Effects of intra- and interindividual variation on the appropriate use of normal ranges. *Clin Chem* 1974;20:1535-42.
74. Harris EK. Some theory of reference values. I. Stratified (categorized) normal ranges and a method for following an individual's clinical laboratory values. *Clin Chem* 1975;21:1457-64.
75. Clinical Laboratory and Standards Institute: Harmonized terminology database. (Accessed March 5, 2020, at [htd.clsi.org](http://htd.clsi.org).)
76. Barnett V, Lewis T. Outliers in statistical data. Chichester, UK: John Wiley; 1994.
77. Hawkins DM. Identification of outliers. London: Chapman and Hall; 1980.
78. Reed AH, Henry RJ, Mason WB. Influence of statistical method used on the resulting estimate of normal range. *Clin Chem* 1971;17:275-84.
79. Healy MJ. Outliers in clinical chemistry quality-control schemes. *Clin Chem* 1979;25:675-7.
80. Horn PS, Feng L, Li Y, Pesce AJ. Effect of outliers and non-healthy individuals on reference interval estimation. *Clin Chem* 2001;47:2137-45.

81. Box GEP, Cox DR. Analysis of transformations. *J R Stat Soc* 1964;B26:211-52.
82. Linnet K. Nonparametric estimation of reference intervals by simple and bootstrap-based procedures. *Clin Chem* 2000;46:867-9.
83. CBStat: a program for statistical analysis in clinical biochemistry. (Accessed February 20, 2020, at <http://www.cbstat.com>.)
84. MedCalc: easy to use statistical software. (Accessed February 20, 2020, at <http://www.medcalc.org>.)
85. Analyse-it: method validation edition. 2020. (Accessed February 20, 2020, at <http://analyse-it.com/products/method-validation/>.)
86. Pavlov IY, Wilson AR, Delgado JC. Reference interval computation: which method (not) to choose? *Clin Chim Acta* 2012;413:1107-14.
87. Mardia KV. Tests of univariate and multivariate normality. In: Krishnaiah PR, ed. *Handbook of statistics: analysis of variance*. Amsterdam: North-Holland Publishing; 1980:279-320.
88. Solberg HE. Statistical treatment of reference values in laboratory medicine: testing the goodness-of-fit of an observed distribution to the Gaussian distribution. *Scand J Clin Lab Invest Suppl* 1986;184:125-32.
89. Shultz EK, Willard KE, Rich SS, Connelly DP, Critchfield GC. Improved reference-interval estimation. *Clin Chem* 1985;31:1974-8.
90. Horn PS, Pesce AJ. Reference intervals: a user's guide. Washington, DC: AACC Press; 2005.
91. Gindler EM. Calculation of normal ranges by methods used for resolution of overlapping Gaussian distributions. *Clin Chem* 1970;16:124-8.
92. Katayev A, Fleming JK, Luo D, Fisher AH, Sharp TM. Reference intervals data mining: no longer a probability paper method. *Am J Clin Pathol* 2015;143:134-42.
93. Katayev A, Balciza C, Seccombe DW. Establishing reference intervals for clinical laboratory test results: is there a better way? *Am J Clin Pathol* 2010;133:180-6.
94. Jones G, Horowitz G, Katayev A, et al. Reference intervals data mining: getting the right paper. *Am J Clin Pathol* 2015;144:526-7.
95. Holmes DT, Buhr KA. Widespread Incorrect Implementation of the Hoffmann Method, the Correct Approach, and Modern Alternatives. *Am J Clin Pathol* 2019;151:328-36.
96. Hemel JB, Hindriks FR, van der Slik W. Critical discussion on a method for derivation of reference limits in clinical chemistry from a patient population. *The Journal of automatic chemistry* 1985;7:20-30.
97. Oosterhuis WP, Modderman TA, Pronk C. Reference values: Bhattacharya or the method proposed by the IFCC? *Ann Clin Biochem* 1990;27 ( Pt 4):359-65.
98. Pottel H, Vrydaghs N, Mahieu B, Vandewynckele E, Croes K, Martens F. Establishing age/sex related serum creatinine reference intervals from hospital laboratory data based on different statistical methods. *Clin Chim Acta* 2008;396:49-55.
99. Hoffmann JJ, van den Broek NM, Curvers J. Reference intervals of reticulated platelets and other platelet parameters and their associations. *Archives of pathology & laboratory medicine* 2013;137:1635-40.
100. DGKL Working Group on Reference Values. (Accessed March 5, 2020, at <https://www.dgkl.de/en/activities/sections/entscheidungsgrenzen-richtwerte/>.)
101. Concorde D, Geffre A, Braun JP, Trumel C. A new approach for the determination of reference intervals from hospital-based data. *Clin Chim Acta* 2009;405:43-8.
102. Tate JR, Sikaris KA, Jones GR, 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* 2014;35:213-35.
103. Ceriotti F. Prerequisites for use of common reference intervals. *Clin Biochem Rev* 2007;28:115-21.
104. Passing H, 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. *J Clin Chem Clin Biochem* 1983;21:709-20.
105. Brewster LM, Mairuhu G, Sturk A, van Montfrans GA. Distribution of creatine kinase in the general population: implications for statin therapy. *Am Heart J* 2007;154:655-61.
106. Karbasy K, Ariadne P, Gaglione S, Nieuwesteeg M, Adeli K. Advances in Pediatric Reference Intervals for Biochemical Markers: Establishment of the Caliper Database in Healthy Children and Adolescents. *J Med Biochem* 2015;34:23-30.
107. Ozarda Y, Ichihara K, Aslan D, et al. A multicenter nationwide reference intervals study for common biochemical analytes in Turkey using Abbott analyzers. *Clin Chem Lab Med* 2014;52:1823-33.
108. Hoq M, Matthews S, Karlaftis V, et al. Reference Values for 30 Common Biochemistry Analytes Across 5 Different Analyzers in Neonates and Children 30 Days to 18 Years of Age. *Clin Chem* 2019;65:1317-26.
109. Yamamoto Y, Hosogaya S, Osawa S, et al. Nationwide multicenter study aimed at the establishment of common reference intervals for standardized clinical laboratory tests in Japan. *Clin Chem Lab Med* 2013;51:1663-72.
110. den Elzen WPJ, Brouwer N, Thelen MH, Le Cessie S, Haagen IA, Cobbaert CM. NUMBER: standardized reference intervals in the Netherlands using a 'big data' approach. *Clin Chem Lab Med* 2018;57:42-56.
111. van der Meulen EA, Boogaard PJ, van Sittert NJ. Use of small-sample-based reference limits on a group basis. *Clin Chem* 1994;40:1698-702.
112. Holmes EW, Kahn SE, Molnar PA, Bermes EW, Jr. Verification of reference ranges by using a Monte Carlo sampling technique. *Clin Chem* 1994;40:2216-22.
113. Jones GR. Validating common reference intervals in routine laboratories. *Clin Chim Acta* 2014;432:119-21.
114. Jones GR, Koetsier SD. RCPAQAP First Combined Measurement and Reference Interval Survey. *Clin Biochem Rev* 2014;35:243-50.
115. Adeli K, Higgins V, Seccombe D, 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;50:925-35.
116. Koerbin G, Sikaris KA, Jones GR, Ryan J, Reed M, Tate J. Evidence-based approach to harmonised reference intervals. *Clin Chim Acta* 2014;432:99-107.
117. Koerbin G, Sikaris K, Jones GRD, Flatman R, Tate JR. An update report on the harmonization of adult reference intervals in Australasia. *Clin Chem Lab Med* 2018;57:38-41.
118. Solberg HE. Presentation of observed values in relation to reference values. *Bull Mol Biol Med* 1983;8:21-6.
119. Flatman R, Legg M, Jones GR, Graham P, Moore D, Tate J. Recommendations for reporting and flagging of reference limits on pathology reports. *Clin Biochem Rev* 2014;35:199-202.

120. Jones GR, Barker A, Tate J, Lim CF, Robertson K. The case for common reference intervals. *Clin Biochem Rev* 2004;25:99-104.
121. O'Connor JD. Reducing post analytical error: perspectives on new formats for the blood sciences pathology report. *Clin Biochem Rev* 2015;36:7-20.
122. Gullick HD, Schauble MK. SD unit system for standardized reporting and interpretation of laboratory data. *Am J Clin Pathol* 1972;57:517-25.
123. Haeckel R, Wosniok W. Quantity quotient reporting. A proposal for a standardized presentation of laboratory results. *Clin Chem Lab Med* 2009;47:1203-6.
124. Ceriotti F. Quantity quotient reporting. Counterpoint. *Clin Chem Lab Med* 2009;47:1207-8.
125. Mayer M, Chou D, Eytan T. Unit-independent reporting of laboratory test results. *Clin Chem Lab Med* 2001;39:50-2.
126. Solberg HE. Discriminant analysis. *CRC Crit Rev Clin Lab Sci* 1978;9:209-42.
127. Haeckel R, Wosniok W. Observed, unknown distributions of clinical chemical quantities should be considered to be log-normal: a proposal. *Clin Chem Lab Med* 2010;48:1393-6.
128. Parvin CA, Gray DL, Kessler G. Influence of assay method differences on multiple of the median distributions: maternal serum alpha-fetoprotein as an example. *Clin Chem* 1991;37:637-42.
129. Rossing RG, Hatcher WE, 3rd. A computer program for estimation of reference percentile values in laboratory data. *Comput Programs Biomed* 1979;9:69-74.
130. Albert A, Heusghem C. Relating observed values to reference values: the multivariate approach. In: Grasbeck R, Alstrom T, eds. *Reference values in laboratory medicine*. Chichester, UK: John Wiley; 1981:289-96.
131. Albert A, Harris EK. *Multivariate interpretation of clinical laboratory data*. New York: Marcel Dekker; 1987.
132. Winkel P. Patterns and clusters—multivariate approach for interpreting clinical chemistry results. *Clin Chem* 1973;19:1329-38.
133. Morrison DF. *Multivariate statistical methods*. New York: McGraw-Hill; 1990.
134. Winkel P, Lyngbye J, Jorgensen K. The normal region—a multivariate problem. *Scand J Clin Lab Invest* 1972;30:339-44.
135. Boyd JC, Lacher DA. The multivariate reference range: an alternative interpretation of multi-test profiles. *Clin Chem* 1982;28:259-65.
136. Hoermann R, Larisch R, Dietrich JW, Midgley JE. Derivation of a multivariate reference range for pituitary thyrotropin and thyroid hormones: diagnostic efficiency compared with conventional single-reference method. *European journal of endocrinology* 2016;174:735-43.
137. Malka R, Brugnara C, Cialic R, Higgins JM. Non-Parametric Combined Reference Regions and Prediction of Clinical Risk. *Clin Chem* 2020;66:363-72.
138. Winkel P. The use of the subject as his own referent. In: Grasbeck R, Alstrom T, eds. *Reference values in laboratory medicine*. Chichester, UK: John Wiley; 1981:65-78.
139. Petersen PH, Fraser CG, Sandberg S, Goldschmidt H. The index of individuality is often a misinterpreted quantity characteristic. *Clin Chem Lab Med* 1999;37:655-61.
140. Aarsand AK, Roraas T, Fernandez-Calle P, et al. The Biological Variation Data Critical Appraisal Checklist: A Standard for Evaluating Studies on Biological Variation. *Clin Chem* 2018;64:501-14.
141. Jones GRD. Estimates of Within-Subject Biological Variation Derived from Pathology Databases: An Approach to Allow Assessment of the Effects of Age, Sex, Time between Sample Collections, and Analyte Concentration on Reference Change Values. *Clin Chem* 2019;65:579-88.
142. Klajnbard A, Szecsi PB, Colov NP, et al. Laboratory reference intervals during pregnancy, delivery and the early postpartum period. *Clin Chem Lab Med* 2010;48:237-48.
143. Carlsson L, Lind L, Larsson A. Reference values for 27 clinical chemistry tests in 70-year-old males and females. *Gerontology* 2010;56:259-65.
144. Non-invasive assessment of hepatic fibrosis: overview of serologic and radiographic tests. 2020. (Accessed February 20, 2020, at <http://www.uptodate.com/contents/noninvasive-assessment-of-hepatic-fibrosis-overview-of-serologic-and-radiographic-tests>.)
145. Stephan C, Vincendeau S, Houlgate A, Cammann H, Jung K, Semjonow A. Multicenter evaluation of [-2]prostate-specific antigen and the prostate health index for detecting prostate cancer. *Clin Chem* 2013;59:306-14.
146. De Grande LA, Goossens K, Van Uytfanghe K, Stockl D, Thienpont LM. The Empower project - a new way of assessing and monitoring test comparability and stability. *Clin Chem Lab Med* 2015;53:1197-204.

## MULTIPLE CHOICE QUESTIONS

1. Which of the following processes should be followed to achieve more homogeneous populations for reference interval testing?
  - a. Obtain specimens from additional subjects
  - b. Repeat analyses on current specimens
  - c. Transform the current data
  - d. Partition the current data
  - e. Find additional outliers among the current data
2. Which of the following characteristics represents an advantage of the nonparametric technique for determining reference intervals?
  - a. It requires a smaller number of specimens than the parametric technique
  - b. There is no need to qualify subjects for inclusion
  - c. There is no need to partition the data
  - d. There is no need to eliminate outliers
  - e. A Gaussian distribution is not required
3. Which of the following represents an advantage of the indirect method of determining reference intervals?
  - a. It can be performed on any population
  - b. It is less expensive
  - c. It requires smaller numbers of specimens
  - d. It does not require partitioning
  - e. It is most useful for specialist referral laboratories
4. Which of the following distinguishes clinical decision limits from conventional reference intervals?
  - a. Outcomes or diagnoses for subjects need to be known
  - b. They are generally method-dependent
  - c. Specific percentages of subjects are above and below thresholds
  - d. They do not require partitioning
  - e. They are easier to establish
5. “Subject-based” reference intervals
  - a. Are obtained from a group of systematically defined reference individuals
  - b. Are the type of reference intervals referred to when the “reference interval” is used with no qualifying words
  - c. Are based on several specimens collected over time in single individuals
  - d. Represent a random sample of all individuals in the parent population
  - e. Are particularly useful for tests where the interindividual variability is similar to the inter-individual variability
6. Ongoing review of the distribution of outpatient data:
  - a. Represents a standard part of a laboratory quality control program
  - b. Is required for lab accreditation
  - c. Is useful to monitor clinical decision limits
  - d. Can detect drift in analytical methods
  - e. Requires transforming the data to a Gaussian distribution
7. The distribution of laboratory data from healthy individuals is often
  - a. Gaussian
  - b. Skewed to the right
  - c. Wider than Gaussian
  - d. Bimodal
  - e. Nonparametric
8. In a patient whose calcium is significantly below the lower reference limit, which of the following tests would be expected to be outside conventional reference limits?
  - a. Alanine aminotransferase
  - b. Bilirubin
  - c. Parathyroid hormone
  - d. Cortisol
  - e. Prothrombin time
9. Which of the following would most likely be used as an exclusion criterion for a reference interval study for serum iron?
  - a. Age
  - b. Sex
  - c. Ethnicity
  - d. Body mass index
  - e. Recent transfusion
10. When reference limits are determined, which of the following methods requires a Gaussian distribution?
  - a. Nonparametric
  - b. Parametric
  - c. Interpercentile
  - d. Bootstrap
  - e. Robust

# Evidence-Based Laboratory Medicine\*

*Patrick M.M. Bossuyt, Paul Glasziou, and Andrea R. Horvath*

## ABSTRACT

### Background

Evidence-based laboratory medicine (EBLM) is an approach to medical practice that integrates the best available research evidence about laboratory investigations with the clinical expertise of clinicians, to improve the health and health care outcomes of individual patients. Practicing EBLM enables laboratory professionals to translate test results to clinically relevant information that helps clinicians in delivering effective and cost-effective patient care.

### Content

This chapter provides an overview on how evidence about laboratory tests is generated, how it is synthesized, and how it

can be applied to questions about diagnosis, screening, prognosis, or monitoring. The topics covered here introduce the reader to the methodological and practical aspects of EBLM. They include (1) the process and methods of practicing EBLM, (2) the key components and types of evidence used in the evaluation of biomarkers, (3) tools for the assessment of the validity and applicability of the evidence, (4) key aspects of synthesizing the evidence in systematic reviews and meta-analyses, (5) basic principles of how EBLM is applied to other purposes of testing than diagnosis, (6) the challenges and tools of implementing the evidence for achieving best laboratory practice, and (7) the history and future challenges of EBLM.

---

\*The full version of this chapter is available electronically on [ExpertConsult.com](#).

## INTRODUCTION

Evidence-based medicine (EBM) was introduced in the 1980s, as “the conscientious, judicious, and explicit use of the best evidence in making decisions about the care of individual patient.”<sup>1,1a</sup> Evidence-based laboratory medicine (EBLM) is the application of principles and techniques of EBM to laboratory medicine: making decisions about different aspects of laboratory testing for individual patients based on the best available evidence from sound research. As such, EBLM aims to improve the value and impact of laboratory testing on health and health care delivery, by providing benefits to patients at acceptable costs, based on solid evidence from scientific research.<sup>1b</sup>

As outlined in Chapter 1, the typical role of the clinical chemist is to “use technology efficiently to derive answers to clinical questions.” If laboratorians wish to provide clinically meaningful answers, their role should not end with just producing high quality measurement data at short turn-around-times. Laboratory professionals should add value to their service by actively translating in vitro testing data to clinically relevant information which clinicians could use in making better informed management choices that lead to the most favorable outcomes for their patients. Practicing EBLM enables laboratory professionals to become such “data translators” and information and knowledge resources.

In this chapter, we provide the reader with an introduction to the practice of EBM, applied to the laboratory. This chapter covers seven key topics that aim to provide a contemporary overview on the methodological and practical aspects of EBLM. The seven sections describe:

1. The key steps that define the process of practicing EBLM.
2. The key components and types of evidence used in the evaluation of the performance and impact of laboratory tests on clinical practice and health and health care outcomes.
3. Tools for the evaluation of the validity of the evidence related to laboratory testing.
4. Key aspects of synthesizing the evidence in systematic reviews and meta-analyses.
5. The specific methodological and practical considerations when EBLM is applied to other purposes of testing than diagnosis.
6. The challenges and tools of implementing the evidence for best laboratory practice.
7. The history and future of EBLM to provide an overview of the evolution of EBM and the challenges and limitations that still need to be surpassed in order to achieve best laboratory practice that is based on sound and high-quality evidence.

This chapter mostly focuses on how evidence is generated, synthesized, and applied when a laboratory test is used for diagnostic purposes. The principles are described in generic fashion; most apply to screening, and prognostic and monitoring tests.

This chapter aims to provide basic understanding and practical knowledge that will help the reader become an EBLM professional. Readers interested in the more intricate details are referred to the massively exploding literature, books, and web-based resources that provide a deeper insight into the topic.<sup>2,3</sup>

## THE EVIDENCE-BASED MEDICINE PROCESS

The process of practicing EBLM starts with a clinical problem, followed by a series of steps, called the “A5” cycle (Fig. 10.1).<sup>2</sup> These steps refer to the following activities:

1. Identify the clinical problem.
2. Ask or formulate the question to help solve the problem.
3. Acquire the evidence that addresses the question.
4. Appraise the evidence.
5. Apply the knowledge gained from the evidence in resolving the problem.
6. Audit the application of the evidence.

Below we discuss each of these steps in more detail.

### Identify the Clinical Problem

The identification of a clinical problem is both the starting point and the foundation of the service provided by the health care professional.

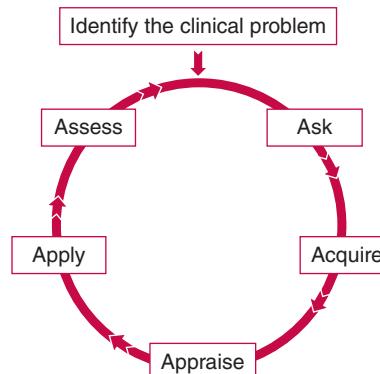
In EBLM, clinical problems are dilemmas about the care of patients for a health professional that involves laboratory tests. A dilemma relates to a choice between actions; between doing one thing rather than another. In laboratory testing these are questions about whether or not to test, or issues about selecting one test rather than another.

An example of a test-related clinical problem is the following:

*A general practitioner, Dr. Ann Brodie, is seeing a patient, Mr. Smith, a 69-year-old male, who presents with shortness of breath during exercise, increasing fatigue and weakness, and some swelling around the ankles. Mr. Smith has been a patient for many years. Dr. Brodie knows the history of Mr. Smith very well, including that he had no cardiovascular events so far.*

*Dr. Brodie suspects congestive heart failure but hesitates about her next steps: sending Mr. Smith to the closest hospital to see a cardiologist, while she knows Mr. Smith does not like hospitals. Mr. Smith's basic risk factors seem well controlled, and Dr. Brodie also considers the downsides associated with overtesting, should the cardiologist decide to run a large battery of tests.*

*On the other hand, a missed diagnosis of congestive heart failure may also have clear consequences for Mr. Smith: if correctly identified as having heart failure,*



**FIGURE 10.1** The “A5” evidence-based laboratory medicine cycle. (Adapted from Price CP, Christenson RH, editors. *Evidence-based laboratory medicine: principles, practice and outcomes*. 2nd ed. Washington, DC: AACC Press; 2007. p. 1–545.) (Copyright permission by AACC).

treatment is likely to improve his symptoms, and reduce his long-term risk of adverse events.

An alternative Dr. Brodie might consider is to order an N-terminal prohormone of brain natriuretic peptide (NT-proBNP) blood test. For this, Dr. Brodie has to consider how the test may guide her next steps. If she plans to send Mr. Smith to the cardiologist, regardless of the test result, then the test is obviously unnecessary. The NT-proBNP test can only be useful if it generates a result that will change her actions, compared to the default option; that is, sending Mr. Smith to the cardiologist. In this case, she may decide not to refer Mr. Smith if the NT-proBNP result is less than 300 ng/L (36 pmol/L).<sup>4</sup>

Now we clearly have a clinical problem: a dilemma, well defined as a choice between two lines of action: the first option is to send the patient to a cardiologist in the nearby hospital for further evaluation, the second is to only do so if the NT-proBNP result is positive, and not to refer Mr. Smith if the result is negative.

Essential for EBM is that the solution for clinical problems like this one is driven by a consideration of the likely consequences of each line of action. What is likely to happen to Mr. Smith if he is sent to the cardiologist? How likely is this to affect his health, now and in the future? What is likely to happen to him if he is not sent to the cardiologist, now at least? How likely is this to affect his health, now and in the future?

In the interest of Mr. Smith, the line of action should be adopted that gives him the best options for a better health, now and in the future. Yet decisions should also take into account the consequences for others involved; for example, for his spouse or family, for society as a whole, which has to cover the costs of testing, further evaluation, treatment, and management of future events.

### Ask the Right Question: The Patients-Intervention-Comparator-Outcomes Format

As a laboratory professional practicing EBLM, you offer your help to Dr. Brodie. To help her solve the clinical problem, you both should be asking the right question.

Doctors and other health care professionals are constantly asking questions, and these can be divided into background questions and foreground questions.

*Background questions*, to date, are more commonly asked by newly qualified professionals by virtue of the ways they have been taught. Such questions typically deal with knowledge (or underlying science) of the pathophysiology or clinical presentation of a condition (e.g., Why is the circulating concentration of troponin I increased in acute coronary syndrome (ACS)? Why and how NT-proBNP is released into the circulation in congestive heart failure?).

*Foreground questions* are related specifically to the application of knowledge, of experience in treating the condition and using tests (e.g., Will a troponin I measurement help me determine whether this patient is suffering from an ACS? Does measurement of NT-proBNP help diagnose heart failure in addition to clinical signs and traditional diagnostic investigations in patients presenting with acute shortness of breath?).

Clinicians tend to ask more foreground (and fewer background) questions as their experience develops. This may change in coming years as a more evidence- and outcomes-based approach to teaching medical students, and training doctors, evolves.

Richardson and coworkers argued that all clinical problems could be expressed in the form of a question and went on to describe the PICO framework for formulating an answerable question<sup>5</sup>:

The four letters P-I-C-O refer to the four elements of the question:

P stands for “Patients”

I stands for the “Index Test,” or the “Intervention” that is considered

C stands for the “Comparator”

O stands for the patient-relevant “Outcomes.”

In laboratory medicine we look at an **Index test**: the test that is being evaluated. This can be an *in vitro medical assay* to measure a **biomarker** (e.g., B-type natriuretic peptide, or high-sensitivity Troponin, or fecal hemoglobin), or a combination of such tests (see definitions and examples in Table 10.1). Alternatively, we can look at a strategy, in which the index test is used to guide further actions, as in a screening strategy.

When the intervention is considered in the context of laboratory medicine, it is worth considering the **purpose** of testing: for example, is the test for (1) screening, (2) diagnosis, (3) prognosis, or (4) monitoring of a condition (see definition and examples in Table 10.1).

An additional relevant question is about the **role** of the test in the clinical pathway, relative to other possible forms of testing (see definition and examples in Table 10.1). Is this test used for *triage*, for example, to prevent patients from undergoing other, more expensive, or invasive tests? In that case, only patients testing positive on the triage test will undergo these more expensive or invasive tests, and the ones who are triage test negative will not. An alternative role of a test considered in the clinical problem is *replacement*: should I do this test, rather than another? A third role of a test is an *add-on* one: should I add this test to one or more tests, already performed, or should I stop after this testing strategy?

In the example of Dr. Brodie, who considers NT-proBNP testing for Mr. Smith, the purpose of testing is clearly a diagnostic one. Mr. Smith decided to see his general practitioner because of the way his shortness of breath was affecting his life; he wanted relief of his symptoms (i.e., health outcomes). Considering the cause of these may be a steppingstone toward effective treatment.

The **outcomes** refer to patient-relevant health outcomes that the intervention is intended to influence. This could be removing symptoms and restoring health or preventing premature death or worsening of symptoms. However, in laboratory medicine we also often consider surrogates for these outcomes, such as a target condition that is being detected (e.g., heart failure), or a target event that could happen in the future (e.g., markers in patients with ACS, to predict the risk of cardiovascular events after discharge).

One crucial observation must be made regarding these well-phrased questions. In all examples testing itself will not directly affect the outcome considered. NT-proBNP testing will not directly reduce the severity of symptoms, and fecal hemoglobin testing will not directly remove any form of colorectal cancer. The effect between testing and outcomes is an indirect one: testing will be used to guide downstream actions, such as starting or stopping interventions, or communicating results to patients. Whether these downstream

**TABLE 10.1 Key Terms and Definitions in Evidence-Based Laboratory Medicine**

<b>Key Term</b>	<b>Related Terms</b>	<b>Explanation (Reference)</b>	<b>Examples</b>
In vitro medical assay	In vitro diagnostic medical device Laboratory assay Measurement procedure	A measurement procedure undertaken on a biological specimen which measures the quantity of the biomarker (see below) intended to be measured; i.e., the measurand. <sup>6</sup>	<ul style="list-style-type: none"> <li>Two-site immunoenzymatic ("sandwich") assay using electrochemiluminescence detection for cardiac troponin (cTn) measurement</li> <li>Cation exchange chromatography or boronate affinity chromatography or latex agglutination immunoassay to measure HbA<sub>1c</sub></li> </ul>
Biomarker	Biological marker	A characteristic that is an indicator of normal biological or pathogenic processes, or pharmacologic responses to a therapeutic intervention. <sup>7</sup>	<ul style="list-style-type: none"> <li>cTns are biomarkers of cardiac diseases associated with myocardial ischemia and necrosis</li> <li>HbA<sub>1c</sub> is a biomarker of altered glycosylation in hyperglycemic states, such as diabetes mellitus</li> </ul>
In vitro medical test	Medical test or testing strategy	In vitro medical tests or testing strategies utilize laboratory assays of biomarkers in a specific clinical context and for a specific clinical purpose (see below), in a specific patient population.	<ul style="list-style-type: none"> <li>Serial cTn testing for diagnosing acute coronary syndrome (ACS) in patients with symptoms of acute chest pain</li> <li>HbA<sub>1c</sub> as a monitoring test to assess treatment effect in type 1 or type 2 diabetic patients</li> </ul>
Clinical pathway	Clinical algorithm Care pathway Critical pathway Care map Guideline	A description of typical processes of care in managing a specific condition in a specific group of patients. <sup>8</sup>	<ul style="list-style-type: none"> <li>Clinical Pathways by NICE in the UK: <a href="http://pathways.nice.org.uk">http://pathways.nice.org.uk</a></li> <li>Clinical pathway for the management of ACS: <a href="https://pathways.nice.org.uk/pathways/chest-pain">https://pathways.nice.org.uk/pathways/chest-pain</a></li> </ul>
Test purpose	Intended use of test Indication for testing Claim (in the context of manufacturer's claim for the intended use of the test)	Test purpose describes the intended clinical application of the test and how the test information will be used to improve clinical management in practice.	<ul style="list-style-type: none"> <li>HbA<sub>1c</sub> as a diagnostic marker of diabetes mellitus</li> <li>HbA<sub>1c</sub> as a monitoring test to assess diabetes control</li> <li>cTn for diagnosing ACS</li> <li>cTn as a prognostic marker for cardiovascular morbidity and mortality</li> </ul>
Test role	Replacement test Triage test Add-on test Reflex testing Reflective testing	<p>Test role describes how the test, used for a specific clinical purpose, will be positioned to alter the existing clinical pathway in a specific condition or target population<sup>9</sup>:</p> <ul style="list-style-type: none"> <li><b>Replacement:</b> When a new test replaces an existing test in the testing pathway.</li> <li><b>Triage:</b> When the new test is used before the existing test or testing pathway, and only patients with a particular result on the triage test continue on the testing pathway.</li> <li><b>Add-on:</b> When a new test is added to the existing testing pathway, either to help interpreting results of analyses when establishing a diagnosis or to assist patient management.</li> </ul>	<p><i>Replacement:</i></p> <ul style="list-style-type: none"> <li>cTn-s replacing creatine kinase myocardial band (CK-MB) as a biomarker of myocardial damage;</li> <li>C-Reactive protein replacing erythrocyte sedimentation rate as marker of acute inflammation</li> </ul> <p><i>Triage:</i></p> <ul style="list-style-type: none"> <li>Natriuretic peptides before echocardiography for congestive heart failure</li> </ul> <p><i>Add-on:</i></p> <ul style="list-style-type: none"> <li>Immunofixation for typing is added when monoclonal gammopathy is found on serum protein electrophoresis</li> <li>HbA<sub>1c</sub> monitoring together with self-monitoring of blood glucose in managing type 1 diabetes patients</li> </ul>

Table adapted from Horvath AR, et al. *Clin Chim Acta* 2014;427:49–57 (Copyright permission by Elsevier).

interventions are effective, and whether they are done in the right patients at the right time, will eventually determine the effect of testing on patient outcome.

In the example of fecal hemoglobin, a positive test result will be used to invite some screening participants for further investigation with colonoscopy. The colonoscopy will be used

to look for cancer, advanced adenoma, or other precursor lesions, and to remove them, if possible.

This implies that it is often necessary, when phrasing the PICO type question, to define the **clinical pathway**: the typical chain of actions in the process of care that is guided by results testing, and able to affect the relevant outcomes.

The **comparator** is an alternative test, or an alternative intervention: what would we do if we do not apply the form of testing considered under the “intervention”? This could be another test (e.g., the GRACE score in the case of unstable angina patients), a clinical action (referral for patients with suspected heart failure, seen by the general practitioner), or no intervention right now (as in screening for colorectal cancer).

In EBLM, questions about testing are never considered in a void, but always for specific types of **patients**, in defined settings (or, perhaps more importantly, at specific points in a diagnostic pathway). In the example of Mr. Smith, again, we must ask ourselves what the defining features are in the clinical question. We know, for example, that he presents himself to a general practitioner with symptoms. That means that studies about NT-proBNP testing in symptomatic patients are relevant, while studies about screening with NT-proBNP in nonsymptomatic elderly people are not. This distinction is relevant because the clinical performance of NT-proBNP and its ability to improve health outcomes is not identical in symptomatic patients, compared to nonsymptomatic persons.

For a specific clinical problem, there may be just one PICO-style question, but it is also possible that the problem leads to two or more PICO style questions. This typically happens if we have multiple outcomes to consider, for example, short-term consequences versus longer-term effects, or positive outcomes (benefits) versus negative ones (harms).

An appropriate definition of the well-phrased question, designed to help solve the clinical problem, will guide the search for relevant evidence from sound research. **Table 10.2** contains several structured questions in this PICO format for these various purposes of testing.

### POINTS TO REMEMBER

- Evidence-based laboratory medicine (EBLM) is a clinical decision support tool that improves health and health care outcomes by integrating the best available research evidence related to laboratory testing with the clinical expertise of the physician and the needs of individual patients.
- Practicing EBLM enables laboratory professionals to translate test results to clinically relevant information which help clinicians deliver the most effective and cost-effective care to their patients.
- The key components of the “5A” cycle of EBLM are ASK-ACQUIRE-APPRAISE-APPLY-ASSESS the evidence.
- A well-phrased clinical question helps searching for the evidence and has the following components: Patients-Intervention-Comparator-Outcomes (PICO).

### Acquire the Evidence

Once we have phrased a well-defined question, we can look for the available evidence. Here a number of options are presented, in decreasing order of strength, according to the “5S” hierarchical structure, described by Brian Haynes.<sup>16</sup> In this hierarchy, (1) original *studies* are at the base, followed by (2) *syntheses* (systematic reviews) of the evidence, (3) *synopses* of studies and syntheses, (4) evidence *summaries* (e.g., guidelines), and (5) the most evolved evidence-based information *systems* at the top (**Fig. 10.2**).

Busy clinicians and laboratorians prefer having information and knowledge readily available right next to the point of health care delivery. With rapidly advancing information technology, electronic decision support systems can now be built and integrated into electronic medical records, but the lack of high-quality evidence and evidence-based guidelines related to laboratory testing still limits the availability of such smart solutions.

A few initiatives, such as a mobile application of a Partial Thromboplastin Time Advisor at the US Centers for Disease Control and Prevention, have already been successfully implemented and a free app has been made accessible by Apple via their iTunes App Store.<sup>17</sup> For more information on clinical decision support systems and their application to diagnostic testing the reader is referred to a collection of useful resources (**Table 10.3**).

Currently the busy health care professional more often turns to evidence-based practice (EBP) guidelines, where professional colleagues have used the principle of EBLM to develop recommendations for practice. We will discuss evidence-based guidelines below.

If such guidelines are not available, you may turn to critically appraised evidence synopses, which can be found in evidence-based journals, such as *BMJ* and *EBM* (see **Table 10.3**).

If you have no success still, move to the next level in the hierarchy as there may be others who have systematically searched and synthesized the available literature for you in form of a systematic review or meta-analysis that answers your PICO-type question. Unfortunately, few databases exist for systematic reviews that address diagnostic testing; a few resources are listed in **Table 10.3**.

If you cannot find evidence-syntheses produced by others, you need to turn to the primary literature and search for reports of individual studies that can help answer the PICO type question. Various tools have been produced to help professionals search the available literature. We discuss a few of these below.

### Tools in Searching for Evidence

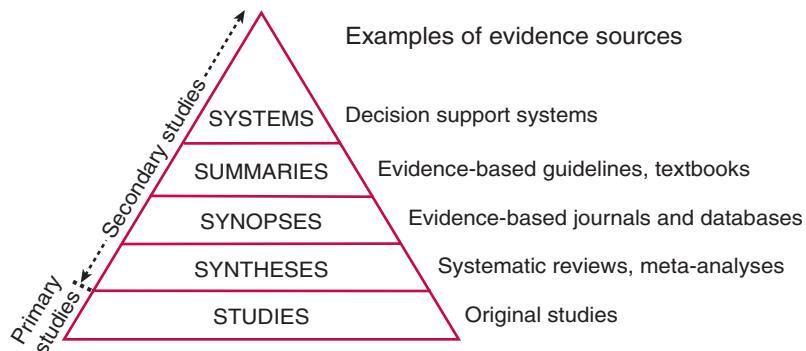
Where you search first will depend on the type of question asked. There are several useful EBM databases, with a few also specific to medical testing and laboratory medicine (see **Table 10.3**). The TRIP database is particularly helpful for quick searches and allows you to specifically search with the PICO elements of your question. This database categorizes the publications found according to the above-mentioned hierarchy and allows filtering for the various types of resources. However, if none of these searches is appropriate, a generic search would use Medline (via PubMed). For diagnostic accuracy or prognostic questions there are some helpful built-in filters in PubMed Clinical Queries that have been developed by the health informatics unit at McMaster University. These attempt to find the higher quality research relevant to your PICO.

Some tips in converting your PICO to search terms are:

1. Initially use the P and I of your search terms rather than the full PICO.
2. Consider the possible synonyms for your search terms and combine these with an OR, so the search will be (P-term-1 OR P-term-2) AND (I-term-1 OR I-term-2).

**TABLE 10.2 Formulating Patients-Intervention-Comparator-Outcomes Questions Related to Laboratory Tests**

Clinical Question	Type of Question	Patient	Index Test/ Intervention	Comparator	Outcome
<b>Purpose of Testing: Screening</b>					
1 To what extent can NT-proBNP measurement detect congestive heart failure in an unselected group of patients? <sup>10</sup>	Screening accuracy	Unselected patients above 40 years of age	NT-proBNP	None	Differentiation of patients with normal and reduced left ventricular systolic function
2 Does fecal hemoglobin screening of asymptomatic individuals between 50 and 75 years of age reduce mortality from colorectal cancer, compared to no screening?	Screening strategy	Asymptomatic adults of 50–75 years of age	Fecal hemoglobin testing	No screening	Mortality from colorectal cancer
<b>Purpose of Testing: Diagnosis</b>					
3 In ambulatory patients presenting to the ER with acute dyspnea, can NT-proBNP diagnose heart failure? <sup>11</sup>	Diagnostic accuracy	Ambulatory adults presenting to the ER with dyspnea	NT-proBNP	None	Heart failure
4 In patients presenting to ER with acute chest pain, does cardiac troponin (cTn) T testing by a point of care testing (POCT) device, compared to high sensitivity cTn T testing in the laboratory, improve mortality of acute myocardial infarction?	Comparative diagnostic accuracy	Patients presenting to the ER with acute chest pain	cTn T testing by POCT device in ER	High sensitivity (hs) cTn T testing in the laboratory	Mortality from acute coronary syndrome
<b>Purpose of Testing: Prognosis</b>					
5 In patients admitted to hospital with heart failure does discharge NT-proBNP concentration predict all-cause mortality or readmission to hospital? <sup>12</sup>	Prognostic accuracy	Adults with heart failure recently discharged from hospital	Discharge NT-proBNP	None	All-cause mortality or readmission to hospital
6 In patients treated for heart failure does NT-proBNP-guided therapy, compared to usual clinical care, reduce all-cause mortality, heart failure-related hospitalization, and all-cause hospitalization? <sup>13</sup>	Prognostic strategy	Adult patients treated for heart failure in primary care	Quarterly NT-proBNP guided therapy	Clinically guided therapy	All-cause mortality, heart-failure related hospitalization, all-cause hospitalization
7 In patients with acute chest pain, diagnosed as unstable angina, does serial hs-Troponin testing, compared to using the GRACE risk score, prevent premature deaths?	Prognostic strategy	Adults with acute chest pain, diagnosed as unstable angina	Serial hs-cTn testing	GRACE risk score	Premature deaths
<b>Purpose of Testing: Monitoring</b>					
8 Does quarterly monitoring with brain natriuretic peptide (BNP) or NT-proBNP help in assessing treatment response and guide therapy to improve symptoms of congestive heart failure? <sup>14,15</sup>	Monitoring strategy	Patients with congestive heart failure	Quarterly BNP or NT-proBNP testing	Routine clinical care	Response to therapy—improved left ventricular function; improved quality of life
9 In patients with type 2 diabetes, does daily self-monitoring of blood glucose, compared to biannual checking of HbA <sub>1c</sub> , improve metabolic control and decrease the progression of secondary complications of diabetes?	Monitoring strategy	Patients with type 2 diabetes	Self-monitoring of blood glucose daily	Biannual checking of HbA <sub>1c</sub>	Secondary complications of diabetes



**FIGURE 10.2** The “5S” system of evidence levels. (Adapted from Haynes BR. Of studies, syntheses, synopses, summaries, and systems: the “5S” evolution of information services for evidence-based healthcare decisions. *Evid Based Med* 2006;11:162–64.) (Copyright by BMJ).

**TABLE 10.3 Evidence-Based Laboratory Medicine/Evidence-Based Medicine Resources for Acquiring the Evidence**

“5S” Hierarchy	Type of Resource	Options/Examples	Links (last accessed May 28th, 2020)
Systems	Decision support	Open Clinical	<a href="https://www.opencds.org/">https://www.opencds.org/</a>
Summaries	Guidelines	NICE Evidence Guidelines International Network (GIN) Agency for Healthcare Research and Quality	<a href="https://www.evidence.nhs.uk">https://www.evidence.nhs.uk</a> <a href="https://www.g-i-n.net">https://www.g-i-n.net</a> <a href="https://www.ahrq.gov">https://www.ahrq.gov</a>
Synopses	EBM journals or portals	<i>BMJ Evidence-Based Medicine</i> UpToDate	<a href="https://ebm.bmj.com">https://ebm.bmj.com</a> <a href="https://www.uptodate.com">https://www.uptodate.com</a>
Syntheses	Systematic reviews	Cochrane Library	<a href="https://www.cochranelibrary.com">https://www.cochranelibrary.com</a>
Studies	Primary studies	PubMed Clinical Queries	<a href="https://pubmed.ncbi.nlm.nih.gov/clinical/">https://pubmed.ncbi.nlm.nih.gov/clinical/</a>
All above	All types	TRIP Database	<a href="https://www.tripdatabase.com">https://www.tripdatabase.com</a>

EBM, Evidence-based medicine; EBLM, evidence-based laboratory medicine.

3. Consider also possible alternative spellings and truncations, for example, using faecal OR fecal, or using colonoscopy\* to cover the options of colonoscope OR colonoscopy OR colonoscopies, etc. For example, when searching for the NT-proBNP diagnostic accuracy question (Question 3) in Table 10.2 various synonyms can be used and combined: for example, [natriuretic peptide, brain OR brain-type natriuretic peptide OR natriuretic factor OR type-b natriuretic peptide OR BNP OR NT-proBNP OR ntproBnp] AND [diagnos\* OR diagnosis]. For an example on a detailed search strategy for screening, diagnostic, prognostic and monitoring questions related to brain natriuretic peptide (BNP), the reader is referred to this excellent systematic review.<sup>15</sup> Finally, with PubMed it is often worth also looking at the Related Articles tab when you find an article that is close to your question.

### Useful Evidence-Based Laboratory Medicine Resources

There are several developments that help laboratory professionals who want to practice EBLM. Much of the relevant evidence is now synthesized in evidence summaries, systematic reviews, or guidelines. Table 10.3 sets out a search sequence (based on Brian Haynes’ idea of the “5S” hierarchy), with some suggested resources. For example, if we searched the UK National Health Service or NHS Evidence for BNP it will

suggest a few hundred resources, but we can further focus this by selecting an option such as systematic reviews or evidence summaries.

However, syntheses and summaries of evidence are likely to be incomplete and/or may be out of date. Therefore for completeness or checking for more recent evidence, it is useful to learn PubMed Clinical Queries which provides excellent search filters for different categories of questions such as diagnostic accuracy, prognosis, or treatment. For a more detailed description of search strategies and resources the reader is referred to a textbook on EBLM.<sup>18</sup>

### Appraise the Evidence

If we have found forms of evidence synthesis, or single studies, we must carefully examine the validity and applicability of the available evidence. This means that we cannot just read the bottom lines of the discussion section of these study reports to find out the answer to the clinical question; we must study the research that has been performed.

### Appraising the Applicability of the Evidence

It is possible that the scientific studies that have been performed were designed to answer a question that is related, but not identical, to the question that we have defined in the PICO format. In that case, we will carefully evaluate to what

extent the evidence generated matches our question. This can range from “not at all” to “completely.”

For example, in [Table 10.2](#), a study addressing Question #4 investigates whether cardiac troponin (cTn) T testing by a point-of-care testing (POCT) device, compared to hs-cTn T testing in the laboratory, improves mortality of acute myocardial infarction (AMI) patients presenting to the emergency room (ER) with acute chest pain. If your laboratory measures cTn I, or an older generation of the cTn T test, the results of this study will not at all be applicable to your question.

Another study investigating Question #2 (i.e., whether fecal hemoglobin screening of asymptomatic individuals between 50 and 75 years of age have reduced mortality from colorectal cancer), may not be completely applicable if your question addresses a different age group or patients with changed bowel habits and thus potentially higher prevalence of underlying gastrointestinal conditions.

It is also possible that your laboratory can only measure BNP but not NT-proBNP. In that case the results of a study that provides an answer to Question #6 related to NT-proBNP guided heart failure therapy and its effect on mortality and hospitalization outcomes will not be directly transferable to your local setting.

While the study for Question #8, which compares BNP and NT-proBNP guided therapy, seems more relevant for answering your question, the outcomes investigated in the study (i.e., improved left ventricular systolic function as a proxy outcome and improved quality of life as a health outcome) are quite different from your originally addressed outcomes of mortality and hospitalization.

### Appraising the Risk of Bias in the Evidence

In addition to an appraisal of the applicability of the generated evidence, we also have to evaluate whether the study that was conducted suffered from limitations. It is now well known that not every study that is reported in the peer-reviewed literature is free from flaws. Peer review is not a perfect system.

In every area of science some studies are performed that have minor or serious shortcomings in the basic study design, in the protocol, in the execution of the protocol, in the statistical analysis, or in the process of reporting itself.

This means that professionals that aim to practice EBLM should be able to distinguish the studies with deficiencies from the ones without, and that they should be capable of evaluating to what extent the flaws affect the results that were generated. Are they fatal, which means that we cannot trust the evidence at all, or are they minor, allowing us to have confidence in the results that were generated?

A key concept here is “bias”: systematic deviations from the truth in study results; that is, either overestimation or underestimation of the true effect of the studied intervention that would have or could have been generated without imperfections in the design and/or conduct of the research trial.

Several critical appraisal tools have been developed that should make it easier for professionals who want to practice EBLM to distinguish strong studies from weaker ones. Tools are available for evaluating the risk of bias in diagnostic accuracy studies, prognostic marker studies, randomized and nonrandomized trials to compare interventions.<sup>19</sup>

It is important to estimate the magnitude and the likely direction of the bias. For example, if the methodological limitations of a trial lead to underestimation of the true effect of the intervention, but the study shows that the intervention is effective, then it is safe to conclude that the intervention is effective, in spite of the presence of potential biases.

### Apply the Knowledge Gained from the Evidence

Once we have found the evidence and have identified any limitations in applicability or validity, we can try to resolve the problem by combining the various pieces of evidence. There may be two or more PICOs addressing multiple outcomes.

Evidence in itself does not always tell us what we have to do. Whether the evidence applies, and whether evidence is sufficient to take or to recommend action, depends on values and preferences. Balancing harms and outcomes, or health gains and costs, can lead to different recommendations, based on differences in individual goals in life, or guided by the availability in scarce resources. What may be sensible for a young man may no longer be recommended for an elderly male, and what seems reasonable in a richer country may seem unattainable in resource-poor settings.

### Audit the Application of Evidence-Based Laboratory Medicine

Assessing or auditing the impact of the application of evidence in practice is the final step in the “A5” cycle. Clinical audit, also called “cooperative audit,” conducted between the laboratory and its clinical customers, is a tool that helps improve the effectiveness of laboratory service. Clinical audit should not be confused with the internal and external audit of the quality management system procedures of the laboratory (see Chapter 3 for details). Clinical audit measures the impact of laboratory testing on various outcomes, including health, organizational, and economic outcomes. A more general role for audit is that it can be used as part of the wider management exercise using key performance indicators for benchmarking performance in comparison with peers.

Clinical audit helps (1) solving problems associated with the clinical pathway or outcome of laboratory services delivered, (2) monitoring test utilization and controlling demand, (3) monitoring the introduction of a new test and/or changes in practice, (4) eliminating redundant tests, (5) monitoring variations between providers, and (6) compliance with best practice guidelines.

For example, the Royal College of Pathologists in the United Kingdom has various clinical biochemistry audit projects and templates that investigate the appropriate utilization of laboratory tests, such as tumor markers, progesterone in pregnancy, coagulation screen requests, d-dimer testing, or the implementation of laboratory medicine related practice guidelines (e.g., for investigation of ovarian cancer in primary care, or laboratory monitoring of diabetes mellitus), or how laboratories investigate hyponatremia and hypokalemia. The College’s website provides useful tools for education and training in clinical audit methods, including how to plan an audit, how to obtain and analyze audit data, and how to write a report. The College also provides a list of key performance indicators to be monitored. For more details the reader is

referred to the College's web site: <https://www.rcpath.org/profession/patient-safety-and-quality-improvement/conducting-a-clinical-audit.html> (last accessed May 28th, 2020).

Other approaches include surveying and monitoring test requesting patterns and case vignettes to explore how clinicians interpret laboratory data and what impact laboratory tests have on medical decisions.<sup>20–22</sup> Results of such analyses can be presented in many forms, including educational sessions, grand rounds, regular feedback to clinicians about their performance compared to peers, etc.<sup>23,24</sup> Such individually tailored clinical audit and feedback approaches could be useful in changing and rationalizing test utilization and contribute to improved effectiveness and cost-effectiveness of laboratory services, especially in areas where common practice significantly deviates from desired practice.

## EVALUATION OF DIAGNOSTIC TESTS

In the first section we have described the steps of EBLM. In this section we summarize how medical tests are typically evaluated, what kind of evidence is generated, and how this can help resolve clinical problems.

The typical journey of an in vitro *medical assay*, measuring a *biomarker*, in becoming a medically useful *test* or testing strategy that yields health benefits, is best described by a dynamic cycle of key elements that revolve around each other like interlocking cogwheels driven by the *clinical pathway* at their core (Fig. 10.3).

Key elements of the test evaluation cycle are defined in Table 10.4 and include (1) analytical performance, (2) clinical performance, (3) clinical effectiveness, (4) cost-effectiveness, and (5) overall impact of testing (ethical, social, psychological, societal, organizational, etc.). In this framework the *clinical purpose*, *role*, and the intended application of the biomarker

in a well-defined clinical pathway drive the other elements of the test evaluation cycle and dictate the study designs for providing the most relevant and highest level of evidence of the test's effectiveness.<sup>25</sup>

Chapter 2 gives further insight into some study design and methodological aspects (such as the “linked evidence approach”) that help assessing the effectiveness and impact of diagnostic tests on medical decisions. Due to the paucity of methodological standards and international agreement, it is unclear what evidence is sufficient for proving the clinical effectiveness of new medical tests. Below we describe a dynamic framework for the evaluation of diagnostic tests and describe in more detail the types of evidence that laboratory professionals should be able to generate and provide before new (and old) biomarkers are used in clinical practice.

### Types of Evidence

In most PICO-type questions, the O reflects the end-result of the clinical pathway and will refer to the health outcomes for individual patients. Yet, evidence that documents the effects of laboratory tests on patient outcomes may not always be available. That does not mean that other types of evidence may never be relevant to the clinical problem—to the contrary. For example, it could be very relevant for achieving better health outcomes if an analytically better performing, more selective screening test, that is less prone to interferences by various endogenous or exogenous metabolites, is used and which results in less false positive or false negative cases.

When considering clinical decisions about laboratory tests, different categories of evidence can be distinguished that relate to the key elements of the test evaluation cycle. Below we discuss the different types of evidence, and the typical sources of the corresponding evidence. Since patient outcomes and clinical effectiveness should guide our investigation choices, we will explain these key components in a patient- and society-centered order: Clinical effectiveness, Cost-effectiveness, Clinical Performance, Analytical Performance.

### Clinical Effectiveness

Clinical effectiveness refers to the ability of a test to improve health outcomes that are relevant to the individual patient (see Table 10.4). The most direct answer to the question whether testing improves patient-relevant outcomes will come from studies that have (1) included patients that fit the “P” description in the PICO, (2) compared the targeted intervention (I) against (3) the comparator (C), in terms of the (4) patient-relevant outcomes (O). The difference, or change, in outcomes between the group of patients undergoing testing (the intervention group) and the group undergoing the comparator (the comparator or control group) reflects the clinical effectiveness of testing: the extent to which the test is able to improve health outcomes.

The most valid results to demonstrate clinical effectiveness will come from pragmatic, randomized comparative trials. “Valid” here means that the findings from the study reflect the actual measure of interest: the change in outcomes in the targeted group of patients. The results will be “biased” (deviate systematically from the actual change) if the study was not designed optimally.



**FIGURE 10.3** Framework for the evaluation of in vitro medical tests. (From Horvath AR, et al. *Clin Chem Lab Med* 2015;53(6):841–48.) (Copyright by De Gruyter).

**TABLE 10.4 Types of Evidence About Medical Tests, Relevant for Clinical Decision-Making**

<b>Types of Evidence</b>	<b>Related Terms</b>	<b>Explanation (Reference)</b>	<b>Examples</b>
Analytical performance	Analytical validity Technical efficacy	Ability of an in vitro medical assay to conform to predefined quality specifications <sup>26</sup>	Universal definition of myocardial infarction recommends that hs-cTn assays must have acceptable imprecision, i.e., $\leq 10\%$ CV, at the 99th percentile of normal. <sup>27</sup> For example, in an analytical performance study, a hs-cTnT assay had a CV of 9% at 13.5 ng/L. <sup>28</sup>
Clinical performance	Clinical validity Test performance Performance or operating characteristics Test accuracy or diagnostic accuracy Diagnostic accuracy efficacy Test efficacy	Ability of a biomarker to conform to predefined clinical specifications in detecting patients with a particular clinical condition or in a physiologic state [adapted <sup>29,30</sup> ].	<i>Diagnostic test:</i> In patients with symptoms of dyspnea, 3 studies on the diagnostic accuracy of NT-proBNP in emergency care showed sensitivities of 74–98%, specificities of 47–93%, and area under the curve (AUC) values of 0.89–0.96. <sup>15</sup>
Clinical effectiveness	Clinical utility Clinical usefulness	Ability of a test to improve health outcomes that are relevant to the individual patient [adapted <sup>26,29,30</sup> ]	The clinical effectiveness of BNP testing for diagnosis of heart failure in patients presenting to emergency room with acute dyspnea were investigated in randomized clinical trials (RCTs) that compared the addition of BNP testing with standard investigations alone followed by routine care. A meta-analysis of these RCTs reported that addition of BNP testing decreased length of hospital stay by ~1 day; possibly reduced admission rates but did not affect 30-day mortality rates. <sup>31</sup>

hs-cTn, High sensitivity-cardiac troponin.

Table adapted from Horvath AR, et al. *Clin Chim Acta* 2014;427:49–57 (Copyright by Elsevier).

For the question whether screening with fecal hemoglobin testing reduces colorectal cancer mortality one can turn to published randomized screening trials. These were large studies, in which eligible subjects were invited to participate, and test positives were invited to colonoscopy, and survival times and causes of deaths registered.

One limitation in practicing EBLM is that the number of randomized trials documenting the effects of testing on patient outcome is very small. In one study, the researchers identified only slightly over 100 such trials in a 4-year period.<sup>32</sup>

The reasons for this scarcity of evidence are manifold. One is the indirect link between testing and patient outcomes, which explains the larger sample sizes—and higher costs—needed for trials of testing. Another is the difference in regulatory requirements. Unlike pharmaceuticals, in vitro diagnostics do not (yet) require evidence of effectiveness from trials before receiving approval for marketing. A third reason is cultural: traditionally, the emphasis in laboratory medicine has mostly been on the analytical measurements themselves, far less on the impact of the clinical performance of the test on medical decisions and patient outcomes.

### Cost-Effectiveness

One may also be interested to explore a wider range of outcomes, not just health outcomes, in comparisons of testing strategies. In addition to the effect on patient-relevant outcomes, decision-makers may also be interested in, for

example, the effects on resource use (cost-effectiveness or efficiency), or in the social, psychological, legal, and ethical implication, and other broader impacts of testing.

In economic evaluations, investigators document the effects of an intervention on health outcomes and on costs and assess whether the test provides value for money. Choices about relevant outcomes and relevant costs are guided by the perspective of the decision-maker. From an individual patient perspective, one only includes the outcomes for that patient, and the resource implications for that very same patient: out-of-pocket expenses, time, etc. For a third-party payer, the perspective leads to a different selection. That payer may be less interested in out-of-pocket costs for the patient. From a societal perspective, all health outcomes and all resources are relevant.

Health technology assessment (HTA) is a form of multidisciplinary research that typically aims at an even wider range of outcomes. HTA studies aim at the “systematic evaluation of the properties and effects of a health technology, addressing the direct and intended effects of this technology, as well as its indirect and unintended consequences, and aimed mainly at informed decision-making regarding health technologies.” (<http://www.inahta.org>; accessed May 28th, 2020.)

Typical for decision problems in laboratory medicine is the large number of potential strategies. With two tests, for example, one can perform one, the other, or both, and define conditional strategies: only perform the second in case of a

particular range of results on the first one, and vice versa. The combinatorial complexity becomes even larger if we realize that, on top of the strategy, action thresholds can be defined in many ways; that is, when will the test result, or combinations of results, start a downstream action?

This combinatorial complexity may be one of the other reasons why direct comparative studies of the clinical effectiveness of medical tests are rare. As an alternative, researchers have turned to modeling (Refer to Chapter 2 for more details). Using mathematical models, they try to estimate the effects on health outcomes from different testing strategies. Hypothetical cohorts of participants “flow” then through the model, and the effectiveness is estimated based on what “happens” to them in the model. Knowledge and understanding of the typical clinical pathway that follows the testing is essential for successfully applying such mathematical models to testing.

In colorectal cancer screening, for example, there have been comparative trials of fecal hemoglobin testing, of endoscopy, of CT-colonography and of some other screening modalities. But no trial has compared all possible screening modalities, let alone the full complexity of all the different age ranges for invitation, screening intervals, and test positivity thresholds for inviting participants for additional colonoscopies.

To provide evidence for decision makers, various groups have built mathematical models. These typically rely on a model of natural history: how polyps advance to cancer in some people, when and how cancer will lead to signs, symptoms, and disease, and how the development of concurrent morbidity means that many people will die from other causes. On top of that model they then build in the mechanisms of screening, estimating how many will participate if invited, and when and how cancer and precursor lesions will be detected.

### Clinical Performance

Although the number of comparative randomized trials of testing that aim to estimate clinical effectiveness is relatively small, the number of trials that document associations between test results and other clinical findings is much larger. In some cases, this type of evidence can also help the laboratory professional to address the clinical problem. Clinical performance of a test refers to the ability of a biomarker to conform to predefined clinical specifications in detecting patients with a particular clinical condition or in a physiologic state (see Table 10.4).

One example of such investigations is the diagnostic accuracy study in which the results from testing are compared against a clinical reference standard. The clinical reference standard is the best available method for establishing the presence, and absence, of the target disease.

In the question regarding NT-proBNP, one could ask “to what extent is NT-proBNP testing able to identify patients with heart failure?” Here the clinical reference standard can be a combination of two trained adjudicators. The results of this comparison can be expressed in statistics to express the clinical performance of the test; in this case it is the diagnostic accuracy. Based on the NT-proBNP measurement, we can make two types of errors: classifying test-positive patients incorrectly as having heart failure (false positives), or classifying test-negative patients incorrectly as not having heart failure (false negatives).

Since the consequences of the two types of errors are typically not comparable, diagnostic accuracy statistics maintain the distinction. A test’s sensitivity expresses to what extent the test is able to identify the ones with the target condition: it is estimated as the proportion of those with the target disease who test positive. With few false negatives, test sensitivity will be high. Sensitivity, as an expression of a test’s clinical performance, is sometimes referred to as “clinical sensitivity,” to distinguish it from analytical sensitivity. Similarly, test specificity expresses to what extent the test is able to detect those without the target disease. It is estimated as the proportion of those without the target disease who tested negative (see Chapter 2).

Other expressions of clinical performance focus on the information in a positive, or a negative test result. The positive predictive value indicates the proportion of those testing positive who actually have the target disease, and the negative predictive value refers to the proportion of those testing negative who did not have the target disease.

The likelihood ratio is another test result-oriented expression of test performance: it indicates how much more likely the result is in those with the target disease, compared to those without the target disease. In case of a dichotomized test results, the positive likelihood ratio—the likelihood ratio of a positive test result—is typically much bigger than unity, and the likelihood ratio for a negative test result much smaller. For more details and calculation of these diagnostic accuracy measures refer to Chapter 2.

Whereas clinical effectiveness studies document whether tests change patient outcomes, clinical performance studies do not: they only evaluate to what extent tests lead to a correct classification of diseased versus nondiseased. When then is clinical performance informative enough to remove the need for studies of clinical effectiveness?

In clinical performance studies, the downstream consequences of testing should guide the evaluation. One is rarely interested in all forms of disease, in all pathologic abnormalities, but only in those for which the management consequences are clear: the effective actions that can help the patient.

The “target condition” is the key term suggested to distinguish between meaningful and inconsequential forms of disease. The target condition represents the spectrum of disease for which there is agreement about the next-step actions. Fecal hemoglobin, for example, is not used in colorectal cancer screening to identify all adenoma, of any size and any pathology. To evaluate its clinical performance in screening, the appropriate definition of the target condition encompasses colorectal cancer and advanced adenomas (larger adenoma, adenoma with villous features or high-grade dysplasia, and sessile serrated polyps).

In some cases, a well-specified level of clinical performance can be sufficient. If a test is being proposed to rule out a condition, then studies of clinical performance can reveal whether it is fit for purpose. It has been suggested, for example, that NT-proBNP concentrations of less than 300 ng/L (36 pmol/L) can rule out acutely decompensated heart failure.<sup>4</sup> That will be the case if there are no or hardly any false negatives. Expressed in another way, this will be the case if the sensitivity is high enough or, alternatively, if the negative predictive value is close to hundred percent.

Another example is related to comparisons of tests. If one test is proposed as an alternative for another, then the test will

**TABLE 10.5 Methodological Tools for Practicing Evidence-Based Laboratory Medicine**

Purpose	Tool	Link (last accessed May 28th, 2020)
<b>Evidence Generation—Diagnostic and Prognostic Accuracy Studies</b>		
Reporting guidelines for diagnostic accuracy studies	STARD	<a href="https://www.equator-network.org/reporting-guidelines/stard">https://www.equator-network.org/reporting-guidelines/stard</a>
Reporting guidelines for multivariable prediction models	TRIPOD	<a href="https://www.tripod-statement.org">https://www.tripod-statement.org</a>
Reporting guidelines for tumor marker prognostic studies	REMARK	<a href="https://www.equator-network.org/reporting-guidelines/reporting-recommendations-for-tumour-marker-prognostic-studies-remark">https://www.equator-network.org/reporting-guidelines/reporting-recommendations-for-tumour-marker-prognostic-studies-remark</a>
<b>Evidence Synthesis and Summary—Systematic Reviews/Meta-Analysis/Guidelines</b>		
Manual for diagnostic accuracy systematic reviews	Cochrane DTA	<a href="https://methods.cochrane.org/sdt/handbook-dta-reviews">https://methods.cochrane.org/sdt/handbook-dta-reviews</a>
Critical appraisal of diagnostic accuracy studies for systematic reviews	QUADAS-2	<a href="https://www.bristol.ac.uk/population-health-sciences/projects/quadas">https://www.bristol.ac.uk/population-health-sciences/projects/quadas</a>
Reporting guidelines for systematic reviews	PRISMA	<a href="http://www.prisma-statement.org">http://www.prisma-statement.org</a>
Reporting guidelines for systematic reviews of diagnostic accuracy studies	PRISMA-DTA	<a href="http://prisma-statement.org/Extensions/DTA">http://prisma-statement.org/Extensions/DTA</a>
Critical appraisal of guidelines	AGREE	<a href="https://www.agreetrust.org">https://www.agreetrust.org</a>
Grading the strength of evidence and recommendations	GRADE	<a href="https://www.gradeworkinggroup.org">https://www.gradeworkinggroup.org</a>

serve its purpose if the clinical performance measures—that is, the sensitivity and specificity—are comparable. It does not necessarily mean that one test is superior to another if its level of clinical performance is higher. If a new test leads to a higher number of positives, then one also must find out that the management outcomes and the improvement in health outcomes for these additional positives are comparable.

For example, when a new generation of cTn assay was introduced with higher analytical sensitivity more patients were identified as being positive than with the previous generation cTn assays. The medical need for such high sensitivity assay was to detect ACS earlier and intervene faster and preserve more myocardial function. It was not demonstrated whether the new test helps to reduce patient mortality related to ACS/AMI, until sufficient data were gathered by using the new test in routine emergency care.

Diagnostic accuracy studies can also be at risk of bias. To reflect best the actual clinical performance, study participants must be sampled using a single set of eligibility criteria. With the clinical reference standard, these study participants are then classified as having the target disease (sometimes called “cases”) or not (“controls”). Everybody who enters the study can become a case, and all those who are noncases, automatically become controls.

If those with the target disease (cases) were sampled with one set of eligibility criteria (severe disease), and those without using a different set of eligibility criteria (e.g., healthy controls, such as first-year medical students), the estimates from that study are likely to lead to overestimation of diagnostic accuracy.

Using two different clinical reference standards for the same condition can also generate bias, even if each of the two reference standards seems valid. An example is the use of pathology for confirming acute appendicitis (first reference standard) and follow-up (second reference standard).

The QUADAS-2 tool has been specifically designed to help EBLM professionals evaluate the risk of bias and to identify reasons for concern about the applicability of study results

(Table 10.5).<sup>19</sup> The tool addresses four domains of a test accuracy study: patient selection, index test, reference standard, and flow and timing. The risk of bias is for each domain, and the first three domains are also assessed in terms of concerns regarding applicability.

Unfortunately, many primary test accuracy studies are poorly reported. Critical elements of study design are missing from the published study report, which makes it difficult for the reader to find out whether the study was free from deficiencies, or whether the study findings apply to the clinical question.

To help authors, editors, reviewers, and readers evaluate the completeness and transparency of reporting diagnostic accuracy studies, the STARD (Standards for Reporting Diagnostic accuracy studies) statement was developed (see Table 10.5). STARD is a one-page list of essential items, which can be used as a checklist. The original STARD statement provides a list of essential items for publishing abstracts of diagnostic accuracy studies, in order to improve reporting as well as capturing such studies by search engines.<sup>33</sup> Other reporting standards exist on the EQUATOR website for multivariable prediction models and for prognostic studies (see Table 10.5).

### Analytical Performance

Analytical performance refers to the ability of a laboratory assay to conform to predefined technical specifications (see Table 10.4). Measures of analytical performance include analytical sensitivity and specificity, limit of detection and quantitation, measurement range, linearity, metrological traceability, measurement accuracy (imprecision and trueness), and consideration of preanalytical variables including interferences and cross-reactions. For more details on each of these components the reader is referred to Chapters 5–7.

The analytical performance of laboratory assays may obviously have an impact on the clinical performance and clinical effectiveness of tests. Quality requirements for analytical performance of laboratory tests should ideally be dictated by clinical needs and the intended application, the clinical consequences and outcomes of testing.

Analytical performance criteria can be defined by using the so-called Milan hierarchy that proposes the following three models<sup>34–36</sup> for setting analytical performance specifications:

*Model 1.* Based on the effect of analytical performance on clinical outcomes

*Model 2.* Based on components of biological variation of the measurand

*Model 3.* Based on the state-of-the-art.

Within Model 1, one distinguishes between direct outcome studies, which investigate the impact of analytical performance of the test on clinical outcomes (Model 1a), and indirect outcome studies, which document the impact of analytical performance of the test on clinical classifications or decisions and thereby on the probability of patient outcomes (Model 1b).

The health outcome and clinical decision-driven analytical performance goals (i.e., Model 1) for cTn assays for diagnosing AMI are determined on the basis of how many diagnostic misclassifications of AMI are acceptable or tolerated by clinicians. A cTn assay, assuming to have zero bias, and a coefficient of variation (CV) of 10 or 6% at the 99th percentile decision limit can result in misclassification rates of 1 and 0.5%, respectively. Systematic errors in analysis can affect diagnostic accuracy even more.<sup>37</sup>

The diabetes mellitus guideline issued by the American Diabetes Association defines analytical performance goals for glucose measurements on the basis of biological variation (i.e., Model 2). To avoid misdiagnosis of patients, the goal for glucose analysis is to minimize total analytical error, and methods should be without measurable bias. This translates to goals for analytical imprecision of  $\leq 2.9\%$ , bias  $\leq 2.2\%$ , and a total error  $\leq 6.9\%$ .<sup>38</sup>

Analytical performance specifications of the same assay could be different for diagnostic or monitoring applications. For example, HbA<sub>1c</sub> is currently used for both diagnosis and monitoring of patients with diabetes mellitus. When a test is used for long-term monitoring assay precision particularly, and bias to a lesser degree is important as small changes of the analyte may be interpreted as clinically significant changes, with consequent management decisions of changing therapy. For HbA<sub>1c</sub> a generally accepted rule is that clinicians interpret an absolute difference of 0.5% (in NGSP units) or 5 mmol/mol (in IFCC units) between successive patient samples as a significant change in glycemic control. This translates to an analytical imprecision goal for monitoring purposes of  $\leq 2\%$ , which will produce a 95% probability that a difference of  $\geq 0.5\%$  HbA<sub>1c</sub> between successive patient samples is due to a significant change in glycemic control (when HbA<sub>1c</sub> is 7% [53 mmol/mol]).<sup>38,39</sup>

Conversely, when a test is used for diagnostic purposes with a predefined single cut-off value for dichotomizing patients, measurement bias is particularly important. It has been shown, for example, that 2% increase in concentration doubles the percentage of false positive diagnoses, both for HbA<sub>1c</sub> and cholesterol.<sup>40</sup> The clinical decision limit for diagnosing diabetes (48 mmol/mol in IFCC units, or 6.5% in NGSP units) is close to the upper limit of the nondiabetic reference interval (i.e., 42 mmol/mol or 6.0%), which means that a small systematic error in HbA<sub>1c</sub> measurement will have a large impact on interpretation and the prevalence of the condition.<sup>41</sup>

## POINTS TO REMEMBER

- Clinical effectiveness of testing refers to the extent to which the test is able to improve health outcomes.
- The purpose and role of testing in a clinical pathway define how tests should be evaluated for effectiveness.
- Case-control studies overestimate the diagnostic accuracy of tests.
- It is rare to find evidence that proves that testing improves patient outcomes directly.
- The best study design to investigate the effectiveness of testing is a diagnostic randomized clinical trial.
- Investigation of the clinical effectiveness of the test includes demonstration of the benefits and harms related to testing.

## Test Evaluation Framework

Now that we have discussed the key evidence types as also main building blocks of the test evaluation cycle (see Fig. 10.3), let us demonstrate how these concepts are linked together and adapted to the development and clinical validation of new biomarkers.<sup>25</sup>

Ideally, new biomarkers should be developed in response to unmet clinical needs, aimed at improving existing clinical care pathways. Troponins, for example, are considered as tests of myocardial damage, but clinicians have long been waiting for noninvasive markers that can predict myocardial infarction before cell damage happens. Circulating endothelial cells, shed from coronary arteries several days to weeks before heart attack, have been proposed as candidates for early prediction of myocardial infarction.

Proof-of-concept studies, usually of “case-control” design, explore the association of the disease or condition with the new potential biomarker. Such designs tend to overestimate the clinical performance of a diagnostic assay, which necessitates further investigations. Once a link is confirmed, the potential use of the new biomarker in the clinical pathway, including the purpose and role of testing, the definition of subsequent actions and the expected outcomes must be considered.

The intended application and expected outcomes of the new test should ideally dictate how good the assay’s analytical performance needs to be. This leads to the next step, that is, the investigation of the actual *analytical performance* of the assay. The intended use of the test and the analytical performance of the assay determine the *clinical performance* of the new biomarker. If the assay is intended to be used for diagnosis, clinical performance is best investigated in a diagnostic accuracy study and expressed as (changes in) diagnostic sensitivity and specificity or other accuracy statistics (see Chapter 2). If the assay is proposed as a prognostic marker, its performance is evaluated in an observational study, expressed as prognostic accuracy, or risk in terms of reclassification.

Investigation of the *clinical effectiveness* of the test includes demonstration of the benefits and harms related to testing to the individual patient, relative to current best practice. The randomized trial is the cornerstone of all evaluations of clinical effectiveness, but randomized clinical trial designs for medical testing are not always efficient, nor are they always needed.<sup>42</sup>

Necessary conditions for the intended application of the new biomarker in the clinical pathway (i.e., purpose and role of testing) in the targeted patient population, in terms of required levels of analytical and clinical performance, should definitely be demonstrated before a new biomarker can be safely released. The possibility of harms from testing should also be considered.

If the test is effective, estimation of *cost-effectiveness* will be key for reimbursement decisions. This may drive the need for comparative clinical studies to better estimate the size of effects, and for economic models to capture long-term consequences of testing and potential uncertainty for both costs and effects. Examples for the above components of the test evaluation cycle are illustrated in Table 10.4.

## EVIDENCE SYNTHESIS

### Systematic Reviews and Meta-Analysis

Searching systematically for the available evidence and critical appraisal of studies are performed by dedicated professionals trained in systematic review methods.

The defining features of systematic reviews include (1) a clear definition of the clinical question to be addressed, (2) an extensive and explicit strategy to find all studies (published or unpublished) that may be eligible for inclusion in the review, (3) criteria by which studies are included and excluded, (4) a mechanism to assess the risk of bias in each study and the information value for the review question, and, in some cases, (5) synthesis of results with the use of statistical techniques of meta-analysis. Systematic reviews thus differ from more traditional, narrative reviews of the evidence.

Meta-analysis is a statistical procedure for generating summary estimates, based on the body of evidence from multiple studies. Two forms of meta-analysis can be found in the literature. One form, called fixed effect meta-analysis, assumes that all available studies have estimated the same statistic, such as the effectiveness of fecal hemoglobin in colorectal screening, or the accuracy of BNP in detecting heart failure. A fixed effect meta-analysis can produce a more precise—and sometimes more valid—summary estimate of that statistic than a single study.

Another form of meta-analysis, called random effects meta-analysis, assumes that there may be many small differences between the different studies, and that it is not justified to assume that the effect is the same in every study. If one accepts this, the meta-analysis does not produce a single summary estimate of the effect, but a distribution, which is typically characterized through its mean and variance.

Systematic reviews can be performed for any type of study, and meta-analysis is possible if the studies result in a quantitative answer. Some tools and resources related to systematic reviews of diagnostic tests are provided in Table 10.5.

Hewitson and colleagues used the findings from four colorectal cancer screening trials with fecal occult blood testing, totaling the experience of over 320,000 participants. Combining the results indicated that screening had a 16% reduction in the relative risk of colorectal cancer mortality (95%-Confidence Interval 10% to 20%); which was 25% when adjusting for screening attendance in the individual studies. There was no difference in all-cause mortality.<sup>43</sup>

Roberts and colleagues summarized the findings from 48 evaluations of serum natriuretic peptide in people presenting with acute heart failure to acute care settings.<sup>44</sup> At the lower thresholds of 300 ng/L (36 pmol/L) for NT-proBNP the summary estimate of sensitivity was 0.99 (95%-CI: 0.97 to 1.00), for a negative predictive value of 0.98 (95%-CI: 0.89 to 1.0) for a diagnosis of acute heart failure.

### Guidelines

Guidelines can also provide a useful form of synthesized evidence. However, guidelines should be treated with some caution as many are not based on the type of rigorous evidence synthesis described above. Before relying on a guideline, you should check that the development process used explicit methods for identifying and assessing research evidence, that the guideline committee had appropriate membership with no or minimal conflicts of interest, and that the evidence is reasonably up to date (which relies on the search date, not the publication date of the guideline).

The AGREE checklist is widely used for critically appraising the methodological quality of guidelines (see Table 10.5). Numerous studies using this checklist have shown that the quality and content validity of guidelines are highly variable. This is particularly true for diagnostic recommendations. These shortcomings are due to the relative weakness of the evidence base of primary diagnostic studies, compared to therapeutic trials and to the heterogeneity of the analytical and clinical performance of various alternative laboratory methods for the same analyte covered in the recommendations.

Diagnostic recommendations are often incorporated in clinical guidelines and developed without the active involvement of laboratory professionals who could provide a deeper insight into the limitations of laboratory assays and a more informed interpretation of test results. The lack of methodological standards for test evaluation and for rating the evidence also contributes to the poor quality of diagnostic recommendations. Numerous evidence-rating systems exist for diagnostic testing, but none of these covers all aspects required for evidence gathering, review, assessment, and linkage to recommendations.<sup>45</sup>

Users of guidelines need to be aware that even when the same sources of evidence are used for making recommendations, sometimes contradictory advice is given by various guideline organizations. This is often due to the fact that value-based judgments on the balance between benefits, harms, risks, and patients' preferences—and the organizational and financial or resource consequences of care—may differ among countries and regions and therefore could influence the final recommendations and their grading.<sup>46</sup> These shortcomings and limitations call for a transparent, well-organized and documented guideline development process, the involvement of multidisciplinary stakeholders, and a public consultation phase before guidelines are released for clinical use.

Acknowledging these issues, numerous guideline organizations have developed standard operating procedures for practice guidelines that meet the AGREE and other contemporary methodological criteria.<sup>47</sup> While no internationally accepted standard exists for processes, the GRADE working group has standardized and developed methods for grading evidence and recommendations and issued a checklist to

assist guideline developers. Relevant tools and papers related to guideline development and grading recommendations are freely downloadable from the web sites listed in Table 10.5.

### POINTS TO REMEMBER

- The evaluation of biomarkers is a complex cyclical process driven by the purpose and role of testing in a clinical pathway.
- In evidence-based laboratory medicine the main types of evidence for assessing the value of testing relate to the analytical performance, clinical performance, clinical and cost-effectiveness, and broader impact of testing on patients and society.
- Systematic reviews of rigorous studies represent the highest level of evidence on the effectiveness of interventions.
- Meta-analysis is a statistical technique that generates summary estimates, based on the combined evidence from multiple studies.
- Practice guidelines are systematically developed statements that assist clinicians to deliver the most appropriate care to their patients.

### OTHER PURPOSES OF TESTING

Above we discussed how the principles of EBLM can be applied to tests used for diagnosis. Although most methods of EBM have been developed for diagnostic testing, many activities of medical laboratories relate to other purposes of testing, such as screening, prognosis, and monitoring. For tests used for purposes other than diagnosis, many of the principles are similar to those outlined above but with some important differences.

In this section we provide a brief overview of the key principles and methodological challenges related to the evaluation and use of screening and prognostic tests. We also offer a more extensive outline of the common but neglected area of monitoring tests.

#### Screening

The WHO defines screening as “the presumptive identification of unrecognized disease in an apparently healthy, asymptomatic population by means of tests, examinations or other procedures that can be applied rapidly and easily to the target population.” (<https://www.who.int/cancer/prevention/diagnosis-screening/screening/en/>, accessed May 28th, 2020.) Examples of large-scale population screening programs for preventable diseases are: screening for raised cholesterol, screening for diabetes, first trimester screening for Down syndrome, colorectal cancer screening by fecal occult blood testing, and neonatal screening for inborn errors of metabolism.

The principles of evaluation of screening are similar to those for diagnostic testing, except that a much higher emphasis is placed on the need for randomized trials to clearly prove that earlier detection is better than the usual clinical detection. Before such a trial is done, evidence is needed of test performance such as analytic and diagnostic accuracy. Because of the earlier stage of disease, the diagnostic sensitivity will generally be lower for screening than for diagnostic testing. Because screening focuses on asymptomatic

individuals, there is also a risk of overdiagnosis, as occurs with prostate-specific antigen (PSA) screening.

#### Prognosis

Many tests are used to help assess a patient's future risk, that is, their prognosis. For example, the level of BNP can help assess the future risks of recurrence of patients with heart failure, HIV viral load predicts progression to AIDS, and increased preoperative carcinoembryonic antigen concentration in resectable colorectal cancer is associated with poorer prognosis. When assessing the accuracy of a prognostic test, one can look at the extent to which the prognostic test reclassifies patients correctly as having—or not having—a future event. However, this is timeframe dependent. Hence a key difference between diagnostic evaluation and prognostic evaluation is the need for longer-term follow-up, and the need to include a specific timeframe.

Prognostic tests may also be predictive of treatment benefits. This prediction may simply be due to higher risk or to more severely affected patients having more potential to benefit. Sometimes the prognostic marker is also predictive of relative response; for example, in breast cancer estrogen receptor positivity predicts benefit from tamoxifen-based chemotherapy.

#### Monitoring

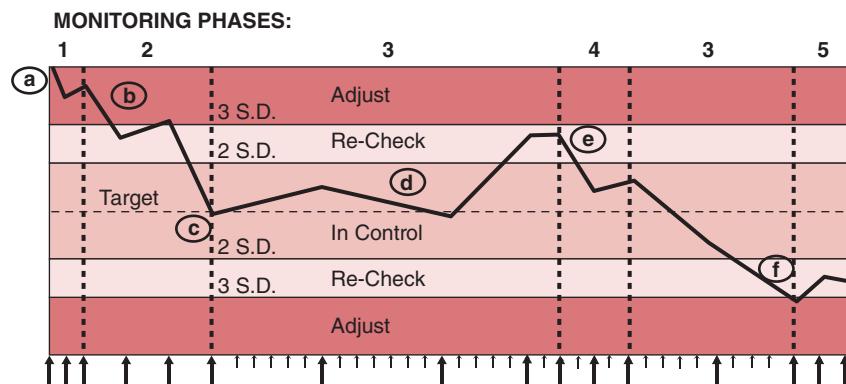
Monitoring is the use of periodic measurements to guide the management of a chronic or recurrent condition. Monitoring probably comprises between one third and one half of all tests requested in general practice and outpatients.

Despite the considerable medical time and resource involved, the principles and strategies for optimal monitoring—choice of test, intervals, and interpretation—have only recently been examined by researchers and methodologists. The objective and methods of monitoring change over the course of treatment. This course of monitoring can be usefully divided into the five phases shown in Fig. 10.4, which also illustrates a control chart for these stages.<sup>48</sup>

At (a) we first note the abnormal measurement and begin a quick series of pretreatment measurements to confirm the abnormal result; (b) then, if appropriate, initiate treatment and monitor at short intervals to check response and achieve control; (c) but once control is achieved, then (d) the intervals may be longer, although they may be supplemented by patient self-monitoring (small arrows); (e) when one measurement is more than three standard deviations (SD) or two measurements are more than two SDs from target, we adjust therapy to re-establish control, and shorten the re-check interval; and finally (f) if treatment becomes unnecessary, a period of postcessation monitoring may be required. We now look at these phases in more detail. In brief, the five phases are listed below.

#### Pretreatment Monitoring

Some monitoring is generally needed before we start treatment. This firm baseline is needed to confirm that the degree of abnormality is beyond the initiation threshold. Many initially abnormal measurements may “normalize” before treatment for several reasons, such as training effects (e.g., with peak flow meter), accommodation to measurement (e.g., with blood pressure measurement), and, perhaps most importantly, regression to the mean (the tendency of repeat tests to be closer to normal).



**FIGURE 10.4** The five phases of treatment monitoring. Large arrows are clinician measurements; small arrows are patient measurements. (For an explanation of the numbers see text.)

### Initial Titration: Establishing Response, Control, and Safety

The initial titration phase has several aims which include: (1) checking the individual's response to treatment, (2) detecting unacceptable adverse effects, and (3) achieving the desired target range.

### Monitoring During Treatment

Once a patient's measurements are within the target range, the objective of monitoring is to ensure that measurements stay within reasonable limits, called "control limits," which are set to ensure that we detect real changes in level while minimizing false positives that are due to short-term measurement variability or technical measurement error. Monitoring is usually much less frequent than during the titration phase.

There are several possible rules to detect out of range tests; for example, one approach illustrated in Fig. 10.4 is to consider that a shift from control has occurred if: (1) a single measurement is outside a 3SD upper and lower control limit, or (2) two or three successive measurements are more than 2SD from the target. Fig. 10.4 shows these two sets of action thresholds; one for action ( $\pm 3\text{SD}$ ) and one for re-measurement ( $\pm 2\text{SD}$ ), with action if the repeat result is also more than 2SD from target.

### Adjustment to Re-Establish Control

If we detect a clear drift beyond the control limits, then adjustment to management is made to re-establish control. As in the titration phase, a shorter measurement interval is generally warranted until control is re-established.

### Cessation of Treatment

As most therapies are not life-long, guidance and down-titration processes are also needed for cessation of treatments. A decision to stop based on current risks and control is made, treatment is withdrawn (perhaps in stages), and, after a "wash out" period, the patient is rechecked to ensure that treatment does not need to be restarted.

The optimal choice of monitoring tests may sometimes change with different phases, but it is usually simpler to stick to a single test. The choice of test is guided by four principles: (1) the *clinical validity*, which is the ability of the test to predict the clinically relevant outcome that we are trying to control or prevent, (2) the *responsiveness*, which is how much

the test changes in response to an intervention relative to background random variation ("signal-to-noise" ratio), (3) the *detectability* of long-term change describes the size of changes in the test over the long term relative to background random variation, and (4) the *practicality*, including the ease of use, invasiveness, and cost of the test.<sup>49</sup>

### FROM EVIDENCE TO ACTION

We have already defined EBLM as a tool that assists clinical management of patients by integrating into clinical decision making the best available research evidence for the use of laboratory investigations with the clinical expertise of the physician and the needs, expectations and concerns of the patients, in order to improve the care and outcomes of individual patients and the effective use of health care resources.<sup>24</sup>

How can these aims be achieved and what is the role of laboratory professionals in getting evidence into practice? Responsibilities of the profession are summed up by Muir Gray, author of the highly acclaimed book entitled *Evidence-Based Health Care (EBHC)*:<sup>50</sup> (1) eliminate poor or useless tests before they become widely available, that is, *stop starting*, (2) remove old tests with no proven benefit or, in fact, harm from the laboratory's repertoire, that is, *start stopping*, and (3) introduce new tests if evidence proves their efficacy and effectiveness, that is, *start starting or stop stopping*.

While the aims are clear and laboratory professionals, clinicians, patients, policymakers, and industry generally endorse such principles, practice data often show the opposite or, at best, wide variations in test utilization. The widening gap between evidence and practice leads to various undesirable scenarios, such as underutilization/overutilization or inappropriate utilization of medical tests that may contribute to underdiagnosis, overdiagnosis, or misdiagnosis of patients with potentially harmful health and other consequences to patients and society.<sup>24</sup>

### Evidence-Based Laboratory Medicine in Everyday Life

For the evidence-based laboratory scientist trying to keep up to date, there are four activities to consider.

1. *Keep a logbook of questions.* We often have questions that arise in consultations with colleagues, discussions with co-workers, from our reading or daily review of laboratory results. A good practice is to keep a simple "logbook" (paper or electronic) of these questions and try to answer a few.

2. For some of these questions, do an explicit literature search for the best evidence to answer this question (we suggest starting with just 1 a week, and doing more as you get faster). This answer might be through an evidence resource such as UpToDate, or a systematic review, or trying to find primary research using the “Diagnostic” filter in PubMed: ClinicalQueries. You should then appraise the key article and keep a record of the conclusions (or you will forget 6 months later when you need it again).
3. Use a literature scanning service, to keep up with important new and valid research. You can try to scan the primary literature yourself, but most of the primary research is not sufficiently valid to be worth changing practice, so the “number needed to read” is usually very large—you might share the load with colleagues.
4. Run a “journal club,” which focuses on some of the best research from the above three processes. We suggest you ask attendees to vote on which are most important to discuss (which helps attendance also). Again, you need to keep a record of the conclusions, and the clinical bottom line. But you should also record any actions that might result, such as a change to local testing guidelines, a change in reporting, the addition of a new test (or deletion of a poor test), etc.

### Implementing Evidence-Based Laboratory Medicine

Many or even most laboratory professionals may not undertake the above processes for themselves, so we need to supplement this with explicit processes to improve the uptake of best practices based on best research. Such evidence implementation is a complex and evolving field, as “leaks” in the process of evidence uptake can occur at many stages as illustrated in Fig. 10.5. Even when good studies, systematic reviews, or guidelines are available their wide scatter in the literature means that practitioners are not aware of them. Even if aware, conflicting information may mean they do not accept the evidence, and even if they accept it, adoption may be complicated by the learning curve, costs, or other barriers.

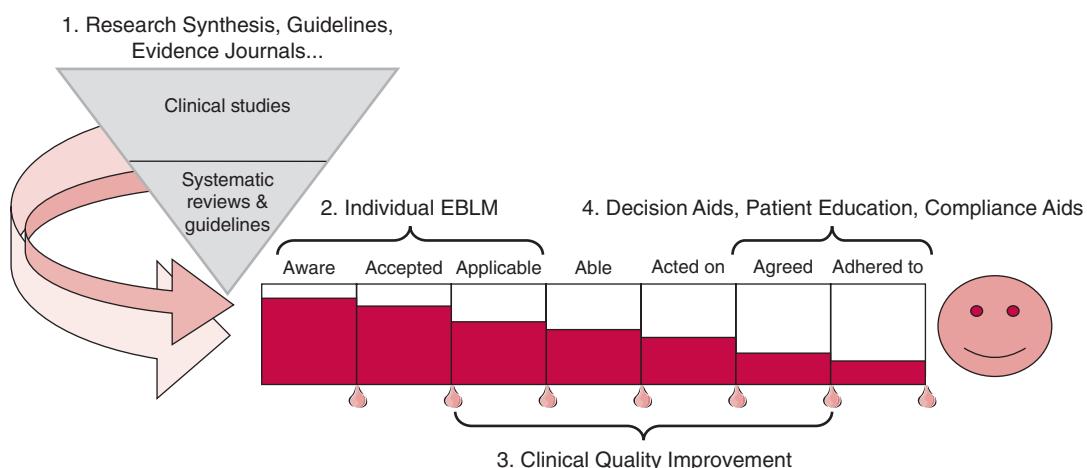
Different strategies are needed for each of these different stages and barriers. Hence, the initial phase of any implementation is to assess the degree of the evidence–practice gap, and the likely barriers to uptake.

There are no “magic bullets” in this process, but some methods shown to have an impact are the use of opinion leaders, small group peer discussion of the evidence, and audit and feedback processes, which can be used in sequence or combination. For laboratory testing, this implementation may be required for both laboratory staff and for the clinicians who use the laboratory.

### The Emerging Problem of “Overdiagnosis”

An increasingly recognized downside of testing is the problem of overdiagnosis: the detection of “abnormalities,” which would never have caused patients any morbidity or mortality.<sup>51</sup> For example, PSA screening may lower prostate cancer mortality, but leads to the detection of an excess of cases who may never have developed clinical prostate cancer.<sup>52</sup> In the 13-year follow-up of the European trial, the cumulative incidence was 6.8% of men in the control group but 10.2% in the PSA screening group: an absolute excess of 3.4%, or a 50% relative increase.<sup>53</sup>

A similar problem arises from the threshold and interpretations of many laboratory tests. One example is the controversy of the appropriate definition of gestational diabetes mellitus: the International Association of the Diabetes and Pregnancy Study Group criteria lead to a higher prevalence than either the older World Health Organization definition or the more recent National Institute for Health and Care Excellence (NICE) definition.<sup>54</sup> A National Institutes of Health consensus conference concluded that there was no clear evidence that the additional cases identified have a net benefit from detection, but the debate rages on. Similar, though less controversial, problems have occurred with testing of vitamin D, cTn, hemoglobin, and even cholesterol. A risk in shifting thresholds for defining diseases is that small changes on the steep part of the normal distribution can mean a large proportion of the population is redefined as



**FIGURE 10.5** The evidence-to-practice pipeline. Transfer of evidence from (1) research production to (2) awareness and acceptance by individual professionals, (3) routine and audited usage, and (4) impact on patients involves several stages, with barriers and losses at each stage resulting in a substantial cumulative “leakage.” *EBLM*, Evidence-based laboratory medicine.

diseased. International standards are needed for when and how to make changes in thresholds and definitions, but do not exist at the time of writing.

## THE HISTORY AND THE FUTURE OF EBLM

EBCM did not come out of thin air. It followed a number of developments in clinical medicine. Below we describe the background and the possible future of EBCM.

### The Development of Evidence-Based Medicine

In the late 1980s, the term “evidence-based medicine” was introduced by an Evidence-Based Medicine Working Group centered in McMaster University, Hamilton, Canada, as a new paradigm for teaching clinical medicine. It de-emphasized traditional authority and, instead, focused on the results from sound scientific research to guide clinical decision-making about the care of patients.<sup>1</sup>

The term first appeared in an editorial in 1991 by the chair of this group, Gordon Guyatt,<sup>55</sup> and the Working Group subsequently went on to produce a portfolio of papers under the title, “Users’ Guides to the Medical Literature” advising clinicians on how to assess information in clinical journals which then can be used as evidence for making more objective clinical decisions.<sup>56</sup>

Out of these discussions came the term critical appraisal of the literature; later, the idea of bringing “critical appraisal to the bedside” was born. Implicit in this thinking was the need for generating good quality evidence and the ability to appraise, synthesize, and present that evidence and to determine whether it was applicable to the problem at hand and the decision(s) to be made. From teaching medicine, the notion that practice should be based on solid evidence permeated to all areas of health care and beyond.

Originally, the label EBM was used to refer to this specific form of practicing medicine. From there, the term rapidly expanded to various subspecialties in medicine, including EBCM, and to other areas of health care such as health care management, practice guidelines, and policy applied to populations, yielding the broader terms of “evidence-based practice” and “evidence-based health care.”

David Sackett and colleagues have defined EBM in terms of “the integration of best research evidence with clinical expertise and patient values.”<sup>57</sup> A key objective of EBM, in its original context, has been “to incorporate the best evidence from clinical research into clinical decisions” to provide care that would lead to the best possible outcomes for the *individual patient*. From this it follows that EBM is a decision support tool to primarily improve health outcomes for the patient.

In the delivery of health care services, it is increasingly important to also consider the operational, economic, and societal impacts of various health care interventions.<sup>2</sup> Therefore the definition of EBM has been broadened to capture these *population*-based concepts of EBP and EBHC. Recognizing the fact that the concept of EBM can be applied to populations as well as individuals, Eddy has offered a broader definition: “EBM is a set of principles and methods intended to ensure that to the greatest extent possible, medical decisions, guidelines, and other types of policies are based on and consistent with good evidence of effectiveness and benefit.”<sup>58</sup>

Many books have now been written on the concepts, teaching, and practice of EBM, which illustrate its origins in

clinical epidemiology, as well as the challenges faced in adopting this approach to the practice of various medical disciplines, including laboratory medicine, and to health care management and policymaking.

### The Future of Evidence-Based Laboratory Medicine

EBCM was primarily developed with general internal medicine, and the first few applications targeted pharmaceutical interventions. For this type of clinical problems, there typically was an abundance of evidence, as most Western countries require sound evidence from strong research to document the effectiveness and safety of new drugs. Large-scale trials in thousands of patients are no exception.

In contrast, the application of the principles of EBM to problems related to medical testing developed more slowly. Unlike drugs, most laboratory tests do not require a large body of evidence about the effect on patient outcomes—in terms of effectiveness and safety—before they can be marketed. As discussed before, current regulation requires the documentation of analytical performance and clinical performance of tests (clinical evidence).

This makes the application of EBM to test-related problems quite a challenge. In the absence of direct evidence of the effect of testing on patient outcomes, EBCM professionals have to rely on other performance measures, or use linked evidence approaches, or modeling (see Chapter 2).

With linked evidence, decision makers piece together evidence from multiple sources. An example could be evidence of a test’s clinical performance (e.g., its ability to detect a specific target condition) and evidence of the effectiveness of treatment (e.g., the change in outcomes from treatment of patients with the very same target condition).

Gradually, groups are questioning the absence of evidence of clinical effectiveness from medical testing. Testing can consume a considerable amount of health care resources, and why should society pay for testing, in the absence of health benefits? If tests have no consequences, if the information generated by them cannot help patients, then it should be discouraged.

As discussed earlier, overtesting and overdiagnosis are also points of discussion. Information from testing can be consequential, but in a negative sense. Test results can start ineffective or even harmful treatment and can cause unnecessary anxiety and distress.

As a result of this, insurance companies, registration authorities and guideline panels now increasingly voice the need for more and better data about the effects of testing. There is increasing regulatory and financial pressure for more evidence to prove that testing and test-treatment interventions are effective and improve patient outcomes. Provision of such information is becoming particularly important in decisions about the market entry, clinical use, and coverage of novel biomarkers (for details see Chapter 12).

In the era of “omics,” nanotechnology, drug discovery, and personalized medicine new potential biomarkers are emerging and “companion diagnostics” are being developed in conjunction with, and for, the safe and effective administration of new medications. Increasing pressure from patients, clinicians, society, policy makers, and regulators for increased effectiveness and efficiency change the landscape of health care, and within that, laboratory medicine, rapidly. All of this will also affect the ways laboratory professionals will practice.

Their role is not just the daily provision of in vitro medical test results in a clinically meaningful way to clinical customers. Laboratory professionals will engage more and more in the evaluation of new, or new versions of old, biomarker assays, and take part in the research and development of emerging biomarkers or testing strategies. Practicing EBLM and using the skills and tools of EBM will assist laboratories in providing value and not just data to their customers.

## SELECTED REFERENCES

2. Price CP, Christenson RH, American Association for Clinical Chemistry. Evidence-based laboratory medicine: principles, practice, and outcomes. 2nd ed. Washington, DC: AACC Press 2007.
3. Glasziou P, Aronson JK, Irwig L. Evidence-based medical monitoring: from principles to practice. Malden, Mass.; Oxford: Blackwell Pub./BMJ Books 2008.
5. Richardson WS, Wilson MC, Nishikawa J, et al. The well-built clinical question: a key to evidence-based decisions. *ACP Journal Club* 1995;123(3):A12-A13.
9. Bossuyt PM, Irwig L, Craig J, et al. Comparative accuracy: assessing new tests against existing diagnostic pathways. *BMJ* 2006;332(7549):1089-92. doi: [332/7549/1089 \[pii\]](https://doi.org/10.1136/bmj.332.7549.1089) 10.1136/bmj.332.7549.1089
16. Haynes RB. Of studies, syntheses, synopses, summaries, and systems: the “5S” evolution of information services for evidence-based health care decisions. *ACP Journal Club* 2006;145(3):A8.
18. Klovning A, Sandberg S. Searching the literature. In: Price CP, Christenson RH, eds. Evidence-based laboratory medicine: principles, practice and outcomes. Washington: AACC Press. 2007:189-212.
19. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011;155(8):529-36. doi: [10.7326/0003-4819-155-8-201110180-00009](https://doi.org/10.7326/0003-4819-155-8-201110180-00009).
24. Horvath AR. From evidence to best practice in laboratory medicine. *Clin Biochem Rev* 2013;34(2):47-60.
25. Horvath AR, Lord SJ, StJohn A, et al. From biomarkers to medical tests: the changing landscape of test evaluation. *Clin Chim Acta* 2014;427:49-57. doi: [10.1016/j.cca.2013.09.018](https://doi.org/10.1016/j.cca.2013.09.018)
32. Ferrante di Ruffano L, Davenport C, Eisinga A, et al. A capture-recapture analysis demonstrated that randomized controlled trials evaluating the impact of diagnostic tests on patient outcomes are rare. *J Clin Epidemiol* 2012;65(3):282-7. doi: [10.1016/j.jclinepi.2011.07.003](https://doi.org/10.1016/j.jclinepi.2011.07.003)
33. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: An Updated List of Essential Items for Reporting Diagnostic Accuracy Studies. *Clin Chem* 2015;61(12):1446-52. doi: [10.1373/clinchem.2015.246280](https://doi.org/10.1373/clinchem.2015.246280)
35. Sandberg S, Fraser CG, Horvath AR, et al. Defining analytical performance specifications: Consensus Statement from the 1st Strategic Conference of the European Federation of Clinical Chemistry and Laboratory Medicine. *Clin Chem Lab Med* 2015;53(6):833-5. doi: [10.1515/cclm-2015-0067](https://doi.org/10.1515/cclm-2015-0067)
42. Bossuyt PM, Lijmer JG, Mol BW. Randomised comparisons of medical tests: sometimes invalid, not always efficient. *Lancet* 2000;356(9244):1844-7. doi: [S0140-6736\(00\)03246-3 \[pii\]](https://doi.org/10.1016/S0140-6736(00)03246-3) 10.1016/S0140-6736(00)03246-3
45. Gopalakrishna G, Langendam MW, Scholten RJ, et al. Guidelines for guideline developers: a systematic review of grading systems for medical tests. *Implement Sci* 2013;8:78. doi: [10.1186/1748-5908-8-78](https://doi.org/10.1186/1748-5908-8-78)
51. Moynihan R, Doust J, Henry D. Preventing overdiagnosis: how to stop harming the healthy. *BMJ* 2012;344:e3502. doi: [10.1136/bmj.e3502](https://doi.org/10.1136/bmj.e3502)
50. Muir Gray JA. Evidence-based healthcare: how to make decisions about health services and public health. Churchill Livingstone 2008.
56. Guyatt GH. Evidence-based medicine. *ACP Journal Club* 1991;114:A16.
57. Guyatt GH, Rennie D. Users’ guides to the medical literature. *JAMA* 1993;270(17):2096-7.
58. Eddy DM. Evidence-based medicine: a unified approach. *Health Aff (Millwood)* 2005;24(1):9-17. doi: [10.1377/hlthaff.24.1.9](https://doi.org/10.1377/hlthaff.24.1.9)

## REFERENCES

1. Evidence-Based Medicine Working Group. Evidence-based medicine. A new approach to teaching the practice of medicine. *JAMA* 1992;268(17):2420-5.
- 2a. Straus SE, Glasziou P, Richardson SW, et al. Evidence-based medicine : how to practice and teach EBM. 5th ed. Edinburgh: Elsevier 2018.
- 1b. Horvath AR. Evidence based laboratory medicine for beginners: getting evidence into practice. *Pathology* 2013;45 doi: [10.1097/01.PAT.0000426765.04291.44](https://doi.org/10.1097/01.PAT.0000426765.04291.44).
2. Price CP, Christenson RH, American Association for Clinical Chemistry. Evidence-based laboratory medicine : principles, practice, and outcomes. 2nd ed. Washington, DC: AACC Press 2007.
3. Glasziou P, Aronson JK, Irwig L. Evidence-based medical monitoring : from principles to practice. Malden, Mass.; Oxford: Blackwell Pub./BMJ Books 2008.
4. Kim HN, Januzzi JL, Jr. Natriuretic peptide testing in heart failure. *Circulation* 2011;123(18):2015-9. doi: [10.1161/CIRCULATIONAHA.110.979500](https://doi.org/10.1161/CIRCULATIONAHA.110.979500)
5. Richardson WS, Wilson MC, Nishikawa J, et al. The well-built clinical question: a key to evidence-based decisions *acp journal club* 1995;123(3):A12-A13.
6. BIPM I, IFCC, ILAC, IUPAC, IUPAP, ISO, OIML. The international vocabulary of metrology—basic and general concepts and associated terms (VIM). JCGM. 3rd ed, 2012.
7. Biomarkers Definitions Working G. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin Pharmacol Ther* 2001;69(3):89-95. doi: [10.1067/mcp.2001.113989](https://doi.org/10.1067/mcp.2001.113989)
8. Kinsman L, Rotter T, James E, et al. What is a clinical pathway? Development of a definition to inform the debate. *BMC Med* 2010;8:31. doi: [10.1186/1741-7015-8-31](https://doi.org/10.1186/1741-7015-8-31)
9. Bossuyt PM, Irwig L, Craig J, et al. Comparative accuracy: assessing new tests against existing diagnostic pathways. *BMJ* 2006;332(7549):1089-92. doi: [332/7549/1089 \[pii\] 10.1136/bmj.332.7549.1089](https://doi.org/10.1136/bmj.332.7549.1089)
10. Bay M, Kirk V, Parner J, et al. NT-proBNP: a new diagnostic screening tool to differentiate between patients with normal and reduced left ventricular systolic function. *Heart* 2003; 89(2):150-4.
11. Mant J, Doust J, Roalfe A, et al. Systematic review and individual patient data meta-analysis of diagnosis of heart failure, with modelling of implications of different diagnostic strategies in primary care. *Health Technol Assess* 2009;13(32):1-207, iii. doi: [10.3310/hta13320](https://doi.org/10.3310/hta13320)
12. Flint KM, Allen LA, Pham M, et al. B-type natriuretic peptide predicts 30-day readmission for heart failure but not readmission for other causes. *J Am Heart Assoc* 2014;3(3):e000806. doi: [10.1161/JAHA.114.000806](https://doi.org/10.1161/JAHA.114.000806)
13. Savarese G, Trimarco B, Dellegrottaglie S, et al. Natriuretic peptide-guided therapy in chronic heart failure: a meta-analysis of 2,686 patients in 12 randomized trials. *PLoS One* 2013;8(3):e58287. doi: [10.1371/journal.pone.0058287](https://doi.org/10.1371/journal.pone.0058287)
14. Don-Wauchope AC, McKelvie RS. Evidence based application of BNP/NT-proBNP testing in heart failure. *Clin Biochem* 2015;48(4-5):236-46. doi: [10.1016/j.clinbiochem.2014.11.002](https://doi.org/10.1016/j.clinbiochem.2014.11.002)
15. Balion C, Santaguida PL, Hill S, et al. Testing for BNP and NT-proBNP in the diagnosis and prognosis of heart failure. *Evid Rep Technol Assess (Full Rep)* 2006(142):1-147.
16. Haynes RB. Of studies, syntheses, synopses, summaries, and systems: the “5S” evolution of information services for evidence-based health care decisions. *acp journal club* 2006;145(3):A8.
17. Savel TG, Lee BA, Ledbetter G, et al. PTT Advisor: A CDC-supported initiative to develop a mobile clinical laboratory decision support application for the iOS platform. *Online J Public Health Inform* 2013;5(2):215. doi: [10.5210/ojphi.v5i2.4363](https://doi.org/10.5210/ojphi.v5i2.4363)
18. Kloving A, Sandberg S. Searching the literature. In: Price CP, Christenson RH, eds. Evidence-based laboratory medicine: principles, practice and outcomes. Washington: AACC Press. 2007:189-212.
19. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011;155(8):529-36. doi: [10.7326/0003-4819-155-8-201110180-00009](https://doi.org/10.7326/0003-4819-155-8-201110180-00009)
20. Skeie S, Perich C, Ricos C, et al. Postanalytical external quality assessment of blood glucose and hemoglobin A1c: an international survey. *Clin Chem* 2005;51(7):1145-53. doi: [10.1373/clinchem.2005.048488](https://doi.org/10.1373/clinchem.2005.048488)
21. Aakre KM, Thue G, Subramaniam-Haavik S, et al. Postanalytical external quality assessment of urine albumin in primary health care: an international survey. *Clin Chem* 2008;54(10): 1630-6. doi: [10.1373/clinchem.2007.100917](https://doi.org/10.1373/clinchem.2007.100917)
22. Kristoffersen AH, Thue G, Ajzner E, et al. Interpretation and management of INR results: a case history based survey in 13 countries. *Thromb Res* 2012;130(3):309-15. doi: [10.1016/j.thromres.2012.02.014](https://doi.org/10.1016/j.thromres.2012.02.014)
23. Willis EA, Datta BN. Effect of an educational intervention on requesting behaviour by a medical admission unit. *Ann Clin Biochem* 2013;50(Pt 2):166-8. doi: [10.1258/acb.2012.012100](https://doi.org/10.1258/acb.2012.012100)
24. Horvath AR. From evidence to best practice in laboratory medicine. *Clin Biochem Rev* 2013;34(2):47-60.
25. Horvath AR, Lord SJ, StJohn A, et al. From biomarkers to medical tests: the changing landscape of test evaluation. *Clinica chimica acta; international journal of clinical chemistry* 2014;427:49-57. doi: [10.1016/j.cca.2013.09.018](https://doi.org/10.1016/j.cca.2013.09.018)
26. Methods Guide for Medical Test Reviews. In: Chang SM, Matchar DB, Smetana GW, et al., eds. AHRQ Publication No 12-EC017 Rockville (MD): Agency for Healthcare Research and Quality (US), 2012.
27. Thygesen K, Alpert JS, Jaffe AS, et al. Third universal definition of myocardial infarction. *Eur Heart J* 2012;33(20):2551-67. doi: [10.1093/euroheartj/ehs184](https://doi.org/10.1093/euroheartj/ehs184)
28. Giannitsis E, Kurz K, Hallermayer K, et al. Analytical validation of a high-sensitivity cardiac troponin T assay. *Clin Chem* 2010;56(2):254-61. doi: [10.1373/clinchem.2009.132654](https://doi.org/10.1373/clinchem.2009.132654)
29. Study Group 5 of the Global Harmonization Task Force. Clinical evidence for IVD medical devices — key definitions and concepts. In: Global Harmonization Task Force, ed. GHTF/SG5/N6; 2012 1–11 2012.
30. Sanderson S, Zimmern R, Kroese M, et al. How can the evaluation of genetic tests be enhanced? Lessons learned from the ACCE framework and evaluating genetic tests in the United Kingdom. *Genet Med* 2005;7(7):495-500. doi: [00125817-200509000-00005 \[pii\]](https://doi.org/10.1089/glm.200509000-00005)
31. Lam LL, Cameron PA, Schneider HG, et al. Meta-analysis: effect of B-type natriuretic peptide testing on clinical outcomes in patients with acute dyspnea in the emergency setting. *Ann Intern Med* 2010;153(11):728-35. doi: [10.7326/0003-4819-153-11-201012070-00006](https://doi.org/10.7326/0003-4819-153-11-201012070-00006)

32. Ferrante di Ruffano L, Davenport C, Eisinga A, et al. A capture-recapture analysis demonstrated that randomized controlled trials evaluating the impact of diagnostic tests on patient outcomes are rare. *J Clin Epidemiol* 2012;65(3):282-7. doi: [10.1016/j.jclinepi.2011.07.003](https://doi.org/10.1016/j.jclinepi.2011.07.003)
33. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Standards for Reporting of Diagnostic Accuracy. *Clin Chem* 2003;49(1):1-6.
34. Panteghini M, Sandberg S. Defining analytical performance specifications 15 years after the Stockholm conference. *Clin Chem Lab Med* 2015;53(6):829-32. doi: [10.1515/cclm-2015-0303](https://doi.org/10.1515/cclm-2015-0303)
35. Sandberg S, Fraser CG, Horvath AR, et al. Defining analytical performance specifications: Consensus Statement from the 1st Strategic Conference of the European Federation of Clinical Chemistry and Laboratory Medicine. *Clin Chem Lab Med* 2015;53(6):833-5. doi: [10.1515/cclm-2015-0067](https://doi.org/10.1515/cclm-2015-0067)
36. Fraser CG, Kallner A, Kenny D, et al. Introduction: strategies to set global quality specifications in laboratory medicine. *Scand J Clin Lab Invest* 1999;59(7):477-8.
37. Sheehan P, Blennerhassett J, Vasikaran SD. Decision limit for troponin I and assay performance. *Ann Clin Biochem* 2002;39(Pt 3):231-6.
38. Sacks DB, Arnold M, Bakris GL, et al. Guidelines and recommendations for laboratory analysis in the diagnosis and management of diabetes mellitus. *Clin Chem* 2011;57(6):e1-e47. doi: [10.1373/clinchem.2010.161596](https://doi.org/10.1373/clinchem.2010.161596)
39. Weykamp C. HbA1c: a review of analytical and clinical aspects. *Ann Lab Med* 2013;33(6):393-400. doi: [10.3343/alm.2013.33.6.393](https://doi.org/10.3343/alm.2013.33.6.393)
40. Hyltoft Petersen P, Klee GG. Influence of analytical bias and imprecision on the number of false positive results using Guideline-Driven Medical Decision Limits. *Clin Chim Acta* 2014;430:1-8. doi: [10.1016/j.cca.2013.12.014](https://doi.org/10.1016/j.cca.2013.12.014)
41. Weykamp C, John G, Gillery P, et al. Investigation of 2 models to set and evaluate quality targets for hb a1c: biological variation and sigma-metrics. *Clin Chem* 2015;61(5):752-9. doi: [10.1373/clinchem.2014.235333](https://doi.org/10.1373/clinchem.2014.235333)
42. Bossuyt PM, Lijmer JG, Mol BW. Randomised comparisons of medical tests: sometimes invalid, not always efficient. *Lancet* 2000;356(9244):1844-7. doi: [S0140-6736\(00\)03246-3 \[pii\] 10.1016/S0140-6736\(00\)03246-3](https://doi.org/10.1016/S0140-6736(00)03246-3)
43. Hewitson P, Glasziou P, Watson E, et al. Cochrane systematic review of colorectal cancer screening using the fecal occult blood test (hemoccult): an update. *Am J Gastroenterol* 2008;103(6):1541-9. doi: [10.1111/j.1572-0241.2008.01875.x](https://doi.org/10.1111/j.1572-0241.2008.01875.x)
44. Roberts E, Ludman AJ, Dworzynski K, et al. The diagnostic accuracy of the natriuretic peptides in heart failure: systematic review and diagnostic meta-analysis in the acute care setting. *BMJ* 2015;350:h910. doi: [10.1136/bmj.h910](https://doi.org/10.1136/bmj.h910)
45. Gopalakrishna G, Langendam MW, Scholten RJ, et al. Guidelines for guideline developers: a systematic review of grading systems for medical tests. *Implement Sci* 2013;8:78. doi: [10.1186/1748-5908-8-78](https://doi.org/10.1186/1748-5908-8-78)
46. Horvath AR. Are guidelines guiding us on how to utilize laboratory tests? *eJIFCC* 2015;26(3):146-57.
47. Kahn SE, Jones PM, Chin AC, et al. Defining the path forward: guidance for laboratory medicine guidelines. *eJIFCC* 2015; 26(3):158-67.
48. Glasziou P, Irwig L, Mant D. Monitoring in chronic disease: a rational approach. *BMJ* 2005;330(7492):644-8. doi: [10.1136/bmj.330.7492.644](https://doi.org/10.1136/bmj.330.7492.644)
49. Bell KJ, Glasziou PP, Hayen A, et al. Criteria for monitoring tests were described: validity, responsiveness, detectability of long-term change, and practicality. *J Clin Epidemiol* 2014;67(2):152-9. doi: [10.1016/j.jclinepi.2013.07.015](https://doi.org/10.1016/j.jclinepi.2013.07.015)
50. Muir Gray JA. Evidence-Based Healthcare: How to Make Decisions About Health Services and Public Health: Churchill Livingstone 2008.
51. Moynihan R, Doust J, Henry D. Preventing overdiagnosis: how to stop harming the healthy. *BMJ* 2012;344:e3502. doi: [10.1136/bmj.e3502](https://doi.org/10.1136/bmj.e3502)
52. Sandhu GS, Andriole GL. Overdiagnosis of prostate cancer. *J Natl Cancer Inst Monogr* 2012;2012(45):146-51. doi: [10.1093/jncimonographs/lgs031](https://doi.org/10.1093/jncimonographs/lgs031)
53. Schroder FH, Hugosson J, Roobol MJ, et al. Screening and prostate cancer mortality: results of the European Randomised Study of Screening for Prostate Cancer (ERSPC) at 13 years of follow-up. *Lancet* 2014;384(9959):2027-35. doi: [10.1016/S0140-6736\(14\)60525-0](https://doi.org/10.1016/S0140-6736(14)60525-0)
54. Bilous R. Diagnosis of gestational diabetes, defining the net, refining the catch. *Diabetologia* 2015;58(9):1965-8. doi: [10.1007/s00125-015-3695-4](https://doi.org/10.1007/s00125-015-3695-4)
55. Guyatt GH. Evidence-based medicine. *ACP Journal Club* 1991;114:A16.
56. Guyatt GH, Rennie D. Users' guides to the medical literature. *JAMA* 1993;270(17):2096-7.
57. Sackett DL, Rosenberg WM, Gray JA, et al. Evidence based medicine: what it is and what it isn't. *BMJ* 1996;312(7023):71-2.
58. Eddy DM. Evidence-based medicine: a unified approach. *Health Aff (Millwood)* 2005;24(1):9-17. doi: [10.1377/hlthaff.24.1.9](https://doi.org/10.1377/hlthaff.24.1.9)

## MULTIPLE CHOICE QUESTIONS

1. Of the below questions, which one is a background question?
  - a. Will screening with the prostate-specific antigen test improve mortality in men above 70 years of age?
  - b. Does self-monitoring of INR help reduce complications of warfarin treatment?
  - c. Why is haptoglobin decreased in hemolytic anemia?
  - d. Does Troponin measurement by point-of-care testing improve survival of patients presenting to emergency with chest pain compared to troponin measured in a central laboratory?
  - e. In patients admitted to hospital with heart failure does discharge BNP concentration predict readmission to hospital?
2. P-I-C-O is a format for expressing an answerable question in evidence-based medicine. What does the C in PICO stand for?
  - a. C stands for clinical
  - b. C stands for comparator
  - c. C stands for comparable
  - d. C stands for confidence
  - e. C stands for conclusion
3. The below are testing-related health outcomes. Which one is a proxy outcome?
  - a. Patient satisfaction
  - b. Removal of symptoms
  - c. Prevention of premature death
  - d. Quality of life
  - e. Reduced HbA<sub>1c</sub> in response to treatment
4. Diagnostic accuracy, as expressed by sensitivity and specificity, is a measure of
  - a. Analytical performance
  - b. Analytical validity
  - c. Clinical performance
  - d. Clinical effectiveness
  - e. Clinical utility
5. The below are key elements of the test evaluation cycle, EXCEPT
  - a. Analytical performance
  - b. Clinical performance
  - c. Clinical effectiveness
  - d. Cost-effectiveness
  - e. Health technology assessment
6. The below describe the main objectives of clinical audit, EXCEPT:
  - a. Solving problems associated with the clinical pathway or outcome of laboratory services delivered
  - b. Monitoring test utilization and controlling demand
  - c. Monitoring variations in practice
  - d. Assessing compliance with quality management procedures
  - e. Measuring the impact of testing on outcomes
7. What is the difference between a systematic review and a meta-analysis?
  - a. There is no difference—the two are synonyms
  - b. Meta-analysis is the statistical technique for providing summary estimates based on the results from multiple studies in a systematic review
  - c. A systematic review is the statistical technique for providing summary estimates based on the results from multiple studies in a meta-analysis
  - d. A systematic review is a more extensive form of meta-analysis
  - e. A meta-analysis is a more extensive form of systematic review
8. Main characteristics of good monitoring tests are listed below, EXCEPT:
  - a. Clinical validity
  - b. Responsiveness
  - c. Detectability
  - d. High biological variation
  - e. Practicality
9. What is the most correct definition of “overdiagnosis”?
  - a. Overdiagnosis is the detection of a condition that would never have caused patients any morbidity or mortality
  - b. Overdiagnosis is testing without subsequent clinical actions
  - c. Overdiagnosis is a situation where too many tests are ordered
  - d. Overdiagnosis is a situation where too many test results are reported
  - e. Overdiagnosis is the diagnosis of a condition more often than it is actually present
10. The term “evidence-based medicine” was introduced in the
  - a. 17th century
  - b. 18th century
  - c. 19th century
  - d. 20th century
  - e. 21st century

## Biobanking\*

*Christina Ellervik and Jim B. Vaught*

### ABSTRACT

#### Background

Biobanks may be established for nonresearch purposes, such as diagnostic, therapeutic, treatment, forensic, transplantation, and transfusion, or for research purposes as part of epidemiologic studies and clinical trials. Biobank planning is essential for biospecimen integrity in support of such research, but also for the establishment, governance, management, operation, access, use, sustainability, and discontinuation of biobanks.

#### Content

We focus on best practices procedures for collection, processing, storage, and retrieval of biospecimens with regard to

downstream analyses for blood, urine, and saliva. Security measures, disaster planning, quality management, accreditation and certification, staff education, chain-of-custody, annotation of data, cost, and sustainability issues are reviewed. Adoption of various internal and external standards is discussed. Ethical, legal, and social issues, as well as administrative issues regarding governance, ownership, stewardship, and access criteria are discussed.

\*The full version of this chapter is available electronically on [ExpertConsult.com](#).

## INTRODUCTION AND HISTORICAL PERSPECTIVE

Biobanks exist in many fields of the natural and medical sciences and may consist of collections of human, environmental, animal, microorganisms, plant, and museum material.<sup>1</sup>

Biobanking involves the collection, processing, transport, storage (biopreservation), and retrieval of biospecimens for future purposes (see At a Glance: Biobanking).<sup>1</sup> Evidence-based practices are critical to the future of biobanking and more research is needed to replace current empirical practices with evidence-based protocols. If one prepares carefully by using standard operating procedures (SOPs), variability due to preanalytical issues may be largely avoided.

Technically biobanks have existed for hundreds of years with historical collections of ancient human, animal, or botanical material, even before the term biobank was created and a definition existed. Within these “biobanks,” conservation, long-term preservation, and collection management were standard practices long before modern human biobanks were established. Pathology collections were the most prevalent in the early era of biobanks over 100 years ago primarily for treatment and diagnostic purposes.<sup>2</sup> However, as a consequence of growing recognition that such collections could also contribute significantly to biomedical research, more extensive epidemiologic and clinical trial collections were initiated 40 years ago. According to a survey of 456 current biobanks in the United States, 17% have existed for the last 20 years, and 60% have been established within the last 10 years.<sup>3</sup> The rapid growth of biobanks during the last 20 years may be explained by the technological (information technology [IT], automation, instrumentation, and advances in methodologies) and scientific developments making it

easier to handle, store, and analyze large and complex sets of data. The total global number of biobanks is unknown.

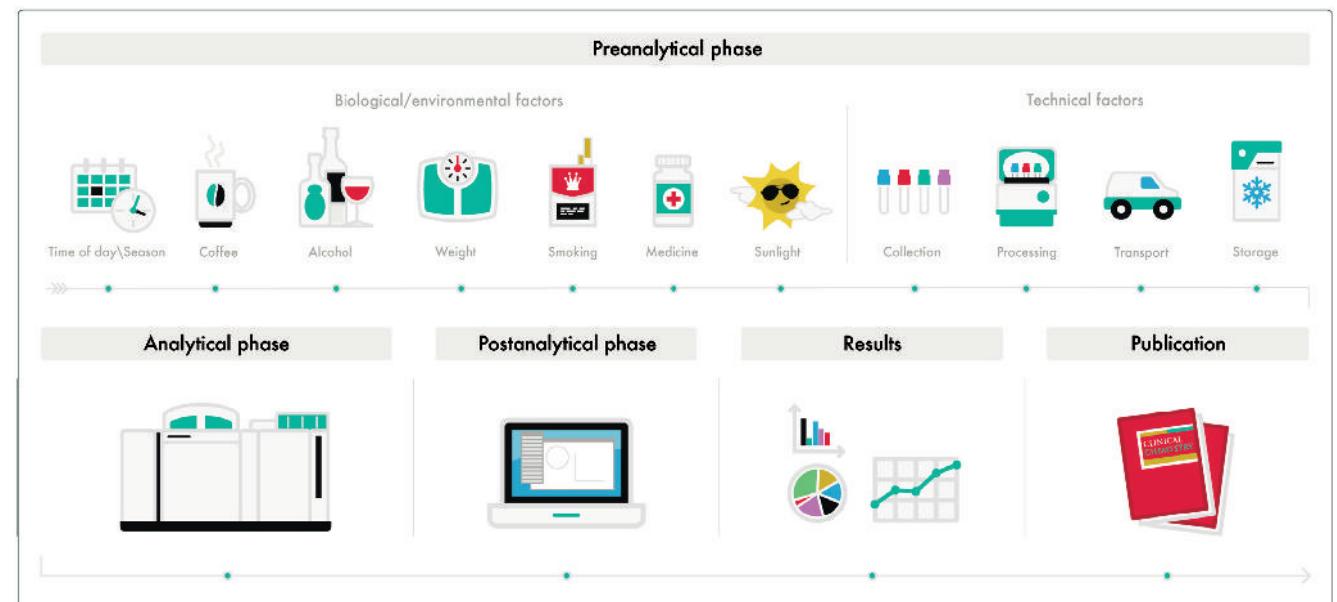
The synonyms for biobank are bank, biological resource center, biospecimen resource, biorepository, and repository. “Biobank” is now the most widely used term, whereas “biorepository” was the first to appear in PubMed in 1994. Since then, the number of publications using either of the terms “biobank” or “biorepository” have gradually increased each year, now reaching 7621 hits from a search in PubMed (on February 23rd, 2020). However, this is probably only the tip of the iceberg, as many human collections and cohorts were established long before the word “biobank” became an established term.

The Organization for Economic Co-operation and Development (OECD) defines a biobank as “A collection of biological material and the associated data and information stored in an organized system, for a population or a large subset of a population.”<sup>4</sup> In a survey of 303 people related to biobanks, 50% agreed that the term “biobanking” describes the collection, processing, and storage of all human, animal, plant, microbial, and environmental materials;<sup>5</sup> 60% agreed that biobank and biorepository have the same meaning; 22% agreed that just one banked biospecimen constitutes a biobank and almost 90% agreed that to be called a “biobank,” the collection must be associated with sample data. There is broad agreement of what constitutes a biobank but the precise definitions of biobanking are many and differ according to different stakeholders and country in which the biobank is located.<sup>6</sup> Being unaware that their collections comprise a biobank will make the owners or researchers of such collections less likely to respond to communications and legal requirements concerning biobanks.<sup>6</sup> Such attitudes can potentially pose hazards to biospecimen integrity and data privacy.

## AT A GLANCE

### Biobanking

#### Preanalytical, Analytical, and Postanalytical Phases of the Biobanking Process



From Ellervik C, Vaught J. Pre-analytical variables affecting the integrity of human biospecimens in biobanking. *Clin Chem* 2015;61:914–34, with permission.

## POINTS TO REMEMBER

### Definitions

**Biobanking/Banking**—The process of storing material or specimens for future use.

**Biobank/Biorepository**—An entity that receives, stores, processes, and/or distributes specimens, as needed. It encompasses the physical location and the full range of activities associated with its operation.

**Biospecimen Resource**—A collection of biological specimens that is acquired for a defined purpose. Management responsibility of the biospecimen resource is led by the custodian for the collection. Biospecimen resources may be stored in a repository or laboratory, depending on the numbers of specimens contained therein.

**Culling**—Reviewing and eliminating specimens in a collection or an entire collection either by destruction or transfer to a new custodian.

**Custodian**—The individual responsible for the management of a biospecimen resource. The custodian works with other key stakeholders in the management of the resource including the tracking of all relevant documentation for the resource and for ensuring that policies regarding access to the resource are in place and implemented according to appropriate guidelines.

**Desiccation**—Excessive loss of moisture; the process of drying up.

**Identifier/Identifying Information**—Information (e.g., name, social security number, medical record or pathology accession number, etc.) that would enable the identification of the subject. For some specimens this information might include the taxon name and collection number.

**Lyophilized**—Dehydrated for storage by conversion of the water content of a frozen specimen to a gaseous state under vacuum. Also called freeze-dried.

**Material Transfer Agreement**—An agreement that governs the transfer of tangible research materials and data between two organizations, when the recipient intends to use it for his or her own research purposes. It defines the rights and obligations of the provider and the recipient with respect to the use of the materials.

**Retrieval**—The removal, acquisition, recovery, harvesting, or collection of specimens.

**Specimen**—A specific tissue, blood sample, etc., taken from a single subject or donor at a specific time. For some biological collections “specimen” may have the same meaning as “individual.”

Definitions from ISBER: International society for biological and environmental repositories. Best practices for Repositories: collection, storage, retrieval and distribution of biological materials for research—4th edition. [www.isber.org](http://www.isber.org), 2018.

The search terms “biospecimen” or “specimen” revealed 127,351 hits (on February 23rd, 2020) in PubMed, with the first publication dating from 1828. The term “biospecimen” appeared in 1965 in PubMed. However, a human biospecimen refers to any material taken from the human body and may also constitute larger parts of tissue samples and whole organs; in that sense, museum collections of normal and diseased tissue; organs or bodies in museums, hospitals, or academic institutions exhibits; and education or research

collections may also constitute early biobanks. This historical perspective of biobanks is largely unexplored. The OECD guidelines do not include a definition of biospecimen, but The National Cancer Institute (NCI) Biorepositories and Biospecimen Research Branch (BBRB) website provides the following definition: “Biospecimens are materials taken from the human body, such as tissue, blood, plasma, and urine that can be used for cancer diagnosis and analysis.”<sup>7</sup>

The science of biobanking is a dynamic field. Recent focus has been on quantity and quality of biospecimens. But an additional current focus is now directed to biobank sustainability socially, operationally, and financially.<sup>8</sup> Biospecimen science is the emerging field of study which aims to quantify and control preanalytical factors.<sup>9</sup> Thus biospecimen science studies evaluate and optimize approaches to biospecimen collection, processing, and storage, and other related procedures.

### Rationale and Objectives for Biobanking

The rationale for initiating biobank collections may be research, diagnostic, therapeutic, transplantation, transfusion, quality assurance, forensic, or archeologic studies (which may be used for exhibits, education, and research). Initiators may be from academic, hospital, governmental, or industrial organizations.

The objectives of biobanks are many and depend on the nature of the intended research. Many biobanks collect data for future research projects for which the aims and technologies are not necessarily well developed at the time when samples and accompanying data are collected. For the biobanks which are primarily storage warehouses and not engaged in research, the objective may simply be to maintain the highest-quality samples possible. In clinical and research biobanks, the objectives may be omics-biomarker research combined with clinical data to predict individual predisposition of disease, target prevention, and personalize treatment (personalized or precision medicine) tailored to each individual person.<sup>10,11</sup> For biobanks focused on quality assurance, the objective may be to use the material for developing or optimizing new diagnostic methods. The objective of therapeutic biobanks is to store viable tissue from donors for future recipients (i.e., semen or oocytes). The objective of forensic biobanks is to store the biological material from which the DNA profiles were analyzed, for documentation and possible testing for legal purposes.<sup>12</sup> For the archeologic collections, the objectives may be the development of exhibits, educational material, and research in evolutionary biology and anthropology.

### Types and Use of Biobanks

Biobanks may be classified according to the funding source, the ownership, the location, the recruitment strategy, the biospecimen type, the administration, the users, the purpose, or membership in a network (Table 11.1).<sup>3,13–17</sup> According to surveys of US and European biobanks, most biobanks are disease-specific compared to general population collections.

Human biobanking takes place in the pharmaceutical industry, commercial labs, government facilities, hospitals, and academia. Biobanks are central tools in basic research, genetic epidemiologic studies, and clinical trials and the results are used in translational research and precision medicine.

A biobank may store samples from other biobanks, from many basic or translational research projects, and from

**TABLE 11.1 Types of Biobanks**

<b>Types of Biobanks</b>	<b>Selected Statistics Derived From Henderson et al. (2013)</b>	<b>Types of Biobanks</b>	<b>Selected Statistics Derived From Henderson et al. (2013)</b>
<b>Based on Funding</b>		<b>Based on Biospecimen Type and Number</b>	
Governmental (state, region, federal)	Examples of funding sources: • Federal: 36% of biobanks	Whole blood, plasma, serum, buffy coat, dried blood spot	22% have less than 1000 biospecimens
Nonprofit	• The parent organization of the biobank: 30%	Urine	52% have less than 10,000 biospecimens
Commercial	• Fees for services: 11%	Feces	23% have more than 100,000 biospecimens
Participant	• Individuals or foundations: 10%	Saliva	77% store serum or plasma
Access fee		Solid tissue	69% store solid tissue
		Hair, nails	30% store urine
			13% store only one specimen type
<b>Based on Ownership</b>		<b>Based on Administration</b>	
Governmental	78% of biobanks are part of academic institution	Storage: A deposit for different research groups	
Hospital	27% of biobanks are part of hospital organization	Research: Each research project has its own biobank	
Academic	Many biobanks are part of more than one institution		
Industrial			
<b>Based on Location</b>		<b>Based on Users</b>	
International		Mono-user	
National		Oligo-user	
State/regional		Multiple users	
Industrial			
<b>Based on Recruitment/Nonrecruitment</b>		<b>Based on Purpose</b>	
Recruitment	44% of biobanks are pediatric	Consent required	
• Population-based	75% of biobanks get biospecimens from participants donating	• Research specific requests for donation	
• Newborn cohort	57% of biobanks get biospecimens from residual/left-over	• Anatomical gifts from deceased donors to education, transplant, or research (by testimony)	
• Adult cohort	specimen from clinical procedures	Consent not required	
• Pediatric cohort	29% of biobanks facilitate general research	• Archeological: exhibits, education, research	
• Family study	53% of biobanks facilitate disease-specific research	• Treatment (stem cell, semen, blood)	
• Twin study		• Clinical, pathological	
• Household		• Legal/forensic	
• Hospital		• Quality control	
• Environmental exposure			
• Other cohorts			
• Disease-based		<b>Based on Network<sup>15</sup></b>	
• Case-control		Storage	
• Case-only		Bring-and-share	
• Hospital recruited		Catalogue	
• Clinical trial recruited		Partnership	
Nonrecruitment		Contribution	
• Residual biospecimen/leftovers		Expertise	
• Pathology			
• Microbiology			
• Chemistry			
• Cadaveric tissue			
• Treatment (e.g., eggs, oocytes, sperm, blood)			

Statistics from Henderson GE, Cadigan RJ, Edwards TP, Conlon I, Nelson AG, Evans JP, et al. Characterizing biobank organizations in the U.S.: results from a national survey. *Genome Med* 2013;5:3.

clinical settings as well, and is thus a storage facility. Research biobanks may process and store samples from specific phenotypes and patients with a specific disease (i.e., cases) and in addition samples from representative disease-free controls from the underlying population, or samples from a general population base. General population studies may collect biospecimens on a single individual basis or on a family basis. Both case-specific biobanks and general population biobanks may have representative samples from only specific age groups (e.g., children vs. adults) or from any age range. Case-specific biobanks are useful for diagnosis, disease stratification, and prognostic purposes. Clinical biobanks may store samples from cases with a given disease, but are primarily organized for clinical purposes, usually as part of a diagnostic process. Some clinical biobanks are also research biobanks, where the participants have provided informed consent for future use of diagnostic samples for research purposes. Both clinical and commercial biobanks may store samples from volunteers for treatment purposes (stem cells, transplant organs and tissues, oocytes). Depending on the type of biobank, different regulations and accreditation procedures exist.

## BEST PRACTICES IN BIOBANKING

Differences in SOPs for collection, processing, and storage of biospecimens within and between studies may introduce differences in quality and results either toward the null or with skewed bias resulting in misclassification of disease, loss of study power, and increased costs. Thus standardization within and between studies is needed. Best practices are guidelines written by experts and issued by organizations such as the International Society for Biological and Environmental Repositories (ISBER),<sup>1</sup> US NCI,<sup>7</sup> OECD,<sup>4</sup> and World Health Organisation International Agency for Research on Cancer (WHO-IARC).<sup>18</sup>

### Preanalytical Variability

The term preanalytical is defined as anything that comes before the analysis phase of a biospecimen sample (see also Chapter 5). Thus in biobanking preanalytical handling is basically all processes that precede the analysis of a biospecimen after it is collected from a donor or removed from storage. Preanalytical variables are factors that affect the integrity of the biospecimens, and later the results of analyses. Assessing and controlling the preanalytical handling of biospecimens is fundamental for the integrity and optimal future use of biospecimens.<sup>19,20</sup> Biobanking involves the collection, processing, transport, storage (biopreservation), and retrieval of biospecimens. Evidence-based practices are critical to the future of biobanking, but more research is needed. Many factors influence the analytical results in clinical biochemistry, that is, preanalytical biological or environmental variability, preanalytical technical variability, analytical variability, and postanalytical variability (see *At a Glance: Biobanking*). Most errors in a clinical chemistry lab are due to preanalytical errors<sup>21,22</sup> and may result in inaccurate test results or systematic biases.<sup>23</sup> The most common preanalytical errors occur in the ordering or collection phase.<sup>24</sup> Preanalytical variables can introduce in vitro modifications, either systematically or randomly, which can adversely affect laboratory results.

### BOX 11.1 Factors Leading to Incomplete Collection

#### Specimen-Related Factors

Prioritization of clinical diagnostics to research (volume of biospecimens is too small to be used for research)  
Unable to obtain specimen

#### Staff-Related Factors

Illness  
Vacation  
Forget to collect  
Missing consent

#### Patient-Related Factors

Forget to collect  
No instruction—wrong self-collection  
Forget appointment  
Illness  
Vacation  
Patient not adequately prepared (diet, medication etc.)

#### Management Related Factors

Forgot to schedule the patient  
Change of scheduling date  
Change of collection location

#### Logistics

Bad weather hindering transport (of patient to collection site, or of staff to patient's home)  
Long distance from patient's home to collection site

Courtesy Christina Ellervik.

### Ordering, Collecting, and Receiving Biospecimens

The collection of human biospecimens is a part of the biobanking process that cannot easily be automated and depends on many factors, such as availability of biospecimens, staff, participants, management, and logistics (Box 11.1).<sup>25</sup> Collecting biospecimens in cohort studies is a balance between biospecimen quantity and type, accrual rate and number, location, costs, transport logistics, and storage requirements. The resulting participation rate may depend on what level of cooperation is reasonable to request from a participant.<sup>26</sup> The collection of biospecimens may be invasive (e.g., blood), less-invasive (e.g., dried blood spot [DBS]), or noninvasive (e.g., urine or saliva). Blood and urine are biospecimens commonly collected for clinical analyses. Less-invasive and noninvasive methods minimize use of valuable blood samples, and may lead to an increased sample size of the study population owing to their reduced costs, ease of collection without specialized staff, and willingness of participants to donate (Table 11.2 and Box 11.2).<sup>27–29</sup>

In a clinical chemistry laboratory, preanalytical variables related to ordering or receiving biospecimens may impact the quantity or quality of biospecimens (e.g., missed, incorrect, or duplicate collection, data entry error, incorrect patient or collector ID, insufficient sample, diluted sample, improper labeling, lost biospecimens) (Table 11.3).<sup>19</sup> If biospecimens are obtained without consent, with a lost consent, or a restricted consent, then their value may be limited.

Biological and environmental factors may also affect downstream analyses (Box 11.3).<sup>30–33</sup> The total variability of

**TABLE 11.2 Advantages and Disadvantages for Major Biospecimen Categories**

<b>Advantages</b>		<b>Disadvantages</b>
Blood and blood components (whole blood, plasma, serum)	Most analyses possible	Patients need to rest Requires trained staff Invasive: Painful collection Number of tubes may affect participation rate Analytes are tube-additive dependent
Dried blood spot	Minimally invasive Easy collection No processing Easy RT transport Less painful Patient self-collection Small blood volume Equal to whole blood No processing Minimal risk Pediatric collection Long-term storage at RT Space-saving Cost-effective	No staff training: risk of disposal of samples due to bad collection technique Low or high hematocrit may interfere with analyses Too small blood volume: requires high sensitivity of analytical method
Urine	Noninvasive Easy collection Patient self-collection Pediatric collection	Transport and short-term storage on ice Contamination
Saliva	Noninvasive Easy collection Patient self-collection Pediatric collection DNA is only the donor's DNA Cost-effective Patients afraid of needles Minimal risk of contracting infections Suitable for large-scale collection Easy transport	Low concentration of analytes

RT, Room temperature; WB, whole blood.

Courtesy Christina Ellervik.

From Ellervik C, Vaught J. Pre-analytical variables affecting the integrity of human biospecimens in biobanking. *Clin Chem* 2015;61: 914–34; with permission.

these factors may impact the concentrations of analytes. Smoking may increase red and white blood cell indices.<sup>34,35</sup> The participants' position during blood collection may affect many molecules' concentrations, which increase from supine, to sitting, to standing position, although the latter position is discouraged.<sup>31</sup> See Chapter 5 for further discussion on this topic. Twenty-four-hour variation may be seen in many chemistry analytes,<sup>32</sup> with peak and low values at

**BOX 11.2 Types of Primary Biospecimen Collection**

<b>Invasive</b>	<b>Less Invasive</b>	<b>Noninvasive</b>
Whole blood	Dried blood spot	Urine
Plasma	Dried serum spot	Saliva
Serum	Cord blood	Buccal cells
Tissue	Placenta	Feces
Pathological		Hair
Normal around pathological		Nail
Normal		Breast milk
Cerebrospinal fluid		Nasal secretions
Amniotic fluid		Tears
Bronco-alveolar lavage		Sweat
Stem cells		Cervico-vaginal excretions
		Semen
		Oocytes

Courtesy Christina Ellervik.

different times of the day. Marked metabolic and hormonal changes occur after food ingestion.<sup>36</sup> The postprandial response varies according to factors such as eating behavior, food composition, fasting duration, time of day, chronic and acute smoking history, and coffee and alcohol consumption.<sup>36</sup> Some biological factors can be controlled in studies by requiring certain conditions for participant inclusion, for example, fasting/nonfasting, abstaining from smoking and strenuous exercise hours before collection (see Box 11.3). Environmental factors include geographic location, altitude, inside or outside temperature, season, humidity, and moisture.<sup>19,33,37</sup> During summer vitamin D concentrations are higher than in winter, and in more sunny geographical locations individuals have higher vitamin D values than in less sunny locations.<sup>38</sup> Direct sunlight may affect concentrations of bilirubin, porphyrins, and vitamin A. The total variability of these factors may impact concentrations of analytes. For the measurement of medication concentrations and hormones, the timing of collection is especially important. Thus these factors should be standardized, documented, and taken into consideration when interpreting results or comparing or pooling the results of multiple studies. Collection of repeat samples from the same individual taken a few days apart may attenuate the effects of preanalytical and analytical variation. Serial measurements may also be taken with longer time intervals between, to measure changes or effects of intervention over time.

All biospecimens should be treated as biohazards and all processes involving biospecimens should adhere to principles of general laboratory safety.

### Processing of Biospecimens

Automated systems for processing samples incorporate barcode reading of primary tubes (collection tubes), decapping, fractionating, aliquoting into predefined secondary tubes or plates, and transferring of labels onto secondary tubes.<sup>39</sup> The automated systems should have complete sample tracking capability. Benefits of automated fractionation systems include fewer errors in sample handling and prevention of endurance related injuries due to repetitive work actions. Such systems are operator

**TABLE 11.3 General Preanalytical Variables, Recommendation, and Documentation Requirements for Biobanking**

Step	Preanalytical Variables	Recommendation	Documentation Requirements
Ordering	Ordering forgotten	Laboratory information system	Date and time of ordering Other annotation to database: clinical tests, diagnoses, socio-demographic, other measurements.
	Consent: none, forgotten, restricted, lost	Secure consent	Consent type
	Typing error	Check spelling	
	Incorrect patient ID	Check IDs	Patient ID, name, gender, birthday, age
	Incorrect collector ID	Scan IDs, avoid manual typing	Reference number Tube ID number Collector ID
	Incorrect identification of patient		Label errors
	Pairing patient ID with primary tube ID	Check pairing	
	Improper labeling, mislabeling, no labeling	Stable adhesive and unique labeling Check labeling	
	Biological and environmental factors (Box 11.3)	Follow use evidence-based literature and guidelines for standardization	Date and time of collection Biological and environmental variability (see also Box 11.3) Fasting/nonfasting Time since last meal, smoking, beverage, alcohol, medication, chewing gum
			Staff collection or patient collection
Collection	Forgotten collection	Educate staff and patients	
	Incorrect collection		
	Duplicate collection		
	Collection device types	Use same tubes throughout a study and between studies	Any information on devices, brands, volume, and types
	Collection device age	Check expiration date for collection device	Anatomical location
	Anatomical location of collection	Sterile collection	Primary tube brand
	Contamination of specimen: microorganisms, tube material, tube additive		
	Empty tube	Check volume	Volume collected
	Insufficient sample volume		Intended or unintended dilution
	Diluted sample		
Receiving	Open container: spill	Secure stopper on tubes	Document spill
	Label removed, label destroyed	Never re-label: re-collect biospecimen or destroy biospecimen	Label errors
Processing	Biospecimen lost after collection	Secure chain-of-custody	Lost biospecimens
	Not received after collection		
	Short-term storage temperature and time until processing	Track temperature and time	Short-term storage temperature and time until processing
	Processing duration	Process rapidly	Date and time of processing
	Aliquot volume	Aliquot to secondary tubes Multiple small volume aliquots instead of few large volume aliquots	Secondary tube brand and type (single tube, plate, matrix, straw) Number of aliquots Volume of aliquots
		Label on secondary tubes: Cryo-stable, readable unique 2D (and 1D label)	Coding with linkage to primary tube number and patient ID
	Improper labeling, mislabeling, detached label	Coded and anonymized	Link between patient ID, primary and secondary tube IDs (1D and 2D labels)
	Pairing primary tube ID with secondary tube ID	Follow short-term or long-term storage temperature recommendations	Temperature during transport (temperature log)
	Environmental exposures (Box 11.3)		Date and time from departure
			Date and time at arrival
Transport/shipping	Sent to wrong laboratory	Schedule shipping according to collection time	Duration from destination A to B
	Receiver not on duty		

**TABLE 11.3 General Preanalytical Variables, Recommendation, and Documentation Requirements for Biobanking—cont'd**

Step	Preanalytical Variables	Recommendation	Documentation Requirements
	Packaging, labeling	Follow packaging guidelines according to type of shipment Gentle transport Pack for stable temperature Use licensed couriers Ship small amounts, not the whole collection at once Keep duplicates apart	Register which biospecimens have been shipped Name of courier Type of packaging, labeling
Long-term storage	Time from processing to storage Storage duration, temperature, and facility Other environmental impact: Sunlight Humidity Moisture Dehydration, evaporation Oxidation Desiccation	If possible: use evidence-based literature, pilot study, or internal biomarkers to determine long-term storage time and temperature impact on stability and recovery Store at $-80^{\circ}\text{C}$ or liquid nitrogen (if RT-stable, store at RT)	Duration Time from processing to storage Detailed storage information: <ul style="list-style-type: none"><li>• Box number and placement in box</li><li>• Rack number and placement in rack</li><li>• Freezer number and rack placement in freezer</li><li>• Back-up freezer number for each freezer</li><li>• Freezer location</li><li>• Freezer brand</li><li>• Freezer temperature (temperature log)</li><li>• Temperature log</li></ul> Freeze-thaw cycles: <ul style="list-style-type: none"><li>• Number</li><li>• Date and time</li><li>• Purpose</li><li>• Staff name</li><li>• Discard or return</li></ul> Type of emergency Attempts to rescue biospecimens
	Freeze-thaw cycles	Avoid multiple freeze-thaw/ single-use aliquots only	
	Especially for emergencies/disasters: Encapsulation in ice after re-freezing Microbiological contamination (yeast, mold, fungus, bacteria, and virus-causing biological hazards) No labeling or destroyed labeling	Have an emergency or disaster plan for transferring biospecimen in case of power outage, flooding, earthquakes, hurricane, fire Have enough back-up freezers Maintain, repair, replace freezers Store in multiple locations Make sure labels are cryo-stable Destroy biospecimens with un-readable labels	
	Missing aliquots Misplaced aliquots	Secure chain-of-custody	An electronic laboratory information system for documentation

1D, One-dimensional; 2D, two-dimensional; ID, identification; RT, room temperature.

From Ellervik C, Vaught J. Pre-analytical variables affecting the integrity of human biospecimens in biobanking. *Clin Chem* 2015;61:914–34; with permission.

independent and ensure proper sample tracking.<sup>39</sup> In laboratories with low throughput or less financial resources, manual handling may be needed, but this approach increases the risk of errors. Multiple aliquots should be created at the beginning of processing a biospecimen rather than delayed until the specific assay is conducted, as repeated freeze-thaw cycles may be detrimental in some cases (e.g., RNA).<sup>40,41</sup> A study confirmed the validity and reliability of a high-throughput, high-density, low-volume biobank sample processing solution for blood fractionation and archiving biospecimens utilizing the 384 aliquoting format sample storage tube system.<sup>40</sup> A study of high-density

scaling allowed for reproducible aliquoting and processing of 70- $\mu\text{L}$  volumes of blood.<sup>40</sup> With this approach the authors introduced the principle of single-use only for samples, circumventing multiple freezing and thawing cycles.

### Storage of Biospecimens

Ideally, a “freeze-thaw stable” fluid biospecimen is not affected by thermal, mechanical, or chemical stress. Thus the goal in storage of biospecimens is to minimize or halt these detrimental processes.<sup>42</sup> Storage encompasses both short-term and long-term storage of biospecimens, depending on

### BOX 11.3 Biological and Environmental Variability Affecting Downstream Analyses Measured in Blood, Urine, and Saliva

Biological Variability		Environmental Variability
Age	Posture	
Gender	Circadian variation	
Ethnicity	Diurnal variation	
Body mass index	Hydration status	
Menstrual cycle	Fever	
Pregnancy	Disease	
Lactation		
Diet	Seasonal changes	
Alcohol	Temperature	
Medication	Humidity	
Caffeine	Moisture	
Smoking	Geographic location	
Fasting/nonfasting	Altitude	
Exercise	Sunlight	

From Ellervik C, Vaught J. Pre-analytical variables affecting the integrity of human biospecimens in biobanking. *Clin Chem* 2015;61:914–34; with permission.

their planned future use. Biospecimens contain degradative molecules (e.g., proteases, lipases, nucleases).<sup>43</sup> Long-term storage may result in aggregation, precipitation, or biochemical degradation of proteins (altering both structure and activity); ice damage; dehydration and increase in salt concentration resulting in osmotic damage; formation of water crystals; recrystallization after thawing; and toxicity from substances that are added to the biospecimens in the freezing state (cryoprotectants) or in the drying state (lyoprotectants) in order to protect the active ingredients.<sup>42</sup> These changes may cause the real biological variations to disappear. There is a considerable variation among biomarkers in stability and recovery; therefore different storage conditions may apply depending on the downstream analyses. Preanalytical variables for long-term storage are listed in Table 11.3. Freeze-thaw cycles are a major concern, and may happen unintentionally during transport of frozen samples or freezer failure, or intentionally because the biospecimens are thawed for analyses and then refrozen.

In order to mitigate possible freeze-thaw cycle problems, controlled rate freezing and thawing methodology may be employed. These technologies are used especially in the case of cell preservation. For example, in 2019, Baboo et al. studied the effects of various rates of controlled freezing and thawing on the viability of human cryopreserved T cells.<sup>44</sup>

It is advised to perform pilot studies and to carefully search the literature before measuring biomarkers on stored biospecimens. It is also important to have some biospecimens available only for quality control purposes, on which the same biomarkers are measured in fresh biospecimens repeatedly on a regular annual basis to monitor any critical changes (Box 11.4). Standard protocols are necessary for reproducible and reliable results.

### Storage Facilities and Equipment

Different types of storage facilities exist (Box 11.5). The choice of facility and equipment depends on:

- Sample size
- Accrual rate

### BOX 11.4 Storage Recommendations for Fluid Biospecimens

Store in the vapor phase of liquid nitrogen or, alternatively, at minimum of  $-80^{\circ}\text{C}$ .  
 Keep a constant cooling rate during freezing.  
 Minimize temperature fluctuations during storage.  
 Minimize repeated freeze-thaw cycles.  
 Fast thawing methods should be utilized.  
 Thawing rates should be monitored and validated.  
 Run a pilot stability/recovery study or study literature carefully for specific potential future biomarkers of interest.

Data from Hubel A, Aksan A, Skubitz AP, Wendt C, Zhong X. State of the art in preservation of fluid biospecimens. *Biopreserv Biobank* 2011;9:237–44.

### BOX 11.5 Storage Type

Liquid nitrogen freezers  

- Vapor  $\text{LN}_2$  ( $\leq -150^{\circ}\text{C}$ )
- Liquid  $\text{LN}_2$  ( $-196^{\circ}\text{C}$ )

 Mechanical freezers  
 Refrigerators  
 Walk-in environmental storage systems  
 Fully automated entry and retrieval systems  
 Ambient temperature storage

$\text{LN}_2$ , Liquid nitrogen storage.  
 Courtesy Christina Ellervik.

- Complexity of collection and processing procedures
- Type and number of specimens to be stored
- Anticipated length of time the specimens will be stored
- Intended use for the specimens
- Volume and number of aliquots (for later use)
- The resources available for purchasing the equipment
- Storage density
- Predictions of future growth
- Quality management
- Number of staff
- Equipment support and maintenance
- Logistics
- Economic factors
- Biobank governance factors
- Sustainability.<sup>45</sup>

Whether to store biospecimens locally in several locations, centrally, or both depends on their anticipated use. If the samples are expected to be used often, it is recommended to have a duplicate set close to the core laboratory for practical reasons. If the samples are planned to be stored for more than a year, it is recommended to store them centrally.

**Liquid nitrogen storage.** Vapor-phase storage ( $\leq 150^{\circ}\text{C}$ ) is preferred over liquid-phase storage ( $-196^{\circ}\text{C}$ ), but both storage formats have advantages and disadvantages.<sup>1,26,46</sup> Use of the vapor-phase avoids risk of transmission of infectious agents but necessitates a readily available supply of liquid nitrogen storage ( $\text{LN}_2$ ). Liquid-phase storage affords better security in case of a shortage of  $\text{LN}_2$ . The design of the tank is critical to maintain  $\text{LN}_2$  in the vapor-phase. The hazards associated with use of liquid nitrogen are extreme cold, evaporation, asphyxiation, oxygen deprivation, and pressure buildup and explosions of storage vials. The extreme cold can cause

frostbites, cold burns, and eye and tissue damage on personnel. Personal protective equipment should be worn when handling biospecimens stored in LN<sub>2</sub> tanks, including face and eye protection, closed-toed shoes, full covering of legs and feet, eye goggles, and heavy gloves. Liquid nitrogen expands to 700 to 800 times its original volume when it vaporizes. Because nitrogen displaces oxygen, there is a risk of oxygen deficiency in the biobank facility, which may cause asphyxiation, unconsciousness, and eventually death. The risk is inversely correlated with the size of the room. Sufficient ventilation and oxygen sensors should be in place. Oxygen may build up around the tanks increasing the flammability of materials; thus combustible materials must be kept away from the tanks. High pressures can build up when nitrogen evaporates, and tanks must be secured with sufficient vents and pressure relief vessels to protect against explosions. Daily LN<sub>2</sub> usage should be recorded and monitored.

**Mechanical freezers.** Mechanical freezers ( $-80^{\circ}\text{C}$ ) vary in size, shape, temperature, and voltage. When using these freezers it is important to ensure adequate ventilation and maintain a sufficient distance between the freezers. Ambient temperature in repositories should not exceed  $22^{\circ}\text{C}$  ( $72^{\circ}\text{F}$ ).<sup>1</sup> In rooms containing multiple mechanical units, this is particularly critical. Excessive heat may shorten compressor life, and insufficient air circulation may lead to growth of microorganisms in biospecimens. With a larger number of samples, it may be a better solution in terms of costs and long-term biospecimen integrity to choose liquid nitrogen storage instead.

**Refrigerators.** Refrigerators ( $+4^{\circ}\text{C}$  [maximum range  $+2$  to  $8^{\circ}\text{C}$ ]) are usually used for short-term temporary storage between collection and processing.

**Walk-in environmental storage systems.** Recommended practice for a  $-20^{\circ}\text{C}$  or colder walk-in environment is to have audible alarms, motion devices, and door releases.<sup>1</sup>

**Ambient storage.** Special technologies for dry storage of DNA and RNA at room temperature (RT) have been developed, enabling easier shipment of these extracts.<sup>47–54</sup> This approach minimizes required storage space, reduces electrical costs and shipping costs, is helpful when mechanical or cryogenic equipment is not available (e.g., during shipping or in rural areas), or may serve as an alternative method for back-up storage. The technology is comparable to cryopreserved DNA or RNA for up to 1 year.

**Fully automated entry and retrieval systems.** The reasons for choosing fully automated systems may be large sample sizes with too many biospecimens to handle manually, rapid accrual rates, sample integrity (minimizing temperature variations), tracking accuracy, audit requirements, speed, safety, and efficient management.<sup>55</sup> In a survey of biobankers from 2007, 8% had automated systems, whereas 46% were not interested in acquiring one, and another 46% would be interested in considering automation in the future.<sup>55</sup> The automated solutions may require automation-compatible plasticware, sample preparation, and laboratory management information systems (LIMS). Temperatures range from ambient,  $-20$ ,  $-80$ , and  $-150^{\circ}\text{C}$ . The smaller the storage volumes to be aliquoted and the larger the need for one-time-use-only instead of repeated freeze-thaw cycles, the larger the need for automation. Different vendors offer solutions with various capacity, temperature, and throughputs (input and retrieval/day).

**For all storage systems.** An automatic defrost feature should keep the biobank free of water, ice, and frost. It is also

recommended to have a duplicate set of samples stored on different power supplies or in two geographically separate locations to protect against equipment failure, fires, or natural disasters. One or more empty back-up freezers should be operating in case of freezer failure; 1.5 to 3% for LN<sub>2</sub> and 10% of the total number of mechanical freezers is recommended.<sup>1</sup> It is advised to conduct continuous temperature monitoring and periodic reviews of storage equipment performance, as well as to keep equipment maintenance records and replace or repair equipment whenever necessary. Storage facilities must be installed with alarm systems (audible sound, email, text messages or SMS, phone, and pager alerts). Alarm management plans must be developed and documented. In case of power loss, back-up power generation should be in place.

### Transport and Shipping of Biospecimens

Transport of biospecimens includes any method from inter-departmental to international shipment. This process is at risk of being the weakest step of the biobanking pipeline, but is still often overlooked. Logistics of transport are costly and require trained personnel.

Any shipment of biological material must comply strictly with all applicable local, state, federal, and international laws governing packing, marking, and labeling, and must be shipped according to applicable government and International Air Transport Association (IATA) and International Civil Aviation Organization (ICAO) regulations.<sup>56</sup> There must be a correct classification of biospecimens and preservatives.<sup>56</sup> Incorrect packaging, marking, or labeling or incorrectly completed shipping documentation may cause delays or refusals by customs officials and will most likely affect the sample and the study integrity. The Centers for Disease Control and Prevention (CDC) has published guidelines for shipment of samples in the US according to the above regulations. There are several CDC guidelines for sample shipping, depending on the type of specimen and level of infectivity (see [www.cdc.gov](http://www.cdc.gov)). Best practices include using licensed couriers and maintaining any shipment documentation and logs.<sup>1</sup> Devices are available to electronically track and monitor temperature during transport. It is advisable to mail smaller fractions of a collection or shipment in order to keep the environment as stable as possible within a shipping container, and in case a shipment is lost or delayed resulting in the destruction of the biospecimens. Furthermore, it is advised always to keep duplicate biospecimens apart within a shipment or between shipments.

Timing of international shipment or interdepartmental transport (holidays, working days in the week) is crucial for the integrity of the biospecimens. It is advisable to perform a pilot test of the conditions, logistics, and packaging material before real shipments take place. Keep copies of all permits, inform the recipient of the shipment, and require a receipt upon arrival of biospecimens.

Variables affecting biospecimen integrity during transport are season, time of day, delays, distance, and method of transportation (aircraft, courier).<sup>1</sup> International shipments especially can experience delays due to customs clearance. It is important to monitor temperature and provide sufficient coolant for the potential for an additional 3-day delay. Biospecimen packaging should have appropriate insulation for extremes of heat and cold temperatures. Refrigerants between sample containers rather than on top or underneath

should be placed, wadded paper used, and empty space filled with Styrofoam. The most frequently used coolants are gel packs or dry ice. Liquid nitrogen dry shippers may also be accepted by special couriers.

Recommendations for shipment conditions include:

- At RT (20 to 30 °C): use insulated packaging, protect from heat or cold and sunlight
  - At 2 to 8 °C: use preconditioned gel packs at –15 °C
  - At –20 °C: use preconditioned gel packs at below –20 °C, or move the whole freezer
  - At –70 °C: use dry ice (hazardous), or move the whole freezer
  - Below –150 °C: use a specialized liquid nitrogen dry shipper
- Dry ice shipments are included in the IATA class 9 dangerous goods regulation, and require that staff are trained at sites for proper handling.<sup>56</sup>

All tubes should be transported vertically in the closure-up position. This is especially important for blood tubes as the upright position reduces hemolysis and in non-anticoagulated tubes it prevents fibrin from attaching to tube closure.<sup>57,58</sup> Gentle handling and transport reduce the risk of shaken samples and subsequent hemolysis. Cushioned transport boxes should be used for long-distance transportation. DBS should be transported to the lab at ambient temperature within 24 hours of collection.<sup>59,60</sup> Desiccants should be used for biospecimens sensitive to humidity. Blood, urine, saliva, DNA, or RNA shipped by commercial courier may encounter extreme seasonal temperatures, and this factor should be accounted for in long-term large-scale multinational studies.<sup>61,62</sup> Biospecimens that must be shipped to a processing laboratory cannot be used for the most unstable biomarkers, or processing should begin in the field to ensure stability during transport.

### **Retrieval of Biospecimens**

Biobank material is precious and often difficult to replace; if its integrity is compromised, then the biospecimens become useless for the intended purpose of the research. Thus when retrieving biospecimens it is important to only retrieve the minimum number necessary. Manual retrieval of biospecimens within traditional freezers may result in thawing and refreezing, which is detrimental to biospecimen integrity and biomarker stability. Furthermore, it is a time-consuming process. The more automated equipment and LIMS systems, which are available and appropriate to use, the easier is the tracking and retrieval of samples. The ease of retrieval also depends on the methods of labeling and barcoding of the biospecimens. Some biospecimens may have a coding system identifying not only the participant but also the contents and conditions of the biospecimens.<sup>63</sup>

For biobanks that collect samples for unspecified purposes and depend on external users, an internet browser system (with biospecimen types, protocols, diseases, etc.) and a request system integrated or compatible with a LIMS are useful.<sup>64</sup>

### **Collection, Processing, and Storage of Blood Biospecimens**

#### **Collection**

Collection of biospecimens should be carried out by trained staff. Blood collection from children requires staff with specialized experience in pediatric phlebotomy.<sup>65</sup> Professionalism of the collection personnel is important to ensure quality of the biospecimens and to avoid discomfort to the participants.

Patients' willingness to participate may be negatively impacted if samples must be retaken.

Depending on the expected downstream analyses, multiple collection tube types involving different additives may be needed depending on the anticipated immediate and future analyses (Table 11.4).<sup>19,66</sup> The composition of blood depends on the order of which blood is drawn, with subsequent draws having higher concentrations of protein and calcium compared to the first draw.<sup>67</sup> Collection tubes are color-coded according to the type of additive. The general clinical chemistry laboratory recommendation is to follow a special order of draw dependent on the additive in the tube, to avoid cross-contamination among the tubes (departments of laboratory medicine have appropriate guidelines), but such a recommendation may have negligible effects on biospecimen integrity in vacutainers.<sup>68,69</sup> The same tube brand, and preferably the same lot number (depending on the length of the study), should be used throughout the study and between studies (cases/controls, collaborations), as different brands may use different additives or anticoagulants, and lot-to-lot variation may introduce bias. Expiration dates on tubes should be checked, as the vacuum in evacuated systems decreases with tube age and can affect the blood draw and the filling of the tube. Blood collection devices and components (tube stoppers, stopper lubricants, tube walls, surfactants, clot activators, and separator gels) may interfere with the endogenous analytes, add extraneous materials, or bind blood components<sup>70</sup> and thus result in a bias in downstream measurements of these elements. Differences in posture (standing, sitting, supine) cause changes in plasma volume resulting in hemoconcentration from supine to sitting to standing, and thereby increasing analyte concentrations.<sup>71</sup> Using a too-thin needle may result in hemolysis, distorting hematologic cell counts and potassium concentrations.<sup>72</sup> Prolonged use of a tourniquet results in hemoconcentration and changes in analyte concentrations.<sup>73</sup> Although thorough mixing is recommended by manufacturers of collection tubes, a study showed that lack of tube mixing did not reveal clinically significant differences in analytes concentration compared to those which were mixed.<sup>74</sup> Maintaining low air to liquid ratio may reduce oxidation of analytes. Hemolysis, icteria, and lipemia may result in spurious and unreliable test results.<sup>75,76</sup> Hemolysis may be biological due to exercise or disease, or may be caused by the collection procedure, transport, and storage of blood samples.<sup>77</sup> For an extensive discussion of the issues related to sample collection, refer to Chapters 4 and 5.

Plasma is defined as whole blood minus cells, and serum as plasma minus fibrinogen. Plasma requires less processing time and a higher yield of the noncellular fraction.<sup>78</sup> Compared to serum, plasma has a higher protein concentration and lacks interference from intracellular components (e.g., potassium) released during clotting. A rule of thumb for common analyses is that Ethylenediaminetetraacetic acid (EDTA) tubes are suitable for DNA extraction (whole blood, buffy coat), hematology (whole blood), HbA<sub>1c</sub> (whole blood), and a range of proteins (plasma). To minimize glycolysis, collection tubes for plasma glucose should be placed “in an ice-water slurry, and the plasma should be separated from the cells within 30 minutes. If that cannot be achieved, a tube containing a rapidly effective glycolysis inhibitor, such as citrate buffer, should be used for collecting the sample. Tubes with only enolase inhibitors, such as sodium fluoride, should not be relied on to prevent glycolysis.<sup>79</sup> Lithium-heparin plasma is suitable for a wide range of assays

**TABLE 11.4 Preanalytical Variables, Recommendation, and Documentation Requirements for Biobanking of Blood and Dried Blood Spots**

Step	Preanalytical Variables	Recommendation	Documentation
<b>Blood</b>			
Collection	Tube lot-to-lot variation Tube brand Tube additive (anticoagulants, clot activators, separator gels) Tube material (stoppers, stopper lubricants, walls, surfactants) Inappropriate blood/additive ratio Order of draw: carry-over	Use trained personnel Vertical, close-up position Inversion of tubes depends on tube additive. For trace elements, special tubes should be used  Fill collection tube as recommended by manufacturer Tube for coagulation or hemostasis: should follow discard tube, and kept at room temperature up to 1 h before centrifugation. Recommended order of draw (CLSI H3 A6): 1. Discard tube 2. Coagulation tube 3. Serum tube 4. Heparin tube 5. EDTA tube 6. Glycolytic inhibitor  Whole blood should not be chilled pre-centrifugation.	Tube brand and lot number Tube additive Tube material (stoppers, stopper lubricants, walls, surfactants)  Appropriate or improper filling Order of draw
	Specimen type: whole blood, plasma, serum Presence of intravenous catheters (IV) Tourniquet time Needle size Mixing Hemolysis Icteria Lipemia Clotting Posture of patient	If possible, take sample at other anatomical location. Avoid taking samples in previously flushed IV lines.	Document if blood taken from IV line Tourniquet time Needle size Mixing Hemolysis Icteria Lipemia Clotting time Posture of patient: standing, sitting, supine
Processing	Centrifugation: Brand Speed Temperature Duration Gravity Number (single or double) Broken tube	Many analytes are stable without centrifugation for 24 h. But some analytes are more labile to temperature and time between collection and centrifugation. For some analytes a contact time of plasma and cells of less than 2 h is recommended. Serum should be allowed clotting at room temperature.  Centrifugation: Separated serum/plasma should not remain at room temperature post-centrifugation for longer than 8 h, otherwise refrigerate. Post-centrifugation many analytes are stable within 48 h at 4 °C with exceptions. If assays are not completed within 48 h from collection (or from separation), serum/plasma should be frozen.  For coagulation or hemostasis: Double centrifugation may be performed. Leave at room temperature up to 4 h post-centrifugation is acceptable.	Centrifugation: Brand of centrifuge Speed Temperature Duration Gravity (centrifuge speed in g-forces) Number (single or double)

*Continued*

**TABLE 11.4 Preanalytical Variables, Recommendation, and Documentation Requirements for Biobanking of Blood and Dried Blood Spots—cont'd**

Step	Preanalytical Variables	Recommendation	Documentation
<b>Dried Blood Spots</b>			
Collection	Untreated or treated cards  Paper thickness  Environmental exposure Hematocrit  Clotting, layering, super-saturation, insufficient volume, wet, serum rings, visible traces of hemolysis, or exogenous contamination Concentration in periphery versus center of spot	Collect from: ear, heel (newborn, infants), fingertips or toe (children or adults) Volume: 50–100 µL  Dry at RT, horizontal position, 3–4 h Avoid air vents, sunlight, heat, contamination, touching, smearing Do not stack Pack when completely dry: sealable plastic bag, desiccant and humidity cards  Reject	Manufacturer of cards Untreated/treated  Drying time Environmental exposure For newborn DBS: Mother's clinical conditions and medication that potentially may affect results on DBS Document rejected biospecimens and reasons for rejection
Transport	Environmental exposure (see Boxes 11.1 and 11.2)	Horizontal position Use glycine paper between DBS cards to prevent cross-contamination Keep at RT Transport of DBS is unregulated	Transport position, temperature, duration
Long-term storage	Environmental exposure (see Boxes 11.1 and 11.2)	Horizontal position in low gas-permeable, zip-closure plastic bags with desiccant and humidity indicator cards. Low humidity (less than 30%)  For DNA punching: use DNA-punch tools	Position and packaging of paper

CLSI, Clinical Laboratory Standards Institute; DBS, dried blood spot; IV, intravenous; RT, room temperature.

See text for detailed references. Also refer to CLSI guidelines: <http://clsi.org> on blood collection (H18 A4 and H21-A5) and DBS collection (NBS01-A6). From Ellervik C, Vaught J. Pre-analytical variables affecting the integrity of human biospecimens in biobanking. *Clin Chem* 2015;61:914–34; with permission.

such as iron parameters, thyroid hormones, kidney function, liver enzymes, C-reactive protein, and other proteins. Citrate-stabilized tubes are preferred for coagulation testing and for culturing lymphocytes.<sup>57</sup> For cytokine analyses, serum tubes are preferred as anticoagulants may cause in vitro cytokine induction, thereby artificially increasing cytokine concentrations.<sup>80</sup> Coagulation testing requires special care. Excessive mixing of the coagulation tubes may result in hemolysis or platelet clumping, leading to erroneous results.<sup>57</sup> Problematic phlebotomy collections may produce spurious activation of the hemostasis system. In addition, hemolytic specimens and prolonged venous stasis may cause hemoconcentration and unreliable variations in many coagulation assays.<sup>81</sup>

DBS are an easy source of biospecimens that can be collected at remote sites in poor resource areas.<sup>37,60,82</sup> DBS consist of small volumes (50 to 100 µL) of capillary blood collected from peripheral anatomic sites (see Table 11.4) and deposited onto dedicated paper cards. Samples should dry at RT in the horizontal position for 3 to 4 hours. DBS should be rejected if they exhibit clotting, layering, super-saturation, insufficient volume, wetness, serum rings, visible traces of hemolysis, or exogenous contamination. The collection paper is inexpensive, relatively easy to manufacture, readily printed, and has good adsorption properties. DBS may be used for

various analyses (Tables 11.4 and 11.5), but there are many analytes for which DBS has not been validated. Differences in paper type, the type of chemical used for treatment of papers (if not untreated), paper thickness, blood volume applied, density, and the viscosity of blood may induce differences in extraction recovery, matrix effects, analyte stability, and chromatography effects in downstream analyses. The advantages of DBS compared to venipuncture are the low cost, the relatively painless procedure, and the ease of sample collection, transport, and storage (see Table 11.2). However, heat, direct sunlight, humidity, and moisture are detrimental to the stability of DBS biospecimens and to analyte recovery (see Box 11.3).

### Processing

Contamination from cell lysis may occur if separation of the cellular component from plasma and serum is delayed. Serum or plasma should be separated from contact with cells as soon as possible. Many analytes are stable without centrifugation for 24 to 48 hours,<sup>58,83–86</sup> but some analytes are more labile to temperature and time between collection and centrifugation. For some analytes, a contact time of less than two hours is recommended.<sup>57,58</sup> For cytokine analyses, cells must be separated from serum immediately after blood collection.<sup>80,87</sup> For viable cells, isolation may be performed immediately after collection and

**TABLE 11.5 Table of Body Fluids for Downstream Analyses**

	<b>Whole Blood</b>	<b>Plasma</b>	<b>Serum</b>	<b>Buffy</b>	<b>ACP</b>	<b>DBS</b>	<b>Urine</b>	<b>Saliva</b>
Chemistry		X	X			X	X	X
Hematology	X					X	X	
Coagulation		X						
Glucose	X	X				X	X	X
HbA <sub>1c</sub>	X					X		
Hormones		X	X			X	X	X
Inflammation		X	X			X	X	X
Cytokines			X			X	X	X
Vitamins			X			X		
Live cells				X				
Proteomics		X	X			X	X	X
Metabolomics		X	X			X	X	X
Genomics/gDNA	X			X	X	X	X	X
ccfDNA		X					X	X
Transcriptomics/mRNA	X			X		X	X	X
miRNA		X	X				X	X

ACP, All cell pellet; ccfDNA, circulating cell free DNA; DBS, dried blood spot; gDNA, germline DNA; HbA<sub>1c</sub>, glycosylated hemoglobin; miRNA, microRNA (circulating).

From Ellervik C, Vaught J. Pre-analytical variables affecting the integrity of human biospecimens in biobanking. *Clin Chem* 2015;61:914–34; with permission.

with the addition of dimethyl sulfoxide or DMSO if the sample is to be frozen rapidly after isolation, or isolation up to 48 hour at RT (but at the expense of stability of other biomarkers).<sup>87</sup> Temperature-controlled centrifuges are recommended. The complete biobanking blood sample preparation workflow can be consolidated in an automated blood fractionating system. The higher the throughput, the larger the need for automated blood fractionating systems, to ensure equal sample aliquoting with respect to volume and equal distribution of sample fraction material. The following fractions may be obtained from 9 to 10 mL whole blood: plasma (6 to 7 mL); lymphocytes and mononuclear cells (1 to 2 mL); erythrocytes and other cells (1 to 2 mL). Mononuclear leucocytes are the only cell type from blood that can be used for developing cell lines, as they are capable of continued viability and growth.<sup>27,87</sup>

Automated systems may be capable of detection of gel separators and buffy layers, as well as fractionation into plasma and serum.<sup>39</sup> Automated blood fractionation systems may also be connected to automated DNA extraction systems, which are preferable for high-throughput biobanks. This approach ensures tracking of samples, normalization, and high quality and high yield of DNA. In laboratories with low throughput or less financial resources, manual handling may be needed, but this approach increases the risk of errors. If the collection site is not close to the laboratory, it may be appropriate to perform simple processes such as on-site centrifugation, aliquoting, fractionation of serum, isolation of buffy coat and plasma, and storage of samples in smaller transportable coolers or freezers. However, more complex processes such as separation into stratified blood cells or cultures require more advanced laboratory equipment and are not suitable for smaller rural on-site processing centers.<sup>27,87</sup> Multiple aliquoting is advised to reduce future freeze-thaw cycles.

A study found that storage of blood biospecimens beyond 24 hours prior to centrifugation caused significant changes in most analytes investigated.<sup>83</sup> Delayed processing may account for a variability of more than 10% in one third of chemistry and

hematologic analytes.<sup>88</sup> Thus it is recommended to immediately separate plasma or serum from cells in order to provide for optimal analyte stability at RT.<sup>83</sup> However, if prolonged contact of plasma or serum with cells cannot be avoided, it is recommended to use serum because of the higher instability of plasma analytes.<sup>83</sup> Preanalytical processing factors for clotted biospecimens include insufficient centrifugation speed and time, an unbalanced centrifuge, and rough handling and pipetting into the cellular layer when removing plasma.<sup>57</sup> Plasma for coagulation testing should preferably be platelet-poor (platelet count  $<10 \times 10^9/L$ ), that is, spun twice at a specified speed. If coagulation testing is to be performed at some future time on frozen biospecimens, they should be stored platelet-free.

Delay before fractionation may impact transcriptomic, metabolomic, and proteomic profiles, whereas storage temperature has a lesser impact.<sup>89</sup>

### Storage

Stability studies of analytes after long-term storage compared to the fresh sample values, with estimation of recovery rates, are important to determine the effects of long-term storage on the original concentration of the analytes. Recovery rates may increase or decrease after long-term storage and thus result in either increased or attenuated risk ratios, respectively, when assessing the associations of an analyte with a disease state. Chemistry, hormone, and protein analytes are stable when serum samples are stored at  $-80^{\circ}\text{C}$  up to 13 months,<sup>90</sup> but various studies of longer-term stability of chemistry, hormone, enzyme, vitamin, and protein analytes have shown different stability patterns depending on the analyte, time, and temperature of storage.<sup>91–97</sup> No systematic influence on –omics analyses (metabolomics, proteomics, transcriptomics, epigenomics) of time-in-storage at  $-80^{\circ}\text{C}$  or in liquid nitrogen has been observed in samples collected in heparin, EDTA, or citrate stored over a period of 13 to 17 years<sup>89</sup> or for metabolomics analytes in DBS stored in  $-20$  and  $-80^{\circ}\text{C}$  for 2 years<sup>98</sup>; however, long-term storage at RT and repeated freeze-thaw cycles should be

avoided.<sup>99</sup> DNA showed sufficient yield, purity, and integrity when extracted from whole blood samples stored at RT (18 °C) using bio-stabilization technology, at low (−20 °C) and at ultra-low (−80 °C) temperatures,<sup>50,100</sup> or buffy coats stored for up to 9 years in a −80 °C.<sup>101</sup> Live cells are stable at RT for up to 48 hours, but must be either cultured or cryopreserved in liquid nitrogen to remain viable.<sup>26</sup> The transfer of thawed EDTA whole blood or buffy coats into RNA preservative offers a method to recover sufficient RNA of acceptable quality for microarray experiments.<sup>89</sup> Plasma or serum for miRNA analysis should be kept at −80 °C in RNA-free cryo-tubes or the miRNA should be extracted immediately.<sup>102</sup>

## Collection, Processing, and Storage of Urine Biospecimens

### Collection

Urine can be collected in many ways: 24-hour, spot, overnight, morning urine, second morning, or other timed collection (see Chapters 4 and 5).<sup>103</sup> Urine collection for biobanking may be used for later measurements of many analytes,<sup>104</sup> including the peptidome and proteome,<sup>105</sup> metabolome,<sup>104</sup> nucleotides, nucleosides,<sup>106</sup> RNA,<sup>102</sup> and DNA<sup>107</sup> (see Table 11.5). Urine miRNA concentrations are usually higher in patients (organ-specific) or individuals exposed to medication than in healthy individuals.<sup>102</sup> If several tests are to be performed, preanalytical requirements for those tests may be conflicting and may require either multiple biospecimen types or aliquoting immediately after collection and before processing.

Although guidelines have been developed for many immediate urine analyses (dipstick, macroscopic, casts and cells, microscopic, albumin/creatinine), optimal preanalytical handling guidelines for most biomarker studies on biobanked material are analyte dependent.<sup>103</sup>

The use of additives may be helpful for preservation of particular urine analytes during 24-hour urine collection.<sup>103</sup> There are many different preservative methods, but a universal preservative allowing complete urinalysis does not exist. Addition of preservatives and the particular type of preservatives may change urine volume and give rise to potential interference with assay methods.<sup>104</sup> Depending on downstream analyses, urine biospecimens may or may not be kept refrigerated or on ice during the collection period, a decision which will be influenced by the time until processing or storage (see Tables 11.3 and 11.6).<sup>103,104,108,109</sup> Centrifugation may result in loss of some analytes.<sup>103,104,108</sup> Urine may be contaminated by dipsticks or bacteria. As is the case for blood collection, leaching of substances into urine may interfere with assays or bind analytes; both are important considerations when working with low concentrations of analytes.<sup>104</sup> The completeness of a 24-hour urine collection is the extent to which the entire 24 hours is covered. Completeness of 24-hour urine collection can be verified using para-aminobenzoic acid (PABA), as it is completely and rapidly excreted in urine.<sup>110</sup> To compensate for the dilution of the urine in spot collections, adjustment for creatinine concentration is most often used.<sup>111</sup> Other preanalytical variables specific for urine<sup>112</sup> are listed in Table 11.6.

### Processing

Before aliquoting, the urine sample must be mixed to ensure homogeneity of specific gravity and composition of the

urine in the aliquots.<sup>104</sup> Depending on downstream analyses, centrifugation may be required.<sup>103,104,108</sup> For proteomics and metabolomics applications mild centrifugation is recommended.<sup>104,108</sup>

### Storage

Long-term storage at a temperature lower than −80 °C without additives is preferred unless otherwise specified for specific downstream analyses.<sup>104,113</sup> High long-term stability and measurement validity for numerous clinical chemistry analytes (including creatinine) stored at −22 °C for 12 to 15 years without addition of any urine preservative has been demonstrated.<sup>114</sup> For proteome and metabolome analyses, urine storage at RT causes progressive degradation of proteins.<sup>105,112</sup> Freeze-thaw cycles have minimal impact on protein profiles but repeated freeze-thaw cycles should be avoided.

## Collection, Processing, and Storage of Saliva Biospecimens

### Collection

Saliva and buccal cell samples have many advantages compared to blood collection (see Table 11.2).<sup>19</sup> Their collection involves noninvasive methods at lower cost, and they can be used in clinically challenging situations (children, handicapped, patients afraid of needles). They are safer to handle, and can be used for self-assessment and thus in cohorts using the return of samples by mail.<sup>115,116</sup> Saliva is used for a variety of analytes also measured in blood, and biomarker-omics studies, such as the metabolome, transcriptome, genome, proteome, miRNA,<sup>102</sup> and microbiome profiling in disease detection (local and systemic) (see Table 11.5). Disadvantages are the low concentration of analytes, that concentrations of analytes can vary several-fold (i.e., the matrix effects may be unpredictable), lack of approved testing for certain analytes in saliva, and lack of understanding of how saliva biomarkers relate to the serum concentrations. The advantage of extracting DNA from saliva is that the germline DNA only represents DNA from the person who donated the sample, as compared to DNA derived from blood which could originate from other persons, such as in case of multiple transfusions or bone marrow transplantation due to chimerism.<sup>117,118</sup> Furthermore, salivary DNA is a useful source of germline DNA in studies of hematologic neoplasias.<sup>119</sup>

Saliva and buccal cells can be collected in tubes or on cards. Saliva may be collected as whole saliva or gland-specific saliva (see Table 11.6).<sup>116</sup> Whole-mouth saliva collection may be obtained by different techniques in a resting mode (draining/pассивное дробление, spitting, suction, swab) or with stimulation (with sugar, paraffin gum, acid).<sup>120–122</sup> Specific glandular collection is invasive, more complex, and requires skilled personnel. Saliva should be collected at least 2 hours after eating and drinking, preceded by a mouth rinse,<sup>123</sup> and kept on ice to minimize bacterial action and proteolysis, which may affect saliva composition.<sup>124</sup> Furthermore, for proteome analyses, a protease inhibitor cocktail may be added.<sup>125</sup> Ambient temperature collection and storage devices are available,<sup>126</sup> reducing storage space, costs, and simplifying transportation requirements. Stimulated saliva produces a more variable (between-and within-subject) but higher analyte recovery compared to no stimulation.<sup>120–122</sup> Total protein concentrations (proteome) seem to be largely similar among various collection techniques, but individual concentrations of analytes are collection dependent.<sup>120–122</sup> Saliva from passive drooling tends to be

**TABLE 11.6 Preanalytical Variables, Recommendation, and Documentation Requirements for Biobanking of Urine and Saliva**

Step	Preanalytical Variables	Recommendation	Documentation Requirements
<b>Urine</b>			
Collection	Collection method: a. 24 h, spot, morning, other timed b. Spot: first void versus mid-stream	Collection depends on immediate and future downstream analyses. For chemistry analytes and future omics-studies: 24-h collection is recommended. Spot or timed urine-specimen may suffice (with creatinine adjustment) Use illustrated instructions for self-collection Vertical, close-up position	Collection method Has the patient intentionally been drinking excessive amounts of water to be able to void?
	Improper sampling technique or volume		Volume
	Environmental exposure	Avoid direct sunlight	Document environmental exposures
	Macroscopic inspection: color, turbidity, casts		Macroscopic inspection
	Diluted urine		Excessive hydration
	Dipstick components contamination of lab analyses	Do not use dipstick for same biospecimen as going to the lab	Dipstick
	pH, specific gravity, urinary tract infection, salt concentration, viscosity, blood		Document chemical measurements
	Preservative (and type) or no preservative	Preservatives may be added depending on downstream analyses, short-term storage temperature, and collection method	Preservative (and type) or no preservative
Processing	Centrifugation speed	Centrifugation: depends on downstream immediate or future analyses	Centrifugation speed, gravity and time
Transport	Temperature	Dependent on collection method and analytes	Temperature and duration
<b>Saliva</b>			
Collection	Amount and composition of secreted saliva depends on: smell, taste, blood type, flow rate, type and size of salivary gland, duration and type of stimulus, collection method	Vertical, close-up position Collect at least 2 h after eating and drinking Depending on downstream analyses: addition of protease inhibitor Collection on ice if not RT stable	Biospecimen collection device (brand, RT stable or not) Biospecimen type: whole saliva, (glandular specific saliva), or buccal cells Unstimulated or stimulated Type of stimulus Addition of protease inhibitor Short-term collection and storage temperature Use of dry ice
Processing	Centrifugation speed	Depending on downstream analyses: centrifugation	Centrifugation speed
Transport	Temperature	Transport temperature depends on collection device: if not RT stable, transport on ice	Temperature and duration

See text for detailed references. Also refer to CLSI guidelines: <http://clsi.org> on urine collection (GP16-A3).

RT, Room temperature.

From Ellervik C, Vaught J. Pre-analytical variables affecting the integrity of human biospecimens in biobanking. *Clin Chem* 2015;61:914–34; with permission.

more viscous and may be difficult to process in the laboratory.<sup>123</sup> The amount and composition of secreted saliva depends on smell, taste, disease status, drugs taken by the donor, age, sex, diet, blood type, physiologic status, flow rate, circadian rhythm, type and size of salivary gland, and duration and type of stimulus.<sup>116,123</sup>

Buccal cells may be collected using oral rinse, swabs, or cytobrushes.<sup>28,127,128</sup> The disadvantage of the oral rinse method is that participants have to swish and spit a solution, which is distasteful, and depending on the solution may give a burning sensation in the mouth.

### Processing

Depending on downstream analyses and the viscosity of the saliva, a centrifugation step may be needed, but has the risk of changing or losing some salivary components.<sup>124</sup> The most optimal centrifugation speed has not been established. Delayed processing may result in increased and decreased protein peaks, likely through digestion of some proteins by salivary proteases. Stabilizing agents may be added for long-term storage of saliva for later RNA extractions. Overall, whole saliva and oral rinse techniques are superior to cytobrushes or swabs as they provide high-quality, high-yield DNA, and provide a higher percentage of high molecular weight DNA,<sup>126,127,129–134</sup> but less than blood.

### Storage

Storage protocols may depend on expected downstream analyses. Protein profiles change with varying storage temperature, storage duration, and freezing rate, whereas freeze-thaw cycles seem to have a minimal impact.<sup>113,125,135</sup> The recommended storage temperature for protein profiling is –80 °C. No differences in mRNA, C-reactive protein, cortisol, and cytokines were noted if saliva samples were split into aliquots and immediately frozen at –80 °C, compared to storage at 4 °C for 24 hours followed by freezing.<sup>135</sup> When the Oragene self-collection kit was stored for 8 months at RT, there was no reduction in either the quantity or quality of DNA extracted.<sup>126</sup>

### Collection, Processing, and Storage for DNA and RNA Derivatives

Biological factors affecting DNA and RNA are gender (higher yields in women than in men), age (decreasing DNA yield with increasing age), body mass index or BMI (increasing yield with increasing BMI), and tobacco consumption (higher DNA yield in current smokers compared to never-smokers).<sup>136,137</sup> Powder contamination from powdered gloves may give rise to sporadic false-negative polymerase chain reactions or PCRs.<sup>138</sup>

Circulating cellular free DNA (ccfDNA) is a promising biomarker in oncology,<sup>139</sup> prenatal diagnosis,<sup>140</sup> and other disease conditions<sup>141</sup> (for additional discussion on ccfDNA, refer to Chapters 71 and 72). Most cell-free genomic DNA in serum samples is generated *in vitro* in the collection tube during clotting process and lysis of cells (contamination with germline DNA), and not *in vivo*.<sup>142</sup> Plasma is a better choice,<sup>143</sup> although serum ccfDNA has been suggested as a useful source of DNA with which to screen for post-transfusion microchimerism.<sup>142</sup> ccfDNA in plasma is not associated with gender and age, and blood donation does not appear to affect the concentrations.<sup>143</sup> K<sub>3</sub>EDTA collection tubes or cell-free DNA blood collection tubes are recommended, and hemolysis should be avoided.<sup>143</sup> ccfDNA concentrations are stable

up to 4 to 6 hours at RT or 4 °C, and there is a significant increase after 4 hours at RT or 0 °C.<sup>143</sup>

RNA is vulnerable to degradation by naturally occurring enzymes. Preanalytical collection factors affecting RNA quantity, quality, and gene expression analyses are tube type,<sup>62,144</sup> sterility,<sup>87</sup> tube additive,<sup>145</sup> biospecimen type, volume of blood, short-term storage temperature until extraction,<sup>137</sup> and lag time until extraction.<sup>146,147</sup> RNA yields when extracted from Tempus-stabilized tubes are higher than from PAXgene-stabilized tubes. However, high-quality RNA may be extracted from both tube types.<sup>62</sup> Higher RNA yields have been obtained from cord blood (three to four times higher) compared to adult blood.<sup>62</sup> Suboptimal blood volumes collected in the tubes may affect gene expression.<sup>62</sup>

Plasma, serum, and other bio-fluids contain very low amounts of RNA. General precautions should be taken to prevent RNase contamination and degradation of the RNA in biospecimens: keep nuclease-free environments for handling, use nuclease-free low nucleic acid binding plasticware and filter barrier pipette tips.<sup>148</sup> Circulating miRNAs are considered to be promising biomarkers in cardiology, oncology, and nephrology. miRNAs are challenging to work with as they are susceptible to many preanalytical variables and differences in the biospecimens and handling can affect their detection and quantitation.<sup>102,149–151</sup> miRNAs are relatively stable in plasma and serum and more stable than messenger RNA. Diet, exercise, age, race, altitude, exercise, drugs, chemicals, and smoking affect miRNA concentrations. miRNAs have large interindividual variability which are even larger in urine samples. miRNAs in fluids have cellular and extracellular (the biomarker) origins. It is the extracellular miRNA which is the biomarker. Therefore in the collection process it is important to prevent cellular contamination and hemolysis as these factors may falsely increase miRNA due to the release of miRNA from cells. EDTA or citrate collection tubes are preferred, whereas heparin is not recommended. Heparin inhibits downstream enzymatic steps (cDNA synthesis) and PCR. The blood sample must be separated immediately into plasma or serum fractions. Coagulation times and temperature differences may lead to variation in concentrations of miRNA. Plasma contains higher miRNA content than serum but there is less variation in miRNA concentrations within a properly sampled serum sample compared to plasma. A sufficient volume should be collected (neither too low nor too high). Failure to detect plasma miRNAs may be due to polymerase inhibitors rather than to the absence of miRNA.

Citrate-stabilized blood yields higher-quality DNA and RNA and lymphocytes collected for culture than EDTA blood.<sup>87</sup> Fresh and frozen whole blood have been shown to yield equal amounts of DNA.<sup>152</sup> An all-cell pellet (ACP) yields 80% of DNA compared to frozen whole blood and 99% of the yield of fresh blood. Frozen buffy coat and residual blood cells each yield only half as much DNA as frozen ACP, and the yields are more variable.<sup>152</sup> The DNA yield from DBS samples is minute, however studies have shown that DNA from DBS can be whole-genome amplified and used for genome-wide studies.<sup>82</sup>

Processing time is an important source of preanalytical variation in DNA yield. It has been shown that decreasing the lag time between blood collection and refrigeration, and between refrigeration and centrifugation, results in a substantial increase in DNA yield.<sup>136</sup> The main factor negatively affecting DNA yield is hemolysis.

Double centrifugation of plasma is recommended for cfDNA as it ensures the absence of any cells (and can be done after long-term storage).<sup>143</sup> If extraction is delayed after blood processing, plasma samples must be stored at  $-80^{\circ}\text{C}$ . If extraction is not delayed, plasma samples may be stored at  $4^{\circ}\text{C}$  for up to 3 hours. Plasma samples are sensitive to freeze-thaw cycles and should be aliquoted prior to freezing. cfDNA extracts may be stored at  $-20^{\circ}\text{C}$  and should not undergo more than three freeze-thaw cycles.<sup>143</sup>

Freezing at  $-80^{\circ}\text{C}$  is still the most common method of storing extracted DNA and RNA.<sup>113</sup> Degradation of DNA increases with dilution, higher storage temperature, multiple suspensions, and repeated freeze-thaw cycles.<sup>47,48,153</sup> Pre-analytical storage factors affecting RNA analyte quantity, quality, or gene expression are temperature, storage time, concentration of RNA, and repeated thaws.<sup>147,154,155</sup> Special technologies for storing DNA<sup>47–49</sup> and RNA<sup>51–53</sup> at RT have been developed, enabling easier shipment of these derivatives. This approach, which minimizes the required storage space and reduces electrical and shipping costs, is helpful when mechanical or cryogenic equipment is not available (e.g., during shipping or in rural areas), or may serve as an alternative method for back-up storage.<sup>1</sup> With this technology, DNA showed no degradation at RT or in accelerated aging experiments at high temperature ( $50$  and  $70^{\circ}\text{C}$ ), during an 8-month storage period.<sup>47,48</sup>

### Downstream Use of Samples

Protocols developed for one type of biospecimen research do not always apply for other types of research. Thus the decision on what type and volume of biospecimens to include in a biobank for specific purposes affects all other downstream analyses and logistical processes. Where possible, there should be consideration for the types of testing which are expected to take place, but the testing methodologies may not always be known in the planning or collection phase, a fact that may limit or impact analytical results. Considerations needed before analysis of banked biospecimens include the minimum volume and type of biospecimen needed for the assay, the stability and recovery of the analyte to be measured in frozen versus fresh state based on pilot studies, recommendations from the literature on temperature of short- and long-term storage, and duration from collection to storage.

### Pediatric Biobanking

Apart from ethical, legal, and social issues (ELSI), several practical considerations differ in pediatric biobanking compared to collection of adult biospecimens. As almost any biospecimen collection in children may be associated with anxiety and distress, careful consideration of required biospecimen volume and type must be taken into account according to assays and potential future studies. Issues concerning patients' willingness to donate are especially important in pediatric biobanking.<sup>65</sup> Other alternatives to blood from venipuncture exist, for example, blood taken from catheters already in place, DBS, saliva, urine, and hair. However, it is important to ensure that the downstream assays are validated for alternative biospecimens. Having decided upon the type of biospecimen, a suitable sample method must be chosen. It is preferable if biospecimen collection can be implemented as part of routine clinical procedures. Biospecimens should be obtained in a friendly environment by staff specially trained in pediatric collection techniques. (For additional information

on this topic, refer to Chapter 4.)<sup>65</sup> Instead of collecting additional biospecimens, the use of leftover samples should be considered. The volume of biospecimens is less than that obtained from adults, and often only limited biospecimen material remains for biobanking.<sup>65</sup> Total blood volume in children is approximately  $80\text{ mL/kg}$  (higher in neonates, and lower in adolescents).<sup>156</sup> Maximum amounts of blood to be drawn from children in one draw and per month have previously been published.<sup>65,156,157</sup> But as a rule of thumb, a volume of approximately  $0.1\%$  of body weight in one draw (except for neonates) is acceptable,<sup>65</sup> for example,  $10\text{ mL}$  for a body weight of  $10\text{ kg}$ . This is also in accordance with a guideline from a World Health Organization (WHO) review, stating that the pediatric blood sample volume limits that are consistent with physiologic "minimal risk" is approximately  $1$  to  $5\%$  of total blood volume.<sup>156</sup> Butterfly needles and pediatric small-volume blood tubes ( $0.5$  to  $3.0\text{ mL}$ ) should be used.<sup>158</sup> Repeated sampling should be avoided.<sup>65</sup> Vacuum blood collection tubes can cause veins to collapse and should only be used on older children.<sup>65,158,159</sup> Blood collection from ill children should not exceed  $3\text{ mL/kg}$  post-neonatally within 24 hours (3.8% of total blood volume).<sup>156,160</sup> In children with anemia, a special caution should be taken. Saliva from newborns may be difficult to obtain as their mouth is small and saliva volume is limited, but older children may be easier to instruct and are more cooperative.

### Security Systems and Emergency Planning

Security systems are needed to prepare for the avoidance or response to an emergency (Table 11.7 and Boxes 11.6 and 11.7). Security systems may never be 100% safe no matter how many situations they are prepared for.

During the last few years, some unfortunate emergencies have provided more insight into how to prepare for, respond to, and recover from emergencies and disasters. Disasters are not isolated events and should be expected and prepared for. Some natural disasters will tend to occur repeatedly in disaster-prone geographical areas. Table 11.7 lists the kinds of disasters that may happen and how to prepare for them.<sup>1,161–167</sup> The types of degradation that may occur with biospecimens in case of disasters are listed in Box 11.6. It is important to have protection systems and to prepare for the response and recovery of the biobank and the biospecimen collections. Researchers expect their biomaterial to be stored in safe and proper conditions. The emergency plan should cover responses to all possible disasters, including regular training of staff members. In order to prevent the complete destruction of all biobank biospecimens, it is recommended to have duplicate and split collections in geographically distant locations. Redundant systems of technology, equipment, and personnel should be maintained such as having redundant alert systems (SMS, telephone calls, email, alarms, monitoring, etc.) and remote monitoring. Back-up freezers should be in near proximity to the existing facility, as this will reduce biospecimen removal time and enable preservation of valuable samples. But it is also advisable to establish relationships internally and externally for temporary long-term storage. Removal of biobank materials is time-consuming, and usually requires two staff members at a minimum. Emptying a mechanical freezer of standard size ( $700$  to  $900\text{ L}$ , containing  $50$  to  $100,000$  biospecimens) may take 1 hour for two people and relocation time should be added depending on the location of the back-up freezers. The facility power supply should be redundant with 24-hour

**TABLE 11.7 Disasters: Natural, Man-Made, and Technical**

<b>Disaster Type</b>	<b>Protection, Preparation, Emergency Plan</b>
<b>Natural</b>	
Storms, hurricanes	Installation of isolation systems. Spare replacement parts.
Heat or drought	Redundant air conditioning, industrial fans. Choose room-temperature storage systems for biospecimen.
Cold, snow, frost	Spare replacement parts for critical or consumable items.
Flooding	Have a basement, tank, or container below lowest part of biobank. Special flooding alarm systems. Drains. Generator on top floor or rooftop.
Earthquake	Earthquake-secure buildings
Epidemic	Sterile secure storage
<b>Man-Made</b>	
Theft, sabotage	Locked doors, logs, camera surveillance
Accidental destruction	Keep duplicates
Strike	Contracts with staff
Resignation (know-how disappearing)	Education, cross-training, SOPs
Bomb threats	Back-up storage
Chemical spills, contamination	Secure containers, freezers
Communication loss	Multiple call systems
Malfeasance	Federal law
Law enforcement action	Law enforcement action
<b>Technical</b>	
Fire	Fire detection and alarms. Fire extinguishers, sprinklers, non-water-based systems. Limit flammable parts and units.
Freezer burn-down	Back-up freezers and storage space, alarm systems, monitor equipment, use different vendors and freezer systems, back-up CO <sub>2</sub> and LN <sub>2</sub>
Water damage	Generator on rooftop
Air condition failure	Industrial fans
Power outage	Back-up generators, 24-h generator support from external contractors
Explosion (possibly leading to serious injury or death)	Keep duplicate samples apart
Nuclear mishaps	No example
Vehicle accident during transport/shipping	Ship only smaller fractions, keep duplicates
Server break-down of data storage	Cloud solutions

LN<sub>2</sub>, Liquid nitrogen storage; SOP, standard operating procedure.

Data from Baird PM, Benes FM, Chan CH, Eng CB, Groover KH, Prodanovic Z, et al. How is your biobank handling disaster recovery efforts? *Biopreserv Biobank* 2013;11:194–201.

### BOX 11.6 Consequences of a Disaster to the Biospecimens

- Concentration differences due to water or evaporation
- Oxidation
- Degradation
- Evaporation
- Desiccation
- Moisture
- Sunlight (strand breaks in DNA)
- Encapsulation in ice after refreezing
- Microbiological contamination: yeast, mold, fungus, bacteria, and virus causing biological hazards
- Destroyed barcodes

Courtesy Christina Ellervik.

back-up emergency generators. A redundant fuel supply for the generators should be ensured with local back-ups or contracts with fuel vendors. In order to better predict or prevent equipment failure or replacement, it is important to maintain service contracts. Room-temperature storage technologies

should be taken into account when planning new biobanks. A prioritized list of the biospecimens and an order in which to remove them during an emergency should be prepared. To protect against theft or sabotage of biospecimens, secure locks, access codes, logs, and camera surveillance systems are recommended.

A framework for the recovery of biospecimens after a natural disaster has been described based on empirical evidence after a flooding disaster in a Danish biobank (Box 11.8).<sup>165</sup> Freeze-drying was used to remove the ice on the biospecimens encapsulated in ice after defrosting-refreezing. Next and most importantly, the integrity of the samples had to be examined to decide whether to destroy or keep the samples. The integrity of the samples may be determined by assessment of microbiological contamination (germ count) and quality testing of the samples. The microbiological analysis was followed by cleaning and reorganizing the samples to avoid both re-contamination of cleaned samples and spreading the contamination outside the biobank, and to protect biobank employees. The quality testing of the samples was done by comparing concentrations of certain

### BOX 11.7 General Security and Emergency Preparedness Recommendations

Prioritized list of samples and a plan to evacuate them  
 Redundant alert systems (sms, calls, email, etc.)  
 A back-up on-call person or 24-h team who lives very close to the biobank  
 Duplicate and split collections  
 Remote monitoring  
 Redundant alarm and monitoring  
 Redundant freezers in near proximity  
 Maintain service contracts  
 Check equipment at fixed intervals  
 Emergency lighting  
 Uninterruptible power supply  
 • Redundant systems  
 • Computer systems  
 • Electronic systems  
 • Monitoring systems  
 • Safety systems (e.g., oxygen sensors, ventilation systems)  
 • Controllers for liquid nitrogen freezers  
 Generators  
 • Have fuel supply to run continuously for a minimum of 48 h and preferably a minimum of 72 h

- Have refill fuel storage supplies
  - Keep list/contracts of vendors for further fuel supplies in case of an emergency
  - Contact vendors when emergency is alerted
- Access**
- Locked doors
  - Controlled keys/codes
  - Surveillance of entry (camera, door entry sensors)
  - Motion detectors
  - Glass break and door entry sensors
- Fire**
- Preventive plan
  - Detection systems
  - Fire extinguishing
  - Sprinkler systems
  - Non-water-based fire retardants
  - Standard operating procedures for emergency response

Courtesy Christina Ellervik.

### BOX 11.8 Biobank Disaster

#### *Flooding of the Danish Diet, Cancer, and Health Biobank*

In July 2011, the Danish Diet, Cancer, and Health Biobank flooded during an extreme rainstorm, exceeding existing safety measures. The biobank was located in the basement below sea level, and the water level reached 1.70 m in half an hour. The biobank included blood, urine, and fat tissue samples from 57,053 Danes. The samples were immediately rescued, but some samples were fully or partially thawed. The restoration of the samples included freeze-drying, cleaning (due to microbiological organisms), repackaging, reorganization, and quality testing. The restoration took 3 years (see references in the text).

Courtesy Christina Ellervik.

analytes in destroyed samples to those obtained in duplicate intact biospecimens stored elsewhere, and by remeasuring the concentrations of analytes in the destroyed samples and comparing the values with those obtained in previous measures. The expected timeframe of the recovery of 900,000 samples was 1 year, but it was actually 3 years. Furthermore, the quality testing is an ongoing process, and must be employed every time the samples are used for research.

Disaster planning should be part of the biobank's budget as well. Even if an insurance policy is maintained, the biobank should keep enough holdings to account for non-reimbursement for any losses. The preparation phase includes expenses for maintenance contracts, alarm systems, monitoring, back-up freezers, back-up storage locations, etc. The recovery expenses are primarily salary for the recovery personnel and for equipment replacement.

### Standardization in Biobanking

A biobank is technically a laboratory, and if it is established by a formal laboratory, usually accreditation, SOPs, and quality

control protocols are inherent procedures, but this is not always the case for biospecimens collected for research only.<sup>168</sup> In a survey, 80% of biobankers agreed that "To be called a 'biobank,' the management of a sample collection must follow professional standards (i.e., standards agreed to by a recognized external group such as an association, society or network)"<sup>5</sup>; the remaining respondents were unsure or disagreed.

Biobanks can adopt various types of standards for internal and external standardization. Internal standardization includes national and international best practices by ISBER<sup>1,169,170</sup> and NCI,<sup>7</sup> SOPs,<sup>171</sup> and self-assessment tools.<sup>172</sup> External standardization includes proficiency testing in comparison to other laboratories,<sup>173</sup> certification programs by the Canadian Tissue Repository Network (CTRNet),<sup>13,174</sup> and accreditation programs by the International Organization for Standardization (ISO) 20387 standard<sup>175</sup> or the College of American Pathologists Biorepository Accreditation Program (CAP BAP).<sup>176</sup> A detailed comparison on how the certification and accreditation programs differ is described in several papers,<sup>177,178</sup> but the programs have overlapping and unique independent features. For pharmaceutical biobanking, additional Good Laboratory Practice (GLP)<sup>179</sup> and Good Manufacturing Practice (GMP)<sup>180</sup> apply.

SOPs should accurately and unambiguously describe the biobanking tasks to be performed (see **Points to Remember** below). Deviations from, differences in, and poorly documented SOPs may affect test results, interpretations, and conclusions, and can potentially also lead to a misdiagnosis, inadequate treatment, and compromising patient safety. The more complex the preanalytical steps are for biospecimens in a biobank, the higher the risk that errors will occur, and the greater the need for automating the process. The smaller the study the larger the effects of not standardizing will be felt.

Biobanking management is related to processes developed for laboratory proficiency. According to health regulations and accreditation requirements, laboratory staff needs to be educated in basic laboratory processes to assure safety and

quality of the biobanking-related processes. For biobank purposes, additional specific biobank training is advised. A series of papers outlined several biobanking education programs.<sup>181</sup> According to ISBER, a training coordinator should be responsible for monitoring, training, and maintaining appropriate training documentation of all employees.<sup>1</sup> ISBER endorses two biobank-specific courses, sponsored by the University of British Columbia, Office of Biobank Education and Research (OBER), and CTRNet, but other courses are also available.<sup>181</sup> A 2-year Master's Degree in Management and Biobanking is available in Lyon France through Catholic University. The course objectives are to communicate the standards to biobank personnel, researchers, clinicians, and trainees in new and existing biobanks in order to enhance quality. The educational programs cover key concepts in establishing, maintaining, and using biobanks, for example, biospecimen handling, data management, ELSI, quality management, and governance. ISBER and the American Society for Clinical Pathology Board of Certification have partnered to create the Qualification in Biorepository Science program (QBRs).<sup>182</sup>

### POINTS TO REMEMBER

- Some key standard operating procedures for biobanks<sup>a</sup>
- Administration
    - Job descriptions, roles, and responsibilities
  - Participants and recruitment management
    - Obtaining informed consent
    - Withdrawal of consent
  - Records and documentation management
    - Information access control
    - Database back-up systems
  - Facilities management/operation
    - Emergency procedure for freezer and refrigerator failure
    - Maintenance of sample storage facility and equipment
  - Quality assurance procedures
    - Assessing quality of nucleic acids
  - Safety
    - Handling hazardous chemical waste
  - Training
    - Education and training
  - Materials handling and documentation
    - Labeling and tracking materials
    - Inventory verification
    - Specific standard operating procedures for collection, processing, aliquoting, and storage depending on specimen type
  - Material release
    - Sample shipping and transportation
    - Completion of material transfer agreement
    - Material request and release
    - Return of biospecimens for clinical purposes

<sup>a</sup>As suggested by the Canadian Tissue Repository Network (CTRNet).<sup>174</sup>

## BIOBANK ADMINISTRATION

### Laboratory Information Management System and Labeling

A laboratory information management system (LIMS) is a software or web-based system that integrates information about various aspects of laboratory informatics. LIMS features related to biobank functionalities are listed in Box 11.9.

LIMS are usually commercially available, but some biobanks develop their own customized systems. The benefit of the LIMS is the automated high-throughput data management allowing efficient and accurate sample organization, traceability, and management thereby improving productivity. Furthermore, it eliminates manual error-prone processes by simplifying and automating data administration and thereby improving data reliability. The LIMS makes it easier for quality management of the biobanking processes with the detailed logs and the ability to produce reports and statistics. These metrics are also important for the assessment of the sustainability of the biobank (see later). In a multicenter biobank, LIMS systems should, if not obtained from the same company, at least be compatible with each other. This is also applicable to when information is gathered from different systems, for example, hospital lab systems. Note that in some biobanks these systems may be referred to as BIMS, biobank information management systems, in recognition of the inclusion of additional historical information about biospecimens, that is, history of collection, processing, and storage.

Labeling and coding are essential for further tracking and retrieval of the samples. Labels should be unique, highly adhesive to the primary and secondary tubes, and cryostable. Orientation, placement, and size of labels, as well as density, resolution, and type (one-dimensional or two-dimensional) of the barcodes should be carefully chosen. Barcodes on secondary tubes should be two-dimensional and preferably should be molded into the tube to ensure the longevity of the coding. Specifying types of identifiers, institution names, and dates, using multiple identifiers and alphanumeric codes should minimize errors.<sup>183</sup> The ISBER Biospecimen Science Working Group has developed a “Standard PREanalytical Code” (SPREC) to facilitate annotation of the main pre-analytical factors in a coded labeling system.<sup>184</sup>

The overall biospecimen rejection rate in a clinical chemistry lab is approximately 0.2 and 7.6% of these are due to improper labeling (incompletely labeled, mislabeled, not labeled, label removed, label destroyed). In these situations, resampling is the only solution. Relabeling is strongly discouraged.<sup>185</sup>

### Annotating Clinical Data to Biospecimens

The need for data annotation to accompany the biospecimens depends on the research design, the coding of the biospecimens, data availability, consent from donors, institutional review board (IRB) and ethical approvals, and the biobank's economic situation. Examples of some types of data annotation are listed in Box 11.10. Coding of biospecimens for obtaining follow-up disease endpoints must be designed in a way that a link from the coded ID to the patient's original ID can be obtained by an authorized data manager. Data availability is different between countries according to various privacy rules and regulations. In Scandinavia, it is possible to link the individuals' unique personal IDs to the national demographic, social, and health registries, including, for example, information on vital status, nationality and birthplace, International Classification of Diseases (ICD)-coding of any disease diagnosed in or out of hospital, causes of death, surgical procedures, births, pathology, medication, genetic diseases, and lab measurements.<sup>186</sup> However, many countries do not have registries and may rely on hospital records or self-reported endpoints from questionnaires. To obtain data annotation, there needs to be consent from the participants, and IRB and ethical approvals.

### BOX 11.9 Laboratory Management Information Systems Features and Benefits Specific for Biobanks

#### Features

Pre-preanalytical variables (scheduling)

- Requests and reception
- Barcoding and labeling (printing and scanning)
- Participant data
  - ID
  - Age and sex

Preanalytical variables

- Biological and environmental variation

• Collection, for example,

- Allows for multiple collection sites
- Date and time of collection
- Specimen type, volume

• Processing

• Storage

- Hierarchical location

• Transport

• Retrieval

• Distribution

Analytical variables

• Plate handling

• Instrument integration

• Method

- Reference materials
- CV% (intra-assay, total)

Postanalytical variables

- Inventory system
- Plate handling

Data management

- Data from laboratory instruments

- Data from other sources (e.g., clinical)

#### Benefits

Workflow efficiency

Higher throughput

Data reliability

Simplifies administration

Logistic and operational support

- Complete and accurate biospecimen traceability

- Biospecimen organization and management

Regulatory compliance

Improves productivity and accuracy

Clinical data annotation

Elimination of manual and error-prone processes

Quality control

Statistics and reports

LIMS, Laboratory management information systems; CV, coefficient of variance.

Courtesy Christina Ellervik.

### BOX 11.10 Type of Clinical Data Annotation

#### Disease Endpoints

Types:

- ICD diagnoses
- Surgical procedures

Way of obtaining the endpoints:

- Electronic patient records
- Registers
- Questionnaire
- Interview

#### Questionnaire

Socio-demographic

Lifestyle

Health

Environmental

Race

Family history of above

#### Laboratory Data

Common clinical chemistry measurements

Genetic (DNA/RNA)

Microbiological

Pathological

Other biomarkers

#### Phenotypic Data (Any Health Examinations)

ECG

Spirometry

Blood pressure

Anthropometric measures (height, weight, waist and hip measurements)

Courtesy Christina Ellervik.

Furthermore, data annotation may be costly due to the time-consuming and staff-dependent collections, administration, and validation of data, the service fees for registry access, and the costs for the server space and for developing the IT infrastructure, data privacy, and security procedures.

#### Publication Standards

Any part of the biobanking process should be documented in detail, to ensure valid, reliable, and comparable results within and between studies.<sup>187</sup> The information specifically concerning biospecimens' preanalytical conditions reported in scientific publications varies considerably.<sup>188</sup> The preanalytical

data should therefore be documented simultaneously with the biobanking process by the investigators, and reported in the literature<sup>189,190</sup> to serve as quality indicators in order to be able to evaluate, interpret, compare, validate, and reproduce the experimental results.<sup>191,192</sup> The *Biospecimen Reporting for Improved Study Quality Guidelines (BRISQ)* and *Reporting Recommendations for Tumor Marker Prognostic Studies (REMARK)* are useful tools for this purpose.<sup>188,190,193,194</sup> CAP has specified a list of 170 semiquantitative preanalytical variables that could be used to annotate biospecimens at the time of collection in order to document sample integrity.<sup>191,192</sup>

## Costs

Establishing, operating, and maintaining a biobank is costly, and different financing models have been published.<sup>195–199</sup> Planning the budget and the financing of a biobank is usually an iterative process with different scenarios that are analyzed to derive the most suitable and desirable model. Financing a biobank is not a static process with a one-time investment but a dynamic process. With technological advancements in biospecimen science, downstream assays, and biorepository infrastructure needs, new investments are needed to keep the biobank processes, equipment, and infrastructure current. Furthermore, with an increasing number of researchers, complex datasets, analytical results, and national and international data sharing, investments in IT security to protect privacy is a matter of concern and utmost importance.

The important components of the biobank expenses are summarized in At a Glance: Biobankconomics and apply to the biobanking processes and equipment. Different financing schemes will apply depending on whether the biobank is project-based (e.g., a biobank for each project) or storage-based (e.g., a storage warehouse with core lab biobank functions for several projects). When planning a biobank, it is important to determine the sources for financing. The approach to financing will depend on the aim of the biobank, the size, the number and type of projects, etc. Funding biobanks is a challenge as their scope is not always defined or too broadly defined, which doesn't fulfill the funders' need to focus on the endpoint of research results. Funders often assume that initial start-up and infrastructure costs and initial operational costs of individual biobanks are sufficient and that the biobanks will become self-sustaining.<sup>200</sup> However, it is crucial for the sustainability that maintenance and other costs are at least partially recovered. Usually, if there isn't a sustainable funding source, costs are included as part of other focused research grant proposals, which makes the sustainability of the biobank a fragmented process. Funding for biobanks may include governmental, hospital, private, and public sources (Box 11.11). Quality in the biobanking process is also important as this will reduce the number of repeat collections and measurements, increasing power, decreasing the need for over-collection of participants and samples, thereby reducing costs. Thus the future aim for academic research studies should be convergence to industry standards for biospecimen quality. The costs associated with additional staffing and infrastructure related to return of research results and incidental findings in genomic studies should also be taken into account.<sup>200</sup>

A business plan is also needed (Box 11.12). Only approximately one third of existing biobanks were started with a business plan in place.<sup>3</sup> A business plan should include a description of the biobank's purpose, the customer base, a description of the products (core laboratory service, etc.), and a market analysis.<sup>200</sup> Different user groups may include academic and industrial partners locally, nationally, or internationally.

The majority of academic biobanks belong to more than one organization and thus financing is usually a patchwork from different sources. In a survey of US biobanks, the largest funding sources for biobanks are the federal government (30% of all biobanks), the parent organization of the biobank (30%), fees for services (11%), and individuals or foundations (10%).<sup>3</sup> Other sources of funding may include state government, clinical awards, sale of specimens or other

## BOX 11.11 Funding Sources

### Fund

- Private
- Public
- Governmental
  - Local
  - State/regional
  - National/federal
  - International (e.g., European Union funding)

### Industry

### Hospital

### Host-institution (may be one of the above)

### Researchers

### Patient associations

### Physician associations

### Individual donations

### Biobank networks

### Biobank activities:

- Providing consultancy service
- Providing laboratory service (processing, analyses)
- Leasing storage space to external researchers
- Providing emergency back-up support for other biobanks
- Fees for biobank services (processing, storage, transport, retrieval)

Courtesy Christina Ellervik.

## BOX 11.12 Business Plan for Biobank Warehouses

### Biobank description

- Biobanks purpose
- The customer base
- What is offered?
- How does the biobank differ from others alike?

### Market analysis

- Who are the stakeholders?
- Who are the competitors in the market?
- Is it a viable and feasible idea?

### Marketing and sales strategy

### Description of the biobank organization and management

### The products and services

- Core lab service
- Specimen request
- Storage service
- consultant service

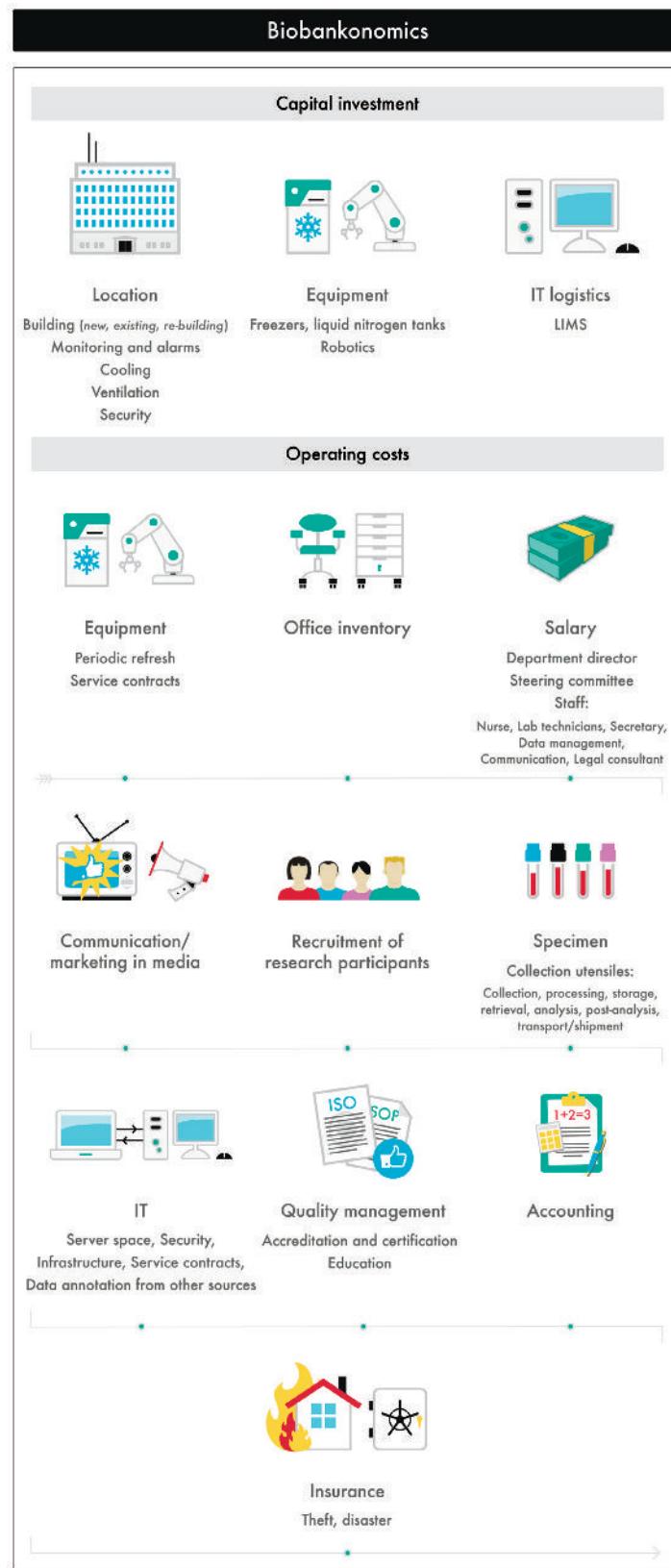
### Funding (what and how)

### Financial projections

Licenses, permits, consents, institutional review board approvals, etc.

Courtesy Christina Ellervik.

products.<sup>3</sup> The majority of biobanks charge commercial customers more than their academic customers (66%) while 30% just recover their expenses.<sup>201</sup> Furthermore, financial sustainability for most academic biobanks is dependent on institutional support as cost recovery models are not aligned with the concepts of these biobanks, which are primarily for the public good (personalized treatments and diagnostics) at no charge.<sup>195</sup>

**AT A GLANCE*****Biobankconomics*****Capital Investment and Operating Costs in Biobanking.**

Courtesy Christina Ellervik.

**BOX 11.13 Bankruptcy of a Biobank**

The Icelandic genomics company deCODE Genetics Inc. was founded in 1996 to identify human genes associated with common diseases and to apply the knowledge in guiding the development of candidate drugs. Their database covers 140,000 people. The mutations that deCODE detected explained only a small fraction of each major disease and could not support the development of diagnostic tests or drugs. In 2009, the company filed for bankruptcy protection after 13 years of failing to make a profit. In 2012, deCODE became a subsidiary of Amgen for \$415 million. In 2013, Amgen spun out its comprehensive clinical data IT infrastructure and interpretation suite as NextCODE Health. In 2015, NextCODE was bought by WuXi PharmaTech and rebranded as WuXi NextCODE Genomics (see references in the text).

Courtesy Christina Ellervik.

Regardless of the funding source, funders of biobanks should always obtain an auditor-signed accounting of finances by the end of the year. Funders should be continuously informed of the financial status and the progress of the biobank (at least once a year). An example of lack of financial sustainability is provided in [Box 11.13](#).<sup>202–204</sup>

### Sustainability

Biobank sustainability can be expressed in financial, operational (resources, technical and management aspects, environmental), and ELSI terms.<sup>201,205,206</sup> Sustainability implies providing the best methods and quality within these frameworks to remain productive with appropriate systems and processes to assure a long-lasting lifetime for the biobank. The opposite of a sustainable biobank is a biobank system that fails in governance, biospecimen quality, and data privacy or other important components of sustainability. Sustainability is important for the public (donors) for their willingness to participate, for funders to be assured of secured investments, and for researchers to ensure specimen quality.<sup>199</sup> Failure of any of these frameworks is a risk for all biobanks, and a sustainability plan should be part of the business plan.

*Financial sustainability* focuses on the long-term financial viability of the biobank. This is achieved by balancing income and costs to the extent possible, and a thoughtful business plan to minimize depletion of economic resources. Biobank metrics is a means of achieving financial sustainability by making biobank activities (performance metrics) transparent for funders and assuring them of the viability of their investments.<sup>207,208</sup>

The metrics may be quantitative and encompass any statistics on collections, processing, storage, distribution, etc., and qualitative metrics may encompass scientific progress reports (publications [number and journals], impact factor of journals, research citations, h-indexes for researchers, grants, collaborations, etc.).

*Operational sustainability* focuses on the operational management of the biobank, that is, any “input, internal, or output components of the biobanking process.”<sup>199</sup> This includes technical processes (preanalytical, analytical, postanalytical, storage, transport) and “green solutions” to help ensure resource sustainability, meaning that the specimen resources

are used without being depleted, degraded, or destroyed. But it also includes staff management (lean processes; for additional discussion on the lean process, refer to Chapters 3 and 6), biobank networks, and consortia. Quality management systems, accreditation, best practices, and guidelines are resources to help ensure operational sustainability. Metrics also provide feedback of operational processes in the operational framework.

*Social sustainability* focuses on the acceptability of the biobank by the public, and by a commitment to accepted standards of practice.<sup>198,209</sup> Donors require transparency about their biospecimens, and the assurance of privacy related to the information derived from and annotated to the specimen, and that the biobank complies with ethical and legal standards.<sup>199</sup>

## ETHICAL, LEGAL, AND SOCIAL ISSUES AND POLICY

The yearly distribution of published and cited articles related to ethics in biobanks has been steadily increasing during approximately the last 20 years.<sup>210</sup> The topics of investigation and discussion are consent, privacy, return of results to participants, public trust, consent of children, commercialization, the role of ethics boards, data sharing and exchange, and ownership. ELSI are continuously evolving and contribute to the persistent lack of international coordination and harmonization of biobanking practices.<sup>45</sup> Several cases have been taken to court due to disputes over ELSI issues.<sup>211–213</sup> In the following, we focus on general ELSI issues and not on those that are country specific.

### Consent

The Declaration of Helsinki was developed by the World Medical Association as a set of ethical principles regarding human experimentation but is not legally binding.<sup>214</sup> It has been revised six times and has greatly influenced national and international laws on ethical research principles involving human subjects research.

Biospecimen collection, biobanking, and research protocols must be approved by an IRB or research ethics committee. The researcher should provide specific, sufficient, and transparent information about the research project (purpose, methods, risks, benefits, biobanking, use of biospecimens, responsibilities of patient and researchers, compensation, withdrawal of consent, contact person, etc.) for the participant to give a voluntary written “informed consent.” Few biobanks use electronic consent, but there seems to be an overall interest among US biobanks to explore the implementation issues, including user preferences and receptivity to such methodology.<sup>215</sup> [Table 11.8](#) lists benefits and challenges for electronic consent among participants and researchers.

There is still much debate on whether the consent form should be:

- Broad and open, permitting all future uses that an ethical review board agrees is scientifically solid and ethically defensible (global consent)
- Specific and limited with consent only to current use of biospecimens in a particular project (specific consent)
- Stratified for different purposes, for example, any future use, future use within original study topic, future use outside original study topic, future use with new consent, use of electronic health records, etc. (tiered consent).<sup>213</sup>

**TABLE 11.8 Electronic Consent**

	<b>Benefits</b>	<b>Challenges</b>
Participants	Convenience No pressure More informed No travel	Consent discussion difficult Questions about confidentiality and privacy
Researchers	Convenience Higher enrollment Less cost Less paperwork Increased capability	Verification of participants Is the electronic consent acceptable and valid?

Courtesy Christina Ellervik.

#### BOX 11.14 Consent

##### *The Havasupai Indian Tribe Case.*

Havasupai Tribe v. Arizona State University,  
Case No. CV2005-013190,

Superior Court of Arizona, Maricopa County

In 1989, a member of the Tribe asked researchers at Arizona State University (ASU) to look into a perceived “epidemic” of diabetes among tribal members. Although the researchers at ASU expressed an interest to broaden the research, the Tribe was likely not interested in a study exploring other issues, but they did not disclose the possibility. In 1990 ASU researchers collected 200 blood samples from the Havasupai Indian Tribe for a study of diabetes. The consent form focused on “behavioral/medical disorders,” but the oral information to the tribe members focused on diabetes. The tribe members objected to the use of biospecimen for research in schizophrenia, inbreeding, evolutionary genetics, and other projects stigmatizing the tribe. In 2004 the tribe filed a lawsuit against ASU for “breach of fiduciary duty, lack of informed consent, fraud, misrepresentation, fraudulent concealment, intentional infliction of emotional distress, negligent infliction of emotional distress, conversion, violations of civil rights, negligence, gross negligence and negligence per se.”

The case was settled out of court on April 20, 2010. ASU agreed to pay \$700,000 to tribe members, provide other forms of assistance, and return the blood samples to settle the legal claims that the university researchers had improperly used tribe members’ blood samples in genetic research (see references in the text).

Courtesy Christina Ellervik.

Consent issues were the major part of a lawsuit filed by the Havasupai Tribe against Arizona State University in 2004 (Box 11.14).<sup>213,216</sup>

In order to facilitate and maximize research and use of biospecimens and data, many funding agencies, researchers, and regulators encourage broad consent, meaning unrestricted use with no specified end date.<sup>3</sup> Reasons for the broad consent are that details about future research projects or technology usually cannot be provided at the time of consent. Requirements that have been suggested to be fulfilled for the broad consent and consent to future research are privacy and safety policies, the right for the donors to withdraw consent, and for each new study to be approved by the IRB.<sup>217</sup>

The specific and limited consent form may impede international sharing of samples and data.<sup>218</sup>

A survey of the Swedish general public revealed that the majority prefer a general consent, are willing to delegate some decisions to the IRBs, and would allow storage of their samples as long as they are useful for research.<sup>219</sup>

Stratified/differentiated/multilayered consent (the terms are many) cover the principles of a document which asks for multiple consents on one form.<sup>220</sup> The questions may cover genomic research, return of results, approval of registries, approval of transferring information to other health authorities for treatment purposes, and future research studies.

The opt-in method of consent is the active provision of permission by the participant. The opt-out method of consent is the opportunity to actively opt-out of a project in which you will be enrolled if you do not actively decline to participate. The opt-in method is used in many cohort studies and trials. The opt-out method is used within many health care systems where biospecimens taken for clinical purposes and already linked to clinical data may also be used for research, for example, residual DBSs in newborn screening. The advantage of the opt-out approach is that it is a simple process often leading to high enrollment, but the disadvantage is the enrollment of participants who did not wish to enroll.<sup>221</sup> In surveys about public opinions to consent, 80% supported opt-in compared to 69% opt-out,<sup>221</sup> 0.1% refused either the storage or use of their samples,<sup>222</sup> and 0.005% of those who had previously consented withdrew their consent.<sup>222</sup> Motivations for nonconsent are usually concerns about integrity, privacy, and about the ethics or purpose of the project.<sup>223</sup>

#### Participants Legally Incapable of Giving Informed Consent

The researcher should provide information as described above, seek the participant’s assent, consider their preferences and best interests, and obtain appropriate permission from a legally authorized person (if a substitute consent is permitted or required by law).

#### Minors

Policies between countries addressing the management of biospecimens from minors vary widely.<sup>224,225</sup> There is an overall agreement that children comprise a temporarily vulnerable research population, as they lack capacity for consent until they become adults.<sup>226</sup> Even though parents may give a proxy consent for their children,<sup>65</sup> there is general agreement that “access to samples and individual DNA sequence data from children included in population biobanks should, when feasible, await their reconsent as adults.”<sup>226</sup> Involving the child providing assent should be considered, but this approach is associated with other ethical issues, as the appropriate age for such an assent is variable, and the capacity for some children to understand the information is limited.<sup>65,227</sup> However, waiting for the children’s own consent as adults may negatively impact research, as DNA samples and data may not be traceable decades later, some participants do not want to reconsent, and some are no longer traceable. The projected logistics and governance can be complex and costly.<sup>226</sup> If additional biospecimens are taken at the same time as routine collection, it should be made clear to potential participants that extra biospecimens are being taken for research.<sup>65</sup>

### BOX 11.15 Definitions of Terms Used for Degrees of Identifiability of Data or Biospecimen

**Anonymous**—Immediately coded with irreversible link to patient data, is not traceable to its original source or archeologic samples.

**Anonymized**—Clinical data are annotated to the biological material, but any identification of the participant is stripped after collection either irreversibly (unlinked anonymized) or reversibly (linked anonymized) identified by a code to which researchers do not have access but data manager does.

**Coded Samples**—The same as linked anonymized but the researchers and users have access to the code.

**Completely Identifiable**—No coding, example pathology departments.

### Waiver of Informed Consent

Use of biospecimens for additional research is usually allowed by the IRB if the participants give new additional informed consent or if the IRB waives informed consent for future research projects.<sup>211</sup> The IRB is more likely to provide approval if the participant originally consented to future research. An IRB can determine to waive consent for research on clinical biospecimens. Under some circumstances, the IRB also has the authority to waive or alter the consent procedure.

### Privacy and Protection of Data

The definition of privacy ranges from irreversible identification to complete identification.<sup>220,228</sup> The terminology employed to describe degrees of identifiability varies in the literature, as some terms are used with different meanings in various guidelines and journal articles. Box 11.15 provides a list of definitions of some of the terms as they are most commonly used. Coding and encryption of identifiable information (e.g., social security number, name, address) is a method to protect privacy. However, identifiability is a balance of individual rights versus research progress.<sup>229</sup> It is generally not recommended to irreversibly strip identifiers as this would decrease the scientific value of the data and biospecimens.

In general, legislation on privacy and protection of data addresses issues of privacy and confidentiality in protocols regarding human subjects,<sup>230</sup> and sets standards for the protection of the privacy of individually identifiable health information, for the security of electronic protected health information, for the notification of a breach of unsecured protected health information, and for the protection of identifiable information being used to analyze patient safety events and improve patient safety.

International data sharing and sample exchange between research teams from different and distinct jurisdictions have made data management and security more complicated. The UNESCO Bioethics Programme was created in 1993 and has settled on a number of rules in bioethics describing human rights when human genetic data are collected, processed, used, and stored in order to ensure dignity and freedom.<sup>231</sup>

### Return of Research Results in General and Return of Incidental Findings

Return of results in research studies encompasses both clinical and research results. Return of results on an individual

basis can be delivered in many ways, by linkage to electronic health records, cloud services (e.g., 23andme), or by postal mail. Return of results on a general basis to all participants can be delivered by email in the form of newsletters with some summary statistics or linkage to research papers. The company 23andme even informs participants in which papers their specific data had contributed.

Three major ethical principles underlie the return of results in any kind of research: (1) justice, (2) beneficence, and (3) respect for persons.<sup>232</sup> Justice is the obligation to share the benefits of the research with society. Beneficence means to maximize the benefits that can be attained by the research participants. The principle of respect for persons “includes recognition of the integral role of participants in research and underlies the responsibility to return research results.”

Genomic sequencing is increasingly used in genetic research due to increasingly affordable techniques.<sup>232</sup> The data generated from these analyses may be interpretable, whereas some are un-interpretable incidental findings, such as data of unproven clinical validity and utility.

A survey of bioethical policies from international organizations (e.g., World Medical Association, Council of Europe, OECD, etc.) on return of results found that there is an ethical duty to return *general* and *individual* research results.<sup>232</sup> However, the phrasing in the various guidelines differ with some presenting return of results as an obligation and others as an option. Also, the survey did not find any consensus among the international organizations regarding the modalities of sharing the research results, for example, how, when, and with whom they should be shared.<sup>232,233</sup> Thus there is no consensus on the length of time that a research result is “returnable.” There is also no consensus on whether research results should be disclosed to family members if the index participant is deceased.

Due to these discrepancies, a simple, flexible framework for the evaluation of return of individual research results based on the ACCE model, described below, has been proposed for genetic tests, but it can be applied in any research setting<sup>232</sup>:

- A, Analytic validity: that is, the laboratory analyses must be of high accuracy and reliability
- C, Clinical validity: that is, there must be sufficient evidence and agreement of consistency and accuracy for the prediction of outcomes
- C, Clinical utility: that is, there must be sufficient evidence and agreement that the results will significantly improve health of the participant (e.g., by prevention or intervention)
- E, ELSI: that is, the ethical, legal, and social issues associated with return of the results should be critically discussed and evaluated.

According to the above model and when it is possible to re-identify participants, a panel of experts has identified the following responsibilities for biobanks in terms of genetic research testing<sup>234</sup>:

- Clarify the criteria for evaluating findings and roster of returnable findings,
- Analyze a particular finding in relation to these criteria,
- Re-identify the individual contributor, and
- Recontact the contributor to communicate the finding.

Return of results from deceased participants to their families raises many ethical questions as the framework for privacy and confidentiality is generally absent in the postmortem

context. Deceased and living individuals may have competing interests in the disclosure of research results, but generally the benefit of disclosing is in favor of the living family members rather than the deceased participant. A framework for the decision to disclose or not has been described.<sup>235,236</sup>

When asking participants' preferences for disclosure of research results to the participants or to the participants' relatives after death, most prefer to disclose in either situation.<sup>237</sup>

Despite these preferences by the participants, most biobanks do not address the return of individual research results and incidental findings in their publicly available documents.<sup>233</sup> The most common barriers to return results are: logistical/methodological, financial, systems, regulatory, and investigator capacity.<sup>238</sup>

### Legal Issues

Legislation on biobanking differs between countries and continents. When setting up a new study or biobank, it is important to comply with international (e.g., EU), national/federal, and state/regional laws.<sup>211</sup>

Biospecimens may be obtained from donations from living individuals for research projects, from residual biospecimens taken for clinical purposes (i.e., diagnosis or treatment) and then subsequently selected for research purposes, and from postmortem biospecimens.<sup>211</sup>

Residual biospecimens taken for clinical purposes (diagnosis or treatment) and then subsequently selected for research do not usually require consent, but do require IRB approval, and the IRB is authorized to waive consent for research if samples are de-identified.<sup>211</sup> Retainment of residual newborn screening dried blood samples are done for various purposes, including program evaluation, quality assurance, and biomedical research, but legislation varies among states in the USA and among countries globally.<sup>224</sup>

In many countries, a uniform anatomical gift act (UAGA) covers donations of body or body parts postmortem by deceased donors for the purpose of education of medical students, transplant, and research by testament (i.e., will).<sup>211</sup>

### Governance, Ownership, Custodianship/Stewardship

The ISBER<sup>1</sup> has published best practices and guidelines on governance, ownership, and stewardship of biobanks. Governance includes the oversight and protection of biospecimens, that is, ethical/legal, scientific, technological, management, and economical issues. Those governing bodies are responsible for participants, researchers, society, sponsors, etc., for the best practice to handle biospecimens and data according to legislation and guidelines. Depending on the size and scope of the biobank, governance may be driven by the principal investigator, a board or committee, the hospital management, etc.

Stewardship or custodianship is the management of the biobank, including storage, access, and release of biospecimens, but the stewards themselves do not have influence on the decision to use and how to use the samples.<sup>239</sup> Principles for stewardship should be defined by those who are responsible for management of the biobanks.

Ownership may include donors, collectors, investigators, researchers, industry, hospitals, steering committees, patient advocacy groups, state/regional/federal governments, or others. The perception that the investigator who is responsible for the collection of specimens for a study "owns" the samples

### BOX 11.16 Ownership

*Washington University, v. William J. Catalona. Appeals from the United States No. 06-2301 District Court, 20 June 2007.*

Over 25 years, Dr. William Catalona set up a biobank including more than 100,000 serum samples, 4400 DNA samples, and 3500 prostate tissue biospecimens at the Washington University (WU) in St. Louis, MO. The samples were collected via informed consent. The court was asked by WU to determine the ownership of biological materials contributed by individuals for the purpose of genetic cancer research and currently housed on the campus of WU. As he was considering accepting a position at another clinical center, Dr. Catalona claimed that the contributing individuals had declared the direct transfer of their biological materials to him. Dr. Catalona also moved for an order prohibiting WU from utilizing, disseminating, transferring, or destroying the biological materials at issue. The district court concluded WU owns the biological materials and neither Dr. Catalona nor any contributing individual has any ownership or proprietary interest in the disputed biological materials. The ruling was upheld by the Supreme Court in 2008 (see references in the text).

Courtesy Christina Ellervik.

is a long-standing issue in biobanking and has been the topic of court cases in the United States.<sup>212,240</sup>

The NCI Best Practices adopted the term custodianship to indicate the biobank's role in maintaining collections, indicating that the specimens are not under the biobank's ownership but instead are in its temporary custody until study goals are met. Thus the owners transfer their biobank property to the custodians, who physically possess the biospecimens for the benefit of the owners. The owners have the right to use and reclaim their biospecimens at any time; the custodians do not. The question about ownership versus custodianship was the issue of a court case between a researcher who had collected biospecimens and the university who held the biospecimens in a biobank (Box 11.16).<sup>211,212,241</sup> Governance, stewardship, and ownership should be transparent for all stakeholders in biobanks, that is, in legal documents, public websites, and information and consent forms.

### Access to Samples and Data, Agreements, and Sharing

Access to biospecimens and data in biobanks depends on the type, purpose, the sustainability plan, and the agreements among the stakeholders (patients/donors, collectors, funders, receivers, institution) of the biobank. Some biobanks have predefined projects and specimens are not available for most requests, whereas other biobanks maintain collections for future unspecified projects and are open to requests. Irrespective of these policies, any biobank has to facilitate research as much as possible but consider the expense of using precious material<sup>242</sup>; thus using or keeping material is a delicate balance. The drawback of using material too early may be misuse of biospecimens for random research focuses without having a sustainability plan. The drawback of keeping material stored for long periods without using it is hindering of research and the risk of decreased quality of biospecimens with long-term storage. Thus to maximize the "return of investment" it is advised to maximize availability within the limits of a sustainability plan.

The UK Biobank is an unparalleled resource of open-access strategy to data upon application.<sup>243</sup> This strategy has resulted in a worldwide publication rate of 1000 articles since the biobank was opened to researchers in April 2012 until January 2020. With numbers steadily growing, there are currently 4000 international and UK researchers who are using the biobank. Researchers mainly apply for data including genetics (95%), whereas only few apply for samples (4%) or recontact of participants (1%).<sup>243</sup> However, other biobanks are more restrictive on access to data and samples for various reasons including sample availability, ethical issues, and privacy issues.<sup>244</sup>

The sustainability plan for usage of biospecimens and data should include agreements on data sharing, access, data preservation, data release, and access agreements for biospecimens and data. Usually an access committee responds to requests; the access committee can be the steering committee, or an independent, scientific advisory board.<sup>242</sup> Specific criteria are usually predefined for the requests, including purpose, scientific merit of the project, *curriculum vitae* and other researcher's credentials, targeted objectives, IRB approvals, funding, etc. Lack of description of these qualifications may all be reasons for not approving access. Further reasons for nonapproval may be lack of biospecimens, research aims which do not address the biobank purpose, ethical concerns, etc.<sup>239</sup>

The access agreement between the biobank and the researchers should include information about material transfer agreements (MTA), data transfer, costs (administration fees), (prohibition of) transfer of data or biospecimens to third parties, intellectual property rights, responsibility of each party, prepublication approval by the steering committee/advisory board, code of conduct, statistical monitoring, return of results created by the project to the biobank for future research use, etc. (Box 11.17).<sup>242</sup> Approximately 40% of biobanks place no requirements on disposition of the specimens, 33% require the return of remaining biospecimens by the end of the project, and 21% require specimens to be destroyed.<sup>239</sup> The specific access to data is usually permitted through a password-protected user account upon acceptance of terms and conditions. Access may be stratified depending on the researcher's professional background and the anticipated use of the materials.<sup>242</sup>

In order to enhance biospecimen and data sharing, the Biobanking and Biomolecular Resources Research Infrastructure (BBMRI) have produced standardized dataset descriptions consisting of attributes describing biobank content.<sup>245</sup> The aim is to facilitate less expensive and time-consuming sharing through harmonization of basic attributes. For already established biobanks, this can be done at a descriptive level, and for future biobanks, these attributes should be taken into consideration at the time of IRB and legal approval of the study.

In case-control studies or multicenter studies, preanalytical collection, processing, transport, and storage should be similar and if possible centralized with only minimal local processing, in order to avoid uncontrolled and unmeasured variation.<sup>45</sup>

### Culling of Biospecimens

“Culling is the process of reviewing and eliminating selected specimens or an entire collection either by destruction or by transfer to a new *custodian*.<sup>1</sup> Different reasons for culling are listed in Box 11.18, all of which are issues that have been discussed in other sections of this chapter.

### BOX 11.17 Examples of Content in Access Agreements Between Biobanks and Researchers

- Biospecimen
  - Type and volume of biospecimen
  - Criteria for use
  - Criteria for transport/shipment
  - Criteria for storage and handling by the researchers
  - Return of used biospecimen
  - Destruction of used biospecimen
  - Date of return/project end
- Data
  - Transfer
  - Access
    - How
    - Type of data
  - Prohibition of transfer of data to third parties
  - Date of return/project end
- Costs (administration fees)
- Responsibilities of the
  - Biobank
  - Researchers
- Code of conduct
- Prepublication
  - Statistical monitoring and review
  - Approval by the steering committee/advisory board
- Return of results created by the project to the biobank for future research use
- Project description
- Approvals

Courtesy Christina Ellervik.

### BOX 11.18 Reasons for Culling of Collections

- Destruction
- Transfer
- Storage space constraints
- Control costs
- Sale of company
- Consent issues
- Regulatory changes
- Protocol modifications
- Scientific purposes have been met
- Compromised specimen integrity
  - Equipment failure
  - Disasters
  - Freeze-thaw cycles
- All identifying information has been lost
- Lack of use
- Potential biohazards associated with the specimen
- Extra specimens were collected or stored in excess of the investigator's protocol
- When the status of a participant changes from “eligible to ineligible” or their case/control status changes

Courtesy Christina Ellervik.

Closure of biobanks (population or direct-to-consumer-genetic-testing companies) is associated with ethical issues such as informed consent, storage, and privacy.<sup>246</sup> What happens if a biobank closes? Will the biospecimens and or data be destroyed or transferred and what are the criteria for these decisions? Should participants be informed and when? Biobanks should have long-term sustainability plans and policies for the collection's disposition (including destruction and transfer) clearly described and transparent for all stakeholders and IRBs.

The biobank LIMS should be capable of tracking biospecimens which have been destroyed or transferred. Biospecimens are potential biohazards, and disposal of biospecimens should be done according to hazardous material rules and regulations. Transfer of collections may require new IRB approval and MTAs and a change of ownership, custodianship, and governance. These issues should be fully addressed.

## CONCLUSION

Biobank planning is time consuming but essential for the overall organization, including considerations concerning quality management, ELSI issues, recruitment of participants, and the preanalytical, analytical, and postanalytical variables associated with biospecimen integrity. In Points to Remember, a checklist is provided for a comprehensive approach to biobanking, based on the details discussed in this chapter.

Most errors in a clinical laboratory are due to preanalytical errors. Preanalytical variability of biospecimens can have significant effects on downstream analyses and controlling such variables is therefore fundamental for the future use of biospecimens in personalized medicine for diagnostics or prognostic purposes. For the purpose of biomarker discovery and development, preanalytical requirements and documentation are as important as analytical requirements for the evaluation of clinical performance of the biomarker, especially when seeking Food and Drug Administration (FDA) approval. When comparing or pooling results from different studies, investigators should take into account the differences in preanalytical conditions. Currently preanalytical variables are not routinely documented in the biospecimen research literature. Future studies using biobanked biospecimens should describe in detail the preanalytical handling of biospecimens, and analyze and interpret the results with regard to the effects of these variables (further discussion of preanalytical variables can be found in Chapter 5).

Governance, ownership, and custodianship issues must be considered and planned before any collection takes place. Access criteria should be easily available and transparent. A business plan including costs, funding, and sustainability is an important document for the stakeholders and sponsors who need to be assured of the biobank's viability. ELSI issues differ between countries, thus making it difficult and challenging to collaborate across borders. The publication standards which have emerged in recent years are important developments toward higher quality documentation in articles on biomarkers from biobanked biospecimens.

Future issues to be resolved are the development of evidence-based methods mitigating the effects of preanalytical variables, development of a biobank-specific ISO certification standard, and the further development of biobank sustainability plans.

## POINTS TO REMEMBER

### **Biobank Checklist**

#### **Planning Phase**

- Study design
- Biospecimen types and volumes
- Measurements and health examination
- Questionnaire
- Type of data annotation
- Infrastructure
- Laboratory information system
- Quality management documents
- Labeling
- Coding
- Traceability of biospecimen
- Accreditation or certification
- Organization
- Policies on access, material transfer, data release
- Governance, ownership, custodianship
- Business plan including expenses and funding
- Sustainability
- Ethical, legal, and social issues
- Invitation, consent-form
- Return-of-results: which results, how to communicate, critical values
- Staff education
- Facilities (for collection, processing, storage)
- Equipment and utensils
- Security and disaster plan
- Culling of collection

#### **Recruitment Phase**

- Communication strategy
- Scheduling visits

#### **Pilot Phase**

- Run all phases from recruitment to retrieval of specimen

#### **Preanalytical Phase**

- Collection
- Processing
- Immediate analyses (return of results to participants)
- Transport/shipment
- Storage
- Retrieval

#### **Analytical Phase**

- Analysis

#### **Postanalytical Phase**

- Data and biospecimen validation

#### **Collaboration Phase**

- Return of "data results" to the biobank
- Return of results to participants
- Material and data transfer agreements

## SELECTED REFERENCES

1. Campbell LD, Astrin JJ, DeSouza Y, et al. The 2018 revision of the ISBER Best Practices: summary of changes and the Editorial Team's Development Process. *Biopreserv Biobank* 2018;16:3–6.
3. Henderson GE, Cadigan RJ, Edwards TP, et al. Characterizing biobank organizations in the U.S.: results from a national survey. *Genome Med* 2013;5:3.
5. Hewitt R, Watson P. Defining biobank. *Biopreserv Biobank* 2013;11:309–15.
7. Institute NC. NCI Biorepositories and Biospecimen Research Branch (BBRB) Best Practices for biospecimen resources, 2016 Edition. [biospecimen.cancer.gov/bestpractices](http://biospecimen.cancer.gov/bestpractices). NCI 2016.
8. Simeon-Dubach D, Watson P. Biobanking 3.0: evidence based and customer focused biobanking. *Clin Biochem* 2014;47:300–8.
19. Ellervik C, Vaught J. Pre-analytical variables affecting the integrity of human biospecimens in biobanking. *Clin Chem* 2015;61:914–34.
56. Gordy D, Tashjian RS, Lee H, Movassaghi M, Yong WH. Domestic and international shipping of biospecimens. *Methods Mol Biol* 2019;1897:433–43.
57. Institute CaLS. Collection, transport, and processing of blood specimens for coagulation testing and general performance of coagulation assays. Approved guideline H21-A5. CLSI: Wayne, PA 2008.
58. Institute CaLS. Procedures for handling and processing of blood specimens for common laboratory tests. H18-A4; approved guideline - 4th ed. CLSI: Wayne, PA 2010.
113. Paskal W, Paskal AM, Debski T, Gryziak M, Jaworowski J. Aspects of modern biobank activity—comprehensive review. *Pathol Oncol Res* 2018;24:771–85.
169. Kozlakidis Z, Seiler C, Simeon-Dubach D. ISBER Best Practices fourth edition: a success story. *Biopreserv Biobank* 2018.
170. Kozlakidis Z. ISBER President's Message: The intent of the ISBER Best Practices fourth edition. *Biopreserv Biobank* 2018;16:64.
172. Betsou F. The ISBER self-assessment tool indicates main pathways for improvement in biobanks and supports international standardization. *Biopreserv Biobank* 2018;16:7–8.
174. Hartman V, Castillo-Pelayo T, Babinszky S, et al. Is Your biobank up to standards? A review of the National Canadian Tissue Repository Network required operational practice standards and the controlled documents of a certified biobank. *Biopreserv Biobank* 2018;16:36–41.
175. Furuta K, Allocca CM, Schacter B, Bledsoe MJ, Ramirez NC. Standardization and innovation in paving a path to a better future: an update of activities in ISO/TC276/WG2 biobanks and bioresources. *Biopreserv Biobank* 2018;16:23–7.
176. McCall SJ, Branton PA, Blanc VM, et al. The College of American Pathologists Biorepository Accreditation Program: results from the first 5 years. *Biopreserv Biobank* 2018;16:16–22.
207. Henderson MK, Goldring K, Simeon-Dubach D. Advancing professionalization of biobank business operations: performance and utilization. *Biopreserv Biobank* 2019;17:213–8.
208. Henderson MK, Goldring K, Simeon-Dubach D. Advancing professionalization of biobank business operations: a worldwide survey. *Biopreserv Biobank* 2019;17:71–5.
211. Allen MJ, Powers ML, Gronowski KS, Gronowski AM. Human tissue ownership and use in research: what laboratorians and researchers should know. *Clin Chem* 2010;56:1675–82.
243. Conroy M, Sellors J, Effingham M, et al. The advantages of UK Biobank's open-access strategy for health research. *J Intern Med* 2019;286:389–97.
238. McElfish PA, Long CR, James LP, et al. Characterizing health researcher barriers to sharing results with study participants. *J Clin Transl Sci* 2019;3:295–301.
205. Vaught J, Rogers J, Carolin T, Compton C. Biobankonomics: developing a sustainable business model approach for the formation of a human tissue biobank. *J Natl Cancer Inst Monogr* 2011;2011:24–31.

## REFERENCES

1. Campbell LD, Astrin JJ, DeSouza Y, et al. The 2018 Revision of the ISBER Best Practices: summary of changes and the Editorial Team's Development Process. *Biopreserv Biobank* 2018;16:3–6.
2. Eiseman E, Haga SB. Handbook of human tissue sources: a national resource of human tissue samples. 1st ed: Rand Publishing Paperback; 2000.
3. Henderson GE, Cadigan RJ, Edwards TP, et al. Characterizing biobank organizations in the U.S.: results from a national survey. *Genome Med* 2013;5:3.
4. OECD. Guidelines for Human Biobanks and Genetic Research Databases (HBGRDs). [www.oecd.org](http://www.oecd.org). OECD 2009.
5. Hewitt R, Watson P. Defining biobank. *Biopreserv Biobank* 2013;11:309–15.
6. Shaw DM, Elger BS, Colledge F. What is a biobank? Differing definitions among biobank stakeholders. *Clin Genet* 2014; 85:223–7.
7. Institute NC. NCI Biorepositories and Biospecimen Research Branch (BBRB) Best Practices for biospecimen resources, 2016 Edition. [biospecimen.cancer.gov/bestpractices](http://biospecimen.cancer.gov/bestpractices). NCI 2016.
8. Simeon-Dubach D, Watson P. Biobanking 3.0: evidence based and customer focused biobanking. *Clin Biochem* 2014;47: 300–8.
9. Vaught J. Biobanking comes of age: the transition to biospecimen science. *Annu Rev Pharmacol Toxicol* 2016; 56:211–28.
10. Jaffe S. Planning for US Precision Medicine Initiative underway. *Lancet* 2015;385:2448–9.
11. Sankar PL, Parker LS. The Precision Medicine Initiative's all of us Research Program: an agenda for research on its ethical, legal, and social issues. *Genet Med* 2017;19:743–50.
12. Tozzo P, Pegoraro R, Caenazzo L. Biobanks for non-clinical purposes and the new law on forensic biobanks: does the Italian context protect the rights of minors? *J Med Ethics* 2010; 36:775–8.
13. Matzke EA, O'Donoghue S, Barnes RO, et al. Certification for biobanks: the program developed by the Canadian Tumour Repository Network (CTRNet). *Biopreserv Biobank* 2012; 10:426–32.
14. Watson PH. Biobank classification: communicating biorepository diversity. *Biopreserv Biobank* 2014;12:163–4.
15. Zika E, Paci D, Braun A, et al. A European survey on biobanks: trends and issues. *Public Health Genomics* 2011;14:96–103.
16. Shickle D, Griffin M, El-Arifi K. Inter- and intra-biobank networks: classification of biobanks. *Pathobiology* 2010;77: 181–90.
17. Devereux L, Watson PH, Mes-Masson AM, et al. A review of international biobanks and networks: success factors and key benchmarks—a 10-year retrospective review. *Biopreserv Biobank* 2019;17:512–9.
18. Common minimum technical standards and protocols for biological resource centres dedicated to cancer research. Switzerland 2007.
19. Ellervik C, Vaught J. Pre-analytical variables affecting the integrity of human biospecimens in biobanking. *Clin Chem* 2015;61:914–34.
20. Kellogg MD, Ellervik C, Morrow D, Hsing A, Stein E, Sethi AA. Preanalytical considerations in the design of clinical trials and epidemiological studies. *Clin Chem* 2015;61:797–803.
21. Carraro P, Zago T, Plebani M. Exploring the initial steps of the testing process: frequency and nature of pre-preanalytic errors. *Clin Chem* 2012;58:638–42.
22. Lippi G, Guidi GC, Mattuzzi C, Plebani M. Preanalytical variability: the dark side of the moon in laboratory testing. *Clin Chem Lab Med* 2006;44:358–65.
23. Makary MA, Daniel M. Medical error—the third leading cause of death in the US. *BMJ* 2016;353:i2139.
24. Szecsi PB, Odum L. Error tracking in a clinical biochemistry laboratory. *Clin Chem Lab Med* 2009;47:1253–7.
25. Meir K, Gaffney EF, Simeon-Dubach D, et al. The human face of biobank networks for translational research. *Biopreserv Biobank* 2011;9:279–85.
26. Elliott P, Peakman TC. The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. *Int J Epidemiol* 2008;37:234–44.
27. Holland NT, Smith MT, Eskenazi B, Bastaki M. Biological sample collection and processing for molecular epidemiological studies. *Mutat Res* 2003;543:217–34.
28. Hansen TV, Simonsen MK, Nielsen FC, Hundrup YA. Collection of blood, saliva, and buccal cell samples in a pilot study on the Danish nurse cohort: comparison of the response rate and quality of genomic DNA. *Cancer Epidemiol Biomarkers Prev* 2007;16:2072–6.
29. Bhatti P, Kampa D, Alexander BH, et al. Blood spots as an alternative to whole blood collection and the effect of a small monetary incentive to increase participation in genetic association studies. *BMC Med Res Methodol* 2009;9:76.
30. Statland BE, Winkel P. Response of clinical chemistry quantity values to selected physical, dietary, and smoking activities. *Prog Clin Pathol* 1981;8:25–44.
31. Tolonen H, Ferrario M, Kuulasmaa K. Standardization of total cholesterol measurement in population surveys—pre-analytic sources of variation and their effect on the prevalence of hypercholesterolaemia. *Eur J Cardiovasc Prev Rehabil* 2005;12:257–67.
32. Sennels HP, Jorgensen HL, Goetze JP, Fahrenkrug J. Rhythmic 24-hour variations of frequently used clinical biochemical parameters in healthy young males—the Bispebjerg study of diurnal variations. *Scand J Clin Lab Invest* 2012;72:287–95.
33. Adil MM, Alam AY. Temperature regulation and standardization practices of clinical laboratories in Karachi. *J Pak Med Assoc* 2005;55:88–90.
34. Jayasuriya NA, Kjaergaard AD, Pedersen KM, et al. Smoking, blood cells and myeloproliferative neoplasms: meta-analysis and Mendelian randomization of 2.3 million people. *Br J Haematol* 2019;189:323–34.
35. Pedersen KM, Colak Y, Ellervik C, Hasselbalch HC, Bojesen SE, Nordestgaard BG. Smoking and increased white and red blood cells. *Arterioscler Thromb Vasc Biol* 2019;39:965–77.
36. Simundic AM, Cornes M, Grankvist K, Lippi G, Nybo M. Standardization of collection requirements for fasting samples: for the Working Group on Preanalytical Phase (WG-PA) of the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM). *Clin Chim Acta* 2014; 432:33–7.
37. Wagner M, Tonoli D, Varesio E, Hopfgartner G. The use of mass spectrometry to analyze dried blood spots. *Mass Spectrom Rev* 2014.
38. Edvardsen K, Engelsen O, Brustad M. Duration of vitamin D synthesis from weather model data for use in prospective epidemiological studies. *Int J Biometeorol* 2009;53:451–9.

39. McQuillan AC, Sales SD. Designing an automated blood fractionation system. *Int J Epidemiol* 2008;37 Suppl 1:i51–i5.
40. Malm J, Vegvari A, Rezeli M, et al. Large scale biobanking of blood - the importance of high density sample processing procedures. *J Proteomics* 2012;76 Spec No.:116–24.
41. Malm J, Fehniger TE, Danmyr P, et al. Developments in biobanking workflow standardization providing sample integrity and stability. *J Proteomics* 2013;95:38–45.
42. Hubel A. Preservation of cells: a practical manual. John Wiley & Sons, Inc.; 2018.
43. Hubel A, Spindler R, Skubitz AP. Storage of human biospecimens: selection of the optimal storage temperature. *Biopreserv Biobank* 2014;12:165–75.
44. Baboo J, Kilbride P, Delahaye M, et al. The impact of varying cooling and thawing rates on the quality of cryopreserved human peripheral blood T cells. *Sci Rep* 2019;9:3417.
45. Vaught J, Abayomi A, Peakman T, Watson P, Matzke L, Moore H. Critical issues in international biobanking. *Clin Chem* 2014;60:1368–74.
46. Mercuri A, Turchi S, Borghini A, et al. Nitrogen biobank for cardiovascular research. *Curr Cardiol Rev* 2013;9:253–9.
47. Colotte M, Coudy D, Tuffet S, Bonnet J. Adverse effect of air exposure on the stability of DNA stored at room temperature. *Biopreserv Biobank* 2011;9:47–50.
48. Ivanova NV, Kuzmina ML. Protocols for dry DNA storage and shipment at room temperature. *Mol Ecol Resour* 2013;13:890–8.
49. Howlett SE, Castillo HS, Gioeni LJ, Robertson JM, Donfack J. Evaluation of DNAsable for DNA storage at ambient temperature. *Forensic Sci Int Genet* 2014;8:170–8.
50. Udtha M, Flores R, Sanner J, et al. The protection and stabilization of whole blood at room temperature. *Biopreserv Biobank* 2014;12:332–6.
51. Seelenfreund E, Robinson WA, Amato CM, Tan AC, Kim J, Robinson SE. Long term storage of dry versus frozen RNA for next generation molecular studies. *PLoS One* 2014;9: e111827.
52. Mathay C, Yan W, Chuaqui R, et al. Short-term stability study of RNA at room temperature. *Biopreserv Biobank* 2012;10:532–42.
53. Hernandez GE, Mondala TS, Head SR. Assessing a novel room-temperature RNA storage medium for compatibility in microarray gene expression analysis. *Biotechniques* 2009;47:667–8, 70.
54. Saragusty J, Anzalone DA, Palazzese L, et al. Dry biobanking as a conservation tool in the Anthropocene. *Theriogenology* 2020;150:130–8.
55. Comley J. Automated biobanking - the next big step for biorepositories. *Drug Discovery World* 2007;Summer:46–70.
56. Gordy D, Tashjian RS, Lee H, Movassaghi M, Yong WH. Domestic and international shipping of biospecimens. *Methods Mol Biol* 2019;1897:433–43.
57. Institute CaLS. Collection, transport, and processing of blood specimens for coagulation testing and general performance of coagulation assays. Approved guideline H21-A5. CLSI: Wayne, PA 2008.
58. Institute CaLS. Procedures for handling and processing of blood specimens for common laboratory tests. H18-A4; approved guideline - 4th ed. CLSI: Wayne, PA 2010.
59. van Amsterdam P, Waldrop C. The application of dried blood spot sampling in global clinical trials. *Bioanalysis* 2010;2: 1783–6.
60. Mei JV, Alexander JR, Adam BW, Hannon WH. Use of filter paper for the collection and analysis of human whole blood specimens. *J Nutr* 2001;131:1631S–6S.
61. Olson WC, Smolkin ME, Farris EM, et al. Shipping blood to a central laboratory in multicenter clinical trials: effect of ambient temperature on specimen temperature, and effects of temperature on mononuclear cell yield, viability and immunologic function. *J Transl Med* 2011;9:26.
62. Duale N, Brunborg G, Ronningen KS, et al. Human blood RNA stabilization in samples collected and transported for a large biobank. *BMC Res Notes* 2012;5:510.
63. Lehmann S, Guadagni F, Moore H, et al. Standard preanalytical coding for biospecimens: review and implementation of the Sample PREanalytical Code (SPREC). *Biopreserv Biobank* 2012;10:366–74.
64. McDonald SA, Ryan BJ, Brink A, Holtschlag VL. Automated web-based request mechanism for workflow enhancement in an academic customer-focused biorepository. *Biopreserv Biobank* 2012;10:48–54.
65. Brisson AR, Matsui D, Rieder MJ, Fraser DD. Translational research in pediatrics: tissue sampling and biobanking. *Pediatrics* 2012;129:153–62.
66. Shabikhani M, Lucey GM, Wei B, et al. The procurement, storage, and quality assurance of frozen blood and tissue biospecimens in pathology, biorepository, and biobank settings. *Clin Biochem* 2014;47:258–66.
67. McNair P, Nielsen SL, Christiansen C, Axelsson C. Gross errors made by routine blood sampling from two sites using a tourniquet applied at different positions. *Clin Chim Acta* 1979;98:113–8.
68. Salvagno G, Lima-Oliveira G, Brocco G, Danese E, Guidi GC, Lippi G. The order of draw: myth or science? *Clin Chem Lab Med* 2013;51:2281–5.
69. Sulaiman RA, Cornes MP, Whitehead SJ, Othonos N, Ford C, Gama R. Effect of order of draw of blood samples during phlebotomy on routine biochemistry results. *J Clin Pathol* 2011;64:1019–20.
70. Bowen RA, Remaley AT. Interferences from blood collection tube components on clinical chemistry assays. *Biochem Med (Zagreb)* 2014;24:31–44.
71. Lippi G, Salvagno GL, Lima-Oliveira G, Brocco G, Danese E, Guidi GC. Postural change during venous blood collection is a major source of bias in clinical chemistry testing. *Clin Chim Acta* 2014;440C:164–8.
72. Lippi G, Salvagno GL, Montagnana M, Brocco G, Cesare GG. Influence of the needle bore size used for collecting venous blood samples on routine clinical chemistry testing. *Clin Chem Lab Med* 2006;44:1009–14.
73. Lippi G, Salvagno GL, Montagnana M, Lima-Oliveira G, Guidi GC, Favaloro EJ. Quality standards for sample collection in coagulation testing. *Semin Thromb Hemost* 2012;38: 565–75.
74. Lima-Oliveira G, Lippi G, Salvagno GL, et al. Processing of diagnostic blood specimens: is it really necessary to mix primary blood tubes after collection with evacuated tube system? *Biopreserv Biobank* 2014;12:53–9.
75. Lippi G, Plebani M, Favaloro EJ. Interference in coagulation testing: focus on spurious hemolysis, icterus, and lipemia. *Semin Thromb Hemost* 2013;39:258–66.
76. Simundic AM, Baird G, Cadamuro J, Costelloe SJ, Lippi G. Managing hemolyzed samples in clinical laboratories. *Crit Rev Clin Lab Sci* 2020;57:1–21.

77. Lippi G, Sanchis-Gomar F. Epidemiological, biological and clinical update on exercise-induced hemolysis. *Ann Transl Med* 2019;7:270.
78. Narayanan S. The preanalytic phase. An important component of laboratory medicine. *Am J Clin Pathol* 2000;113:429–52.
79. Sacks DB, Arnold M, Bakris GL, et al. Guidelines and recommendations for laboratory analysis in the diagnosis and management of diabetes mellitus. *Clin Chem* 2011;57:e1–e47.
80. House RV. Cytokine measurement techniques for assessing hypersensitivity. *Toxicology* 2001;158:51–8.
81. Lippi G, Franchini M, Montagnana M, Salvagno GL, Poli G, Guidi GC. Quality and reliability of routine coagulation testing: can we trust that sample? *Blood Coagul Fibrinolysis* 2006;17:513–9.
82. Hollegaard MV, Grove J, Grauholm J, et al. Robustness of genome-wide scanning using archived dried blood spot samples as a DNA source. *BMC Genet* 2011;12:58.
83. Boyanton BL, Jr., Blick KE. Stability studies of twenty-four analytes in human plasma and serum. *Clin Chem* 2002;48:2242–7.
84. Oddoze C, Lombard E, Portugal H. Stability study of 81 analytes in human whole blood, in serum and in plasma. *Clin Biochem* 2012;45:464–9.
85. Jackson C, Best N, Elliott P. UK Biobank Pilot Study: stability of haematological and clinical chemistry analytes. *Int J Epidemiol* 2008;37 Suppl 1:i16–i22.
86. Tanner M, Kent N, Smith B, Fletcher S, Lewer M. Stability of common biochemical analytes in serum gel tubes subjected to various storage temperatures and times pre-centrifugation. *Ann Clin Biochem* 2008;45:375–9.
87. Holland NT, Pfleger L, Berger E, Ho A, Bastaki M. Molecular epidemiology biomarkers—sample collection and processing considerations. *Toxicol Appl Pharmacol* 2005;206:261–8.
88. Gaye A, Peakman T, Tobin MD, Burton PR. Understanding the impact of pre-analytic variation in haematological and clinical chemistry analytes on the power of association studies. *Int J Epidemiol* 2014;43:1633–44.
89. Hebels DG, Georgiadis P, Keun HC, et al. Performance in omics analyses of blood samples in long-term storage: opportunities for the exploitation of existing biobanks in environmental health research. *Environ Health Perspect* 2013;121:480–7.
90. Brinc D, Chan MK, Venner AA, et al. Long-term stability of biochemical markers in pediatric serum specimens stored at -80 degrees C: a CALIPER Substudy. *Clin Biochem* 2012;45:816–26.
91. Gislefoss RE, Grimsrud TK, Morkrid L. Stability of selected serum proteins after long-term storage in the Janus Serum Bank. *Clin Chem Lab Med* 2009;47:596–603.
92. Hannisdal R, Gislefoss RE, Grimsrud TK, Hustad S, Morkrid L, Ueland PM. Analytical recovery of folate and its degradation products in human serum stored at -25 degrees C for up to 29 years. *J Nutr* 2010;140:522–6.
93. Hostmark AT, Glattre E, Jellum E. Effect of long-term storage on the concentration of albumin and free fatty acids in human sera. *Scand J Clin Lab Invest* 2001;61:443–7.
94. Reed AB, Ankerst DP, Leach RJ, Vipraio G, Thompson IM, Parekh DJ. Total prostate specific antigen stability confirmed after long-term storage of serum at -80C. *J Urol* 2008;180:534–7.
95. Holl K, Lundin E, Kaasila M, et al. Effect of long-term storage on hormone measurements in samples from pregnant women: the experience of the Finnish Maternity Cohort. *Acta Oncol* 2008;47:406–12.
96. Hernestal-Boman J, Jansson JH, Nilsson TK, Eliasson M, Johansson L. Long-term stability of fibrinolytic factors stored at -80 degrees C. *Thromb Res* 2010;125:451–6.
97. Rolandsson O, Marklund SL, Norberg M, Agren A, Hagg E. Hemoglobin A1c can be analyzed in blood kept frozen at -80 degrees C and is not commonly affected by hemolysis in the general population. *Metabolism* 2004;53:1496–9.
98. Prentice P, Turner C, Wong MC, Dalton RN. Stability of metabolites in dried blood spots stored at different temperatures over a 2-year period. *Bioanalysis* 2013;5:1507–14.
99. Yin P, Peter A, Franken H, et al. Preanalytical aspects and sample quality assessment in metabolomics studies of human blood. *Clin Chem* 2013;59:833–45.
100. Nederhand RJ, Droog S, Kluft C, Simoons ML, de Maat MP. Logistics and quality control for DNA sampling in large multicenter studies. *J Thromb Haemost* 2003;1:987–91.
101. Mychaleckyj JC, Farber EA, Chmielewski J, et al. Buffy coat specimens remain viable as a DNA source for highly multiplexed genome-wide genetic tests after long term storage. *J Transl Med* 2011;9:91.
102. Becker N, Lockwood CM. Pre-analytical variables in miRNA analysis. *Clin Biochem* 2013;46:861–8.
103. Institute CaLS. Urinanalysis; approved guideline GP16-A3; third edition. CLSI: Wayne, PA 2009.
104. Delanghe J, Speeckaert M. Preanalytical requirements of urinalysis. *Biochem Med (Zagreb)* 2014;24:89–104.
105. Papale M, Pedicillo MC, Thatcher BJ, et al. Urine profiling by SELDI-TOF/MS: monitoring of the critical steps in sample collection, handling and analysis. *J Chromatogr B Analyt Technol Biomed Life Sci* 2007;856:205–13.
106. Henriksen T, Hillestrom PR, Poulsen HE, Weimann A. Automated method for the direct analysis of 8-oxo-guanosine and 8-oxo-2'-deoxyguanosine in human urine using ultraperformance liquid chromatography and tandem mass spectrometry. *Free Radic Biol Med* 2009;47:629–35.
107. Bali LE, Diman A, Bernard A, Roosens NH, De Keersmaecker SC. Comparative study of seven commercial kits for human DNA extraction from urine samples suitable for DNA biomarker-based public health studies. *J Biomol Tech* 2014.
108. Bernini P, Bertini I, Luchinat C, Nincheri P, Staderini S, Turano P. Standard operating procedures for pre-analytical handling of blood and urine for metabolomic studies and biobanks. *J Biomol NMR* 2011;49:231–43.
109. Veljkovic K, Rodriguez-Capote K, Bhayana V, et al. Assessment of a four hour delay for urine samples stored without preservatives at room temperature for urinalysis. *Clin Biochem* 2012;45:856–8.
110. Bingham SA, Cummings JH. Creatinine and PABA as markers for completeness of collection of 24-hour urine samples. *Hum Nutr Clin Nutr* 1986;40:473–6.
111. Miller RC, Brindle E, Holman DJ, et al. Comparison of specific gravity and creatinine for normalizing urinary reproductive hormone concentrations. *Clin Chem* 2004;50:924–32.
112. Schaub S, Wilkins J, Weiler T, Sangster K, Rush D, Nickerson P. Urine protein profiling with surface-enhanced laser-desorption/ionization time-of-flight mass spectrometry. *Kidney Int* 2004;65:323–32.
113. Paskal W, Paskal AM, Debski T, Gryziak M, Jaworowski J. Aspects of modern biobank activity - comprehensive review. *Pathol Oncol Res* 2018;24:771–85.
114. Remer T, Montenegro-Bethancourt G, Shi L. Long-term urine biobanking: storage stability of clinical chemical parameters

- under moderate freezing conditions without use of preservatives. *Clin Biochem* 2014;47:307–11.
115. Yoshizawa JM, Schafer CA, Schafer JJ, Farrell JJ, Paster BJ, Wong DT. Salivary biomarkers: toward future clinical and diagnostic utilities. *Clin Microbiol Rev* 2013;26:781–91.
  116. Pfaffe T, Cooper-White J, Beyerlein P, Kostner K, Punyadeera C. Diagnostic potential of saliva: current state and future applications. *Clin Chem* 2011;57:675–87.
  117. Liesveld JL, Rothberg PG. Mixed chimerism in SCT: conflict or peaceful coexistence? *Bone Marrow Transplant* 2008;42:297–310.
  118. Yunis EJ, Zuniga J, Romero V, Yunis EJ. Chimerism and tetragametic chimerism in humans: implications in autoimmunity, allorecognition and tolerance. *Immunol Res* 2007;38:213–36.
  119. Rasi S, Bruscaggin A, Rinaldi A, et al. Saliva is a reliable and practical source of germline DNA for genome-wide studies in chronic lymphocytic leukemia. *Leuk Res* 2011;35:1419–22.
  120. Navazesh M, Christensen CM. A comparison of whole mouth resting and stimulated salivary measurement procedures. *J Dent Res* 1982;61:1158–62.
  121. Golatowski C, Salazar MG, Dhople VM, et al. Comparative evaluation of saliva collection methods for proteome analysis. *Clin Chim Acta* 2013;419:42–6.
  122. Topkas E, Keith P, Dimeski G, Cooper-White J, Punyadeera C. Evaluation of saliva collection devices for the analysis of proteins. *Clin Chim Acta* 2012;413:1066–70.
  123. Amado FM, Ferreira RP, Vitorino R. One decade of salivary proteomics: current approaches and outstanding challenges. *Clin Biochem* 2013;46:506–17.
  124. Schipper RG, Silletti E, Vingerhoeds MH. Saliva as research material: biochemical, physicochemical and practical aspects. *Arch Oral Biol* 2007;52:1114–35.
  125. Schipper R, Loof A, de GJ, Harthoorn L, Dransfield E, van HW. SELDI-TOF-MS of saliva: methodology and pre-treatment effects. *J Chromatogr B Analyt Technol Biomed Life Sci* 2007;847:45–53.
  126. Nunes AP, Oliveira IO, Santos BR, et al. Quality of DNA extracted from saliva samples collected with the Oragene DNA self-collection kit. *BMC Med Res Methodol* 2012;12:65.
  127. Garcia-Closas M, Egan KM, Abruzzo J, et al. Collection of genomic DNA from adults in epidemiological studies by buccal cytobrush and mouthwash. *Cancer Epidemiol Biomarkers Prev* 2001;10:687–96.
  128. Heath EM, Morken NW, Campbell KA, Tkach D, Boyd EA, Strom DA. Use of buccal cells collected in mouthwash as a source of DNA for clinical testing. *Arch Pathol Lab Med* 2001;125:127–33.
  129. Mulot C, Stucker I, Clavel J, Beaune P, Loriot MA. Collection of human genomic DNA from buccal cells for genetics studies: comparison between cytobrush, mouthwash, and treated card. *J Biomed Biotechnol* 2005;2005:291–6.
  130. King IB, Satia-Abouta J, Thornquist MD, et al. Buccal cell DNA yield, quality, and collection costs: comparison of methods for large-scale studies. *Cancer Epidemiol Biomarkers Prev* 2002;11:1130–3.
  131. Rogers NL, Cole SA, Lan HC, Crossa A, Demerath EW. New saliva DNA collection method compared to buccal cell collection techniques for epidemiological studies. *Am J Hum Biol* 2007;19:319–26.
  132. Feigelson HS, Rodriguez C, Robertson AS, et al. Determinants of DNA yield and quality from buccal cell samples collected with mouthwash. *Cancer Epidemiol Biomarkers Prev* 2001;10:1005–8.
  133. Nemoda Z, Horvat-Gordon M, Fortunato CK, Beltzer EK, Scholl JL, Granger DA. Assessing genetic polymorphisms using DNA extracted from cells present in saliva samples. *BMC Med Res Methodol* 2011;11:170.
  134. Rylander-Rudqvist T, Hakansson N, Tybring G, Wolk A. Quality and quantity of saliva DNA obtained from the self-administrated oragene method—a pilot study on the cohort of Swedish men. *Cancer Epidemiol Biomarkers Prev* 2006;15:1742–5.
  135. Pramanik R, Thompson H, Kistler JO, et al. Effects of the UK Biobank collection protocol on potential biomarkers in saliva. *Int J Epidemiol* 2012;41:1786–97.
  136. Caboux E, Lallemand C, Ferro G, et al. Sources of pre-analytical variations in yield of DNA extracted from blood samples: analysis of 50,000 DNA samples in EPIC. *PLoS One* 2010;7:e39821.
  137. Kim SJ, Dix DJ, Thompson KE, et al. Effects of storage, RNA extraction, genechip type, and donor sex on gene expression profiling of human whole blood. *Clin Chem* 2007;53:1038–45.
  138. de Lomas JG, Sunzeri FJ, Busch MP. False-negative results by polymerase chain reaction due to contamination by glove powder. *Transfusion* 1992;32:83–5.
  139. Chan KC, Jiang P, Zheng YW, et al. Cancer genome scanning in plasma: detection of tumor-associated copy number aberrations, single-nucleotide variants, and tumoral heterogeneity by massively parallel sequencing. *Clin Chem* 2013;59:211–24.
  140. Chiu RW, Chan KC, Gao Y, et al. Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma. *Proc Natl Acad Sci U S A* 2008;105:20458–63.
  141. Jiang P, Lo YMD. The long and short of circulating cell-free DNA and the ins and outs of molecular diagnostics. *Trends Genet* 2016;32:360–71.
  142. Lee TH, Montalvo L, Chrebtow V, Busch MP. Quantitation of genomic DNA in plasma and serum samples: higher concentrations of genomic DNA found in serum than in plasma. *Transfusion* 2001;41:276–82.
  143. Messaoudi SE, Rolet F, Mouliere F, Thierry AR. Circulating cell free DNA: preanalytical considerations. *Clin Chim Acta* 2013;424:222–30.
  144. Duale N, Lipkin WI, Briese T, et al. Long-term storage of blood RNA collected in RNA stabilizing Tempus tubes in a large biobank—evaluation of RNA quality and stability. *BMC Res Notes* 2014;7:633.
  145. Hebel D, van Herwijnen MH, Brauers KJ, et al. Elimination of heparin interference during microarray processing of fresh and biobank-archived blood samples. *Environ Mol Mutagen* 2014;55:482–91.
  146. Palmirotta R, De Marchis ML, Ludovici G, et al. Impact of preanalytical handling and timing for peripheral blood mononuclear cells isolation and RNA studies: the experience of the Interinstitutional Multidisciplinary BioBank (BioBIM). *Int J Biol Markers* 2012;27:e90–e8.
  147. Pazzaglia M, Malentacchi F, Simi L, et al. SPIDIA-RNA: first external quality assessment for the pre-analytical phase of blood samples used for RNA based analyses. *Methods* 2013;59:20–31.
  148. Exiqon. Profiling of microRNA in serum/plasma and other biofluids. [www.exiqon.com](http://www.exiqon.com). Accessed September 2020.

149. Kim DJ, Linnstaedt S, Palma J, et al. Plasma components affect accuracy of circulating cancer-related microRNA quantitation. *J Mol Diagn* 2012;14:71–80.
150. McDonald JS, Milosevic D, Reddi HV, Grebe SK, Algeciras-Schimmin A. Analysis of circulating microRNA: preanalytical and analytical challenges. *Clin Chem* 2011;57:833–40.
151. Weber JA, Baxter DH, Zhang S, et al. The microRNA spectrum in 12 body fluids. *Clin Chem* 2010;56:1733–41.
152. Gail MH, Sheehy T, Cosentino M, et al. Maximizing DNA yield for epidemiologic studies: no more buffy coats? *Am J Epidemiol* 2013;178:1170–6.
153. Shao W, Khin S, Kopp WC. Characterization of effect of repeated freeze and thaw cycles on stability of genomic DNA using pulsed field gel electrophoresis. *Biopreserv Biobank* 2012;10:4–11.
154. Olivieri EH, Franco LA, Pereira RG, Mota LD, Campos AH, Carraro DM. Biobanking practice: RNA storage at low concentration affects integrity. *Biopreserv Biobank* 2014;12: 46–52.
155. Zhang H, Korenkova V, Sjoberg R, et al. Biomarkers for monitoring pre-analytical quality variation of mRNA in blood samples. *PLoS One* 2014;9:e111644.
156. Howie SRC. Blood sample volumens in child health research: review of safe limits. *Bulletin of the World Health Organization* 2011;89:46–93.
157. Buckbee KM. Implementing a pediatric phlebotomy protocol. *Medical Laboratory Observer* 1994;4:32–5.
158. Willock J, Richardson J, Brazier A, Powell C, Mitchell E. Peripheral venepuncture in infants and children. *Nurs Stand* 2004;18:43–50.
159. Caws L, Pfund R. Venepuncture and cannulation on infants and children. *J Child Health Care* 1999;3:11–6.
160. Broder-Fingert S, Crowley WF, Jr., Boepple PA. Safety of frequent venous blood sampling in a pediatric research population. *J Pediatr* 2009;154:578–81.
161. Baird PM, Benes FM, Chan CH, et al. How is your biobank handling disaster recovery efforts? *Biopreserv Biobank* 2013;11:194–201.
162. Bjugn R, Hansen J. Learning by Erring: fire! *Biopreserv Biobank* 2013;11:202–5.
163. Henderson MK, Simeon-Dubach D, Zaayenga A. When bad things happen: lessons learned from effective and not so effective disaster and recovery planning for biobanks. *Biopreserv Biobank* 2013;11:193.
164. Mintzer JL, Kronenthal CJ, Kelly V, et al. Preparedness for a natural disaster: how Coriell planned for hurricane Sandy. *Biopreserv Biobank* 2013;11:216–20.
165. Roswall N, Halkjaer J, Overvad K, Tjonneland A. Measures taken to restore the Danish Diet, Cancer and Health Biobank after flooding: a framework for future biobank restorations. *Biopreserv Biobank* 2013;11:206–10.
166. Simeon-Dubach D, Zaayenga A, Henderson MK. Disaster and recovery: the importance of risk assessment and contingency planning for biobanks. *Biopreserv Biobank* 2013;11:133–4.
167. Morrin HR, Robinson BA. Sustaining a biobank through a series of earthquake swarms: lessons learned from our New Zealand experience. *Biopreserv Biobank* 2013;11:211–5.
168. Vaught J, Lockhart NC. The evolution of biobanking best practices. *Clin Chim Acta* 2012;413:1569–75.
169. Kozlakidis Z, Seiler C, Simeon-Dubach D. ISBER Best Practices fourth edition: a success story. *Biopreserv Biobank* 2018.
170. Kozlakidis Z. ISBER President's Message: The Intent of the ISBER Best Practices fourth edition. *Biopreserv Biobank* 2018;16:64.
171. Barnes R, Albert M, Damaraju S, et al. Generating a comprehensive set of standard operating procedures for a biorepository network-The CTRNet experience. *Biopreserv Biobank* 2013;11:387–96.
172. Betsou F. The ISBER self-assessment tool indicates main pathways for improvement in biobanks and supports international standardization. *Biopreserv Biobank* 2018;16:7–8.
173. Gaignaux A, Ashton G, Coppola D, et al. A Biospecimen proficiency testing program for biobank accreditation: four years of experience. *Biopreserv Biobank* 2016;14:429–39.
174. Hartman V, Castillo-Pelayo T, Babinszky S, et al. Is your biobank up to standards? a review of the National Canadian Tissue Repository Network required operational practice standards and the controlled documents of a certified biobank. *biopreserv biobank* 2018;16:36–41.
175. Furuta K, Allocca CM, Schacter B, Bledsoe MJ, Ramirez NC. Standardization and innovation in paving a path to a better future: an update of activities in ISO/TC276/WG2 biobanks and bioresources. *Biopreserv Biobank* 2018;16:23–7.
176. McCall SJ, Branton PA, Blanc VM, et al. The College of American Pathologists Biorepository Accreditation Program: results from the first 5 years. *Biopreserv Biobank* 2018;16: 16–22.
177. Barnes RO, Shea KE, Watson PH. The Canadian Tissue Repository Network Biobank Certification and the College of American Pathologists Biorepository Accreditation Programs: two strategies for knowledge dissemination in biobanking. *Biopreserv Biobank* 2017;15:9–16.
178. Tarling T, O'Donoghue S, Barnes R, et al. Comparison and analysis of two internationally recognized biobanking standards. *Biopreserv Biobank* 2020.
179. Cho KH, Kim JS, Jeon MS, Lee K, Chung MK, Song CW. Basic principles of the validation for good laboratory practice institutes. *Toxicol Res* 2009;25:1–8.
180. Allison G, Cain YT, Cooney C, et al. Regulatory and quality considerations for continuous manufacturing. May 20–21, 2014 Continuous Manufacturing Symposium. *J Pharm Sci* 2015;104:803–12.
181. Castellanos-Uribe M, Gormally E, Zhou H, Matzke E, P HW. Biobanking education. *Biopreserv Biobank* 2020;18:1–3.
182. Schacter B, Sieffert N, Hill K, Tanabe P, Simeon-Dubach D. A new qualification for the new year: ISBER and American Society of Clinical Pathology Board of certification announce new qualification in biorepository science examination for biobank technicians. *Biopreserv Biobank* 2020;18:43–4.
183. Kay AB, Estrada DK, Mareninov S, et al. Considerations for uniform and accurate biospecimen labelling in a biorepository and research environment. *J Clin Pathol* 2011;64:634–6.
184. Betsou F, Bilbao R, Case J, et al. Standard PREanalytical Code version 3.0. *Biopreserv Biobank* 2018.
185. Karcher DS, Lehman CM. Clinical consequences of specimen rejection: a College of American Pathologists Q-Probes analysis of 78 clinical laboratories. *Arch Pathol Lab Med* 2014;138:1003–8.
186. Nguyen-Nielsen M, Svensson E, Vogel I, Ehrenstein V, Sunde L. Existing data sources for clinical epidemiology: Danish registries for studies of medical genetic diseases. *Clin Epidemiol* 2013;5:249–62.

187. Poste G. Biospecimens, biomarkers, and burgeoning data: the imperative for more rigorous research standards. *Trends Mol Med* 2012;18:717–22.
188. Cheah S, Dee S, Cole A, Matzke L, O'Donoghue S, Watson PH. An online tool for improving biospecimen data element reporting. *Biopreserv Biobank* 2012;10:501–10.
189. Rifai N, Annesley TM, Berg JP, et al. An appeal to medical journal editors: the need for a full description of laboratory methods and specimen handling in clinical study reports. *Clin Chim Acta* 2012;413:653–5.
190. Moore HM, Kelly A, McShane LM, Vaught J. Biospecimen reporting for improved study quality (BRISQ). *Transfusion* 2013;53:e1.
191. Robb JA, Gulley ML, Fitzgibbons PL, et al. A call to standardize preanalytic data elements for biospecimens. *Arch Pathol Lab Med* 2014;138:526–37.
192. Robb JA, Bry L, Sluss PM, Wagar EA, Kennedy MF. A call to standardize preanalytic data elements for biospecimens, part II. *Arch Pathol Lab Med* 2015.
193. Sauerbrei W, Taube SE, McShane LM, Cavenagh MM, Altman DG. Reporting Recommendations for Tumor Marker Prognostic Studies (REMARK): an abridged explanation and elaboration. *J Natl Cancer Inst* 2018;110:803–11.
194. Meredith AJ, Simeon-Dubach D, Matzke LA, Cheah S, Watson PH. Biospecimen data reporting in the research literature. *Biopreserv Biobank* 2019;17:326–33.
195. Albert M, Bartlett J, Johnston RN, Schacter B, Watson P. Biobank bootstrapping: is biobank sustainability possible through cost recovery? *Biopreserv Biobank* 2014;12:374–80.
196. Clement B, Yuille M, Zaltoukal K, et al. Public biobanks: calculation and recovery of costs. *Sci Transl Med* 2014;6:261fs45.
197. Barnes RO, Schacter B, Kodeeswaran S, Watson PH. Funding sources for Canadian biorepositories: the role of user fees and strategies to help fill the gap. *Biopreserv Biobank* 2014;12:300–5.
198. Warth R, Perren A. Construction of a business model to assure financial sustainability of biobanks. *Biopreserv Biobank* 2014;12:389–94.
199. Watson PH, Nussbeck SY, Carter C, et al. A framework for biobank sustainability. *Biopreserv Biobank* 2014;12:60–8.
200. Henderson MK, Goldring K, Simeon-Dubach D. Achieving and maintaining sustainability in biobanking through business planning, marketing, and access. *Biopreserv Biobank* 2017;15:1–2.
201. Simeon-Dubach D, Henderson MK. Sustainability in biobanking. *Biopreserv Biobank* 2014;12:287–91.
202. Kayser J. Cash-Starved deCODE is looking for a rescuer for its biobank. *Science* 2009;325:1054.
203. Baker M. Big biotech buys iconic genetics firm. *Nature* 2012;492:321.
204. Proffitt A. deCODE Publishes largest human genome population study. <http://www.bio-itworld.com>. Bio-IT World 2015.
205. Vaught J, Rogers J, Carolin T, Compton C. Biobankonomics: developing a sustainable business model approach for the formation of a human tissue biobank. *J Natl Cancer Inst Monogr* 2011;2011:24–31.
206. Vaught J, Rogers J, Myers K, et al. An NCI perspective on creating sustainable biospecimen resources. *J Natl Cancer Inst Monogr* 2011;2011:1–7.
207. Henderson MK, Goldring K, Simeon-Dubach D. Advancing professionalization of biobank business operations: performance and utilization. *Biopreserv Biobank* 2019;17:213–8.
208. Henderson MK, Goldring K, Simeon-Dubach D. Advancing professionalization of biobank business operations: a worldwide survey. *Biopreserv Biobank* 2019;17:71–5.
209. Barbareschi M, Cotrupi S, Guarnera GM. Biobanks: instrumentation, personnel and cost analysis. *Pathologica* 2008;100:139–48.
210. Budimir D, Polasek O, Marusic A, et al. Ethical aspects of human biobanks: a systematic review. *Croat Med J* 2011;52: 262–79.
211. Allen MJ, Powers ML, Gronowski KS, Gronowski AM. Human tissue ownership and use in research: what laboratorians and researchers should know. *Clin Chem* 2010;56: 1675–82.
212. Sandor J, Bard P, Tamburini C, Tannsjo T. The case of biobank with the law: between a legal and scientific fiction. *J Med Ethics* 2012;38:347–50.
213. Mello MM, Wolf LE. The Havasupai Indian tribe case—lessons for research involving stored biologic samples. *N Engl J Med* 2010;363:204–7.
214. World Medical A. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. *JAMA* 2013;310:2191–4.
215. Simon CM, Klein DW, Schartz HA. Traditional and electronic informed consent for biobanking: a survey of U.S. biobanks. *Biopreserv Biobank* 2014;12:423–9.
216. Havasupai Tribe of Havasupai Reservation v. Arizona Bd. of Regents. 204 P 3d 1063 (Ariz App Div 1) 2008.
217. Hansson MG, Dillner J, Bartram CR, Carlson JA, Helgesson G. Should donors be allowed to give broad consent to future biobank research? *Lancet Oncol* 2006;7:266–9.
218. Dove ES, Tasse AM, Knoppers BM. What are some of the ELSI challenges of international collaborations involving biobanks, global sample collection, and genomic data sharing and how should they be addressed? *Biopreserv Biobank* 2014;12:363–4.
219. Kettis-Lindblad A, Ring L, Viberth E, Hansson MG. Genetic research and donation of tissue samples to biobanks. What do potential sample donors in the Swedish general public think? *Eur J Public Health* 2006;16:433–40.
220. Elger BS, Caplan AL. Consent and anonymization in research involving biobanks: differing terms and norms present serious barriers to an international framework. *EMBO Rep* 2006;7:661–6.
221. Kaufman D, Bollinger J, Dvoskin R, Scott J. Preferences for opt-in and opt-out enrollment and consent models in biobank research: a national survey of Veterans Administration patients. *Genet Med* 2012;14:787–94.
222. Johnsson L, Hansson MG, Eriksson S, Helgesson G. Patients' refusal to consent to storage and use of samples in Swedish biobanks: cross sectional study. *BMJ* 2008;337:a345.
223. Melas PA, Sjoholm LK, Forsner T, et al. Examining the public refusal to consent to DNA biobanking: empirical data from a Swedish population-based study. *J Med Ethics* 2010;36:93–8.
224. Lewis MH, Goldenberg A, Anderson R, Rothwell E, Botkin J. State laws regarding the retention and use of residual newborn screening blood samples. *Pediatrics* 2011;127:703–12.
225. Botkin JR, Goldenberg AJ, Rothwell E, Anderson RA, Lewis MH. Retention and research use of residual newborn screening bloodspots. *Pediatrics* 2013;131:120–7.
226. Gurwitz D, Fortier I, Lunshof JE, Knoppers BM. Research ethics. Children and population biobanks. *Science* 2009;325:818–9.

227. John T, Hope T, Savulescu J, Stein A, Pollard AJ. Children's consent and paediatric research: is it appropriate for healthy children to be the decision-makers in clinical research? *Arch Dis Child* 2008;93:379–83.
228. Knoppers BM, Saginur M. The Babel of genetic data terminology. *Nat Biotechnol* 2005;23:925–7.
229. Maschke KJ. Navigating an ethical patchwork—human gene banks. *Nat Biotechnol* 2005;23:539–45.
230. Kulynych J, Korn D. The new HIPAA (Health Insurance Portability and Accountability Act of 1996) Medical Privacy Rule: help or hindrance for clinical research? *Circulation* 2003;108:912–4.
231. Langlois A. The UNESCO Bioethics Programme: a review. *New Bioeth* 2014;20:3–11.
232. Levesque E, Joly Y, Simard J. Return of research results: general principles and international perspectives. *J Law Med Ethics* 2011;39:583–92.
233. Johnson G, Lawrenz F, Thao M. An empirical examination of the management of return of individual research results and incidental findings in genomic biobanks. *Genet Med* 2012;14:444–50.
234. Wolf SM, Crock BN, Van NB, et al. Managing incidental findings and research results in genomic research involving biobanks and archived data sets. *Genet Med* 2012;14:361–84.
235. Tassey AM. The return of results of deceased research participants. *J Law Med Ethics* 2011;39:621–30.
236. Tassey AM. Biobanking and deceased persons. *Hum Genet* 2011;130:415–23.
237. Allen NL, Karlson EW, Malspeis S, Lu B, Seidman CE, Lehmann LS. Biobank participants' preferences for disclosure of genetic research results: perspectives from the OurGenes, OurHealth, OurCommunity project. *Mayo Clin Proc* 2014;89:738–46.
238. McElfish PA, Long CR, James LP, et al. Characterizing health researcher barriers to sharing results with study participants. *J Clin Transl Sci* 2019;3:295–301.
239. Henderson GE, Edwards TP, Cadigan RJ, et al. Stewardship practices of U.S. biobanks. *Sci Transl Med* 2013;5:215cm7.
240. Washington University, v. William J. Catalona. Appeals from the United States No 06–2301 District Court 2007.
241. Kaiser J. Biomedical research. Court decides tissue samples belong to university, not patients. *Science* 2006;312:346.
242. Fortin S, Pathmasiri S, Grintuch R, Deschenes M. 'Access arrangements' for biobanks: a fine line between facilitating and hindering collaboration. *Public Health Genomics* 2011;14: 104–14.
243. Conroy M, Sellors J, Effingham M, et al. The advantages of UK Biobank's open-access strategy for health research. *J Intern Med* 2019;286:389–97.
244. Colledge F, Elger B, Howard HC. A review of the barriers to sharing in biobanking. *Biopreserv Biobank* 2013;11:339–46.
245. Norlin L, Fransson MN, Eriksson M, et al. A minimum data set for sharing biobank samples, information, and data: MIABIS. *Biopreserv Biobank* 2012;10:343–8.
246. Zawati MH, Borry P, Howard HC. Closure of population biobanks and direct-to-consumer genetic testing companies. *Hum Genet* 2011;130:425–32.

**MULTIPLE CHOICE QUESTIONS**

1. What is desiccation?
  - a. Excessive loss of heat
  - b. Excessive loss of cold
  - c. Excessive loss of air pressure
  - d. Excessive loss of dry matter
  - e. Excessive loss of moisture
2. What is the definition of platelet poor plasma?
  - a. A platelet count  $<10 \times 10^9/L$
  - b. A platelet count  $>10 \times 10^9/L$
  - c. A platelet count  $<10 \times 10^8/L$
  - d. A platelet count  $<10 \times 10^7/L$
  - e. A platelet count  $<10 \times 10^6/L$
3. What is the acceptable approximate maximum amount of blood to be drawn from children in one draw (not neonates)?
  - a. A volume of approximately 0.01% of body weight
  - b. A volume of approximately 0.1% of body weight
  - c. A volume of approximately 0.12% of body weight
  - d. A volume of approximately 0.2% of body weight
  - e. A volume of approximately 1.0% of body weight
4. What does SOP stand for?
  - a. Standard operation problem
  - b. Scientific operating procedure
  - c. Standard ownership procedure
  - d. Standard operating procedure
  - e. Steward ownership problem
5. What is the definition of accreditation?
  - a. “The procedure by which a third party gives written assurance that a product, process, or service conforms to specific requirements”
  - b. “The procedure by which a third party gives formal recognition that a body or person is competent to carry out specific tasks”
  - c. “The procedure by which an authoritative body gives written assurance that a product, process, or service conforms to specific requirements”
  - d. “The procedure by which an authoritative body gives written assurance that a product or service conforms to specific requirements”
  - e. “The procedure by which an authoritative body gives formal recognition that a body or person is competent to carry out specific tasks”
6. What is a custodian?
  - a. The individual responsible for the management of the biobank
  - b. The individual collecting samples to the biobank
  - c. The individual funding the biobank
  - d. The individual owning the biobank
  - e. The individual requesting samples from the biobank
7. What is a material transfer agreement (MTA)?
  - a. An agreement that governs the transfer of data between two organizations
  - b. An agreement that governs the transfer of tangible research materials and data between two organizations
  - c. An agreement that governs the transfer of privacy information between two organizations
  - d. An agreement that governs the transfer of tangible research materials between two organizations
  - e. An agreement that governs the transfer of consent between two organizations
8. What was one of the issues with the “*The Havasupai Indian Tribe Case*”?
  - a. Lack of informed consent
  - b. Biobank ownership
  - c. Biobank bankruptcy
  - d. Biobank custodianship
  - e. Lack of identifiability of data
9. Why is dry ice considered hazardous during transportation?
  - a. Explosion, suffocation, and allergic hazard
  - b. Explosion, suffocation, and ingestion hazard
  - c. Explosion, suffocation, and contact hazard
  - d. Explosion, suffocation, and cutting hazard
  - e. Electrical, suffocation, and contact hazard
10. How much does liquid nitrogen expand when it vaporizes compared to its original volume?
  - a. 70 to 80 times
  - b. 700 to 800 times
  - c. 7 to 8 times
  - d. 7000 to 8000 times
  - e. 0.7 to 8 times

# Laboratory Support of Pharmaceutical, In Vitro Diagnostics, and Epidemiologic Studies\*

Omar Fernando Laterza, Amar Akhtar Sethi, Theresa Ambrose Bush, and Nader Rifai<sup>a</sup>

## ABSTRACT

### Background

Biomarkers are used in the clinical laboratory for routine patient care, and in the pharmaceutical industry during drug development, and in establishing safety and efficacy of a candidate drug. Biomarkers are also used in clinical and epidemiologic research to gain a better insight into pathophysiology, to identify predictors of disease, and to refine treatment strategies. The in vitro diagnostic (IVD) industry develops most of the biomarker assays and makes them commercially available. The pharmaceutical and IVD industries, as well as epidemiologic researchers, often seek the help of clinical laboratories in their biomarker studies, thus providing a mutually beneficial and rewarding relationship.

### Content

This chapter describes, in detail, the various areas in which the pharmaceutical and IVD industries and epidemiologic and clinical researchers use biomarkers and illustrates the

ways in which the clinical laboratory can be involved in providing such services, which can be both financially and intellectually rewarding. However, these opportunities have their own challenges, including the need for strict regulatory rules, extensive documentation requirements, and particular data access and storage specifications. The regulatory requirements for performing biomarker testing are described in this chapter; results may be used in premarket submissions to governmental agencies, for both drugs and assay kits. The relevant documents for analytical and clinical evaluations of biomarkers are identified and discussed. Due to the daunting task of summarizing worldwide regulations, the regulatory requirements in the United States are primarily referred to as examples. The reader should refer to their local agencies when assessing the exact needs applicable to their situation. Overall, the goal of this chapter is to provide a general overview to those in the clinical laboratory who are interested in biomarker research collaborations.

\*The full version of this chapter is available electronically on [ExpertConsult.com](#).

<sup>a</sup>The authors acknowledge the contributions of Mark J. Sarno to the earlier version of the chapter in the 6th edition. The authors also acknowledge that a small portion of this chapter was based on Cole TG, Warnick GR, Rifai N. Providing laboratory support for clinical trials, epidemiologic studies, and in vitro diagnostic evaluations. In: Rifai N, Warnick GR, Dominiczak MH, editors. *Handbook of Lipoprotein Testing*. AACC Press, with permission.

## INTRODUCTION

Biochemical markers are routinely and extensively used in patient care to confirm or exclude a diagnosis, screen for a disease, monitor compliance with or a response to a treatment, or assess a prognosis. Over the last several decades, genetic markers have been added to the armamentarium of clinical laboratory tests, with similar uses. To test for these markers, clinical laboratories use reagents and kits that have been developed by the in vitro diagnostic (IVD) industry and that have been approved for clinical use by the US Food and Drug Administration (FDA) in the United States or an equivalent agency, depending on the country. IVD is an industry with more than \$69 billion in sales worldwide.<sup>1</sup> On rare occasions, when a commercial assay is not available for a particular marker, or when an assay needs to be modified to better meet the needs for clinical decision making, clinical laboratory professionals may develop and/or adapt, validate, and offer an assay within a single laboratory for clinical use; such a test is referred to as a laboratory-developed test (LDT).

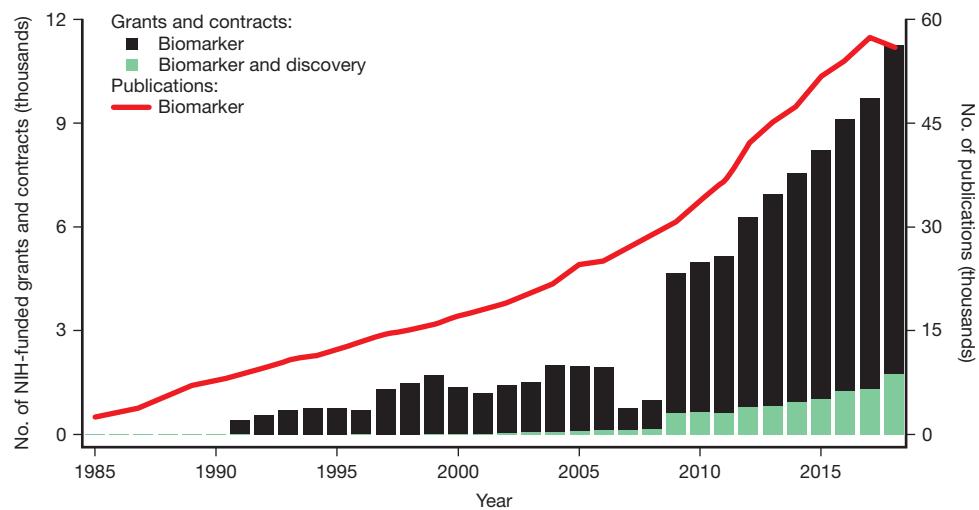
The use of both biochemical and genetic markers is certainly not restricted to clinical laboratories. The pharmaceutical industry uses biomarkers extensively during drug development to determine whether to pursue or abort the efforts to develop a drug, and to establish the safety and efficacy of a candidate drug. As part of drug development, the pharmacokinetics (PK) and pharmacodynamics (PD) of the candidate drug must be examined, thus necessitating the development of an assay for the measurement of the parent drug and its metabolites, as well as biomarkers of PD response. Furthermore, biomarkers can be used as companion diagnostics to identify those individuals who will benefit from a drug or those who are likely to experience adverse effects from a drug (personalized medicine), and on rare occasions, as surrogate markers for a primary endpoint in clinical trials.

In addition to their use in routine patient care and by the pharmaceutical industry, biomarkers are extensively used by clinical and epidemiologic researchers to gain a better insight into pathophysiology, identify predictors of disease, refine

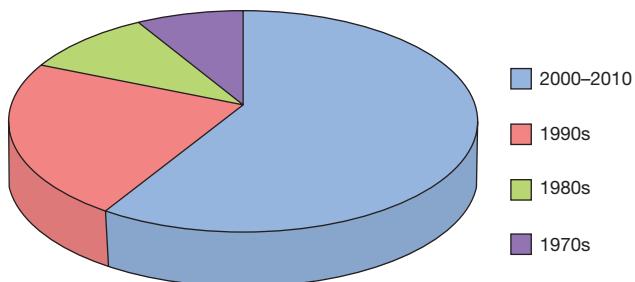
treatment strategies, and develop prognostic indicators. The National Institutes of Health (NIH) have heavily supported biomarker research; according to the NIH RePORTER database, approximately \$36.5 billion were invested in biomarker research from 2009 to 2018, a staggering increase from the \$4.6 billion spent during the 1999 to 2008 period. This investment resulted in a significant intellectual output; a simple search for the term “biomarker” on PubMed found more than 875,000 publications as of February 2020. Fig. 12.1 illustrates the increase in the number of NIH-funded grants that contain the term biomarker in the title and the intellectual output that is reflected by the number of resulting publications over the past two and a half decades (Figure adapted from reference 2).<sup>2</sup>

Clinical chemists are laboratory professionals who understand all aspects of biomarker testing, including the preanalytical, analytical, and postanalytical issues, and who are trained in the de novo development and validation of tests. They have a good understanding of regulatory requirements, appreciate the clinical context of a laboratory test, and practice their profession in a systematic and methodical manner. This combination of skills makes them desirable not only to clinical laboratories but also to the pharmaceutical and IVD industries and to laboratories performing testing for large trial cohorts. For additional discussion on the training, role, and career paths of the clinical chemist, refer to Chapter 1.

An IVD company often needs to partner with a clinical laboratory to validate and characterize the performance of its assay in a real-life setting using a specific patient population. A pharmaceutical company may prefer to contract the development of an assay to measure the concentration of a candidate drug to an outside laboratory rather than performing it in-house. Biomarker testing for safety and efficacy during a clinical trial or in subsequent postmarketing or substudies is usually done in a clinical laboratory setting. These scenarios present financial and intellectual opportunities to clinical chemists practicing in either hospital-based or freestanding clinical laboratories. However, these opportunities certainly



**FIGURE 12.1** Trend in the number of biomarker-related National Institutes of Health (NIH)-funded grants and resulting publications over the past three decades. (Updated from Anderson NL, Ptolemy AS, Rifai N. The riddle of protein diagnostics: future bleak or bright? *Clin Chem* 2013;59:194–7. Courtesy Dr. Adam Ptolemy.)



**FIGURE 12.2** Growth in total (prehuman and clinical) cost per approved new drug. (Derived from DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: new estimates of R&D costs. *J Health Economics* 2016;47: 20–33.)

do not come without their own challenges. Strict regulatory rules, variable workflow, extensive documentation requirements, particular data access, and storage specifications are some of the difficulties. The rewards, however, can be substantial. According to the Tufts Center for Drug Development, the capitalized research and development cost for an approved drug has increased from \$179 million in the 1970s to a staggering \$2.5 billion in the 2000s to early 2010s (Fig. 12.2). It was estimated that in 2006, \$25.6 billion was spent on clinical trials in the United States alone, with a projection of more than \$32 billion for 2011.<sup>3</sup> The global spending on clinical trials is estimated to reach \$69 billion a year by 2025.<sup>4</sup> The growing use of real-world data, such as that gathered from electronic health records systems, and wearable devices, and their analysis based on artificial intelligence such as sophisticated machine-learning algorithms, may aid in the identification of suitable patients for recruitment and optimization of study design that could reduce the cost of clinical trials. However, such an approach is still in its infancy and must be proven; until then the cost of clinical trials remains staggering.<sup>5</sup> Even a small percentage of 5 to 10% that is devoted to laboratory testing can translate into huge sums. In addition to the financial reward, the collaboration between clinical laboratories and the pharmaceutical and IVD industries results in intellectual benefit, in which clinical chemists become involved in the reporting of the findings and publication of the manuscripts resulting from these endeavors.

This chapter describes the nature of the studies in the pharmaceutical and IVD industries and those of pharmaceutical substudies and epidemiologic investigations, the expectations from the clinical laboratory to support such activities, and the involved regulatory requirements. Although the regulatory requirements differ among countries and regions, they tend to be similar in spirit. The regulatory aspects of this chapter are not intended for use as a blueprint to those seeking regulatory guidance or for them to be comprehensive but rather to provide the reader with an overview of the regulatory challenges and requirements in this sphere. Since the US regulations are among the most comprehensive, mature, and well understood, the FDA requirements were used as an example throughout the book to illustrate various points. However, because of the scope and focus of this textbook, both the American and the European Union regulatory requirements for IVD products have been discussed.

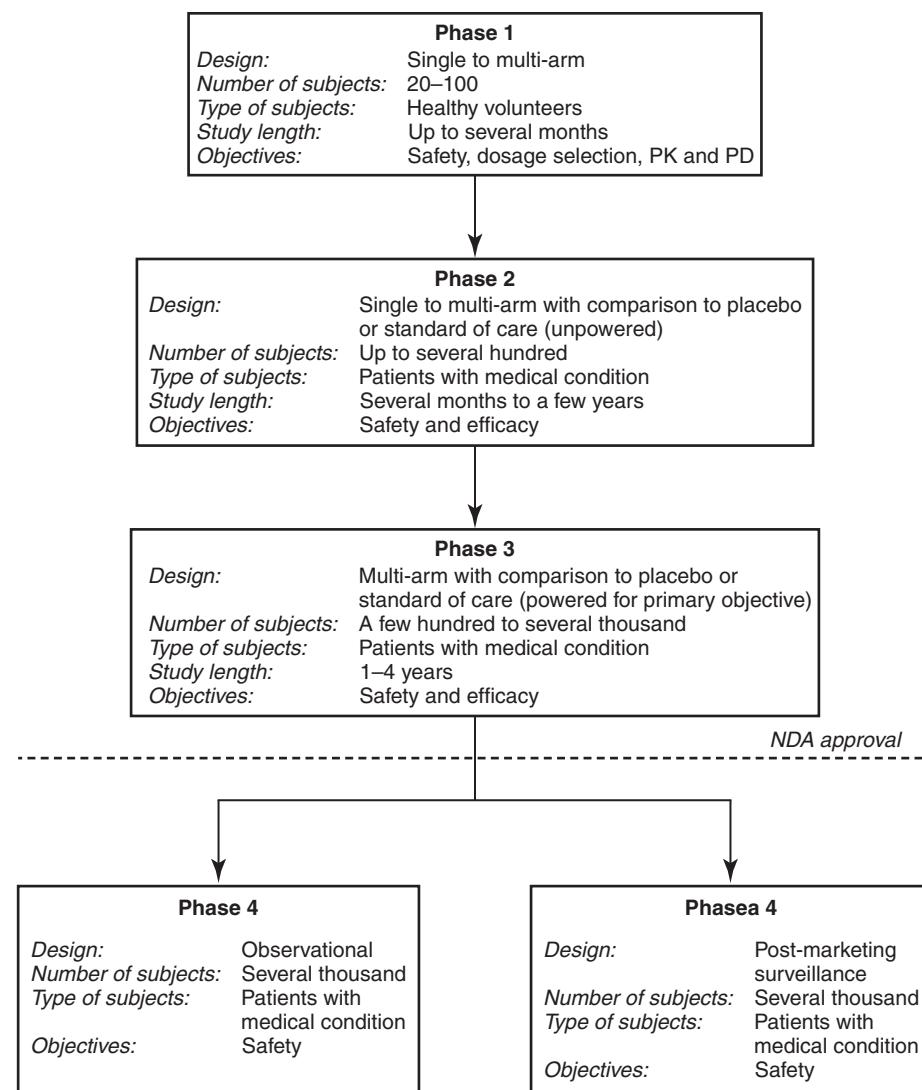
## PHARMACEUTICAL STUDIES

Pharmaceutical investigational studies provide ample opportunity for clinical laboratories to participate in clinical research of profound importance. These studies build upon early research and development of new molecular entities (NMEs), variations of known compounds and formulations, as well as new clinical uses of established drugs, in each case seeking to demonstrate safety and efficacy of the drug candidate for its proposed intended use. Because of the potential risks involved in such clinical research, every individual and organization engaging in these investigations should understand the fundamental principles of the investigational process and the ethical and statutory requirements applicable to their efforts. The following is an overview of the drug development process.

### Drug Development

Most new drug candidates proceed through two major phases of development: preclinical and clinical trials. Preclinical trials (also known as nonclinical studies) are performed to assess the safety and tolerability of the drug in animals. These studies provide insights into the maximum tolerated dose (the maximum dose that an animal species can tolerate for a major portion of its lifetime without significant impairment or toxic effect other than carcinogenicity), the type and frequency of adverse events experienced in the framework of the animal model(s), effects on reproduction, and mutagenicity. Pharmacologic properties of the drug in animal models are also investigated. These include PK effects (the examination of the process, magnitude, and rates of absorption, distribution to organs, metabolism, and elimination of the drug [ADME]) and PD effects (the biochemical and physiologic effects of the drug). Preliminary observations of efficacy may be made in animal models of a particular disease and the inference is subsequently drawn that similar benefits may accrue in humans. However, it is generally acknowledged that observed safety, pharmacologic, and efficacy effects in animals do not always translate to humans. Completely different effects may be observed in humans than in those previously observed in animals, or the magnitude of observed effects may differ significantly between species. Variables that influence the correlation of effects between animal and human investigations include the species of animals used in the study, drug dosing concentrations and method of administration, time of observation, measurement methods, and inherent differences of biochemistry and physiology. In many cases, these variables cannot be adequately controlled to enable the accurate prediction of effects in humans based on experiences in animals. Nevertheless, the preclinical phase represents the basic foundation of the *in vivo* drug development effort.

Following the completion of preclinical trials, the sponsoring organization prepares an investigational new drug (IND) application for submission to the FDA (when filed in the United States; procedures vary among countries). The IND includes the results of the preclinical trials, information on the drug manufacturing process, physico-chemical characterization of the drug (molecular size, structure, stability in certain environments, and so on), and a proposed plan for clinical investigation in humans. The clinical investigational plan is divided into multiple phases (Fig. 12.3). The studies



**FIGURE 12.3** Drug development clinical study phases. NDA, New drug application; PD, pharmacodynamics; PK, pharmacokinetics.

constituting these phases typically are intended to satisfy “adequate tests and substantial evidence necessary for New Drug Application (NDA) approval.”

The intent of the Phase 1 PK/PD study program is to “determine the metabolism and pharmacologic actions of the drug in humans, the side effects associated with increasing doses, and, if possible, to gain early evidence on effectiveness.”<sup>6</sup> The design of a Phase 1 study generally involves administration of the drug in a small number of healthy subjects (perhaps as few as 20); however the Phase 1 studies in oncology are normally done in cancer patients given the high toxicities associated with these drugs. The assumption is that healthy subjects are an acceptable preliminary human model for the patient population for which the drug is intended, and that the general health of the studied subjects may attenuate any adverse effects of the drug. The so-called “mechanism of action” of the drug may also be determined in this phase, although the sponsor must only report the mechanism “if known.”<sup>7</sup> The mechanism of action may be, for instance, a binding of a cell receptor, an inhibition of an enzyme, or perhaps an effect on nucleic acid or protein synthesis. Some

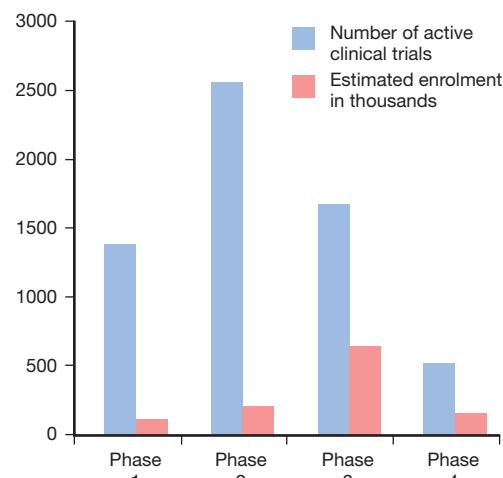
drugs are well characterized in this area. For example, proton pump inhibitors such as omeprazole (Prilosec OTC, Procter and Gamble, Cincinnati, OH) and related compounds are a particularly well-characterized class of drugs. These drugs act by irreversibly blocking the hydrogen/potassium adenosine triphosphate enzyme system (the gastric proton pump) of gastric parietal cells, thereby blocking secretion of acid into the gastric lumen. This effect can be clearly demonstrated in healthy individuals and patients with gastrointestinal (GI) disease. In the affected population, omeprazole promotes healing of GI tissue damage that results from active duodenal ulcers, gastroesophageal reflux disease, erosive esophagitis, simple heartburn, and other related conditions. According to the FDA, approximately 70% of drugs under evaluation successfully complete the Phase 1 investigation and advance to the next phase.

Phase 2 studies include “controlled clinical studies conducted to evaluate the effectiveness of the drug for a particular indication or indications in patients with the disease or condition under study and to determine the short-term side effects and risks associated with the drug. Phase 2 studies are

typically well controlled, closely monitored, and conducted in a relatively small number of patients, usually involving no more than several hundred subjects.”<sup>6</sup> Although Phase 2 studies are well controlled, they are not statistically powered to determine efficacy. The logic of limiting the study size is to limit unneeded exposure of a novel drug candidate with limited development knowledge. Phase 2 studies may evaluate subjects over a few months to a few years, and approximately 30% of drugs in this phase move on to the next phase of investigation. If any additional safety or preliminary efficacy concerns are revealed, a follow-up Phase 2 study may be performed. If primary safety and efficacy is established, then the Phase 3 investigation can proceed.

Phase 3 studies are the final pre-NDA phase. These studies are “intended to gather the additional information about effectiveness and safety that is needed to evaluate the overall benefit–risk relationship of the drug and to provide an adequate basis for physician labeling.”<sup>6</sup> These studies are statistically powered to directly address the question of efficacy through recruitment of up to several thousand subjects, if needed. In clinical areas such as cardiovascular disease (CVD), Phase 3 studies often require long-term follow-up of subjects to determine hard endpoint outcomes such as overall survival, or alternatively as in the case of various cancers, progression-free survival. As such, Phase 3 studies are the costliest of all the investigations, both as a result of the requirement for statistical powering and for the long-term evaluation of certain drugs and intended uses. According to the FDA, approximately 30% of drugs completing Phase 3 studies receive subsequent approval. However, these statistics take into account NDAs for known compounds in new dosage forms and generic drugs. Obviously, there are fewer questions of safety and efficacy in minor modifications of known drugs and generics, and thus a higher likelihood of receiving an approval to the subsequent NDA. When looking purely at NMEs, the cumulative rate of passing all three phases of new drug development and achieving FDA approval is less than 15%.

Clinical research for new drugs sometimes involves another phase. Phase 4 studies are postapproval studies, which may be viewed through two lenses. In one aspect, Phase 4 studies provide additional evidence of safety and “real-world” effectiveness of a drug as evaluated in an observational, non-interventional trial, which is generally run under a formalized protocol. In another aspect, Phase 4 studies represent formal postmarketing surveillance evaluations intended to gather information through an adverse event monitoring system. In both views, the intent is to detect a sign that might necessitate a regulatory action, such as a change in labeling or initiation of a risk management system. These studies may be mandated under various statutes in the United States, including the Food and Drug Administrative Amendments Act of 2007<sup>8</sup> and under accelerated approval requirements<sup>9,10</sup> (often for drugs approved under the “fast track” designation<sup>11</sup> for particularly life-threatening diseases, such as HIV), deferred pediatric studies,<sup>12,13</sup> for which such studies are required under the Pediatric Research Equity Act,<sup>8</sup> and for studies in humans for drugs originally approved under the Animal Efficacy Rule.<sup>14,15</sup> Beyond these statutory requirements, pharmaceutical sponsors may make voluntary postmarketing commitments to perform Phase 4 studies and are expected to complete these studies in good faith. Fig. 12.4 presents the



**FIGURE 12.4** Total number of clinical trials (6119) and participants (1,145,118) in the United States in 2013. (Data courtesy Dr. William Chin.)

numbers of the studies, Phases 1 to 4, in the United States in 2013 and the numbers of participants.

### Use of Biomarkers in Pharmaceutical Studies

Biomarkers have long been used in pharmaceutical research to assess toxicity, but starting in the early 1990s, these markers received increased attention due to the sharply increasing costs of drug development. Therefore biomarkers were considered tools that could help increase the efficiency of the drug development process. Currently, biomarkers are used throughout the drug development process, including leading compound selection, determining dosage, defining and understanding the mechanism of action, and identifying the patients who are most likely to benefit from the drug. Well-qualified biomarkers, especially those used as surrogate endpoints, may have a broader usefulness in regulatory context and clinical practice.

The usefulness of biomarkers and surrogate endpoints to further enhance the drug development process has been the subject of regulatory emphasis. In 2004, the FDA launched the Critical Path Initiative (CPI) with the release of a report titled “Innovation/Stagnation: Challenge and Opportunity on the Critical Path to New Medical Products.”<sup>16</sup> The CPI was the strategy adopted by the agency to “drive innovation in the scientific processes through which medical products are developed, evaluated, and manufactured.” In this first report, there was an emphasis on how biomarkers and surrogate endpoints could be used to optimize the drug development process, putting biomarkers at the forefront in pharmaceutical research. More recently, the Center for Drug Evaluation Research (CDER) within the FDA released a Guidance for Industry on the qualification of “Drug Development Tools,” which specifically presents a process for qualification of a biomarker or biomarkers for a given “context of use” in drug studies.<sup>17</sup> Once qualified, the biomarker(s) may be used within context in any drug study.

There are two main areas in which biomarkers play a strategic role in drug development: (1) in enabling decision making at the early stages of drug development (whether to pursue, abort, or accelerate the development of the drug and in dose selection), and (2) in patient selection. Most of the

biomarker research findings from the early strategic stage remain confidential and inaccessible to the public. However, biomarker-enabled decisions are extremely valuable to the pharmaceutical industry because they lead to significant cost and time reductions. These early decisions could be based on lack of efficacy, presence of toxicity, off-target effects of the candidate drug, or an unfavorable PK profile of the candidate drug. For biomarkers to play a strategic role in this space, good translational models and well-validated and robust clinical assays are required. Biomarkers used in patient selection tend to be used in late-phase clinical trials (Phase 3). These biomarkers are more commonly reported in the public domain and often become companion diagnostics (covered in more detail later in this chapter).

### Clinical Laboratory Opportunities

Because the framework of drug development and some background on the use of biomarkers has now been examined, the many opportunities for clinical laboratories to participate in the process can now be discussed.

#### Animal Studies

Clinical laboratories are not particularly suited to animal safety studies for two reasons: (1) these studies require strict conformance to Good Laboratory Practices (GLP), and (2) the same facility housing the animals must perform the laboratory analyses. Thus these studies are often contracted to large laboratories associated with contract research organizations (CROs) that specialize in animal safety studies. In contrast, pharmaceutical sponsors may award analytical studies to clinical laboratories with recognized specialization in certain clinical or methodological areas. Such laboratories may provide expertise in complex methods such as mass spectroscopy or ultra-high-performance liquid chromatography. These laboratories offer services in methods development for animal models for specific compounds and associated metabolites, provision of kits for sample/specimen collection and shipping, sample analysis and retention, and calculation of results. Thus a clinical laboratory with experience in methods development and validation will find opportunities in the preclinical/animal study phase of drug development. Specifically, the laboratory may participate in:

1. Development and validation of new biomarker methods in specific animal models;
2. Transfer and validation of existing methods for human biomarkers in one or more animal models;
3. Development and validation of methods for measurement of parent drug and drug metabolites for PK assessment.

#### Human Studies

Pharmaceutical studies in humans require an extra level of diligence on the part of the clinical laboratory. At a minimum, sponsors screen potential laboratories for their abilities to perform testing on general organ panels, hematology parameters, serum chemistries, and urinalysis. These results often form the basis of baseline testing and ongoing monitoring of the health of study subjects during, and sometimes after, the drug trial. The pharmaceutical sponsor will assess the ability of the laboratory to receive, to accession and store samples, to perform testing with rapid turnaround and accurate results, to issue reports on a timely basis in accordance with the study protocol, to maintain blinding if required by

the study protocol, and to retain records for later review by sponsor and regulatory resources; the sponsor will also monitor the quality of testing over the study period.

Beyond basic testing, sponsors may select laboratories in much the same way for clinical studies as for preclinical studies, that is, for recognized specialization in certain clinical and methodological areas, as well as the ability to develop and validate methods. Methods development and specialty testing is of great importance in five particular situations:

1. In PK and PD assessments
2. A situation in which a biomarker or a panel of biomarkers provides critical information for the drug safety assessment
3. In which a biomarker or panel of biomarkers provides an indication of the dose-response relationship of the drug
4. In which a biomarker or panel of biomarkers represents a primary, secondary, or exploratory endpoint of the study, particularly when the marker(s) act as a surrogate for clinical endpoint(s)
5. In which a biomarker is used as a companion diagnostic for personalized medicine.

Each of these areas is discussed here in turn.

#### Pharmacokinetic and Pharmacodynamic Assessments

It is generally recognized that PK assessments are indicative of what the body does to the drug, whereas PD assessments are indicative of what the drug does to the body.

The PK assessment involves dosing of subjects with a drug under investigation through any one of several routes (intravenous, subcutaneous, intramuscular, topical/transdermal, or per oral/sublingual). This is generally followed by peripheral blood sampling at various time intervals as appropriate for the drug. In certain circumstances, some drugs may require analysis in unusual biological fluids (e.g., cerebrospinal fluid, saliva, sweat, and so on). Analytical methods are used to measure the presence and concentrations of the parent drug and its metabolites of interest from which PK parameters are calculated. The objective is to evaluate the ADME characteristics of the drug and its metabolites. Liquid chromatography-mass spectroscopy is considered the gold standard for the measurement of small molecule drugs (nonbiologics), whereas immunoassays are commonly used for the measurement of biological drugs (e.g., monoclonal antibodies). For study designs that require characterization of both parent drug and metabolites, mass spectrometry often provides the most accurate and efficient approaches for simultaneous determination of multiple compounds. The methods used may be suitable for direct analysis in blood, plasma, or serum, or may require various sample pretreatments before analysis.

Methods already approved by regulatory agencies may exist for the measurement of the compounds of interest and be available at a CRO laboratory. For NMEs, an assay is normally developed and deployed internally before it is transferred to a CRO. In any case, the clinical laboratory should follow the Bioanalytical Methods Validation Guidance issued by the FDA<sup>18</sup> or similar guidance issued by its counterpart in Europe, the European Medicines Agency (EMA) or other local agencies, depending on the country. This guidance provides the established framework for validation, including validation parameters for “chemical assays” and for microbiological and ligand-binding assays. For chemical assays, the document specifies investigations of analyte selectivity,

accuracy, precision, recovery, calibration robustness, and analyte stability in short- and long-term storage, as well as stability under multiple freezing and thawing cycles. For microbiological and ligand-binding assays, the document discusses considerations of selectivity and quantification issues. The guidance further describes the desired components of an assay validation report and retention of data for later auditing and inspection. The sponsor, and therefore the clinical laboratory performing the analyses, must supply documentation to the FDA or EMA as part of the NDA, including summary information, method development and establishment information, and bioanalytical reports of the application of the method to routine analysis.

Once the method is adopted into routine analysis, accurate and thorough records of each sample analysis must be maintained. Assay runs should meet prespecified acceptance criteria for quality control to ensure validity, and individual patient results should be inspected for errors or omissions. Verified and secured data sets containing the PK results for each study subject will then be analyzed by a kineticist. Often, a pharmaceutical sponsor will have an in-house kineticist or will contract the PK analysis to a consulting kineticist; however, some laboratories offer PK analysis services.

In general, the PK data will be plotted for each individual subject with time postdose on the  $x$ -axis versus drug or metabolite concentration on the  $y$ -axis. Data are fit by either model-dependent and/or compartmental analysis or model-independent and/or noncompartmental analysis. Typical parameters calculated and reported include maximum concentration ( $C_{max}$ ), the corresponding time ( $t_{max}$ ), half-life ( $t_{1/2}$ ), elimination rate constant ( $K$ ), area under the curve from time zero to last sampling ( $AUC_{0-t}$ ) and/or from time zero to infinity ( $AUC_{0-\infty}$ ), volume of distribution ( $V_d$ ), and clearance (CL). For additional discussion on these parameters refer to Chapter 42. For drugs administered over long periods of time, the kineticist may determine various parameters characterizing drug accumulation and circulating concentrations at steady state. Furthermore, for combination drugs or drugs often used as part of a multidrug regimen in a particular health condition, the kineticist may model drug–drug

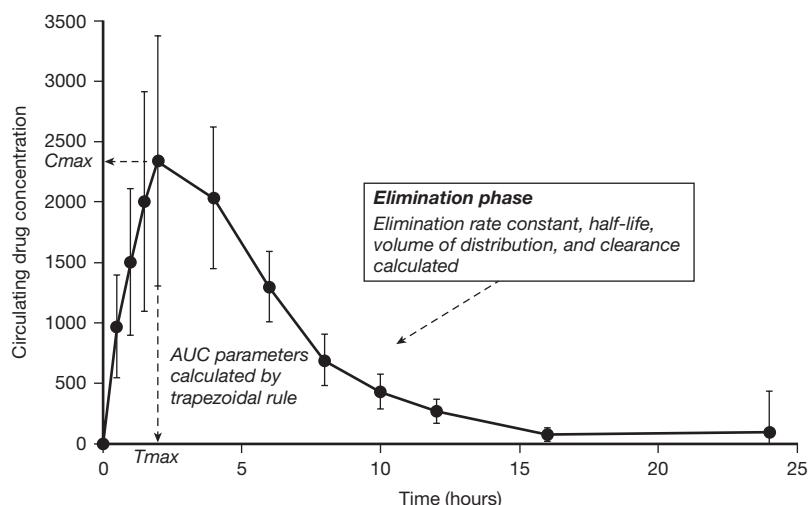
interactions to determine whether safety or efficacy varies depending on the presence and magnitude of an interaction. PK studies and specific analyses may also be conducted in populations of special interest (e.g., hepatically or renally impaired populations). As a typical summary of a PK study, the average concentrations across all subjects at the selected time points are usually presented as a line plot with population error bars. Fig. 12.5 provides an example time-concentration PK plot for a chemotherapeutic agent dosed subcutaneously in an advanced cancer study population.

The PD assessments capture the biochemical and physiologic effects of the drug and its metabolites. These effects run the spectrum from simple drug-induced variations of vital signs, such as heart rate and blood pressure, to complex modeling of drug-target binding and any downstream effects as a result of this interaction. In this regard, the PD assessment attempts to characterize the indirect effects on physiologic parameters and the specific effects on the tissue target (if known). Within this spectrum, various safety assessments may be performed, such as drug effect on liver enzymes or renal function. As with PK studies, the burden will be on the clinical laboratory to accurately test and report any parameter associated with a planned PD analysis.

### Biomarkers for Drug Safety Assessment

Biomarkers to assess drug safety have been extensively used for decades, in both preclinical and clinical research studies. Historically, to detect drug-induced toxicity, most of the safety biomarkers used during drug development were routine clinical laboratory tests that were used to assess tissue and/or organ injury and/or function. These tests include those used to assess liver function (e.g., transaminases, bilirubin, alkaline phosphatase), kidney function (e.g., serum creatinine, creatinine clearance, cystatin C), skeletal muscle (e.g., myoglobin) or cardiomyocyte injury (e.g., creatine kinase-myocardial band, troponins I and T), and bone biomarkers (e.g., bone-specific alkaline phosphatase).

It is generally accepted that preclinical toxicology models perform well in predicting clinical toxicity. Approximately 70% concordance was found in retrospective studies.<sup>19,20</sup> The best



**FIGURE 12.5** A typical pharmacokinetic time–concentration plot: average circulating drug concentration values and variation from a population of advanced cancer patients ( $n = 21$ ). AUC, Area under the curve.

concordance was observed in hematological, GI, and cardiovascular toxicities, whereas hepatic and hypersensitivity and/or cutaneous reactions had the poorest concordance.<sup>21</sup> However, approximately 25% of drug failures in Phase 2 studies are still due to drug toxicity. For this reason, there has been a great deal of effort in drug development to discover novel and perhaps better biomarkers of organ toxicity. For instance, the Predictive Safety Testing Consortium of the Critical Path Institute has been spearheading efforts for the qualification of new biomarkers of drug-induced tissue and/or organ injury. The Predictive Safety Testing Consortium leads an extensive collaboration between a large number of pharmaceutical companies and regulatory agencies, such as the FDA, the EMA, and the Japanese Pharmaceuticals and Medical Devices Agency, and has successfully qualified several acute kidney injury (AKI) biomarkers (albumin, β2-microgobulin, clusterin, cystatin C, kidney injury molecule-1, trefoil factor-3, and total urinary protein) in pre-clinical animal studies.<sup>22</sup> This qualification was endorsed by the FDA and EMA. Although the qualification of some of these biomarkers for clinical usefulness is currently ongoing in a systematic fashion, it has been concluded that the use of these biomarkers in clinical trials should be considered on a case-by-case basis.

There are also efforts to develop novel toxicity biomarkers for organs such as the liver, skeletal muscle, the heart, muscles, and the vasculature.<sup>23,24</sup> A great deal of these efforts are being spearheaded by the TransBioLine project (<https://transbioline.com/>). The goal of this project, which is composed of academic, industry, and government organizations, is to develop novel safety biomarkers that will reliably indicate organ injury for drug development purposes. Currently, there are (1) five organ-specific work packages (WP) including liver, kidneys, pancreas, blood vessels, and central nervous system, (2) a liquid biopsy WP, (3) a sample and assay development WP, and (4) a data management and analysis WP. It is expected that this project will create the needed infrastructure and processes for long-term research and development of safety biomarkers. It would not be surprising if some biomarkers initially developed to monitor drug-induced injury eventually find their way into clinical practice,

thus reversing the historical flow of safety biomarkers mentioned previously. Most of these biomarkers are summarized in Table 12.1.

The emergence of the fields of proteomics, genomics, and metabolomics has also signaled a new era of biomarker use. These more novel methods may be applied as companion diagnostics (covered in more detail later in this chapter) to determine which patients receiving a given drug may benefit or encounter specific adverse events, and potentially, to what magnitude they may experience such effects. As an example, pharmacogenomic tests have been used to determine which patients receiving an antipsychotic drug may experience suicidal thoughts based on variants of two receptor genes.<sup>25</sup> (For additional information of pharmacogenomics, refer to Chapter 73.) Another example, pharmacogenetic tests for cytochrome P450 2D6 (CYP450 2D6) polymorphisms, have been cleared by the FDA (e.g., AmpliChip CYP450 test, Roche Molecular Diagnostics, Pleasanton, California) or validated as an LDT to characterize individual metabolic characteristics that might contribute to variances in safety or efficacy of drugs metabolized by this enzyme.

Also, in some cases established biomarkers have been investigated for new indications. For instance, cardiac troponin has been investigated as a risk marker for cardiotoxicity for women taking trastuzumab for advanced breast cancer and for children who receive doxorubicin for the treatment of acute lymphoblastic leukemia.<sup>26</sup> Similar potential has been demonstrated for C-reactive protein (CRP), when measured with a high-sensitivity assay in the former population, with the marker displaying more than 90% sensitivity for detection of reduced left ventricular ejection fraction as a result of long-term trastuzumab therapy.<sup>27</sup>

As in the foregoing discussion on PK and PD, laboratories with method development and validation skills will attract pharmaceutical sponsors to support drug safety assessments. A laboratory that also possesses a particular expertise in the emerging diagnostic fields, such as pharmacogenomics with a broad menu of established tests, will find themselves in high demand for participation in drug studies.

**TABLE 12.1 Representative Biomarkers Used in Safety Studies**

Organ/Tissue Toxicity	Classical Biomarkers	Biomarkers Evaluated in Preclinical Animal Models
Acute kidney injury	sCr, urea, urinary albumin	Cystatin C, NGAL, KIM-1, Clusterin, Trefoil factor 3, GST-α, GST-π, fibrinogen, miRNAs (all in urine)
Liver	AST, ALT, GGT, 5'NT	GLDH, MDH, PON-1, PNP, ARG-1, SDH, GST-α, miRNA122
Muscle	AST, CK, myoglobin	sTnI, sTnT, CK protein M, Pvalb, Myl3, Fabp3, Aldoa
Heart	cTnI, cTnT	Natriuretic peptides, interleukin 6, myeloperoxidase, sCD40 Ligand
Pancreas	Amylase, lipase	TAP
Vasculature		VEGF, GRO/CINC-1, TIMP-1, vWGpp, NGAL, TSP-1, smooth muscle actin, calponin, transgelin

5'NT, 5' nucleotidase; Aldoa, aldolase A; ALT, aspartate aminotransferase; ARG-1, arginase; AST, alanine aminotransferase; CK, creatine kinase; cTnI, cardiac troponin I; cTnT, cardiac troponin T; Fabp3, fatty acid-binding protein 3; GGT, γ-glutamyl transpeptidase; GLDH, glutamate dehydrogenase; GRO/CINC-1, growth-regulated gene product/cytokine-induced neutrophil chemoattractant-1; GST, glutathione-S-transferase; KIM-1, kidney injury molecule-1; MDH, malate dehydrogenase; miRNA, microRNA; Myl3, myosin light chain 3; NGAL, neutrophil gelatinase-associated lipocalin; PNP, purine nucleoside phosphorylase; PON-1, paraoxonase/arylesterase 1; Pvalb, parvalbumin; sCr, serum creatinine; SDH, sorbitol dehydrogenase; sTnI, skeletal troponin I; sTnT, skeletal troponin T; TAP, trypsin activation peptide; TIMP-1, metalloprotease inhibitor-1; TSP-1, thrombospondin-1; VEGF, vascular endothelial growth factor; vWGpp, von Willebrand propeptide.

### Biomarkers for Dose–Response Relationships

In drug development, the preclinical phases of research often involve measurement of a broad list of biomarkers for target engagement, PD response, and safety. Some of these biomarkers can be used as safety and efficacy indicators in early clinical development to give early insight into how the compound is working in humans and to inform dose selection for Phase 2 studies. The most useful biomarkers for this purpose are those that offer a clear dose response in PK/PD assessments (see section on Pharmacokinetic and Pharmacodynamic Assessments). Many Phase 1 study designs use a multiarm format, with the study arms defined by dose level. The resulting safety and PK and/or PD data are therefore analyzed individually by study arm. As part of this investigation, the selected biomarkers will be scrutinized for indications of both beneficial effects and safety risks. Similarly, Phase 2 studies often use multiple dose levels as a final attempt to select the final drug dose to evaluate in Phase 3 studies. As such, specific biomarkers may serve as one criterion in dose selection.

As an example of biomarkers used in this fashion, the CYP450 2D6 pharmacogenetic test is examined again. The test categorizes individuals into four groups of metabolic status: poor (little or no CYP450 2D6 function), intermediate (function between poor and normal), extensive (normal function), and ultra-rapid (multiple copies of the CYP450 2D6 gene are expressed, and therefore yield a greater than normal function). These categories have profound impact in the setting of opioid therapy for pain management. For drugs metabolized by this enzyme, such as hydrocodone, safety and efficacy will vary by metabolizer status. Specifically, CYP450 2D6 converts hydrocodone to hydromorphone, a more potent opioid. Thus ultra-rapid metabolizers will experience overexposure to the narcotic effects, which may cause serious adverse events or even death. Conversely, poor metabolizers will experience a reduced analgesic benefit due to underdosing of the more potent metabolite.

Thus clinical laboratories with broad and specialized menus of tests and methodologies provide critical resources for assessment of dose–response relationships in drug development. In some cases, as with the CYP450 2D6 example, the biomarker(s) will be used to categorize individual study subjects at baseline based on the potential safety and efficacy profiles of the drug in subpopulations of interest. In other cases, the biomarker(s) may be used to assess safety and efficacy by dose level after the completion of the study as a reflection of drug exposure.

The clinical development of sitagliptin is also a good example of biomarkers helping to expedite drug development through a strong dose–response relationship. Sitagliptin, a first-in-class dipeptidyl peptidase-4 (DPP-4) inhibitor, benefited from a robust engagement between the drug and its target (target engagement) and use of well-qualified disease-related biomarkers (i.e., glucose, glycosylated hemoglobin [HbA<sub>1c</sub>]) to achieve proof-of-concept, facilitate the design of clinical efficacy trials, and streamline dose focus and optimization studies. The DPP-4 enzyme inactivates GLP-1, which is a gut hormone involved in the regulation of blood glucose concentrations. Thus target engagement biomarkers such as DPP-4 activity, and biomarkers that demonstrated a close relationship to the target disease (proximal biomarkers), such as active and inactive GLP-1, were used in clinical development. These were clearly characterized in preclinical

species and used for the translation strategy of sitagliptin. Specifically, preclinical experiments demonstrated that approximately 80% inhibition of DPP-4 activity was associated with maximal lowering of glucose concentrations.<sup>28</sup> This also correlated with an increase in plasma GLP-1 concentration. PK and/or PD modeling revealed that the concentration that yielded 80% of the maximum effective response of plasma DPP-4 inhibition corresponded to a plasma sitagliptin concentration of approximately 100 nmol/L. It was also determined that a single dose of 200 mg provided DPP-4 inhibition (>80%) for 24 hours.<sup>29</sup> This finding allowed the rapid determination of the ideal dose to move to the next steps in clinical development. This success was a result of the excellent usefulness of these biomarkers in preclinical species and the development of a careful translational strategy.

Adiponectin is yet another example. Specifically, it was used in the development of peroxisome proliferator-activated receptor (PPAR) agonists in patients with type 2 diabetes. A number of PPAR agonists in the thiazolidinedione family have been developed for the treatment of diabetes. Unfortunately, at the time of development of the initial drugs such as troglitazone, rosiglitazone, and pioglitazone, there were no target engagement biomarkers that could be used to help establish dose selection. Since the development of the initial thiazolidinediones, the biomarker adiponectin has been identified<sup>30</sup> because it increased in a dose-dependent manner after thiazolidinedione administration.<sup>31</sup> Years later, the Biomarkers Consortium (<https://fnih.org/what-we-do/biomarkers-consortium>), a public–private platform for precompetitive collaboration specific to biomarker research, endorsed adiponectin as a predictor of metabolic responses to PPAR agonists in patients with type 2 diabetes, and adiponectin became a putative target engagement biomarker for PPAR agonists.

### Biomarkers as Surrogate Clinical Endpoints

The FDA generally prefers a clinical endpoint for a drug trial. Such endpoints may include death, disease progression, a disease-related event such as myocardial infarction, or a clinical conversion of a predisease state to the disorder of interest such as conversion from prediabetes to type 2 diabetes. Generally, the endpoints are compared between the patients randomized to the investigational drug arm and the patients randomized to a placebo or standard-of-care control arm. However, diseases that slowly evolve from risk states to clinical endpoints pose economic and operational problems for sponsors. The situation is further exacerbated by conditions of low population prevalence. How can a sponsor design a study to follow a statistically powered sampling for an outcome of low prevalence that may take 10 or more years to develop? By the time the study is completed, the company may have ceased activities due to financial issues; other drugs may have entered the market, thereby closing the window of opportunity for the drug under study; or intellectual property rights may have lapsed. For that matter, by devoting long-term resources to a given drug, the sponsor may sacrifice opportunities to explore other drug candidates. For all the foregoing reasons, pharmaceutical companies propose biomarkers as surrogate outcomes.

CVD represents a long-established condition for application of surrogate endpoints. As an example, the statin class of lipid-lowering drugs, including pravastatin, lovastatin,

fluvastatin, atorvastatin, rosuvastatin, pitavastatin, and simvastatin, have generally demonstrated efficacy in studies in which low-density lipoprotein cholesterol (LDL-C) provides a surrogate for clinical outcomes, such as myocardial infarction, revascularization, or death due to coronary heart disease. Drugs that significantly reduce circulating LDL-C are inferred to reduce risk for the clinical outcomes of interest. However, some linkage of biomarkers such as LDL-C to actual outcomes has been observed. As an example, the Scandinavian Simvastatin Survival Study<sup>32</sup> provided evidence in this regard. Over 5.4 years of median follow-up, simvastatin produced a 35% decrease in LDL-C, which was associated with a 42% reduction in risk of coronary death and a 37% reduction in risk of revascularization procedures. Similarly, studies of anti-hyperglycemic medications of various classes, including insulins, glinides, thiazolidinediones, sulfonylureas,  $\alpha$ -glucosidase inhibitors, amylin analogs, incretin mimetics, DPP-4 inhibitors, selective sodium-glucose transporter-2 inhibitors, and metformin often use changes in HbA<sub>1c</sub> as surrogate measures of improving or worsening glycemic status.

Even with use of surrogate measures, longitudinal studies in conditions such as CVD and diabetes nevertheless require ongoing testing at multiple time points, perhaps over several years. Thus a clinical laboratory contracted to perform study-related testing will be required to maintain accurate and reproducible methods over a long period of time. In this regard, the laboratory will benefit from participation in standardization programs such as the Centers for Disease Control and Prevention (CDC) Lipid Standardization Program (LSP) or the National Glycohemoglobin Standardization Program (NGSP). The goals of these programs are generally to provide external quality assurance measures to help maintain consistency and reliability of results. For example, the NGSP seeks to standardize HbA<sub>1c</sub> test results to those of the Diabetes Control and Complications Trial<sup>33</sup> and United Kingdom Prospective Diabetes Study,<sup>34</sup> which established the direct relationships between HbA<sub>1c</sub> values and outcome risks in patients with diabetes. If such external programs do not exist for a given biomarker, the clinical laboratory should develop internal quality control measures to achieve the same objectives. The importance of quality control and quality assurance cannot be overstated in the situation in which biomarkers provide the primary evidence of efficacy in a drug trial.

### Biomarkers as Companion Diagnostics and Personalized Medicine

Personalized medicine has been the subject of much attention in the past few years, and different government organizations have attempted to define the paradigm. For instance, it was defined as “Providing the right treatment to the right patient, at the right dose at the right time” by the European Union, and as “The tailoring of medical treatment to the individual characteristics of each patient” by the US President’s Council of Advisors on Science and Technology.<sup>35</sup> Additional definitions by the Personalized Medicine Coalition (PMC), American Medical Association, and the NIH were also suggested. All of these definitions share the concept of delivering the right medical treatment to the right patient at the right time, based on the patient’s characteristics, needs, and preferences.

The FDA has long shown a strong commitment to personalized medicine, as evidenced by the publication in 2013

of a comprehensive report entitled “Paving the way for personalized medicine: FDA’s role in a new era of medical product development.”<sup>36</sup> In this report, the FDA highlights ways in which the agency has worked “to respond, anticipate and help drive scientific developments in personalized therapeutics and diagnostics” and uses Kalydeco (Vertex Pharmaceuticals, Cambridge, MA) as a prime example of personalized medicine. Kalydeco was developed for the treatment of cystic fibrosis and designed to address the underlying cause of the disease in a patient population that carries the G551D mutation in the cystic fibrosis transmembrane regulator gene. The use of a genetic test to identify the G551D mutation has streamlined and expedited the review process of the drug. A “Progress and Update Report” published in 2018 indicates that one of every three drugs the agency approved over the past 2 years is in some way personalized medicine (e.g., mentions a pharmacogenomic marker in its labeling). In 2018 alone, 25 of the 59 (44%) NMEs approved by CDER were classified by the PMC as personalized medicine drugs.<sup>37</sup> This is a significant increase from 10 years earlier, when only 10% of NMEs approved annually were considered personalized medicine agents by that organization.

Although the term *personalized medicine* has often been used to refer to *companion diagnostics*, the latter term is narrower in scope and has certain development and regulatory implications. In 2014 guidance,<sup>11</sup> the FDA defined an IVD companion diagnostic device as “an *in vitro* diagnostic device that provides information that is essential for the safe and effective use of a corresponding therapeutic product.” This guideline helped clarify the definition and approval requirements of companion diagnostics. The FDA definition carries certain regulatory requirements, such as the inclusion of the instructions of use of the IVD companion diagnostic device in the labeling of both the diagnostic device and the corresponding therapeutic product. Also, the approval of a new targeted drug and the diagnostic device should ideally occur simultaneously. This can be challenging for the agency however, because it requires coordination between different review branches of the agency: the Center for Devices and Radiological Health (CDRH), CDER, and/or Center for Biologics Evaluation and Research (CBER).

An IVD companion diagnostic device can be used to identify patients who are most likely to either benefit from a treatment or to experience adverse events. A companion diagnostic device could also be used to monitor response to treatment and to “identify patients in the population for whom the therapeutic product has been adequately studied, and found safe and effective.”<sup>11</sup> The first companion diagnostic to gain regulatory approval was HercepTest in 1998. This device consisted of an *in vitro* assay for the detection of HER2 protein expression in breast cancer cells. The assay was manufactured by DAKO (Denmark) and is used to select HER2-positive metastatic breast cancer patients who were likely to benefit from treatment with trastuzumab (Herceptin). Today, HER2 testing is a routine aspect of the clinical diagnosis of breast cancer patients for the purpose of therapy selection. There are at least 10 different commercial devices approved as companion diagnostics, produced by at least 7 different companies, and based on different technologies, such as immunohistochemistry, fluorescence or chromogenic *in situ* hybridization, and next-generation sequencing (NGS).

In the last few years, a large number of new companion diagnostic assays have been cleared or approved by the FDA, bringing the total at time of press to over three dozen. Many were approved for more than one drug and for multiple indications. This compares to just 17 approved drugs in 2015. A comprehensive and up-to-date List of Cleared or Approved Companion Diagnostic Devices (In Vitro and Imaging Tools) with helpful links to the premarket approvals and NDAs/Biologics License Applications can be found at <https://www.fda.gov/medical-devices/vitro-diagnostics/list-cleared-or-approved-companion-diagnostic-devices-vitro-and-imaging-tools>. The majority of the FDA-approved companion diagnostics to date are for oncology and are based on genetic (i.e., polymerase chain reaction or probes) or immunohistochemistry assays; one of them is based on imaging. They use either peripheral blood mononuclear cells, tissue biopsies, or circulating cell-free DNA. Furthermore, almost all currently approved companion diagnostics are for patient selection, as opposed to therapy monitoring or other uses.

A recently approved assay includes the FoundationOne CDx assay (Foundation Medical, Inc Boston, MA), which is based on NGS using DNA isolated from formalin fixed paraffin embedded (FFPE) tumor tissue specimens and detects substitution, insertion, and deletion alterations, and copy number alterations in 324 genes. It also reports microsatellite instability and tumor mutational burden. The test is intended as a companion diagnostic to identify patients with non-small cell lung carcinoma (NSCLC), melanoma, breast cancer, colorectal cancer, and ovarian cancer who may benefit from treatment with 18 different drugs. Another NGS-based companion diagnostic test is Oncomine Dx Target Test (Life Technologies Corp, Carlsbad, CA). The ability to measure multiple markers in one sample is extremely beneficial given the usually limited amount of available biopsied tissue.

Also of interest are assays recently developed as companion diagnostics for new immune-oncology drugs. The PD-L1 IHC 22C3 pharmDx, manufactured by Dako North America, Inc. (Carpinteria, CA), is a qualitative immunohistochemical assay using monoclonal mouse anti-PDL1, clone 22C3 antibody used in the detection of PD-L1 protein in FFPE tumor tissue. The read-outs of the assay are either Tumor Proportion Score (TPS) or Combined Proportion Score (CPS), depending on the tumor type. TPS is the percentage of viable tumor cells showing partial or complete membrane staining. CPS is the number of PD-L1 staining cells (tumor cells, lymphocytes, macrophages) divided by the total number of viable tumor cells, multiplied by 100. The test is used as an aid to identify patients for treatment with pembrolizumab (Keytruda) in the following indications: NSCLC, gastric or gastroesophageal junction adenocarcinoma, cervical cancer, urothelial carcinoma, head and neck squamous cell carcinoma (HNSCC), and esophageal squamous cell carcinoma (ESCC).

The Ventana PD-L1 (SP142) is manufactured by Ventana Medical Systems, Inc (Tucson, AZ) and measures PD-L1 in tumor cells and tumor-infiltrating immune cells in formalin-fixed paraffin-embedded specimens. The assay is indicated as an aid for identifying patients for treatment with atezolizumab (Tecentriq) in urothelial carcinoma and triple negative breast carcinoma. Different cut-off points may also be used in different indications for both assays.

Other important companion diagnostic assays which have been widely used in oncology include tests for BRAF and

ALK. The cobas 4800 BRAF V600E mutation assay (Roche Molecular Diagnostics), was the first companion diagnostic test to be approved under the FDA's new companion diagnostics paradigm. It is approved for the selection of patients with metastatic or unresectable melanoma for treatment with vemurafenib (Zelboraf) alone or in combination with cobimetinib (Cotellic). The THxID BRAF (bioMérieux Clinical Diagnostics, Durham, NC) is indicated for the selection of patients for treatment with dabrafenib, tafinlar, and trametinib, which are indicated for advanced or unresectable melanoma. The ALK FISH probe (Abbott Laboratories, Abbott Park, IL) is used for the selection of patients who are more likely to benefit from crizotinib (Xalkori) treatment, which is indicated for NSCLC, while the ALK IHC assay (Ventana Medical Systems) is approved as a companion diagnostic for crizotinib (Xalkori), alectinib (Alecensa), and ceritinib (Zykadia).

## Regulatory Considerations

### Animal Studies

21 Code of Federal Regulations (CFR) § 58 of the Food, Drug, and Cosmetic Act provides the statutory requirements for GLP for nonclinical (i.e., preclinical) laboratory studies. Compliance with GLP is required for research for any product regulated by the FDA, including human and animal drugs (the European Union, Japan, and other countries and regions have their own regulatory bodies and requirements that may differ from those of the United States). Furthermore, testing facilities must comply with GLP. The statutory language defines a “testing facility” as a “person” (includes a corporation or scientific establishment) who “actually conducts a nonclinical laboratory study, that is, actually uses the test article [drug] in a test system [animal].”<sup>38</sup> This language most directly applies to an organization that houses animals, doses them with test and control materials, and draws biological samples for analysis; however, the requirements can extend to clinical laboratories that analyze collected samples. For this reason, pharmaceutical companies ask laboratories to comply with the spirit and concepts of GLP, particularly with regard to personnel requirements, laboratory operational areas, specimen and data storage facilities, equipment maintenance and calibration, standard operating procedures (SOPs), and records and reports. These items may be found in subparts B, C, D, E, and J of 21 CFR § 58.<sup>38</sup>

### Human Studies

In the United States, laboratory accreditation under the Clinical Laboratory Improvements Amendments of 1988 (CLIA) is a minimum requirement for the measurement of samples used outside of the research setting—that is, in interventional drug trials. CLIA also dictates the requirements for laboratory personnel who perform the analytical work for clinical trials. Among the personnel records that should be maintained are professional certifications, résumés or curriculum vitae for all laboratory personnel, and training records. In addition, a variety of other study-related information must be retained when appropriate: relevant agreements and contracts, case report forms for study subjects, as well as all study-related worksheets and computer printouts. Compliance with ISO 15189 is required in the European Union.

The concepts of Good Clinical Practice (GCP) also apply to clinical laboratories taking part in drug studies.

The International Conference on Harmonization of Technical Requirements for Pharmaceuticals for Human Use identifies several essential documents required for testing performed in support of a drug clinical trial. Among these, clinical laboratories should generate and maintain the following documents:

1. § 8.2.11: range intervals for medical/laboratory/technical procedures and/or tests included in the protocol<sup>†</sup>
2. § 8.2.12: medical/laboratory/technical procedures/tests (certification or, accreditation or, established quality control and/or external quality assessment, or other validation where required)
3. § 8.3.6: updates to range intervals for medical/laboratory/technical procedures and/or tests included in the protocol
4. § 8.3.7: updates of medical/laboratory/technical procedures/tests
5. § 8.3.25: records of retained body fluids/tissue samples.

Clinical laboratories that are involved in these activities are also subject to inspection by the FDA. As an example, representatives of the Division of Scientific Investigations within CDER inspect clinical facilities and analytical laboratories conducting bioequivalence studies.<sup>39</sup> In general, inspections occur if (1) a facility has no inspection history, was classified “Official Action Indicated” on its last inspection, or has not been inspected within the previous 3 years; (2) a facility that performs a nonconventional bioequivalence study (e.g., a study using PD rather than PK endpoints for bioequivalence) and has not been inspected; or (3) the Office of Generic Drugs requests a directed inspection due to questions about the quality or integrity of data submitted to the FDA (e.g., missing data points, errors in calculation, or inadequate documentation).

Beyond statutory regulation and inspection through CLIA, GCP, or by the FDA, the sponsor of a clinical study may send representatives to perform quality audits before, during, and after completion of a clinical study. The intent of these audits is to evaluate record keeping and processing procedures, verify presence and conformance to laboratory SOPs, as well as to review any study testing already performed at the time of the audit. As with the regulatory agencies, the sponsor representatives will cite any deficiencies in a debriefing

meeting and subsequently work with the laboratory to implement corrective actions. In this light, laboratories should view sponsor auditing as an opportunity to improve internal processes rather than a risk of punishment.

Overall, the intent of regulatory oversight in a drug trial setting is to ensure that accurate data are captured and recorded to enable the scientific assessment of safety and effectiveness of the new drug. Laboratories substantively in compliance with statutes and their own internal procedures will enjoy long and profitable relations with pharmaceutical sponsors.

## IN VITRO DIAGNOSTIC STUDIES

### Background

IVD devices are used to gather evidence-based information from patient specimens, for clinical use by health care providers in the management of the patient. The FDA defines IVD devices as “those reagents, instruments, and systems intended for use in diagnosis of disease...intended for use in the collection, preparation, and examination of specimens taken from the human body.”<sup>40</sup> The information gathered from an IVD device is used to diagnose or otherwise manage patients to eventually cure, treat, or prevent disease.<sup>41</sup>

## IN VITRO DIAGNOSTIC STUDIES IN THE UNITED STATES

In the United States, regulation of IVD devices used in diagnostic testing spans several federal agencies, including the FDA and the Centers for Medicare and Medicaid Services (CMS). The main regulations relevant to IVDs are covered in Code of Federal Regulations 21 CFR § 809.<sup>42</sup> CMS, with the support of the CDC generally regulate all clinical diagnostic laboratory activities through the authority provided in the CLIA of 1988 and subsequent guidance documents. Thus although the FDA regulates the safety, effectiveness, design, and manufacturing of the IVD device, CMS regulates the quality of clinical laboratories and the clinical testing process, which includes the use of all IVD devices. All submissions to the FDA for IVD device approval are reviewed by either the CDRH or the CBER. The latter is used when IVDs fall under the definition of a biological product (e.g., blood donor screening tests for infectious agents).

### Classification

In the United States, IVDs are classified into three categories based on their intended use and any associated risks to the patient and public health.<sup>43</sup> Thus classes I, II, and III designate low, moderate, and high risk, respectively, with 50% of all new IVDs falling into the first category, 42% into the second, and only 8% falling into the third category.<sup>44</sup> Low-risk IVDs include immunohistochemical reagents used in conjunction with a diagnosis, and clinical chemistry-based tests for liver enzymes; immunologic tests for cardiac markers or molecular tests for a pathogenic germline genetic mutation, like factor V Leiden or cystic fibrosis, would generally be considered as a moderate-risk IVD device; and finally any IVD device associated with excluding or diagnosing infection with a high-risk virus (e.g., hepatitis B or HIV) or cancer, such as PSA, would fall under the third, high-risk category. Although

### POINTS TO REMEMBER

- Clinical laboratories may participate in both preclinical and clinical studies
- Opportunities include standard test methods and recognized references, as well as specialized and esoteric methods development and validation for pharmacokinetics, pharmacodynamics, and other evaluations
- Biomarker results may be used for critical safety and efficacy assessments, including primary objective and/or surrogate endpoints
- Laboratories are subject to regulation under Clinical Laboratory Improvements Amendments of 1988, Good Laboratory Practices, Good Clinical Practice, and other equivalent statutes and best industry practices
- Laboratories may be inspected by both US Food and Drug Administration and sponsor quality auditors before, during, and after the drug study

several low-risk devices are exempt from premarket review (i.e., the manufacturers do not need to submit an application to the regulatory agency), most moderate- to high-risk devices need a premarket review and approval.<sup>45</sup> The number of tests receiving FDA approval per year had remained surprisingly constant over the past almost three decades despite the marked improvement in discovery technology (Fig. 12.6).

There are three options for filing with the FDA, which are based on the classification of the device. If a similar or “predicate device” exists, the manufacturer will be required to demonstrate that the new device is substantially equivalent by submitting a 510(k) notification.<sup>46</sup> However, if no predicate device exists, the manufacturer must file a PMA or pre-market approval application, which requires more extensive evaluation to provide satisfactory evidence that the device is safe and effective.<sup>47</sup> In cases in which no predicate diagnostic exists, but the device can be deemed to be at lower risk, the FDA allows for a de novo submission classification.<sup>46</sup> The number of diagnostics using the de novo process is slowly increasing, among which B-type natriuretic peptide was the first to be re-classified from a class III to a class II before FDA clearance.<sup>39</sup> Costs to develop PMAs may run in the range of \$10 to \$20 million versus \$2 to \$10 million for 510(k) products, generally taking into account internal development costs, analytical and clinical validity trials, and regulatory review fees. A PMA costs considerably more to maintain due to additional manufacturing process validation work and annual reporting costs. Finally, an additional type of marketed type of IVDs are LDTs, which are developed (or adapted), validated, and offered within a single laboratory for clinical use, also commonly called “home-brew” tests,” most of which are not reviewed by the FDA.<sup>46</sup> All four IVD types are discussed separately in this section.

### Premarket Approval

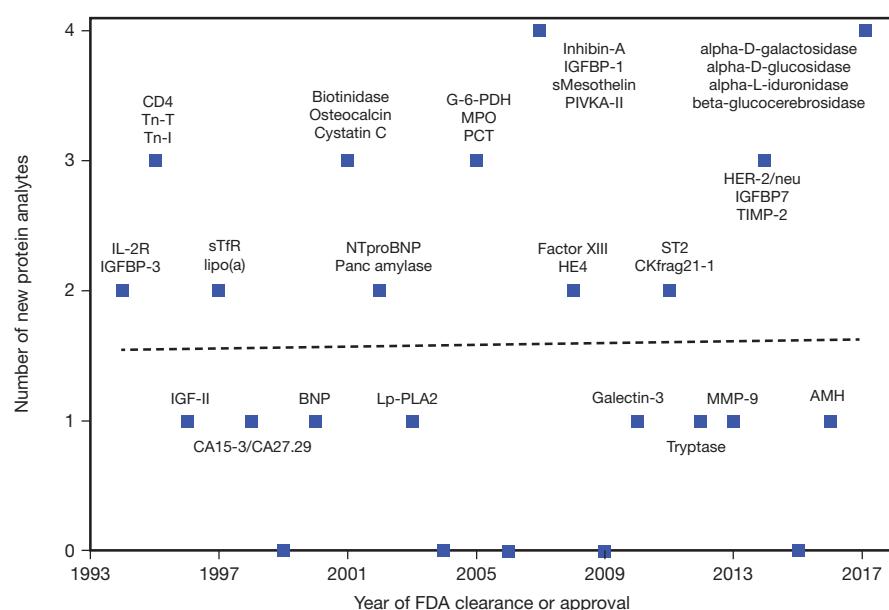
The PMA process is normally used for novel and high-risk devices. Manufacturers are recommended to apply for a PMA

if their device has the potential for an excessive patient risk of illness or injury (class III).

The full PMA application consists of protocols and information to fully explain the design and manufacturing of the device and all procedures and test results required to show its safety and efficacy. The studies to support such claims include evaluations of the analytical and clinical performances of the device. The latter supports the intended use of the device tested in subjects who fulfill the specified intended purpose of the device for its use in patient care. When the FDA or an institutional review board determines that the device presents a significant risk to research subjects, the clinical evaluation requires an investigational device exemption (IDE) before a study can be initiated, which allows the manufacturer to test an unapproved device in a clinical study for a PMA.<sup>46</sup> Examples of device studies typically requiring IDEs include those for interventional devices such as cardiac pacemakers and orthopedic implants. Most IVD studies do not require an IDE unless the diagnostic tool directly drives a patient management decision. After filing, the FDA has 180 days to review a PMA application. On average, the total number of calendar days spent by the FDA and the manufacturer reviewing and creating supplementary information is approximately 9 months for PMAs.<sup>43</sup>

### De Novo Reclassification

If the device is for a new analyte or analytical instrument with no predicate device, or if a new diagnostic claim is desired for a device which may not otherwise be considered high risk, a de novo application can be filed, requesting immediate down-classification of the device on the basis of a risk-benefit argument along with demonstrations of safety and effectiveness. In these cases, if the FDA agrees with the benefit-risk assessment that the device should not fall into Class III, the FDA reviews the device against a reference standard in a manner more similar to a 510(k) than to a PMA. As indicated earlier, the de novo reclassification approach was created for



**FIGURE 12.6** Number of protein analytes cleared or approved by the US Food and Drug Administration (FDA) for almost three decades. BNP, Brain natriuretic peptide; MPO, myeloperoxidase; PCT, procalcitonin; AMH, anti-Müllerian hormone. (Courtesy Dr. Leigh Anderson.)

novel IVD devices with no predicate device to provide flexibility around devices, which although being novel, are assessed as having lower risks, that can be mitigated with either general or special controls, thus fulfilling conditions for classification into class I or II. Under this route, the FDA allows the manufacturer to either submit a "direct" de novo 510(k) application or re-submit a petition for down-classification after a nonsubstantially equivalent or NSE letter. The review is similar to a 510(k) notification.

### 510(k)

As mentioned earlier, most new diagnostic devices are cleared by the FDA through the 510(k) process.<sup>44</sup> This requires demonstrating substantial equivalence of a new device to an existing one that is already on the market.<sup>45</sup> A 510(k) notification is required under federal law before any diagnostic manufacturer can market a new device for sale in the United States. The FDA requires this notification 90 days before intended sales, and the manufacturer must receive FDA clearance prior to marketing. Similar notification would also be required if a manufacturer changes the device's intended use or if the device has been modified with potential implications to its performance characteristics or risk profile.

Several Class I and II devices do not require premarket notification. If this is the case, it will be indicated clearly in the classification regulations. Exempt devices include analyte-specific reagents which are building blocks of complete assays, general laboratory equipment such as instrumentation, and reagents. On average, the total time spent by the FDA on reviewing a 510(k) notification is more than 3 months.<sup>43</sup>

### Laboratory-Developed Tests

Although most analyte-specific IVDs are developed and manufactured by a particular manufacturer and then distributed to multiple laboratories as commercial test kits, the LDTs follow a different paradigm because they are developed, validated, and used in a single laboratory. Under the current LDT pathway, the laboratory conducting LDTs is prohibited by CLIA from releasing any test results before establishing performance characteristics verifying the analytical validity of the LDT.<sup>46</sup> The CLIA program does not address the clinical validity, that is, the ability of the test to achieve its intended use related to a disease or clinical condition, typically described by performance characteristics such as clinical sensitivity or specificity. However, LDTs are considered tests of high complexity and thus laboratories performing them must meet applicable CLIA requirements for high-complexity testing. Until now, LDTs have especially proliferated in areas of genomic testing for cancer.<sup>48</sup> The total number of active LDTs is unknown, although a registry of more than 8000 clinical genetic tests, of which only 15 have been submitted for FDA premarket review, was recently compiled voluntarily, which suggests a much larger true number of active LDTs.<sup>49</sup>

The FDA has, until recently, exercised enforcement discretion for LDTs (i.e., selective action to regulate certain IVDs but not others). This is because the agency historically has only focused on the diagnostic test kits that are marketed broadly, while most LDTs are small, local, and no other test kit is available. Over the last decade or so, the LDT market has grown to include tests for higher-risk indications, high-volume testing

situations by reference labs, and tests that already have FDA-cleared equivalent. Due to concerns about varying testing quality, the FDA issued a draft framework for regulatory oversight of LDTs in October 2014.<sup>46</sup> However, this framework was later abandoned in November 2016 due to significant stakeholder concerns and a change in the administration. In December 2018, US lawmakers released a draft bill of the Verifying Accurate, Leading-edge IVCT Development (VALID) Act providing a regulatory framework for in vitro clinical tests (IVCTs) that applies equally to both LDTs and commercially manufactured IVDs.<sup>50</sup> It is suggested that oversight of IVCTs would continue to fall under CDRH; however, different registration and review pathways would be implemented (as opposed to 510(k)s and PMAs). This proposed regulation has undergone informal rounds of public feedback from various stakeholders prior to its formal introduction in Congress.

### Specific Studies for In Vitro Diagnostic Submission

#### The Outsourced Laboratory

It is a requirement of the FDA for the manufacturer to show that the performance of their IVD device can be reproduced in multiple laboratories, and that they undergo clinical performance testing and/or comparison to their predicate devices in the environments in which they will eventually be used. Thus certain registrational studies are predominantly outsourced, that is, conducted by an organization other than the manufacturer. These independent laboratories are used to characterize the reproducibility of the device and its performance in a real-life setting (e.g., hospital or reference laboratories) and in the target patient populations. In addition, the clinical laboratories may already be using the predicate method and have access to well-characterized specimens. This makes outsourcing of clinical evaluation of IVD devices a natural process.

It is normal for device sponsors to require evidence of professional competence from outsourced laboratories; for example, in the United States, current laboratory certifications by organizations such as the College of American Pathologists (CAP), CLIA, the Joint Commission for Hospital Accreditation, or the Commission on Office Laboratory Accreditation are required. Outside the United States, accreditation under the ISO 15189 standard is generally required.

Any outsourced laboratory needs to apply meticulous practices when working with IVD device manufacturers by maintaining not only appropriately reviewed laboratory protocols and SOPs, but also keeping records of study-specific protocols, which may later need to be submitted to the FDA by the diagnostic manufacturer. Industry sponsors generally conduct onsite audits before selecting outsourcing laboratories or before initiating a study. Additional audits or monitoring visits during and after completion of the clinical trials may also be conducted. The FDA can conduct unannounced site visits and has inspected laboratories performing studies for regulatory submissions, especially those involving a PMA. Thus in the United States, all analytical methods and procedures must be fully documented, audited, and updated as necessary to comply with all federal and laboratory procedures, including CLIA and CAP. During conduct, all aspects of a study should be fully documented with reference to technical protocols, specimen processing and management, patient records, and data processing. An important aspect of record keeping is a document control system, which indexes

all relevant protocols and records and when and what changes have been made in them.

Communication among laboratory staff and with the sponsor personnel is critical, with regularly scheduled meetings and progress reports. Laboratory personnel must be clear about the sponsor's expectations during each step of the development process. Detailed study protocols that outline specific experimental procedures, including prespecified acceptance criteria, are critical. The laboratory's designated study manager must remain in close communication with the sponsor's study manager, the clinical research associate, or other assigned liaison.

### Research Studies

The research and development process for IVDs generally follows a strategic progression after assessment of the target market and intended use. The first stage is designated discovery, that is, conceptualizing a new testing strategy. Discoveries occur as a result of experiments conducted in a research or clinical laboratory, or in the diagnostic manufacturer's own research laboratory. Recent advances in proteomics, genomics, and metabolomics provide excellent tools for identifying suitable novel diagnostic-device targets.<sup>51</sup> Associated with the discovery should be a theoretically feasible biologic reasoning or evidence suggesting that the development of a new IVD can improve diagnostic, predictive, or prognostic capabilities. Proof-of-concept studies, often together with additional confirmation studies, would follow to substantiate the new concept before making a decision to undertake the costs associated with the formal development process.

Initial studies would generally be undertaken by the diagnostic manufacturer's laboratory to demonstrate that the IVD device can be produced consistently and reliably, and with acceptable cost parameters. Pilot scale production helps to realistically determine the cost of production and eventual pricing of the device. Following initial pilot scale production and in-house performance analyses, the manufacturer would produce subsequent devices for external analytical and clinical evaluations for regulatory studies. The registrational studies for submission to the FDA and other regulatory bodies

typically must include at least three separate production lots to demonstrate lot-to-lot consistency. The design control process has been outlined by the FDA, specifying a Quality System Regulation (QSR) and requiring each diagnostic manufacturer to maintain a quality system for the design and manufacturing of the IVD device.<sup>52</sup> Each manufacturer is required to perform quality audits to ensure compliance with QSR and maintain procedures to control the design and production processes of the device. International quality system requirements must be followed for devices intended for global use; ISO 13485 is commonly used.

### Analytical Performance Studies

A full validation of the analytical performance is required using controlled performance analyses. The manufacturer may conduct these in-house, or they may work with a CRO to transfer its IVD device to an independent laboratory. Minimally, this includes assessment of result imprecision, trueness, robustness (including lot-to-lot variation), linearity, and matrix effects. Refer to Chapter 2 for additional information. In addition, analytical sensitivity, reference interval determination or verification, stability of analyte reagents and samples, and potential interferences are determined. As discussed previously, external laboratories carrying out these studies in the United States need to be certified by CLIA, accredited by CAP or an equivalent recognized body, to justify the submission of the resulting data in the application to the FDA. Usually all of these analyses are conducted in accordance with the various guidelines set forth by the Clinical and Laboratory Standards Institute (CLSI) (Table 12.2). The CLSI is a not-for-profit organization that facilitates the development of clinical laboratory testing standards after gathering input and consensus from industry, government, and health care providers ([www.clsi.org](http://www.clsi.org)). In the United States, it is recommended to incorporate the CLSI documents and procedures into the SOPs of the laboratory involved in servicing an IVD manufacturer's FDA submission. Several of these documents are referenced in the CAP accreditation checklists and serve as critical elements for regulatory submissions; they are also maintained by the FDA in a database of recognized consensus standards.<sup>53,54</sup> If the studies are

**TABLE 12.2 Clinical and Laboratory Standards Institute Method Evaluation Documents**

CLSI Code	Subject	Title
EP05-A2	Precision	Evaluation of Precision Performance of Quantitative Measurement Methods
EP06-A	Linearity	Evaluation of the Linearity of Quantitative Measurement Procedures: A Statistical Approach
EP07-A2	Interference	Interference Testing in Clinical Chemistry
EP09-A3	Method Comparison	Measurement Procedure Comparison and Bias Estimation Using Patient Samples
EP14-A2	Matrix Effects	Evaluation of Commutability of Processed Samples
EP17-A2	Sensitivity	Evaluation of Detection Capability for Clinical Laboratory Measurement Procedures
EP24-A2	Accuracy	Assessment of the Diagnostic Accuracy of Laboratory Tests Using Receiver Operating Characteristic Curves
EP25-A	Stability	Evaluation of Stability of In Vitro Diagnostic Reagents
EP26-A	Between-Lot Variation	User Evaluation of Between-Reagent Lot Variation; Approved Guideline
EP28A3	Reference Intervals	Defining, Establishing, and Verifying Reference Intervals in the Clinical Laboratory
C56AE	Sample Integrity	Hemolysis, Icterus, and Lipemia/Turbidity Indices as Indicators of Interference in Clinical Laboratory Analysis

conducted externally, it is recommended that the manufacturer and the contracted laboratory create a comprehensive agreement before initiating the studies that outlines performance characteristic expectations and the precise deliverables; for example, what subjects will be sampled (apparently healthy or specific patient populations), what matrix will be tested, what test system is required to support the dynamic range for the intended purpose in the clinics (upper and lower limit of quantification and upper limit of clinical range), what stability information has already been characterized by the manufacturer, and what exactly may be required to further support the submission.

The laboratories supporting a 510(k) notification study would usually perform analytical performance studies designated by the manufacturer (see Table 12.2) and provide any additional evidence necessary to support equivalence. The FDA summarizes this as:

1. Explanation of how the device functions and its scientific concept
2. Its intended use (i.e., what diagnoses or treatment it will support and what patient population is targeted)
3. Description of technical characteristics as they compare to the predicate device
4. Analytical performance characteristics of the device to support substantial equivalence with predicate device
5. Clinical performance characteristics (see later) that describe the subjects examined and provide a description of the safety and effectiveness, with reference to adverse effects and complications (which are rare for IVD evaluations)

Equivalence for analytical specificity and sensitivity is normally demonstrated by using clinical samples or samples to which the analyte in question has been added. The specimens used to address this question might be from previously sampled populations or newly obtained from recruited subjects meeting certain established criteria for disease condition and/or sample collection. The outsourced laboratory may be called upon to provide access to patient populations to establish reference intervals for new analytes or methods.

Analytical performance studies for the support of a PMA submission are generally identical to those for a 510(k) notification, though more details up to and including submission of raw data may be expected in the submission.

### Clinical Performance Studies

Clinical performance studies are required for PMA, certain 510(k)s, and de novo submissions. As mentioned previously, an IDE, if required, must be obtained before the clinical study is initiated to allow the unapproved IVD to be used in a clinical study for the purpose of collecting performance data for the subsequent submission. The main purpose of clinical studies is to ensure that the device is safe and effective for its intended use. The FDA describes in detail how this can be determined in 21 CFR Part 860.7.<sup>54</sup> The intended population and specific conditions for which the device will be used need to be specified, along with an assessment of the reliability and benefit compared with probable injury or harm due to the device.<sup>55</sup> The scientific evidence to substantiate such claims would optimally be generated from well-designed and valid investigations. Initial trials may involve feasibility or pilot studies that are in general exploratory and limited in scope and size. These studies provide preliminary evidence of device performance. This information is later

used for optimizing the design for subsequent and much larger trials that are intended to support specific claims of safety and efficacy. These trials can be prospective or retrospective, depending on the intended claim or prevalence of the disease.

The clinical trials report should provide appropriate evidence that the subjects are suitable for the study, meaning they represent the population of intended use of the device, that is, they qualify to be tested for the condition in question, and that each group of subjects (patients and control subjects, if applicable) are well specified and comparable. The reliability of the data in such investigations is dependent on the device used having already been standardized with regard to design and performance, which is normally carried out in preceding analytical performance studies (see earlier section). Review of data for a PMA also involves review of the manufacturing processes, with an inspection of the device manufacturing facility and an audit of any or all of the clinical trial sites used in gathering the clinical data.<sup>52</sup>

For IVD products, the safety of the device is directly related to the impact of the device's performance, and in particular, its ability to distinguish between patients with and those without the condition, to monitor the condition of interest, or to convey prognostic information such as the risk of development or progression of a particular condition. The device intended use(s) must be incorporated into the study design and analysis methods, and acceptance criteria must be prespecified in a statistical analysis plan. Analysis methods may include the following: receiver-operating characteristic (ROC) curves in which the sensitivity and specificity of the new analyte are graphed and compared with a reference marker<sup>56,57</sup>; calculation of other diagnostic accuracy parameters (e.g., positive and negative likelihood ratios, and positive and negative predictive values); univariate and multivariate logistic regression methods for discrimination of cases and controls; survival analyses relying on Kaplan-Meier methods and univariate and multivariate Cox proportional hazards regression; or other appropriate methods (please see Chapter 2 for additional information on diagnostic accuracy and predictability). For comparison of a new method with a reference or predicate, the ROC curves allow easy comparison of clinical and analytical performance among two or more devices and help to determine a test's performance; this is a standard basic expectation when investigating new diagnostic tools. However, it is important to note that the calculation of the sensitivity and specificity of a device is dependent on the spectrum of both the disease and the examined patient population. Similarly, positive and negative predictive values depend upon the prevalence of disease in the study population. Principal investigators must understand the intricacies of these statistical tools.<sup>58</sup> For further discussion on this topic, refer to Chapter 2. In general, the manufacturer must provide enough evidence to ensure that the device gives the expected results in defined patient populations.

Often, a 510(k) notification only includes data referred to as "preclinical," meaning nonclinical analytical performance studies and a comparison to the predicate, perhaps through regression analysis of values generated on both devices. However, in some cases, the FDA may require that the diagnostic manufacturer produces direct clinical performance data in

addition to the analytical performance investigation.<sup>59</sup> This is normally the case when the link between the analytical and clinical performance is ill-defined or when only a clinical outcome can serve to adequately demonstrate the intended use of the new device. These requests from the FDA have occurred in less than 10% of all 510(k) submissions<sup>60</sup> and have required the manufacturer to perform clinical studies in which the device is investigated in a representative patient population, similar to a PMA submission. However, this is rarely required in more elaborate prospective clinical trials. In contrast, the FDA may require prospective studies for CVD risk factors, such as high-sensitivity CRP and for cancer prognostic tests, although these tests are considered only moderate risk. It should also be noted, however, that the FDA recommends premeetings to discuss study design, including choice of the predicate or comparative device to satisfy the correct classification requirements (i.e., 510[k] vs. PMA). An example of such a situation is kidney injury molecule-1, a new test for diagnosing AKI, in which the predicate method, creatinine or a sum of current clinical variables used to diagnose AKI, is compared with the novel and more sensitive biomarker.<sup>61</sup> Clinical trials would be initiated with patients who are prone to kidney injury due to invasive procedures (major surgery), drugs, or other comorbidities like sepsis. The new marker would be tested against the predicate diagnostic method to examine potential improved classification of patients compared with control subjects. The same study could also examine the prognostic value of the marker in diagnosing future AKI events if the study design allows it, such as being a prospective clinical trial.<sup>62</sup> Again, the manufacturer would need to substantiate the intended use with appropriate scientific evidence and decide if the submission is for a diagnostic or prognostic application, or both. A new device would be cleared by the FDA if it exhibits similar performance characteristics to the predicate device.

### POINTS TO REMEMBER

- In the United States, in vitro diagnostics (IVDs) are classified into three categories based on their intended use and risks to the patient.
- IVDs are filed to the US Food and Drug Administration through a 510(k), de novo, or PMA pathway depending on their classification, or existence of a predicate device.
- The 510(k) submission is required to show substantial equivalence to a predicate device, whereas the PMA submission must include evidence to support that the novel device is safe and effective. A de novo submission includes an argument why the risk-benefit profiles of the device support its classification into a lower risk class, despite lack of a predicate.
- Laboratories can support IVD manufacturers with analytical performance studies and with clinical studies ranging from feasibility analyses to final clinical validation trials.

### IN VITRO DIAGNOSTIC STUDIES IN THE EUROPEAN UNION

As of May 22, 2022, IVDs in the European Union will be regulated under the In Vitro Diagnostics Regulation (IVDR) that was adopted on April 5, 2017. Under the IVDR, manufacturers must maintain a Quality System and prove that

devices demonstrate conformity to applicable requirements—indicated by a CE mark—before they can be placed into clinical use. Relative to the previous regulatory scheme under the IVD Directive 98/79 EC (IVDD), the IVDR provides a more uniform set of requirements across countries, a renewed emphasis on analytical and clinical evidence, and stronger rules for certain high-risk performance studies.

Under the IVDR, IVDs are classified into Class A, B, C, or D using classification principles based on risk to patient and public health. The registration pathway depends on class. Class A products are considered at the lowest risk and can be self-certified by the manufacturer, with no requirement for a preview. For classes B and C, representative devices will go through review by notified bodies; a notified body is an organization designated by an EU country to assess the conformity of certain products before being placed on the market and these bodies carry out tasks related to conformity assessment procedures set out in the applicable legislation, when a third party is required. The European Commission publishes a list of such notified bodies ([https://ec.europa.eu/growth/single-market/goods/building-blocks/notified-bodies\\_en](https://ec.europa.eu/growth/single-market/goods/building-blocks/notified-bodies_en)). For the highest-risk products in Class D, each device will go through an extensive assessment by both a notified body and a specially designated reference laboratory.

The IVDR requires clinical evidence to support the intended purpose of the device to be collected and analyzed throughout the lifetime of the device. A performance evaluation report must be constructed covering the three elements of clinical evidence: analytical performance, scientific validity, and clinical performance of the device.<sup>63</sup> Analytical performance refers to the ability of the device to correctly measure a particular analyte and is assessed through familiar IVD studies such as precision, limit of detection, etc. Scientific validity refers to the association of an analyte with a particular clinical condition and is often supported by published literature. Clinical performance measures such as diagnostic sensitivity and specificity or hazards ratio demonstrate the ability of the device to yield results associated with the clinical condition when used by the intended user in its specified target population. Actual clinical studies are not always required in order to demonstrate clinical performance, which can be demonstrated using scientific peer-reviewed literature, and experience gained through routine diagnostics testing (e.g., real world data from clinical laboratories).

When clinical performance studies are required for clinical evidence, they should be conducted in accordance with recognized ethical principles, designed to minimize any potential bias, and performed in circumstances similar to the intended conditions of use. Certain clinical performance studies involving risks to subjects (i.e., interventional studies) are subject to additional requirements, including prior authorization by the national authority where the study is conducted.

Relative to the IVDD, IVDR will also change the requirements for in-house tests or LDTs, which are referred to as devices manufactured and used only within health institutions. Such tests are subject to the same general safety and performance requirements as other IVDs and must be manufactured and used under a quality management system. Furthermore, the health institution must justify that its target users' needs cannot be met by a device currently available on the market.

## EPIDEMIOLOGIC STUDIES

Epidemiologic studies involve thousands, and sometimes tens of thousands, of individuals who are followed for decades, primarily to establish associations between indicators, such as biomarkers (cholesterol, blood pressure, body mass index, and so on), nutritional intake, and/or vitamin supplements and disease conditions. Such an example is the Physicians' Health Study (PHS), which was a randomized, double-blind, placebo-controlled trial of aspirin and β-carotene for the primary prevention of heart disease and cancer that was conducted among 22,071 U.S. male physicians between the ages of 40 and 84 years. Participants did not have a history of CVD or cancer, and were randomly assigned to one of four treatments: aspirin, β-carotene, both, or neither. Before randomization, baseline blood samples were obtained from participants, and questionnaires were sent annually to elicit information on risk factors and incident health events. The obtained blood samples were carefully archived and stored in small aliquots. Analyte stability signifies a profound consideration in such studies, with care taken to ensure that analyte frequency distributions in stored samples represent the distributions in fresh samples. In the PHS, plasma samples were stored at  $-80^{\circ}\text{C}$ ; however, in contemporary studies, samples are usually stored in liquid nitrogen ( $-196$  to  $-210^{\circ}\text{C}$ ) to better maintain their integrity. During the follow-up period, some participants developed chronic conditions, including diabetes, cancer, and CVD. At this point, questions such as "what biomarker measured at baseline could have identified the individual who developed the disease later in life?" could be asked. By using nested case-control designed study, in which cases were the subjects who developed the disease and control subjects were the age- and sex-matched individuals who remained free of that disease during the same follow-up period, the examination of biomarkers in baseline samples could lead to the identification of a new predictive marker of that disease. For example, lipids, lipoproteins, and apolipoproteins have been traditionally used to assess risk of future myocardial infarction. Because our understanding of atherosclerosis has evolved in the last several decades, and inflammation was shown to play a pivotal role in the inception and development of the atherothrombotic disease, the PHS cohort was used to test the hypothesis of whether an inflammatory marker could predict future myocardial infarction. This particular hypothesis was proven correct; the baseline concentration of the inflammatory marker CRP, when measured by a high-sensitivity assay, was able to predict future myocardial infarction, placing those men in the highest quartile of CRP at almost three times the risk of developing the disease compared with those in the lowest quartile.<sup>64</sup> This observation was instrumental in opening the door to a completely new area of research that enabled a better understanding of the etiology of coronary heart disease and led to the use of novel therapeutic modalities to reduce risk of coronary events. Furthermore, this finding implicated inflammation in a variety of other physiologic and pathologic conditions, including obesity, hypertension, and diabetes, and led to the inception of the "common soil hypothesis," which showed inflammation as a common culprit in these diseases.

### Why Use the Clinical Laboratory for Epidemiologic Testing?

Historically, the concentrations of biochemical markers in epidemiologic studies were measured in small research laboratories.

For example, samples for the measurement of interleukin-6, zinc, and methylmalonic acid were split into aliquots and sent to three different researchers in various institutions who were known experts in these respective areas. Although this practice was deemed acceptable in terms of the quality of the generated data, a significant portion of the samples was wasted due to the splitting of the specimens and the use of instrumentations that required large sample volumes. Epidemiologic study samples are very precious and irreplaceable; every effort must be made to preserve them and maintain their integrity.

Instrumentations in the clinical laboratory use microvolume for analysis; for example, only  $15\text{ }\mu\text{L}$  is required for the determination of a complete lipoprotein profile (total cholesterol, triglycerides, high-density lipoprotein cholesterol, and LDL-C). Special cups to minimize dead volume (approximately  $35\text{ }\mu\text{L}$ ) and evaporation are also commercially available. In addition, equipment that was once only used in research laboratories, such as liquid chromatography-tandem mass spectrometry, isoelectric focusing, atomic absorption spectrometry, and enzyme-linked immunosorbent assay, are currently available in many clinical laboratories. This represents an opportunity for clinical laboratories that possess such technologies and are used to handling samples with small volumes to provide research epidemiologists with a one-stop-shop where the concentrations of a diverse group of markers, such as cholesterol, selenium, adhesion molecules, carotenes, cotinine, and estriol, can be determined using the most sophisticated equipment, and certified medical technologists can apply rigid clinical quality control and quality assurance practices. Furthermore, clinical laboratory professionals are trained to develop and validate chromatographic methods and immunoassays in the absence of commercially available ones, thus enabling research epidemiologists to explore new frontiers. In the last decade, major advancements have been made in the fields of proteomics, metabolomics, transcriptomics, and lipidomics, particularly in terms of enhancing the sensitivity and specificity of mass spectrometers, improving sample preparation procedures, and enhancing awareness of logistical issues. Those clinical laboratories involved in such research programs can be instrumental in beginning to explore these powerful tools in answering questions in large cohorts. Clearly, such testing is not restricted to biochemical markers; clinical laboratories also provide a myriad of genetic tests that enable the association of particular mutations with specific phenotypes.

Clinical laboratorians can be active participants in the research activities by influencing the choice of examined markers and interpreting laboratory findings by delineating the limitations of these tests and potential confounders. As indicated earlier, these activities are financially and intellectually rewarding to the clinical laboratory professional. The results of these studies are usually published in prestigious journals and often lead to changes in clinical practice or public health policies.

### What Are the Logistical and Regulatory Requirements for the Clinical Laboratory?

The requirements for providing testing for epidemiologic studies are simpler than those supporting ongoing clinical trials; there is no need for a sophisticated laboratory information system, extensive staffing, or a strong adherence to rigid

FDA requirements. Testing for epidemiologic studies is done in large batches and is usually accomplished in weeks or months. Therefore it is not a strict requisite to maintain a long-term stability in assay and equipment characteristics. The results are reported to researchers in a simple format at the completion of analyses; individual subject reports or communication with various centers is not necessary.

However, the laboratory must maintain accreditation and participate in specialty programs such as the LSP and NGSP (discussed earlier), as well as proficiency testing surveys. The latter presents a particular challenge considering many of the analytes measured in these studies are in transition from research to clinical laboratories, and may not be included in proficiency testing surveys. An alternative practice would be for clinical laboratories measuring the same analytes to exchange patient samples and/or serum pools to confirm the performance quality of their assays.

Another important issue is to ensure that the integrity of the stored samples is intact and the concentration of the analyte of interest is not compromised as a result of the storage conditions. The stability can either (1) be determined from the literature, (2) be assessed by comparing the frequency distribution of the analyte in stored samples ( $n \approx 100$ ) to that of an equal number of freshly collected samples from similar population, or (3) be determined by traditional stability studies.

## PHARMACEUTICAL SUBSTUDIES

As indicated earlier, clinical trials are designed to determine the safety and efficacy of a candidate drug for particular clinical indications. Once the trial is completed and the original questions are answered, the available clinical database and collected blood samples provide a unique opportunity to explore novel hypotheses and answer new questions; these sorts of studies are called substudies. The Cholesterol and Recurrent Event (CARE) trial was designed to determine whether pravastatin decreases the incidence of recurrent myocardial infarction and coronary death in subjects with average cholesterol concentrations. Participants were randomized to either pravastatin or placebo and followed for 5 years. After completing the study and answering the original question, a substudy was designed in a nested case-control fashion to explore the interaction between pravastatin and inflammatory markers. This substudy revealed that inflammatory markers could predict recurrent coronary events, and the group with the highest risk was the group with an increased concentration of inflammatory markers, which was randomly assigned to placebo. This observation suggested that pravastatin had anti-inflammatory characteristics, and therapy with this drug might reduce risk in such subjects. This finding is perhaps more clinically and scientifically significant than the original objectives of the CARE trial.<sup>65</sup> This information was learned in a remarkably short period of time, with a minimal expense because the original trial had already been completed; only access to blood samples and the clinical database were necessary.

As a result of these findings, AstraZeneca launched the JUPITER (Justification for the Use of statins in primary Prevention: an Intervention Trial Evaluating Rosuvastatin) trial to determine whether rosuvastatin, their brand of statin, could reduce the incidence of major cardiovascular

events in individuals with a low LDL-C of less than 130 mg/dL (3.4 mmol/L) and CRP concentrations of more than 2 mg/L.<sup>66</sup> This trial demonstrated that patients who received rosuvastatin had a 44% reduction in the combined primary endpoints of myocardial infarction, stroke, arterial revascularization, hospitalization for unstable angina, or death from cardiovascular causes compared with those who received placebo. As a result, the FDA granted AstraZeneca the clinical claim of using rosuvastatin to reduce the risk of myocardial infarction, stroke, revascularization procedures in individuals with normal LDL-C, and no clinically evident coronary heart disease, but who have an increased risk based on age, CRP, and the presence of at least one additional CVD risk factor. Both the Canadian Cardiovascular Society<sup>67</sup> and the American Heart Association/American College of Cardiology guidelines now recommend taking CRP into consideration when assessing CVD risk and in targeting statin therapy.<sup>68</sup>

Pharmaceutical companies are increasingly aware of the importance and potential benefit of substudies. Therefore particular attention is paid to the appropriate storage and archival of clinical study samples.

## What Are the Requirements for the Clinical Laboratory?

In contrast to the original clinical trial, testing for substudies is done in a batch mode and in a similar fashion to that described previously for epidemiologic studies. Because the cost of substudies is minimal compared to those of the original trial, investigators and sponsors can afford to ask more daring questions to explore the prognostic usefulness of a novel marker or to examine the role of another in the initiation and development of a disease process. Therefore the laboratory must be able to measure or to add to its armamentarium of tests novel markers to enable investigators to explore new hypotheses. This type of translational research, the transfer of assays from the research laboratory bench to the routine clinical laboratory, is embraced by both the NIH and the pharmaceutical industry. The same concerns regarding the small sample volume in epidemiologic studies, as expressed previously, may apply here as well. Pharmaceutical companies, however, may request more than just accreditation from the clinical laboratory; a more extensive documentation system, comprehensive SOPs, and a particular security method for data handling and storage may be required. These processes are similar to those described previously for clinical trials testing.

## OTHER CONSIDERATIONS

### Registration of Clinical Trials

When discussing clinical trials, it is important to address, albeit briefly, the issue of registration. Comprehensive registration of clinical trials and public disclosure of study results have been instituted to alleviate concerns over selective reporting of clinical trial results.<sup>69</sup> This action came about after the New York State Attorney General filed a lawsuit against GlaxoSmithKline on June 2, 2004, alleging the withholding of information regarding the efficacy and safety of paroxetine, a selective serotonin reuptake inhibitor, in children with depression. Shortly after this, the American Medical Association's House of Delegates called on the US Department of Health

and Human Services to establish a comprehensive national registry. The greatest impetus to registering clinical trials, however, resulted from the International Committee of Medical Journal Editors (ICMJE) decision requiring all clinical trials to register as a condition for consideration for publication, starting July 1, 2005.<sup>70</sup> At the time, the ICMJE only recognized clinicaltrials.gov as an acceptable registry, which had been authorized by the FDA Modernization Act of 1997 and sponsored by the US National Library of Medicine of the NIH. In addition to clinicaltrials.gov, there are currently 17 other international primary registries that have met the World Health Organization registry criteria for content, quality and validity, accessibility, unique identification, technical capacity, and administration (<http://www.who.int/ictrp/network/primary/en>). At present, more than 320,000 clinical trials have been registered in clinicaltrials.gov alone (Fig. 12.7). Although this may be viewed as a great success story, the actual intention of registering a trial was not only to make the existence of the trial known publicly and the original design transparent but also to share the outcome of the trial when results are available so findings can be examined by patients and their physicians to see whether new treatments are safe and effective; this latter intention has not yet been fully realized. In 2017 NIH and FDA clarified the early law of 2007 and issued a “final rule” on the expectations and penalties for failing to disclose trial results. A recent article examined the reporting of over 4700 trials posted on clinicaltrials.gov 2 years after the 2017 ruling demonstrating a significant lack of compliance; less than 45% of the trials reported results on time or early, ~24% reported late, and ~32% never reported any results. Ironically, the worst offenders were academic and nonprofit organizations; the performance of federal organizations including NIH was also lackluster.<sup>71</sup> Clearly, more work is needed to fully implement the goals of this endeavor.

It is important to note that not only interventional clinical trials can be registered but also observation studies. Clinicaltrials.gov is mandated by the US Congress to be available for registration of all studies involving human subjects. As the conduct and reporting of scientific research become more transparent, clinical laboratory professionals might need to

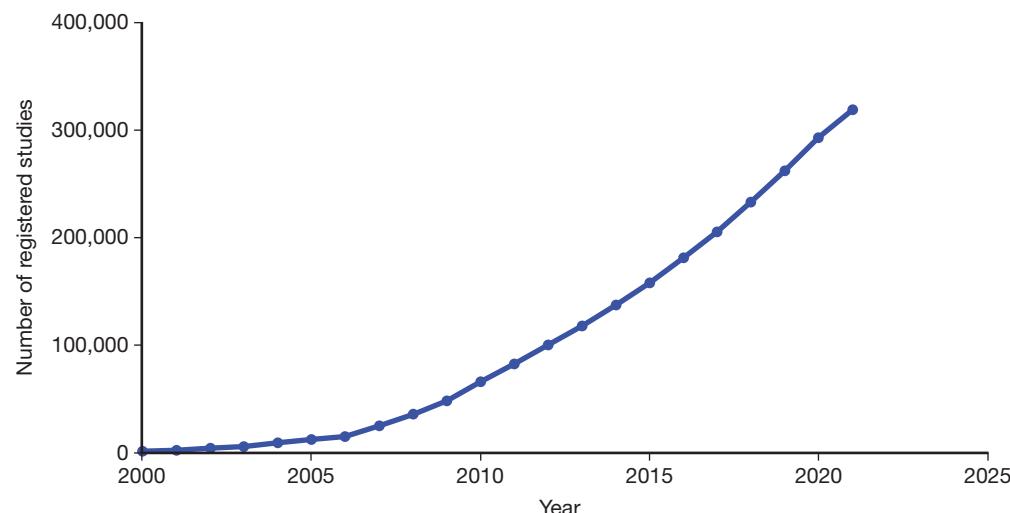
register their diagnostic and prognostic studies, which are often the subjects of pharmaceutical substudies. Such a step would be a positive development that assures the general public and funding bodies of the integrity of the process. For additional discussion on this point, refer to Chapter 1.

### Biobanking

A large and growing number of biological specimens are being stored for future biomarker research. In the United States alone, the number of stored samples was estimated to be approximately 300 million in the year 2000 and has been increasing by 20 million a year. The efforts around the collection, storage, and handling of these specimens have created a new field that is referred to as “biobanking.”

Biobanking efforts are driven by different entities and entail the collection of a variety of specimen types, such as plasma/serum, urine, tissues, and extracted DNA/RNA. Although a large number of specimens are generated from pharmaceutical clinical trials, other groups, including for-profit and government organizations such as the National Cancer Institute’s Office of Biorepositories and Biospecimen Research and the UK Biobank Initiative, are also collecting a huge number of characterized specimens.

There are many regulatory and logistical issues that need to be addressed in biobanking. Collection and future use of samples require a special approval from regulatory committees and consent from study participants. In some cases, particularly in genetic studies, a double-coding of the samples is required to ensure the protection of the patient’s identity. For additional discussion on confidentiality issues refer to Chapters 1 and 11. Good physical and information technology infrastructures and sophisticated quality management processes are needed in biobanking. Carefully prepared and comprehensive sample collection procedures are essential to minimize preanalytical variation; these processes must be implemented and enforced. The involved personnel must be properly trained and should undergo a routine assessment of their proficiency. Samples that are improperly collected should be clearly marked and preferably discarded because it may be problematic when used. Tracking of samples is also critical, and freezers should be continuously monitored and the



**FIGURE 12.7** Trend of registered clinical studies in [www.clinicaltrials.gov](http://www.clinicaltrials.gov) over a 20-year period.

records archived. Sample handling (e.g., freezing/thawing, aliquots, and so on) must also be tracked and recorded; personnel must strictly follow SOPs. Clearly, sample stability in biobanking is a major issue that has been discussed in great detail in Chapter 11. In cases in which samples from a biobank are used in the validation of an LDT during transitioning to an IVD assay, records are required by regulatory agencies. The annotation of the clinical information associated with each sample is also highly important. For a more comprehensive discussion on biobanking, refer to Chapter 11.

## SUMMARY

Clinical and research laboratories of all types will find many opportunities for involvement in clinical trials testing. Government-sponsored clinical research, pharmaceutical drug development, and development of IVD methods all require reliable laboratory testing. Clinical laboratory professionals cannot only benefit their institutions by participating in such research but also gain personal and professional benefit and satisfaction through their support of these research initiatives.

## SELECTED REFERENCES

2. Anderson NL, Ptolemy AS, Rifai N. The riddle of protein diagnostics: future bleak or bright? *Clin Chem* 2013;59:194–7.
  11. US Food and Drug Administration. In vitro companion diagnostic devices: guidance for industry and food and drug administration staff. <<http://www.fda.gov/ucm/groups/fdagov-public/@fdagov-meddev-gen/documents/document/ucm262327.pdf>>. (Accessed October 23, 2020).
  16. US Food and Drug Administration. Innovation or stagnation: challenge and opportunity on the critical path to new medical products. <https://www.fda.gov/media/77780/download>. (Accessed October 23, 2020)
  18. US Food and Drug Administration. Guideline for industry: bioanalytical method validation. <<http://www.Fda.Gov/downloads/drugs/guidances/ucm070107.Pdf>>. (Accessed October 23, 2020).
  23. Muller PY, Dieterle F. Tissue-specific, non-invasive toxicity biomarkers: translation from preclinical safety assessment to clinical safety monitoring. *Exp Opin Drug Metab Toxicol* 2009;5:1023–38.
  25. Olson SRS, Giffin R. Accelerating the development of biomarkers for drug safety: Workshop summary. Washington, DC: The National Academies Press; 2009.
  36. Paving the way for personalized medicine. FDA's role in a new era of medical product development. <https://www.fda.gov/downloads/ScienceResearch/SpecialTopics/PrecisionMedicine/UCM372421.pdf>. (Accessed October 23, 2020).
  41. US Department of Health and Human Services. Update on emerging genetic tests currently available for clinical use in common cancers. Evidence report/technology assessment no. Gend0511. <<http://www.cms.gov/Medicare/Coverage/DeterminationProcess/Downloads/id92TA.pdf>>. (Accessed October 23, 2020).
  44. US Food and Drug Administration. FDA/CDRH public meeting on oversight of laboratory developed tests (LDTs). <https://www.fda.gov/medical-devices/vitro-diagnostics/laboratory-developed-tests> (Accessed October 23, 2020).
  50. <https://www.natlawreview.com/article/device-modernization-series-vitro-clinical-tests>. (Accessed October 23, 2020).
  59. US Food and Drug Administration. IVD regulatory assistance—overview of IVD regulation. <<http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/IVDRegulatoryAssistance/ucm123682.htm>>. (Accessed October 23, 2020).
  60. US Food and Drug Administration. The 510(k) program: evaluating substantial equivalence in premarket notifications [510(k)]. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/510k-program-evaluating-substantial-equivalence-premarket-notifications-510k>. (Accessed October 23, 2020).
  63. 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. <http://www.ce-mark.com/IVD%20Regulation.pdf>. (Accessed October 23, 2020).
- [>.](http://www.Fda.Gov/downloads/drugs/guidances/ucm070107.Pdf) (Accessed October 23, 2020).

## REFERENCES

1. Carlson B. The \$69 billion-dollar market for in vitro diagnostics. <http://bioinfoinc.com>. (Accessed October 23, 2020).
2. Anderson NL, Ptolemy AS, Rifai N. The riddle of protein diagnostics: future bleak or bright? *Clin Chem* 2013;59:194–7.
3. Fee R. The cost of clinical trials. *Drug Discovery Develop Mag* 2007;10:32.
4. May M. Clinical trial costs go under the microscope. *Nature Med News* March 6, 2019.
5. Woo M. Trial by artificial intelligence. *Nature* 2019;573:S100–2.
6. US Food and Drug Administration. Code of federal regulations title 21. 21 cfr § 312.21. <<http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?fr=312.21>>. (Accessed October 23, 2020).
7. US Food and Drug Administration. Code of federal regulations title 21. 21 cfr § 312.23. <<http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/cfrsearch.cfm?fr=312.23>>. (Accessed October 23, 2020).
8. US Food and Drug Administration. Food and Drug Administration Amendments Act (FDAAA) of 2007. <https://www.fda.gov/regulatory-information/selected-amendments-fdc-act/food-and-drug-administration-amendments-act-fdaaa-2007>. (Accessed October 23, 2020).
9. US Food and Drug Administration. Code of federal regulations title 21. 21 cfr § 314.510. <<http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/cfrsearch.cfm?fr=314.510>>. (Accessed October 23, 2020).
10. US Food and Drug Administration. Code of federal regulations title 21. 21 cfr § 601.41. <<http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?CFRPart=601&showFR=1&subpartNode=21:7.0.1.1.2.5>>. (Accessed October 23, 2020).
11. US Food and Drug Administration. In vitro companion diagnostic devices: guidance for industry and food and drug administration staff. <<http://www.fda.gov/ucm/groups/fdagov-public/@fdagov-meddev-gen/documents/document/ucm262327.pdf>>. (Accessed October 23, 2020).
12. US Food and Drug Administration. Code of federal regulations title 21. 21 cfr § 314.55(b). <<http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?fr=314.55>>. (Accessed October 23, 2020).
13. US Food and Drug Administration. Code of federal regulations title 21. 21 cfr § 601.27(b). <<http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?fr=601.27>>. (Accessed October 23, 2020).
14. US Food and Drug Administration. Code of federal regulations title 21. 21 cfr § 314.610(b)(1). <<http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?fr=314.610>>. (Accessed October 23, 2020).
15. US Food and Drug Administration. Code of federal regulations title 21. 21 cfr § 601.91(b)(1). <<http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?CFRPart=601&showFR=1&subpartNode=21:7.0.1.1.2.8>>. (Accessed October 23, 2020).
16. US Food and Drug Administration. Innovation or stagnation: challenge and opportunity on the critical path to new medical products. <https://www.fda.gov/media/77780/download>. (Accessed October 23, 2020).
17. US Food and Drug Administration. Guidance for industry and FDA staff: qualification process for drug development tools. <https://www.fda.gov/Drugs/GuidanceComplianceRegulatory-Information/Guidances/default.htm>. (Accessed October 23, 2020).
18. US Food and Drug Administration. Guideline for industry: bioanalytical method validation. <<http://www.Fda.Gov/downloads/drugs/guidances/ucm070107.Pdf>>. (Accessed October 23, 2020).
19. Olson H, Betton G, Robinson D, et al. Concordance of the toxicity of pharmaceuticals in humans and in animals. *Regul Toxicol Pharmacol* 2000;32:56–67.
20. Sistare FD, DeGeorge JJ. Preclinical predictors of clinical safety: opportunities for improvement. *Clin Pharmacol Therap* 2007;82:210–14.
21. Shanks N, Greek R, Greek J. Are animal models predictive for humans? *Philos Ethics Humanit Med* 2009;4:2.
22. US Food and Drug Administration. Predictive Safety Testing Consortium. <https://c-path.org/programs/pst/> (Accessed October 23, 2020).
23. Muller PY, Dieterle F. Tissue-specific, non-invasive toxicity biomarkers: translation from preclinical safety assessment to clinical safety monitoring. *Exp Opin Drug Metab Toxicol* 2009;5:1023–38.
24. Philips BJ, Lane K, Dixon J, et al. The effects of acute renal failure on drug metabolism. *Exp Opin Drug Metab Toxicol* 2014;10:11–23.
25. Olson SRS, Giffin R. Accelerating the development of biomarkers for drug safety: Workshop summary. Washington, DC: The National Academies Press; 2009.
26. Lipshultz SE, Rifai N, Sallan SE, et al. Predictive value of cardiac troponin T in pediatric patients at risk for myocardial injury. *Circulation* 1997;96:2641–8.
27. Onitilo AA, Engel JM, Stankowski RV, et al. High-sensitivity C-reactive protein (hs-CRP) as a biomarker for trastuzumab-induced cardiotoxicity in HER2-positive early-stage breast cancer: a pilot study. *Breast Cancer Res Treat* 2012;134:291–8.
28. Kim D, Wang L, Beconi M, et al. 2r)-4-oxo-4-[3-(trifluoromethyl)-5,6-dihydro[1,2,4]triazolo[4,3-a]pyrazin-7(8h)-yl]-1-(2,4,5-trifluorophenyl)butan-2-amine: a potent, orally active dipeptidyl peptidase IV inhibitor for the treatment of type 2 diabetes. *J Med Chem* 2005;48:141–51.
29. Herman GA, Bergman A, Stevens C, et al. Effect of single oral doses of sitagliptin, a dipeptidyl peptidase-4 inhibitor, on incretin and plasma glucose levels after an oral glucose tolerance test in patients with type 2 diabetes. *J Clin Endocrinol Metabol* 2006;91:4612–19.
30. Combs TP, Wagner JA, Berger J, et al. Induction of adipocyte complement-related protein of 30 kilodaltons by ppar gamma agonists: a potential mechanism of insulin sensitization. *Endocrinology* 2002;143:998–1007.
31. Wagner JA. Early clinical development of pharmaceuticals for type 2 diabetes mellitus: from preclinical models to human investigation. *J Clin Endocrinol Metab* 2002;87:5362–6.
32. Kjekshus J, Pedersen TR. Reducing the risk of coronary events: evidence from the Scandinavian Simvastatin Survival Study (4s). *Am J Cardiol* 1995;76:64C–8C.
33. The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. The Diabetes Control and Complications Trial Research Group. *N Engl J Med* 1993;329:977–86.
34. Stratton IM, Adler AI, Neil HA, et al. Association of glycaemia with macrovascular and microvascular complications of type 2 diabetes (UKPDS 35): prospective observational study. *BMJ* 2000;321:405–12.

35. President's council of advisors on science and technology. Priorities for personalized medicine, vol. Washington, DC: U.S. Department of Health and Human Services; 2014.
36. Paving the way for personalized medicine. FDA's role in a new era of medical product development. <https://www.fda.gov/downloads/ScienceResearch/SpecialTopics/PrecisionMedicine/UCM372421.pdf>. (Accessed October 23, 2020).
37. Personalized medicine at the FDA: a progress & outlook report. [http://www.personalizedmedicinecoalition.org/Userfiles/PMC-Corporate/file/PM\\_at\\_FDA\\_A\\_Progress\\_and\\_Outlook\\_Report.pdf](http://www.personalizedmedicinecoalition.org/Userfiles/PMC-Corporate/file/PM_at_FDA_A_Progress_and_Outlook_Report.pdf). (Accessed October 23, 2020).
38. US Food and Drug Administration. Code of federal regulations title 21. 21 cfr § 58, et seq. <<http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?CFRPart=58>>. (Accessed October 23, 2020).
39. US Food and Drug Administration. Manual of Policies and Procedures (CDER). <https://www.fda.gov/about-fda/center-drug-evaluation-and-research-cder/cder-manual-policies-procedures-mapp>. (Accessed October 23, 2020).
40. US Food and Drug Administration. Code of federal regulations title 21. 21 cfr § 809.3. <<http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?FR=809>>. (Accessed October 23, 2020).
41. US Department of Health and Human Services. Update on emerging genetic tests currently available for clinical use in common cancers. Evidence report/technology assessment no. Gend0511. <<http://www.cms.gov/Medicare/Coverage/DeterminationProcess/Downloads/id92TA.pdf>>. (Accessed October 23, 2020).
42. US Food and Drug Administration. Code of federal regulations title 21. 21 cfr § 809 C. <<http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?CFRPart=809>>. (Accessed October 23, 2020).
43. Mansfield E, O'Leary TJ, Gutman SI. Food and drug administration regulation of in vitro diagnostic devices. *J Mol Diagnos* 2005;7:2–7.
44. US Food and Drug Administration. FDA/CDRH public meeting on oversight of laboratory developed tests (LDTs). <https://www.fda.gov/medical-devices/vitro-diagnostics/laboratory-developed-tests> (Accessed October 23, 2020).
45. Rome BN, Kramer DB, Kesselheim AS. Approval of high-risk medical devices in the US: implications for clinical cardiology. *Curr Cardiol Rep* 2014;16:489.
46. Testimony of AdvaMedDX Executive Director Andrew Fish, in U.S. Congress, House Committee on Energy and Commerce, Subcommittee on Health, 21st Century Cures: Examining the regulation of laboratory developed tests, hearings, 113th Cong., 2nd sess. 2014:8.
47. Jarow JP, Baxley JH. Medical devices: US medical device regulation. *Urol Oncol* 2015;33:128–32.
48. Hornberger J, Doberne J, Chien R. Laboratory-developed test-synframe: an approach for assessing laboratory-developed tests synthesized from prior appraisal frameworks. *Gen Test Mol Biomark* 2012;16:605–14.
49. Testimony of the American Clinical Laboratory Association (Acla) President Alan Mertz, in U.S. Congress, House Committee on Energy and Commerce, Subcommittee on Health, 21st Century Cures: Examining the regulation of laboratory-developed tests. <<http://docs.house.gov/meetings/IF/IF14/20140909/102625/HHRG-113-IF14-Wstate-MertzA-20140909.pdf>>. (Accessed October 23, 2020).
50. <https://www.natlawreview.com/article/device-modernization-series-vitro-clinical-tests>. (Accessed October 23, 2020).
51. Heffner KM, Hizal DB, Kumar A, et al. Exploiting the proteomics revolution in biotechnology: from disease and antibody targets to optimizing bioprocess development. *Curr Opin Biotechnol* 2014;30:80–6.
52. US Food and Drug Administration. Code of federal regulations title 21. 21 cfr § 820. <<http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?CFRPart=820>>. (Accessed October 23, 2020).
53. College of American Pathologists. Accreditation checklists. <[http://www.cap.org/web/home/lab/accreditation?\\_adf.ctrl-state=1bgoqgw3m\\_4&\\_afrLoop=489632211342544](http://www.cap.org/web/home/lab/accreditation?_adf.ctrl-state=1bgoqgw3m_4&_afrLoop=489632211342544)>. (Accessed October 23, 2020).
54. US Food and Drug Administration. Code of federal regulations title 21. 21 cfr § 860.7. <<http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?FR=860.7>>. (Accessed October 23, 2020).
55. US Food and Drug Administration. Premarket approval (PMA) – clinical studies. <<http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/HowtoMarketYourDevice/PremarketSubmissions/PremarketApprovalPMA/ucm050419.htm>>. (Accessed October 23, 2020).
56. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin Chem* 1993;39:561–77.
57. Zhou X, Obuchowski NA, McClish D. Statistical methods in diagnostic medicine. 2nd ed. New York: Wiley and Sons; 2002.
58. Pepe M. The statistical evaluation of medical tests for classification and prediction. New York: Oxford University Press; 2004.
59. US Food and Drug Administration. IVD regulatory assistance—overview of IVD regulation. <<http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/IVDRegulatoryAssistance/ucm123682.htm>>. (Accessed October 23, 2020).
60. US Food and Drug Administration. The 510(k) program: evaluating substantial equivalence in premarket notifications [510(k)]. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/510k-program-evaluating-substantial-equivalence-premarket-notifications-510k>. (Accessed October 23, 2020).
61. Dieterle F, Marrer E, Suzuki E, et al. Monitoring kidney safety in drug development: emerging technologies and their implications. *Curr Opin Drug Discovery Dev* 2008;11:60–71.
62. Bihorac A, Chawla LS, Shaw AD, et al. Validation of cell-cycle arrest biomarkers for acute kidney injury using clinical adjudication. *Am J Respir Crit Care Med* 2014;189:932–9.
63. 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. <http://www.ce-mark.com/IVD%20Regulation.pdf>. (Accessed October 23, 2020).
64. Ridker PM, Cushman M, Stampfer MJ, et al. Inflammation, aspirin, and the risk of cardiovascular disease in apparently healthy men. *N Engl J Med* 1997;336:973–9.
65. Ridker PM, Rifai N, Pfeffer MA, et al. Inflammation, pravastatin, and the risk of coronary events after myocardial infarction in patients with average cholesterol levels. *Cholesterol And Recurrent Events (care) investigators. Circulation* 1998;98:839–44.

66. Ridker PM, Danielson E, Fonseca FA, et al. Rosuvastatin to prevent vascular events in men and women with elevated C-reactive protein. *N Engl J Med* 2008;359:2195–207.
67. Genest J, McPherson R, Frohlich J, et al. 2009 Canadian Cardiovascular Society/Canadian Guidelines for the diagnosis and treatment of dyslipidemia and prevention of cardiovascular disease in the adult—2009 recommendations. *Can J Cardiol* 2009;25:567–79.
68. Goff DC Jr, Lloyd-Jones DM, Bennett G, 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. *J Am Coll Cardiol* 2014;63:2935–59.
69. Rifai N, Altman DG, Bossuyt PM. Reporting bias in diagnostic and prognostic studies: time for action. *Clin Chem* 2008;54:1101–3.
70. De Angelis C, Drazen JM, Frizelle FA, et al. Clinical trial registration: a statement from the International Committee of Medical Journal Editors. *N Engl J Med* 2004;351:1250–1.
71. Piller C. Transparency of trial. *Science* 2020;367:240–3.

**MULTIPLE CHOICE QUESTIONS**

1. What is the objective of the Phase 3 study?
  - a. Postmarketing surveillance
  - b. Safety and efficacy
  - c. Pharmacodynamics
  - d. Determination of the mechanism of action
2. In the US, methods validation for biomarkers in pharmaceutical research should follow the principles described in?
  - a. Good Clinical Practice
  - b. Good Laboratory Practices
  - c. Internal lab standard operating procedures
  - d. The U.S. Food and Drug Administration Guidance for Industry
3. Biomarkers are used in drug development for which of the following reasons?
  - a. To understand pharmacokinetics/pharmacodynamics relationships
  - b. To assess safety
  - c. As surrogate endpoints and companion diagnostics
  - d. All of the above
4. Most US in vitro companion diagnostic devices must obtain regulatory approval from which of the following?
  - a. Centers for Medicare and Medicaid Services
  - b. Center for Devices and Radiological Health (CDRH)
  - c. Center for Drug Evaluation and Research (CDER)
  - d. Center for Biologics Evaluation and Research (CBER).
5. What is a typical example of a class III in vitro diagnostic device?
  - a. Immunohistochemical reagent
  - b. Factor V Leiden test
  - c. Pap smear
  - d. Sodium
6. What is a standard statistical method for comparing diagnostic accuracy of a new in vitro diagnostic device with a predicate method?
  - a. Kaplan-Meier curve analyses
  - b. Cox proportional hazard regression analyses
  - c. Logistical regression analyses
  - d. Receiver operating characteristic curve analyses
7. In the United States, which guidelines should the laboratory use to support general analytical performance studies for an in vitro diagnostic device?
  - a. Internal standard operating procedures
  - b. Clinical and Laboratory Standards Institute guidelines
  - c. U.S. Food and Drug Administration guidance documents
  - d. College of American Pathologists checklists
8. The majority of companion diagnostics approved by the US Food and Drug Administration are for drugs used in:
  - a. Cardiovascular diseases
  - b. Oncology
  - c. Neurologic diseases
  - d. Infectious diseases
9. Which of the following statements is true about the regulation of in vitro diagnostics (IVDs) in the EU under the In Vitro Diagnostics Regulation?
  - a. Clinical evidence for IVDs must be collected and analyzed throughout the lifetime of the device
  - b. Notified body review is required for all IVDs
  - c. Every IVD will require a clinical performance study to be conducted in a laboratory
  - d. There are no requirements for IVDs manufactured and used only within health institutions (also known as in-house tests or laboratory-developed tests)
10. What is the purpose of an audit conducted by the sponsor (drug-developer) of a laboratory performing human sample testing?
  - a. To ensure the laboratory is Good Laboratory Practices certified
  - b. To ensure the laboratory has an appropriate documentation practice for the study undertaken and that it meets the intended purpose of validation and testing of samples
  - c. To make sure the laboratory is certified by the American College of Pathologists
  - d. To ensure the laboratory has the biomarker of interest validated with an associated standard operating procedure

# Machine Learning and Big Data in Laboratory Medicine\*

*Stephen R. Master, Randall K. Julian, Jr., and Shannon Haymond*

## ABSTRACT

### Background

The large number of test results generated by clinical laboratories has led to challenges in data management and analytics. Because of the potential diagnostic value of examining these results in aggregate, it is important to utilize emerging tools for the analysis of high-dimensional data. Machine learning uses a variety of computational algorithms to analyze complex datasets and make robust predictions.

### Content

This chapter discusses the varied definitions of *big data* and their application to laboratory medicine. It also presents workflows, concepts, common algorithms, infrastructure, and applications related to the use of machine learning in the clinical laboratory. The chapter is a more technical and extensive version of one previously authored on these topics.<sup>1</sup>

Because each biomarker measured in a patient can be plotted as a single number, the collection of biomarkers measured in this patient can be represented in a high-dimensional space. Unsupervised learning methods are used to find patterns in this high-dimensional space, and supervised learning methods use known outcomes from a set of subjects to develop a model to predict the outcome in a new, unknown subject. A variety of different algorithms, each with different advantages and disadvantages, has been used for these machine learning tasks. Implementing machine learning in the laboratory requires not only understanding the basic algorithmic concepts, but also deploying an appropriate computational infrastructure. Machine learning has been successfully deployed in laboratory medicine settings using a variety of underlying datasets, including traditional laboratory values, next-generation sequencing data, and images.

\*The full version of this chapter is available electronically on [ExpertConsult.com](http://ExpertConsult.com).

## POINTS TO REMEMBER

- Laboratory data associated with an individual patient can be considered a point in a high-dimensional space
- Developing machine learning models involves a process of data collection, training, and testing
- Machine learning methods can be used to address both classification and regression problems
- Unsupervised learning methods look for inherent structure within high-dimensional data
- Supervised learning methods train a classifier based on data associated with a known target outcome

## INTRODUCTION

The modern clinical laboratory routinely generates large amounts of patient data, and this rise in volume has been driven by both increases in the number of processed samples and by technological developments in high-throughput analyzers and in highly multiplexed assays. In addition to data sources within the laboratory, increasing amounts of data from other diagnostic modalities (such as noninvasive imaging) and clinical phenotype data have become readily accessible through the electronic medical record (EMR). As a result, laboratorians have more comprehensive information than ever before on a larger group of cases, and the increase in data has led to an increasing reliance on computational tools to understand and interpret the data.

The traditional laboratory has made great progress in handling the operational aspects of high-throughput laboratory medicine, including monitoring patient samples and identities, running validated assays and providing guidance on their use, monitoring for factors that may change the validity of results, and delivering results to the treating clinician in a timely fashion. However, the increasing volume of data often means that there is insufficient real-time analysis and diagnostic integration of these complex results. Further, despite the increasing amount of data potentially available for any given patient from the clinical laboratory, the task of integrating multivariate results often falls to the treating clinician.

The advent of computational approaches, such as machine learning, provides an opportunity for the clinical laboratory to improve the nature of the diagnostic information provided for patient care. In this chapter, we will discuss the nature of large data sources within the clinical laboratory and provide an overview of the tools of machine learning along with their use.

### Univariate versus Multivariate Results

It is important to consider diagnostic implications in order to understand the potential value of integrating multiple laboratory results (or, more generally, multiple biomarkers) into a single result.

A single biomarker that completely separates two diagnostic populations (say, presence vs. absence of disease) can achieve perfect sensitivity and specificity (Fig. 13.1A). However, once the values of this biomarker overlap between diagnostic categories (Fig. 13.1B), the sensitivity and specificity drop, and the area under the curve (AUC) decreases to less than 1. If two single biomarkers have similar

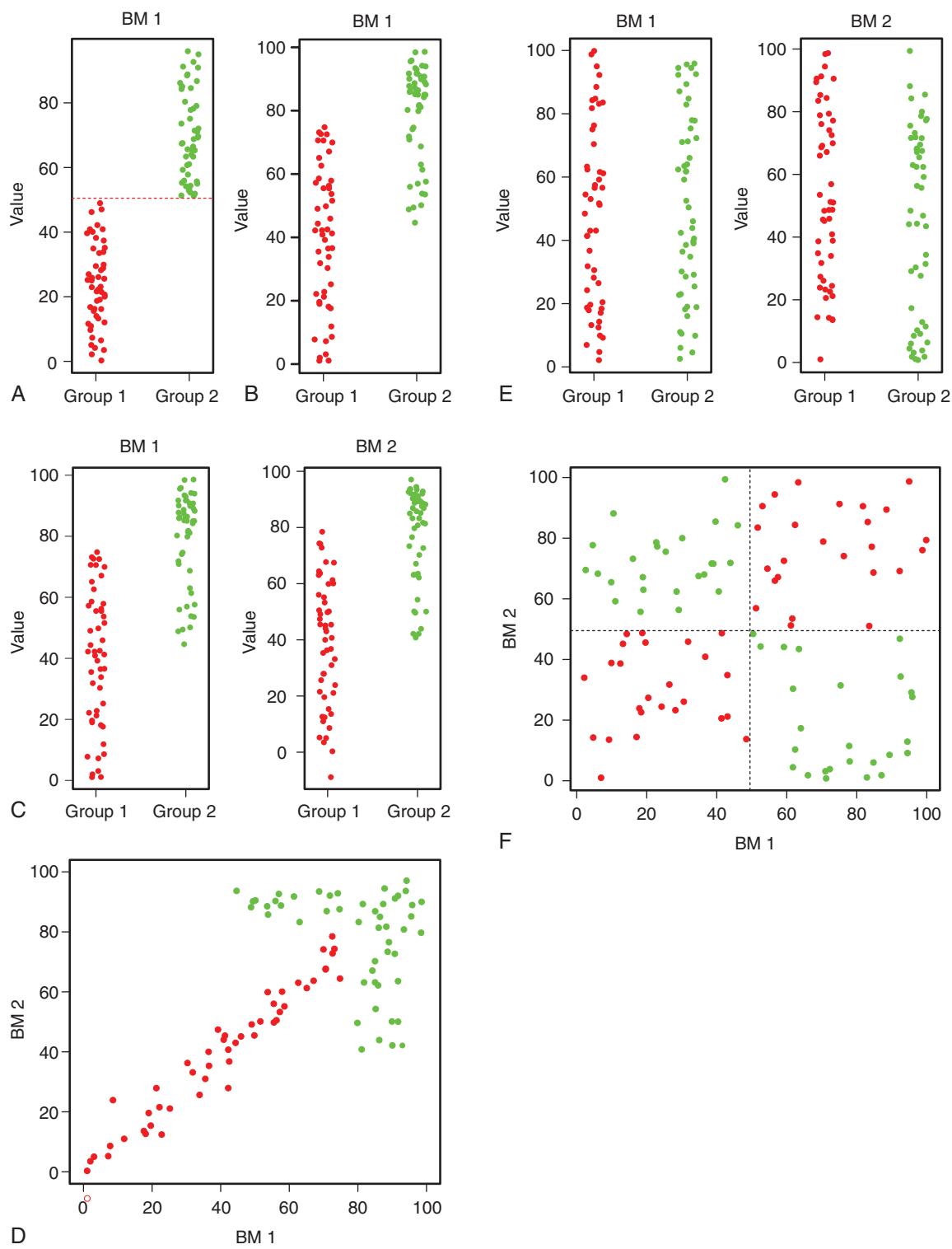
overlaps between categories (Fig. 13.1C), then it would appear that neither one adds significantly to the diagnostic value compared with the other. However, it may be the case that when considered in two dimensions, these biomarkers are individually similar in performance, can provide an aggregate diagnosis that fully separates presence versus absence of disease (Fig. 13.1D). Further, even biomarkers that appear completely uninformative on their own (Fig. 13.1E) may be informative when assessed in a multidimensional way (Fig. 13.1F).

This multidimensional view may be extended with increasing numbers of biomarkers. As previously shown, a single biomarker may be graphed in one dimension, and two biomarkers can be graphed in two dimensions ( $x$ - and  $y$ -axis). More generally, any number of biomarkers can be represented in a high-dimensional space, where the number of dimensions is equal to the number of biomarkers. This is adequate for visualization when the number of dimensions is three or fewer, but for higher numbers of dimensions it is a challenge to understand the relationship between cases that are represented this way. In a subsequent section, we will discuss computational techniques such as dimensional reduction and clustering that can address this problem. For the moment, it is important to understand the reason that large numbers of biomarkers imply a high-dimensional space and to recognize that the lack of human intuition about high-dimensional data leads to a need for computer-based tools.

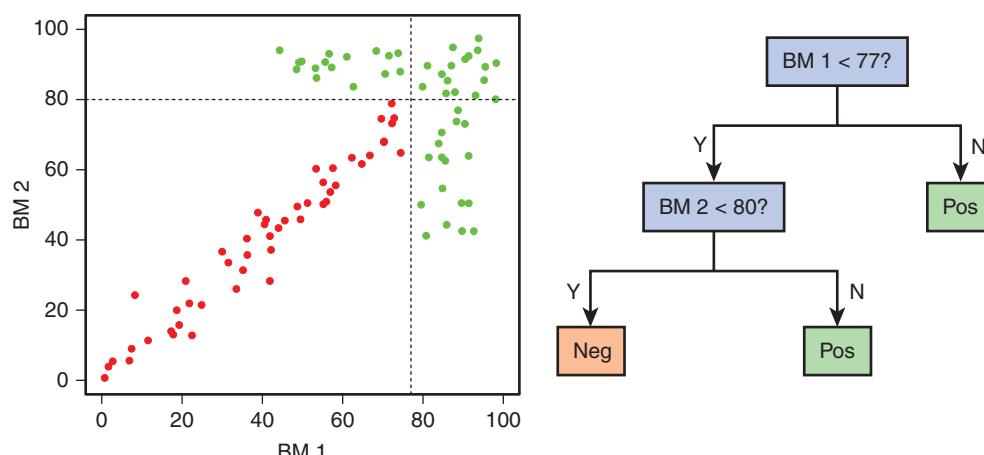
### Machine Learning

Because of the complex nature of high-dimensional data, it is often necessary to utilize various computational algorithms to understand the diagnostic information. *Machine learning* refers to the processes that are used to take a set of training cases (typically including the known outcome that is to be predicted), create a computational classifier, and use that classifier to accurately predict the outcome of a new case or set of cases that were previously unknown (note that throughout this chapter we will use the term “cases”; this could refer to patient “samples,” and it is typical in machine learning to refer to a case as an “observation”). As a simple example, the two-dimensional data that we previously described (Fig. 13.1C and D) can be encoded using a decision tree (Fig. 13.2), and a new case can accurately be classified. In this instance, machine learning (a term that is sometimes interchangeable with artificial intelligence) takes the observed training data from Fig. 13.1D, creates the appropriate decision tree, and applies it to classify new cases. Later in the chapter, we will discuss a variety of different machine learning algorithms, along with their various strengths and weaknesses.

There are two basic types of problems addressed by machine learning: classification and regression. Classification problems take the high-dimensional input data set and use it to predict one of a discrete set of outputs. Typically, this is a binary decision such as disease versus no disease, although it is equally possible to have multiple classes. Regression, in contrast, takes the set of input variables and predicts a continuous output. For example, a set of input laboratory values that predicted a patient’s tumor size or length of hospital stay would both use regression strategies. The relationship between regression and classification will be further discussed later in the chapter.

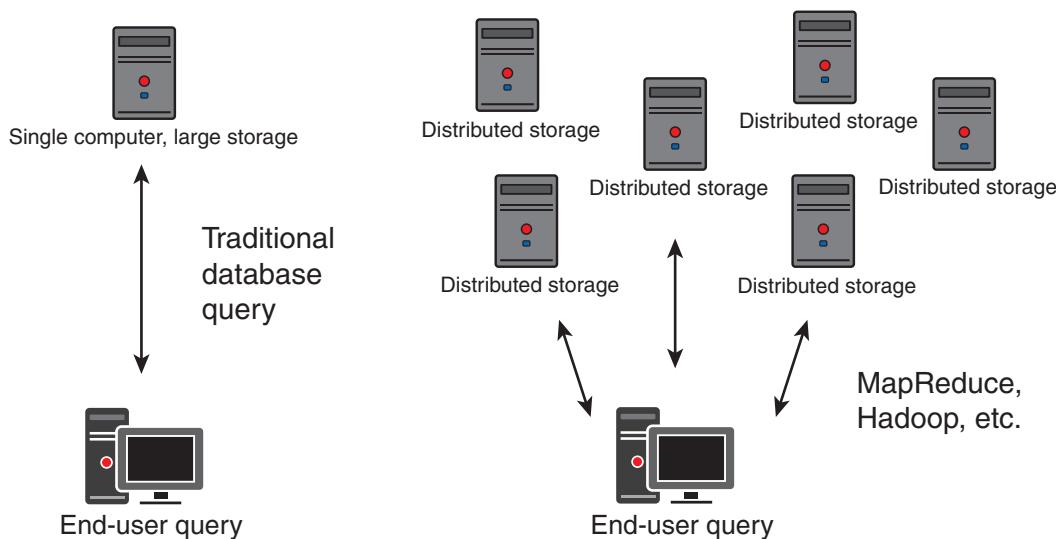


**FIGURE 13.1** Univariate versus multivariate biomarkers. Green indicates disease, and red indicates no disease. **A**, Fully informative univariate biomarker. **B**, Partially informative univariate biomarker. **C**, Two partially informative univariate biomarkers. **D**, Biomarkers from (C), plotted in two dimensions. **E**, Two uninformative univariate biomarkers. **F**, Biomarkers from (E), showing that in two dimensions a combination of these biomarkers fully separates two classes. (Modified from Haymond S, Julian RK, Gill EL, Master SR. Machine learning and big data in pediatric laboratory medicine. In: Dietzen DJ, Wong ECC, Bennett MJ, Haymond S, editors. *Biochemical and molecular basis of pediatric disease*. 5th ed. Cambridge, MA: Academic Press; 2021.)



**FIGURE 13.2** A composite biomarker in two dimensions that fully separates disease (green) from no disease (red). A decision tree that classifies cases based on biomarker 1 and 2 is shown. (Modified from Haymond S, Julian RK, Gill EL, Master SR. Machine learning and big data in pediatric laboratory medicine. In Dietzen DJ, Wong ECC, Bennett MJ, Haymond S, editors. *Biochemical and molecular basis of pediatric disease*. 5th ed. Cambridge, MA: Academic Press; 2021.

## Traditional Database      Distributed, “Big Data” Model



**FIGURE 13.3** Traditional database models versus distributed database models in the technology-driven definition of *big data*.

### Big Data

Discussions of machine learning and its application often occur in conjunction with references to *big data*. At its simplest level, *big data* simply refers to the large datasets, such as medical datasets, that we described in the introduction and will further elucidate in subsequent sections. However, there are subtle differences in the way that the term *big data* has been used, and these can be broken into at least three categories: a technology-driven definition, a property-driven definition, and an analysis-driven definition.

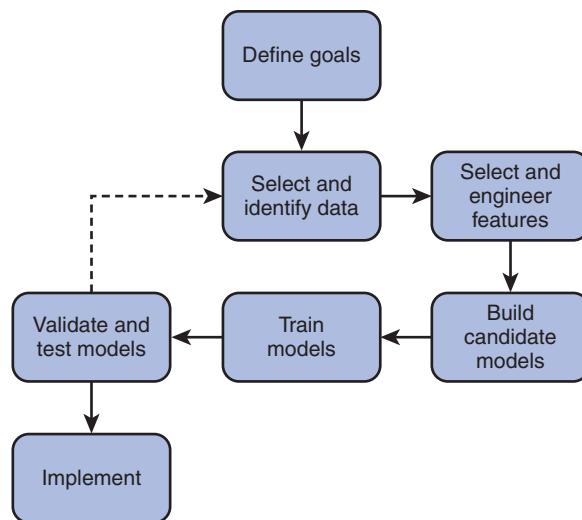
The *technology-driven definition* refers to the challenges with having a dataset that is too large to reside on a single computer system or a dataset that is too complex for a traditional relational database management system. Big data requires

distribution over a larger, networked area (Fig. 13.3). Traditional datasets contain structured data that are easily modeled using a relational format where data is organized into rows and columns, making up one or more tables. Relational databases are manipulated using a standard programming language, SQL (Structured Query Language). In contrast, *big data* may contain a variety of data types, including those that are unstructured or semi-structured and, therefore not amenable to tabular database structures. This poses fundamental difficulties, compared with traditional data architectures, since any query or computation with the data must coordinate resources that are spread over a larger domain than a single server or storage device and must be able to do so for relational and nonrelational data models. These distributed

jobs require parallel processing to increase efficiency, given the number of computational processes typically required of *big data*. Several software tools (MapReduce, Hadoop, etc.) were specifically designed to overcome the challenges raised by these sorts of distributed processes.<sup>2,3</sup> Similarly, nonrelational data structures such as NoSQL have become a critical component of *big data* architectures because they are well-suited for high throughput use of unstructured and semi-structured data types commonly encountered in *big data* applications.<sup>4</sup> NewSQL databases have also gained popularity as a new type of relational database structure with the scalable properties of NoSQL.<sup>5</sup> When the technology-driven definition of *big data* is being used, it is typically in the context of these or similar software tools and data architectures that solve the particular technical challenges imposed by *big data*. It is certainly true that large health system databases can outstrip traditional storage and computing paradigms, and in this sense the technology-driven definition is relevant. However, these issues are now typically transparent to the end-user, as they are handled by enterprise IT groups, and thus the clinical laboratorian does not ordinarily have to consider the details of the technical solutions that have been developed.

The *property-driven definition* of *big data* refers to a number of inherent characteristics of data that are being acquired in modern commercial and other settings. These properties are often summarized by the 4 (or 5) “Vs”: volume, velocity, variety, and veracity (with some commentators adding “value”).<sup>6,7</sup> Volume refers to the increasingly large amounts of data that are being produced; in the health care context, the amount of laboratory and clinical data now available is notable. Velocity refers to the rate at which information is being recorded; in a major academic medical center, it is not unusual to generate over 15 million tests per year in a core laboratory alone. Variety refers to the different modalities of data; the combination of numeric laboratory data, textual interpretive data, imaging data from pathology and radiology, and other physiologic measurements contribute to the variety of data within the modern EMR. Veracity refers to the potentially unreliable nature of the quantities of information, often with the assumption that volume, velocity, and variety can help to overcome the limitations posed by issues with veracity. While the clinical laboratory has focused on high-quality data streams, it remains possible that errors have crept into the medical record, and this possibility should be accounted for when assessing the overall patient picture. The property-based definition of *big data* provides important insights into the nature and quality of data gathered at high scale, and it is important for laboratorians to assess the extent to which their data fit this description.

Finally, the *analysis-driven definition* of *big data* uses the term to refer to high-dimensional datasets that are not amenable to human interpretation in the absence of computational tools for efficient data visualization and machine learning. This definition appears to unite the more formal, property-driven definition, which has characterized much discussion in the internet-driven commercial sector, with a recognition of challenges imposed by smaller high-dimensional datasets. These laboratory datasets may be extracted from the laboratory information system (LIS), or they may be genomic, proteomic, and metabolomic datasets which report large number of analytes from a single patient. Interestingly,



**FIGURE 13.4** Overview of the machine learning process. (Modified from Haymond S, Julian RK, Gill EL, Master SR. Machine learning and big data in pediatric laboratory medicine. In Dietzen DJ, Wong ECC, Bennett MJ, Haymond S, editors. *Biochemical and molecular basis of Pediatric disease*. 5th ed. Cambridge, MA: Academic Press; 2021.)

the analysis-driven definition emphasizes the common set of computational tools that are required in both cases, whether a metabolomics interpretive problem or a traditional *big data* set that exhibits the 4 Vs. Whether or not the clinical laboratorian finds the property-driven definition directly applicable, the tools of machine learning are critical to the ability to make intelligent use of the data.

### Overview of the Machine Learning Process

An overview of the machine learning process is shown in Fig. 13.4. Machine learning approaches start with defining the goals for the model and framing the relevant question as a machine learning problem that will be used to generate a model that is a functional representation of the available data with some amount of error. A suitable source of data must be identified for use in the model. Typical sources within the laboratory include records from the LIS; measurement systems with large numbers of potential data elements, such as mass spectrometry; other high-dimensional data typically associated with the “-omics”; and digital images. Exploratory data analysis and data preparation steps are conducted to help identify or engineer potentially informative predictors (or features) for the candidate models. A portion of the available data needs to be designated as a “training set” that will be used to shape the model parameters as it learns to associate input data with a correct output. There are a large number of actual machine learning algorithms that can be utilized, and each one has slightly different parameters and performance for certain applications. The remaining data may be used as a “validation set” used to select the best machine learning algorithm and associated set of parameters for the problem. Model training and validation results may necessitate reiterating through the process to find a seemingly useful model. Once the best performing model and its parameters are selected, the model must be tested, ideally on an independent “test set” of data, which may be held out from the original data set. In the best case, test data can be independently

collected from a variety of settings in order to determine whether the model is able to generalize its predictive performance in new situations. Implementation of a machine learning model requires consideration of end-user needs and available infrastructure. We will address each of these steps in subsequent sections of this chapter.

## DATA SOURCES

The increased volume and variety of available data are among the major factors that have driven interest and successful application of machine learning methods in various industries outside of laboratory medicine. Within the laboratory, data have also become more plentiful both at a single time point (with increased numbers of conventional tests or newer, highly multiplexed testing formats) and longitudinally as patients are monitored over time. In turn, we are experiencing growing enthusiasm for using machine learning methods to aid in medical and operational decision making in clinical laboratories. Typical clinical laboratory data for machine learning will be discussed based on their source.

### Laboratory Information System Data

One of the most abundant data sources for machine learning in laboratory medicine is the LIS. LIS databases contain millions of records that are coded and available as highly structured and standardized data elements. Increasingly, organizations are integrating laboratory data with clinical features and outcome data and are aggregating large amounts of this medical data across health systems, creating data repositories ideal for use with machine learning. LIS data repositories also contain information that may not be transmitted to electronic health records for patient care, but are valuable features for predictive analytics in laboratory operations. Examples include documentation of physician notifications of critical results and time stamps for intermediate steps in laboratory processes (in addition to collection, order, receipt, and result times).

LIS data are often obtained through reports that are exported as formatted text (typically comma-separated values; .csv files) or in spreadsheet formats (e.g., .xls). The columns of this format reflect different variables, such as patient name, medical record number, test name, test result, units, upper and lower reference limits, flags, and others. Every row lists a different result, and the format can be easily imported into a number of data analytics packages for analysis and machine learning. A similar form of data can be acquired by directly querying the database of the lab information system using SQL or another equivalent structured language. Results of the database query can also be placed in the same format and equivalently processed by most data analytics packages.

At this time, many laboratorians experience limitations with access to LIS data, in both timeliness and breadth. Results from ad hoc queries enable development and validation of machine learning models using retrospective data. However, data access must become timely and automated to facilitate implementation of machine learning pipelines, as discussed in a later section. Though of benefit due to its large case mix, aggregated laboratory data in multisite repositories is likely subject to method bias related to the different measurement procedures used. Such data veracity issues should be examined during exploratory data analysis and addressed

prior to modeling.<sup>8</sup> Finally, because not every patient gets the same panel of lab tests, some LIS datasets may be sparse (that is, many values are not available for many analytes). Sparseness is discussed in a later section of this chapter.

### Mass Spectrometry Data

The data produced by mass spectrometers can come in several forms depending on the instrument and how the instrument acquisition method is designed. In this section, some of the most common types of mass spectrometry data will be described based on the type of instrument. For a more comprehensive background introduction to the principles of mass spectrometry, see Chapter 20.

To optimally utilize mass spectrometry data, it is important to understand both the various scales and dimensions of mass spectrometry data and the methods by which mass spectrometry data can be accessed for use in machine learning applications. As a rule, the raw binary files created by instrument acquisition software systems cannot be read directly. Instrument vendors have many reasons for using a binary format, which is sometimes even encrypted, as the native format for a mass spectrometer. There are two ways to access raw data from mass spectrometers. One way is to use a software library supplied by the vendor to extract data directly from the native binary format. The second is to use one of several file formats that have been created as either formal or *de facto* standard file formats. Some instrument vendors provide a feature in the acquisition software to generate an export file. In this section, the most common export file formats will be discussed. Direct access to the vendor binary files will not be covered. For information on using instrument vendor software libraries, refer to the instrument vendor documentation.

Mass spectrometry data can take a 1-D (intensity of a single ion) or 2-D (intensity of multiple ions forming a spectrum) form. Instruments often report data using flat files; this is especially true for 1-D data, but flat files can be used for 2-D data as well. The most common flat file for spectra is the format called JCAMP-DX used by groups like National Institute of Standards and Technology (NIST) and the US Environmental Protection Agency (EPA).<sup>9,10</sup> It is relatively straightforward to read flat text files containing either 1-D or 2-D data from the types of instruments mentioned for purposes of machine learning. One of the most common applications using only 2-D spectra is the look-up of an unknown from a library. Library search requires selecting a similarity measure and is subject to the curse of dimensionality described below in the section on additional principles and limitations of machine learning. Recently, medical devices that perform microbial identification have been constructed using matrix assisted laser desorption time-of-flight (MALDI-TOF) single-stage analyzers, which use libraries of standardized spectra from a wide range of organisms to perform a probabilistic identification.<sup>11,12</sup>

When some form of chromatographic separation is introduced into a mass spectrometry system, 3-D data (spectra over time) are available. The mass spectrometer is then usually configured to repeat an acquisition for a period of time. Again, the intensity of a single mass-to-charge ratio ( $m/z$ ) value could be monitored for a period of time, which would produce a 2-D data set (sometimes called a chronogram, or a chromatogram if there is chromatographic separation) where

now the *x*-axis is time. However, it is also typical to monitor several *m/z* values during the sample introduction period, which produces a three-dimensional data set representing a collection of 2-D chronograms. In this 3-D data, the *x*-axis is usually time, and the *y*-axis is used to represent the individual *m/z* values monitored, and the *z*-axis represents the intensity observed. Common applications of chronograms include inductively coupled plasma mass spectrometry (ICP-MS) analysis of metals, flow injection, more recently ambient ionization systems such as desorption electrospray ionization (DESI),<sup>13</sup> and the various derivatives of paper-spray.<sup>14</sup> Chronograms are routinely used in clinical applications for a variety of analytes (see Chapter 20).

Some instrument software can export 3-D data in a flat-file format. While JCAMP-DX does not formally support the extra time dimension, there have been extensions proposed to allow the standard to be used for chromatography data.<sup>15</sup> To represent the time dimension, a format called “analytical data interchange format for mass spectrometry” (ANDI-MS) was created by the Analytical Instruments Association (AIA) based on the Network Common Data Format (netCDF) developed and maintained by the Unidata program at the University Corporation for Atmospheric Research (UCAR). AIA/ANDI-MS was adopted as a standard by the ASTM International organization.<sup>16</sup> The format has been known variously as ANDI-MS, AIA, or netCDF. The ASTM standard uses netCDF because it is a machine-independent binary file format that can hold almost any array-like data. The format can be used to hold single spectra, chronograms, and full chromatographic data. Programming libraries are needed to access netCDF, but they are not dependent on an instrument vendor. There are many language bindings for netCDF data, and the format is extremely well documented and supported by UCAR. The specific variables stored in the netCDF file according to the ASTM specification, however, are limited. The most significant limitation is the lack of a way to store higher dimensions of mass spectrometry. This is not an actual limitation of the UCAR netCDF format, but rather the way the format is used to store mass spectrometry data.

It is now common to collect clinical data using one or more liquid chromatographic stages combined with two stages of mass spectrometry. This has typically been called LC-MS/MS. For machine learning applications, the extra stage of mass spectrometry requires keeping track of the *m/z* value used by the first mass analyzer (precursor ion). The second stage of mass analysis can then be set, as described above, to either collect a full product ion spectrum or monitor a single product ion. When one precursor/product pair is monitored, it's called single-reaction monitoring (SRM). When multiple *m/z* pairs are monitored, it is commonly called multiple-reaction monitoring (MRM). In both cases, chromatograms are generated, representing the intensity of a precursor ion fragmenting into a product ion. The ASTM netCDF format makes no provision for tracking the extra information of the product/precursor pair, so it cannot be used to store data for SRM or MRM measurements. New formats have been created, primarily based on the XML standard,<sup>17</sup> to represent tandem mass spectrometry data. XML is essentially a well-defined specification for a flat text file which improves machine readability.

The first XML format to be developed for mass spectrometry data was mzXML,<sup>18</sup> which is well supported and is an extremely practical format for all types of mass spectrometry

data analysis. The mzXML format was developed at the Institute for Systems Biology (ISB) to support the first proteomics applications. As such, it is under the control of the ISB, which has the benefit of a very long-running support team and widespread use through the distribution of ISB open-source tools. It does, however, change as needed to support the ISB tools. It also uses some non-XML features to improve performance. To establish a more robust *exchange* data format, the Human Proteome Organization (HUPO) created an open, multi-vendor standards body along with processes to approve changes to the format. The standard produced is now called mzML and has the support of vendors, developers (including the ISB), and journal editors. The mzML format deliberately trades performance for completeness.<sup>19</sup> Unlike mzXML, which uses a fixed collection of elements and attributes to describe complex mass spectrometry measurement, mzML uses a small set which must be combined with a controlled vocabulary. The schema for mzML has not changed in many years, ensuring that software designed to read the data will not be broken by changes. The controlled vocabulary, called psi-ms, however, is being constantly updated as new instruments and new experiments are invented.<sup>20</sup> Both the format and the controlled vocabulary are under the control of an editorial board, and both are supported by a very active community. The choice to use mzXML or mzML should be determined by the application. When performance is critical, mzXML is a clear choice. When long-term interchangeability is important, mzML has a better chance of being readable over the long term.

The XML formats described above can be used for almost any mass spectrometry experiment. They are designed to handle very high-dimensional data such as the type collected for data-dependent and independent acquisition, as well as multiple stages of mass spectrometry (sometimes referred to as MS<sup>n</sup>) and spectra collected with a variety of system settings such as multiple collision energies and other parameters.

Finally, many mass spectrometry measurements are designed to produce a quantity or ratio. For flow injection, this may be a quantity based on the intensity of a single *m/z* value. When collecting a single spectrum, the result could be the ratio between various *m/z* intensities or the presence or absence of an ion signal at a given *m/z* value. In chromatography measurements, the result could be everything from the area of a single SRM peak or the combination of multiple SRM peaks further processed into a quantity through the use of a calibration curve. All of these results have potential use in machine learning applications. The most common data format for mass spectrometry results is a vendor-specific flat file. These usually require complex parsing because many times, flat files are designed to be read by humans and not machines. One solution to improving the accessibility of results data is a format called mzTab-M developed by a consortium including HUPO, the Metabolomics Society, and the Metabolomics Standards Initiative.<sup>21</sup> mzTab-M is a tab-delimited format that is designed to hold the results of complex mass spectrometry measurements and be easily readable by both humans and machines.

The data produced by mass spectrometers in clinical applications can range from single numbers to very high dimensional data. Through the work of many groups, and with the collaboration from instrument vendors, there are now several ways to access everything from raw spectral data to computed results. Using the tools provided by these

informatics groups, it is now possible to develop sophisticated machine learning systems from almost any instrument vendor's data at any level of complexity.

### Omics Data

The general term “omics” refers to the high-throughput methods to comprehensively characterize and quantify a large number of molecules, grouped by their structural or functional similarities. Mass spectrometry (discussed above) is one method of generating omics-scale data; however, there are many others that also provide rich sources of data for machine learning. For example, next-generation sequencing can provide not only genomic data but also quantitative transcriptomic information through RNA-seq.<sup>22</sup> In this case, every subject or sample will have a numerical result for each transcribed gene that is measured indicating its abundance.

Omics-scale data can create technological challenges because their size and complexity exceed the capacity of traditional infrastructure components used for data storage, data transfer, and computational power. An additional challenge with these datasets is that often the number of variables ( $p$ ) measured vastly exceeds the total number of cases ( $N$ ), and situations with  $p \gg N$  can be particularly prone to overfitting (see below) and other problems.<sup>23</sup> Care is required when using machine learning approaches in this setting, with a very high number of dimensions, to ensure that robust models are created.

## TRAINING AND VALIDATION

### Bias Variance Tradeoff

It is a fact of statistical machine learning that there is no one best type of model for all problems. This result has come to be known as the *No Free Lunch Theorem*.<sup>24</sup> All models have three sources of error that add up to give the total estimation error: *Bias*, *Variance*, and the *Irreducible error*.<sup>25</sup> The first two are directly related and must be balanced to achieve a sound machine learning system. The third is independent of the first two and requires a separate approach to reduce its impact.

1. **Bias:** This is the error made by the model due to generalization. Bias is the result of the model being simpler than the real-world problem being solved or the features used by the model are not informative enough. A biased model is not totally without value. For example, in regression, we often choose a simple linear model to approximate nonlinear, or noisy data. The downside of a biased model is that, regardless of the amount of training data used, the error will not be reduced beyond some limit. Thus for highly nonlinear data, a linear model will have *high bias*. High bias models *underfit* the training data. The bias of a model can be lowered by increasing the complexity of the model or increasing the number or information content of features. An *unbiased model* is one for which increasing the amount of data will result in making the bias lower, ultimately to zero. However, counter-intuitively, zero bias is not always desirable. For example, in data with noise, a regression model with zero bias would draw a curve that went through every data point, which means the model has simply *memorized* or *overfit* the training data, not actually modeled it. For classification problems, this is the same as getting all of the labels on the training data

precisely correct. This type of model will not perform well when new data are introduced.

2. **Variance:** This is the error caused by a model being sensitive to variations in the training data. Variance is the result of the model being more complex than the real-world problem being solved, or containing too many features for the size of the training set (see the discussion of  $p \gg N$  in the previous section). A model with *high variance* is overly sensitive to small fluctuations in the training data and can be the source of overfitting mentioned above. It is this relationship between bias and variance that must be considered when building a model. A low-bias model with high variance will overfit the training data and perform poorly on data that is not part of the training set. A *low-variance* model is one that is not sensitive to fluctuations in the training data. A low variance model with high bias will underfit the training data and also perform poorly on new data.
3. **Irreducible error:** This error is due to the actual noise in the data. Noise can take many forms and should be characterized in order to design a process to lower it. There are many ways to “clean up” data before processing. Signal processing can help reduce the noise in measurement data. Care must be taken to apply methods such as signal averaging and filtering. Distinguishing signal from noise is a nontrivial task, and simplistic methods can distort data in a way that makes models perform worse than they did without noise reduction. The other primary source of irreducible error is outliers. An *outlier* is an observation that is an abnormal distance from other values in a random sample from a population. In statistical machine learning, many measures of model performance can be highly distorted by outliers. For example, when using least squares for either regression or classification, the mean squared error (MSE) is used, and the mean is known to be especially sensitive to outliers.<sup>26</sup>

The ideal machine learning model has both low bias and low variance, and finding the right tradeoff between these two types of error is the purpose of model validation. The data available for training a model can be used in various ways to maximize the performance of a particular model. Constructing a robust validation process combined with good data cleaning is a requirement for building a model that performs well with real-world data.

### Regression versus Classification

Machine learning, at its core, is about building models that make predictions based on data. There are two basic kinds of predictions that can be made: real valued and categorical. Regression models make numeric predictions, and classification models make categorical predictions. Both types of models take a set of features in a training data set and produce a response variable. Regression and classification both fall in the type of machine learning known as supervised learning, which simply means that the training data set contains known values for the response variable. The job of supervised machine learning is to fit the parameters of a model to the training set so the difference between the predicted response variable and the known value is minimized.

For each example in a dataset, the training algorithm calculates the difference between the true response values in the dataset and the values computed with the model, i.e., it thus

evaluates how closely the model output fits the true values. The function to compute this error value is the *loss function* (or *cost function*). This loss function is used to either compute the weights assigned to the factors in the model or search for the weights that give the minimum error. In this way, the algorithm “learns” how to optimally match its output to the training output.

The loss function is simply a way to quantify how close the predicted response values from a model are to the true response values from the dataset. For linear regression models, the most common loss function is the *mean squared error*. The MSE is, however, sensitive to outliers and can be prone to overfitting, which will be discussed further later in this section. Other measures of error can also be used depending on the application. Regardless of how the error is determined, training a regression model involves minimizing the error between the predicted values and the observed values for a given input.

For purely linear models (including polynomial regression and multiple regression), there happens to be a closed-form analytic solution for the values of the coefficients that give the minimum error. This solution is so useful it is called the Normal Equation of linear regression. Most linear regression algorithms use this approach because it is easy to code and very fast. For nonlinear models like the exponential:  $y = \exp(ax)$  or the logistic function:  $y = 1/(1 - \exp(-ax))$ , there is no analytic solution to compute the coefficients with the minimum cost function value. Worse, some models do not have so-called *convex* cost functions at all. Convex cost functions can be thought of as “bowl-shaped” with smooth sides and a definite bottom. They have the property that a simple search method called gradient descent (GD) is guaranteed to find the bottom, and that value is optimal solution to minimize cost, or error. For more complex cost functions, more sophisticated optimization method can be used to search for the minimum error. However, if there is no definite “bottom” to the error surface, then there is no guarantee that the “best” solution will be found. It is important to note that during training, many important models run the risk of falling into what appears to be the lowest point in the error surface. This so-called “local minima” will produce a model with suboptimal performance.

There is a strong relationship between regression and classification. One way to think about it is that regression attempts to train a model which goes through data points in the training data when the desired result is a number. At the same time, classification uses the same techniques to train a model that goes between data points in training data when the desired result is the separation of classes. With classification, “error” is how far off the model’s prediction is for the class of an input, instead of the model’s error in predicting a number from the input. Just like regression problems can be solved by fitting a straight line through a data set, classification problems can have linear decision boundaries. The idea that both regression and classification both compute functions means that almost all machine learning algorithms can be used for both regression and classification.

### Assessing the Performance of Machine Learning

For classification problems, the most common approach for quantifying the closeness of the results of a model to the dataset is to compute the fraction of incorrect classifications.

Classification algorithms optimize the weights of predictors to minimize the fraction of misclassified cases, and like regression, decision boundaries between classes are subject to over- and underfitting. Unlike regression problems, classification problems are more complicated because, in classification, it is useful to know how well the model placed cases into various classes.

### Confusion Matrix

A confusion matrix is a table that gives a summary of how well a model placed each example into the classes. The term “confusion matrix” is commonly used within machine learning to describe a results table such as is used within laboratory medicine to report the outcomes of diagnostic tests. The rows are the actual (true) labels for the classes in the dataset, and the columns are the predicted labels for the classes:

	<b>CKD (predicted)</b>	<b>Not CKD (predicted)</b>
CKD (actual)	49 (true positive TP) 1 (false negative FN, Type II Error)	1 (false negative FN, Type II Error)
Not CKD (actual)	5 (false positive FP, Type I Error)	95 (true negative TN)

CKD, Chronic kidney disease.

This confusion matrix shows the results of an imaginary classifier for chronic kidney disease (CKD). The classifier is binary, meaning it simply classifies an example between the two classes. A confusion matrix can represent any number of classes and can have as many rows and columns as there are class labels. This example shows that of 50 cases that were actually CKD, the model predicted 49 correctly, or 49 true positives. The model misclassified one case as Not CKD, which is a false negative. Of the 100 cases that were actually Not CKD, the classifier correctly predicted 95 cases. These classifications are called true negatives. The classifier misclassified five cases as having CKD when they were actually negative. These are false-positive classifications. Depending on the application, different importance is placed on each of the four outcomes (TP, FP, FN, and TN). A classifier meant to diagnose a disease may be “better” if it has a low false-negative rate, since those examples may mean that a patient has a disease and the classifier missed it. However, this is not always the case. If a disease is known to be slow acting, and further tests will be performed, the classifier will get additional attempts to get a true positive.

If the treatment for a disease has the risk of harm, then a false positive may be worse than a false negative. In this way, patterns in the confusion matrix can be used to diagnose problems with a model. It is also common to combine individual outcomes to compute summary metrics that give a more global picture of classifier performance.<sup>27,28</sup> It is worth noting that the terms that follow (accuracy, precision, sensitivity, etc.) all have slightly different definitions in the field of machine learning than in other areas of science.

### Accuracy

Accuracy is the number of correctly classified examples divided by the total number of examples:

$$\text{accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

When all classes are equally important, accuracy gives a good measure of how well a classifier performs. As discussed above, this is not always the case. In the diagnostic example given above, the *cost-sensitive accuracy* metric can be used. This is done by estimating a cost for each type of misclassification and multiplying its cost (one for FP and one for FN) before computing accuracy. Accuracy is also most informative when the number of cases in each class is reasonably balanced. When this is not the case, as is often the case in health care datasets, using just the accuracy to describe a classifier can be misleading. To illustrate this point using an extreme case of a binary classifier, consider a dataset with a 95% normal and 5% abnormal case mix. One could achieve an accuracy of 95% simply by using a classifier that predicts every case will be normal.

### Precision

Precision is the number of correct positive classifications (true positives) divided by the total number of positive classifications:

$$\text{precision} = \text{TP}/(\text{TP} + \text{FP})$$

Precision is a term from the information retrieval domain and originally described the number of relevant records in the list of all records returned. In the classification context, high precision means avoiding the mistake of classifying a subject as having a disease when they do not. Another example often used is classifying a legitimate email message as spam.

### Sensitivity (Recall)

Sensitivity and recall are terms used for the number of correct positive classifications divided by the total number of positive cases in the dataset being used for testing:

$$\text{sensitivity} = \text{TP}/(\text{TP} + \text{FN})$$

Sensitivity is a measure of a classifier's ability to avoid false negatives. In the example of a classifier used to identify a disease, sensitivity is the proportion of people who are predicted to be positive for the disease among all those who have the disease. The term recall comes from the domain of information retrieval, where the objective is to identify data records in a search. In that context, it describes the classifier's ability to not overlook records that match the search criteria.

### Specificity (Selectivity)

Specificity is the number of true negatives divided by the total number of negative cases in the dataset:

$$\text{specificity} = \text{TN}/(\text{TN} + \text{FP})$$

Specificity is a measure of a classifier to avoid false positives. In the discussion of weighted accuracy, it becomes clear why metrics like sensitivity and specificity are used in a medical context. Rarely are false positives and false negatives equally weighted. Further, there is a tradeoff between sensitivity and specificity that is important to note. A classifier that always labels cases positive has 100% sensitivity, but since it ignores false positives, it cannot be used to rule out a disease condition. Specificity measures the ability of the classifier to correctly label healthy patients without a condition.

### Receiver Operating Characteristic Curves

Receiver operating characteristic (ROC) curves are commonly used to assess binary classifiers. The term is a historical artifact from radar and communications research (thus the use of the word receiver). A ROC curve is a plot of the true-positive rate (TPR) versus the false-positive rate (FPR). The ROC curve can only be computed from classifiers, which give either a confidence score or a probability for a case classification. The model is used to make a prediction, and the TPR and FPR are computed using threshold values in the range from 0 to 1 (at any discrete increment desired).

The TPR is:

$$\text{TPR} = \text{TP}/(\text{TP} + \text{FN})$$

and the FPR is:

$$\text{FPR} = \text{FP}/(\text{FP} + \text{TN})$$

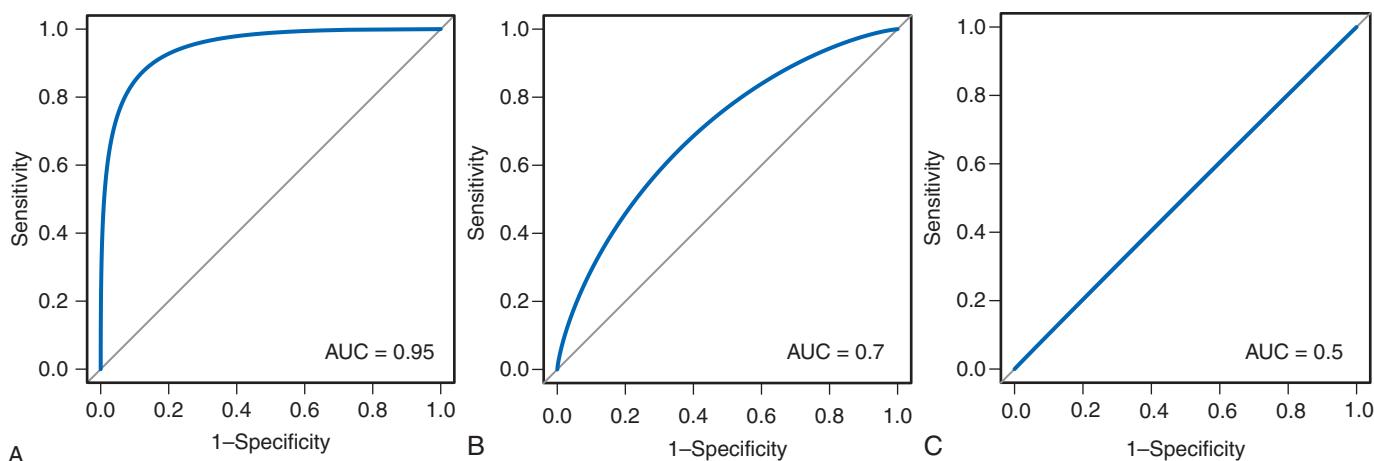
For each case, the probability threshold for assigning a case to the positive class is set to a value starting at 0, and then the TPR and FPR are computed. In Fig. 13.5, it's easy to see that if the threshold for classification is 0, all the cases will be labeled as positive, and both the TPR and the FPR will be 1. When the threshold is 1, then all cases will be classified as negative, and the TPR and FPR will both be 0. If at every threshold value the TPR and FPR are equal, then the classifier is no better than random guessing. The metric which can be drawn from this process is the area under the ROC curve (AUC). For the random guessing case, the AUC would be 0.5. What is sought is a classifier which keeps the TPR as high as possible while keeping the FPR as low as possible. The AUC in this instance would be as close to 1 as possible.

The AUC of the ROC curve is not a perfect measure of classifier performance, and care must be taken, especially if the ROC curves of two classifiers cross, or if there is a small sample size and the curves become noisy. Despite this, the AUC has held up as a pragmatic choice for measuring classifier performance over a large range of model types and classification problems.

### Model Testing

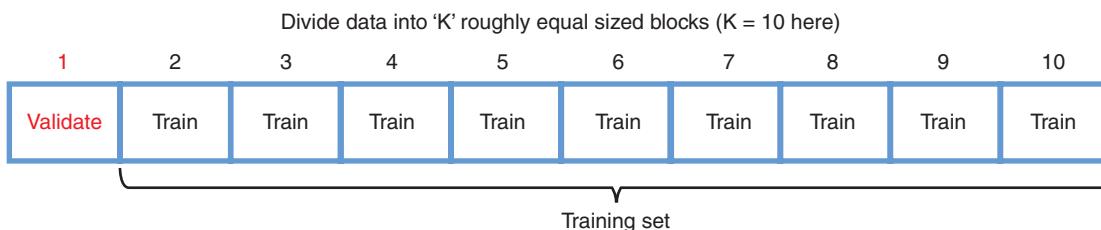
A good machine learning model is one that makes good predictions on data it has never seen before. That is to say, it works with data that was not part of any dataset available when the model was constructed. To build a model that generalizes well, it is necessary to estimate how well the model will perform on data not used as part of training. The typical method for estimating model performance is to use the *validation set* approach. The idea is to split the dataset into parts, and each part is used for a different task. Depending on the size of the dataset, varying degrees of complexity can be used in the splitting process.

The basic testing process begins with splitting the labeled dataset into two parts: the *training set* and the *validation set*. First, the data are randomized. Next, 70 to 80% of the data are assigned to the training set and the rest to the validation set. The validation set is called a *hold-out set*. This procedure can be further enhanced by splitting the hold-out set further into two sets, one as the validation set and a *test set*. The training set is used to train the model, while the validation set is used to select the best machine learning algorithm for the problem. The validation set is also useful in finding the best values for any parameters that are not variables in the



**FIGURE 13.5** Receiver operating characteristic (ROC) curves. Sensitivity is plotted versus  $1 - \text{Specificity}$  as the assay cutoff is changed. Plots are arranged from most informative (A) to least informative (C). A, ROC = 0.95; B, ROC = 0.7; C, ROC = 0.5, which is completely uninformative (the same as chance). (Modified from Haymond S, Julian RK, Gill EL, Master SR. Machine learning and big data in pediatric laboratory medicine. In Dietzen DJ, Wong ECC, Bennett MJ, Haymond S, editors. *Biochemical and molecular basis of pediatric disease*. 5th ed. Cambridge, MA: Academic Press; 2021.)

#### Cross Validation in Detail “K-fold”



**FIGURE 13.6** *k*-Fold cross-validation. For each iteration, a portion of the data is removed, and the remaining data are used for training. The trained model is tested on the held-out data, and the process is then repeated by removing the next portion of the data. (Modified from Haymond S, Julian RK, Gill EL, Master SR. Machine learning and big data in pediatric laboratory medicine. In Dietzen DJ, Wong ECC, Bennett MJ, Haymond S, editors. *Biochemical and molecular basis of pediatric disease*. 5th ed. Cambridge, MA: Academic Press; 2021.)

model, but rather settings associated with the type of model used. This type of parameter is called a *hyperparameter*, and it must be selected to optimize the fit to balance the bias and variance. (called *hyperparameter tuning*). If a test set is created, then it is used as a final check of the performance of the selected model and hyperparameters before applying the model to real-world data.

There are many choices for dividing data between training, validation, and test sets, depending on the application. If the model has already been selected, then a test set is not needed since the validation set will not have been used to make any decisions about the model. The validation set can be used to determine model performance before implementation. Second, if the original dataset is small (relative to the number of factors), then the splitting process may leave too little data in both the training and validation sets. A very common solution to this problem is a resampling process called *cross validation* (CV), which will be covered below.

The validation set approach to training and testing machine learning algorithms uses two steps: train the model on the training set and then perform an evaluation of the model

using a quality of fit as described earlier. If the quality of fit indicates that the model is performing poorly, then a search can be performed to optimize any hyperparameters and then reevaluated. Further, it is common to use the results of a quality of fit measure to select between alternative algorithms. This is why it can be useful to hold out a final test set to evaluate the final algorithm and hyperparameters before using the model.

Cross-validation was developed because of situations where the number of observations is so small that splitting leaves too little data for either effective training or testing. The problem with small datasets is that—if they are randomized and 70% is taken as the training set—it could be the case that there is higher variability in the training set than the overall dataset purely by accident. The CV process is performed as follows: first, the dataset is randomized and then split into  $K$  parts. One part is held out, and the rest of the data are used as a training set as shown in Fig. 13.6. The held-out data are used as the validation set, and the results are recorded. Next, another part is held out as a validation set, and the remaining data are again used for training. The process is

repeated until each part has been used as the validation set. The cross-validation estimate of the prediction error is simply the average of all the individual validation errors.

When building training and validation sets it is important to consider the distribution of classes and avoid class imbalance. The training, validation, and test sets should all contain roughly the same proportion of classes. This also goes for selecting cases in each fold for cross-validation. Up- or down-sampling strategies may be used to mitigate imbalanced classes.<sup>29,30</sup>

It turns out that the bias-variance tradeoff comes into play when performing cross-validation and determining the optimal value of  $K$  (the number of parts). Typical values for  $K$  are 5 and 10. These are arbitrary, but empirically seem to produce reasonable results. If  $K$  were equal to the number of cases, the process would be known as *leave one out cross validation* (LOOCV). Although LOOCV is approximately unbiased for the actual prediction error, it can have high variance. This is because, by taking only one case out, the remaining training sets all are roughly the same. As  $K$  is lowered and the number of cases in the validation set increases, the estimate will have lower variance but could become highly biased. Therefore the choice of the number of folds should be chosen based on the shape of the learning curve. If adding cases to a training set does not change the validation outcome by much, then the learning curve does not have much of a slope at that training set size. If the training curve has a significant slope at the training set size for a particular value of  $K$ , then CV will overestimate the true prediction error due to bias. Generally, using 5- or 10-fold cross-validation represents a good compromise between bias and variance for this technique.

### Overfitting

As has been previously introduced, one reason why cross-validation and independent validation sets are significant relates to the phenomenon known as *overfitting*. An overfit model has been constructed to fit the training data very closely, but it does not generalize well to new datasets. As an example, consider a simple classification case with two classes: disease and no disease. The disease class has biomarkers with values less than 5, and the nondisease class has biomarkers with values  $\geq 5$ . Suppose also that the training set has two cases with biomarker values of 1 and 3, and two nondisease cases with biomarker values equal to 8 and 9. Under these circumstances, it is possible to create a perfect classifier that says, "If the biomarker is equal to 1 or 3, you have a disease case; otherwise, you have a nondisease case." This rule classifies the training set with perfect accuracy. However, given that the actual underlying rule that we were hoping to capture is related to a cutoff of 5, our model will not generalize well; specifically, it will not correctly predict new cases with biomarker values of 0, 2, or 4. By focusing too closely on details of the training data set, the model loses its ability to generalize.

While this simplified example of overfitting may seem like an obvious mistake, it is much more difficult to avoid this issue when dealing with a high-dimensional data set where the "real" rule that underlies the difference between diagnostic classes is not known. Consider that as the number of dimensions (variables, or biomarkers) increases, it may become easier to find a combination of these dimensions that gives a close-to-perfect fit for a categorical diagnostic problem. In

fact, if there are enough dimensions and a small number of cases, it becomes increasingly likely that a good fit can be derived purely by chance. For this reason, it is never appropriate to report the performance (AUC, etc.) of a classifier on the full original training set data to infer the actual performance of a model.

Cross-validation is a critical tool to help overcome overfitting by ensuring that the full data set is not used all at once when estimating parameters of the model. More importantly, it is critical that a truly independent test set be used to validate any model, because this ensures that the model performance has not been skewed by overfitting. A number of the algorithms described below in the section on common machine learning algorithms have been designed to combat overfitting.

Another way to reduce overfitting is to change the loss function associated with fitting the model. *Regularization* is a common method for reducing overfitting in all types of machine learning. Regularization in regression changes the loss function to add a term whose value is higher when the model is more complex. These are called penalty terms and are based on the idea of changing the way distance is measured. Regularized loss functions introduce a hyperparameter into the model, by requiring a parameter to be specified in order to measure the quality of fit.

### Feature Engineering

In machine learning terminology, datasets have two parts:

1. The *response variable* which is to be predicted. These are either numbers (in regression problems) or class labels (in classification problems)
2. The *features* from which the prediction is to be made. Features are also called factors or predictors. They can be either numerical values or categorical values. If there is more than one feature in a dataset, then the collection is called a *feature vector*

It is easy to think that the raw data available to build a model will constitute a suitable dataset.<sup>31</sup> Unfortunately, this is rarely the case. It often takes a great deal of effort and usually significant domain expertise to turn raw data into a good dataset. The labor-intensive part of the job is usually dealing with data formats and simply putting the data into what has become commonly referred to as a "tidy" condition. Having tidy data starts with recognizing that machine learning algorithms can only deal with data that can be represented as a square table.

Once raw data are in a table, cleaning the data is almost always required. Examples of data cleaning include ensuring that:

- Categorical features are all spelled or coded consistently
- Features (columns) are consistent in format (numeric or text)
- A value was entered into each column properly during the data import, and no missing values result in an incorrect assignment of other values to features (for example, an "off-by-one" shift of data)
- Outliers are identified and handled as appropriate
- When multiple raw data files are combined, the rows match up exactly

After these high-level problems have been cleaned up, it is then important to evaluate the features both in terms of quantity and quality. Univariate and multivariate exploratory

data analyses, using both qualitative and quantitative approaches, are critical for examining a data set prior to modeling. This allows a data scientist to assess the data integrity and gain insight into the data.

Pairwise comparisons may reveal useful relationships. For example, a large number of highly correlated features will not add much information to the model fit. Alternatively, too many uninformative features can make the model worse than if there were fewer, more informative features. If there are many more features than cases, the problem is overdetermined, and overfitting is likely to result. The pattern of missingness may itself be an informative feature, whereas a feature with too many missing values may be of little value. Skewed distributions may suggest the need for scaling or transformation prior to modeling. Domain expertise can help with preparing and selecting features. Techniques for performing these steps will be covered in the next sections.

Most learning algorithms can only deal with numerical data, so if there are categorical features, they need to be encoded in some fashion. In some cases, the categories can be replaced by numeric values. This, however, can have some problems. If there are only two categories, then coding them as 0 and 1 can work, although numerically, it implies that there is an order to the categories. One common way of eliminating the idea of order when changing categorical features to numerical values is called *one-hot encoding*. This encoding method creates a new feature for each value of a categorical variable. This new feature is called a *dummy feature*. For example, if a feature contained values like “arms,” “legs,” “torso,” “head,” one-hot encoding would transform the feature into four new dummy features, each named for the unique values in the original feature. Then, the new feature would be given a value of 1 if the categorical feature for that case had that value and 0 if it did not. Creating new features or changing features to make them more suitable for use in machine learning algorithms is at the heart of feature engineering.

Another example of feature engineering is putting numeric values into a condition to work with specific algorithms. Since many machine learning algorithms use GD to optimize the weights assigned to features, the range of values and their distribution can greatly affect GD’s ability to find the minimum error. To help algorithms like GD, it is sometimes helpful to convert the numeric values of a feature into a uniform range. This process is called *normalization* or *centering*, and it involves simply subtracting the minimum value of the feature from each case and dividing by the average value. This puts all the data points in the range of 0 to 1 and can dramatically improve the GD algorithm.

Another helpful feature engineering tool is called *standardization*. Standardization adjusts the numeric values of a feature, so they appear to have been drawn from a normal distribution with a mean of 0 and a standard deviation of 1. This is done by computing the mean and standard deviation of the cases for a feature and, like normalization, subtracting the mean and dividing by the standard deviation for each case. This results in replacing each numeric value of the feature with its z-score. Again, this can help various algorithms like GD.

The feature engineering tools discussed so far come with caveats. First, creating a large number of dummy variables using encoding may create an overfitting situation depending

on how many cases are in the dataset. This creates more work during feature selection. Second, the numeric data may have outliers that distort the calculation of the mean and standard deviation. Finally, the data in a feature may not be drawn from a normal distribution, so scaling it as if it were can change the data in ways that make model performance worse. As noted earlier, doing some exploratory analysis on a feature will provide a good idea of which techniques might be justified and helpful.

Another problem faced in almost all machine learning projects is missing data. A feature may have no recorded value for some cases. Some algorithms, like linear regression, cannot handle missing values and will give an error when they are processed. Because of this, several methods have been developed to deal with missing data ranging from the simple to the complex. There are machine learning algorithms that are able to handle missing data, like tree-based methods that ignore missing data, or linear mixed models which have particular strategies for estimating missing values. If the feature is not particularly informative, based on domain expertise, the whole feature can just be dropped from the data set. If the number of cases is large and the number of cases with missing data is relatively small, the cases can be removed from the data set. Each method should be tested to determine the effects on model performance. If a brute force method results in poor performance, then it may be worth attempting to fill in the missing data using *data imputation*.

The idea behind data imputation is relatively simple, although the implementation can be complicated. Imputation involves building a model of the feature and then using the model to compute a likely value for a missing element. One simple model is to use the average (or median or most frequent) value of the feature and replace the missing values with that value. Other, more sophisticated methods involve estimating the distribution of values and randomly drawing values from the distribution.

There are other more sophisticated methods for working with missing data, but the general advice is to try several methods, build several models, and use the method which gives the best test results.

## Feature Selection

The goal of creating any dataset is to build it from only informative features and for there to be many more cases than features. This is often not possible. In any real-world dataset, some features are redundant in any number of ways. Further, even in a world of *big data*, the number of features can often be overwhelming compared to the number of cases. An example would be gene expression data of a set of patients. With the growing size of expression panels and new sequencing techniques, the number of possible features measured can far exceed the number of cases. This situation calls for some method of selecting the most informative available features from those that are available. This is desirable not only because it leads to a simpler model, but also because some models like support vector machines and neural networks perform poorly when irrelevant features are included in the model. Other models like linear and logistic regression perform poorly when features are correlated.

The goal of feature selection is, therefore<sup>31</sup> reduce the number of features as far as possible without compromising the performance of the model.

There are two primary methods for feature selection: explicit removal by the analyst and implicit removal by the model. Explicit removal by the analyst involves the analyst doing exploratory data analysis on features and dropping features that are either highly correlated with another feature, or noninformative based on other factors such as the distribution of values. It is common to select features that are correlated with the response variable and drop uncorrelated features. Care must be taken when using this approach because when there are a large number of features and a small number of cases, the odds of there being a highly correlated feature by random is high. Doing correlation analysis on cross-validation folds can help reduce the odds of picking features that are correlated with the response by accident.

The other main method for performing feature selection is to choose a model that selects features as part of the training process. Some machine learning algorithms are particularly well suited for this, including tree methods, and so-called regularization models (see the following section for a more detailed discussion of these methods).

## COMMON MACHINE LEARNING ALGORITHMS

Machine learning algorithms are typically divided into two classes: supervised and unsupervised learning (Boxes 13.1 and 13.2). Supervised learning begins with a set of data collected from cases whose “correct” answer or diagnosis is included in the data set. Based on these training data, the supervised learning algorithm is intended to use the relationships it has learned to predict the answer or diagnosis using data from a new set of cases. For example, if a set of protein levels has been measured in each of a set of patient samples and are shown to be effective predictors of the patient’s classification as “disease” or “no disease,” a supervised learning algorithm may be designed to use protein levels from a new, unknown patient sample and to predict whether this case should be classified as “disease.” In contrast, unsupervised learning begins with data

### BOX 13.1 Unsupervised Learning

- Dimensional reduction
- Principal components analysis (PCA)
- Multi-dimensional scaling (MDS)
- t-distributed stochastic neighbor embedding (t-SNE)
- Clustering
- k-means
- Self-organizing maps (SOM)
- Hierarchical

### BOX 13.2 Supervised Learning

- Linear regression
- Logistic regression
- k-nearest neighbors (kNN)
- Support vector machines (SVM)
- Partial least squares discriminant analysis (PLS-DA)
- Decision trees
- Random forest (RF)
- Boosted trees (Adaboost, xgboost)
- Neural networks

collected from cases *without* any preexisting set of answers or diagnoses. In the unsupervised case, the machine learning algorithm is designed to identify prominent features or patterns in the data that may be of interest. For example, unsupervised analysis of gene expression data has been used to show that diffuse large B-cell lymphomas (DLBCL) reflect two predominant, distinguishable patterns of gene expression.<sup>32</sup> Despite the fact that this difference was not previously identified, the two groups were subsequently shown to be associated with significantly different survival times, confirming their biological and clinical relevance.

### Unsupervised Learning (Box 13.1)

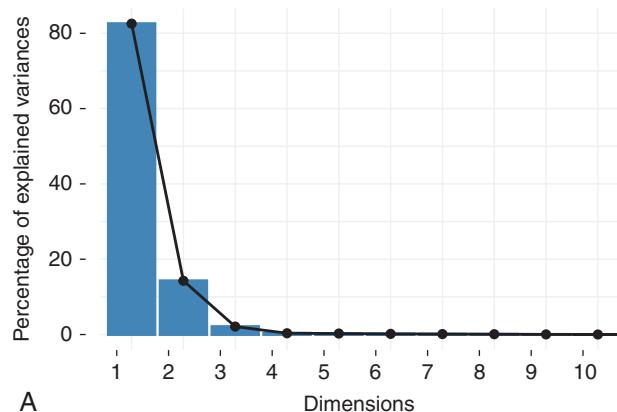
Unsupervised methods may themselves be subdivided into two general groups: dimensional reduction tools and clustering algorithms.

### Dimensional Reduction

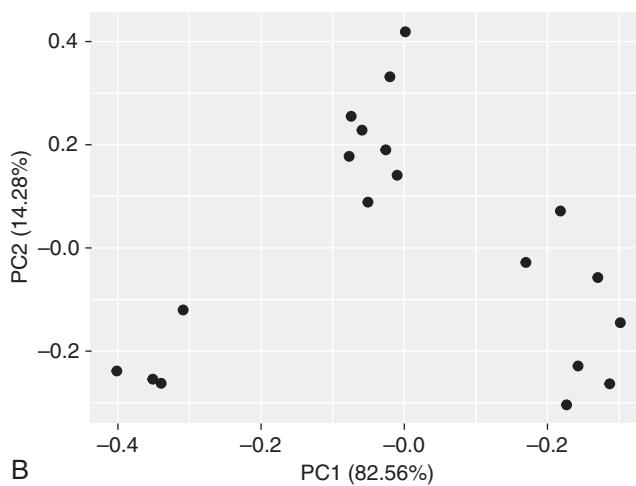
Simple datasets with only two variables can be visualized by plotting on a two-dimensional surface. For example, if a set of serum samples have been analyzed for levels of free thyroxine (T4) and thyroid stimulating hormone (TSH), each case can be represented on a scatterplot to reveal the well-known inverse relationship between these two analytes. In this case, each analyte (T4 or TSH) corresponds to a dimension, and one can be assigned to the x axis while the other is assigned to the y axis. However, when more variables are present, such as with a highly multiplexed assay, it is no longer possible to represent all of the dimensions in a convenient form within the two or three dimensions that the human brain can comfortably assess. Thus for high-dimensional datasets, some form of “dimensional reduction” can be useful. The general principle of these approaches is that predominant relationships between cases that are found in the high-dimensional data are reflected in a two- or three-dimensional representation, while more subtle high-dimensional relationships may be lost.

The most common form of dimensional reduction is principal components analysis (PCA), which relies on capturing the greatest variance in the data set.<sup>33</sup> Briefly, this method first looks in the original high-dimensional space in order to determine how to select the single axis that captures the greatest variance (spread) of the cases. After selecting this first axis, it then determines how to select an orthogonal (perpendicular) axis to the first axis that captures the maximum remaining spread in the data. It continues in this way, selecting axes, until it has created a full representation with as many dimensions as the original data. These axes, called principal components, represent linear combinations of the original features and are the way that PCA reduces dimensionality. Instead of using individual features, a model may now use the principal components as features.

Typically, the results of PCA are represented in two ways. First, a plot of the amount of variation captured by each dimension can be used to determine how many of the dimensions contain useful information (Fig. 13.7A). Because the first dimension always contains the greatest variance and the remaining dimensions are correspondingly lower, this graph always appears as a decreasing series, and it is possible to, for example, determine how many dimensions are required to capture a certain percentage (for example, 70%) of the total variation in the data set. Second, individual dimensions can be selected for plotting in two or three dimensions (Fig. 13.7B).



A



**FIGURE 13.7** Principal components analysis. Nineteen points in 10-dimensional space were processed with PCA. **A**, Graph of the amount of total variance in the data set explained by each principal component. Note that most of the variance in the data set is captured by the first two principal components, which are shown in **B**. **B**, Plot of cases using the values of first two principal components, plotted on the x (Principal Component 1) and y (Principal Component 2) axes. Notice that a clustering (three distinct groups) of the cases is visually apparent in this lower-dimensional representation. (Modified from Haymond S, Julian RK, Gill EL, Master SR. Machine learning and big data in pediatric laboratory medicine. In Dietzen DJ, Wong ECC, Bennett MJ, Haymond S, editors. *Biochemical and molecular basis of pediatric disease*. 5th ed. Cambridge, MA: Academic Press; 2021.)

For example, plotting the first and second principal components (dimensions) on the x and y axis of a graph gives an optimal representation in terms of the amount of variation in the original data set that can be represented in two dimensions. PCA plots are often used to determine (by eye) if there are natural grouping in cases based on the data, or if a given high-dimensional data set contains information that is consistent with previously known or suspected groupings of the cases.

There are several points to consider when using or evaluating the results of PCA analysis. First, the scaling of the original data may significantly affect. For example, if one analyte naturally ranges from 0 to 1000, while another analyte only ranges from 0 to 10, the first will have a greater influence on the variance than the second. One method of addressing this problem is normalizing each variable to a common range prior to analysis. Second, while it is common to display the principal components that explain the most variation (e.g., principal components 1 and 2 for a two-dimensional display, or 1 through 3 for a three-dimensional display), there is no reason to limit the display to just those highest-variance components. There are clear examples in the gene expression literature, for example, where biologically relevant differences between cases that were not apparent in principal components 1 through 3 were clearly seen in principal components 4 through 6.<sup>34</sup> Finally, different principal components may represent different sources of variation, and not all of these necessarily reflect underlying biology. For example, systematic differences between laboratories may be contained in specific principal components, and if these are identified then they may be excluded from visualization or further analysis.

A related technique to PCA, known as multidimensional scaling (MDS), attempts to explicitly maintain, to the extent possible, the “distances” between cases in high-dimensional

space in a lower-dimensional representation.<sup>35,36</sup> The intent of this visualization, which is typically to identify clusters of similar cases, is the same as PCA. The results of MDS are dependent on the way that the “distance” between cases in high-dimensional space is defined. Under certain defined conditions (Euclidean distances between cases in high-dimensional space, utilizing the top principal components), MDS and PCA return identical results. As with PCA, this technique has been used in biomedical applications such as the interpretation of gene expression signatures.<sup>37,38</sup>

A third common dimensional reduction technique, known as t-SNE (t-distributed stochastic neighbor embedding), has been widely used for visualizing the relationship between cell types identified through single-cell sequencing.<sup>39,40</sup> t-SNE uses a Gaussian distribution in high-dimensional space to determine how “probable” it is that two points are close together, and it attempts to create a similar mapping of the relationship between points in a low-dimensional representation. Because it is a nonlinear technique (unlike PCA or MDS), it has some theoretical advantages in ensuring that nearby points in the high-dimensional space are grouped together in the low-dimensional space.

### Clustering Algorithms

A second class of unsupervised algorithms focuses on arranging cases in such a way that prominent groups can be distinguished. These groups are known as clusters, and determining which cases cluster together can identify prominent patterns in the data and potential disease categories in clinical samples. Clustering techniques are often grouped into “top-down” and “bottom-up” methods, depending on whether the number of clusters is defined (“top-down”) or whether the cases are first grouped according to how close they are to each other.

Two prominent top-down methods are  $k$ -means clustering and self-organizing maps (SOMs). In both cases, it is typical to begin by choosing the number of clusters that are estimated to be appropriate for the data set. The  $k$ -means algorithm begins by randomly placing the centers of the  $k$  clusters (centroids) in within the region of the data in high-dimensional space and then assigning each case to the nearest centroid.<sup>41</sup> The centroids are then adjusted to be the mean of the cases that are assigned to them, and this process (assign points to nearest centroid, adjust centroid to be the mean of those points) is repeated until the centroids have reached a stable position (Fig. 13.8A). Effectively, a high concentration of cases in a certain region of space will “drag” a centroid to that region and place it at the mean of those points. Thus the final centroids are representative of clusters of cases in the high-dimensional space, and cases are assigned to clusters based on their nearest centroid.

SOMs are similar in that they allow a specified number of centroids to move through the high-dimensional space until they settle in regions with high concentrations of points.<sup>42</sup> However, a SOM also begins with a relationship between the clusters that allows for the results to be more easily interpreted. For example, if four clusters are chosen to summarize a given set of cases, a SOM might arrange these in a chain (a  $1 \times 4$  configuration). The clusters in this chain not only move toward each other, but adjacent clusters also attract each other to maintain their relative arrangement (by analogy, it is possible to think of a chain of clusters as being connected by pieces of “virtual elastic”). Thus not only are cases assigned to clusters, but adjacent clusters are more similar to each other than nonadjacent clusters. This technique was used in early gene expression profiling work examining the relationship between acute leukemia samples.<sup>43</sup> Using a  $1 \times 4$  SOM, it was shown that acute myeloid leukemia samples were found in one end cluster, while the acute lymphoblastic leukemias were found in the other three clusters. Further, within the three acute lymphoblastic leukemias (ALL) clusters, one cluster was predominantly B-cell ALL, while the other two clusters, which were adjacent to each other, contained T-cell ALL samples. Thus adjacent clusters were shown to be more similar to each other. More complex gene expression datasets have been clustered using two-dimensional SOMs (e.g.  $6 \times 8$  grids), where similar clusters can be adjacent both horizontally and vertically in the connected grid.<sup>38,44</sup>

In contrast to top-down techniques such as  $k$ -means clustering and SOMs, bottom-up clustering begins by considering the relationships between individual cases. The most widely used bottom-up technique is hierarchical clustering, which has been widely used in the literature to organize both cases and their biomarkers into interpretable groups.<sup>45</sup> Hierarchical clustering begins by finding the two closest cases (or biomarkers, when looking at the pattern of expression across all cases) and grouping them. The next most similar pair is then grouped, and so on until the closest case is an existing group. Each case is represented as the terminal branch in a tree structure (dendrogram), and grouped cases are connected in the tree with a height that represents the distance between the nodes (Fig. 13.8B). When existing groups are connected to other existing groups (or to an individual case), there are a number of choices for determining the distance, including the distance between the nearest cases in each group, the distance between the farthest cases in each group,

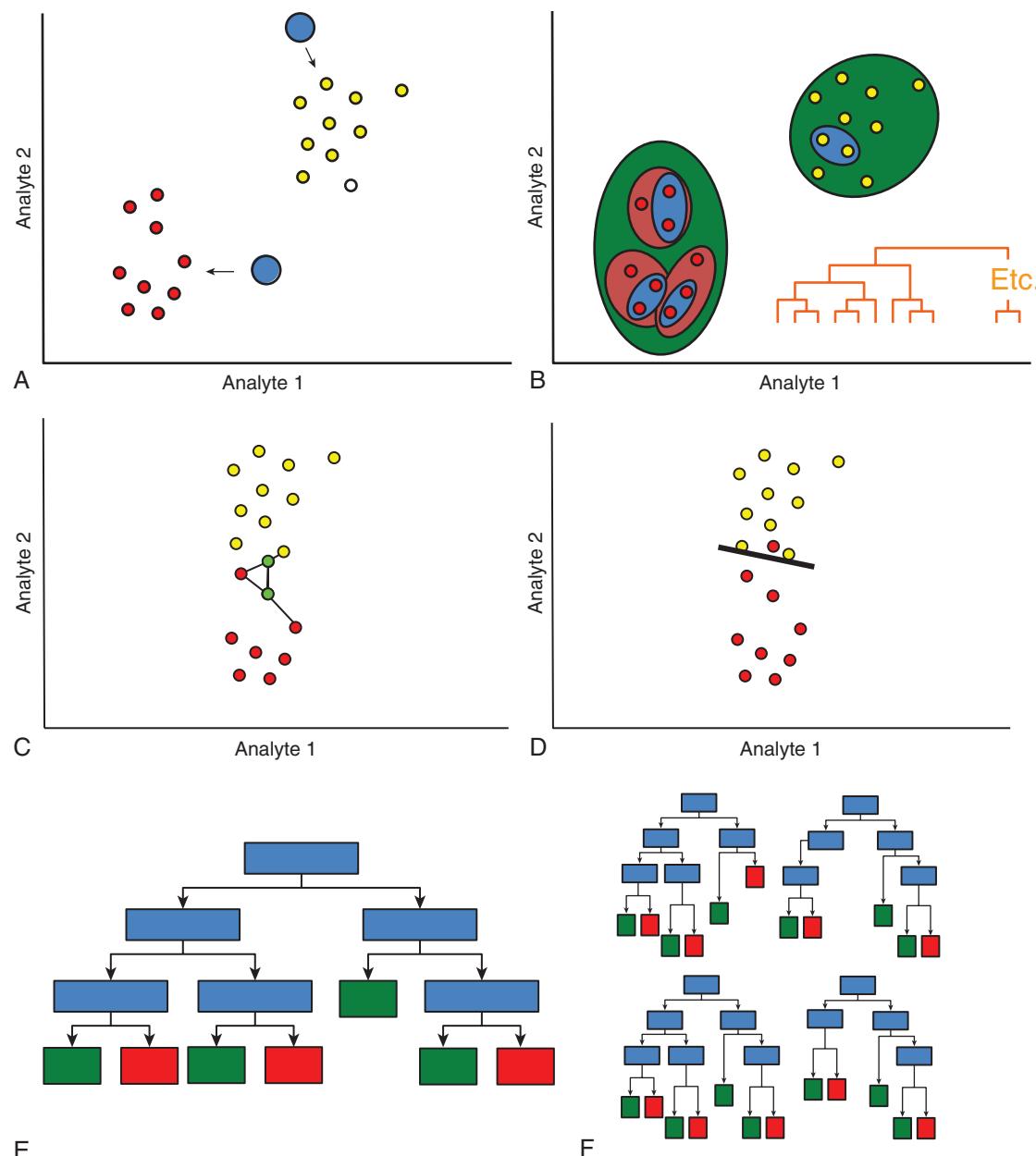
and the difference between the group averages. When the full tree is constructed, it is often possible to look at large, highly separated branches to identify natural clusters of cases in the data. As discussed previously, this technique was applied to gene expression data derived from a variety of lymphocyte and DLBCL samples, and on the basis of the clustering pattern two previously unknown, clinically relevant DLBCL subtypes were identified.<sup>32</sup> It is important to also note that applying this type of clustering simultaneously to both cases (based on the similarity of biomarker patterns) and biomarkers (based on the similarity across cases) will arrange the cases and biomarkers into a format that can be visualized and used to quickly identify patterns of, for example, groups of cases that all have high expression of a set of biomarkers. Such clustered heat maps are commonly used when reporting results of highly multiplexed assays across large and diverse case sets. As with all distance-based clustering methods, the results may vary widely depending on the particular complement of analytes or biomarkers that are utilized for the clustering.

### Supervised Learning (Box 13.2)

Supervised methods are used to create classifiers based on a training set of cases that have a known label. For example, a set of cases that are known to be disease positive or negative may be used to identify a biomarker signature that predicts the presence of disease in an unknown case. A variety of these machine learning algorithms exist, and development of novel techniques has proceeded rapidly over the past 20 to 30 years.<sup>33,46</sup>

Logistic regression is a well-known technique for using multivariable (high-dimensional) data to classify cases into two or more groups.<sup>47,48</sup> While linear regression attempts to find an optimal line to describe the numeric relationship between input variables and an output, logistic regression uses the inverse logit function  $P = \frac{e^x}{e^x + 1}$ , so that a function that goes from  $-\infty$  to  $+\infty$  can map naturally to a probability  $p$  (say, the probability of disease) that ranges from 0 to 1. Like linear regression, logistic regression models can include a number of variables and interactions, and it is possible to assess the statistical significance of any given variable as it contributes to a logistic regression model. Because of this important characteristic, it is very straightforward to interpret the results of logistic regression models and to understand which variables contribute to the output and how (i.e., in which direction and relative magnitude). Interpretable models are particularly important in clinical settings in order to apply medical judgment to the validity of the model.

The  $k$ -nearest neighbor algorithm (kNN) uses a reference library of cases that are directly compared with a test case that we wish to classify (Fig. 13.8C).<sup>49</sup> In multidimensional space, the “nearest” neighbor is the closest (see above for a discussion of “distance”) case in the reference library to the test case. However, simply picking the single nearest reference case will not perform well if there is a region of overlap. For example, if a single, outlier “nondisease” library case exists in a region surrounded by many “disease” cases, we likely want to classify a test case in this region as “disease,” regardless of whether its closest library case is nondisease. Thus we actually use the “ $k$ ” nearest neighbors, where  $k$  is typically greater than 1 (3, 5, 9, etc.). When  $k = 5$ , we consider the five closest cases in the library and let them vote on the classification of the test case.



**FIGURE 13.8** Selected unsupervised and supervised machine learning algorithms shown in two-dimensional space. Cases from different classes are shown in red and yellow. **A**, K-Means Clustering. Cluster-defining centroids (blue) are iteratively moved to high-density regions of space. **B**, Hierarchical Clustering. Nearest cases in space are connected, and the results are encoded in a dendrogram (right). As previously joined cases are found to be closest to other single or grouped cases, this is represented at higher levels of the tree. The height of the dendrogram branches indicate the distance between cases or case groups. **C**, k-Nearest Neighbors. A green test point is interrogated using the k-nearest neighbors algorithm, where  $k = 3$ . Two of the three nearest points are red, so the test point is also said to be in the red class. **D**, Support Vector Machines. In two dimensions, a support vector machine equivalent creates a boundary between classes and can handle overlap between points in the training set. **E**, Decision Tree. Beginning at the top, each blue box represents a choice (such as "Biomarker 1 < 15"), and the correct branch from is decision node is chosen based on the answer. At the end (bottom) of the tree, various classes (represented by green and red boxes in this case) are possible outcomes and may be reported by the classifier. **F**, Random Forest. A random forest is an ensemble of decision trees, each of which is trained using a different random subset of the training data and random set of decision variables. The ensemble of trees (the "forest") votes to determine the reported outcome.

kNN is a conceptually very simple classifier, and it has been used as the basis of a US Food and Drug Administration (FDA)-cleared classifier of autoimmune disease using a panel of measurements of autoantibodies.<sup>50</sup> One potential limitation of kNN, however, is that its classification (particularly at the border between classes) is very dependent on the precise points that exist in the library. Using the terminology that we defined earlier in this chapter, a kNN implementation can have high variance. Nonetheless, its conceptual simplicity is an advantage, and it has been used in commercial clinical laboratory assays.

Support vector machines (SVM) construct a boundary in the high-dimensional space that separates different classes (Fig. 13.8D).<sup>51</sup> A simple analogy in two dimensions would be a line that separates the two classes (while allowing for some overlap in the classes). However, because the cases exist in a high-dimensional space, the separator of an SVM is actually a hyperplane. Further extensions of the SVM also permit the boundary to be curved rather than linear in order to better capture the shape of separation between the cases. As noted in the later section on applications, SVM models are particularly popular and effective for machine learning in laboratory medicine.

It is possible to combine unsupervised techniques with supervised learning in an attempt to simplify overall classification task. Techniques such as PCA(see above) and partial least-squares (PLS) have been used to preprocess data, retaining only a certain number of transformed dimensions prior to processing.<sup>52</sup> Partial least squares discriminant analysis (PLS-DA) extends this idea beyond traditional regression into discriminant analysis, which aims to optimize the separation of cases based on categorical groups. PLS-DA is a well-suited and popular tool for metabolomic analyses.<sup>53</sup> These techniques are examples of a more general principle: namely, that appropriate preprocessing or transformation of data may simplify certain machine learning problems.

Decision trees are a familiar feature in traditional medical diagnostics, and they have been widely used to summarize clinical decision algorithms. For example, a decision tree classifier for simple acid/base disorders might begin by asking whether the pH is less than 7.6, and—if the answer is “yes”—might continue to a subsequent decision node that splits based on the value of bicarbonate and ends with a diagnosis of either metabolic alkalosis or respiratory alkalosis. Similarly, a general decision tree classifier using laboratory values begins at a top node that asks about the value of a single biomarker and then splits into branches that are based on the value of other biomarkers, finally ending in one or more different diagnoses (Fig. 13.8E). In this way, an arbitrary high-dimensional separation between different diagnostic classes can be encoded in a decision tree.<sup>54</sup> There are efficient algorithms for recursively training a subset of decision trees, and in fact decision trees can be used to encode both classification problems and regression estimates (thus decision trees are often referred to as CART, “Classification and regression trees”).

The ease of training a tree, along with its human-readable interpretability, are significant advantages to CART approaches. Furthermore, the familiarity of decision trees within a medical context makes them a natural choice for certain problems in the clinical laboratory. However, there are several disadvantages to decision trees. First, complex CART trees have a tendency to overfit the training data set,

and thus it may be useful to artificially limit the size and complexity of a tree. Second, trees have high variance and may be sensitive to the specific training set used.<sup>55</sup> One way to address this issue is to train a number of different decision trees using different random samples of the original training data. The different subsets of the training data are known as “bootstrap” samples, since they are created by sampling with replacement (i.e., a sample may be drawn more than once). In this example of a so-called “ensemble” method, the group of trees are each allowed to classify a new case of interest, and the vote of the full ensemble determines the classification of the new case. By training ensemble trees, each one using a different bootstrap sample, we ensure that a single feature of the full training set is less likely to dominate the full ensemble of trees, thus lessening the likelihood of overfitting. This technique is known as “bagging” (named for “Bootstrap AGgregation”).<sup>56</sup>

An extension of bagging leads to the widely used “random forest” technique (Fig. 13.8F).<sup>57</sup> While bagging trains an ensemble of trees based on all the variables from a bootstrap sample, random forests extend this idea also randomly choosing only a subset of variables to consider for node in the tree. Thus the random forest further ensures the diversity of trees that make up the forest (ensemble). Random forests have been shown to be robust classifiers, and by training each tree on a subset of the cases it is possible to estimate their performance by assessing the error rate if that single tree were used to classify the rest of the training cases that were not used to construct it. By combining these error rates from all the trees in the forest, the “out-of-the-bag” (OOB) estimate provides a sense of the performance of the random forest prior to its actual validation with an independent test set of data.

Because of its robustness to overfitting, OOB error estimate, and overall strong performance, the random forest has been successfully used in a number of medical contexts, including disparate areas such as the diagnosis of acute kidney injury using clinical and laboratory data and the detection of myelodysplastic syndrome using a traditional hematology analyzer.<sup>58,59</sup> However, there are drawbacks to random forests and other ensemble methods. Specifically, although it is easy to interpret any individual decision tree within the forest, it is not straightforward to determine the behavior of the full ensemble of trees. Random forest packages commonly include tools to determine the overall importance of a given variable by determining the extent to which removing or changing the variable would alter the output of the forest. However, the full lack of human interpretability of the model may still be an issue in certain clinical applications, and determining the validation required for clinical use is an important consideration.

More recently, another extension known as “boosting” has emerged.<sup>60</sup> Rather than independently training each tree on a subset of the data, boosted trees concentrate on fitting portions of the data that have not been well fit by earlier trees. In this way, new trees that are added to the ensemble can focus on reducing the remaining error in the model. Various implementations of boosting are available, including AdaBoost and xgboost.<sup>61</sup> Of note, boosting has been highly successful in a variety of contexts and has become a default “best choice” for many classification and regression problems similar to those most commonly encountered in the clinical laboratory.

Neural networks are named for their conceptual similarity to a simplified version of a network of biological neurons.<sup>62,63</sup> In the same way that a neuron has many axons that receive input signals and a dendrite that delivers an output signal, a computational neural network has multiple inputs, a set of weights that control how those inputs are used, a way of calculating an output based on the inputs and their weights, and a method to deliver that output to many other virtual axons. Typical neural networks, which are the basis of “deep learning” approaches, use many layers of interconnected neurons that transform the complex, high-dimensional input signal into a classification or regression output. During training, the error between the output of the network and the correct training answer is “backpropagated” through the network, such that the weights on inputs are changed. Each layer of the trained neural network abstracts certain features of the data, and subsequent layers of the network can use these transformed representations to build increasingly high-order representations of features of the original data. Deep learning approaches have become more straightforward with the advent of efficient, open-source libraries and interfaces such as TensorFlow, Theano, and Keras.<sup>25,64</sup>

While deep learning and neural networks have been deployed in many contexts, they are primarily used for the analysis of image data and textual data. Most notably, this technique forms the basis of many automated analysis packages that are designed to classify histologic and radiologic images. As described above, different layers of a neural network can create different transformations and abstractions of the data. For example, the first layer of a neural net used in image analysis might, upon inspection, perform edge detection. Edges in this layer might be grouped into high-order constructs in subsequent layers, and so on until a full diagnosis or identification is made. These abstractions arise naturally out of the training of a neural network and do not require specification by the user. However, it is often difficult to interpret the precise way in which a complex deep learning model is using data at each layer, and neural networks represent a classic case of the “black box” classifier that is diagnostically effective but not amenable to human interpretation. These characteristics must be taken into account when validating a deep learning model in the clinical laboratory in order to ensure that a sufficient variety of cases have been examined and that common (or known but uncommon) artifacts do not confuse the classifier.

## ADDITIONAL PRINCIPLES AND LIMITATIONS OF MACHINE LEARNING

### Distance

Previous sections described the importance of measuring error between predicted values and training values for both regression and classification. In these supervised learning methods, either the correct  $y$  value for a given  $x$  is known, or the correct class  $c$  for a given  $x$  is known. A common measure of the error in regression problems, and by extension, classification problems is the root mean squared error (RMSE). At the heart of this equation is Euclid’s measure of distance. RMSE corresponds to the Euclidean norm. It is also called the L2 norm. When outliers are a significant component of a data set, Euclidean distance is not always a useful measure of error.

An alternative is called the mean absolute error (MAE), which corresponds to the L1 norm and is sometimes called the Manhattan norm because it measures the distance between two points as if you could only move between points laid out in a grid-like fashion like you would in a city. Norms can be generalized as L $K$ .

The index of the norm is not limited. The higher the value of  $K$ , the more the distance depends on large values than it does on small ones. Thus RMSE is more sensitive to outliers than MAE. However, when working with data that are normally distributed, outliers are exponentially rare, and RMSE works extremely well and is very commonly used.

Distance measures also play an important role in  $K$ -Nearest Neighbor algorithms, as well as all unsupervised algorithms where the distance of a data point determines membership in a cluster to the center of a cluster. Interestingly, distance measures are usually described as dissimilarity measures in the context of clustering applications.

For categorical variables, it is common to give a distance of 1 if the features are different and a distance of 0 if the features are the same. This approach is called the Hamming distance.<sup>65</sup>

Distance will come up again when considering methods for avoiding overfitting, and in datasets with many features. The goal is often to reduce the number of features (dimensions). Two popular methods for reducing the overfitting seen in linear regression is called regularization. Regularized models constrain the weights assigned to features to “smooth out” the model. Two common regularized models are the ridge regression and the Lasso (least absolute shrinkage and selection operator regression). Both methods add an extra term to the cost function used to minimize error. Ridge regression adds an L2 norm distance term of the coefficients to the cost function.

In contrast, the Lasso adds an L1 norm distance term of the coefficients to the cost function. The goal of these extra terms is to constrain the values of computed coefficients, and they differ primarily in the selection of their distance measure. The practical effect of constraining coefficient values is to drive the coefficient (sometimes called weight) of unimportant features to zero. Lasso (because it is an L1 norm) does this aggressively, effectively turning a high-dimensional model into a low-dimensional model. Ridge regression is more sensitive to outliers (as described above) and allows more features in the model and increases the dimensionality of the model. As will be described in the next section, distance metrics (especially Euclidean L2 norms) can fail at high dimensionality. If manual feature engineering and selection can be done effectively, the dimensionality of the model can be reduced using the context of the problem. If not, then choosing the right distance measure for model coefficient constraints can be extremely helpful.

### The Curse of Dimensionality

The difficulties with datasets with a high number of features are so problematic that professionals in the machine learning community have gone so far as to declare the situation cursed. The challenges must be extraordinary to have received such a name. The truth is actually much worse than one might imagine.

First, having a high number of features means that even if you have an exceptional number of cases (many more than features), you may have a true *big data* problem on your hands, and need infrastructure far beyond most computing

environments. However, today, this is a simple technical challenge, and there are techniques described in the infrastructure section to help. The real problems with high-dimensional data are much more insidious.

To start with, think about training a model for gene expression data. If you have a database with 1000 features, which are binary (0 meaning no expression or 1 representing expression), then for any single case, there are 21,000 possible combinations. Even if there were 1012 rows in a database, it would only contain  $10^{-286}$  of 1% of all possible examples. In effect, your training algorithm will have a sample size so close to zero that training a real model is impossible. In the prior section on training and validation some practical tools for reducing the number of dimensions based on statistical analysis were discussed, and there are also feature selection techniques based on correlation analysis combined with cross-validation, which can be helpful.

There is another very serious problem with high-dimensional data. It is highly nonintuitive because humans have no way to develop an intuition regarding more than the three (or four, if you count time) dimensions we experience every day. The problem arises from the way distance is measured, described in the previous section. The Euclidean L2 norm has a property that makes every data point the same distance from every other point when the number of dimensions exceeds approximately 10 to 20. That means regression, classification problems with more than 10 or 20 features that rely on the L2 norm will measure all distances to be the same. The L1 norm behaves better at higher dimensions, and even fractional norms have been proposed, which provide even better performance at high dimensionality.<sup>66</sup>

The failure of the Euclidean norm to provide meaningful information above 20 or so dimensions is related to another problem with high-dimensional space: every datapoint is also very far from every other datapoint. Picking two random points in a unit square will produce an average distance of about 0.5. Doing the same experiment in a unit cube will give an average distance of about 0.7. However, if you pick two random points in a 1,000,000-dimensional hypercube, they will give an average distance of approximately 400.<sup>25</sup> This means that high-dimensional datasets are very likely to be very sparse. When you add a new case, it is likely to be very far from all of the training cases, which will make any predictions based on the training set much less reliable. All of this adds up to the fact that the higher the number of dimensions, the higher the chance of overfitting.

## Sparse Data

The issue of sparsity in data happens not only in high-dimensional situations, it can also happen at low dimensionality. The issue dates back to the seventeenth century, when the philosopher Hume challenged the logic of predicting the future based on the past. This leads to an unsolved problem which is sometimes called the *Black Swan Paradox*.<sup>67</sup> It is a notable problem since all machine learning algorithms predict the future based on the past (that is, on known data).

The issue of low numbers of observed cases can be thought about in terms of Bayesian decision theory.<sup>68</sup> An informal statement of Bayes formula is:

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

For classification, the term *posterior* simply means the probability of a case actually being in a particular class given the values of the features observed. The term *likelihood* is used to express how likely it is that the set of feature values would result given that the case belongs to a particular class. The larger the likelihood, the more probable it is that the case belongs to that class. A key term in the formula is the *prior* which is the probability that, with *no new information*, the case belongs to a particular class—in other words the distribution of a particular class in the entire population. Interestingly the *evidence* term, which is the probability of seeing particular feature values, acts as a scaling factor to make sure that the value of the posterior is between 0 and 1 which is required to make it a valid probability. It should be clear that the prior is critically important to the prediction.

The key assumption in machine learning is that the true distribution of classes can be estimated from the training data, and that this can be done reasonably. In other words, an estimate of the prior probability of a class occurring in the overall population can be computed and used to calculate the probability that the new case is of class Y, given the observation of feature X. Since it is almost never possible to know the true prior probability of observing a class, an estimate has to be made using training data.

The issue with sparse data is it becomes very difficult to estimate the prior probability using a small training set. If the prior probability has a high error, then performance of a classifier will be poor. The best estimate for the prior probability is the maximum likelihood estimate (MLE). One way to gain an intuition for the problem is to consider estimating the outcome of flipping a fair coin. This situation is a two-class problem ("heads" or "tails"). The prior for the outcome "heads" ( $x = 1$ ) is the same as the outcome "tails" ( $x = 0$ ), which is 0.5.

The mathematics can become tedious,<sup>69</sup> however the most common estimate for the posterior probability for the toss of a fair coin is to use the *mode* of the MLE:

$$\text{posterior} = \frac{N_1}{N_1 + N_0}$$

If a coin is flipped three times ( $N = 3$ ), and all three results are tails ( $N_1$  is number of times heads shows up, and  $N_0$  is the number of times the result is tails, and  $N$  is the total number of tosses) then  $N_1 = 0$  and the *posterior* probability is 0/3, suggesting that the probability of getting heads on any future flip is 0, which cannot be right.

One way classification algorithms can be improved when dealing with small numbers of cases is called *add-one smoothing*. The idea is to simply change the estimate from the *mode* of the MLE to the *mean* of the MLE. By using the *mean*, the probability of heads in a single future trial is:

$$\text{posterior} = \frac{N_1 + 1}{N_1 + N_0 + 2}$$

For our example of three tosses of a fair coin resulting in all heads, *add-one smoothing* gives a probability of 1/5 rather than 0. Using the mean of the MLE rather than the mode gives a better estimate for small sample sizes. It is the justification for adding 1 to empirical counts, normalizing, and then using those values in the posterior calculation.

This theoretical discussion is intended to help when working on classification problems where data partitioning can

result in small sample sizes even from *big data* sources. An example is applying a filter like “the number of times a specific person has a specific health care event.” Filters like this can reduce the sample set to a size where Bayesian approaches may be the only way to achieve reasonable predictions.<sup>67</sup>

### Interpretability

When an algorithm makes a prediction, it is often important to be able to explain why the prediction was made and how confident the predicted result is. In linear regression, for example, it is easy to understand why a particular  $y$  value was predicted from an input  $x$ . The same is true for many classical statistical machine learning algorithms for both regression and classification. There are, however, many machine learning algorithms that have more predictive power but lack explanatory power. These so-called black-box algorithms are increasing in predictive power and are increasingly being used. The first models to fall into this category were multi-layer artificial neural networks (described in the section on common machine learning algorithms). In the case of artificial neural networks, it is possible to make excellent predictions, but the *reason* for a specific output cannot be determined by looking at the weights assigned to the neural connections. Artificial neural networks have become so deep (using many layers) and have such complicated architectures that any idea of interpreting how the system arrived at a specific conclusion has been sacrificed for predictive power. Even extremely easy to interpret models, like decision trees, are now the basis of powerful ensemble methods like Boosting, Random Forests, and Gradient Boosting. These ensemble methods are very difficult to even visualize, let alone describe why they give specific outputs. Even models that use linear decision boundaries, like Support Vector Machines, are often coupled with kernel methods. Kernels transform data in  $n$ -dimensional space into a much higher-dimensional space. In this higher-dimensional space, a linear boundary is computed and transformed back into the original space as a nonlinear boundary. The nonlinear boundary computed for the original space can be visualized, but the meaning of the higher-dimensional space cannot be interpreted.

This is not to say that these new, complex models cannot be checked for computational correctness, but simply using the parameters of the model to understand how predictions are made is difficult and an active area of research and discussion in the machine learning community. Especially in high-stakes predictions, stakeholders need to know why a model can be trusted.

When building a machine learning system, the context in which it is used must be considered. If explanatory power is a critical factor, then black-box models are at a clear disadvantage and may even be prohibited from use. In cases where black-box predictive power is valued over a lower-performing, transparent model, there are techniques for getting information on both global and local factors that can help explain how a black-box model works. *Explainers* perform various tests on black-box models and produce detailed information on both a global (whole model) and local (case) level. Global information includes the importance of features, how a change in a feature affects the prediction, and how features interact with each other. Local information is explored to understand how a specific case was treated by the model. These local factors may use either simplified versions of

the overall model to explain a decision boundary or may visualize parts of the overall model that most impacted the prediction.

Tools for describing global information exist for most machine learning algorithms but analyzing the impact of predictors on individual cases is more complex. One tool for interpreting black-box models is LIME (Local Interpretable Model-agnostic Explanations).<sup>70</sup> The basic idea behind LIME is to find an interpretable model that is *locally faithful* to the classifier. LIME makes the simplifying assumption that a linear model can be used in the local region near an individual case. It does this by generating a linear *surrogate* model from cases selected from the region around the case being tested. Because the LIME algorithm produces a linear surrogate model, this model is inherently interpretable, which is extremely useful. LIME does have the weakness that the linear model produced might not represent the region near a particular case very well; thus local accuracy might suffer. Further, in the LIME algorithm, there is no guarantee of obtaining a unique value for a case, depending on how much of the region around the case is sampled.

Another approach to building an interpretable surrogate model is the SHAP (SHapley Additive exPlanations) algorithm. SHAP uses the Shapley regression value developed in 1953 by Lloyd S. Shapley.<sup>71</sup> Shapley regression is a way of measuring the importance of contributions from multiple actors in cooperative games. By treating predictors as contributors to a final result, Lundberg and Lee used the approach to create SHAP values for model interpretation.<sup>72</sup> SHAP uses the Shapley regression value to evaluate the difference in the prediction for a given case with and without each predictor. In Shapley regression, all permutations of predictors are considered, and the result of each prediction valuation is then used to compute the SHAP value. This makes versions of SHAP (TreeSHAP) particularly useful for tree-based methods since the order of predictors in a tree changes the output. Covering all permutations is time-intensive, but the SHAP algorithm results in the same type of linear surrogate as LIME. SHAP, however, guarantees local accuracy and a unique value for the effects of each predictor on the classification of a case.

## COMPUTATIONAL INFRASTRUCTURE

As the adoption of machine learning increases in clinical laboratories, so will the need for computing technologies and software tools that support its development and pipelines for implementation.

### Computing Technologies

Computing options include personal workstations and servers. Either may be deployed physically within a local IT system (“on premises”), on a private cloud environment managed by local IT (“cloud-hosted”), or on a public cloud environment managed by the cloud provider (“cloud-based”). Laptops, personal computers, and virtual desktop machines are types of local workstations that are routinely used for development work. Compared to workstations, servers have specifically designed operating systems, faster processor speeds, and more memory and storage capacity. Cloud computing is the web-based delivery of servers, storage, databases, and other applications in a flexible manner dependent on user demand.

Servers (on-premises or cloud) configured for distributed computing are most effective for performing large-scale machine learning experiments and for running models in production. In distributed systems, multiple, connected computers work together on the same task in order to enable computing processes to occur independently and in parallel, maximizing efficiency. This can be further extended with parallel processing, where two or more central processing units (CPUs) are deployed simultaneously to handle individual components of an overall task. Another type of computing engine, graphical processing units (GPUs), have become critical for the advancement of deep learning applications. Google developed their own version of accelerated parallel processors, known as tensor processing units (TPUs).<sup>73</sup> GPU and TPU hardware components have been purposely configured for parallel processing, and they excel at efficiently performing the matrix operations and calculations required for deep learning.

### Software Tools and Frameworks

Machine learning workflows are reliant on various software frameworks and computer programming scripts that are commonly connected through application programming interfaces (APIs). The most popular coding languages for data science and machine learning are R and Python. Both are available in free, open-source versions. These languages are often used with software libraries, toolkits, or user interfaces that aid in executing the machine learning process and facilitating reproducible workflows.<sup>74</sup> For example, the Scikit-learn machine learning library is considered a fundamental tool for Python programmers doing predictive modeling.<sup>75</sup> Similarly, R programmers commonly rely on the caret package and the more recently developed tidymodels suite of packages for machine learning.<sup>76,77</sup> Keras is a widely used library for specifying and training deep learning models.<sup>78</sup> It was developed for use in Python, but also now has an R interface. Popular among students and beginners is Weka (Waikato Environment for Knowledge Analysis), an open-source graphical user interface (GUI)-based application that requires minimal code to perform machine learning.<sup>79</sup> TensorFlow and PyTorch are open-source machine learning platforms designed to facilitate rapid experimentation and scale the production of building complex models, including deep learning, while providing highly flexible system architectures.<sup>80,81</sup> TensorFlow and PyTorch underlie many commercial machine learning applications used by companies such as Google, Facebook, IBM, Twitter, and Uber. Amazon, Microsoft, and Google also have machine learning applications (Amazon Machine Learning, Azure ML Studio, Google Cloud ML Engine) that provide user interfaces for building, training, testing, and deploying models by leveraging APIs to and from data sources, distributed computing platforms, machine learning libraries, and other resources.

### Machine Learning Pipelines

Machine learning development may be conducted on an individual workstation or within a “sandbox”-type computing cluster environment. Deploying a machine learning model to production involves additional considerations, including the architecture of a machine learning pipeline. In production, the code describing the machine learning model is a single, relatively small component of the necessary infrastructure

and code base.<sup>82</sup> Considerable resources are devoted to collecting input data, verifying it, and extracting features from it. Other tools are needed to analyze data and manage and automate workflows. A serving infrastructure is required to put the predictions into use. Finally, the components of the system must be monitored for reliability, efficiency, speed, cost, and accuracy, for example.

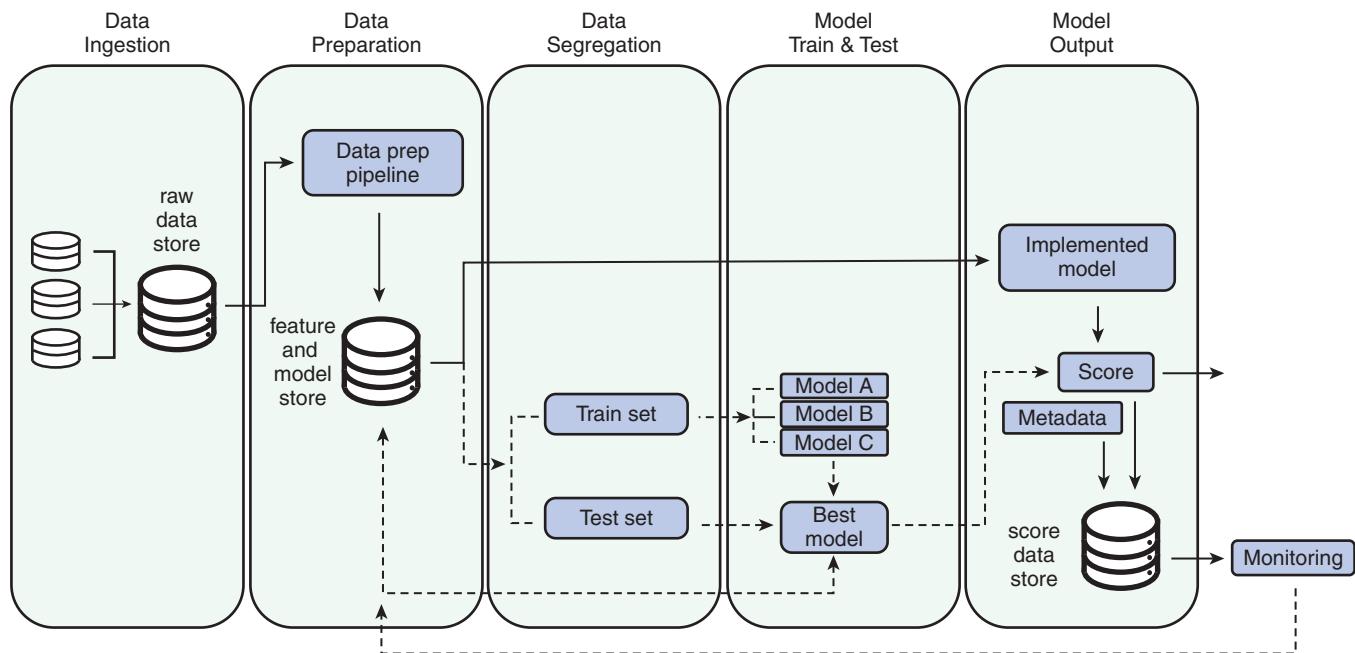
Pipelines may be configured for batch or real-time modes for both prediction and learning. In batch prediction mode, data are periodically collected (based on a schedule or a trigger) and multiple records are sent for processing through the pipeline to provide a set of predictions that will be used at some point in the future. In real-time predictions, results and features are time-sensitive, and the pipeline must operate on a faster scale, processing each record on demand. The deployed model can be static or dynamic, depending on how it learns. Static algorithms use offline learning, meaning these models are trained periodically using a batch of data, commonly on a predefined schedule or in response to performance issues. Once the model is trained and optimized for production, it is no longer updated unless it is re-trained. In contrast, a dynamic (adaptive) model uses online learning and continually updates as new data enters the system and feedback from the prediction results are incorporated into the model. Static models are easier to deploy and maintain than dynamic models. Both approaches require post-implementation monitoring, as even a static model is subject to time-related changes in input data and the overall code base. An example diagram for a machine learning pipeline is shown in Fig. 13.9.

The first step in the process is data ingestion. Incoming data in its original form is brought into a data store, creating an immutable record of a raw dataset. Optimal selection of the data store type and configuration of parallel and/or distributed workflows contribute to the data processing speed. The next step is data preparation. This step involves critical processes of data exploration to assess the condition and validity of the data, data transformation to correct any issues and prepare it for modeling, and feature engineering to generate the required inputs for modeling.<sup>83</sup> Prepared data containing the selected features goes to the algorithm for scoring (prediction) by the deployed model and then to the serving infrastructure for presentation of results to end users. Scores and associated metadata are sent to a data store for use by a monitoring service. Prepared data and scores may also be sent for model training and evaluation for either online or offline learning systems. Underlying these steps are systems for scheduling and orchestrating tasks within jobs, logging metadata, and facilitating data requests and transfers.

### REGULATORY ENVIRONMENT

Machine learning can revolutionize clinical chemistry through improvements in research, assay design, data analysis, and decision support. However, like all the tools used in the clinical laboratory, machine learning software must be evaluated in the context of the applicable regulations. There are several key considerations when developing machine learning software in health care.

- Sources of data: privacy regulations, sources of potential bias
- Algorithm type: locked-down models, adaptive models
- Risk: the need for interpretation, verification, and validation



**FIGURE 13.9** Schematic representation of an example machine learning pipeline in production. The black dashed lines represent additional processes and feedback used for an adaptive model design or for the offline, intermittent development and validation of an implemented model. The initial development and validation of the implemented model are not shown. (Modified from Haymond S, Julian RK, Gill EL, Master SR. Machine learning and big data in pediatric laboratory medicine. In Dietzen DJ, Wong ECC, Bennett MJ, Haymond S, editors. *Biochemical and molecular basis of pediatric disease*. 5th ed. Cambridge, MA: Academic Press; 2021.)

The US FDA has published a detailed discussion paper,<sup>84</sup> which describes a framework for evaluating machine learning systems as medical devices. Laboratories producing results that are used in the practice of medicine in the United States are under the jurisdiction of the US FDA and covered by the Clinical Laboratory Improvement Amendments (CLIA) law. Meeting the requirements for CLIA certification involves validating systems used for the production of laboratory results. Systems that are either cleared or approved by the US FDA still require validation to ensure that the system (software and hardware) are fit for purpose and operate properly in the laboratory environment in which they are used. This section will highlight key issues and summarize the FDA's current and proposed regulatory framework.

Machine learning systems are ultimately software programs which run on some form of computing hardware. These can range from embedded systems controlling cardiac devices to large-scale decision support systems running on cloud computing infrastructure. Currently, the FDA treats most machine learning applications that are part of a medical device as a single system. A new software category has recently emerged: Software As A Medical Device (SaMD). These are programs which evaluate data from a signal or an image or are applications which interact with patients and make decisions or recommendations based on user input.

The current position of the FDA is that once a machine learning model has been trained, it is treated the same as any other piece of software used in the delivery of health care. There are validation requirements and depending on risk, premarket and possibly postmarket surveillance requirements. However, as long as the model cannot be changed without a software update, it is evaluated by the same rules as other software applications.

Because training machine learning systems involves using data, there are steps that need to be taken to ensure that privacy laws are followed. For example, facial recognition systems trained from large-scale databases of images, especially of children, taken without the consent of the people photographed, have recently come under scrutiny and even generated legal action against the software developers.<sup>85,86</sup> Issues with gender, ethnicity, and other forms of bias in training data have also been found in machine learning systems. These problems have led to calls for more careful consideration of how training datasets are assembled.<sup>87</sup> All machine learning development should take into consideration privacy, bias, and training set balance issues, but in some situations, those steps are a legal or regulatory requirement.

A new issue being faced by the FDA is the development of adaptive algorithms. Unlike locked-down models, adaptive systems update weights or other model parameters based on data collected while operating (see Fig. 13.9). The FDA has proposed treating adaptive machine learning systems as having an automated change process designed to achieve a specific clinical goal. This automated change process would be treated the same way a manual change process is. In other words, if a machine learning system developer decides that better outcomes can be achieved by retraining a model, for example, that change process would have to be validated and documented. An automated update system would also have to be validated and documented. If a given set of inputs generated a given set of outputs at the time of clearance, then any change in the outputs after clearance by a process would have to be evaluated and determined to either be accepted or not. The same would hold for adaptive systems. If the risk of the adaptation or the adaptation *process* itself is deemed to be too high, then the device would not be approved.

The FDA has cleared multiple machine learning-based systems over the past several years. They have made great strides to improve and speed the pace of both rulemaking and device evaluation. The FDA, and all medical device developers, must balance innovation with patient safety. It is clear that it is in the broad public interest to make the best use of new software technology as rapidly as safely possible to improve health care outcomes for patients.

## APPLICATIONS

The opportunities for machine learning in laboratory medicine are vast. Algorithms can be used for prediction in both clinical and operational decision making. This section describes recent examples of machine learning in laboratory medicine. These selected examples demonstrate the variety of problems that have been addressed with machine learning approaches and the potential for these algorithms to improve laboratory medicine workflows.

In medical decision making, machine learning has been applied for identifying at-risk populations and aiding with diagnosis and prognosis in clinical scenarios. It is common for laboratory measurements to be included as features in such models. Early warning models for sepsis and cardiac events and predictors of readmission risk are among the first types of machine learning algorithms to be implemented within hospital systems. Other artificial intelligence systems have been implemented in pathology and radiology departments to augment image review workflows. At this time, there are several examples of such tools that have been approved or cleared by the US FDA for use as clinical diagnostics, several following the regulatory framework described in the section above. There is strong interest in applying machine learning to critical care, given the complexity of illnesses and the large amount of data generated from multiple monitoring systems used in intensive care units. These data have been integrated with other clinical data and structured for mining and incorporation into models to predict conditions such as sepsis, acute kidney injury, and circulatory failure.<sup>88-90</sup> Similar efforts are underway in every specialty and for a wide variety of problems in health care.

More specifically, as mentioned throughout this text, there is great opportunity for machine learning within laboratory medicine. This chapter has discussed many of the types of data routinely available from clinical laboratories and the attributes that make them suitable for machine learning. Among these are large, retrospective datasets integrating laboratory results, patient demographics, and clinical labels. Such datasets have been used to predict test results from other values in an effort to reduce redundant testing, specifically due to similar tests in multi-analyte panels. Lidbury and colleagues focused on liver panel testing and used decision trees and support vector machine models to predict (overall accuracy 74.5%) if gamma-glutamyl transferase (GGT) values would be within or above the reference interval, dependent upon alkaline phosphatase (ALP) and alanine aminotransferase (ALT) values.<sup>91</sup> This highlights concerns about the utility of such multi-analyte panel tests, given the redundancy of information provided to requesting clinicians when ordering liver function panels versus individual tests, and it suggests that machine learning may be applied to improve test utilization in such cases. A similar study looked at iron panel testing and

showed ferritin values and their classification (normal vs. abnormal) could be accurately predicted from patient demographics and other iron-related laboratory tests.<sup>92</sup> Following a random forest approach for missing value imputation, Lasso regression showed a correlation of 0.729 for predicted vs observed ferritin values, and logistic regression yielded the best classifier with AUC = 0.97.

Other utilization-related efforts have used machine learning models to triage test orders based on the expected diagnostic value determined from associated multivariate data. Zhang and colleagues found decision tree and logistic regression models could be applied to decrease unnecessary utilization of peripheral blood flow cytometry by 35 to 40% at their institution.<sup>93</sup> Using patient history of hematologic malignancy and available parameters from complete blood count and differential tests, both algorithms predicted whether or not the peripheral blood flow cytometry results showed an abnormal blast or lymphoid population (B or T cell). Richardson and Lidbury investigated basic and ensemble decision tree classifiers to predict results of hepatitis B virus surface antigen and anti-hepatitis C virus (HCV) antibody immunoassays from demographics and other laboratory results.<sup>94</sup> Analysis of variance of mean accuracy rates showed dependency upon the test outcome (positive or negative) and on interactions between the classifier method and the outcome and between the virus and the outcome.

Retrospective laboratory datasets have also been used to train models to aid in autoverification and quality control processes during result review. Autoverification of results has become common in clinical laboratories, particularly those with high throughput. These systems are typically built on a rules-based logic framework that examines single or multiple factors to assess result quality and release or reject results. Demirci and colleagues evaluated the feasibility of training a neural network to classify validated versus rejected results compared to a Boolean logic-based autoverification scheme.<sup>95</sup> Compared to a laboratory specialist's decision to verify a result, the sensitivity of the model was 91% and specificity was 100%. A support vector machine-based multi-analyte model outperformed single-analyte delta checks for detecting wrong blood in tube errors in a simulated data set, achieving AUC of 0.97.<sup>96</sup> In addition, results of this study suggest that such a model may be useful in identifying patient events that lead to clinically significant changes across sets of test results. Some laboratory tests require review of instrument data and machine learning may be applied to streamline these decisions as well. Mass spectrometry-based assays require time-consuming manual review of data to identify quality-related issues prior to reporting results. In an attempt to automate this process, Min and colleagues developed a support vector machine algorithm using parameters collected for each sample during a LC-MS/MS run (e.g., ion ratio, concentration, peak shape features) to identify analytically unacceptable results.<sup>97</sup> The four-feature classifier yielded 100% recall and 81% precision and reduced the manual review requirement by about 87% in their test set.

Another promising application of machine learning is augmenting interpretation workflows in clinical laboratories. As test panel interpretation efforts are largely dependent upon recognizing patterns within multivariate data, machine learning models are well suited for such problems. This proof of concept was demonstrated using a weighted-subspace random forest binary classifier to predict the interpretation of

"No significant abnormality" or "Abnormal profiles" for urine steroid profile tests. The best performing model had a mean area under the ROC curve of 0.955 (95% CI, 0.949 to 0.961).<sup>98</sup> In the same study, a multiclass classifier did a reasonable job predicting the individual abnormal profile interpretation with a mean balanced accuracy of 0.873 (0.865 to 0.880). This type of approach could be applied across various assays where a manual review of multi-analyte data yields a professional interpretation of the results, including organic acid and amino acid panels, serum protein electrophoresis, and hemoglobin electrophoresis.

Data from next-generation sequencing assays require processing via bioinformatics pipelines to produce variant call format files. These pipelines perform alignment of the sequencing reads with detection and annotation of variants to provide a heuristically filtered list of potential variants for review by a genomic specialist for final classification and reporting. It is common practice to combine results from multiple variant callers to increase sensitivity; however, this results in a large number of variants requiring review for quality and clinical relevance. Whether from constitutional or somatic genetic tests, the complexity and size of variant files generated in clinical laboratories have surpassed capacity for efficient review. Deep learning applications for genomic analyses have been recently reviewed.<sup>99</sup>

Machine learning methods have been applied to reduce the time, cost, and subjectivity of genomic analyses and to improve their scalability. A random forest three-class (real, artifact, and uncertain) model was developed to systematically identify valid variants from artifacts in pediatric non-FFPE tumors.<sup>100</sup> The model demonstrated 100% specificity and 97% sensitivity, definitively labelling 96.6% of the variants in the test set, exempting them from manual review. Implementation of the optimal model enabled the laboratory to reduce the turnaround time and increase variant review capacity by 42% with the same number of genome scientists. Random forest and logistic regression have also been applied in genomics workflows to automate pathologist decision making in variant reporting.<sup>101</sup> The authors also highlight several important features in user interfaces for presenting machine learning model results, including transparency to the underlying reasoning for a given prediction. The importance of interpretability and two commonly used approaches (LIME and SHAP) to improve model transparency were discussed in a section above. Another study in cancer genomics showed a deep learning model could accurately classify variants when compared to labels from manual review and to results from orthogonal validation sequencing.<sup>102</sup> This study also demonstrated the value of independent validation sets and the need for retraining using ,5% manual review for optimal performance when the model was applied to a new data set. Another common question in NGS workflows is whether or not a variant should be confirmed through an orthogonal method, such as Sanger sequencing. van den Akker and colleagues developed a proprietary machine learning model that uses reference sequence characteristics and variant call quality signals to accurately differentiate between high-confidence calls that do not require confirmation and low-confidence variants that require further testing.<sup>103</sup> The opacity of this particular report makes it impossible to understand the underlying model or to effectively evaluate its performance, but does offer an interesting use case for machine learning in

bioinformatic pipelines. In addition to the FDA regulatory framework described earlier, scientific literature provides example guidelines for the critical components needed to sufficiently report on model development and validation.<sup>104</sup>

Microscopic image analysis for cell or particulate classification and counting has been automated using machine learning methods. Several of these methods have been cleared by US FDA for use as clinical diagnostics. The Cellavision DM96 and DM1200 are machine learning-based applications for automated image analysis and cell classification and counting of white blood cells, red blood cells, and platelets in blood and body fluids.<sup>105,106</sup> Prepared slides are converted to digital images, which are analyzed by proprietary neural networks using morphologic features extracted from the images and a large database of labelled cells. All images and the results from the classifier are presented in a user interface overview screen. Trained operators can then reclassify cells to other classes, as needed, and verify results.

Similarly, urine sediment analysis has also been automated through the use of neural networks. The Iris Diagnostics automated microscopy system (iQ200) uses a proprietary flow-cell to specifically orient urine particulates for digital capture.<sup>107</sup> The Auto Particle Recognition software then uses size, shape, contrast, and texture as features in a proprietary neural network to classify individual microscopic particles in a urine sample. This system is also cleared for microscopic examination of body fluid specimens. Results and images are presented in a user interface for review, verification, and additional sub-classification. An added advantage of both the Cellavision and iQ200 systems is their capacity for electronic interfacing and autoverification. These two systems reflect some of the potential for automated image analysis within the modern clinical laboratory.

## SELECTED REFERENCES

21. Hastie T, Tibshirani R, Friedman JH. *The elements of statistical learning: data mining, inference, and prediction*, 2nd Ed. New York: Springer; 2009.
23. Géron A. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems*. Sebastopol, CA: O'Reilly Media, Inc.; 2019.
24. James G, Witten D, Hastie T, Tibshirani R. *An introduction to statistical learning: with applications in R*. New York: Springer; 2013.
25. Burkov A. *The hundred-page machine learning book*. Quebec, Canada: Andriy Burkov; 2019.
27. Murphy KP. *Machine learning: a probabilistic perspective*. Cambridge, MA: MIT Press; 2012.
29. Kuhn M, Johnson K. *Feature engineering and selection: a practical approach for predictive models*. Boca Raton, FL: CRC Press; 2020.
44. Kuhn M, Johnson K. *Applied predictive modeling*. New York: Springer; 2013.
72. Wickham H, Grolemund G. *R for data science*. Sebastopol, CA: O'Reilly Media, Inc.; 2017.
74. Kuhn M. Building predictive models in R using the caret package. *J Stat Softw* 2008;28:1–26.
82. FDA Center for Devices and Radiological Health (2020). Artificial Intelligence and Machine Learning in Software as a Medical Device. FDA. <http://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device> (Accessed March 2020).

## REFERENCES

1. Haymond S, Julian RK, Gill EL, Master SR. Machine Learning and Big Data in Pediatric Laboratory Medicine. In Dietzen DJ, Wong ECC, Bennett MJ, Haymond S, eds. Biochemical and Molecular Basis of Pediatric Disease, 5<sup>th</sup> Ed. Cambridge, MA: Academic Press; 2021.
2. Mayer-Schönberger V, Cukier K. Big Data: A revolution that will transform how we live, work, and think. New York: Houghton Mifflin Harcourt; 2014.
3. Apache Hadoop. <http://hadoop.apache.org> (accessed March 2020).
4. Stein L. Creating databases for biological information: an introduction. *Curr Protoc Bioinformatics* 2013;9.1.
5. Khasawneh TN, Al-Sahlee MH, Safia AA. SQL, NewSQL, and NOSQL databases: a comparative survey. In: 2020 11<sup>th</sup> International Conference on Information and Communication Systems (ICICS). Irbid, Jordan; 2020:13–21.
6. Kelleher JD, Tierney B. Cambridge, MA: The MIT Press; 2018.
7. NIST big data interoperability framework: volume 1, definitions version 3. <https://doi.org/10.6028/NIST.SP.1500-1r2> (accessed March 2020).
8. Obstfeld AE, Master SR, Miller WG. Using Big Data to Determine Reference Values for Laboratory Tests. *JAMA* 2015; 320:1495.
9. NIST standard reference database 1A v17. <https://www.nist.gov/srd/nist-standard-reference-database-1a-v17> (accessed March 2020).
10. EPA Tandem MS Data Library, <https://chemdata.nist.gov/dokuwiki/doku.php?id=chemdata:epatandem-library> (accessed March 2020).
11. van Veen SQ, Claas ECJ, Kuijper EJ. High-throughput identification of bacteria and yeast by matrix-assisted laser desorption ionization-time of flight mass spectrometry in conventional medical microbiology laboratories. *J Clin Microbiol* 2010; 48:900–7.
12. Welker M, Van Belkum A, Girard V, et al. An update on the routine application of MALDI-TOF MS in clinical microbiology. *Expert Rev Proteomics* 2019;16:695–710.
13. Takáts Z, Wiseman JM, Gologan B, et al. Mass spectrometry sampling under ambient conditions with desorption electrospray ionization. *Science* 2004;306:471–3.
14. Liu J, Wang H, Manicke NE, et al. Development, characterization, and application of paper spray ionization. *Anal Chem* 2010;82:2463–71.
15. Hau J, Lampen P, Lancashire R, et al. JCAMP-DX v.6.00 for chromatography and mass spectrometry hyphenated methods. *Pure Appl Chem* 2005;77.
16. ASTM E2078-00(2016), Standard Guide for Analytical Data Interchange Protocol for Mass Spectrometric Data, ASTM International, West Conshohocken, PA, 2016, [www.astm.org](http://www.astm.org)
17. Extensible Markup Language (XML) 1.0 (fifth edition). <https://www.w3.org/TR/2008/REC-xml-20081126/> (Accessed March 2020).
18. Pedrioli PGA, Eng JK, Hubley R, et al. A common open representation of mass spectrometry data and its application to proteomics research. *Nature Biotechnol* 2004;22:1459–66.
19. Binz PA, Barkovich R, Beavis RC, et al. Guidelines for reporting the use of mass spectrometry informatics in proteomics. *Nature Biotechnol* 2008;26:862.
20. Controlled Vocabularies, HUPO Proteomics Standards Initiative. <http://www.psdev.info/controlled-vocabularies> (Accessed March 2020).
21. Hoffmann N, Rein J, Sachsenberg T, et al., mzTab-M: a data standard for sharing quantitative results in mass spectrometry metabolomics. *Anal Chem* 2019;91:3302–10.
22. Finotello F, Di Camillo B. Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis. *Brief Funct Genomics* 2015;14:130–42.
23. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: data mining, inference, and prediction, 2nd Ed. New York: Springer; 2009.
24. Wolpert DH, Macready WG. No free lunch theorems for optimization. *IEEE Trans Evol Comput* 1997;1:67–82.
25. Géron A. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems. Sebastopol, CA: O'Reilly Media, Inc.; 2019.
26. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning: with applications in R. New York: Springer; 2013:96–97.
27. Burkov A. The hundred-page machine learning book. Quebec, Canada: Andriy Burkov; 2019.
28. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning: with applications in R. New York: Springer; 2013:37.
29. Murphy KP. Machine learning: a probabilistic perspective. Cambridge, MA: MIT Press; 2012:209.
30. Khalilia M, Chakraborty S, Popescu M. Predicting disease risks from highly imbalanced data using random forest. *BMC Med Inform Decis Mak* 2011;11:51.
31. Kuhn M, Johnson K. Feature engineering and selection: a practical approach for predictive models. Boca Raton, FL: CRC Press; 2020.
32. Alizadeh AA, Eisen MB, Davis RE, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*. 2000;403:503–11.
33. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: data mining, inference, and prediction, 2nd Ed. New York: Springer; 2009:534ff.
34. Yeoh EJ, Ross ME, Shurtleff SA, et al. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell* 2002;1:133–43.
35. Venables WN, Ripley BD. Modern Applied Statistics with S-Plus, 2nd Ed. New York: Springer; 1997:385ff.
36. Campbell MJ. Statistics at Square Two: Understanding modern statistical applications in medicine, 2nd Ed. Malden, MA: Blackwell Publishing; 2006:32ff.
37. Bittner M, Meltzer P, Chen Y, et al. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* 2000;406:536–40.
38. Master SR, Hartman JL, D'Cruz CM, et al. Functional microarray analysis of mammary organogenesis reveals a developmental role in adaptive thermogenesis. *Mol Endocrinol* 2002; 16:1185–203.
39. van der Maaten LJP, Hinton GE. Visualizing high-dimensional data using t-SNE. *J Mach Learn Res* 2008;9:2579–605.
40. Zhou B, Jin W. Visualization of single cell RNA-seq data using t-SNE in R. *Methods Mol Biol* 2020;2117:159–167.
41. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: data mining, inference, and prediction, 2nd Ed. New York: Springer; 2009:460ff.
42. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: data mining, inference, and prediction, 2nd Ed. New York: Springer; 2009:528ff.

43. Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999;286:531–7.
44. Tamayo P, Slonim D, Mesirov J, et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci USA* 1999;96:2907–12.
45. Eisen MB, Spellman PT, Brown PO, et al. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 1998;95:14863–8.
46. Kuhn M, Johnson K. Applied predictive modeling. New York: Springer; 2013.
47. Rosner B. Fundamentals of Biostatistics, 6<sup>th</sup> Ed. Belmont, CA: Thompson; 2006:668ff.
48. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: data mining, inference, and prediction, 2nd Ed. New York: Springer; 2009:119ff.
49. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: data mining, inference, and prediction, 2nd Ed. New York: Springer; 2009:463ff.
50. Binder SR, Genovese MC, Merrill JT. Computer-assisted pattern recognition of autoantibody results. *Clin Diagn Lab Immunol* 2005;12:1353–7.
51. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: data mining, inference, and prediction, 2nd Ed. New York: Springer; 2009:417ff.
52. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: data mining, inference, and prediction, 2nd Ed. New York: Springer; 2009:79ff.
53. Gromski PS, Muhamadali H, Ellis D, et al. A tutorial review: metabolomics and partial least-squares discriminant analysis—a marriage of convenience or a shotgun wedding. *Anal Chim Acta* 2015;879:10–23.
54. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: data mining, inference, and prediction, 2nd Ed. New York: Springer; 2009:305ff.
55. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: data mining, inference, and prediction, 2nd Ed. New York: Springer; 2009:312.
56. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: data mining, inference, and prediction, 2nd Ed. New York: Springer; 2009:282ff.
57. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: data mining, inference, and prediction, 2nd Ed. New York: Springer; 2009:587ff.
58. Chiofalo C, Chbat N, Ghosh E, et al. Automated continuous acute kidney injury prediction and surveillance: a random forest model. *Mayo Clin Proc* 2019;94:783–792.
59. Raess PW, van de Geijn GJ, Njo TL, et al. Automated screening for myelodysplastic syndromes through analysis of complete blood count and cell population data parameters. *Am J Hematol* 2014;89:369–74.
60. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: data mining, inference, and prediction, 2nd Ed. New York: Springer; 2009:337ff.
61. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: KDD ‘16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: Association for Computing Machinery; 2016:785–794.
62. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: data mining, inference, and prediction, 2nd Ed. New York: Springer; 2009:389ff.
63. Chollet F, Allaire JJ. Deep Learning with R. New York: Manning Publications; 2018.
64. Al-Rfou R, Alain G, Almahairi A, et al. Theano: a Python framework for fast computation of mathematical expressions. *arXiv preprint* 2016;*arXiv:1605.02688*.
65. Yamada T. Principles of error detection and correction. In: Imai H, ed. Essentials of error-control coding techniques. San Diego: Academic Press; 1990:11–37.
66. Aggarwal CC, Hinneburg A, Keim DA. On the surprising behavior of distance metrics in high dimensional space. In: Van den Bussche J, Vianu V, eds. Database theory — ICDT 2001. 1973:420–434.
67. Murphy KP. Machine learning: a probabilistic perspective. Cambridge, MA: MIT Press; 2012:77.
68. Duda RO, Hart PE, Stork DG. Pattern Classification, 2<sup>nd</sup> Ed. New York: John Wiley & Sons; 2001:20–83.
69. Murphy KP. Machine learning: a probabilistic perspective. Cambridge: MIT Press; 2012:76–78.
70. Ribeiro MT, Singh S, Guestrin C. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. *arXiv preprint* 2016;*arXiv:1602.04938*.
71. Shapley LS. A value for n-person games. In: Kuhn HW, Tucker AW, eds. Contributions to the Theory of Games 2.28. Princeton, NJ: Princeton University Press; 1953:307–317.
72. Lundberg S, Lee S-I. A Unified Approach to Interpreting Model Predictions. *arXiv preprint* 2017;*ArXiv:1705.07874*.
73. Jouppi NP, Young C, Patil N, et al. In-datacenter performance analysis of a tensor processing unit. In: Proceedings of the 44th Annual International Symposium on Computer Architecture. New York: ACM Press; 2017:1–12.
74. Wickham H, Grolemund G. R for data science. Sebastopol, CA: O’Reilly Media, Inc.; 2017.
75. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011; 12:2825–30.
76. Kuhn, M. Building predictive models in R using the caret package. *J Stat Softw* 2008, 28:1–26.
77. Tidymodels package. <https://CRAN.R-project.org/package=tidymodels> (Accessed March 2020).
78. Chollet F. (2015) keras. <http://keras.io> (Published 2015. Accessed March 2020).
79. Garner SR. Weka: The waikato environment for knowledge analysis. In: Proceedings of the New Zealand computer science research students conference 1995:57–64.
80. Abadi M, Agarwal A, Barham P, et al. Tensorflow: large-scale machine learning on heterogeneous distributed systems. *arXiv preprint* 2016;*arXiv:1603.04467*.
81. Paszke A, Gross S, Chintala S, et al. Automatic differentiation in pytorch. In: 31st Conference on Neural Information Processing Systems. 2017.
82. Sculley D, Holt G, Golovin D, et al. Hidden technical debt in machine learning systems. In: Jordan MI, LeCun Y, Solla SA, eds. Advances in neural information processing systems. Cambridge: MIT Press; 2015:2503–2511.
83. Breck E, Polyzotis N, Roy S, et al. Data validation for machine learning. In Conference on Systems and Machine Learning (SysML). <https://www.sysml.cc/doc/2019/167.pdf> (2019).
84. FDA Center for Devices and Radiological Health (2020). Artificial Intelligence and Machine Learning in Software as a Medical Device. FDA. <http://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device> (accessed March 2020).

85. Hill K. The Secretive Company That Might End Privacy as We Know It. The New York Times. <https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html>. (Published January 18, 2020. Accessed March 2020).
86. See case number '20 CU0370 BAS MSB, Filed Feb 27, 2020 US District Court, Southern California: SEAN BURKE and JAMES POMERENE, Individually and on Behalf of All Others Similarly Situated, Plaintiffs, v. CLEARVIEW AI, INC.
87. Buolamwini J, Gebru T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Friedler SA, Wilson C, eds. Proceedings of the 1st Conference on Fairness, Accountability and Transparency. New York, NY: PMLR; 2018:77–91.
88. Fleuren LM, Klausch TLT, Zwager CL, et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. *Intensive Care Med* 2020; 46:383–400.
89. Sun M, Baron J, Dighe A, et al. Early prediction of acute kidney injury in critical care setting using clinical notes and structured multivariate physiological measurements. *Stud Health Technol Inform* 2019;264:368–372.
90. Hyland SL, Falty M, Hüser M, et al. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nat Med* 2020;26:364–373.
91. Lindbury BA, Richardson AM, Badrick T. Assessment of machine-learning techniques on large pathology sets to address assay redundancy in routine liver function test profiles. *Diagnosis* 2015;2:41–51.
92. Luo Y, Szolovits P, Dighe AS, et al. Using machine learning to predict laboratory test results. *Am J Clin Pathol* 2016;145:778–88.
93. Zhang ML, Guo AX, Kadauke S, et al. Machine learning models improve the diagnostic yield of peripheral blood flow cytometry. *Am J Clin Pathol* 2020;153:235–242.
94. Richardson AM, Lidbury BA. Infection status outcome, machine learning method and virus type interact to affect the optimised prediction of hepatitis virus immunoassay results from routine pathology laboratory assays in unbalanced data. *BMC Bioinformatics* 2013;14:206–213.
95. Demirci F, Akan P, Kume T, et al. Artificial neural network approach in laboratory test reporting: learning algorithms. *Am J Clin Pathol* 2016;146:227–37.
96. Rosenbaum MW, Baron JM. Using machine learning-based multianalyte delta checks to detect wrong blood in tube errors. *Am J Clin Pathol* 2018;150:555–566.
97. Yu M, Bazdylo LAL, Bruns DE, et al. Streamlining quality review of mass spectrometry data in the clinical laboratory by use of machine learning. *Arch Pathol Lab Med* 2019;143:990–998.
98. Wilkes EH, Rumsby G, Woodward GM. Using machine learning to aid the interpretation of urine steroid profiles. *Clin Chem* 2018;64:1586–1595.
99. Zou J, Huss M, Abid A, et al. A primer on deep learning in genomics. *Nat Genet* 2019;51:12–18.
100. Wu C, Zhao X, Welsh M, et al. Using machine learning to identify true somatic variants from next-generation sequencing. *Clin Chem* 2020;66:239–246.
101. Zomnir MG, Lipkin L, Pacula M, et al. Artificial intelligence approach for variant reporting. *JCO Clin Cancer Inform* 2018;2:CCI.16.00079. doi:10.1200/CCI.16.00079.
102. Ainscough BJ, Barnell EK, Ronning P, et al. A deep learning approach to automate refinement of somatic variant calling from cancer sequencing data. *Nat Genet* 2018;50:1735–43.
103. van den Akker J, Mishne G, Zimmer AD, et al. A machine learning model to determine the accuracy of variant calls in capture-based next generation sequencing. *BMC Genomics* 2018;19:263.
104. Luo W, Phung D, Tran T, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *J Med Internet Res* 2016;18:e323.
105. Kratz A, Bengtsson H-I, Casey JE, et al. Performance evaluation of the CellVision DM96 system: WBC differentials by automated digital image analysis supported by an artificial neural network. *Am J Clin Pathol* 2005;124:770–81.
106. Swolin B, Simonsson P, Backman S, et al. Differential counting of blood leukocytes using automated microscopy and a decision support system based on artificial neural networks: evaluation of DiffMaster Octavia. *Clin Lab Haematol* 2003; 25:139–147.
107. Brunzel N. Fundamentals of Urine & Body Fluid Analysis, 4th Edition. St. Louis: Elsevier; 2016:342–5.

## MULTIPLE CHOICE QUESTIONS

1. Which of the following is an example of an ensemble learning method?
  - a. Decision tree
  - b. K-nearest neighbors
  - c. Random forest
  - d. Neural network
  - e. Logistic regression
2. Which of the following methods is appropriate for unsupervised learning?
  - a. Logistic regression
  - b. Principal components analysis
  - c. Boosted decision trees
  - d. Support vector machines
  - e. Random forest
3. Which of the following machine learning goals is a classification (as opposed to regression) problem?
  - a. Predicting the presence or absence of lung cancer
  - b. Predicting the estimated survival time from lung cancer
  - c. Predicting the numeric risk of acquiring lung cancer
  - d. Predicting the size (in cm) of a lung tumor
  - e. Predicting the age of recurrence of a lung tumor
4. Excessive tuning of a machine learning model to match a training data set can often lead to worse predictive performance on a new data set. This phenomenon is known as:
  - a. The curse of dimensionality
  - b. Irreducible error
  - c. Data imputation
  - d. Overfitting
  - e. Feature engineering
5. Which of the following is true of a model with low variance and high bias?
  - a. It will overfit the training data and perform well on new data.
  - b. It includes too many variables to accurately model the data.
  - c. It will underfit the training data and perform poorly on new data.
  - d. It has fit too close to the errors in the training data.
  - e. It has the ideal characteristics for a machine learning model.
6. Sensitivity is a commonly used metric for model performance. Which of the following represents the calculation of sensitivity for a classifier?
  - a.  $TP / (TP + FP)$
  - b.  $TP / (TP + FN)$
  - c.  $FP / (TN + FP)$
  - d.  $TN / (TN + FP)$
  - e.  $FP / (TP + FP)$
7. Which of the following describes the receiver operator characteristic (ROC) curve?
  - a. A plot of the true positive rate versus the false positive rate.
  - b. A plot of the recall versus the precision of a classification model.
  - c. A plot showing the residual error for predicted versus observed results.
  - d. A plot showing the bias and variance for a training set.
  - e. A plot of the loss function for a machine learning algorithm.
8. Which is true of a response variable in machine learning?
  - a. The response variable is the part of a dataset used to predict the feature variable.
  - b. The response variable represents a number to be predicted in a regression problem.
  - c. The response variable describes the difference between the observed and predicted data.
  - d. The response variable should not be highly correlated with feature variables.
  - e. The response variable is a hyperparameter that is tuned in regularization models.
9. Which of the following is an example of feature engineering?
  - a. Dropping features based on exploratory data analysis.
  - b. Selecting features with a regularization model, such as Lasso regression.
  - c. Calculating the error of the feature variables.
  - d. Encoding features as dummy variables.
  - e. Choosing features based on correlation with the dependent variable.
10. Which of the following models would be considered the least complex?
  - a. Convolutional neural network
  - b. Random forest
  - c. Logistic regression
  - d. Support vector machine
  - e. XGBoost

# Laboratory Stewardship and Test Utilization\*

*Gary W. Procop, Ronald B. Schifman, and Peter L. Perrotta*

## ABSTRACT

Laboratory tests substantially impact clinical care, but their overuse, misuse, or underuse can cause patient harm and dissatisfaction, suboptimal patient care, and increased costs. Traditionally, the first consideration of pathology and laboratory medicine has been centered on optimizing the analytic phase of testing, which, of course, is critical for producing reliable and meaningful test results. Subsequently, it has become clear that specimen acquisition, transport and processing (preanalytics) and the accurate and timely reporting of results (postanalytics) substantially affect the reliability of test results. Finally, proper test selection (pre-preanalytics), and accurate interpretation of results (post-postanalytics) are the other important components of laboratory testing that impact value and outcomes. All of these functions are within the scope of practice for pathologists and clinical laboratorians,

who are responsible for the overall quality of laboratory testing. This chapter outlines the reasons for suboptimal test use, lists potential outcomes of improper test use, describes how to create and maintain a test utilization or laboratory stewardship program, provides insights on the management of test utilization projects, and provides examples of specific interventions that can improve test utilization patterns. Finally, this chapter will underscore the importance of collaborative engagement of pathologists and clinical laboratorians with stakeholders to promote the optimal use and performance of laboratory testing from the moment testing is considered through to the clinical response to the results. In addition, because the area of Laboratory Stewardship is a relatively new and evolving field, the authors include examples of implementation of these practices.

---

\*The full version of this chapter is available electronically on [ExpertConsult.com](#).

## INTRODUCTION AND HISTORICAL PERSPECTIVE

Health care costs are a major concern in all parts of the world with rates of expenditure growth exceeding available resources. As newer technologies and medications become available, many of them with major costs involved, there is a need for health care providers to consider the value of all health interventions.

In the United States, a transition from volume-based health care reimbursement to value-based models is well underway.<sup>1–4</sup> These changes essentially limit the amount that will be paid for any particular procedure or treatment for particular diagnoses. The Medicare Access and CHIP Reauthorization Act (MACRA) of 2018 was implemented in part to incentivize clinicians for participating in value-based health care delivery.<sup>5</sup> Health care providers, in fact, have already adjusted to this type of model with inpatient billing in the form of diagnosis-related group (DRG) payments. Although the billing and coding requirements are more complicated than presented here, for a particular DRG, the hospital is paid a set amount for taking care of a patient with a particular condition (e.g., hip replacement surgery). The concern with such a model is that health care organizations might reduce services to decrease costs and increase profits at the expense of quality of care. Payors have countered by not covering complications secondary to what they may view as substandard care. Examples include not covering hospital-acquired infections and readmissions that occur soon after discharge.

Diminishing the use of unnecessary laboratory tests, which increases costs without adding value, is one way health care leadership can decrease health care costs while improving patient care, satisfaction, and overall quality.<sup>6–8</sup> Omission of needed testing is another concern that can impact the delivery of appropriate care and may be more common than overtesting. The main focus of quality systems in laboratory medicine has primarily involved optimizing analytical performance and to some degree improving the preanalytical and postanalytical phases of testing.<sup>9</sup> Much less attention has been directed at addressing the quality of test ordering practices (pre-preanalytics) and the optimal use of results for best patient outcomes (post-postanalytics).<sup>10</sup> Laboratory stewardship and test utilization management are terms used to describe these latter components of laboratory quality performance, which are described in this chapter with a focus on the pre-preanalytical component.

Laboratory-based professionals contribute significantly to patient care by providing high-quality, reliable test results on which diagnostic and therapeutic decisions are made. This is a considerable contribution to patient care, but the truth is that substantially more can be accomplished by increasing laboratory engagement in activities that precede and follow the analytic phase of testing.<sup>11</sup> Greater engagement in pre-analytics helps to ensure the correct specimen is collected in the proper manner, transported in a manner to maintain specimen integrity, processed promptly, and tested accurately. For example, it has been demonstrated that the pre-analytic transit time is the most important factor influencing time-to-positivity of blood cultures; this is not because of the growth characteristics of the bacteria, but rather the amount of time the specimen is delayed before being placed on the instrument.<sup>12</sup> In such an instance, time-to-positivity could be

improved by decreasing transit time, a controllable preanalytic variable. Another example is the investigation of preanalytic factors that could be addressed to decrease hemolysis in blood specimens collected in the emergency department.<sup>13–17</sup> Similarly, attention to the postanalytic reporting of test results can help to ensure that results are accurately reported in a prompt, timely, and user-friendly manner. Over the past decade, much has been done in the both the preanalytic and postanalytic space to improve test performance, reporting, and, subsequently, patient care.

The areas of more recent focus center on the reasons for testing (i.e., pre-preanalytics) and result interpretation and subsequent clinical action (i.e., post-postanalytics). For example, there may be substantial opportunities in population health management by the coordinated use of laboratory data in the postanalytic space.<sup>18</sup> Engagement in these areas will require individuals with appropriate medical knowledge beyond that of test performance and, perhaps even more importantly, excellent communication, professionalism, and team-working skills. Further participation in the pre-preanalytic and post-postanalytic space serves to more completely engage the pathologist and laboratorian in patient care in a manner that is needed to further optimize health-care delivery into the next decades.

### POINTS TO REMEMBER

- Health care costs continue to increase.
- There are pressures to deliver high-quality care at a lower cost.
- Laboratory stewardship initiatives can maintain or improve quality while lowering health care costs.
- Laboratory-based professionals can improve health care delivery and lower health care costs through active participation in test utilization initiatives.

## CAUSES OF OVERUTILIZATION AND UNDERUTILIZATION

Inappropriate test ordering is common. One meta-analysis reported overall mean rates of overtesting and undertesting of 20.6% and 44.8%, respectively.<sup>19</sup> Another study involving outpatients reported that only 49% of hemoglobin A<sub>1c</sub> orders were appropriate, with 21% having been ordered too frequently (overtesting) and 30% having been ordered too infrequently (undertesting).<sup>20</sup> The Choosing Wisely campaign was established by the American Board of Internal Medicine (ABIM) Foundation to promote better utilization of diagnostic testing. More than 80 medical societies have contributed more than 500 recommendations, many of which involve laboratory testing. However, there remains limited evidence-based support for optimal selection and frequency of testing for most analytes, and guidelines so far developed involve complex variability in test frequency recommendations which are dependent on specific clinical factors.<sup>21</sup> This makes it difficult or impossible for testing guidelines to be effective. Even when policies are developed for optimal test ordering, implementation is difficult to support due to lack of functionality in laboratory information systems for expertly managing the flow of orders.

The reasons for underutilization, overutilization, and misutilization of laboratory tests are legion but, importantly,

can be addressed in a variety of ways.<sup>22–24</sup> Why would someone order a test that was not needed (i.e., overutilization)? From our experience, one of the most important reasons is that the provider did not know that a test was already ordered. This is particularly common in hospitalized patients who are being seen by both their primary team and consultants. A test previously ordered by the primary team may also be considered necessary by a consultant, who then reorders the same test. One may ask: Why didn't they check to see what tests have already been ordered? As it turns out, complexities of ordering systems and the limited time physicians have to spend with each patient make this difficult. Although it is feasible to determine through computer searching what tests have been ordered, the fact of the matter is that clinicians are often too busy to stop and check. Instead, they order what they need for that patient and move on, assuming that duplicate orders will be addressed downstream. Unfortunately, in many instances duplicate orders are not detected and canceled, which leads to unnecessary duplicate testing. Although manual interventions may be used to reduce unnecessary testing, these incur a substantial toll in human labor, diverting highly qualified resources (e.g., medical technologists) from patient testing to the work of reviewing logs for duplicate studies. This is operationally inefficient and adds cost to an already overburdened health care system. Furthermore, when tests are canceled "downstream," such as after receipt in the laboratory, the patient has already suffered an unnecessary phlebotomy or another collection procedure. This highlights the importance of stopping unnecessary testing at the point of computerized order entry *before* a specimen is collected.

Laboratory tests, like other diagnostic procedures, may also be overordered for fear of litigation.<sup>25,26</sup> Although this has been stated, it has not been encountered as a common reason in the experience of the authors of this chapter. Excessive ordering may also occur because orders are unnecessarily embedded in order sets or protocols developed by organization to facilitate ordering or to standardize practices.<sup>27</sup> We have experienced this with the complete blood count (CBC) with differential (CBC w/ Diff). For example, a CBC w/ Diff may not be needed in a patient for whom a "Rule Out Myocardial Infarction" order set is being used; yet, if a CBC w/ Diff is included in that order set, then the unneeded differential will be performed for every patient with chest pain for whom this order set is used. Standing orders for testing to be performed at set intervals also contribute to excessive, unnecessary testing. Medical trainees have admitted that they will place standing orders (e.g., a daily order for CBC) for tests commonly performed in the inpatient setting so they do not forget to place the order each day the test is truly needed. It is a major failing if we are training new physicians who practice this way, resulting in the overphlebotomization of patients and a waste of health resources, rather than pausing to consider which tests are truly needed for patient care.

Overutilization, underutilization, and misutilization of laboratory tests can each be caused by a provider's lack of understanding of the test.<sup>23</sup> A thorough understanding of the "how," "when," "how often," and "for whom" a test should be performed is critical for appropriate utilization.<sup>11</sup> Recognition of the complexity of every test offered in the laboratory is simply not feasible for the new intern who is now responsible for placing patient orders. Therefore it is crucial to

provide resources for these individuals in teaching hospitals, which may include readily available laboratorian consultation. There also needs to be a sea change in medical education regarding the traditional approach of many attending physicians who will deal harshly with residents who fail to order a test. The approach of residents and fellows ordering whatever they think may even be remotely needed in order to avoid criticisms of an attending is a paradigm that needs to change, and relies significantly on the approach of attending providers to trainees. Sedrak and colleagues have summarized these and other reasons residents perform unnecessary testing.<sup>28</sup> A succinct attending/resident discussion after a patient encounter regarding what tests are needed could both remedy this cause of overutilization and provide many "teachable moments." The lack of understanding of testing indicates the need for pathologists and clinical laboratorians to become even more engaged in medical education outside their department and to more fully participate in care delivery.<sup>29</sup>

### POINTS TO REMEMBER

- Inappropriate test ordering is common.
- There are many reasons for inappropriate test ordering; determining the root cause is important for designing interventions.
- Advances in hospital informatics systems are needed to optimize test utilization.
- An evaluation of order sets and standing orders, including daily orders, often disclose opportunities for improvement.
- Readily available laboratory consultations may improve test utilization.

## THE LABORATORY STEWARDSHIP/TEST UTILIZATION COMMITTEE

The Laboratory Stewardship or Test Utilization Committee is a hospital- or health system-based committee that is concerned with the optimal utilization of laboratory tests and services. It should be as much focused on underutilization as overutilization. The objective of Laboratory Stewardship committee members is to develop and endorse sound policies and procedures for their institution that promote effective laboratory testing practices among its various stakeholders. The Pharmacy Formulary committee and Antimicrobial Stewardship committee are similar groups that serve as a model for committee structure and function. The following section contains guidance regarding the establishment and management of a successful Laboratory Stewardship committee.

### The Philosophy and Charge of the Committee

It is important for individuals of any committee to understand the charge of the committee and to determine how they can contribute. Although Mission and Vision statements are not mandatory, the development and periodic review of these help committee members to recall their reason for taking time out of an already busy day to participate. The presence of the committee *charge* denotes external support from senior leadership, which helps to lend credibility to the committee (see later). It is our belief that the underlying reasons for the existence of the committee should *not* be solely to

reduce the cost of health care. Reducing costs is not an internal motivator for many, and focusing solely or predominantly on money “saved” may result in a loss of participant engagement. However, we recognize that reduced health care costs are an important by-product of eliminating waste and optimizing care delivery pathways and can be viewed as an ethical imperative because it frees up health dollars for other purposes.

Most health care professionals entered medicine to improve the lives of the patients they serve. These altruistic reasons should be the primary drivers for forming a Laboratory Stewardship or Test Utilization committee.<sup>11</sup> It should be recognized that excessive phlebotomy is painful and stressful to patients, particularly those who are hospitalized and enduring this day after day. This practice may also lead to iatrogenic anemia, which has untoward consequences, such as poor wound healing, increased infections, and, in patients with underlying disease, cardiopulmonary compromise.<sup>30,31</sup>

Another fact that should be considered is that performing even technically sound tests (i.e., tests with high sensitivity and specificity) in low-prevalence populations leads to extremely poor positive predictive values.<sup>11</sup> Otherwise stated, there will be false-positive test results that may lead to additional testing, which could include radiologic studies or other expensive or more-invasive procedures. The avoidance of these deleterious effects on patient care, and the improvement of patient satisfaction and clinical outcomes should drive the need to intervene. Improving patient care, optimizing care delivery, and working to ensure best practices are used will always find a receptive audience and will generate enthusiasm in health care providers.

### Leadership Support and Reporting Structure

The Laboratory Stewardship committee should be sanctioned by the leadership of the institution and should have a defined reporting structure. This is important for a number of reasons. Foremost, it demonstrates that the appropriate use of laboratory tests is important to the leadership of the institution, which lends credibility to the initiatives. However, what if test utilization is not yet on the radar of institutional leadership? In this case, the starting point is to convince leadership, in an evidence-based manner, that these initiatives are important for high-quality patient care, satisfaction, and safety, while contributing to cost-savings in health care delivery. Leadership support is also important because the committee will often be instituting test ordering changes that affect the entire health system. Once the reporting structure is established, it is recommended that regular meetings are scheduled with the individual or committee to whom the committee will report. Laboratory Stewardship committees commonly report into the hospital quality structure or medical operations. There is often a dual reporting to the Chair of Pathology and Laboratory Medicine or equivalent lead of laboratory services. This, too, is important because contributions from testing content experts will be needed when changes are proposed.

The engagement of physicians and other caregivers is also critical. Although “mandates” from hospital or medical leadership may eventually be implemented, they may be resisted. The goal is to introduce and/or ensure that best practices in test utilization are undertaken. Although it would be optimal if every test were used correctly every time, this is unrealistic.

We are foremost concerned with aberrant utilization that could cause patient harm. After this, we are concerned with optimizing utilization to enhance patient care outcomes, increasing patient satisfaction, and decreasing unnecessary health care costs.

The solicitation of individuals from different departments, through their departmental chairs/directors, who are interested in laboratory stewardship/test utilization initiatives, is a good way to form the core group of committee members (see “Committee Composition” later). One of the lessons we have learned over the years, which has been confirmed by others, is that physicians on the committee are often unlikely to make decisions about the practice of other colleagues outside of their subspecialties (e.g., a rheumatologist is unlikely to be prescriptive to a surgical oncologist), and upon reflection, they should not do so. However, any physician should be able to raise questions regarding why certain tests are needed or why they are repeated at certain intervals (i.e., what is the evidence?).<sup>32</sup> A recommended solution for addressing subspecialty issues is to form subspecialty task forces of content experts to address issues that arise within a subspecialty. For example, representatives from Infectious Diseases, Clinical Microbiology, and Immunology/Immunopathology would be core members of a task force assembled to address inappropriate testing for Lyme disease. In such a situation, a member of the committee, either with or without content expertise, can act as a facilitator to work matters to a conclusion.

### Organizational Structures/Committee Composition

The physicians on the committee function in a number of roles. Foremost, physician leadership on the committee and/or subteams is important because a physician-to-physician conversation is often necessary. The committee may be led by a single physician chair, or there may be co-chairs. There are benefits to having a clinical co-chair and a pathology co-chair because each brings a different set of competencies and perspectives to the committee and initiatives. It is also important to have physicians and clinical laboratory scientists as part of the Laboratory Stewardship committee core team. These individuals will contribute to issues encountered daily within their practices and will lend a global (systems-based) perspective to the committee. As mentioned elsewhere, it is beneficial to have providers who may act as ad hoc members, participating with standing committee members in their areas of content expertise.

Nonphysician caregivers are critically important for the success of the Laboratory Stewardship Committee. These include, but are not limited to, genetic counselors, members of the informatics team, statisticians, a representative from finance, nurses, medical administrators, and physician assistants, among others.

Genetic counselors perform a number of activities related to test utilization at institutions wherein there is a substantial amount of genetic testing and should be included if these tests are being considered.<sup>33</sup> These individuals assist patients in obtaining medically necessary genetic testing and provide appropriate pretest counseling while remaining sensitive to the high cost of these tests. More recently, subspecialization has occurred within this group, with individuals concentrating on laboratory-based genetic counseling. These individuals provide heightened expertise in the review of genetic tests

and are highly effective in both ensuring the patient receives the appropriate test while decreasing costs by intervening on unnecessary genetic testing that is often sent to reference laboratories.<sup>34</sup> Moreover, genetic tests may require preauthorization from payors. In some practices, genetic counselors have assumed responsibility for obtaining preauthorization, whereas in other settings they serve as resources to members of the preauthorization team that resides in finance or the clinical laboratory.

Members of the informatics team are critical partners in laboratory stewardship for a variety of reasons. Foremost, these individuals are often responsible for acquiring the data that will be used by the stewardship team to justify a new test utilization initiative. There is need for a good information technology (IT)-clinical interface, and this may be through clinicians who understand the IT system and/or IT specialists who understand the clinical use of the data. A failure to properly understand the data can jeopardize a project. IT is needed at all stages of a project, usually including implementation, as well as ongoing review for effectiveness. Once high-quality data are obtained, the clinical meaningfulness of the data should be determined. This is undertaken by content experts who take into account the clinical scenarios in which the data were obtained. For example, one may ask: Is more than one CBC per day necessary for an inpatient? Such a question cannot be answered without understanding the clinical context. Repetitive CBCs are probably not needed in a patient with uncomplicated pneumonia but may be in a bleeding trauma patient. Such a thorough and thoughtful review of the data will aid in the decision as to whether or not an intervention is warranted. If an intervention is deemed necessary, then high-quality data will serve as a baseline against which the effectiveness of the intervention will be measured.

Once an intervention has been completed and substantial data are available, a review of the effectiveness of the intervention should be undertaken. These data will again be provided by collaborators in the hospital informatics area, reinforcing the importance of a close working relationship with this group. Once sufficient and high-quality postinterventional data are obtained, it must be analyzed. It is important to have access to individuals who are competent in statistical methods. Similarly, it is important to have individuals from finance, either available when needed or as members of the team who can determine the financial impact of the interventions. There are a number of pitfalls in determining the financial impact of interventions, and it is important to realistically calculate cost-savings.<sup>11</sup>

Many of the other nonphysician members of the team will be critical to operationalizing the interventions. These include administrators, nurses, and advanced practice providers (APPs; e.g., physician assistants and nurse practitioners). Administrative representation is particularly important if additional personnel or other resources are needed, to remove barriers, or if new assignments must be made to an employee's job duties. These individuals, working with finance, help to estimate the cost of the undertaking and the impact of intervention. Nurses and APPs are particularly helpful for understanding the current state (i.e., what is actually happening on the floors and clinics) and to assist with implementing certain interventions. These individuals represent champions on the floors, who can present information at local huddles and provide onsite training as necessary. Similarly, individuals with

high level communication and change management skills can help to ensure a smooth implementation of committee interventions and avoid surprises and even potentially angry responses from affected individuals who were not properly notified of changes that affect the way they practice.

## POINTS TO REMEMBER

- The Laboratory Stewardship Committee should have institutional support and a reporting structure.
- Improving the health of patients should be the primary driver of the committee; cost savings is a secondary gain.
- Clinical and laboratory subspecialty experts form a knowledgeable and effective taskforce when addressing test utilization within their scope practice.
- A multidisciplinary Laboratory Stewardship committee, consisting of physician and nonphysician care providers, administrators, informaticists, and others, is recommended.

## PROJECT LIFE CYCLE

Once a Laboratory Stewardship committee has been active for a while there will be a number of ongoing projects in various phases of accomplishment. This complexity underscores the need for an efficient and well-organized project manager. The following section highlights the three major phases and important subcomponents in the project life cycle of a test utilization committee. Giving due consideration to each of these will help to ensure a successful program. These phases are:

- Project Identification and Confirmation
  - Project identification
  - Data acquisition and review
  - Interventional strategies review and selection
- Interventional Planning and Execution
  - Communication and change management
  - Exit strategy
- Impact Analysis

### 1. Project Identification and Confirmation

There are countless projects that could be undertaken by a Test Utilization committee. Therefore the question often becomes: Where to start? When first starting a Laboratory Stewardship committee, it is important to choose initial projects that are likely to succeed, the so-called low-hanging fruit. Success, which can be demonstrated by outcomes (see later), is important in the early stages of a new program to build credibility. Success also builds confidence within the team that they can accomplish meaningful change. This credibility, from a growing track record of success, will also prove helpful in continuing current support from institutional leadership or even obtaining additional administrative or project manager support.

Beware of the team members who are replete with ideas and suggestions that the committee should do but do not expend energy to assist with these projects. It can be useful to ask how they would like to approach the proposed project. If their response is unclear, consider thanking them for the recommendations and place these in a "parking lot" until such a time when resources are available to address their recommendation. Members of the committee should be there to contribute, as well as generate ideas for the committee. So then, how to identify projects?

### A. Project Identification

There are a number of ways to identify projects. Identifying projects is not difficult when committees are first formed. More important is the assessment of the value of the project, the feasibility of task completion, the ranking of it with other projects, and the evaluation of the resources that will be needed to bring the project to its successful completion. Brainstorming with the members of the committee is a good way to start and to identify potential projects. If this is done properly, the same group can also be asked to rate these with respect to both medical importance, feasibility, and the other aforementioned characteristics.

Another way to identify projects is to ask medical leaders of the different clinical departments for their opinions as to which areas are ripe for utilization projects. An added advantage of this approach is that if one of their projects is selected, it is reasonable to return to the person making the proposal and ask them to assist with identifying a clinical champion for the proposal. For example, when we received requests from the emergency department to decrease hemolyzed specimens, an individual from that department served as the clinical champion. They have since led several projects through to completion.

Finally, it is helpful to critically assess the overall direction and goals of the institution and to consider projects that are congruent with these growth strategies. For example, if a new Department of Medical Genetics is being established, then working with the new leadership in this area to optimize genetic testing would likely be embraced. Once a potential intervention is selected, which sometimes occurs due to anecdotal evidence or the input of a more insistent member, then it is important to gather and review data to truly understand the nature of the issue and the validity of the request.

It is also important to recognize the growing literature on laboratory stewardship in general and also of specific projects. It is a vital component for the team to be up to date on relevant literature on any project that is planned, so that achievable projects are selected, good ideas are adopted, and potential errors may be avoided.

### B. Data Acquisition and Review

As mentioned earlier, data are needed to document baseline utilization rates, to determine who is ordering which laboratory tests, to identify the nature of utilization problems, and, after an intervention is made, to determine the effectiveness of the intervention. Therefore sound processes are needed to acquire high-quality data from the laboratory and hospital information databases.<sup>35</sup> Individuals are needed who have the skills to verify the validity of utilization data and to organize large amounts of complex data in useful forms (e.g., graphs and summarized tables) for the committee.

The members of the leadership team should review the data to determine if a test utilization issue exists and if the scope of the issue is sufficient to warrant the time, expense, and energy needed to intervene. If these criteria are met, then the data should be retained because they represent an important preintervention current state against which postintervention data will be compared to assess the efficacy of the intervention. The next question becomes: how do we fix this problem?

### C. Interventional Strategies Review and Selection

A thorough discussion regarding the interventional strategy options should be undertaken. The team should not prematurely arrive at a solution strategy without reviewing all suitable options. This approach helps to ensure the use of the optimal interventional strategy, rather than trying to force a project into a predetermined strategy. If an electronic intervention is considered, then it is recommended to meet with the informatics contact on the team and discuss the end goal of what the group would like to accomplish. The individual from the informatics team usually understands the capabilities of the electronic system(s) in use and can recommend interventional options based on the strengths and limitations of each. This small amount of background work can help to ensure that the most appropriate intervention is undertaken, thus limiting the amount of rework.

### 2. Interventional Planning and Execution

Once a number of projects have been identified and prioritized as described earlier, the team can begin working on the most important topics. It is advisable to initially limit active projects to no more than three, unless there are sufficient resources to address more; this is usually not the case for newly formed utilization groups. The proposed projects should be reviewed with senior leadership to ensure they align with institutional priorities and that they will support the efforts. This review also provides an opportunity to discuss with senior leadership what resources are needed to accomplish the projects and how they might remove potential barriers.

### A. Communication and Change Management

Why are you modifying *my* order set? Who are you, and why are you changing or interfering with the way I practice? These are the types of questions that can be averted through thoughtful communication and the execution of a change management strategy.

Foremost, it is important to understand the primary stakeholders of the test targeted for intervention. Ideally, there has already been a clinical champion identified who is working with the team on this project. If there is not, seek to identify one as part of the intervention. One of the best ways to find a clinical champion is to demonstrate the need to leadership in the stakeholder area using baseline data. It has been our experience that, once baseline data demonstrating inappropriate utilization are shared, individuals within the stakeholder area will offer opinions regarding how to intervene. This is the perfect time to ask for a clinical champion to join the team. The importance of the peer-to-peer conversations within the group that will be affected cannot be overstated.

There should be broad communication of the upcoming changes to the entire medical staff. If the test is predominantly used by a particular group, then additional focused communication, which can often be achieved by email, should be considered. Communicating the changes that are forthcoming and the reason for the change should be done in advance of any infrastructure builds or program changes. This provides the medical staff the opportunity to provide feedback, which will occasionally disclose a nuance that had not been considered. If the interventional implementation proceeds, then there should be another communication near to when the change will be implemented. If only a single communication is made or if the change is communicated

too far in advance of the intervention, then individuals may forget about the impending change, which can lead to questions and/or complaints.

### B. Exit Strategies

Even the best planned interventions sometimes will have unintended consequences. A way to avoid this, in part, if one is utilizing an electronic intervention (i.e., a clinical decision support tool), is to allow the intervention to run “silently” in the background for a period of time before going “live.” This can determine the scope of the impact including how many individuals or groups will be affected by the change and who among the medical staff is placing the orders. This information also affords one a chance to optimize communications, address issues that may not have been considered, and hopefully avert disaster.

It is important to consider how the team will respond if the changes lead to significant problems. Foremost, a response team needs to be present, so do not make major changes on a Friday afternoon, a weekend, or holiday, and there needs to be a process of oversight, for example, by monitoring requests in real time or directly communicating with users. If untoward complications arise, one of the easiest approaches to take is to return to the former state of operations (rollback). These types of setbacks are often disheartening to the entire utilization group. However, a rapid rollback is usually preferable to pushing on with a failing system. Thoughtful planning, excellent communications, and piloting of the intervention can minimize such occurrences. If possible, the group should persevere under adverse situations by identifying the adverse issue and proposing, in real time, potential workarounds in an effort to salvage the project. In the worst situations it may be necessary to restart the project. As an example, although all approvals were obtained from our primary stakeholders for an intervention that addressed the excessive ordering of serum and urine protein electrophoresis assays, it was not until after we instituted a 2-week hard stop to deter duplicate testing that we discovered this intervention seriously interfered with research protocols that required more frequent testing. The intervention was temporarily discontinued while we determined a workaround to allow physicians with patients enrolled in clinical trials to obtain the more frequent testing dictated by the research protocol. The intervention otherwise remained active for all other providers. It is often found that the environments in which we practice are very complex, and unintended consequences are not infrequent.

### 3. Determination of Impact

The time and energy expended by a number of valued employees, in addition to other resources, will be invested in initiatives to optimize test utilization. It is reasonable for administration and senior leaders to expect progress reports describing the impact of these initiatives. Prior to initiating any intervention, there should be a thorough knowledge of the baseline activities of ordering providers. This information, which the committee used to justify the initiative, will also serve as the baseline against which the change produced by the intervention may be measured in a preintervention/postintervention analysis.

Both short-term and long-term analyses can be used to quantify the impact of utilization interventions. The short-term

analyses function largely to determine if the interventions are having their intended impact and to detect any unexpected consequences of the interventions. Any significant interruptions in care delivery must be brought to the attention of the committee and senior leadership sponsors. Although this is undesirable because it may disenfranchise caregivers and make leadership question their confidence in the committee, openness to criticism and a rapid response are key components for a project. The importance of thorough preintervention planning and an excellent and well-executed communicated change management plan is therefore essential.

Once an intervention has been in place for a number of months (e.g., 6 months to a year), it is important to analyze its impact. Did the intervention function as planned, or did caregivers eventually discover a way to circumvent the activity you were trying to discourage? Many individuals and health system groups are interested in the outcomes of such interventions, as we learn from one another in this space. As large projects are completed, consideration should be given to sharing your findings with the health care community through presentations at national meetings and peer-reviewed publications.

Detailed how to analyze the effects of the sundry interventions is beyond the scope of this chapter because they vary depending on the type of intervention undertaken. However, appropriate analyses are important to avoid serious errors and misleading claims made through miscalculations. A common error is using test charge (i.e., patient charge) instead of test cost (i.e., cost to perform or refer a test) when determining the financial impact of decreasing laboratory costs.<sup>11</sup> Similarly, the fixed costs within the laboratory are usually unaffected by utilization management changes, so these are usually not included in cost-savings calculations. The details of these undertakings are covered thoroughly in the Analytics and Measures section of GP49 *Developing and Managing a Medical Laboratory (Test) Utilization Management Program*.<sup>11</sup>

### POINTS TO REMEMBER

- Organization is the key to project management.
- Thoughtful project identification and ranking will avoid the exhaustion of resources.
- Baseline utilization rates should be determined to inform project progression and to measure against once an intervention has occurred.
- Consider all possible intervention strategies before selecting a course of action.
- Communicate broadly prior to implementing any interventions.
- Measure and report the impact of the test utilization intervention.
- Have a Plan B, in the event the intervention is problematic.

### INTERVENTIONAL STRATEGIES

There are a large number of interventional strategies that can help to optimize the use of laboratory resources.<sup>36</sup> These vary from silent interventions working behind the scenes to highly intrusive interventions. The effectiveness of each intervention varies, so they should be carefully considered prior to implementation and measurements made after implementation to assess their efficacy. There should be a reasonable

balance between intrusiveness and effectiveness. It is important to consider the workflow of our health care providers in an era of mounting stress and burnout and not to add unnecessarily to an already troubling situation. We should strive to minimize the intrusiveness of our interventions and help providers whenever possible. The perfect balance would be interventions that are silent or minimally intrusive, providing assistance to clinical providers while improving patient safety.

There are two major categories of interventions that are described as follows:

- Provider feedback and education
- Test order control

The first of these involves providing education and feedback to providers on test utilization practices. Test utilization education may be provided just prior to a test being offered (i.e., going live). Although purely educational interventions are of limited efficacy, these provide useful material for teaching students and capture the rationale for the intervention. Educational material also provides valuable content for long-standing resources (e.g., test directories and test ordering guides) used by clinicians and laboratory staff when testing questions arise. Information may also reside in real-time computerized decision support systems (CDSTs; e.g., in the form of “best practice alerts”). Retrospective feedback may be presented to ordering providers to show them their test ordering practices. The other major interventional strategy considered here involves methods of test order control. These interventions may include optimizing the test menu to facilitate appropriate ordering, stopping unnecessary testing using CDSTs, and restricting testing to certain clinical groups (e.g., only certain privileged providers can order a specific test).

### 1. Education and Feedback

Three of the major forms of feedback provided to clinicians who utilize laboratory tests are discussed here: prospective educational information, real-time content delivery, and retrospective analysis and feedback.

#### A. Prospective

Prospective notification regarding the optimal use of testing is tantamount to education. There are a number of studies demonstrating that purely educational interventions are ineffective in changing test ordering behavior.<sup>37</sup> If this is true, then why do it? There are a number of reasons. The most important is the training of clinical providers and fostering in them good test ordering practices. Phillips and colleagues demonstrated that individuals who trained in programs that did not adhere to good test ordering practices continued these wasteful utilization habits well into their careers, even if they subsequently practiced in groups that adhered to good stewardship practices.<sup>38</sup> Conversely, they also found that those who were trained in good stewardship practices continued as good stewards, even if they practiced in an otherwise wasteful group. This *imprinting* during the formative years of medical education molded the future practice of the physician. Educating the next generation of practitioners thus appears fertile ground to improve the utilization practices of future providers.

Changes, and communications regarding these changes, are commonly made to test menus. Standardized communication of such changes, perhaps in the form of a monthly “Laboratory Updates” newsletter, provides an opportunity to

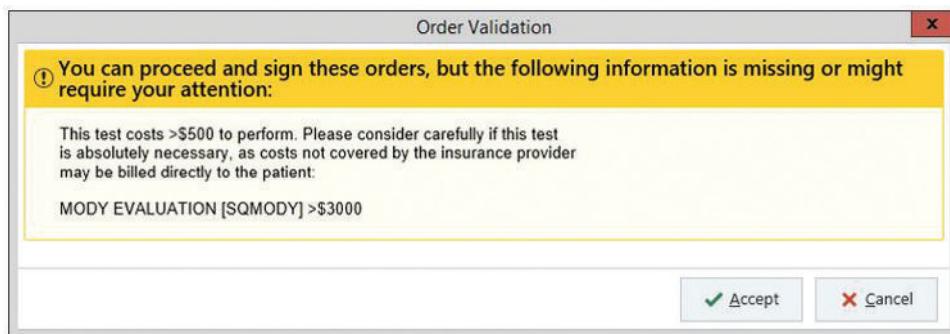
provide education regarding the optimal use of an assay. This can be done succinctly by describing how a laboratory test can be used to diagnose a particular disease, specifically stating the indications for use, circumstances where the test is not indicated, and advice on a retesting interval if repeat testing may be indicated. The material included in such newsletters can be readily repurposed for student teaching for individuals at academic centers. When a substantial number of these newsletters have been created, they may be presented in toto to the incoming class of interns as a resource for test ordering. The challenge with educational initiatives like this is that they are often missed, not read, or readily forgotten by the busy clinical team. Therefore one should consider archiving and curating these, given the effort it takes to create them and their importance as an educational resource.

Most laboratories have an electronic laboratory test menu or formulary. Linking the laboratory test menu to the laboratory utilization guide provides an easily accessible resource for providers that is just “a click away.” Curation is an important component of the life cycle of test utilization menus and guides, so a process for periodic review is recommended. When a current laboratory utilization guide exists, providers with questions can be directed to the particular page or forwarded a link that answers their question. In so doing, in addition to having their question answered, they are being made aware of the resource that they will hopefully use in the future.

#### B. Real-Time Feedback

Real-time feedback may be manual or electronic. The manual intervention usually involves a telephone call to the provider who is ordering a potentially inappropriate test. There are a number of challenges associated with telephone-based interventions. The foremost question is: Who is doing the calling? There are residency programs wherein faculty consider this type of intervention a learning experience. It can be, or it can be viewed with fear and trepidation and become one of the worst experiences of residency. Residents involved in a test utilization review program should have time to read about the test in question and should have access to a pathologist or laboratorian with considerable expertise in the area and preferably implementing an agreed protocol. In this manner, they can have a meaningful conversation with the provider. Otherwise, the result is a resident trying to refuse a test to an insistent provider, who likely knows more about the subject area and eventually overpowers the resident and gets the desired test. This can be an unpleasant and unproductive experience for all involved. If residents are to be employed in test utilization management, the experience should be truly beneficial to their education while improving patient care.

In some instances, the disagreement regarding whether or not to perform a test must be escalated to the pathologist or laboratorian, who will speak directly with the provider. This is important to do if there are misunderstandings about the scientific details of the testing. This is less productive from a medical operations perspective, because productivity is optimal when clinical providers are seeing patients and pathologists or laboratorians are doing their work rather than debating the appropriateness of testing. Nevertheless, intervening is a responsibility of individuals with laboratory oversight and is occasionally necessary. However, minimizing this type of interaction is of benefit to all. Herein lies the advantages of electronic real-time feedback.



**FIGURE 14.1** This Best Practice Alert informs the provider that they are ordering an expensive test. It is considered a “Soft Stop” because he/she can override the intervention at the point of order entry by choosing the Accept option.

The advent of computerized physician/provider order entry (CPOE), which has become available in most electronic health records (EHRs), affords the unprecedented opportunity to interact with the ordering provider at the point of order entry. These interventions, which utilize clinical decision support tools (CDSTs), come in a variety of forms and are in many ways a two-edged sword. One major advantage of CDSTs is that the laboratory has the opportunity to guide providers or even block test ordering that is inappropriate. Although these interventions may seem obviously beneficial, disadvantages may include interruption of medical care delivery, and provider fatigue and dissatisfaction. Therefore these must be used very judiciously, and it is recommended that they are not implemented until the potential impact has been vetted by a committee that includes some members who are end-users.

Here we see again the importance of a close and collegial working relationship with members of the informatics team. A thorough understanding of the capabilities of the informatics system at the institution is needed, so that as many “behind the scenes” interventions can be undertaken to decrease the interventions that need to be seen and/or acted upon by the provider. As an example, at one of the institutions of the author (GWP), the Associate Chief Medical Informatics Officer has formed and convenes a committee that reviews and approves all interventions prior to build, testing, and deployment. This has been a great addition for a number of reasons. Foremost, the proposed intervention must be thought out, presented, and explained to the committee. This provides an opportunity for discussion and debate, which in many instances results in improvements to the initial proposal. This is in contrast to what often seems to happen, wherein the loudest voice receives an intervention to address a complaint, and in some instances there are untoward consequences. Following agreement by the committee, the interested parties work with the informatics team, who thoroughly understand the capabilities of the system, to build the intervention. Next, the interventions are tested “in silent mode” behind the scenes to determine who will be affected by the intervention and the frequency with which the intervention will be activated. This is important to understand the scope of the intervention, particularly if it requires manual override. Having such a process to thoroughly vet and address intervention requests serves to ensure the validity of the requests, works

to build the best intervention possible, and forecasts who will be affected and to what degree.

There are a variety of clinical decision support tools.<sup>35</sup> The routine availability of these varies somewhat by the vendor of the hospital information system, but there are many similarities.<sup>39-41</sup> In addition, custom programming is possible if vendor provided solutions are not available. A drawback to custom programming is the maintenance, because these must be individually upgraded with each system upgrade. Two of the common types of CDSTs, which will be discussed here in greater detail, include the best practice alert (i.e., a soft stop) and the hard stop (i.e., interruptive).

The best practice alert, as the name implies, is a means of alerting a provider and conveying important information. These alerts can usually be bypassed with a single click at the point of order entry (Fig. 14.1). Unfortunately, they may not be read for a number of reasons, one of the most important of which is alert fatigue. When providers are bombarded with too many best practice alerts or other types of notifications, they simply stop reading them and “click through.” This is unfortunate, as there is an electronic record of notification, which is discoverable and conceivably could have medicolegal implications. Therefore best practice alerts should be used judiciously to diminish alert fatigue and are not the intervention of choice for life-threatening issues.

Best practice alerts have a variety of applications and are commonly used by pharmacy services and blood utilization committees, among others. Examples include a best practice alert to disclose drug-drug interactions, which found a wide range of variability in the responses to the alert<sup>42</sup>; and a best practice alert to decrease unnecessary plasma transfusion, which was automatically inactivated in appropriate clinical situations such as massive transfusions which demonstrated both a high acceptance rate by providers and a substantial reduction in unnecessary transfusions.<sup>43</sup> Engagement of clinical colleagues to develop a CDST for familial hypercholesterolemia has shown the benefits of a collaborative approach<sup>44</sup> and the use of a large-scale quality improvement initiative to promote adherence to guidelines using an interruptive soft stop CDST, led to an improvement of guideline compliance from 32.3 to 58% ( $P < .001$ ).<sup>45</sup> Our experiences, which are similar to those described earlier, have shown that, although soft stops are not as effective as hard stops in preventing unnecessary tests, they can deter some unnecessary testing, contribute to health care cost savings and are less intrusive than hard stops.<sup>41,46,47</sup>

### C. Retrospective Feedback

Retrospective feedback involves a review and analysis of provider ordering patterns. This review may focus on the use of a test, or all tests used by a particular group of providers. This is usually done to detect outlier behavior.<sup>48</sup> The display of results may be done in a manner that keeps the ordering providers anonymous, or it may be done with a key, so that an individual provider can view their ordering pattern within the group without knowing the identity of the other ordering providers. Full disclosure of the identity of all ordering providers for all to see is discouraged; however, the key could be provided to departmental leadership, so that consideration may be given to both underutilizers and overutilizers.

One important recommendation regarding retrospective feedback is to compare providers only in like practices. For example, the monitoring of hemoglobin A<sub>1c</sub> would likely be very different between family physicians and orthopedic surgeons, given their different scopes of practice. Family practitioners commonly care for patients with chronic diseases, such as diabetes mellitus, so comparing the use of this test among such a group makes more sense.

Demonstrating the range of usage of a test within a group of end-users usually evokes a time of reflection. The providers consider how they use the test in comparison with their peers who have a similar practice. It also provides an opportunity for open communication regarding the group's thoughts concerning the optimal use of this test in their setting. This review also provides an opportunity for a literature review to ensure everyone is up to date regarding the optimal use of a test or testing in a particular situation. These discussions may also suggest the subject matter as a topic for grand rounds or other educational venues.

### POINTS TO REMEMBER

- Education should occur prior to implementing new tests or associated with testing changes.
- Electronic real-time feedback is possible through best practice alerts that are available in most hospital information systems.
- The overuse of best practice alerts or other "pop-ups" results in alert fatigue, wherein the ordering provider ignores the alert.
- Retrospective feedback is useful for demonstrating the ordering patterns of a group and initiating conversations concerning best practices.

## 2. Test Order Management

Two broad categories of test order management are test menu control and test order restrictions. The latter category can also be divided into real-time test order restrictions and provider privileging.

### A. Test Menu Control

The structure and presentation of tests in a test menu have a great influence on ordering patterns. Three aspects that will be discussed here: order sets, elimination of outdated tests, and test presentation.

An order set is a combination of tests that are usually collated for the convenience of the provider, which is important for provider satisfaction and operational efficiency. The

caveats with order sets are that they (1) must be appropriately designed, (2) must be appropriately used, (3) must be reviewed on a regular basis, and (4) should have input from individuals with laboratory expertise.

The tests that are contained in an admissions order set for an individual with a particular disease (e.g., a suspected myocardial infarction) should contain the initial tests needed for this type of patient admission. It would be inappropriate to use this order set to *monitor* this same patient. Another order set is needed for monitoring patients, if the ordering of individual tests (i.e., *à la carte* ordering) is not desired. It would be important to clearly name these two order sets to distinguish one from another and provide clear advice on when each set should be used.

What happens when the wrong test is added to an order set? If this is not detected, or if the presence of the unneeded test result is simply ignored, then the unnecessary test may continue to be used for months or years, depending on the review-cycle for order sets. The time when order set construction is requested and the time of order set reviews are good opportunities to review the contents of an order set, assess the need for each test present, and review, optimally with a pathologist or laboratorian, if the tests included are the best tests to detect or monitor the disease state under consideration. In some EHRs, providers can create their own order sets to quickly order tests they commonly use, which can present challenges to utilization efforts.

Another opportunity to improve test utilization is the removal of outdated tests or tests of little clinical utility. The latter is particularly important when a test may be confused with a more medically relevant test. For example, we have seen providers order cryptococcal antibody testing, which has little-to-no clinical use, when they in fact meant to order the cryptococcal antigen test. We have observed similar confusion with serologic tests for influenza and *Bordetella*. These issues were eliminated by removing the less clinically relevant tests from the test menu. A standing review of the test order menu usually discloses opportunities for further optimization.

The order in which tests are presented in the test menu influences test selection. Years ago, our laboratory validated a newly released US Food and Drug Administration (FDA)-approved assay to detect human immunodeficiency virus type 2 (HIV-2). This was done to maintain an optimal test menu, although HIV-2 was uncommon in the United States and the degree of future spread was unclear. We were astonished when the laboratory experienced a deluge of orders for HIV-2. Communications with the providers soon disclosed they were not interested in HIV-2, but rather the more prevalent type, HIV-1. The erroneous orders were tracked back to unit clerks who were placing the orders at the time; these individuals were unaware of the differences between HIV-1 and HIV-2 and chose the HIV-2 test simply because it was listed first in the test menu. A rearrangement of the test order sequence resulted in the virtual elimination of HIV-2 orders—test names and their presentation in the ordering menu matter.

Similarly in a study of two interventions designed to direct clinicians to order a CBC rather than a "CBC w/ Diff" in instances wherein the differential white cell count is not needed, one intervention was purely educational, whereas the other changed test order presentation. There was no statistically significant difference in the ordering pattern after the initial educational intervention but changing the order of these two

tests in the test menu resulted in a statistically significant reduction in “CBC w/ Diff” orders, which was the desired effect.<sup>14</sup>

## B. Test Order Restrictions

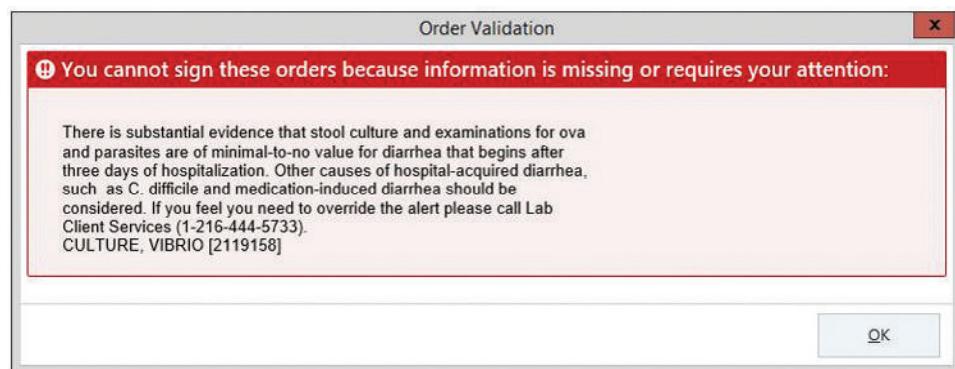
**Real-time test order restrictions.** The hard stop is a CDST that prevents an order from being completed in the order entry system. There are variations on this theme, but usually the provider needs to acknowledge that the test is truly needed before proceeding. This may be accomplished by calling the laboratory to request an override or in certain instances obtaining a specialty consult to obtain approval. These types of interventions are potentially disruptive to care delivery and, even when used appropriately, are often not popular with care providers.

These types of interventions are best employed to stop potential patient harm events (e.g., transfusion of an ABO incompatible blood unit), therapeutic errors (e.g., administration of a penicillin derivative to someone with a penicillin allergy), or, as we have used it, to decrease unnecessary phlebotomy for tests that are truly duplicative and unnecessary (Fig. 14.2).<sup>49</sup> Not surprisingly, hard stop interventions are more effective than best practice alerts or soft stops wherein

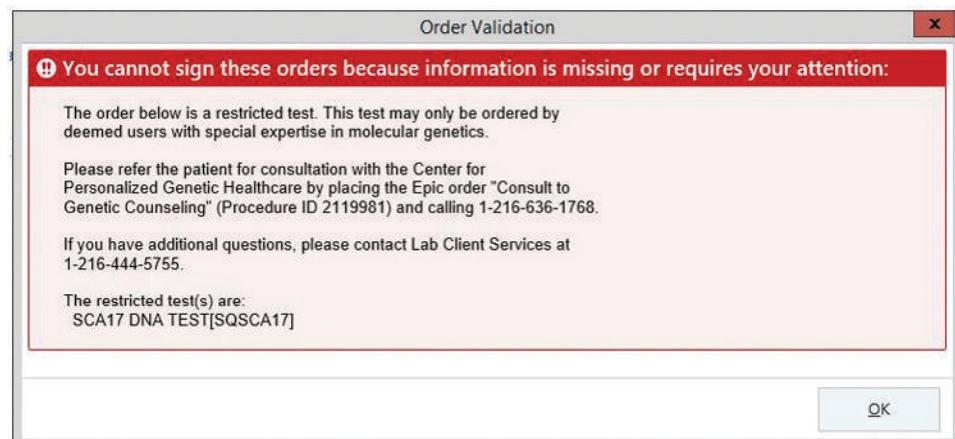
the provider can bypass the intervention at the point of order entry.<sup>46</sup>

**Provider privileging.** Should every physician in your medical group be allowed to order every test? How about every intern or resident? Although providers may have the “legal” right to order a test, group practices may impose local restrictions on tests to those providers who commonly use particular tests in their practice based on the profile of the patients for whom they care. This is a delicate matter because it involves imposed limitations in the practice of medicine. A tiered approach to testing, with reasonable restrictions and avenues for providers to obtain what is needed, has been found to be an acceptable approach.<sup>50</sup>

Hard stop CDSTs may also be used to restrict the ordering of tests that are outside the scope of practice of a physician or other provider (Fig. 14.3).<sup>50</sup> For example, an orthopedic surgeon likely has no reason to order a paraneoplastic panel for the assessment of encephalopathy. In such an instance, it would be appropriate to restrict such testing to the physicians, such as neurologists, who have such a scope of practice. This type of intervention could also be programmed to restrict expensive test orders by residents or fellows and require the order to be filed by staff physicians. It could also be used



**FIGURE 14.2** This is an example of a hard stop that is used to intervene on a test request that is likely unnecessary. Although providers cannot override the intervention at the point of order entry, they are provided with an option to telephone Laboratory Client Services in the event they would like to override this intervention.



**FIGURE 14.3** This is an example of a hard stop that is used to enforce the test ordering restriction for molecular genetic tests. Although providers cannot override the intervention at the point of order entry, they are provided with an option to obtain a Genetics consult, if they feel the test is necessary.

for certain very expensive tests that hospital or laboratory leadership have deemed should not be sent out or performed without an appropriate internal review.

### POINTS TO REMEMBER

- Test Order Control can be broadly categorized as test menu control and test order restrictions.
- The presentation of orders within the ordering menu can influence test selection.
- Clinical decision support tools within the electronic medical record allow for real-time test order restriction, but these must be used judiciously.
- Privileging is a method of test order control that limits the ordering of select tests to care providers with particular expertise.

### CONCLUSION

Laboratory stewardship plays an important role in modern health care delivery and is anticipated to grow in importance as health care delivery continues to evolve. Waste should be taken out of the system but in a thoughtful manner that engages the most knowledgeable stakeholders with the goal of optimizing patient care and not just reducing costs. This is a group activity, a true team effort, and system-based approach that includes pathologists, laboratorians, clinical providers, administrators, finance, informatics, nursing, and more. By maintaining a focus on evidence-based best practices for patient care, the team will remain engaged and leadership receptive to implementing your plans. By monitoring and reporting the effectiveness of your interventions, you and your team will gain credibility, and support for future endeavors will more likely be forthcoming. Most importantly, addressing underutilization will help to ensure patients get the testing that they need, whereas addressing overutilization will help to ensure optimal patient safety and satisfaction while reducing costs. These interventions, *in toto*, often result in cost-savings while maintaining or improving quality of care, which is a formula with which no one can argue.

### SELECTED REFERENCES

2. Conrad DA, Vaughn M, Grembowski D, et al. Implementing value-based payment reform: a conceptual framework and case examples. *Med Care Res Rev* 2016;73:437–57.
4. Balog JE. The meaning of health in the era of value-based care. *Cureus* 2017;9:e1042.
6. Benson ES. Initiatives toward effective decision making and laboratory use. *Hum Pathol* 1980;11:440–8.
8. Kim JY, Dzik WH, Dighe AS, et al. Utilization management in a large urban academic medical center: a 10-year experience. *Am J Clin Pathol* 2011;135:108–18.
9. Howanitz PJ, Perrotta PL, Bashleben CP, et al. Twenty-five years of accomplishments of the College of American Pathologists Q-probes program for clinical pathology. *Arch Pathol Lab Med* 2014;138:1141–9.
10. Lundberg GD. Adding outcome as the 10th step in the brain-to-brain laboratory test loop. *Am J Clin Pathol* 2014;141:767–9.
11. Procop GW, Daley AT, Baron J, et al. GP49 developing and managing a medical laboratory (Test) utilization management program. 1st ed. Wayne, PA: Clinical Laboratory Standards Institute; 2017.
18. Ducatman BS, Ducatman AM, Crawford JM, et al. The value proposition for pathologists: a population health approach. *Acad Pathol* 2020;7:2374289519898857.
21. Lang T, Croal B. National minimum retesting intervals in pathology: a final report detailing consensus recommendations for minimum retesting intervals for use in pathology. 2015.
22. Ashton CM, Petersen NJ, Soucek J, et al. Geographic variations in utilization rates in Veterans Affairs hospitals and clinics. *N Engl J Med* 1999;340:32–9.
23. Cabana MD, Rand CS, Powe NR, et al. Why don't physicians follow clinical practice guidelines? A framework for improvement. *JAMA* 1999;282:1458–65.
24. Daniels M, Schroeder SA. Variation among physicians in use of laboratory tests. II. Relation to clinical productivity and outcomes of care. *Med Care* 1977;15:482–7.
34. Wakefield E, Keller H, Mianzo H, et al. Reduction of health care costs and improved appropriateness of incoming test orders: the impact of genetic counselor review in an academic genetic testing laboratory. *J Genet Couns* 2018;27:1067–73.
35. Baron JM, Dighe AS. The role of informatics and decision support in utilization management. *Clin Chim Acta* 2014;427:196–201.
36. Baird G. The laboratory test utilization management toolbox. *Biochem Med (Zagreb)* 2014;24:223–34.
37. Rubinstein M, Hirsch R, Bandyopadhyay K, et al. Effectiveness of practices to support appropriate laboratory test utilization: a laboratory medicine best practices systematic review and meta-analysis. *Am J Clin Pathol* 2018;149:197–221.
39. Elwyn G, Scholl I, Tietbohl C, et al. “Many miles to go ...”: a systematic review of the implementation of patient decision support interventions into routine clinical practice. *BMC Med Inform Decis Mak* 2013;13(Suppl 2):S14.
41. Procop GW, Weathers AL, Reddy AJ. Operational aspects of a clinical decision support program. *Clin Lab Med* 2019;39:215–29.
45. Ducatman AM, Tacker DH, Ducatman BS, et al. Quality improvement intervention for reduction of redundant testing. *Acad Pathol* 2017;4:2374289517707506.
50. Riley JD, Procop GW, Kottke-Marchant K, Wyllie R, Lacbawan FL. Improving molecular genetic test utilization through order restriction, test review, and guidance. *J Mol Diagn* 2015;17:225–9.

## REFERENCES

1. Conrad DA. The theory of value-based payment incentives and their application to health care. *Health Serv Res* 2015; 50(Suppl 2):2057–89.
2. Conrad DA, Vaughn M, Grembowski D, et al. Implementing value-based payment reform: a conceptual framework and case examples. *Med Care Res Rev* 2016;73:437–57.
3. Nash DB. The dream of value-based care. *Am Health Drug Benefits* 2017;10:5–6.
4. Balog JE. The meaning of health in the era of value-based care. *Cureus* 2017;9:e1042.
5. Chen SL, Coffron MR. MACRA and the changing medicare payment landscape. *Ann Surg Oncol* 2017.
6. Benson ES. Initiatives toward effective decision making and laboratory use. *Hum Pathol* 1980;11:440–8.
7. Robinson A. Rationale for cost-effective laboratory medicine. *Clin Microbiol Rev* 1994;7:185–99.
8. Kim JY, Dzik WH, Dighe AS, et al. Utilization management in a large urban academic medical center: a 10-year experience. *Am J Clin Pathol* 2011;135:108–18.
9. Howanitz PJ, Perrotta PL, Bashleben CP, et al. Twenty-five years of accomplishments of the College of American Pathologists Q-probes program for clinical pathology. *Arch Pathol Lab Med* 2014;138:1141–9.
10. Lundberg GD. Adding outcome as the 10th step in the brain-to-brain laboratory test loop. *Am J Clin Pathol* 2014;141: 767–9.
11. Procop GW, Daley AT, Baron J, et al. GP49 developing and managing a medical laboratory (Test) utilization management program. 1st ed. Wayne, PA: Clinical Laboratory Standards Institute; 2017.
12. Procop GW, Nelson SK, Blond BJ, et al. The impact of transit times on the detection of bacterial pathogens in blood cultures: a college of American Pathologists Q-Probes study of 36 institutions. *Arch Pathol Lab Med* 2019.
13. Phelan MP, et al. Impact of interventions to change CBC and differential ordering patterns in the emergency department. *Am J Clin Pathol* 2018.
14. Phelan MP, Nakashima MO, Good DM, et al. Impact of interventions to change CBC and differential ordering patterns in the emergency department. *Am J Clin Pathol* 2019;151:194–7.
15. Phelan MP, Reineks EZ, Hustey FM, et al. Does pneumatic tube system transport contribute to hemolysis in ED blood samples? *West J Emerg Med* 2016;17:557–60.
16. Phelan MP, Reineks EZ, Schold JD, et al. Preanalytic factors associated with hemolysis in emergency department blood samples. *Arch Pathol Lab Med* 2018;142:229–35.
17. Phelan MP, Reineks EZ, Schold JD, et al. Estimated national volume of laboratory results affected by hemolyzed specimens from emergency departments. *Arch Pathol Lab Med* 2016;140:621.
18. Ducatman BS, Ducatman AM, Crawford JM, et al. The value proposition for pathologists: a population health approach. *Acad Pathol* 2020;7:2374289519898857.
19. Zhi M, Ding EL, Theisen-Toupal J, et al. The landscape of inappropriate laboratory testing: a 15-year meta-analysis. *PLoS One* 2013;8:e78962.
20. Driskell OJ, Holland D, Hanna FH, et al. Inappropriate requesting of glycated hemoglobin (Hb A1c) is widespread: assessment of prevalence, impact of national guidance, and practice-to-practice variability. *Clin Chem* 2012;58:906–15.
21. Lang T, Croal B. National minimum retesting intervals in pathology: a final report detailing consensus recommendations for minimum retesting intervals for use in pathology. 2015.
22. Ashton CM, Petersen NJ, Soucek J, et al. Geographic variations in utilization rates in Veterans Affairs hospitals and clinics. *N Engl J Med* 1999;340:32–9.
23. Cabana MD, Rand CS, Powe NR, et al. Why don't physicians follow clinical practice guidelines? A framework for improvement. *JAMA* 1999;282:1458–65.
24. Daniels M, Schroeder SA. Variation among physicians in use of laboratory tests. II. Relation to clinical productivity and outcomes of care. *Med Care* 1977;15:482–7.
25. Hindorff LA, Burke W, Laberge AM, et al. Motivating factors for physician ordering of factor V Leiden genetic tests. *Arch Intern Med* 2009;169:68–74.
26. Pelaccia T, Plotnick LH, Audetat MC, et al. A scoping review of physicians' clinical reasoning in emergency departments. *Ann Emerg Med* 2020;75:206–17.
27. Rubins D, Boxer R, Landman A, Wright A. Effect of default order set settings on telemetry ordering. *J Am Med Inform Assoc* 2019;26:1488–92.
28. Sedrak MS, Patel MS, Ziembra JB, et al. Residents' self-report on why they order perceived unnecessary inpatient laboratory tests. *J Hosp Med* 2016;11:869–72.
29. Faisal A, Andres K, Rind JAK, et al. Reducing the number of unnecessary routine laboratory tests through education of internal medicine residents. *Postgrad Med J* 2018;94:716–9.
30. Lyon AW, Chin AC, Slotsve GA, et al. Simulation of repetitive diagnostic blood loss and onset of iatrogenic anemia in critical care patients with a mathematical model. *Comput Biol Med* 2012.
31. Tosiri P, Kanitsap N, Kanitsap A. Approximate iatrogenic blood loss in medical intensive care patients and the causes of anemia. *J Med Assoc Thai* 2010;93(Suppl 7):S271–6.
32. Ambasta A, Pancic S, Wong BM, et al. Expert recommendations on frequency of utilization of common laboratory tests in medical inpatients: a canadian consensus study. *J Gen Intern Med* 2019;34:2786–95.
33. Waltman L, Runke C, Balcom J, et al. Further Defining the Role of the Laboratory Genetic Counselor. *J Genet Couns* 2016;25:786–98.
34. Wakefield E, Keller H, Mianzo H, et al. Reduction of health care costs and improved appropriateness of incoming test orders: the impact of genetic counselor review in an academic genetic testing laboratory. *J Genet Couns* 2018;27:1067–73.
35. Baron JM, Dighe AS. The role of informatics and decision support in utilization management. *Clin Chim Acta* 2014; 427:196–201.
36. Baird G. The laboratory test utilization management toolbox. *Biochem Med (Zagreb)* 2014;24:223–34.
37. Rubinstein M, Hirsch R, Bandyopadhyay K, et al. Effectiveness of practices to support appropriate laboratory test utilization: a laboratory medicine best practices systematic review and meta-analysis. *Am J Clin Pathol* 2018;149:197–221.
38. Phillips Jr RL, Petterson SM, Bazemore AW, Wingrove PW, Puffer JC. The effects of training institution practice costs, quality, and other characteristics on future practice. *Ann Fam Med* 2017;15:140–8.
39. Elwyn G, Scholl I, Tietbohl C, et al. "Many miles to go ...": a systematic review of the implementation of patient decision support interventions into routine clinical practice. *BMC Med Inform Decis Mak* 2013;13(Suppl 2):S14.

40. Graham MM, James MT, Spertus JA. Decision support tools: realizing the potential to improve quality of care. *Can J Cardiol* 2018;34:821–6.
41. Procop GW, Weathers AL, Reddy AJ. Operational aspects of a clinical decision support program. *Clin Lab Med* 2019;39: 215–29.
42. Cho I, Lee Y, Lee JH, Bates DW. Wide variation and patterns of physicians' responses to drug-drug interaction alerts. *Int J Qual Health Care* 2018.
43. Shah N, Baker SA, Spain D, et al. Real-time clinical decision support decreases inappropriate plasma transfusion. *Am J Clin Pathol* 2017;148:154–60.
44. Hasnie AA, Kumbamu A, Safarova MS, Caraballo PJ, Kullo IJ. A clinical decision support tool for familial hypercholesterolemia based on physician input. *Mayo Clin Proc Innov Qual Outcomes* 2018;2:103–12.
45. Ducatman AM, Tacker DH, Ducatman BS, et al. Quality improvement intervention for reduction of redundant testing. *Acad Pathol* 2017;4:2374289517707506.
46. Procop GW, Keating C, Stagno P, et al. Reducing duplicate testing: a comparison of two clinical decision support tools. *Am J Clin Pathol* 2015;143:623–6.
47. Riley JD, Stanley G, Kottke-Marchant K, Procop GW. The impact of an electronic expensive test notification. *Am J Clin Pathol* 2018;149:530–5.
48. Hirsch O, Donner-Banzhoff N, Schulz M, Erhart M. Detecting and visualizing outliers in provider profiling using funnel plots and mixed effects models—an example from prescription claims data. *Int J Environ Res Public Health* 2018;15:2015.
49. Procop GW, Yerian LM, Wyllie R, Harrison AM, Kottke-Marchant K. Duplicate laboratory test reduction using a clinical decision support tool. *Am J Clin Pathol* 2014;141:718–23.
50. Riley JD, Procop GW, Kottke-Marchant K, Wyllie R, Lacbawan FL. Improving molecular genetic test utilization through order restriction, test review, and guidance. *J Mol Diagn* 2015;17:225–9.

## MULTIPLE CHOICE QUESTIONS

1. Of the following, which is the most important characteristic of a Test Utilization Committee that is charged with overseeing an institution's Laboratory Stewardship Program?
  - a. The committee meets frequently
  - b. The committee reports to the chief medical officer (or equivalent)
  - c. The committee includes engaged members from different specialties (e.g., medicine, cardiology, surgery)
  - d. The committee is led by a pathologist or doctoral level scientist
  - e. The committee has direct authority over the clinical laboratory
2. Which of the following is the most likely potential beneficial effect of reducing unnecessary testing in hospitalized patients?
  - a. Hospital-acquired anemia
  - b. Increased length of stay
  - c. Improved patient outcomes
  - d. Reduced phlebotomy and patient discomfort
  - e. Increased testing costs
3. Which of the following is the most effective way to reduce the number of repetitive tests (e.g., repeat CBC and basic metabolic panel [BMP]) performed on inpatients?
  - a. Publicize testing guidelines through laboratory newsletters and educational conferences
  - b. Display turnaround time information at time of ordering
  - c. Establish a satellite laboratory near intensive care units
  - d. Implement local policies for routine inpatient phlebotomy schedules
  - e. Schedule annual educational conferences related to test utilization
4. Your institution has implemented best practice alerts or "pop-ups" in your electronic test ordering system with the goal of reducing duplicate testing. Which of the following is the most common reason why such alerts are not accepted by health care organizations?
  - a. They can lead to user fatigue and frustration
  - b. They are difficult to maintain in electronic ordering systems
  - c. They are difficult to build in most electronic ordering systems
  - d. Health care practitioners do not accept assistance with test ordering
  - e. Many other techniques are more effective to reduce duplicate testing
5. Which of the following interventions will be most effective in decreasing the number of high-cost molecular genetic tests ordered by health care providers who do not commonly order such testing?
  - a. Display alternatives to the requested molecular test at time of order entry
  - b. Display the cost of the test at order entry
  - c. Display an "alert" at order entry asking the provider to confirm the test request
  - d. Retrospectively review requests for molecular genetic tests
  - e. Restrict ordering of certain molecular genetic tests to providers who commonly order these tests
6. Which of the following interventions is most likely to reduce the number of red blood cell (RBC) folate tests performed to identify folate deficiency?
  - a. Remove the RBC folate test from the test formulary
  - b. Restrict use of RBC folate to primary care providers
  - c. Reflexively perform RBC folate testing for patients with an elevated mean corpuscular volume (MCV)
  - d. Add an electronic "alert" at time of order entry suggesting that serum folate testing is preferred
  - e. Place RBC folate testing on admission order sets
7. From a clinical perspective, which of the following is the most important reason to optimize the utilization of laboratory tests?
  - a. To improve patient outcomes
  - b. To establish accurate diagnoses in a timely manner
  - c. To decrease laboratory costs
  - d. To decrease patient costs
  - e. To reduce false negative laboratory results
8. Which of the following is a frequent cause of unnecessary duplicate testing?
  - a. Providers from different services order the same test at about the same time
  - b. Laboratory analyzers repeat tests with abnormal results
  - c. Providers order repeat testing because the initial test results are unexpected
  - d. Intentional duplicate ordering to increase revenue
  - e. Patients are transferred to another facility where tests are repeated
9. Which of the following interventions used to change test ordering practices can be effective, but is often only transiently so?
  - a. Removing antiquated tests from a test menu
  - b. Restricting the ability of ordering tests to specific hospital services
  - c. Educating ordering providers on test utilization
  - d. Providing testing costs on the laboratory website
  - e. Developing local testing guidelines for specific diseases
10. Which of the following is the most frequent cause of diagnostic test *overutilization* in hospitalized patients?
  - a. Daily ordering of routine tests (e.g., CBC, BMP)
  - b. Increasing use of specialized diagnostic tests
  - c. Increasing use of advanced imaging techniques
  - d. Accidental ordering of duplicate tests by different providers
  - e. Patient demand for repetitive testing
11. A genetics laboratory has testing specialists discuss all orders for genomic microarray testing with the ordering provider at the time of test request. This test utilization intervention is best classified as of which type?
  - a. Retrospective order review
  - b. Real-time best practice alert
  - c. Prospective order review
  - d. Test menu restriction
  - e. "Soft" electronic ordering alert

12. Which of the following philosophies can help test utilization committees to remain focused on the most important aspects of test utilization?
- Focus on both overutilization *and* underutilization of diagnostic tests
  - Focus on laboratory cost reduction as the highest priority
  - Focus on the diagnostic accuracy of tests
  - Focus mostly on high-cost testing
  - Focus on reducing reference laboratory testing
13. Which of the following activities of a test utilization committee related to oversight of a hospital's test formulary (i.e., test menu) is most likely to reduce the number of unnecessary tests?
- Providing turnaround time goals for testing in the laboratory formulary
  - Providing up-to-date reference ranges within the test formulary
  - Providing indications for testing in the test formulary
  - Using the most currently accepted names for specific tests
  - Removing antiquated tests from the test menu

# Principles of Basic Techniques and Laboratory Safety\*

*Tahir S. Pillay and Stanley F. Lo<sup>a</sup>*

## ABSTRACT

### Background

To appropriately interpret clinical laboratory test results and adequately validate assays, the basic principles and techniques of analytical chemistry need to be understood. These techniques should be used by laboratory professionals in a safe testing environment.

### Content

Factors that affect the analytical process and operation of the clinical laboratory are described in this chapter. The concepts of solute and solution, and the international system of units used to standardize their expression and reporting are described. These solutions are composed of various types of chemicals used in the development of clinical laboratory assays. The importance of water purity, appropriate reagent

preparation, and the different types of reference materials are addressed. The principles of basic techniques in the clinical laboratory, including pipetting, centrifugation, radioactivity, gravimetry, and thermometry, are also discussed. These techniques are used in a variety of laboratory tasks such as making buffers, performing dilutions, evaporation, lyophilization, and filtration. Safety is a constant and crucial concern for laboratory personnel. Each laboratory must create a comprehensive safety program. Plans for the handling of chemicals, exposure to blood-borne pathogens, tuberculosis, and other highly infectious agents are necessary components of a safety plan. The training of laboratory personnel to identify various types of biological, chemical, and electrical hazards and to react appropriately to fire must be addressed in such a plan.

\*The full version of this chapter is available electronically on [ExpertConsult.com](#).

<sup>a</sup>The author gratefully acknowledges the original contributions of Drs. Edward W. Bermes, Stephen E. Kahn, Donald S. Young, Edward R. Powsner, and John C. Widman, on which portions of this chapter are based.