CLINICAL AND LABORATORY STANDARDS INSTITUTE®

November 2011

EP24-A2

Assessment of the Diagnostic Accuracy of Laboratory Tests Using Receiver Operating Characteristic Curves; Approved Guideline—Second Edition

This document provides a protocol for evaluating the accuracy of a test to discriminate between two subclasses of subjects when there is some clinically relevant reason to separate them. In addition to the use of receiver operating characteristic curves and the comparison of two curves, the document emphasizes the importance of defining the question, selecting the sample group, and determining the "true" clinical state.

A guideline for global application developed through the Clinical and Laboratory Standards Institute consensus process.

Clinical and Laboratory Standards Institute

Setting the standard for quality in clinical laboratory testing around the world.

The Clinical and Laboratory Standards Institute (CLSI) is a not-for-profit membership organization that brings together the varied perspectives and expertise of the worldwide laboratory community for the advancement of a common cause: to foster excellence in laboratory medicine by developing and implementing clinical laboratory standards and guidelines that help laboratories fulfill their responsibilities with efficiency, effectiveness, and global applicability.

Consensus Process

Consensus—the substantial agreement by materially affected, competent, and interested parties—is core to the development of all CLSI documents. It does not always connote unanimous agreement, but does mean that the participants in the development of a consensus document have considered and resolved all relevant objections and accept the resulting agreement.

Commenting on Documents

CLSI documents undergo periodic evaluation and modification to keep pace with advancements in technologies, procedures, methods, and protocols affecting the laboratory or health care.

CLSI's consensus process depends on experts who volunteer to serve as contributing authors and/or as participants in the reviewing and commenting process. At the end of each comment period, the committee that developed the document is obligated to review all comments, respond in writing to all substantive comments, and revise the draft document as appropriate.

Comments on published CLSI documents are equally essential, and may be submitted by anyone, at any time, on any document. All comments are addressed according to the consensus process by a committee of experts.

Appeals Process

If it is believed that an objection has not been adequately addressed, the process for appeals is documented in the CLSI Administrative Procedures.

All comments and responses submitted on draft and published documents are retained on file at CLSI and are available upon request.

Get Involved—Volunteer!

Do you use CLSI documents in your workplace? Do you see room for improvement? Would you like to get involved in the revision process? Or maybe you see a need to develop a new document for an emerging technology? CLSI wants to hear from you. We are always looking for volunteers. By donating your time and talents to improve the standards that affect your own work, you will play an active role in improving public health across the globe.

For further information on committee participation or to submit comments, contact CLSI.

Clinical and Laboratory Standards Institute
950 West Valley Road, Suite 2500
Wayne, PA 19087 USA
P: 610.688.0100
F: 610.688.0700
www.clsi.org
standard@clsi.org

ISBN 1-56238-777-4 (Print)
ISBN 1-56238-778-2 (Electronic)
ISSN 1558-6502 (Print)
ISSN 2162-2914 (Electronic)

EP24-A2
Vol. 31 No. 23
Replaces GP10-A
Vol. 15 No. 19

Assessment of the Diagnostic Accuracy of Laboratory Tests Using Receiver Operating Characteristic Curves; Approved Guideline—Second Edition

Volume 31 Number 23

Martin H. Kroll, MD
Bipasa Biswas
Jeffrey R. Budd, PhD
Paul Durham, MA
Robert T. Gorman, PhD
Thomas E. Gwise, PhD
Abdel-Baset Halim, PharmD, PhD, DABCC
Aristides T. Hatjimihail, MD, PhD
Jørgen Hilden, MD
Kyunghee Song, PhD

Abstract

Clinical and Laboratory Standards Institute document EP24-A2—*Assessment of the Diagnostic Accuracy of Laboratory Tests Using Receiver Operating Characteristic Curves; Approved Guideline—Second Edition* provides guidance for laboratorians and manufacturers who assess clinical test accuracy. It is not a recipe; rather, it is a set of concepts to be used to design an assessment of test performance or to interpret data generated by others. In addition to the use of ROC curves and comparison of two curves, the document emphasizes the importance of defining the question, selecting a sample group, and determining the "true" clinical state. The statistical data generated can be useful whether one is considering replacing an existing test, creating or adding a new test, or eliminating a current test.

Clinical and Laboratory Standards Institute (CLSI). *Assessment of the Diagnostic Accuracy of Laboratory Tests Using Receiver Operating Characteristic Curves; Approved Guideline—Second Edition.* CLSI document EP24-A2 (ISBN 1-56238-777-4 [Print]; ISBN 1-56238-778-2 [Electronic]). Clinical and Laboratory Standards Institute, 950 West Valley Road, Suite 2500, Wayne, Pennsylvania 19087 USA, 2011.

Suggested Citation

CLSI. *Assessment of the Diagnostic Accuracy of Laboratory Tests Using Receiver Operating Characteristic Curves; Approved Guideline—Second Edition.* CLSI document EP24-A2. Wayne, PA: Clinical and Laboratory Standards Institute; 2011.

Proposed Guideline
March 1987

Tentative Guideline
December 1993

Approved Guideline
December 1995

Approved Guideline—Second Edition
November 2011

ISBN 1-56238-777-4 (Print)
ISBN 1-56238-778-2 (Electronic)
ISSN 1558-6502 (Print)
ISSN 2162-2914 (Electronic)

Committee Membership

Consensus Committee on Evaluation Protocols

Greg Cooper, CLS, MHA
Chairholder
W. Gregory Cooper LLC
Denton, Texas, USA

James F. Pierson-Perry
Vice-Chairholder
Siemens Healthcare Diagnostics
Newark, Delaware, USA

J. Rex Astles, PhD, FACB, DABCC
Centers for Disease Control and Prevention
Atlanta, Georgia, USA

Jeffrey R. Budd, PhD
Beckman Coulter
Chaska, Minnesota, USA

Jonathan Guy Middle, PhD
University Hospital Birmingham NHS Trust
Birmingham, United Kingdom

Mitchell G. Scott, PhD
Washington University School of Medicine
St. Louis, Missouri, USA

Lakshmi Vishnuvajjala, PhD
FDA Center for Devices and Radiological Health
Silver Spring, Maryland, USA

Document Development Committee on Assessment of Diagnostic Accuracy of Laboratory Tests Using ROC Curves

Martin H. Kroll, MD
Chairholder
Boston Medical Center
Boston, Massachusetts, USA

Jeffrey R. Budd, PhD
Beckman Coulter
Chaska, Minnesota, USA

Robert T. Gorman, PhD
Siemens Healthcare Diagnostics Inc.
Newark, Delaware, USA

Jørgen Hilden, MD
University of Copenhagen
Copenhagen, Denmark

Kyunghee Song, PhD
FDA Center for Devices and Radiological Health
Silver Spring, Maryland, USA

Staff

Clinical and Laboratory Standards Institute
Wayne, Pennsylvania, USA

Luann Ochs, MS
Vice President, Standards Development

Ron S. Quicho
Staff Liaison

Patrice E. Polgar
Project Manager

Megan P. Larrisey, MA
Editor

Acknowledgment

CLSI and the Consensus Committee on Evaluation Protocols gratefully acknowledge the following individuals for their help in preparing this document:

Bipasa Biswas
FDA Center for Devices and Radiological Health
Silver Spring, Maryland, USA

Paul Durham, MA
Culver City, California

Thomas E. Gwise, PhD
FDA Center for Drug Evaluation and Research
Silver Spring, Maryland, USA

Abdel-Baset Halim, PharmD, PhD, DABCC
Daiichi Sankyo Pharma Development
Edison, New Jersey, USA

Aristides T. Hatjimihail, MD, PhD
Hellenic Complex Systems Laboratory
Drama, Greece

Contents

Foreword

Laboratorians, investigators, *in vitro* diagnostic manufacturers, and clinicians are often interested in how well a test performs clinically. This is true whether considering replacing an existing test with a newer one, adding a new test to the laboratory's menu, eliminating tests where possible, or evaluating the diagnostic power of a laboratory test relative to another clinical or diagnostic tool. This project was originally intended to make recommendations about assessing the clinical performance of diagnostic tests. The concepts of Swets and Pickett[1] were adopted, whereby clinical performance is divided into (1) a discriminatory or diagnostic element (diagnostic accuracy) and (2) a decision or efficacy element. Laboratory tests are ordered to help answer questions about patient management. How much help an individual test result provides is variable and, in any case, a highly complicated issue. Management decisions and strategies are complex activities that require the physician to consider probabilities of disease, quality of the data available, effectiveness of various treatment/management alternatives, probability of outcomes, and value (and cost) of outcomes to the patient. Many types of clinical data (including laboratory results) are usually integrated into a complex decision-making process. Most often, a single laboratory test result is not the sole basis for a diagnosis or a patient-management decision.

Therefore, some have criticized the practice of evaluating the diagnostic performance of a test as if it were used alone. However, each clinical tool (eg, a clinical laboratory test, an electroencephalogram, an electrocardiogram, a nuclide scan, an X-ray, a biopsy, a pulmonary function test, or a sonogram) is meant to make some definable discrimination. It is important to know just how inherently accurate each test is as a diagnostic discriminator. *Note that assessing diagnostic accuracy, without engaging in comprehensive clinical decision analysis, is a valid and useful activity for the clinical laboratory.* Diagnostic accuracy is the most fundamental characteristic of the test itself as a classification device; it measures the ability of the test to discriminate among alternative states of health. In the simplest form, this property is the ability to distinguish between just two states of health or circumstances. Sometimes this involves distinguishing health from disease; other times it might involve distinguishing between benign and malignant disease, categorizing subjects as responding to therapy vs those not responding, or predicting who will become ill vs who will not. This ability to distinguish or discriminate between two states among subjects is a property of the test itself.

Indeed, the ability of the test to distinguish between the relevant alternative states or conditions of the subject (ie, diagnostic accuracy) is the most basic property of a laboratory test as a device to help in decision making. Note that this basic property cannot be separated from the clinical problem being addressed and the spectrum effect of the mix of subject states on which the test system is based. This property is the place to start when assessing the value of a test in the patient-management process.

Exploration of the usefulness of medical information, such as test data, involves a number of factors or parameters that are not properties of the test system; rather, they are properties of the circumstances of the clinical application. These include the probability or prevalence of disease, the possible clinical outcomes and the relative values of diagnostic outcomes, the costs to the patient (and others) of incorrect information (false-positive and false-negative classifications), and the costs and benefits of various treatment options. These characteristics or properties form the context in which test information is used, but are not properties of the test system. These factors interact with test results to affect the usefulness of the test, but do not affect test accuracy.

In summary, diagnostic accuracy is defined as the basic ability to discriminate between two subclasses of subjects when there is some clinically relevant reason to separate them. This concept of diagnostic accuracy refers to the quality of the information (classification) provided by the test, which should be distinguished from the practical usefulness of the information.[1] Both are aspects of test performance. The assessment of diagnostic accuracy is the place to start in evaluating test performance. If a test cannot discriminate between clinically relevant subclasses of subjects, then there is little incentive to further explore a possible clinical role. If, on the other hand, a test does exhibit a substantial ability to

discriminate, then by examining the degree of accuracy of the test and/or by comparing its accuracy to that of other tests, one can decide whether to delve into a more complex assessment of its role in patient management (decision analysis). This document addresses the assessment of diagnostic accuracy but not the analysis of usefulness or the role of the test in the patient-management process.

In this second edition of the guideline, the document development committee has provided more details on the construction and interpretation of receiver operating characteristic (ROC) curves. Many more examples are included to help the reader assess an individual curve and its associated area under the curve, as well as to compare two curves. Sample size calculations are provided for the first time.

NOTE: Although a step-by-step technique for generating ROC curves has been presented in EP24, it is assumed that most users of this guideline will access commercially available software for this task.

Key Words

Area under the curve, diagnostic accuracy, false-negative fraction, false-positive fraction, medical decision level, receiver operating characteristic curve, sensitivity, specificity, true-negative fraction, true-positive fraction

Assessment of the Diagnostic Accuracy of Laboratory Tests Using Receiver Operating Characteristic Curves; Approved Guideline—Second Edition

1 Scope

This guideline outlines the steps and principles of prospectively planned and retrospective studies to evaluate the intrinsic diagnostic accuracy of a clinical laboratory test, defined as its fundamental ability to discriminate correctly among alternative states of health. It is not intended to help determine how best to use a diagnostic test in clinical practice, but instead to determine how accurate a laboratory test is in terms of diagnostic sensitivity and specificity.

Receiver operating characteristic (ROC) curve methodology arose in response to needs in electronic signal detection and problems with radar in the early 1950s.[2] It is derived from conditional probabilities, as originally formulated by Bayes.[3] This guideline aims to define ROC curves and to explain how to design, construct, interpret, and apply the information from ROC studies to evaluate diagnostic tests. For simplicity, only continuous scales, such as those typical for *in vitro* diagnostic tests, are discussed. The clinical condition that the test is intended to detect must be verifiable through some means other than the test under investigation. In other words, there must be an independent clinical reference standard against which one can compare the test. By selecting cutoffs between positive and negative diagnoses along the continuous scale of the test, the diagnostic outcomes for these decision levels are compared to the true clinical condition, which, in turn, generates the ROC curve.

This guideline will be of value to a wide variety of possible users, including:

- Investigators who are developing new tests for specific applications
- Manufacturers of reagents and devices for performing tests who are interested in assessing or validating test performance in terms of diagnostic accuracy
- Regulatory agencies interested in establishing requirements for claims related to diagnostic accuracy
- Clinical laboratorians who are reviewing data or the literature, and/or generating their own data, to make decisions about which tests to employ in their laboratories
- Health care or scientific workers interested in critical evaluation of data being presented on clinical test performance

2 Introduction

An ROC curve provides the following advantageous properties:

- It visually displays the performance of one or more diagnostic markers or tests across the entire measuring interval.
- By plotting unitless values (sensitivity vs specificity or sensitivity vs 1 − specificity), one can compare the diagnostic performance of two or more diagnostic markers or tests regardless of:
  – Units of expression of different markers or tools (eg, mg/dL, mmol/L, U/L)
  – Type of diagnostic test (eg, a clinical laboratory test, pulmonary function test, radiography)
  – Type of biological sample analyzed (eg, serum vs urine, saliva vs blood)

- It gives a clinician flexibility to select the appropriate medical decision level depending upon the medical situation and the clinical setting. (**NOTE:** In a pivotal study, selecting the optimal cutoff and evaluating the diagnostic accuracy in the same study leads to the biased estimation [overestimation] of the diagnostic accuracy. These issues are discussed in detail in the literature.[4-6])

By evaluating (or examining) ROC based on a marker, the clinician could choose a decision level offering high sensitivity but lower specificity. In another situation using this marker, the clinician could choose a different decision level offering high specificity but lower sensitivity to reduce false positives (FPs).

3 Standard Precautions

Because it is often impossible to know what isolates or specimens might be infectious, all patient and laboratory specimens are treated as infectious and handled according to "standard precautions." Standard precautions are guidelines that combine the major features of "universal precautions and body substance isolation" practices. Standard precautions cover the transmission of all known infectious agents and thus are more comprehensive than universal precautions, which are intended to apply only to transmission of blood-borne pathogens. Standard and universal precaution guidelines are available from the Centers for Disease Control and Prevention.[7] For specific precautions for preventing the laboratory transmission of all known infectious agents from laboratory instruments and materials and for recommendations for the management of exposure to all known infectious diseases, refer to CLSI document M29.[8]

4 Terminology

4.1 A Note on Terminology

CLSI, as a global leader in standardization, is firmly committed to achieving global harmonization wherever possible. Harmonization is a process of recognizing, understanding, and explaining differences while taking steps to achieve worldwide uniformity. CLSI recognizes that medical conventions in the global metrological community have evolved differently in the United States, Europe, and elsewhere; that these differences are reflected in CLSI, International Organization for Standardization (ISO), and European Committee for Standardization (CEN) documents; and that legally required use of terms, regional usage, and different consensus timelines are all important considerations in the harmonization process. In light of this, CLSI's consensus process for development and revision of standards and guidelines focuses on harmonization of terms to facilitate the global application of standards and guidelines.

Essentially, new documents are obliged to adhere to the most current version of the *International Vocabulary of Metrology – Basic and General Concepts and Associated Terms* (VIM)[9] whenever an ambiguity occurs in the interpretation or understanding of terms. In the latest edition, many definitions have become more explicit and understandable, but the language of the VIM is difficult and compact. VIM deals with general metrology and terminology that should be useful for most disciplines that measure quantities.

The understanding of a few terms has changed during the last decade as the concepts have developed. *Precision* (measurement precision) is defined as closeness of agreement between indications or measured quantity values obtained by replicate measurements on the same or similar objects under specified conditions. The term *measurand* is used when referring to the quantity intended to be measured, instead of *analyte* (component represented in the name of a measurable quantity) when its use relates to a biological fluid/matrix. Additionally, *clinical accuracy* has been changed to *diagnostic accuracy* because the term "clinical" has a regulatory connotation in Europe and elsewhere.

4.2 Definitions

clinical state – a state of health or disease that has been defined by either a clinical definition or some other independent reference standard; **NOTE:** Examples of clinical states include "no disease found," "disease 1" (where 1 represents the first clinical state under consideration), "disease 2" (where 2 represents the second clinical state under investigation), and so on.

decision level//decision threshold//decision point//cutoff level – a test value or statistic that marks the upper (or lower) boundary between diagnostic categories, ie, between negative (acceptable or unaffected) results and positive (unacceptable or affected) results.

diagnostic accuracy (clinical accuracy) – the ability of a diagnostic test to discriminate between diseased and nondiseased subjects, or between two or more clinical states; **NOTE:** An example would be discrimination between rheumatoid arthritis and systemic lupus erythematosus.

diagnostic test – a measurement or examination used to classify subjects into a particular class or clinical state; **NOTE:** Laboratory tests are often called "*in vitro* diagnostic" tests.

distribution-free (statistical procedure) – one that does not presuppose that the data arise from a distribution of a particular kind, such as the normal (gaussian) family of distributions; **NOTE 1:** A near-synonym is "nonparametric" (see definition for **nonparametric,** below); **NOTE 2:** For example, drawing a histogram is a simple distribution-free operation, as is any "local" maneuver aimed at smoothing the histogram or smoothing a trend. Any procedure exclusively based on an ordering (ranking) of observations, rather than on their numerical values, is also distribution-free; **NOTE 3:** "Distribution-free" does not mean "assumption-free." Assumptions of representative (fair) sampling and independence (independent observations), for instance, are universal.

false-negative fraction (FNF) – ratio of subjects who have the disease, but who have a negative test result, to all subjects who have the disease; FN / (FN + true positive [TP]); equivalent to (1 − sensitivity).

false-negative (FN) result – negative test result for a subject in whom the disease or condition of interest is present.

false-positive fraction (FPF) – ratio of subjects who do not have the disease, but who have a positive test result, to all subjects who do not have the disease; FP / (FP + true negative [TN]); same as (1 − specificity).

false-positive (FP) result – positive test result for a subject in whom the disease or condition of interest is absent.

measurand – quantity intended to be measured (JCGM 200:2008)[9]; **NOTE 1:** The specification of a measurand requires knowledge of the kind of quantity, description of the state of the phenomenon, body, or substance carrying the quantity, including any relevant component, and the chemical entities involved (JCGM 200:2008)[9]; **NOTE 2:** The term "measurand" and its definition encompass all quantities, while the commonly used term "analyte" refers to a tangible entity subject to measurement. For example, "substance" concentration is a quantity that may be related to a particular analyte.

nonparametric (statistical procedure) – a "distribution-free" (see definition for **distribution-free,** above) statistical procedure is also called nonparametric because, unlike a parametric procedure, it does not assume a particular distribution.

parametric (statistical procedure) – one that involves an assumption as to the kind of distribution underlying the data and focuses on estimating a small number of characterizing quantities, called the parameters of the distribution; **NOTE 1:** For example, a normal (gaussian) distribution is specified by just

two parameters, that is, its mean and its standard deviation; **NOTE 2:** See definitions for **nonparametric** and **distribution-free,** above.

precision (measurement) – closeness of agreement between indications or measured quantity values obtained by replicate measurements on the same or similar objects under specified conditions (JCGM 200:2008)[9]; **NOTE 1:** Measurement precision is usually expressed numerically by measures of imprecision, such as standard deviation, variance, or coefficient of variation under the specified conditions of measurement (JCGM 200:2008)[9]; **NOTE 2:** The "specified conditions" can be, for example, repeatability conditions of measurement, intermediate precision conditions of measurement, or reproducibility conditions of measurement (see ISO 5725-3:1994)[10] (JCGM 200:2008)[9]; **NOTE 3:** Measurement precision is used to define measurement repeatability, intermediate measurement precision, and measurement reproducibility (JCGM 200:2008)[9]; **NOTE 4:** Sometimes "measurement precision" is erroneously used to mean measurement accuracy (JCGM 200:2008).[9]

prevalence – the probability of a particular clinical state in a specified population or subpopulation at a given point in time; **NOTE 1:** One can expect the prevalence to change, depending upon the population under study; **NOTE 2:** Prevalence is a frequency, not a rate.

receiver operating characteristic (ROC) curve – a graphical description of test performance representing the relationship between the true-positive fraction (sensitivity) and the false-positive fraction (1 − specificity); **NOTE:** Alternate terms are "ROC plot," "receiver operat*or* characteristic curve," "receiver opera*ting* characteristic plot," and "receiver operat*or* characteristic plot."

sensitivity (diagnostic) – the ability of a test to give a positive result for subjects who have the disease or condition for which they are being tested; **NOTE 1:** It is measured as the ratio of positive test results in those who have the condition to the total number who have the condition, and is often expressed as a percentage; **NOTE 2:** Formerly, the term "clinical sensitivity" was used in CLSI documents.

specificity (diagnostic) – the ability of a test to give a negative result for subjects who do not have the disease or condition for which they are being tested; **NOTE 1:** It is measured as the ratio of negative test results in those unaffected by the condition to the total number of condition-free subjects, and is often expressed as a percentage; **NOTE 2:** Formerly, the term "clinical specificity" was used in CLSI documents.

spectrum (of the condition) – various presentations of the condition of interest due to expected dissimilar manifestations with respect to various matrix characteristics; **NOTE 1:** In a testing situation, the condition of interest is usually defined to be binary, ie, it is either present or absent; **NOTE 2:** Notwithstanding this generalization, the condition of interest can often be expected to manifest itself differently with respect to various conditions (eg, stage of disease, severity of disease, genetic background, body composition, comorbidity, lifestyle, and demographics) that cannot be captured when dichotomizing its continuum.

spectrum bias – given a test and its intended application, the bias between estimated test performance and true test performance when the sample used for evaluating an assay does not properly represent the entire disease spectrum over the target (intended-use) population; **NOTE 1:** Spectrum bias is due to spectrum effect when one is not careful with the study design, ie, when the subject selection method departs substantially from true random sampling. When this occurs, biased estimates of sensitivity, specificity, receiver operating characteristic curves, and their summaries will result; **NOTE 2:** Spectrum bias can be summarized, in mathematical terms, as a problem created by improper sampling.

spectrum effect – effect that sampling different condition substrata in a population will have on test performance estimators.

true-negative fraction (TNF) – see **specificity.**

true-negative (TN) result – negative test result in a subject in whom the disease is absent.

true-positive fraction (TPF) – see **sensitivity.**

true-positive (TP) result – positive test result in a diseased or affected subject.

uncertainty (of measurement) – non-negative parameter characterizing the dispersion of the quantity values being attributed to a measurand, based on the information used (JCGM 200:2008)[9]; **NOTE:** Measurement uncertainty comprises, in general, many components. Some of these may be evaluated by Type A evaluation of measurement uncertainty from the statistical distribution of the quantity values from series of measurements and can be characterized by standard deviations (SDs). The other components, which may be evaluated by Type B evaluation of measurement uncertainty, can also be characterized by SDs, evaluated from probability density functions based on experience or other information (JCGM 200:2008).[9]

4.3 Abbreviations and Acronyms

AMI	acute myocardial infarction
AUC	area under the curve
CDA	cumulative distribution analysis
CEN	Comité Européen de Normalisation (European Committee for Standardization)
CI	confidence interval
CK-MB	creatine kinase MB fraction
FN	false negative
FNF	false-negative fraction
FP	false positive
FPF	false-positive fraction
ISO	International Organization for Standardization
LD	lactic dehydrogenase
LDL	low-density lipoprotein
LR	likelihood ratio
OxLDL	oxidized low-density lipoprotein
PoC	probability of concordance
RIA	radioimmunoassay
ROC	receiver operating characteristic
SD	standard deviation
SE	standard error
STARD	Standards for the Reporting of Diagnostic Accuracy Studies
TN	true negative
TNF	true-negative fraction
TP	true positive
TPF	true-positive fraction
VIM	*International Vocabulary of Metrology – Basic and General Concepts and Associated Terms*

5 Designing the Basic Evaluation Study

The sequence of events in the ROC curve analysis is explained in Figure 1.

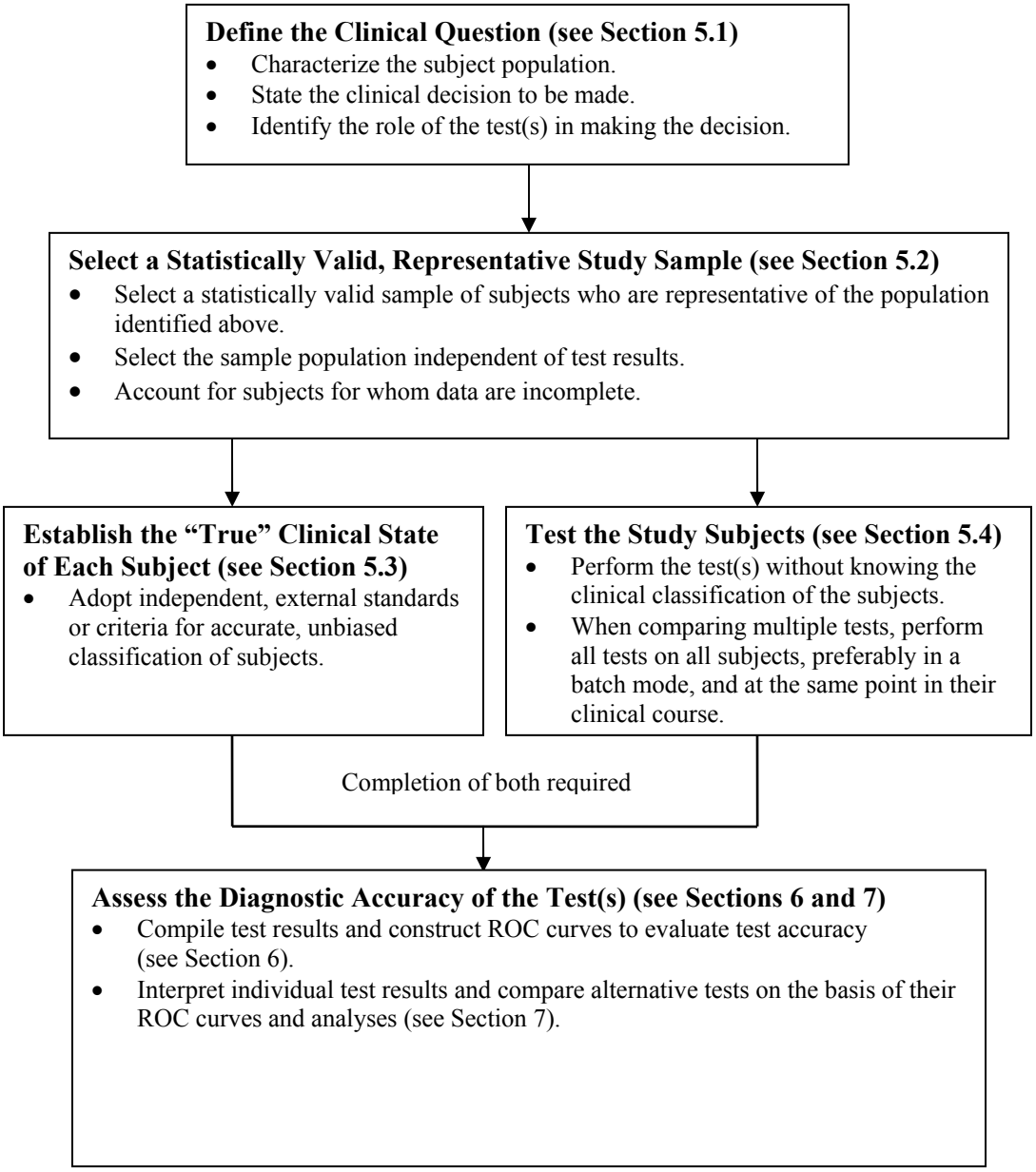


Figure 1. Flow Chart for the Evaluation Procedure

NOTE: Guidelines for the presentation of diagnostic test evaluations have been given by the international STARD (Standards for the Reporting of Diagnostic Accuracy Studies) group.[11,a]

[a] The initial version of STARD was published simultaneously in major journals in 2003. The current version is available at www.stard-statement.org. Their recommendations have been incorporated in the present Section 5 without further reference to STARD.

5.1 Define the Clinical Question

The goal of any laboratory test is to provide valuable information for patient care. There is always a relevant clinical question that must be defined because it establishes the particular patient-care issue to be addressed by the evaluation. For example, can troponin concentrations be used to discriminate between acute myocardial infarction (AMI) and other causes of chest pain in subjects who present to an emergency department with a history suggestive of AMI? Or, which, among several tests, is the best to use for discriminating between those subjects with breast cancer who will respond to a particular chemotherapy and those who will not?

Usually, the clinical question or goal involves a population of apparently similar subjects (grouped together on the basis of information available before the test under evaluation is done) that should be subdivided into relevant management subgroups. The results of the test should indicate to which management subgroup individual subjects belong. For example, a radioimmunoassay (RIA) for serum angiotensin-converting enzyme activity might be expected to answer the following question: "Among subjects with hypercalcemia, which ones have sarcoidosis?" The apparently similar subjects share the characteristic of hypercalcemia. The test helps to divide them into subgroups: those with sarcoidosis and those with some other cause of hypercalcemia (such as malignancy or hyperparathyroidism), each of which would receive different management.

In all cases, the target population must be well defined, including the nature, duration, and magnitude of the qualifying conditions. For example, this might include a serum calcium concentration greater than "X" on two occasions at least one week apart, as well as age range, sex, and other findings (eg, chest X-ray) that are required for including and excluding subjects from the population. By requiring a rigorous definition for the target population of an ROC study, one also defines the clinical question, and, in turn, the future population to which the diagnostic accuracy results of the test in question will apply.

5.2 Select a Statistically Valid, Representative Study Sample

The process of clearly defining the clinical question serves to identify the population relevant to the test evaluation. From this target population, a sample of subjects is chosen for the study. These subjects should be selected to represent the larger population of clinical interest about which conclusions are to be drawn.

When the accuracy of a test as a screening tool is being assessed, then the study sample should be representative of the population to be screened. Consider, for example, fecal occult blood testing for colon cancer. If the goal is to evaluate the accuracy of the test in discovering colon cancer in middle-aged subjects with no specific signs or symptoms suggestive of the disease, then the sample studied should be taken entirely from such a population. Studying a group of cancer-free, unaffected young volunteers or a group already known to have carcinoma of the colon is not appropriate.

The same principles apply when a test is not being used for screening, but for differentiating between disease states in symptomatic subjects. If a test is to be used to identify acute pancreatitis in subjects with a history and presentation indicating the possibility of pancreatitis, the sample should comprise such persons. Because the test is not intended to distinguish between unaffected volunteers and subjects with well-defined pancreatitis, a study sample containing unaffected subjects is not appropriate. Conclusions based on such a sample would not serve the purpose of the study.

5.2.1 Selection Bias

Selection bias occurs when subjects do not properly represent the relevant target population. To avoid selection biases that could compromise the study's validity or relevance to the question being posed, choose only subjects who fit within well-defined inclusion/exclusion criteria. Make certain that these

criteria define the clinical question and thus the relevant target population. Using only subjects with well-established or clinically apparent disease, for example, can exclude a large subgroup of subjects who may be more difficult to diagnose, especially those with occult or early disease. Likewise, using young, unaffected volunteers can be inappropriate to the presumptive application of the test. Insistence on creating a consecutive series from a well-defined stream of (candidate) cases, with documentation of reasons for exclusion ("reject log"), is mandatory in clinical "field trials" because it is the only effective safeguard against subjectivity and arbitrariness in the data collection (see Section 5.2.3); analogous rules apply in other phases of test development.

Measures of accuracy are influenced by the spectrum effect of varied medical conditions in the target population and, therefore, in the sample. Diagnostic tests must be evaluated in a clinically relevant population. However, test performance often varies across medical conditions. Failure to recognize and address such heterogeneity in a population will lead to estimates of test performance that are not generalizable to the relevant clinical question. This spectrum effect can be addressed by taking all subjects in order from a population that is typical of the populations in which this test will be used or by stratifying the results into identifiable medical conditions, a strategy that may prove difficult in some cases because of sample size considerations.[12] The importance of the proper spectrum of subjects is discussed in detail in the literature.[13-17]

5.2.2 Data Collected or Changed Retrospectively

Do not allow the test results or the testing procedures to affect the selection of subjects. Excluding subjects with unexpected, equivocal, or discordant results is likely to make the test appear more useful than it is. A retrospective study with only subjects who had their test results reported excludes subjects who could not be successfully tested for various reasons, again possibly distorting the performance of the test.

5.2.3 Selection Before Testing

Choosing subjects before testing acts as a precaution against the biases introduced when the test results directly or indirectly influence the selection of subjects. To avoid any biases, include in the test all subjects who meet the definition of the target (intended-use) population until a predetermined number of subjects is obtained. Once chosen, subjects should not be dropped from the study. If some subjects do not complete the study (because of technical errors, analytical interferences, death, or lack of follow-up), they should be accounted for in the final report (or they should be tabulated along with the other data and the ensuing consequences should be discussed). The same applies to indeterminate test results, unless "indeterminate" can be treated as a test result in its own right. Further, blinding is required for the personnel responsible for determination of clinical status from test results to ensure objectivity.

5.2.4 Sample Size

When determining the ROC curve, the diagnostic accuracy of detecting affected subjects at different decision levels is independent of the diagnostic accuracy of detecting unaffected subjects at the same decision levels. The uncertainty of this estimation, and of the ROC curve, decreases with increasing sample sizes. To minimize the uncertainty of the estimate of diagnostic accuracy for both affected and unaffected subjects, it is often desirable to have approximately equal numbers of subjects who are truly affected and truly unaffected. For sample size considerations related to sensitivity and specificity, see Section 7.1.3. For sample size considerations related to area under the curve (AUC), see Section 7.2.4.

5.2.5 Consult a Statistician

When the conditions of the study are complex, consult a statistician.

5.3 Establish the "True" Clinical State of Each Subject

An objective assessment of diagnostic accuracy requires comparing the results provided by the test with some independent, external definition of truth. The clinical question, defined above, establishes the categories of "truth" (states of health) that are relevant to the evaluation. Criteria or standards are applied to place individual persons in their respective categories of truth. The criteria or standards may include biopsy data, surgical or autopsy findings, imaging data, long-term clinical observation, or a more definitive laboratory test. The criteria or standards are adequate for practical purposes if they are substantially reliable and established independently of the diagnostic system (test) undergoing evaluation.[17]

Because a subject's condition may change over time, either spontaneously or in response to treatment, the diagnostic truth should be established simultaneously with the testing. Prognostic truths should be established according to a defined protocol that avoids censoring and other biases.

5.3.1 Validity of Evaluation

When evaluating the diagnostic accuracy of a test, the validity of the evaluation is limited by the accuracy with which the subjects are classified. A perfect test can appear to perform poorly simply because the "truth" was not established accurately for each patient and, therefore, the test results disagree with the apparent "true" diagnosis. On the other hand, when test results do agree with an inaccurate classification, the test will appear to perform *better* than it actually does. It is important, then, to attempt to classify individual persons as correctly as possible, as well as to consider the possible biases in the results caused by the classification scheme. The closer the classifications are to the truth, the less distortion there will be in the apparent performance of any test being evaluated.

5.3.2 True Clinical Subgroup

Routine clinical diagnoses are likely to be inadequate for evaluation studies. Determining a patient's true clinical subgroup can require procedures such as biopsy, surgical exploration, autopsy examination, angiography, or long-term follow-up of response to therapy and clinical outcome. Although such procedures can add to the financial cost of the evaluation, a less expensive, routine clinical evaluation can prove quite costly in the long term if its erroneous conclusions lead to improper test use or improper patient management.

5.3.3 Approaches to Classification

In many clinical situations, obtaining an independent, accurate classification of the patient's true clinical condition is difficult. Several approaches have been developed to deal with the difficulties in identifying true states of health. One approach is to define the diagnostic problem (diagnostic classification or category) in terms of measurable clinical outcomes.[18] A second approach is to employ a consensus, majority rule, or expert review to arrive at a less error-prone identification process.[19] For an in-depth examination of the topic of misclassification, see Fleiss,[20] Bross,[21] or Goldberg.[22] Misclassification affects the determination of diagnostic accuracy of a diagnostic test whether the context is classification (diagnosis in the narrow sense), eventual outcome (prognosis), or reaction to treatment (responsiveness/response potential).

5.3.4 Independent Classification

To avoid bias in evaluating the diagnostic accuracy of a test, the true clinical state should also be determined independent of the test(s) under investigation or used for comparison. Of course, the new test should not be included in the criteria used to classify the subjects; nor should a closely related test be included in these criteria. For example, if an RIA for creatine kinase MB fraction (CK-MB) is being

evaluated for the diagnosis of AMI, neither CK-MB by electrophoresis or by immunoinhibition should be included in the "gold standard" workup for classifying the study subjects. Furthermore, if the performance of the CK-MB assay is to be compared directly to the performance of the lactic dehydrogenase (LD) Type 1/Type 2 isoenzyme ratio, then LD isoenzyme results should also not be included in the diagnostic criteria because the apparent performance will be biased in favor of any test that is part of the gold standard.

5.3.5 Masked Evaluation

To ensure that the classification is not influenced by the result of the test under evaluation, it should be performed masked (or blinded), that is, without knowing the results of the test. Furthermore, the criteria for classifying each patient into a management subgroup should be as objective as possible. When the classification rests on subjective evaluation of clinical or morphological patterns, such as radionuclide scans or bone marrow smears, the decision for each patient should reflect the votes cast by experts who each interpret the material masked, and independent of the others, supplemented by a process that reconciles the differences and thus achieves a consensus decision.

5.4 Test the Study Subjects

5.4.1 Conduct a Masked Study

The person performing the test under evaluation should do so masked, that is, without knowing the clinical status of the subject. Ideally, the testing should be completed before the clinical question is answered. Knowing the answer to the clinical question can introduce bias.

5.4.2 Identical Specimens

When comparing two or more tests, it is preferable that the subjects be identical for all tests.[b] Failure to use the identical subjects for evaluating each test may result in misleading conclusions because biases can affect the selection of subjects. Thus, apparent differences in test performance can simply reflect differences in the composition of the groups tested. If some subjects have more advanced and, presumably, more easily detectable disease and are tested by only some of the tests, those tests could appear to have better sensitivity than the others. Conversely, inclusion of subjects with minimal disease, which might be harder to detect, would tend to diminish the apparent sensitivity of tests performed on these subjects, as compared with tests not done on these subjects. Performing all tests on all subjects at the same point in the course of each subject's illness ensures that differences in sensitivity and specificity are not simply due to differences between the patient materials or to unnoticed differences in the application of diagnostic criteria.

5.4.3 Testing Environment

Assaying all samples in one batch is suggested, when possible, to minimize intermediate imprecision. However, attention should be given to maintaining measurand stability through proper storage conditions.

As part of defining the clinical question to be answered, it is mandatory to *define the testing environment.* This includes issues such as who obtains the specimen and where and when the specimen is obtained (eg, "in the emergency department immediately on arrival"), preparation of the patient, storage of specimens, and technical monitoring of laboratory processes.

One should not perform repeat testing if the measurand is unstable.

[b] That is, for the same sample matrix, except with different types of specimens, at least obtained at the same time.

6 Construction of a Receiver Operating Characteristic Curve

6.1 Assess the Diagnostic Accuracy of the Test

The diagnostic accuracy of a test is assessed by examining its ability to correctly classify individual persons into two subgroups, eg, a subgroup of persons affected by some disease or condition (and therefore needing treatment) and a second subgroup of unaffected persons. The condition being assessed has only binary states: presence or absence. If there is no overlap in test results from these two subgroups, then the test can identify all persons correctly and discriminate between the two subgroups perfectly. However, if there is some overlap in the test results for the two subgroups, the ability of the test to discriminate is not perfect. In either case, it is desirable to have a way to represent and measure this ability to discriminate (ie, diagnostic accuracy).

6.1.1 Diagnostic (Clinical) Sensitivity and Specificity

The probability that a test will be positive or identify the presence of a target condition in a diseased or affected group is its diagnostic sensitivity. The probability that a test will be negative or identify the absence of a target condition in a nondiseased or unaffected group is its diagnostic specificity.

Diagnostic sensitivity (true-positive fraction [TPF]) is defined as follows:

$$\frac{Number\ of\ true\text{-}positive\ results}{Number\ of\ true\text{-}positive\ results + Number\ of\ false\text{-}negative\ results}, \quad (1)$$

or $TP/(TP+FN)$. This is the fraction of persons who are truly affected by the disease or condition whose test results are positive.

Diagnostic specificity (true-negative fraction [TNF]) is defined as follows:

$$\frac{Number\ of\ true\text{-}negative\ results}{Number\ of\ true\text{-}negative\ \ results + Number\ of\ false\text{-}positive\ results}, \quad (2)$$

or $TN/(TN+FP)$. This is the fraction of persons who are truly unaffected by a disease or condition whose test results are negative.

6.1.2 Receiver Operating Characteristic Curves

The choice of decision level implies a tradeoff between sensitivity and specificity. The range of tradeoffs between sensitivity and specificity is conveniently represented by the ROC curve.[23] ROC methodology was developed in the context of electronic signal detection and issues surrounding the behavior and use of radar receivers in the middle of the twentieth century.[17] The first known use of this analysis occurred in medicine when an ROC-type curve was used in the 1950s to characterize the ability of an automated Pap smear analyzer to discriminate between smears with and without malignant cells.[24]

The ROC curve graphically displays the entire range of a test's performance for a particular sample group of affected and unaffected subjects. It is, then, a "test performance curve," representing the fundamental diagnostic accuracy of the test by plotting all the sensitivity (1 − specificity) (or TPF-FPF) pairs resulting from repeatedly varying the decision threshold over the entire measuring interval of results observed. On the y-axis, sensitivity, or TPF, is plotted. On the x-axis, false-positive fraction (FPF) (or 1 − specificity) is plotted.

6.1.2.1 Generating the Receiver Operating Characteristic Curve

Example 1. Suppose an investigator would like to construct an ROC curve on a new assay (Assay X). For eight subjects, the investigator collects both assay results and a determination of true clinical status determined by an independent clinical reference standard. These are given in Table 1.

Table 1. Determinations of Assay X for Eight Subjects

Patient ID	Assay X Concentration	Clinical Status (Target Condition "Present" or "Absent")
1	1.6 ng/mL	Absent
2	2.1 ng/mL	Absent
3	6.4 ng/mL	Present
4	7.0 ng/mL	Absent
5	9.5 ng/mL	Present
6	15.1 ng/mL	Present
7	15.1 ng/mL	Absent
8	24.8 ng/mL	Present

Abbreviation: ID, identification.

There are seven distinct values for eight subjects. The investigator should count the number of subjects who would fall into the four categories (true positive [TP], true negative [TN], FP, and false negative [FN]) depending on where one imagines the cutoff for diagnosis to be placed. With seven distinct values, there are six intervals to examine in addition to the two outer intervals (below 1.6 and above 24.8 ng/mL in Table 1). Once the frequencies are known for those four categories, the sensitivity and specificity (or 1 − specificity) can be obtained. This expansion of the dataset is illustrated in Table 2.

Table 2. Computation of Clinical Performance Measures for Assay X

Assay X (cutoff concentration, ng/mL)	TP	TN	FP	FN	Sensitivity	Specificity	1 − Specificity
Cutoff < 1.6	4	0	4	0	100%	0%	100%
1.6–2.1	4	1	3	0	100%	25%	75%
2.1–6.4	4	2	2	0	100%	50%	50%
6.4–7.0	3	2	2	1	75%	50%	50%
7.0–9.5	3	3	1	1	75%*	75%*	25%*
9.5–15.1	2	3	1	2	50%	75%	25%
15.1–24.8	1	4	0	3	25%	100%	0%
Cutoff > 24.8	0	4	0	4	0%	100%	0%

* See text for explanation.

The first row records the fact that, if a concentration < 1.6 ng/mL were chosen as the cutoff, then all eight subjects would be considered "positive," so the estimated sensitivity would be 100% and specificity would be 0%. For each of the subsequent rows in Table 2, subjects are defined as "positive" if they have assay concentrations above a cutoff that one imagines to be chosen between the concentration of the present row and that of the next. For instance, the *-marked percentages refer to any cutoff > 7.0 but < 9.5 ng/mL. If one of these values itself is chosen as the cutoff, it is for the user to decide to which of the neighboring intervals it belongs. See Section 6.1.2.2.

This leaves the investigator with eight pairs of numbers (sensitivity and [1 − specificity]) that can be plotted. Once these points are plotted, the empirical ROC plot is generated by connecting the points. Various options exist regarding how the plot can be drawn (eg, smooth, piecewise linear, step function), according to personal preferences and preferred small sample properties. For the present illustration,

however, the data points will be connected in a piecewise linear manner (ie, the eight observations are taken completely literally). This allows the ROC graph to be drawn (see Figure 2), but a shortcut exists. First, sort the subjects in order of their laboratory test results from low on the left to high on the right, regardless of their categorization (see Figure 3). The simple sketch in Figure 3 then holds all the information needed for drawing the ROC, and the user may go directly from the raw data to a sketch of this kind, bypassing the detailed tabulation of Table 2.

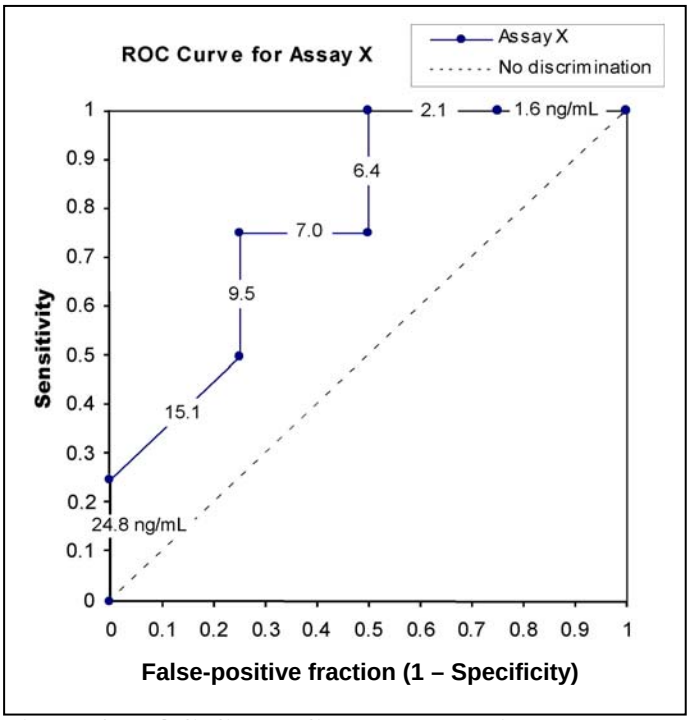


Figure 2. ROC Curve Constructed Using the Dataset in Table 1 (with underlying measurement marked on each step)

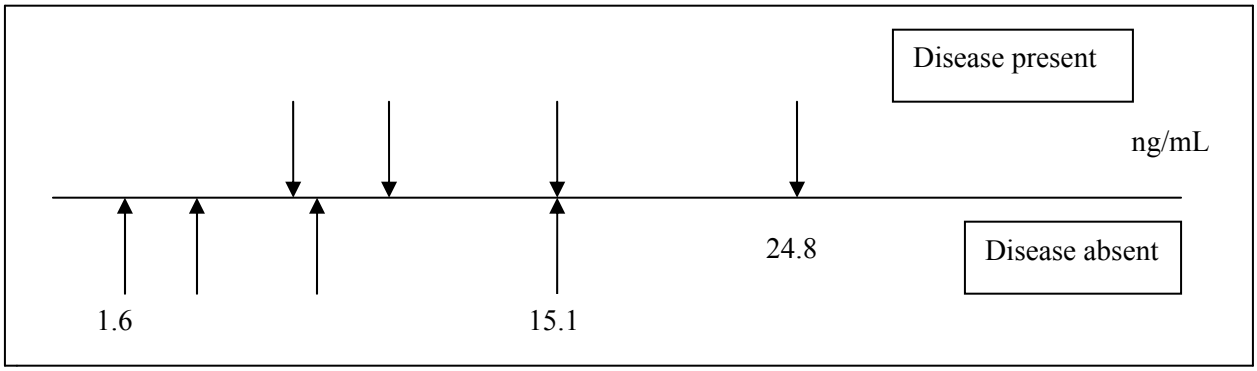


Figure 3. Auxiliary Sketch for Quick Construction of the ROC by Hand

To use the data in Figure 3 in this way, move a pointer (cutoff) from left to right starting to the left of all the data points. At each space between subjects, determine the proportion of affected subjects that are to the right of the cutoff (TP) and the proportion of unaffected that are to the right of the cutoff (FP). Plot these points on the graph with TPF on the y-axis and FPF on the x-axis.[c] The first ROC point is by definition (1.0, 1.0) in the upper right-hand corner of the graph, because all affected and all unaffected subjects are to the right of the cutoff. If the laboratory test is a good indicator of disease, the first set of

[c] The document development committee has chosen the most common way to plot an ROC curve. Some investigators plot sensitivity vs specificity, or FPF vs FNF, to improve clarity. The decision regarding how to draw the plot depends upon the type of data and personal preferences.

subjects selected by the cutoff will be only unaffected subjects, which means the TPF will remain 1.0 while the FPF will begin to decrease. This will appear on the ROC curve as a straight line moving horizontally leftward from the (1.0, 1.0) point. Eventually, some affected subjects will appear to the left of the cutoff and the ROC curve will start to bend downward to the left. Again, if the laboratory test is a good indicator of disease, at some point only unaffected subjects will remain to the right of the pointer. This is where the ROC curve will hit the y-axis, because FP has reached 0.0 and cannot go lower. If this exercise were to continue, the ROC trajectory would move vertically down the y-axis until, by definition, it reaches the point (0.0, 0.0) when all subjects (both affected and unaffected) are to the left of the cutoff.

In the example, the size of each horizontal step is $1/4 = 0.25$ because there are four unaffected individuals. The vertical step size also happens to be 1/4 because there are also four affected individuals. At 15.1 ng/mL (see Tables 1 and 2 and Figure 3), there are "tied" measurements, and a horizontal step coinciding with a vertical step forms a slanting line segment.

The resulting ROC curve is given in Figure 2. The reader should note the following: (1) moving through the points from *left to right* in Figure 3 corresponds to moving from *top to bottom* in Table 2 and from *upper right to lower left* in the ROC diagram in Figure 2; (2) each line segment of the ROC trajectory represents a particular observed concentration value; and (3) their junction points represent intervals on the ng/mL scale.

Note, in Figure 2, that if Assay X were uninformative, it would provide no discrimination between those with and without the disease (ie, it would be no better than chance). The AUC can be used to describe how informative a test is. This technique is covered in Section 7.2, and the AUC results from Example 1 are provided in Section 7.2.1. The area under the dotted diagonal is 0.5, so, to be useful, a test must produce an AUC substantially > 0.5.

6.1.2.2 Decision Thresholds

One can use ROC curves to select the appropriate medical decision level, depending upon the medical situation and the clinical setting. In the ROC curve, the various combinations of sensitivity and specificity possible for the test in a given setting are readily apparent. Also apparent, then, are the tradeoffs inherent in varying the decision threshold for that test. As the decision level changes, sensitivity improves at the expense of specificity, or vice versa. This can be observed directly from the plot. Note that the decision thresholds, though known, are not part of the plot. However, selected decision thresholds can be displayed at the point on the plot where the corresponding sensitivity and specificity appear.

When a decision threshold is chosen, the user must decide the neighboring interval in which it should reside, ie, the cut is to be made "just to the left of" or "just to the right of" the chosen value.

A number of mathematical strategies exist for deciding on a cutoff level to use in making decisions. One strategy attempts to minimize the distance (in a suitable sense) to the upper left corner of the graph (the point [0.0, 1.0] in the plot), representing a perfect test (see Figure 6). Another maximizes the sum of sensitivity and specificity. For Example 1, this would provide two cutoff levels. These criteria, while objectively providing cutoff options, do not take into account the clinical utility of the test or the costs associated with a decision.

Such costs, however, are a major concern when the clinician is faced with treatment decisions. The difficulty in quantifying the risks and costs of incorrect diagnoses often precludes the development of a cost function that can be used to derive an "optimum" cutoff. In this case, the responsibility for deciding the relevant tradeoff in clinical performance rests squarely with the clinician, after careful consideration of the benefits or detriments associated with positive or negative diagnoses. For example, suppose a subject presents with signs and symptoms of a myocardial infarction. A false-positive result would result in the administration of antithrombolytic drugs, risking hemorrhagic stroke. On the other hand, an FN

result would allow the damage to the heart to progress, causing more severe heart damage and possibly death. The tradeoff to be made here is clearly outside the responsibility and expertise of the statistician or the test vendor and must be made by the clinician.

One can minimize FN results by keeping the sensitivity at 100%. Usually, one needs to confirm the positive results because one has sacrificed specificity for sensitivity. D-dimer testing for venous thromboembolism is a case in point. Negative subjects are sent home, whereas positive subjects are sent to radiology for more accurate testing. The physician wants high probability that a negative test result can be trusted in order to comfortably send the patient home. An FN test could result in a patient's dying at home. An FP result will be correctly identified by the radiological examination, so the only cost is the extra use of hospital resources before discharging the patient home.

In contrast, consider human immunodeficiency virus testing. An FP test could result in antiviral therapy with medications that could have adverse effects, including toxicity to the kidneys, liver, and pancreas, as well as lipid profile changes that could result in cardiovascular disease. An FN result would likely be picked up later as the subject's T-cell count would decrease before the disease progressed further.

Similar considerations apply to clinical situations with three or more management options; hence, two or more cutoff points to be chosen on the ROC curve. Think of a decision problem involving not only a "positive" range ("treat as affected") and a "negative" range ("treat as unaffected"), but also a middle range in which the patient is routed to further, possibly invasive, testing. Here, overall diagnostic performance is determined not only by the ROC configuration, but obviously also by the properties of the second-stage diagnostic tools; details of this scenario are beyond the scope of the present text. (Note that this three-option situation is different from a three-category clinical problem involving subjects belonging to three diagnostic categories; the latter would call for a *three-dimensional* ROC diagram because the ordinary ROC diagram would be of little help.) Situations calling for more than one cutoff point[25] will not be discussed further.

Because TPF and FPF are calculated entirely separately, using the test results from two different subgroups of persons (affected and unaffected, respectively), the ROC curve is independent of the prevalence in the sample of the disease or condition of interest. However, as mentioned above, the TPFs and FPFs, and thus the ROC curve, are still influenced by the type of subjects (spectrum effect) included in the sample.

Although ROC curves are useful in determining the threshold for the test and to understand what happens when this threshold is changed, comparing AUCs to compare tests may ignore clinical consequences. AUC is a measure of discrimination and not necessarily a measure of diagnostic accuracy of the test. Two tests with the same AUC may not have the same clinical consequences even though all points on the curve are considered equivalent; however, they are not necessarily "clinically" equivalent.

6.1.3 Sample Selection Considerations in Establishing Decision Levels

In the example in Figure 4, discrimination was almost perfect, with nearly no overlap between the two samples of measurements (and the sample sizes were large enough to indicate that this was not just a lucky coincidence). When this happens, it is typically because there is a wide open concentration interval with nearly no observations between the unaffected and the affected populations. Provided that the data reflect the composition of a realistic target population, it means that sensitivity and specificity will be close to 100%, no matter how the cutoff is chosen within that open concentration interval.

Note that, other than the near-perfect discrimination, the shape of the ROC is difficult to discern in Figure 4. It is difficult to visually compare two well-performing tests. In this case, it may be useful to plot the false-negative fraction (FNF) (1 − sensitivity) on the vertical axis, and use a log-log scale. Furthermore, some tests involve the choice of parameters, eg, an algorithm that combines several test results to make a

clinical prediction. In this case, the parameters may be optimized based on the AUC. Optimization involves observing small changes in the target quantity. This is computationally more stable when the quantity is small, as with the plot in Figure 2, than when near unity, as with the usual AUC.

In an additional ROC study, one could sample the population of interest intentionally including nonextreme cases as well as extreme cases, as long as they can be properly diagnosed by the reference test (gold standard) in order to further refine where the cutoff point should be set within the wide open concentration interval.

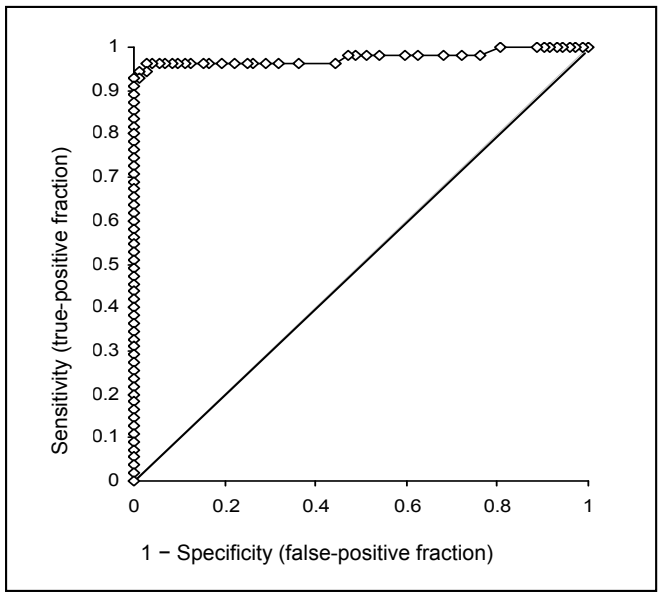


Figure 4. Antibody Test in a Target Population (n = 127, AUC = 0.977)

6.2 Generating the Receiver Operating Characteristic Curve: Ties

Usually, clinical data occur in one of two forms: discrete or continuous. Most clinical laboratory data are continuous, being generated from a measuring device with sufficient resolution to provide observations on a continuum. Measurements of electrolyte, therapeutic drug, hormone, enzyme, and tumor marker concentrations are essentially continuous. As in Example 1 above, when there are ties in continuous data at an Assay X concentration of 15.1 ng/mL, both the TPF and the FPF change simultaneously, resulting in a point that is displaced both horizontally and vertically from the last point. Connecting such adjacent points produces diagonal (nonhorizontal and nonvertical) lines on the plot, as seen in Figure 2 in Section 6.1.2.1. Slanting segments in the ROC curve, then, indicate ties.

Urinalysis dipstick results, on the other hand, are discrete data, as are rapid pregnancy testing devices, which give positive/negative results. Scales in diagnostic imaging also generally provide discrete data with rating categories such as “definitely affected,” “presumed affected,” “equivocal,” “presumed unaffected,” and “definitely unaffected.” In this case, the ROC will consist of five line segments, all of which will typically be slanting because even the bins labeled “definitely affected” or “unaffected” will end up containing a few misdiagnosed cases. A binary test (target condition present or absent) provides just two line segments connecting (0, 0) and (1, 1) to the test’s (sensitivity, 1 − specificity) point. Further discussion of discrete test results[26] is beyond the scope of this guideline.

6.3 Construction of the Receiver Operating Characteristic Curve When the Quantification Range Is Restricted

ROC curves extend "corner to corner," that is, from (0, 0), or the corner of the square plotting region representing 0% sensitivity and 100% specificity, to the diagonally opposite corner, (1,1), which represents the other extreme, namely 100% sensitivity and 0% specificity. Frequently, however, there is a measuring interval at the lower end of the scale below which it is considered inadvisable to trust the number produced (see CLSI document EP17[27]). The higher end of the test interval is usually not an issue because the sample being tested can often be diluted to be within the assay interval. Because no such adjustment is possible for the "low" interval, all persons in this low interval must be treated as a group. For example, in the case of an interval "below the limit of quantitation" (see Figure 5), which is also the affected end of the scale, suppose 24% of affected and also 3% of unaffected persons have a value in that low interval. In this case, the first part of the ROC curve is a line segment from (0, 0) to (0.03, 0.24); for explanation, it could simply be labeled "low." The line is fairly steep in this example (slope = 8), reflecting the fact that a "low" result is fairly good evidence of disease. In fact, whenever this line segment has a slope > 45 degrees, it implies that having a measurand level below quantitation is not an uninformative test result, but speaks in favor of the target disease. "High" results may have to be treated in an analogous manner.

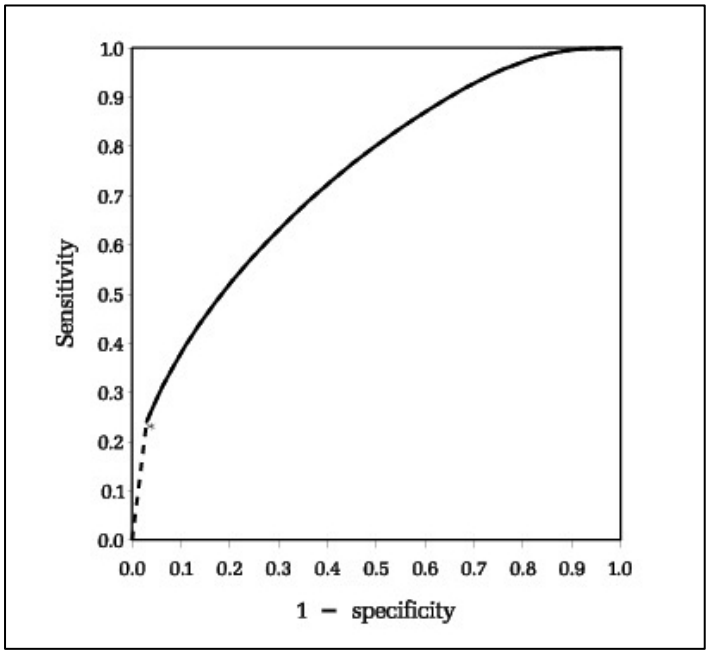


Figure 5. ROC for a Moderately Discriminative Quantitative Test. A fair proportion of the diseased (24%) have values below (accurate) determination. Just 3% of the nondiseased also have values below (accurate) determination. The graph serves to illustrate that values below determination must be treated on a par with any other reportable laboratory result. The point (0.03, 0.24) corresponds to the lower limit of determination. The dashed line below the asterisk in Figure 5 represents the entire range below determination (limit of quantitation) ie, it represents the (useful) information the clinician obtains when the laboratory replies "below determination" (and it reveals how common this reply will be).

7 Interpretation

Two different frameworks can be used to interpret ROC curves. First, the plot itself can be used to find the sensitivity-specificity pair that will best meet the needs of the clinical problem being addressed. Second, an overall measure of diagnostic accuracy (ie, AUC) can be assessed for any ROC curve. Within

both of these frameworks, a single plot can be interpreted by itself. However, each framework also lends itself to comparing curves and the tests that underlie them.

7.1 Relating the Receiver Operating Characteristic Curve to Sensitivity and Specificity

7.1.1 Using a Receiver Operating Characteristic Curve to Determine a Decision Level

When a sensitivity-specificity pair is determined from an ROC curve, the underlying data table used to generate the plot will also specify the decision level that generated that pair. A decision level should be chosen based on the intended use of the test and/or with respect to the type of device. Some devices may require unique approaches. One common way to select such a sensitivity-specificity pair is to find the point on the ROC curve that is, in a suitable sense, closest to the upper left corner. This optimizes sensitivity and specificity. Often this can be achieved by drawing a line from the lower right to the upper left and finding the point of intersection with the ROC plot, as in Figure 6. By referring to Table 2 in Section 6.1.2.1, one sees that this point corresponds to a decision level being chosen anywhere from 7.0 to just below 9.5 ng/mL, which results in sensitivity and specificity that are both 0.75. Note that this method does not account for the relative costs of FNs and FPs. However, further discussion of this topic is beyond the scope of EP24.

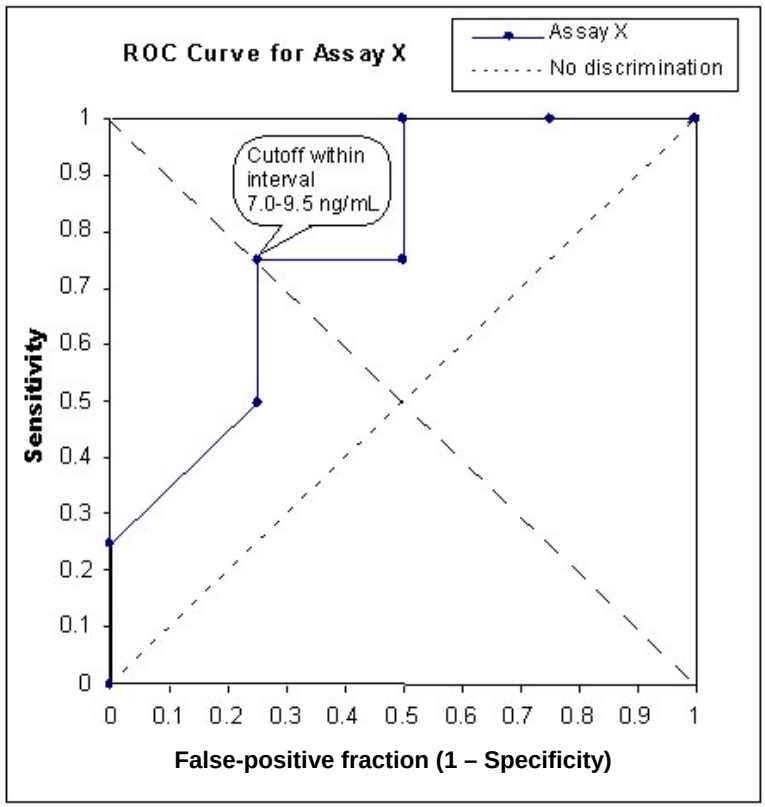


Figure 6. Example of a Constructed ROC Plot

7.1.2 Using Sensitivity-Specificity Pairs to Compare Receiver Operating Characteristic Curves

Tests can be compared to one another at a single, observed or theoretical, sensitivity or specificity.[28-31] The closeness of the ROC curve to the upper left corner is commonly used to determine how discriminating the test is as a diagnostic test and is often used (as in Figure 6) to compare two diagnostic tools. Using this criterion, Figure 7 shows that Test A is more discriminating than Test B because its curve lies above the curve of Test B across the graph's domain. Figure 8, on the other hand, shows two diagnostic tests that appear similar in discriminating properties but differ in their sensitivities and

specificities at different decision levels, except at the point where the two curves cross. Test A shows greater sensitivity than Test B at high specificity, but Test B shows greater sensitivity than Test A at lower specificity.

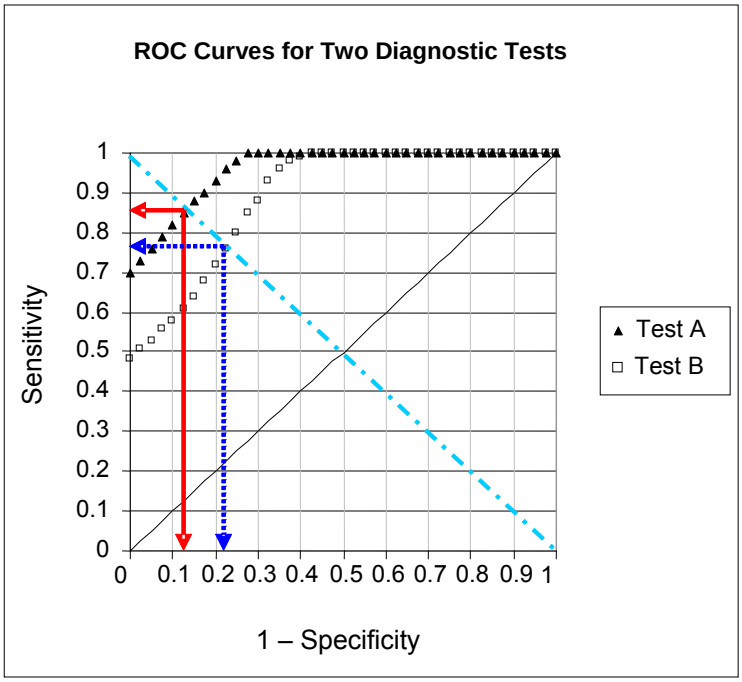


Figure 7. Test A Is Superior to Test B by Any Criterion (see Figures 8 and 9)

Note that the ROC curves in Figures 7, 9, and 10 do not show the curve continuing down to the (0, 0) point. Even though this line segment is not shown, such a continuation is assumed, by definition.

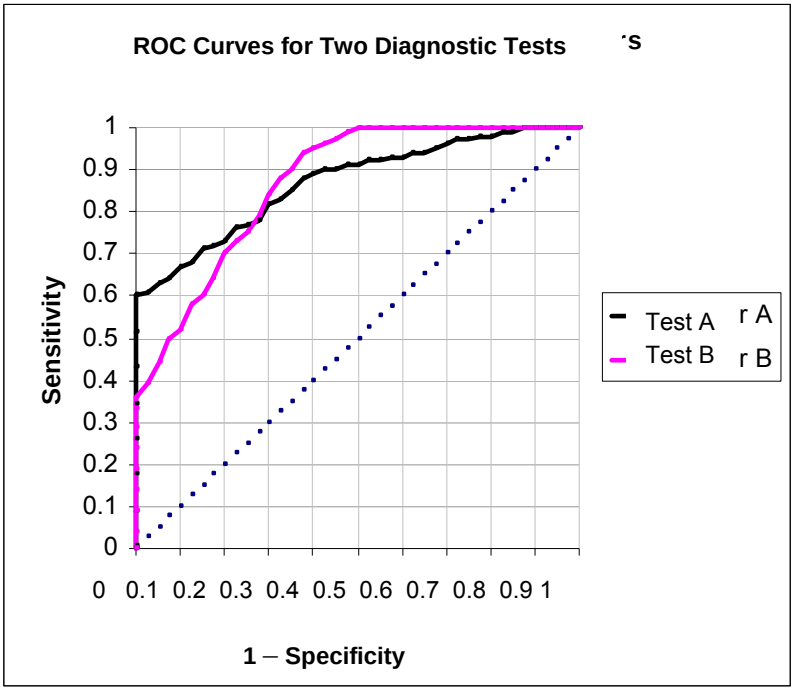


Figure 8. Test A Is Superior to Test B Only When a High Specificity Is Required

One may not always want to compare tests based on closeness to the upper left corner. An alternative approach would be to fix the sensitivity or specificity at a predetermined level. For the purpose of ruling out serious pathology in a patient with unexplained symptoms, one may ensure a low incidence of FNs by specifying a sensitivity value and examining the corresponding specificity. In Figure 9, at a predetermined sensitivity of 0.80, Test A has a much higher specificity than Test B. Conversely, one may ensure a low incidence of FPs by preselecting a specificity and then examining the corresponding sensitivity. In Figure 10, at a preselected specificity of 0.90, Test A has a much higher sensitivity than Test B.

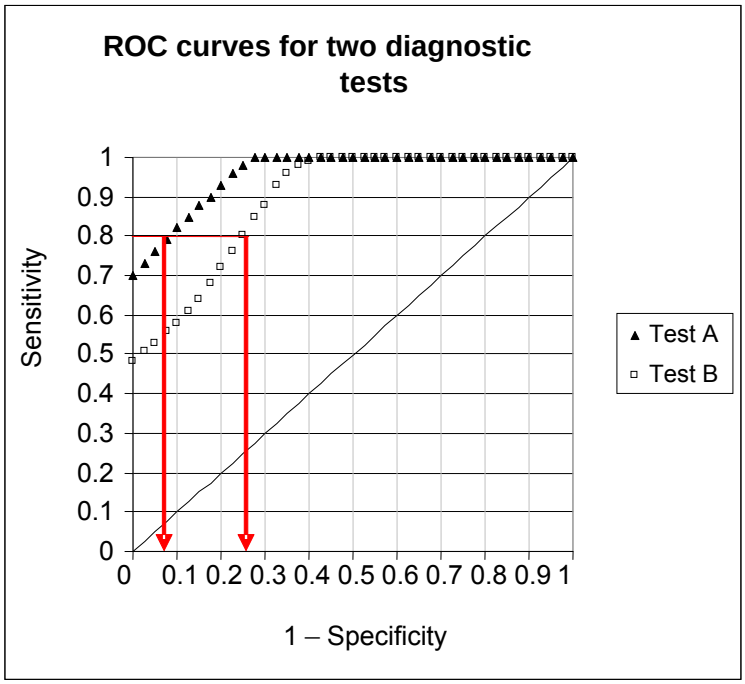


Figure 9. Given the Sensitivity of 0.80, Test A Has Higher Specificity

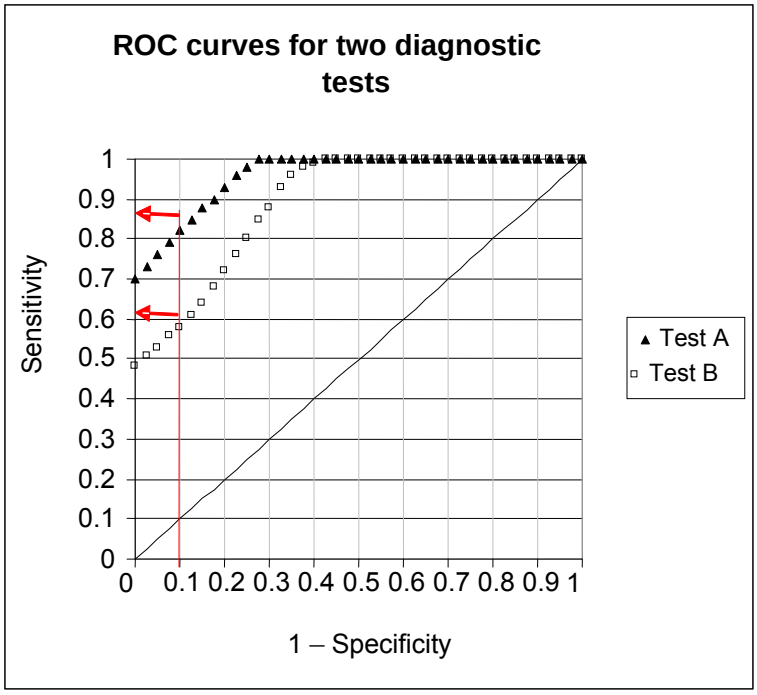


Figure 10. Given the Specificity of 0.90, Test A Has Higher Sensitivity

One needs to pay particular attention to the curves when comparing two or more tests in situations in which the curves cross two or more times. Depending on the purpose of the test, one may choose the test that maximizes either sensitivity or specificity.

7.1.3 Sample Size for Sensitivity and Specificity

To determine the sample size to estimate a single test's sensitivity or specificity, the following formulas (3) and (4) are generally used.[32] Let n_D represent the number of required subjects with disease and $n_{\bar{D}}$ represent the number of required subjects without disease.

$$n_D = \frac{\left(G(1-\alpha/2)\sqrt{TPF(1-TPF)}\right)^2}{L^2} \quad (3)$$

and

$$n_{\bar{D}} = \frac{\left(G(1-\alpha/2)\sqrt{FPF(1-FPF)}\right)^2}{L^2} \quad (4)$$

where L is the desired width of one half of the confidence interval (CI) for either sensitivity or specificity, $G(1-\alpha/2)$ is the $1-\alpha/2$ percentile of the standard normal distribution and α is the desired confidence level of the estimate. These equations can be used when the decision threshold is prespecified.

Table 3 shows some examples of sample size estimates using a 95% CI ($\alpha = 0.05$) and other parameter values that might typically be used.

Table 3. Sample Sizes Required to Obtain Desired Precision

TPF* or *FPF	***L***	***n****
0.8	0.05	246
0.85	0.05	196
0.9	0.05	139
0.95	0.05	73
0.7	0.1	81
0.75	0.1	73
0.8	0.1	62
0.85	0.1	49

* Sample size calculations are rounded up to the next whole number.

The above equations are based on a normal approximation of the binomial distribution, an assumption that breaks down as TPF or FPF approaches 1.0. In addition, as the sample sizes fall below those shown above, the results of these equations often no longer match the results derived using an exact binomial calculation. An alternative approach to sample size calculation requires estimates of expected and minimally acceptable TPF or FPF.[33]

7.2 Area Under a Receiver Operating Characteristic Curve

One common measure to quantify the diagnostic accuracy of a laboratory test with a single number is the AUC. Values range from 1.0 (perfect separation of the test values of the two groups with no misclassification) down to zero (theoretically, at least; perfect separation but 100% misclassification). When there is no diagnostic information at all (ie, the test results for the two populations have identical distributions and the ROC curve runs along the diagonal), then the area is 0.5. All tests of practical value have areas well above this. The main attraction of the area calculation is that it does not focus on a

particular portion of the curve, such as the region closest to the upper left corner or the sensitivity at some chosen specificity, but reflects the entire curve.

7.2.1 Measuring the Area Under the Curve

The statistician readily recognizes the ROC area as the Mann-Whitney version of the nonparametric two-sample statistic[34,35] introduced by the chemist Frank Wilcoxon. An area of 0.8, for example, means that a randomly selected person from the affected group has a laboratory test value higher (when the affected group tends to have higher test values than the unaffected group) or lower (when the affected group tends to have lower test values than the unaffected group) than that for a randomly chosen person from the unaffected group 80% of the time. The relation between AUC for an ROC curve and rank-sum statistics is discussed in Appendix C.

When there are no ties between the affected and unaffected groups, this area is easily computed from the curve as the sum of the rectangular areas under this graph. Analytical formulas to calculate the area appear in reports by Bamber[34] and Hanley and McNeil.[35] Alternatively, the area can be obtained from the Wilcoxon rank-sum statistic.[36]

Parametric approaches to calculating area (ie, those employing some model for fitting a curve) have also been described. Both parametric and nonparametric methods are discussed and compared in published reviews.[2,37]

When using summary indices such as AUC, sensitivity, or specificity, there is a loss of information. Therefore, one should always visually examine the ROC curve itself, as well. Example 1 has been used to generate an ROC curve (see Figure 2) and to determine cutoff levels (see Figure 6). The AUC measured from Example 1 is shown in Table 4.

Table 4. Area Assessment in Example 1

Assay X	AUC	95% CI	SE	z	*p*	Patient Diagnosis = Present
Concentration (ng/mL)	0.78	0.42 to 1.00	0.182	1.54	0.0614	Have higher values
H_0: AUC ≤ 0.5. H_1: AUC > 0.5.						

In Table 4, not only is the AUC listed (0.78), but also an approximate 95% CI (0.42 to 1.00) is given. This range is quite wide because of the small sample size and encompasses 0.5. Therefore, the possibility that Assay X is no better than chance (0.5) cannot be excluded (one-sided approximate $p = 0.06$).

The AUC measurement can also be seen as the average sensitivity over all specificities (the range of specificities is 1.0) or the average specificity over all sensitivities (the range of sensitivities is 1.0). The average sensitivity can also be determined over a defined interval of specificities by computing the partial AUC and dividing by the width of this interval of specificities, as shown in Figure 11. In a similar fashion, the average specificity can be determined over a defined range of sensitivities, as shown in Figure 12.

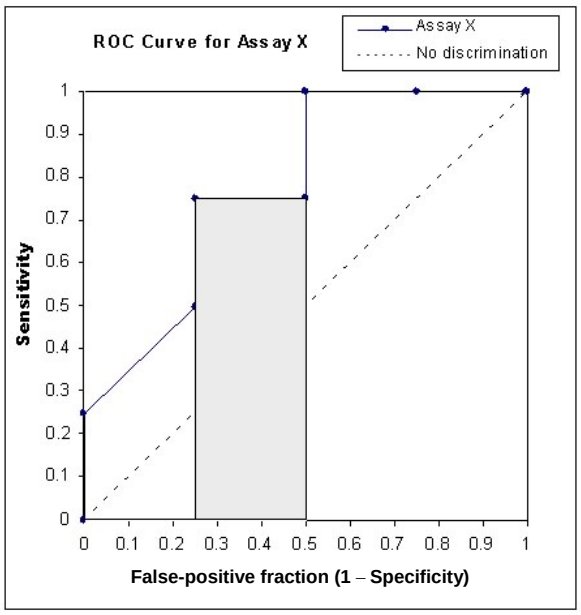


Figure 11. Average Sensitivity Over Specificity Interval From 0.5 to 0.75

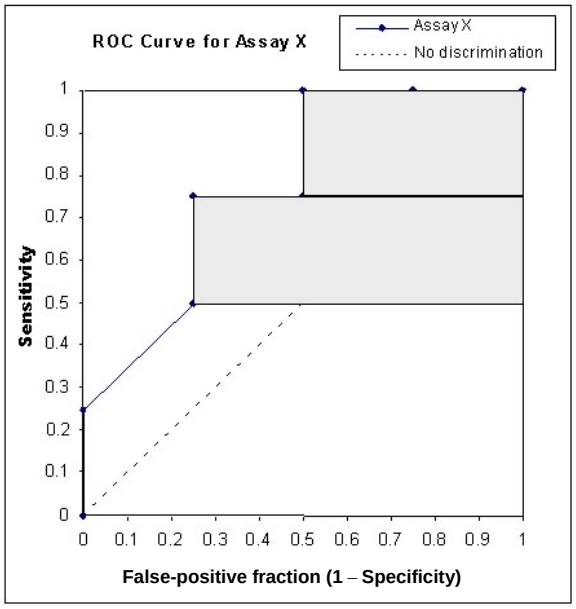


Figure 12. Average Specificity Over Sensitivity Interval From 0.5 to 1.0

Such average measurements would be appropriate in cases in which only a defined range of accuracy (sensitivity or specificity) is clinically acceptable.

7.2.2 Comparing the Area Under the Curve of Two Tests

Direct statistical comparison of multiple diagnostic tests is common in clinical laboratories. Usually, two (or more) tests are performed on the same subjects (or specimens), as in a split-specimen comparison. This is often called "paired design."

A global approach is to compare entire ROC curves by using an overall measure, such as AUC. This can be performed either nonparametrically or parametrically, ie, with or without a model that supplies, and constrains, the shape of distribution(s) of the measurand. This is especially attractive to laboratories because the comparison does not rely on the selection of a particular decision threshold. However, the user should always inspect the ROC plot visually when comparing tests, rather than rely *only* on summary measures that condense all the information into a single number. A good example is presented in Figure 8 in Section 7.1.2. In this example, two tests have similar AUCs, but one is skewed to the right while the other is skewed to the left, which makes Test A more sensitive than Test B at high degrees of specificity.

A seemingly natural choice of statistics for comparing two AUCs is the difference in the AUCs divided by the standard error (SE) of the difference. The null hypothesis, H_0: $AUC_1 = AUC_2$, is tested by comparing the value of z (see equation 5, below) with a standard normal distribution[33,38] because the z statistic has approximately a standard normal distribution. If $|z|$ is > 1.96, then the two AUCs are significantly different at a significance level of $\alpha = 0.05$.

$$z = \frac{AUC_1 - AUC_2}{\sqrt{Var(AUC_1 - AUC_2)}} = \frac{AUC_1 - AUC_2}{SE(AUC_1 - AUC_2)} \quad (5)$$

7.2.2.1 Calculating Test Statistics

The AUC as well as its SE can be calculated using either parametric or nonparametric methods. One parametric method is based on the binormal assumption,[d,39] and the nonparametric method is based on Mann-Whitney U statistics.[35]

If the samples used to generate the two ROC curves are independent, then the denominator can be obtained by taking the square root of the sum of the two variances. However, if the samples are not considered independent, such as in a paired design, the SE of the differences should include an additional term for correlation because the two AUCs will be correlated.

7.2.2.2 Comparing Correlated Areas Under the Curve

A method for comparing two AUCs (ie, testing for equality) in a paired design is discussed by Hanley and McNeil.[38] This method uses the Dorfman and Alf[34] approach to calculate the AUC as well as its SE. There is also a nonparametric approach to compare two AUCs in a paired design (DeLong et al.).[40]

The test statistic for a paired design has an additional term that includes a correlation coefficient, *r*.

$$z = \frac{AUC_1 - AUC_2}{\sqrt{Var(AUC_1 - AUC_2)}} = \frac{AUC_1 - AUC_2}{\sqrt{Var(AUC_1) + Var(AUC_2) - 2rSE(AUC_1)SE(AUC_2)}} \quad (6)$$

Table 5 lists the correlation coefficients, *r,* for different values of average correlation and average area. The row value is obtained by taking an average of two correlations, $(r_N + r_A)/2$. Here, r_N is a correlation coefficient from unaffected subjects by the two different tests, and r_A is obtained from affected subjects. The column value is an average of the two AUCs, $(A_1 + A_2)/2$. The degree of correlation depends on the types of diagnostic tests. However, the correlation is likely to be positive because the specimens are collected from the same subjects. The larger the correlation between the two diagnostic tests, the more powerful (sensitive) the statistical comparison becomes and the more likely it is to declare the difference as statistically significant. Hanley and McNeil[38] have discussed this in terms of sample size and statistical power and suggested a table of reasonable *r* values (ie, Table 5). A calculated example involving real data is given in Appendix D.

[d] The binormal assumption states that the measurand, when expressed on a suitable scale (with transformation in some cases), has the following property: in each of the two populations (with disease, without disease), its distribution is normal (gaussian); both the means and the variances are allowed to differ.

Table 5. Correlation Coefficients Between Two ROC Areas.* From Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology.* 1983;148(3):839-843. Used with permission.

Average Correlation Between Ratings†	Average Area‡											
	0.700	**0.725**	**0.750**	**0.775**	**0.800**	**0.825**	**0.850**	**0.875**	**0.900**	**0.925**	**0.950**	**0.975**
0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.01	0.01	0.01	0.01
0.04	0.04	0.04	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.02	0.02	0.02
0.06	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.04	0.04	0.04	0.03	0.02
0.08	0.07	0.07	0.07	0.07	0.07	0.06	0.06	0.06	0.06	0.05	0.04	0.03
0.10	0.09	0.09	0.09	0.09	0.08	0.08	0.08	0.07	0.07	0.06	0.06	0.04
0.12	0.11	0.11	0.11	0.10	0.10	0.10	0.09	0.09	0.08	0.08	0.07	0.05
0.14	0.13	0.12	0.12	0.12	0.12	0.11	0.11	0.11	0.10	0.09	0.08	0.06
0.16	0.14	0.14	0.14	0.14	0.13	0.13	0.13	0.12	0.11	0.11	0.09	0.07
0.18	0.16	0.16	0.16	0.16	0.15	0.15	0.14	0.14	0.13	0.12	0.11	0.09
0.20	0.18	0.18	0.18	0.17	0.17	0.17	0.16	0.15	0.15	0.14	0.12	0.10
0.22	0.20	0.20	0.19	0.19	0.19	0.18	0.18	0.17	0.16	0.15	0.14	0.11
0.24	0.22	0.22	0.21	0.21	0.21	0.20	0.19	0.19	0.18	0.17	0.15	0.12
0.26	0.24	0.23	0.23	0.23	0.22	0.22	0.21	0.20	0.19	0.18	0.16	0.13
0.28	0.26	0.25	0.25	0.25	0.24	0.24	0.23	0.22	0.21	0.20	0.18	0.15
0.30	0.27	0.27	0.27	0.26	0.26	0.25	0.25	0.24	0.23	0.21	0.19	0.16
0.32	0.29	0.29	0.29	0.28	0.28	0.27	0.26	0.26	0.24	0.23	0.21	0.18
0.34	0.31	0.31	0.31	0.30	0.30	0.29	0.28	0.27	0.26	0.25	0.23	0.19
0.36	0.33	0.33	0.32	0.32	0.31	0.31	0.30	0.29	0.28	0.26	0.24	0.21
0.38	0.35	0.35	0.34	0.34	0.33	0.33	0.32	0.31	0.30	0.28	0.26	0.22
0.40	0.37	0.37	0.36	0.36	0.35	0.35	0.34	0.33	0.32	0.30	0.28	0.24
0.42	0.39	0.39	0.38	0.38	0.37	0.36	0.36	0.35	0.33	0.32	0.29	0.25
0.44	0.41	0.40	0.40	0.40	0.39	0.38	0.38	0.37	0.35	0.34	0.31	0.27
0.46	0.43	0.42	0.42	0.42	0.41	0.40	0.39	0.38	0.37	0.35	0.33	0.29
0.48	0.45	0.44	0.44	0.43	0.43	0.42	0.41	0.40	0.39	0.37	0.35	0.30
0.50	0.47	0.46	0.46	0.45	0.45	0.44	0.43	0.42	0.41	0.39	0.37	0.32
0.52	0.49	0.48	0.48	0.47	0.47	0.46	0.45	0.44	0.43	0.41	0.39	0.34
0.54	0.51	0.50	0.50	0.49	0.49	0.48	0.47	0.46	0.45	0.43	0.41	0.36
0.56	0.53	0.52	0.52	0.51	0.51	0.50	0.49	0.48	0.47	0.45	0.43	0.38
0.58	0.55	0.54	0.54	0.53	0.53	0.52	0.51	0.50	0.49	0.47	0.45	0.40
0.60	0.57	0.56	0.56	0.55	0.55	0.54	0.53	0.52	0.51	0.49	0.47	0.42
0.62	0.59	0.58	0.58	0.57	0.57	0.56	0.55	0.54	0.53	0.51	0.49	0.45
0.64	0.61	0.60	0.60	0.59	0.59	0.58	0.58	0.57	0.55	0.54	0.51	0.47
0.66	0.63	0.62	0.62	0.62	0.61	0.60	0.60	0.59	0.57	0.56	0.53	0.49
0.68	0.65	0.64	0.64	0.64	0.63	0.62	0.62	0.61	0.60	0.58	0.56	0.51
0.70	0.67	0.66	0.66	0.66	0.65	0.65	0.64	0.63	0.62	0.60	0.58	0.54
0.72	0.69	0.69	0.68	0.68	0.67	0.67	0.66	0.65	0.64	0.63	0.60	0.56
0.74	0.71	0.71	0.70	0.70	0.69	0.69	0.68	0.67	0.66	0.65	0.63	0.59
0.76	0.73	0.73	0.72	0.72	0.72	0.71	0.71	0.70	0.69	0.67	0.65	0.61
0.78	0.75	0.75	0.75	0.74	0.74	0.73	0.73	0.72	0.71	0.70	0.68	0.64
0.80	0.77	0.77	0.77	0.76	0.76	0.76	0.75	0.74	0.73	0.72	0.70	0.67
0.82	0.79	0.79	0.79	0.79	0.78	0.78	0.77	0.77	0.76	0.75	0.73	0.70
0.84	0.82	0.81	0.81	0.81	0.81	0.80	0.80	0.79	0.78	0.77	0.76	0.73
0.86	0.84	0.84	0.83	0.83	0.83	0.82	0.82	0.81	0.81	0.80	0.78	0.75
0.88	0.86	0.86	0.86	0.85	0.85	0.85	0.84	0.84	0.83	0.82	0.81	0.79
0.90	0.88	0.88	0.88	0.88	0.87	0.87	0.87	0.86	0.86	0.85	0.84	0.82

* Correlation coefficient r between two ROC areas A_1 and A_2 as a function of the average correlation between ratings (rows) and average area (columns).

† $(r_N + r_A)/2$, where r_N = correlation coefficient from unaffected subjects by the two tests; r_A = correlation coefficient from affected subjects by the two tests.

‡ $(A_1 + A_2)/2$.

7.2.3 Comparing Average Sensitivities or Specificities: Partial Area Under the Curve Approach

The AUC is not always a useful summary statistic when one wants to compare two curves. The most obvious example is when two ROC curves cross but have equal AUCs, as in Figure 8. The fact that the curves cross tells us that one test performs better than the other for certain clinical settings, and vice versa for other clinical settings. In cases such as these, clinical requirements should influence how the tests are compared. For example, if one must choose between two tests and their ROC curves cross but the AUCs are nearly equal, one might consider comparing test sensitivities at only high levels of specificity. One approach would be to decide what levels of specificity would be acceptable and then compare the tests' average sensitivities restricted to that specificity (or FP) interval.

Therefore, the average sensitivities for two tests can be compared, but only over identical FPF intervals, as seen in Figure 11 in Section 7.2.1. Conversely, two tests can be compared by using average specificity over a selected sensitivity interval, as seen in Figure 12 in Section 7.2.1. In either case (see the example in Figure 8 in Section 7.1.2), two tests could be ranked differently depending on which intervals served as the basis for their comparison.

7.2.4 Sample Size for Evaluating Area Under the Curve

Two classic references for determining sample size requirements for ROC analysis are the papers by Hanley and McNeil.[35,38] The techniques described in these papers cover three different cases: the one-sample case, the two-sample case with independent samples, and the two-sample case with both measurements performed on the same subjects (as in a paired design).

7.2.4.1 One-Sample Case

When the investigator is interested only in the diagnostic accuracy of a single device, two approaches can be taken to determine a sample size. One is to determine the sample size required for a specified width on the CI for the AUC. Alternatively, in a preliminary evaluation of the test, one may want to determine the sample size necessary to obtain statistical significance for testing that the AUC is > 0.50.

The case of a specified CI width will be addressed first. Assume that:

n_A = the number of affected subjects, and
n_N = the number of unaffected subjects.

Using an anticipated value for AUC, the quantities Q_1 and Q_2 are obtained, as follows:

$$Q_1 = AUC / (2 - AUC), \text{ and} \tag{7}$$

$$Q_2 = 2AUC^2 / (1 + AUC). \tag{8}$$

Now, the SE of the AUC is given by[36]:

$$SE(AUC) = \sqrt{\frac{AUC(1 - AUC) + (n_A - 1)(Q_1 - AUC^2) + (n_N - 1)(Q_2 - AUC^2)}{n_A \bullet n_N}}. \tag{9}$$

Using a realistic trial value for the two n's, one can vary the n's until obtaining an acceptable CI width. If one has data from a pilot study already analyzed, a simpler method is to realize that the SEs vary inversely with the square root of the sample size, so that if:

n_1 = the total sample size from the pilot study,
$SE(AUC_1)$ = the SE of the AUC from the pilot study,
n_2 = the total sample size for the proposed study, and
$SE(AUC_2)$ = the desired SE of the AUC for the proposed study,

then, under the condition that the prevalence remains the same from the pilot study population to the population in the proposed study, n_2 can be computed simply by solving the following:

$$\frac{\sqrt{n_1}}{\sqrt{n_2}} = \frac{SE(AUC_2)}{SE(AUC_1)}, \text{ or} \tag{10}$$

$$\sqrt{n_2} = \sqrt{n_1} \bullet \frac{SE(AUC_1)}{SE(AUC_2)}. \tag{11}$$

In the second approach to determining the sample size, the investigator's focus is on the hypothesis test that demonstrates the efficacy of the diagnostic assay. A noninformative test (ie, one that is not different from random choice) would have an AUC of 0.50. Thus, the hypothesis to be tested is AUC = 0.50, vs the one-sided alternative, AUC >0.50. The condition that must be satisfied to meet the above conditions is as follows:

$$n \geq \frac{\left(SE\{AUC\}\right)^2 \left(Z_\beta + Z_\alpha\right)^2}{\delta^2}, \text{ where} \tag{12}$$

n = total number of subjects,
δ = the difference stipulated (hoped for) between the AUC and 0.50,
$Z_\alpha = G(1-\alpha)$ = the critical value for the (one-sided) hypothesis test at significance level α, and
$Z_\beta = G(1-\beta)$ = the constant defined by the required power level: 0.84, 1.28, or 1.645 for 80%, 90%, or 95% power, respectively.[41] G here stands for the standard gaussian (normal) cumulative distribution function as defined in Section 7.1.3. See equations (3) and (4).

7.2.4.2 Two-Sample Case

The accuracy of two tests can be compared by detecting differences in two AUCs. In the example below, the comparison will determine whether either test has a larger AUC than the other and is therefore a two-sided comparison. A similar example using a one-sided comparison can be found in the paper by McNeil and Hanley.[29]

Let: AUC_1 = the area under the curve for Test 1, which is the predicate or standard against which Test 2 is to be compared;
AUC_2 = the area under the curve for Test 2,
$Z_\alpha = G(1-\alpha/2)$ = the critical value for the (two-sided) hypothesis test at significance level α, and
$Z_\beta = G(1-\beta)$ = the constant defined by the required power level: 0.84, 1.28, or 1.645 for 80%, 90%, or 95% power, respectively. G here stands for the standard gaussian (normal) cumulative distribution function as defined in Section 7.1.3. See equations (3) and (4).

To compute the required sample size, one first chooses realistic tentative values for the two AUCs that differ by an amount, δ, that one hopes to be able to document if it exists. One then computes the intermediate quantities V_1 and V_2, given below, namely:

$$V_1 = \frac{AUC_1}{2 - AUC_1} + \frac{2AUC_1^2}{1 + AUC_1} - 2AUC_1^2, \text{ and} \tag{13}$$

$$V_2 = \frac{AUC_2}{2 - AUC_2} + \frac{2AUC_2^2}{1 + AUC_2} - 2AUC_2^2. \tag{14}$$

Now the sample size required from each of the two tests and from each of the two states (unaffected, affected) for a test of significance is $n_1 = n_2 =$

$$n_{unpaired} = \left[\frac{Z_\alpha \sqrt{2V_1} + Z_\beta \sqrt{V_1 + V_2}}{AUC_2 - AUC_1} \right]^2 \tag{15}$$

Savings in sample size can be achieved if the same subjects can be tested with Test 1 and Test 2. The savings are related to the level of correlation between the two AUCs. More precisely,

$$n_{paired} = (1 - r) \cdot n_{unpaired}, \text{ where} \tag{16}$$

$n_{unpaired}$ = the total number of subjects required using two independent samples,
n_{paired} = the total number of subjects required using paired samples, and
r = the correlation of the two AUCs, which can be interpolated from Table 5.[38]

Although the studies used for initial estimates may not exactly match the ultimate protocol for the final study, they still provide approximate estimates for the problem at hand. The process of generating sample size requirements is always fraught with risk because one needs to provide parameter estimates that are unknown, and are the purpose of commissioning the study. Thus, one should be prepared to revise these estimates as data are collected that verify or contradict the assumed values used in the computations. For a nonparametric approach to variance estimation, applicable to sample size calculation, see DeLong et al.[40]

8 Application of Receiver Operating Characteristic Curves

The ROC curve is simple, graphical, and easily appreciated visually on a universal scale. It represents inherent diagnostic accuracy, which is the discriminating ability over the entire measuring interval of the test. It does not require selection of a particular decision threshold because the whole range of possible decision thresholds is included. Because the ROC curve is generated independent of prevalence, obtaining samples with representative prevalence is not important as long as cases are collected in a nonselective fashion. In fact, it is usually preferable to have approximately equal numbers of subjects with both conditions. In the resultant curve, both specificity and sensitivity and the tradeoff between them are readily accessible.

The properties of ROC curves and functions of ROC curves, such as AUC, permit a number of uses, including the following:

- Determining whether a test is better than chance
- Finding an optimal point on the curve for a clinical application
- Judging whether the test is better suited for proving or for disproving disease or target condition present (attains a high specificity for reasonable sensitivity or vice versa)
- Assessing diagnostic accuracy

- Comparing two tests regardless of the units they use

- Evaluating whether two tests for the same process have the same or dissimilar diagnostic potential

Reference interval studies, as detailed in CLSI document C28,[42] have been used to generate cutoff points. However, such studies, by their nature, include only one population, so quantification of the tradeoff between sensitivity and specificity is unavailable. Such a cutoff only determines a point where results from the reference population are unlikely to fall. Such cutoffs do not answer a clinical question with regard to a specific medical condition.

The ROC curve compiles much information in a simple construct. This is its strength and leads to the many uses detailed above. However, because of this simplification, there are many things ROC curves cannot provide. First, the decision threshold and the number of subjects with and without the clinical conditions at each point along the curve are lost in the graphical representation. The curve can be annotated at selected places if desired, but it is impractical to do so across the entire curve. It is common to provide a complete table of the results used to generate the curve. Another option is to use a cumulative distribution analysis (CDA) plot, which is a joint depiction of sensitivity and specificity as a function of cutoff values.[43] More information on CDA plots is provided in Appendix B.

Some common mistakes observed in studies using ROC analysis are failure to provide a CI for the AUC to evaluate its margin of error and the lack of a formal hypothesis test when comparing two curves. A common incorrect practice is to state that the CIs of the AUCs overlap; however, this is not equivalent to a hypothesis test. As with any summary statistic, the AUC must be considered as one part of a comprehensive analysis of a candidate diagnostic device or test. Both sensitivity and specificity (and hence AUC) can be influenced by case mix, disease severity, and concomitant risk factors for the clinical state under consideration. A single ROC and AUC calculation cannot take these influences into account.

As with any statistic, the AUC has its strengths and weaknesses, and it should be used carefully. For instance, the algebraic equivalence of the AUC statistic to the Mann-Whitney U and Wilcoxon rank-sum test makes it insensitive to the risk level of the subjects being studied. Inclusion of novel risk factors in well-functioning risk prediction equations rarely improves the risk stratification to a clinically appreciable extent. Even when regression statistics or significance tests suggest a large and indisputable “effect,” decision analysis most often shows that the clinical impact will be small. Although the AUC is not, strictly speaking, a measure of clinical benefit in the sense of decision analysis, the AUC behaves like one, in that it duly portrays the smallness of the impact. Unfortunately, when faced with a large regression effect and a minuscule AUC increase, investigators have been inclined to trust the former, leading to the frequent claim[44-47] that the AUC is unduly insensitive to added diagnostic or prognostic information. The claim is false; the AUC should be regarded as trustworthy (“pessimistic but not unduly so”), whereas regression output sometimes inspires false hope of clinical gain.

An analogous situation arises when a well-performing diagnostic test is made more informative (eg, by reduction of measurement error); even when the reduction is pronounced (and statistically significant), the AUC changes little. This fact is illustrated in Appendix A.

References

[1] Swets JA, Pickett RM. *Evaluation of Diagnostic Systems*. New York, NY: Academic Press Inc.; 1982:1-6.

[2] Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin Chem.* 1993;39(4):561-577.

[3] Kullback S. *Information Theory and Statistics.* Mineola, NY: Dover Publications; 1997.

[4] Linnet K, Brandt E. Assessing diagnostic tests once an optimal cutoff point has been selected. *Clin Chem*. 1986;32(7)1341-1346.

[5] Kondratovich M, Yousef W. Evaluation of accuracy and 'optimal' cutoff of diagnostic devices in the same study. Joint Statistical Meeting. ASA Section on Statistics in Epidemiology. 2005:2547-2551.

[6] Leeflang MM, Moons KG, Reitsma JB, Zwinderman AH. Bias in sensitivity and specificity caused by data-driven selection of optimal cutoff values: mechanisms, magnitude, and solutions. *Clin Chem*. 2008;54(4):729-737.

[7] 2007 Guideline for Isolation Precautions: Preventing Transmission of Infectious Agents in Healthcare Settings. http://www.cdc.gov/hicpac/pdf/isolation/isolation2007.pdf. Accessed November 16, 2011.

[8] CLSI. *Protection of Laboratory Workers From Occupationally Acquired Infections; Approved Guideline—Third Edition*. CLSI document M29-A3. Wayne, PA: Clinical and Laboratory Standards Institute; 2005.

[9] Bureau International des Poids et Mesures (BIPM). *International Vocabulary of Metrology – Basic and General Concepts and Associated Terms* (VIM, 3rd edition, JCGM 200:2008) and *Corrigendum* (May 2010). http://www.bipm.org/en/publications/guides/vim.html. Accessed November 16, 2011.

[10] ISO. *Accuracy (trueness and precision) of measurement methods and results – Part 1: Intermediate measures of the precision of a standard measurement method.* ISO 5725-3. Geneva, Switzerland: International Organization for Standardization; 1994.

[11] Bossuyt PM, Reitsma JB, Bruns DE, et al. Toward complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Acad Radiol.* 2003;10(6):664-669.

[12] Mulherin SA, Miller WC. Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. *Ann Intern Med*. 2002;137(7):598-602.

[13] Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med.* 1978;299(17):926-930.

[14] Robertson EA, Zweig MH, Van Steirteghem AC. Evaluating the clinical efficacy of laboratory tests. *Am J Clin Pathol.* 1983;79(1):78-86.

[15] Zweig MH, Robertson EA. Why we need better test evaluations. *Clin Chem.* 1982;28(6):1272-1276.

[16] Lachs MS, Nachamkin I, Edelstein PH, Goldman J, Feinstein AR, Schwartz JS. Spectrum bias in the evaluation of diagnostic tests: lessons from the rapid dipstick test for urinary tract infection. *Ann Intern Med.* 1992;117(2):135-140.

[17] Metz CE. ROC methodology in radiologic imaging. *Invest Radiol.* 1986;21(9):720-733.

[18] Valenstein PN. Evaluating diagnostic tests with imperfect standards. *Am J Clin Pathol.* 1990;93(2):252-258.

[19] Revesz G, Kundel HL, Bonitatibus M. The effect of verification on the assessment of imaging techniques. *Invest Radiol.* 1983;18(2):194-198.

[20] Fleiss JL. *Statistical Methods for Rates and Proportions.* New York, NY: Wiley; 1981:188-211.

[21] Bross I. Misclassification in 2x2 tables. *Biometrics.* 1954;10:478-486.

[22] Goldberg JD. The effects of misclassification on the bias in the difference between two proportions and the relative odds in the fourfold table. *J Am Stat Assoc*. 1975;70(351):561-567.

[23] Metz CE. Basic principles of ROC analysis. *Semin Nucl Med.* 1978;8(4):283-298.

[24] Lusted LB. ROC recollected [editorial]. *Med Decis Making.* 1984;4:131-135.

[25] Lu ZX, Walker KZ, O'Dea K, Sikaris KA, Shaw JE. A1C for screening and diagnosis of type 2 diabetes in routine clinical practice. *Diabetes Care.* 2010;33(4):817-819.

[26] Glasziou P, Hilden J. Threshold analysis of decision tables. *Med Decis Making.* 1986 Jul-Sep;6(3):161-168.

[27] CLSI/NCCLS. *Protocols for Determination of Limits of Detection and Limits of Quantitation; Approved Guideline.* CLSI/NCCLS document EP17-A. Wayne, PA: NCCLS; 2004.

[28] Beck JR, Shultz EK. The use of relative operating characteristic (ROC) curves in test performance evaluation. *Arch Pathol Lab Med.* 1986;110(1):13-20.

[29] McNeil BJ, Hanley JA. Statistical approaches to the analysis of receiver operating characteristic (ROC) curves. *Med Decis Making.* 1984;4(2):137-150.

[30] Greenhouse SW, Mantel N. The evaluation of diagnostic tests. *Biometrics.* 1950;6:399-412.

[31] Qin G, Hsu YS, Zhou XH. New confidence intervals for the difference between two sensitivities at a fixed level of specificity. *Stat Med.* 2006;25(20):3487-3502.

[32] Zhou XH, Obuchowski NA, McClish DK. *Statistical Methods in Diagnostic Medicine*. New York, NY: John Wiley & Sons; 2002:196-198.

[33] Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. New York, NY: Oxford University Press; 2003:218-220.

[34] Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *J Math Psychol.* 1975;12(4):387-415.

[35] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology.* 1982;143(1):29-36.

[36] Hollander M, Wolfe DA. *Nonparametric Statistical Methods*. New York, NY: John Wiley; 1973:67-78.

[37] Hanley JA. Receiver operating characteristic (ROC) methodology: the state of the art. *Crit Rev Diagn Imaging.* 1989;29(3):307-335.

[38] Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology.* 1983;148(3):839-843.

[39] Dorfman DD, Alf E. Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals—rating method data. *J Math Psychol.* 1969;6:487-496.

[40] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44(3):837-845.

[41] Kahn HA, Sempos CT. *Statistical Methods in Epidemiology.* New York, NY: Oxford University Press; 1989:34-35.

[42] CLSI. *Defining, Establishing, and Verifying Reference Intervals in the Clinical Laboratory; Approved Guideline—Third Edition.* CLSI document C28-A3c. Wayne, PA: Clinical and Laboratory Standards Institute; 2008.

[43] Krouwer JS. Cumulative distribution analysis graphs—an alternative to ROC curves. *Clin Chem.* 1987;33(12):2305-2306.

[44] Cook NR. Assessing the incremental role of novel and emerging risk factors. *Curr Cardiovasc Risk Rep*. 2010;4(2):112-119.

[45] Melander O, Newton-Cheh C, Almgren P, et al. Novel and conventional biomarkers for prediction of incident cardiovascular events in the community. *JAMA*. 2009;302(1):49-57.

[46] Moons KGM. Criteria for scientific evaluation of novel markers: a perspective. *Clin Chem.* 2010;56(4):537-541.

[47] Uno H, Tian L, Cai T, Kohane IS, Wei LJ. Comparing risk scoring systems beyond the ROC paradigm in survival analysis. *Harvard University Biostatistics Working Paper Series.* Working Paper 107, 2009.

Appendix A. Effect of Measurement Uncertainty on Receiver Operating Characteristic Curves

To properly define the total allowable measurement uncertainty for each measurand, consider its effect on the receiver operating characteristic (ROC) curve and on the diagnostic accuracy indices, including the area under the ROC curve (AUC).

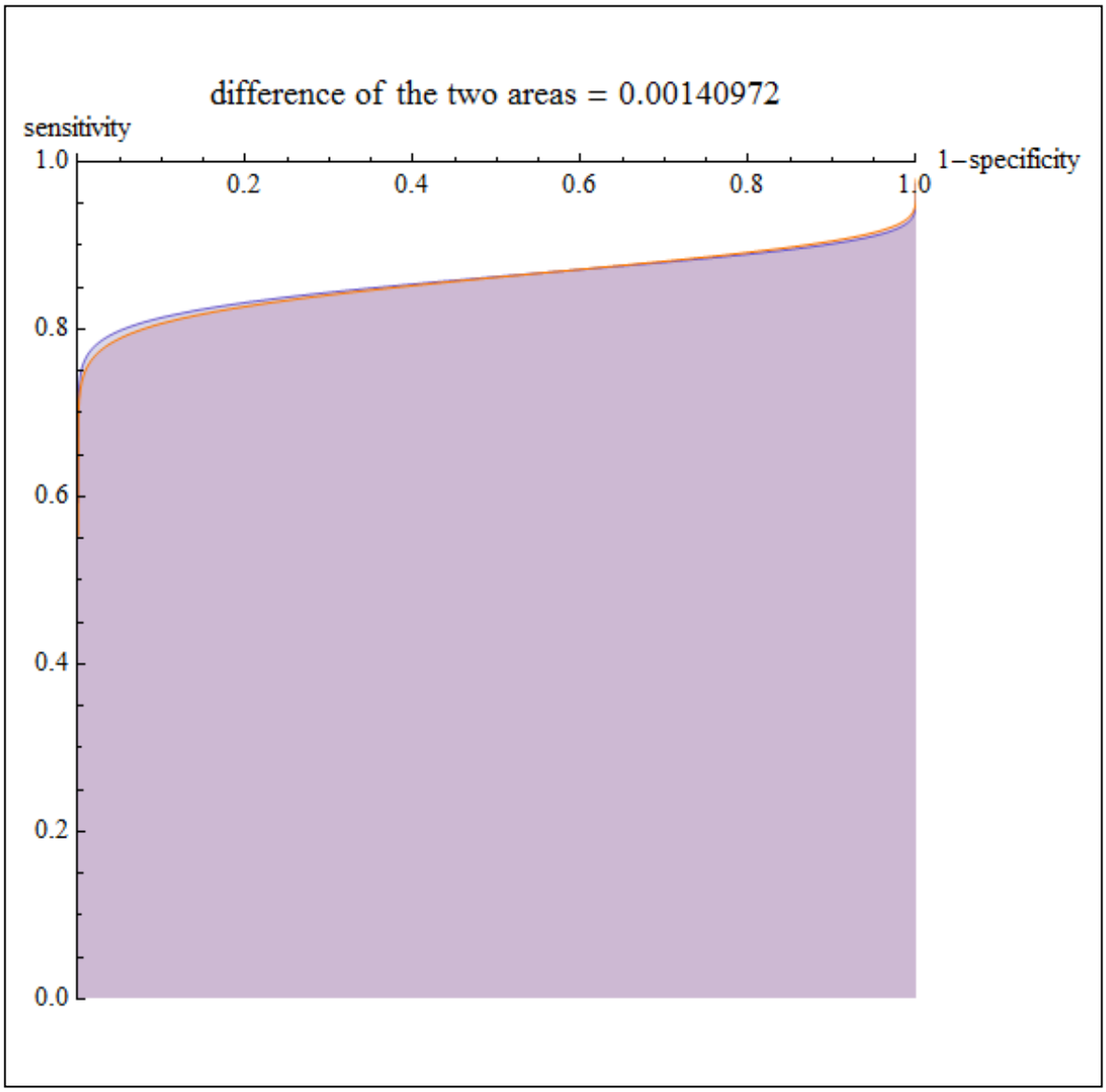


Figure A1. Comparison of the ROC Curves of Two Diagnostic Tests Measuring the Serum Glucose of Two Normally Distributed Populations. The uncertainty of the first test (blue curve) is 0.05 standard deviation (SD), while the uncertainty of the second test (orange curve) is 0.50 SD. The graph serves to show that, despite the large difference in measurement SD, the curves are virtually indistinguishable.

Appendix A. (Continued)

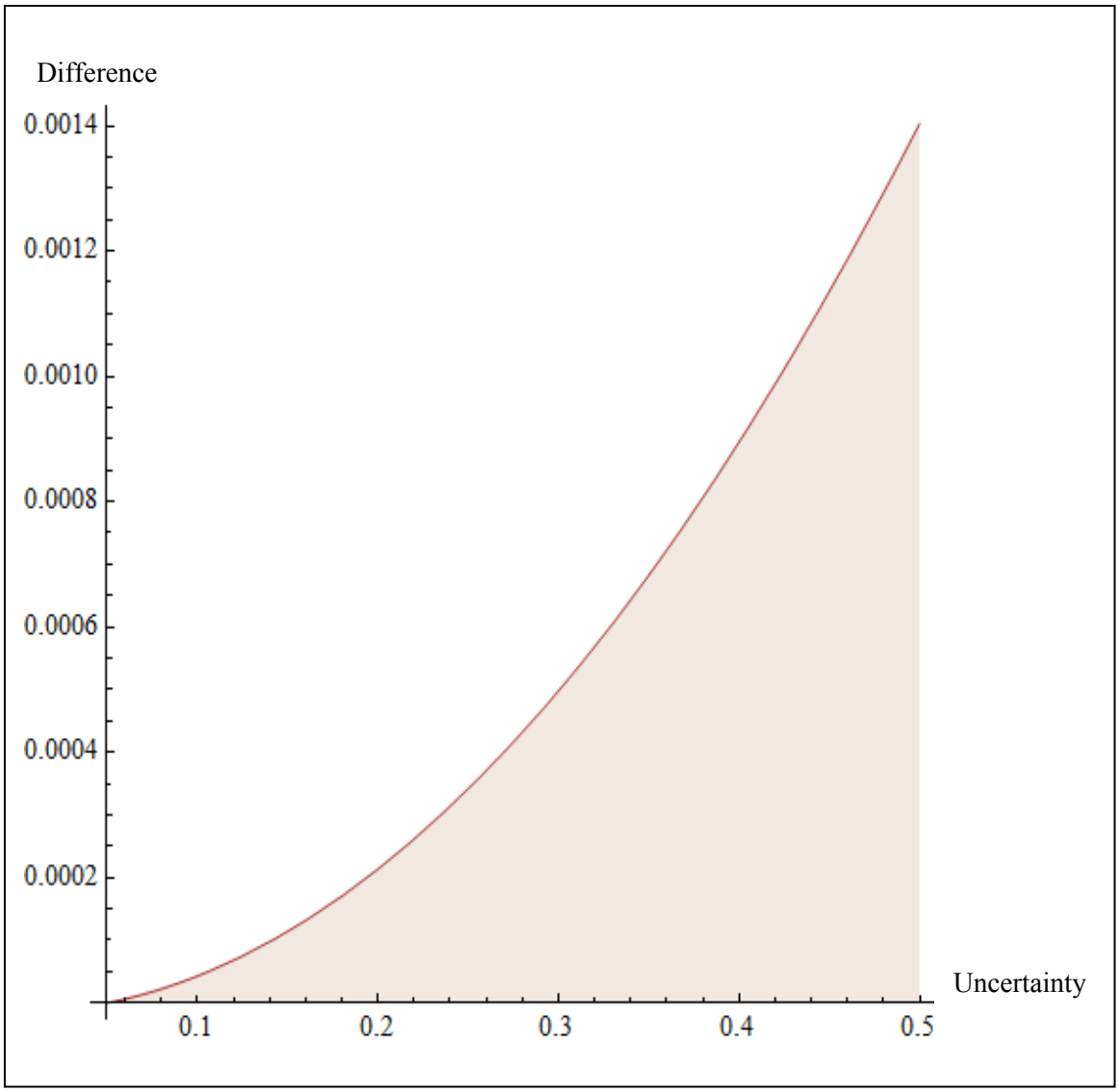


Figure A2. Difference in the Surfaces Under the ROC Curves of Two Diagnostic Tests Measuring the Serum Glucose of Two Normally Distributed Populations. The uncertainty of the first test is 0.15 SD, while the uncertainty of the second test varies from 0.05 SD to 0.50 SD.

Table A1. Explanation of Figures A1 and A2 of the ROC Curves

		Figure A1	Figure A2
Unaffected population	Distribution	Normal	Normal
	Mean	0.00	0.00
	SD	1.00	1.00
Affected population	Distribution	Normal	Normal
	Mean	7.00	7.00
	SD	6.40	6.40
Uncertainty of measurement	First test	0.05	0.05
	Second test	0.50	0.05–0.50
Difference in the two areas under the curves (AUC of the first test − AUC of the second test)		0.0014	0.0000–0.0014

Appendix A. (Continued)

The effect of the measurement error on the AUC has been studied since 1997.[1-4] It is possible to compare the ROC curves of two diagnostic tests measuring the same measurand[5] with different uncertainties of measurement.[6] For the comparison, the difference in the areas under the respective curves can be used, for either the specificity interval (0, 1) or any subinterval of interest.

Figures A1 and A2 illustrate two applications of this approach. Both figures are based on a German population including known diabetic subjects, with a bimodal distribution of fasting plasma glucose,[7] assuming binormality. In Figure A1, the ROC plot of a test with an uncertainty of 0.05 SD, which could be considered as an example of state-of-the-art performance, is compared with the ROC plot of an alternative test with uncertainty that is ten times greater. In Figure A2, the plot shows the difference between the areas under the ROC curve of a test with an uncertainty of 0.05 SD and a second test with uncertainty that varies from 0.05 SD to 0.50 SD. Table A1 explains the data for the two figures.

Both figures show that the uncertainty of measurement has relatively little effect on the areas under the ROC curves of the tests, in accordance with previous findings on the imprecision effects on ROC curves of cardiac markers.[8]

References for Appendix A

1 Coffin M, Sukhatme S. Receiver operating characteristic studies and measurement errors. *Biometrics*. 1997;53(3):823-837.

2 Faraggi D. The effect of random measurement error on receiver operating characteristic (ROC) curves. *Stat Med.* 2000;19(1):61-70.

3 Reiser B. Measuring the effectiveness of diagnostic markers in the presence of measurement error through the use of ROC curves. *Stat Med.* 2000;19(16):2115-2129.

4 Schisterman EF, Faraggi D, Reiser B, Trevisan M. Statistical inference for the area under the receiver operating characteristic curve in the presence of random measurement error. *Am J Epidemiol.* 2001;154(2):174-179.

5 Hatjimihail AT. Receiver Operating Characteristic Curves and Uncertainty of Measurement. Wolfram Demonstrations Project. Wolfram Research Institute; 2009. http://www.demonstrations.wolfram.com/ReceiverOperatingCharacteristicCurvesAndUncertaintyOfMeasure. Accessed November 16, 2011.

6 Bureau International des Poids et Mesures (BIPM). *Evaluation of measurement data – Guide to the expression of uncertainty in measurement* (JCGM 100:2008). http://www.bipm.org/en/publications/guides/gum.html. Accessed November 16, 2011.

7 Vistisen D, Colagiuri S, Borch-Johnsen K; DETECT-2 Collaboration. Bimodal distribution of glucose is not universally useful for diagnosing diabetes. *Diabetes Care.* 2009;32(3):397-403.

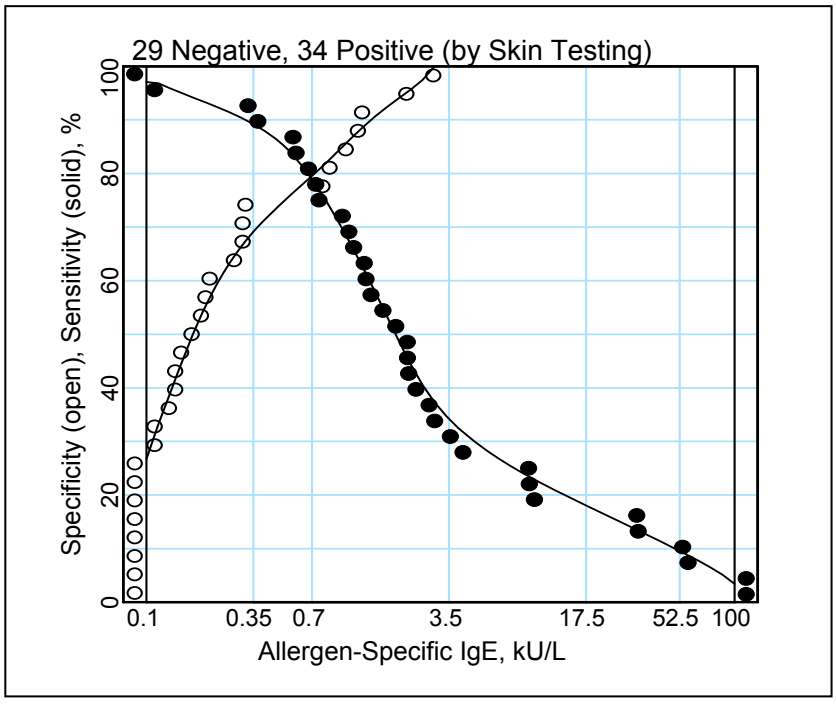
8 Kupchak P, Wu AH, Ghani F, Newby LK, Ohman EM, Christenson RH. Influence of imprecision on ROC curve analysis for cardiac markers. *Clin Chem.* 2006;52(4):752-753.

Appendix B. Cumulative Distribution Analysis Plots: Their Nature, Construction, and Practical Application

Some investigators find cumulative distribution analysis (CDA) plots helpful.

A CDA plot is a joint depiction of sensitivity and specificity as a function of cutoff values.[1] It consists of two plots in a coordinate space based on values ranging from 0 to 1 (fractiles) or from 0 to 100 (centiles) on the vertical axis and, on the horizontal axis, measurand concentrations, or whatever units are appropriate to the continuous diagnostic indicator in question.

For comparing two assay methods, the classic receiver operating characteristic (ROC) representation has the advantage of requiring just two curves, not the four that would be needed using CDA plots. However, the ROC curve achieves this economy of expression by sacrificing information, namely, the concentrations associated with its data points (sensitivity and specificity pairs). This loss can be remedied by annotating all or some of the nodes in the empirical ROC representation with the cutoff values, but this is an awkward solution that causes visual clutter and yields a visually distorted sense of scale for the cutoff values. Accordingly, when it is important not to lose sight of analytical characteristics—primarily the effective limits of quantitation, but also variance functions—CDA plots can serve as a valuable complement to ROC curves in studies of diagnostic accuracy. See Figure B1.



Abbreviation: IgE, immunoglobulin E.

Figure B1. Example of a CDA Plot. Quantitative assay for allergen-specific immunoglobulin E with a reportable range of 0.35 to 100 kU/L. Results outside this range are plotted in vertical "gutters" left and right, with status determined by skin testing. Curves represent empirical smoothings (by the Harrell-Davis method) of the cumulative sensitivity and specificity trajectories.

Appendix B. (Continued)

In particular, this practice can safeguard against excessive idealization in studies intended to evaluate an assay (or compare it to another) in its state of development. An assay's measuring interval typically implies a limit to either the sensitivity or specificity that it can achieve, or limits to both capabilities. In the most common scenario, in which values above a given cutoff are construed as positive for the clinical state in question, and in which indefinitely high concentrations can be measured under dilution, there may be no practical limit to the achievable specificity.

The CDA plot is not simply a method for representing information complementary to that in an ROC curve; it is arguably more fundamental, because an (unannotated) ROC curve—and a likelihood ratio (LR) plot—can be constructed from a CDA plot, but not vice versa. An ROC curve is, after all, "just a plot of the cumulative distribution function of x [a one-dimensional continuous diagnostic indicant] in the affected against that of the unaffected subpopulation. The ratio of the corresponding probability densities (LR) becomes its local slope."[2]

From this perspective, the conceptually natural way to construct an ROC curve is first to build a CDA plot, smooth the two trajectories, and determine the sensitivity and specificity pairs from these plots at a grid of potential cutoffs. The resulting plot in ROC (or LR) space can be made as smooth as desired by increasing the density of the concentration value grid.

This approach can take advantage of well-developed univariate techniques: parametric distribution fitting, nonparametric kernel density estimation, local regression methods, and so on. In the study by Ollert et al.,[3] smoothing was accomplished at this stage by using the Harrell-Davis estimator, which is arguably the nonparametric method of choice in clinical chemistry for centile estimation because of its endorsement in *Statistical Bases of Reference Values in Laboratory Medicine.*[4]

More importantly, by displaying both the experimental observations and the smoothed trajectories in the CDA representation, it is possible to judge goodness of fit by eye in a familiar space, taking advantage of the laboratorian's feel for the significance of analytical errors, that is, errors expressed in the assay's native concentration scale. This may well outweigh the (usually) minor loss of efficiency implied by combining the sensitivity and specificity information after smoothing—as opposed to smoothing the combined sensitivity and specificity results directly (ie, in ROC or LR space).

In summary, CDA plots serve as a valuable complement to traditional ROC and LR plots by maintaining contact with the natural analytical measurement characteristics of the assay(s) being evaluated, thus helping to prevent unjustifiable ascription of idealized capabilities to the assays as reflected in published ROC curves. Furthermore, smooth ROC curves can be constructed nonparametrically by combining sensitivity and specificity values read off smoothed CDA trajectories. Curves constructed in this manner offer a check on the reasonableness of assumptions underlying model-based ROC curve construction.

References for Appendix B

1 Krouwer JS. Cumulative distribution analysis graphs—an alternative to ROC curves. *Clin Chem.* 1987;33(12):2305-2306.

2 Hilden J. The area under the ROC curve and its competitors. *Med Decis Making.* 1991;11(2):95-101.

Appendix B. (Continued)

[3] Ollert M, Weissenbacher S, Rakoski J, Ring J. Allergen-specific IgE measured by a continuous random-access immunoanalyzer: interassay comparison and agreement with skin testing. *Clin Chem*. 2005;51(7):1241-1249.

[4] Harris EK, Boyd JC. *Statistical Bases of Reference Values in Laboratory Medicine.* New York, NY: Marcel Dekker, Inc.; 1995.

Appendix C. Receiver Operating Characteristic Curve Areas and Rank-Sum Statistics

There is a close link between the area under the receiver operating characteristic (ROC) curve and the probability of concordance (PoC), defined here as the probability that someone affected with the target clinical condition will have a more pathological test result than someone without it. It is also known as Harrell's *c* index. These notions are in turn closely linked to the rank sums employed in the Wilcoxon-Mann-Whitney test procedure.[1]

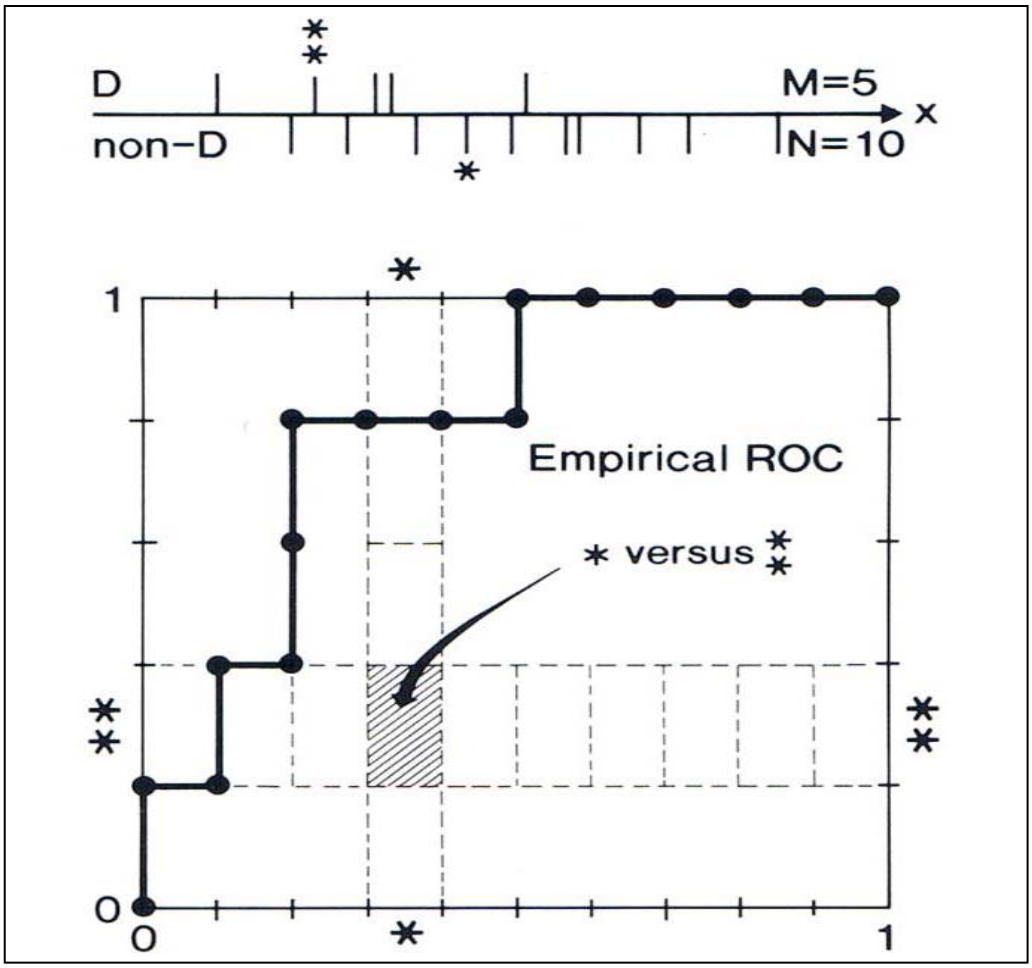


Figure C1. Geometrical Illustration of the Link Between AUC and Rank Tests.[2] (Reprinted with permission from Hilden J. The area under the ROC curve and its competitors. *Med Decis Making*. 1991;11(2):95-101.)

Appendix C. (Continued)

In the small dataset of Figure C1, low values denote those affected by the clinical condition. The sample comprises five individuals affected by the target clinical condition (D) and 10 unaffected by the target clinical condition (non-D), so 50 (5 • 10) D~non-D pairs can be formed. In $U = 40$ (or 80%) of the 50 pairs, the D subject has a more "affected" value of measurand x than his non-D counterpart; that is, the frequency of concordance is $U / (5 \bullet 10) = 0.80$. Thus, $U = 40$, which is the key statistic employed in the Mann-Whitney version of the rank test; this defines the frequency of concordance, which in turn is also the area under the ROC curve (AUC).

To show this, divide the ROC square into small rectangles whose areas are 1/50 each. One observes that there is one rectangle for each of the 50 pairs, 40 below and 10 above the ROC. For instance, the non-D observation marked * exceeds four observations in the D group. Therefore, its contribution to the AUC is four small rectangles, located along the vertical strip that corresponds to observation *. Conversely, the horizontal strip associated with the D case marked ** reflects the fact that nine of the 10 non-D subjects had higher x values.

The rank sums employed in Wilcoxon's version of the test equal U apart from a constant term; therefore, all the procedures mentioned here capitalize on the same idea (ie, PoC).

NOTE: For readers with mathematical training, the proof that the area under the population ROC equals the PoC is simple. Let $f(x)$ be the density of the unaffected distribution, with cumulative function $F(x)$; the analogous notation for the affected population is $g(y)$, $G(y)$. When low values are taken as pathological, $(F(c), G(c))$ is the ROC point generated by choosing cutoff = c. Because the PoC refers to an independently sampled (x, y) pair, one has:

$$\text{PoC} = \iint_{x>y} g(y) f(x)\,dy\,dx = \int_x \{\int_{y<x} g(y)\,dy\}\, f(x)\,dx = \int G(x)\,dF(x)$$

$$= \int (\text{ROC ordinate})\ d(\text{ROC abscissa}) = (\text{area under the ROC}).$$

References for Appendix C

1 Moses L. Wilcoxon-Mann-Whitney test. In: Armitage P, Colton T, eds. *Encyclopedia of Biostatistics*. Vol 6. New York, NY: Wiley; 1998:4742-4745.

2 Hilden J. The area under the ROC curve and its competitors. *Med Decis Making.* 1991;11(2):95-101.

Appendix D. A Receiver Operating Characteristic Curve Comparison Example

Although a step-by-step technique for generating receiver operating characteristic (ROC) curves has been presented, it is assumed that most readers will use off-the-shelf software for this task. Example 1, with the corresponding data in Table 2, the ROC curve in Figure 2, and the area under the curve (AUC) result in Table 4 (in the main text of this guideline), provides an example of a single ROC curve for users against which to test such software. The following is an example of a two-sampled paired test that can be similarly used.

The two assays considered are low-density lipoprotein (LDL) and oxidized low-density lipoprotein (OxLDL). OxLDL is thought to be the active molecule in the process of atherosclerosis, so its proponents believe that its serum concentration should provide more accurate risk stratification than the traditional LDL assay. Table D1 below lists the unsorted values for diagnosis, OxLDL, and LDL for each patient. Table D2 gives the one-sample computations of the AUCs and their respective standard errors (SEs),[1] along with the results associated with a test of significance for the difference in AUCs for the two assays. Table D2 provides the results of the test of significance for this difference using the method of Hanley and McNeil.[2] These computations reflect the formulae given in equations (5) and (6) in Section 7.2.2 and equations (7) to (9) in Section 7.2.4 of the main text of the guideline.[2] The ROC curves for the two assays are illustrated in Figure D1.

Table D1. OxLDL and LDL Assay Values (in U/L) for 50 Subjects

Diagnosis	OxLDL	LDL
0	37	2.1
0	44	2.35
0	42	3.91
0	62	5.4
0	42	3.31
0	61	3.9
0	77	4.38
0	51	2.85
0	52	3.67
0	60	1.48
0	74	2.6
0	73	3.25
0	70	3.76
0	64	3.5
0	54	2.66
0	66	4.45
0	63	5.27
0	54	3.57
0	66	3.74
0	48	2.78
0	59	3.15
0	22	3.01
1	83	5.88
1	86	4.05
1	57	3.75
1	76	3.21
1	96	4.11
1	77	4.15
1	72	2.31
1	71	2.57

Appendix D. (Continued)

Table D1. (Continued)

Diagnosis	OxLDL	LDL
1	41	2.6
1	95	4.22
1	116	7.55
1	60	2.74
1	77	4.57
1	66	3.51
1	76	3.08
1	60	2.95
1	143	5.71
1	88	3.92
1	64	3.38
1	73	1.29
1	78	3.71
1	53	3.22
1	60	3.4
1	78	3.4
1	82	4.03
1	66	3.09
1	76	3.47
1	45	3.57

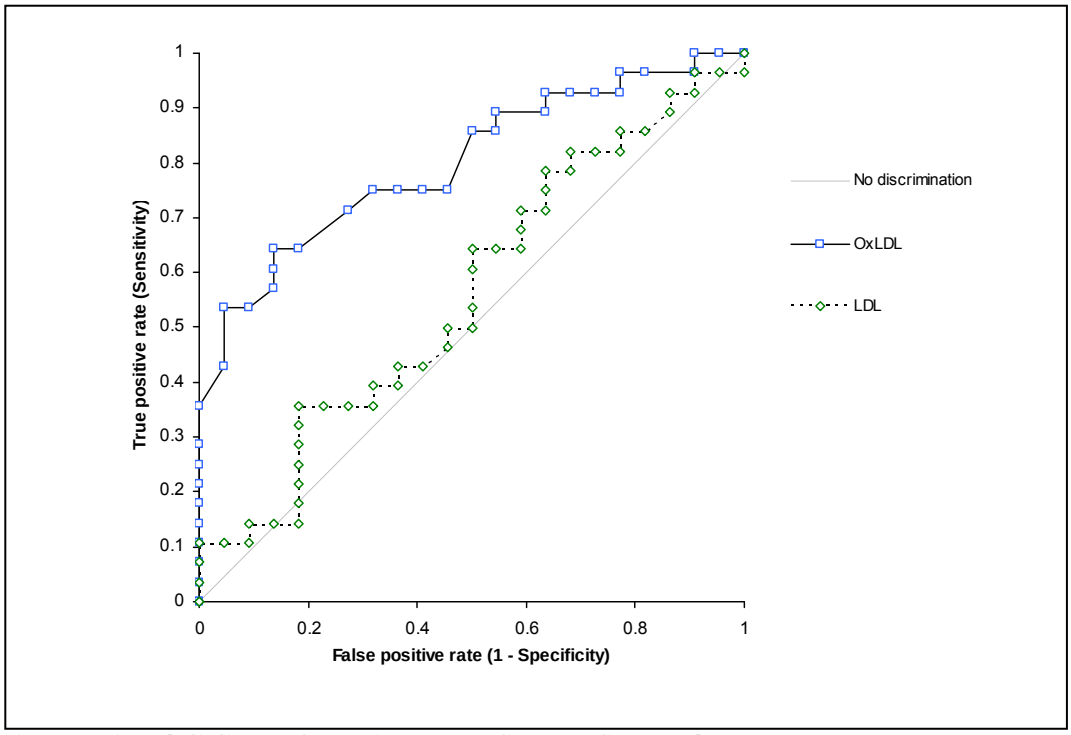


Figure D1. ROC Curve for Paired Two-Sample Study (OxLDL and LDL)

Appendix D. (Continued)

Table D2. AUC and Summary Statistics[1,2]

Test	AUC	SE	95% CI for AUC	Z	Hypothesis Test	*p* Value
OxLDL	0.80	0.062	0.68–0.92	4.83	H_0: Area ≤ 0.5. H_1: Area >0.5.	<0.0001
LDL	0.56	0.082	0.40–0.72	0.76	H_0: Area ≤ 0.5. H_1: Area >0.5.	0.2245
Difference	0.24	0.075	0.09–0.39	3.16	H_0: Difference between areas $= 0$. H_1: Difference between areas $\neq 0$.	0.0016

Abbreviations: AUC, area under the curve; CI, confidence interval; LDL, low-density lipoprotein; OxLDL, oxidized low-density lipoprotein; SE, standard error.

References for Appendix D

1 Zhou XH, Obuchowski NA, McClish DK. *Statistical Methods in Diagnostic Medicine.* New York, NY: Wiley & Sons; 2002.

2 Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*. 1983;148(3):839-843.

This page is intentionally left blank.

The Quality Management System Approach

Clinical and Laboratory Standards Institute (CLSI) subscribes to a quality management system approach in the development of standards and guidelines, which facilitates project management; defines a document structure via a template; and provides a process to identify needed documents. The quality management system approach applies a core set of "quality system essentials" (QSEs), basic to any organization, to all operations in any health care service's path of workflow (ie, operational aspects that define how a particular product or service is provided). The QSEs provide the framework for delivery of any type of product or service, serving as a manager's guide. The QSEs are as follows:

Organization	Personnel	Process Management	Nonconforming Event Management
Customer Focus	Purchasing and Inventory	Documents and Records	Assessments
Facilities and Safety	Equipment	Information Management	Continual Improvement

EP24-A2 addresses the QSE indicated by an "X." For a description of the other documents listed in the grid, please refer to the Related CLSI Reference Materials section on the following page.

Organization	Customer Focus	Facilities and Safety	Personnel	Purchasing and Inventory	Equipment	Process Management	Documents and Records	Information Management	Nonconforming Event Management	Assessments	Continual Improvement
		M29				X C28 EP17					

Related CLSI Reference Materials*

C28-A3c **Defining, Establishing, and Verifying Reference Intervals in the Clinical Laboratory; Approved Guideline—Third Edition (2008).** This document contains guidelines for determining reference values and reference intervals for quantitative clinical laboratory tests. A CLSI-IFCC joint project.

EP17-A **Protocols for Determination of Limits of Detection and Limits of Quantitation; Approved Guideline (2004).** This document provides guidance for determining the lower limit of detection of clinical laboratory methods, for verifying claimed limits, and for the proper use and interpretation of the limits. An NCCLS-IFCC joint project.

M29-A3 **Protection of Laboratory Workers From Occupationally Acquired Infections; Approved Guideline—Third Edition (2005).** Based on US regulations, this document provides guidance on the risk of transmission of infectious agents by aerosols, droplets, blood, and body substances in a laboratory setting; specific precautions for preventing the laboratory transmission of microbial infection from laboratory instruments and materials; and recommendations for the management of exposure to infectious agents.

* CLSI documents are continually reviewed and revised through the CLSI consensus process; therefore, readers should refer to the most current editions.

Active Membership
(As of 1 November 2011)

Sustaining Members

Abbott
American Association for Clinical Chemistry
AstraZeneca Pharmaceuticals
Bayer Healthcare, LLC Diagnostic Division
BD
Beckman Coulter, Inc.
bioMérieux, Inc.
College of American Pathologists
Diagnostica Stago
GlaxoSmithKline
National Institute of Standards and Technology
Ortho-Clinical Diagnostics, Inc.
Pfizer Inc
Roche Diagnostics, Inc.

Professional Members

AAMI
American Association for Clinical Chemistry
American Association for Laboratory Accreditation
American Medical Technologists
American Society for Clinical Laboratory Science
American Society for Clinical Pathology
American Society for Microbiology
American Type Culture Collection
Association of Public Health Laboratories
Associazione Microbiologi Clinici Italiani (AMCLI)
British Society for Antimicrobial Chemotherapy
Canadian Society for Medical Laboratory Science
COLA
College of American Pathologists
College of Medical Laboratory Technologists of Ontario
College of Physicians and Surgeons of Saskatchewan
Critical Path Institute
ESCMID
Family Health International
Hong Kong Accreditation Service Innovation and Technology Commission
International Federation of Biomedical Laboratory Science
International Federation of Clinical Chemistry
Italian Society of Clin. Biochem. and Clin. Molec. Biology
JCCLS
National Society for Histotechnology, Inc.
Nova Scotia Association of Clinical Laboratory Managers
Ontario Medical Association Quality Management Program-Laboratory Service
RCPA Quality Assurance Programs Pty Limited
SIMeL
Sociedad Española de Bioquímica Clínica y Patología Molec.
Sociedade Brasileira de Análises Clínicas
Sociedade Brasileira de Patologia Clínica
The Joint Commission
The Korean Society for Laboratory Medicine
World Health Organization

Government Members

Armed Forces Institute of Pathology
BC Centre for Disease Control
Canadian Science Center for Human and Animal Health
Centers for Disease Control and Prevention
Centers for Disease Control and Prevention - Ethiopia
Centers for Disease Control and Prevention - Namibia
Centers for Disease Control and Prevention - Nigeria
Centers for Disease Control and Prevention - Tanzania
Centers for Disease Control and Prevention - Zambia
Centers for Medicare & Medicaid Services
Centers for Medicare & Medicaid Services/CLIA Program
Chinese Committee for Clinical Laboratory Standards
Chinese Medical Association (CMA)
Clalit Health Services
Department of Veterans Affairs
DFS/CLIA Certification
Diagnostic Accreditation Program
Ethiopian Health and Nutrition Research Institute
FDA Center for Veterinary Medicine
FDA Ctr. for Devices/Rad. Health
Health Canada
Institute of Tropical Medicine Dept. of Clinical Sciences
MA Dept. of Public Health Laboratories
Malaria Research Training Center
Marion County Public Health Department
Meuhedet Central Lab
Ministry of Health and Social Welfare - Tanzania
Mongolian Agency for Standardization and Metrology
Namibia Institute of Pathology
National Cancer Institute, OBBR, NIH
National Food Institute Technical University of Denmark
National Health Laboratory Service C/O F&M Import & Export Services
National HIV & Retrovirology Lab Public Health Agency of Canada
National Institute of Health-Maputo, Mozambique
National Institute of Standards and Technology
National Pathology Accreditation Advisory Council
New York State Dept. of Health
Ontario Agency for Health Protection and Promotion
Pennsylvania Dept. of Health
SA Pathology
Saskatchewan Health-Provincial Laboratory
Scientific Institute of Public Health
State of Alabama
State of Wyoming Public Health Laboratory
The Nathan S. Kline Institute
University of Iowa, Hygienic Lab
US Naval Medical Research Unit #3
USAMC - AFRIMS

Industry Members

3M Medical Division
AB Diagnostic Systems GmbH
Abbott
Abbott Diabetes Care
Abbott Point of Care Inc.
Access Genetics
Aderans Research
AdvaMed
Akonni Biosystems
Ammirati Regulatory Consulting
Anapharm, Inc.
AspenBio Pharma, Inc.
Astellas Pharma
AstraZeneca Pharmaceuticals
Astute Medical, Inc.
Ativa Medical
Axis-Shield PoC AS
Bayer Healthcare, LLC Diagnostic Division
BD
BD Biosciences - San Jose, CA
BD Diagnostic Systems
BD Vacutainer Systems
Beaufort Advisors, LLC
Beckman Coulter Cellular Analysis Business Center
Beckman Coulter, Inc.
Beth Goldstein Consultant
Bio-Rad Laboratories, Inc.
Bio-Rad Laboratories, Inc. - France
Bioanalyse, Ltd.
Biocartis
BioDevelopment S.r.l.
Biohit Oyj.
Biomedia Laboratories SDN BHD
bioMérieux, Inc.
Blaine Healthcare Associates, Inc.
BRI Consultants Limited
Calloway Laboratories
Canon U.S. Life Sciences, Inc.
CBI Inc.
Cempra Pharmaceuticals, Inc.
Cepheid
Cerilliant Corp.
Compliance Insight, Inc.
Constitution Medical Inc
Controllab
Copan Diagnostics Inc.
Crescendo Bioscience
Cubist Pharmaceuticals, Inc.
Dahl-Chase Pathology Associates PA
Diagnostica Stago
DX Assays Pte Ltd.
Eiken Chemical Company, Ltd.
Elanco Animal Health
Elkin Simson Consulting Services
Emika Consulting
EndPoint Associates, LLC
Enigma Diagnostics
Eurofins Medinet
Evidia Biosciences Inc.
EXACT Sciences Corporation
Gen-Probe
Genefluidics
GlaxoSmithKline
Greiner Bio-One Inc.
Himedia Labs Ltd
HistoGenex N.V.
Hospital Sungai Buloh
Icon Laboratories, Inc.
Innovotech, Inc.
Instrumentation Laboratory
Integrated BioBank
IntelligentMDx, Inc.
Intuity Medical
ITC Corp
Japan Assn. of Clinical Reagents Industries
Johnson & Johnson Pharmaceutical Research & Develop., L.L.C.
Kaiser Permanente
KoreaBIO
Krouwer Consulting
Lab PMM
Laboratory Specialists, Inc.
LifeLabs
LifeScan, Inc.
Liofilchem SRL
LipoScience, Inc.
Maine Standards Company, LLC
Marketing MicroScan & Molecular Korea-Siemens Healthcare
Masimo Corp.
Masimo Labs
Mbio Diagnostics, Inc.
MDxHealth SA
Medical Device Consultants, Inc.
Merck & Company, Inc.
Merial Limited
Meso Scale Diagnostics, LLC.
Micromyx, LLC
Molecular Response
Moscow Antidoping Agency
Nanosphere, Inc.
Nihon Kohden Corporation
Nissui Pharmaceutical Co., Ltd.
NJK & Associates, Inc.
NorDx - Scarborough Campus
Nova Biomedical Corporation
NovaBiotics
Novartis Institutes for Biomedical Research
Optimer Pharmaceuticals, Inc.
Ortho-Clinical Diagnostics, Inc.
Ortho-McNeil, Inc.
Oxyrase, Inc.
Paratek Pharmaceuticals, Inc.
PathCare Pathology Laboratory
PerkinElmer Genetics, Inc.
Pfizer Animal Health
Pfizer Inc
Pfizer Italia Srl
Phadia AB
Philips Healthcare Incubator
PPD
ProteoGenix, Inc.
QML Pathology
Quotient Bioresearch Ltd.
R-Biopharm AG
Radiometer America, Inc.
Roche Diagnostics GmbH
Roche Diagnostics, Inc.
Roche Molecular Systems
RPL Laboratory Solutions, Inc. DBA RPL Compliance Solutions
Sanofi Pasteur
Sarstedt, Inc.
Sekisui Diagnostics
Seventh Sense Biosystems
Siemens Healthcare Diagnostics Inc.
Siemens Healthcare Diagnostics Products GmbH
Soloy Laboratory Consulting Services, Llc
SomaLogic
Sphere Medical Holding Limited
Streck Laboratories, Inc.
Super Religare Laboratories Ltd
Sysmex America, Inc.
Sysmex Corporation - Japan
Tetraphase Pharmaceuticals
The Clinical Microbiology Institute
The Medicines Company
TheraDoc
Theranos
Theravance Inc.
Thermo Fisher Scientific
Thermo Fisher Scientific, Oxoid Products
Thermo Fisher Scientific, Remel
Transasia Bio-Medicals Limited
Trek Diagnostic Systems
Tulip Group
Ventana Medical Systems Inc.
Veracyte, Inc.
Vivacta
Watson Pharmaceuticals
Wellstat Diagnostics, LLC
XDx, Inc.

Associate Active Members

31st Medical Group SGSL (AE)
3rd Medical Group (AK)
48th Medical Group/MDSS RAF Lakenheath (AE)
55th Medical Group/SGSAL (NE)
59th MDW/859th MDTS/MTL Wilford Hall Medical Center (TX)
82 MDG/SGSCL Sheppard AFB (TX)
Academisch Ziekenhuis-VUB UZ Brussel (Belgium)
ACL Laboratories (IL)
ACL Laboratories (WI)
Adams County Hospital (OH)
Adena Regional Medical Center Hospital (OH)
Affiliated Laboratory, Inc. (ME)
Akron Children's Hospital (OH)
Al Ain Hospital (Abu Dhabi, United Arab Emirates)
Al Hada Armed Forces Hospital/TAIF/KSA (Saudi Arabia)
Al Noor Hospital (United Arab Emirates)
Al Rahba Hospital (United Arab Emirates)
Alameda County Medical Center (CA)
Albany Medical Center Hospital (NY)
Albemarle Hospital (NC)
Alberta Health Services (AB, Canada)
Alexandra Hospital (Singapore)
All Children's Hospital (FL)
Allegiance Health (MI)
Alpena Regional Medical Center (MI)
Alta Bates Summit Medical Center (CA)
Alverno Clinical Laboratories, Inc. (IN)
American Esoteric Laboratories (AEL) (TN)
American University of Beirut Medical Center (NJ)
Anand Diagnostic Laboratory (India)
Anne Arundel Medical Center (MD)
Antech Diagnostics (CA)
Antelope Valley Hospital District (CA)
Appalachian Regional Healthcare System (NC)
Arkansas Children's Hospital (AR)
Arkansas Dept of Health Public Health Laboratory (AR)
Arkansas Methodist Medical Center (AR)
Arnot Ogden Medical Center Laboratory (NY)
ARUP Laboratories (UT)
Asan Medical Center (Korea, Republic Of)
Asante Health System (OR)
Ashley County Medical Center (AR)
Asiri Group of Hospitals Ltd. (Sri Lanka)
Aspen Valley Hospital (CO)
ASPETAR (Qatar Orthopedic and Sports Medicine Hospital) (Qatar)
Aspirus Wausau Hospital (WI)
Auburn Regional Medical Center (WA)
Augusta Health (VA)
Aultman Hospital (OH)
Avera McKennan Hospital & University Health Center (SD)
AZ Sint-Jan (Belgium)
Azienda Ospedale Di Lecco (Italy)
Baptist Hospital of Miami (FL)
Baptist Memorial Health Care Corporation - Hospital Laboratories Works (TN)
Barnes-Jewish Hospital (MO)
Bassett Healthcare (NY)
Baton Rouge General (LA)
Baxter Regional Medical Center (AR)
BayCare Health System (FL)
Baylor Health Care System (TX)
Bayou Pathology, APMC (LA)
Baystate Medical Center (MA)
BC Biomedical Laboratories (BC, Canada)
Beloit Memorial Hospital (WI)
Berg Diagnostics (MA)
Beth Israel Medical Center (NY)
Bio-Reference Laboratories (NJ)

Blanchard Valley Hospital (OH)
Bon Secours Health Partners (VA)
Bonnyville Health Center (AB, Canada)
Boston Medical Center (MA)
Boulder Community Hospital (CO)
Boyce & Bynum Pathology Labs (MO)
Brant Community Healthcare System/Brant General Hospital (Ontario, Canada)
Bremerton Naval Hospital (WA)
Brian All Good Community Hospital/121 Combat (AP)
Bridgeport Hospital (CT)
Brooke Army Medical Center (TX)
Broward General Medical Center (FL)
Cadham Provincial Laboratory-MB Health (MB, Canada)
Calgary Laboratory Services (AB, Canada)
California Pacific Medical Center (CA)
Cambridge Health Alliance (MA)
Cape Fear Valley Medical Center Laboratory (NC)
Capital Coast Health (New Zealand)
Capital Health System Mercer Campus (NJ)
Caritas Norwood Hospital (MA)
Carl R. Darnall Army Medical Center Department of Pathology (TX)
Carolina Medical Laboratory (NC)
Carolinas Healthcare System (NC)
Carpermor S.A. de C.V. (D.F., Mexico)
Catholic Health Initiatives (KY)
Cedars-Sinai Medical Center (CA)
Cenetron Diagnostics (TX)
Central Baptist Hospital (KY)
Central Kansas Medical Center (KS)
Centre Hospitalier Anna-Laberge (Quebec, Canada)
Centre Hospitalier Regional De Trois Riveras (PQ, Canada)
Centro Médico Imbanaco (Colombia)
Chaleur Regional Hospital (NB, Canada)
Chang Gung Memorial Hospital (Taiwan)
Changhua Christian Hospital (Taiwan)
Changi General Hospital (Singapore)
Chatham - Kent Health Alliance (ON, Canada)
Chesapeake General Hospital (VA)
Chester County Hospital (PA)
Children's Healthcare of Atlanta (GA)
Childrens Hosp.- Kings Daughters (VA)
Children's Hospital & Research Center At Oakland (CA)
Childrens Hospital Los Angeles (CA)
Children's Hospital Medical Center (OH)
Children's Hospital of Central California (CA)
Children's Hospital of Orange County (CA)
Children's Hospital of Philadelphia (PA)
Childrens Hospital of Wisconsin (WI)
Children's Hospitals and Clinics (MN)
Children's Medical Center (OH)
Children's Medical Center (TX)
Christiana Care Health Services (DE)
CHU - Saint Pierre (Belgium)
CHU Sainte-Justine (Quebec, Canada)
CHUM Hospital Saint-Luc (Quebec, Canada)
CHW-St. Mary's Medical Center (CA)
City of Hope National Medical Center (CA)
Cleveland Clinic (OH)
Clinica Alemana De Santiago (Chile)
Clinical and Laboratory Standards Institute (PA)
Clinical Labs of Hawaii (HI)
College of Physicians and Surgeons of Alberta (AB, Canada)
Collingwood General & Marine Hospital (ON, Canada)
Commonwealth of Kentucky (KY)
Commonwealth of Virginia (DCLS) (VA)
Community Hospital (IN)
Community Hospital of the Monterey Peninsula (CA)
Community Medical Center (NJ)
Community Memorial Hospital (WI)
Complexe Hospitalier de la Sagamie (Quebec, Canada)
CompuNet Clinical Laboratories Quest Diagnostics JV (OH)
Concord Hospital (NH)
Consultants Laboratory of WI LLC (WI)
Contra Costa Regional Medical Center (CA)
Cook Children's Medical Center (TX)
Cookeville Regional Medical Center (TN)
Cornwall Community Hospital (ON, Canada)
Covance CLS (IN)
Covenant Medical Center (TX)
Creighton Medical Lab (NE)
Crozer-Chester Medical Center (PA)
Cumberland Medical Center (TN)
Darwin Library NT Territory Health Services (NT, Australia)
David Grant Medical Center (CA)
Daviess Community Hospital (IN)
Deaconess Hospital Laboratory (IN)
Dean Medical Center (WI)
Denver Health & Hospital Authority (CO)
DHHS NC State Lab of Public Health (NC)
DiagnoSearch Life Sciences Inc. (Maharashtra, India)
Diagnostic Laboratory Services, Inc. (HI)
Diagnostic Services of Manitoba (MB, Canada)
Dimensions Healthcare System Prince George's Hospital Center (MD)
DMC University Laboratories (MI)
Drake Center (OH)
Driscoll Children's Hospital (TX)
DUHS Clinical Laboratories Franklin Site (NC)
Dynacare Laboratory (WI)
Dynacare NW, Inc - Seattle (WA)
DynaLIFE (AB, Canada)
E. A. Conway Medical Center (LA)
East Georgia Regional Medical Center (GA)
East Kootenay Regional Hospital Laboratory-Interior Health (BC, Canada)
East Texas Medical Center-Pittsburg (TX)
Eastern Health - Health Sciences Centre (NL, Canada)
Eastern Health Pathology (Victoria, Australia)
Easton Hospital (PA)
Edward Hospital (IL)
Effingham Hospital (GA)
Elmhurst Hospital Center (NY)
Emory University Hospital (GA)
Evangelical Community Hospital (PA)
Evans Army Community Hospital (CO)
Exeter Hospital (NH)
Exosome Diagnostics, Inc. (MN)
Federal Medical Center (MN)
First Health of the Carolinas Moore Regional Hospital (NC)
Fletcher Allen Health Care (VT)
Fleury S.A. (Brazil)
Florida Hospital (FL)
Fox Chase Cancer Center (PA)
Fraser Health Authority Royal Columbian Hospital Site (BC, Canada)
Gamma-Dynacare Laboratories (ON, Canada)
Garden City Hospital (MI)
Garfield Medical Center (CA)
Gaston Memorial Hospital (NC)
Geisinger Medical Center (PA)
Genesis Healthcare System (OH)
George Washington University Hospital (DC)
Gestión de Calidad (Argentina)
Gettysburg Hospital (PA)
Ghent University Hospital (Belgium)
Good Shepherd Medical Center (TX)
Grana S.A. (TX)
Grand Marquis Co., LTD (Taiwan)
Grand River Hospital (ON, Canada)
Grand Strand Regional Medical Center (SC)
Grey Bruce Regional Health Center (ON, Canada)
Group Health Cooperative (WA)
Gundersen Lutheran Medical Center (WI)
Guthrie Clinic Laboratories (PA)
Hôtel-Dieu de Lévis (PQ, Canada)
Halton Healthcare Services (ON, Canada)
Hamad Medical Corporation (Qatar)
Hamilton Regional Laboratory Medicine Program - St. Joseph's (ON, Canada)
Hanover General Hospital (PA)
Harford Memorial Hospital (MD)
Harris Methodist Fort Worth (TX)
Harris Methodist Hospital Southwest (TX)
Hartford Hospital (CT)
Health Network Lab (PA)
Health Sciences Research Institute (Japan)
Health Waikato (New Zealand)
Heartland Health (MO)
Heidelberg Army Hospital (AE)
Helen Hayes Hospital (NY)
Helix (Russian Federation)
Henry Ford Hospital (MI)
Henry M. Jackson Foundation for the Advancement of Military Medicine-MD (MD)
Hi-Desert Medical Center (CA)
Highlands Medical Center (AL)
HJF Naval Infectious Diseases Diagnostic Laboratory (MD)
Hoag Memorial Hospital Presbyterian (CA)
Hoboken University Medical Center (NJ)
Holy Cross Hospital (MD)
Holy Name Hospital (NJ)
Holy Spirit Hospital (PA)
Hôpital de la Cité-de-La-Santé De Laval (Quebec, Canada)
Hôpital du Haut-Richelieu (PQ, Canada)
Hôpital Maisonneuve-Rosemont (PQ, Canada)
Hôpital Santa Cabrini Ospedale (PQ, Canada)
Horizon Health Network (NB, Canada)
Hospital Albert Einstein (SP, Brazil)
Hospital Sacre-Coeur de Montreal (Quebec, Canada)
Hôtel-Dieu Grace Hospital Library (ON, Canada)
Hunter Area Pathology Service (Australia)
Hunter Labs (CA)
Huntington Memorial Hospital (CA)
Imelda Hospital (Belgium)
Indian River Memorial Hospital (FL)
Indiana University Health Bloomington Hospital (IN)
Indiana University Health Care-Pathology Laboratory (IN)
Inova Central Laboratory (VA)
Institut fur Stand. und Dok. im Med. Lab. (Germany)
Institut National de Santé Publique Du Quebec Centre de Doc. - INSPQ (PQ, Canada)
Institute Health Laboratories (PR)
Institute of Clinical Pathology and Medical Research (Australia)
Institute of Laboratory Medicine Landspitali Univ. Hospital (Iceland)
Institute of Medical & Veterinary Science (SA, Australia)
Intermountain Health Care Lab Services (UT)
International Health Management Associates, Inc. (IL)
Irwin Army Community Hospital (KS)
Jackson County Memorial Hospital (OK)
Jackson Memorial Hospital (FL)
Jackson Purchase Medical Center (KY)
Jessa Ziekenhuis VZW (Belgium)
John C. Lincoln Hospital - N.MT. (AZ)
John F. Kennedy Medical Center (NJ)
John H. Stroger, Jr. Hospital of Cook County (IL)
John Muir Health (CA)
John T. Mather Memorial Hospital (NY)
Johns Hopkins Medical Institutions (MD)
Johns Hopkins University (MD)
Johnson City Medical Center Hospital (TN)
JPS Health Network (TX)
Kailos Genetics (AL)
Kaiser Permanente (MD)
Kaiser Permanente Medical Care (CA)
Kenora-Rainy River Reg. Lab. Program (ON, Canada)
King Abdulaziz Hospital, Al Ahsa Dept. of Pathology & Laboratory Medicine (Al-hasa, Saudi Arabia)
King Fahad National Guard Hospital KAMC - NGHA (Saudi Arabia)
King Fahad Specialist Hospital-Dammam, K.S.A. (Eastern Region, Saudi Arabia)
King Faisal Specialist Hospital & Research Center (Saudi Arabia)
King Hussein Cancer Center (Jordan)
Kingston General Hospital (ON, Canada)
Laboratória Médico Santa Luzia Ltda (Brazil)
Laboratory Alliance of Central New York (NY)
Laboratory Corporation of America (NJ)
Laboratory Medicin Dalarna (Dalarna, Sweden)
LabPlus Auckland District Health Board (New Zealand)
LAC/USC Medical Center (CA)
Lafayette General Medical Center (LA)
Lakeland Regional Medical Center (FL)
Lancaster General Hospital (PA)
Landstuhl Regional Medical Center (Germany)
Langley Air Force Base (VA)
LeBonheur Children's Hospital (TN)
Legacy Laboratory Services (OR)
Letherbridge Regional Hospital (AB, Canada)
Lewis-Gale Medical Center (VA)
Lexington Medical Center (SC)
L'Hotel-Dieu de Québec (PQ, Canada)
Licking Memorial Hospital (OH)
LifeBridge Health Sinai Hospital (MD)
LifeLabs Medical Laboratory Services (BC, Canada)
Lifeline Hospital (United Arab Emirates)
Loma Linda University Medical Center (LLUMC) (CA)
Long Beach Memorial Medical Center-LBMMC (CA)
Long Island Jewish Medical Center (NY)
Louisiana Office of Public Health Laboratory (LA)
Louisiana State University Medical Ctr. (LA)
Lower Columbia Pathologists, P.S. (WA)
Lower Mainland Laboratories (BC, Canada)
Lyndon B. Johnson General Hospital (TX)
Maccabi Medical Care and Health Fund (Israel)
Mafraq Hospital (United Arab Emirates)
Magnolia Regional Health Center (MS)
Main Line Clinical Laboratories, Inc. Lankenau Hospital (PA)
Makerere University Walter Reed Project Makerere University Medical School (Uganda)
Marquette General Hospital (MI)
Marshfield Clinic (WI)
Martha Jefferson Hospital (VA)
Martin Luther King, Jr./Drew Medical Center (CA)
Martin Memorial Health Systems (FL)
Mary Hitchcock Memorial Hospital (NH)
Mary Washington Hospital (VA)
Mater Health Services - Pathology (Australia)
Maxwell Air Force Base (AL)
Mayo Clinic (MN)
MCG Health (GA)
Meadows Regional Medical Center (GA)
Medical Center Hospital (TX)
Medical Center of Louisiana At NO-Charity (LA)
Medical Centre Ljubljana (Slovenia)
Medical College of Virginia Hospital (VA)
Medical University of South Carolina (SC)
Memorial Hermann Healthcare System (TX)
Memorial Medical Center (PA)
Memorial Medical Center (IL)
Memorial Regional Hospital (FL)
Mercy Franciscan Mt. Airy (OH)
Mercy Hospital & Medical Center (IL)
Methodist Dallas Medical Center (TX)
Methodist Hospital (TX)
Methodist Hospital Park Nicollet Health Services (MN)
Methodist Hospital Pathology (NE)
MetroHealth Medical Center (OH)
Metropolitan Hospital Center (NY)
Metropolitan Medical Laboratory, PLC (IA)
Miami Children's Hospital (FL)
Mid Michigan Medical Center - Midland (MI)
Middelheim General Hospital (Belgium)
Middlesex Hospital (CT)
Mike O'Callaghan Federal Hospital (NV)
Minneapolis Medical Research Foundation (MN)
Mississippi Baptist Medical Center (MS)
Mississippi Public Health Lab (MS)
Monongalia General Hospital (WV)
Montreal General Hospital (Quebec, Canada)
Morehead Memorial Hospital (NC)
Mouwasat Hospital (GA, Saudi Arabia)
Mt. Carmel Health System (OH)
Mt. Sinai Hospital (ON, Canada)
Mt. Sinai Hospital - New York (NY)
Naples Community Hospital (FL)
Nassau County Medical Center (NY)
National B Virus Resource Laboratory (GA)
National Cancer Center (Korea, Republic Of)
National Institutes of Health, Clinical Center (MD)
National Naval Medical Center (MD)
National University Hospital Department of Laboratory Medicine (Singapore)
National University of Ireland, Galway (NUIG) (Ireland)
Nationwide Children's Hospital (OH)
Naval Hospital Oak Harbor (WA)
Naval Medical Center Portsmouth (VA)
Naval Medical Clinic Hawaii (HI)
NB Department of Health (NB, Canada)
New England Baptist Hospital (MA)
New England Sinai Hospital (MA)
New Lexington Clinic (KY)
New York City Department of Health and Mental Hygiene (NY)
New York Presbyterian Hospital (NY)
New York University Medical Center (NY)
Newark Beth Israel Medical Center (NJ)
Newfoundland Public Health Laboratory (NL, Canada)

North Carolina Baptist Hospital (NC)
North District Hospital (China)
North Mississippi Medical Center (MS)
North Shore Hospital Laboratory (New Zealand)
North Shore-Long Island Jewish Health System Laboratories (NY)
Northridge Hospital Medical Center (CA)
Northside Hospital (GA)
Northside Medical Center (OH)
Northwest Texas Hospital (TX)
Northwestern Memorial Hospital (IL)
Norton Healthcare (KY)
Ochsner Clinic Foundation (LA)
Ohio State University Hospitals (OH)
Ohio Valley Medical Center (WV)
Onze Lieve Vrouwziekenhuis (Belgium)
Ordre Professionnel Des Technologistes Médicaux Du Quebec (Quebec, Canada)
Orebro University Hospital (Sweden)
Orlando Health (FL)
Ospedale Casa Sollievo Della Sofferenza - IRCCS (Italy)
Our Lady's Hospital For Sick Children (Ireland)
Palmetto Baptist Medical Center (SC)
Pamela Youde Nethersole Eastern Hospital (Hong Kong East Cluster) (Hong Kong)
Pathgroup (TN)
Pathlab (IA)
Pathology and Cytology Laboratories, Inc. (KY)
Pathology Associates Medical Lab. (WA)
Pathology Inc. (CA)
Penn State Hershey Medical Center (PA)
Pennsylvania Hospital (PA)
Peterborough Regional Health Centre (ON, Canada)
PHS Indian Hospital - Pine Ridge (SD)
Piedmont Hospital (GA)
Pitt County Memorial Hospital (NC)
Potomac Hospital (VA)
Prairie Lakes Hospital (SD)
Presbyterian Hospital - Laboratory (NC)
Presbyterian/St. Luke's Medical Center (CO)
Prince of Wales Hospital (Hong Kong)
Princess Margaret Hospital (Hong Kong, China)
Providence Alaska Medical Center (AK)
Providence Health Services, Regional Laboratory (OR)
Provincial Laboratory for Public Health (AB, Canada)
Queen Elizabeth Hospital (P.E.I, Canada)
Queen Elizabeth Hospital (China)
Queensland Health Pathology Services (Australia)
Queensway Carleton Hospital (ON, Canada)
Quest Diagnostics, Incorporated (CA)
Quintiles Laboratories, Ltd. (GA)
Rady Children's Hospital San Diego (CA)
Ramathibodi Hospital (Thailand)
Redington-Fairview General Hospital (ME)
Regions Hospital (MN)
Reid Hospital & Health Care Services (IN)
Reinier De Graaf Groep (Netherlands)
Renown Regional Medical Center (NV)
Research Medical Center (MO)
Response Genetics, Inc. (CA)
RIPAS Hospital (Brunei-Maura, Brunei Darussalam)
Riverside County Regional Medical Center (CA)
Riverside Health System (VA)
Riverside Methodist Hospital (OH)
Riyadh Armed Forces Hospital, Sulaymainia (Saudi Arabia)
Rockford Memorial Hospital (IL)
Royal Victoria Hospital (ON, Canada)
SAAD Specialist Hospital (Saudi Arabia)
Sacred Heart Hospital (WI)
Sacred Heart Hospital (FL)
Sahlgrenska Universitetssjukhuset (Sweden)
Saint Francis Hospital & Medical Center (CT)
Saint Mary's Regional Medical Center (NV)
Saints Memorial Medical Center (MA)
Salem Memorial District Hospital (MO)
Sampson Regional Medical Center (NC)
Samsung Medical Center (Korea, Republic Of)
San Francisco General Hospital-University of California San Francisco (CA)
Sanford USD Medical Center (SD)
Santa Clara Valley Medical Center (CA)
SARL Laboratoire Caron (France)
Scott & White Memorial Hospital (TX)
Seattle Children's Hospital/Children's Hospital and Regional Medical Center (WA)
Sebastian River Medical Center (FL)
Seoul National University Hospital (Korea, Republic Of)
Seoul St. Mary's Hospital (Korea, Republic Of)
Seton Healthcare Network (TX)
Seton Medical Center (CA)
Sharp Health Care Laboratory Services (CA)
Sheik Kalifa Medical City (United Arab Emirates)
Shore Memorial Hospital (NJ)
Singapore General Hospital (Singapore)
South Bend Medical Foundation (IN)
South Eastern Area Laboratory Services (NSW, Australia)
South Miami Hospital (FL)
Southern Community Laboratories (Canterbury, New Zealand)
Southern Health Care Network (Australia)
Southwest Healthcare System (CA)
Southwestern Medical Center (OK)
Spectra East (NJ)
Spectra Laboratories (CA)
St. Agnes Healthcare (MD)
St. Anthony Hospital (OK)
St. Barnabas Medical Center (NJ)
St. Elizabeth Community Hospital (CA)
St. Eustache Hospital (Quebec, Canada)
St. Francis Hospital (SC)
St. Francis Memorial Hospital (CA)
St. John Hospital and Medical Center (MI)
St. John's Episcopal Hospital (NY)
St. John's Hospital & Health Center (CA)
St. John's Mercy Medical Center (MO)
St. John's Regional Health Center (MO)
St. Jude Children's Research Hospital (TN)
St. Luke's Hospital (IA)
St. Luke's Hospital (PA)
St. Mary Medical Center (CA)
St. Mary's Hospital (WI)
St. Michael's Medical Center, Inc. (NJ)
St. Tammany Parish Hospital (LA)
Stanford Hospital and Clinics (CA)
Stanton Territorial Health Authority (NT, Canada)
State of Connecticut Department of Public Health (CT)
State of Ohio/Corrections Medical Center Laboratory (OH)
State of Washington Public Health Labs (WA)
Stillwater Medical Center (OK)
Stony Brook University Hospital (NY)
Stormont-Vail Regional Medical Ctr. (KS)
Strong Memorial Hospital (NY)
Sudbury Regional Hospital (ON, Canada)
Sunnybrook Health Sciences Centre (ON, Canada)
Sunrise Hospital and Medical Center (NV)
Swedish Edmonds Hospital (WA)
Swedish Medical Center (CO)
Sydney South West Pathology Service Liverpool Hospital (NSW, Australia)
T.J. Samson Community Hospital (KY)
Taichung Veterans General Hospital (Taiwan)
Taiwan Society of Laboratory Medicine (Taiwan)
Tallaght Hospital (Ireland)
Tartu University Clinics (Estonia)
Temple Univ. Hospital - Parkinson Pav. (PA)
Tenet Healthcare (PA)
Texas Children's Hospital (TX)
Texas Department of State Health Services (TX)
Texas Health Presbyterian Hospital Dallas (TX)
The Brooklyn Hospital Center (NY)
The Charlotte Hungerford Hospital (CT)
The Children's Mercy Hospital (MO)
The Cooley Dickinson Hospital, Inc. (MA)
The Credit Valley Hospital (ON, Canada)
The Hospital for Sick Children (ON, Canada)
The Medical Center of Aurora (CO)
The Michener Inst. for Applied Health Sciences (ON, Canada)
The Naval Hospital of Jacksonville (FL)
The Nebraska Medical Center (NE)
The Ottawa Hospital (ON, Canada)
The Permanente Medical Group (CA)
The Toledo Hospital (OH)
The University of Texas Medical Branch (TX)
Thomas Jefferson University Hospital, Inc. (PA)
Timmins and District Hospital (ON, Canada)
Tokyo Metro. Res. Lab of Public Health (Japan)
Touro Infirmary (LA)
TriCore Reference Laboratories (NM)
Trident Medical Center (SC)
Trinity Medical Center (AL)
Tripler Army Medical Center (HI)
Tuen Mun Hospital, Hospital Authority (China)
Tufts Medical Center Hospital (MA)
Tulane Medical Center Hospital & Clinic (LA)
Turku University Central Hospital (Finland)
Twin Lakes Regional Medical Center (KY)
UCI Medical Center (CA)
UCLA Medical Center Clinical Laboratories (CA)
UCSD Medical Center (CA)
UCSF Medical Center China Basin (CA)
UMC of El Paso- Laboratory (TX)
UMC of Southern Nevada (NV)
UNC Hospitals (NC)
Unidad de Patología Clínica (Mexico)
Union Clinical Laboratory (Taiwan)
United Christian Hospital (Kowloon, Hong Kong)
United States Air Force School of Aerospace Medicine / PHE (TX)
Unity HealthCare (IA)
Universitair Ziekenhuis Antwerpen (Belgium)
University College Hospital (Ireland)
University Hospital (GA)
University Hospital Center Sherbrooke (CHUS) (Quebec, Canada)
University Medical Center at Princeton (NJ)
University of Alabama Hospital Lab (AL)
University of Chicago Hospitals Laboratories (IL)
University of Colorado Health Sciences Center (CO)
University of Colorado Hospital (CO)
University of Illinois Medical Center (IL)
University of Iowa Hospitals and Clinics (IA)
University of Kentucky Medical Center (KY)
University of Maryland Medical System (MD)
University of Minnesota Medical Center-Fairview (MN)
University of Missouri Hospital (MO)
University of Pennsylvania Health System (PA)
University of Pittsburgh Medical Center (PA)
University of Texas Health Center (TX)
University of the Ryukyus (Japan)
University of Virginia Medical Center (VA)
UPMC Bedford Memorial (PA)
US Naval Hospital Naples ( )
UZ-KUL Medical Center (Belgium)
VA (Asheville) Medical Center (NC)
VA (Bay Pines) Medical Center (FL)
VA (Central Texas) Veterans Health Care System (TX)
VA (Chillicothe) Medical Center (OH)
VA (Cincinnati) Medical Center (OH)
VA (Dallas) Medical Center (TX)
VA (Dayton) Medical Center (OH)
VA (Indianapolis) Medical Center (IN)
VA (Iowa City) Medical Center (IA)
VA (Miami) Medical Center (FL)
VA (San Diego) Medical Center (CA)
VA (Tampa) Hospital (FL)
VA (Wilmington) Medical Center (DE)
Valley Health / Winchester Medical Center (VA)
Vancouver Island Health Authority (SI) (BC, Canada)
Vanderbilt University Medical Center (TN)
Verinata Health, Inc. (CA)
Via Christi Regional Medical Center (KS)
Viracor-IBT Reference Laboratory (MO)
Virginia Regional Medical Center (MN)
Virtua - West Jersey Hospital (NJ)
WakeMed (NC)
Walter Reed Army Medical Center (DC)
Warren Hospital (NJ)
Washington Hospital Center (DC)
Washington Hospital Healthcare System (CA)
Waterbury Hospital (CT)
Waterford Regional Hospital (Ireland)
Wayne Memorial Hospital (NC)
Weirton Medical Center (WV)
West China Second University Hospital, Sichuan University (China)
West Jefferson Medical Center (LA)
West Penn Allegheny Health System-Allegheny General Hospital (PA)
West Shore Medical Center (MI)
West Valley Medical Center Laboratory (ID)
Westchester Medical Center (NY)
Western Baptist Hospital (KY)
Western Healthcare Corporation (NL, Canada)
Wheaton Franciscan Laboratories (WI)
Wheeling Hospital (WV)
White Memorial Medical Center (CA)
Whitehorse General Hospital (YT, Canada)
William Beaumont Army Medical Center (TX)
William Beaumont Hospital (MI)
William Osler Health Centre (ON, Canada)
Winchester Hospital (MA)
Winn Army Community Hospital (GA)
Wishard Health Sciences (IN)
Womack Army Medical Center Department of Pathology (NC)
York Hospital (PA)
Yukon-Kuskokwim Delta Regional Hospital (AK)

CLINICAL AND LABORATORY STANDARDS INSTITUTE®

Explore the Latest Offerings from CLSI!

As we continue to set the global standard for quality in laboratory testing, we're adding initiatives to bring even more value to our members and customers.

THE Key TO Quality™
from the Clinical and Laboratory Standards Institute (CLSI)

Power Forward with this Official Interactive Guide

Fundamentals for implementing a quality management system in the clinical laboratory.

Find Membership Opportunities

See the options that make it even easier for your organization to take full advantage of CLSI benefits and our unique membership value.

CLSI U
Laboratory Standards Education Center

Visit the CLSI U Education Center

Where we provide the convenient and cost-effective education resources that laboratories need to put CLSI standards into practice, including webinars, workshops, and more.

Shop Our Online Products

Including eCLIPSE Ultimate Access™, CLSI's cloud-based, online portal that makes it easy to access our standards and guidelines—*anytime, anywhere.*

For more information, visit www.clsi.org today.

CLINICAL AND LABORATORY STANDARDS INSTITUTE®

950 West Valley Road, Suite 2500, Wayne, PA 19087 USA

P: 610.688.0100 **Toll Free (US):** 877.447.1888 **F:** 610.688.0700

E: customerservice@clsi.org **www.clsi.org**

PRINT ISBN 1-56238-777-4
ELECTRONIC ISBN 1-56238-778-2