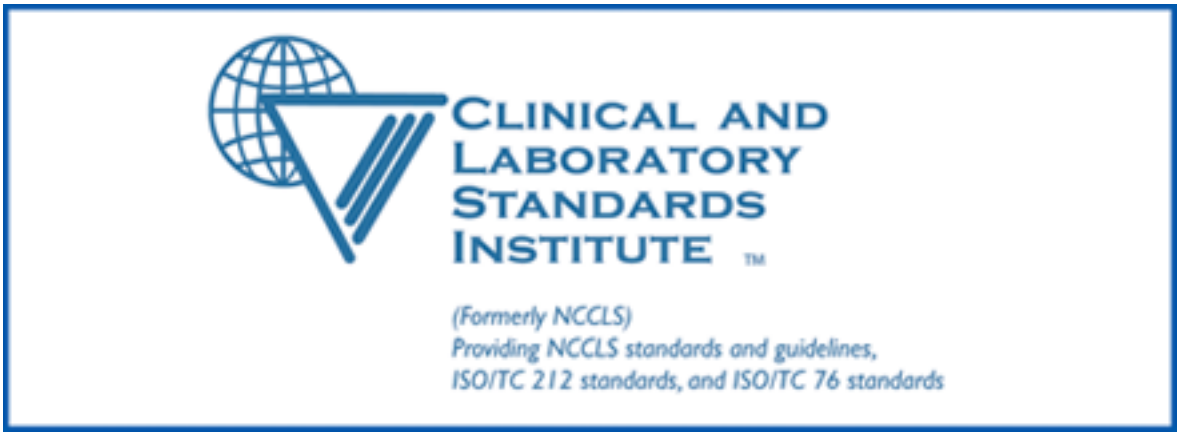
GP10-A
Vol. 15 No. 19
Replaces GP10-T
Vol. 13 No. 28

December 1995

Assessment of the Clinical Accuracy of Laboratory Tests Using Receiver Operating Characteristic (ROC) Plots; Approved Guideline



This document provides a protocol for evaluating the accuracy of a test to discriminate between two subclasses of subjects where there is some clinically relevant reason to separate them. In addition to the use of ROC plots, the importance of defining the question, selecting the sample group, and determining the "true" clinical state are emphasized.

NCCLS

GP10-A
THIS NCCLS DOCUMENT HAS BEEN
REAFFIRMED
WITHOUT CHANGE
AS AN APPROVED CONSENSUS DOCUMENT
EFFECTIVE MAY 2001

NCCLS...

Serving the World's Medical Science Community Through Voluntary Consensus

NCCLS is an international, interdisciplinary, nonprofit, standards-developing and educational organization that promotes the development and use of voluntary consensus standards and guidelines within the healthcare community. It is recognized worldwide for the application of its unique consensus process in the development of standards and guidelines for patient testing and related healthcare issues. NCCLS is based on the principle that consensus is an effective and cost-effective way to improve patient testing and healthcare services.

In addition to developing and promoting the use of voluntary consensus standards and guidelines, NCCLS provides an open and unbiased forum to address critical issues affecting the quality of patient testing and health care.

PUBLICATIONS

An NCCLS document is published as a standard, guideline, or committee report.

Standard A document developed through the consensus process that clearly identifies specific, essential requirements for materials, methods, or practices for use in an unmodified form. A standard may, in addition, contain discretionary elements, which are clearly identified.

Guideline A document developed through the consensus process describing criteria for a general operating practice, procedure, or material for voluntary use. A guideline may be used as written or modified by the user to fit specific needs.

Report A document that has not been subjected to consensus review and is released by the Board of Directors.

CONSENSUS PROCESS

The NCCLS voluntary consensus process is a protocol establishing formal criteria for:

- The authorization of a project
- The development and open review of documents
- The revision of documents in response to comments by users
- The acceptance of a document as a consensus standard or guideline.

Most NCCLS documents are subject to two levels of consensus–"proposed" and "approved." Depending on the need for field evaluation or data collection, documents may also be made available for review at an intermediate (i.e., "tentative") consensus level.

Proposed An NCCLS consensus document undergoes the first stage of review by the healthcare community as a proposed standard or guideline. The document should receive a wide and thorough technical review, including an overall review of its scope, approach, and utility, and a line-by-line review of its technical and editorial content.

Tentative A tentative standard or guideline is made available for review and comment only when a recommended method has a well-defined need for a field evaluation or when a recommended protocol requires that specific data be collected. It should be reviewed to ensure its utility.

Approved An approved standard or guideline has achieved consensus within the healthcare community. It should be reviewed to assess the utility of the final document, to ensure attainment of consensus (i.e., that comments on earlier versions have been satisfactorily addressed), and to identify the need for additional consensus documents.

NCCLS standards and guidelines represent a consensus opinion on good practices and reflect the substantial agreement by materially affected, competent, and interested parties obtained by following NCCLS's established consensus procedures. Provisions in NCCLS standards and guidelines may be more or less stringent than applicable regulations. Consequently, conformance to this voluntary consensus document does not relieve the user of responsibility for compliance with applicable regulations.

COMMENTS

The comments of users are essential to the consensus process. Anyone may submit a comment, and all comments are addressed, according to the consensus process, by the NCCLS committee that wrote the document. All comments, including those that result in a change to the document when published at the next consensus level and those that do not result in a change, are responded to by the committee in an appendix to the document. Readers are strongly encouraged to comment in any form and at any time on any NCCLS document. Address comments to the NCCLS Executive Offices, 940 West Valley Road, Suite 1400, Wayne, PA 19087, USA.

VOLUNTEER PARTICIPATION

Healthcare professionals in all specialities are urged to volunteer for participation in NCCLS projects. Please contact the NCCLS Executive Offices for additional information on committee participation.

Assessment of the Clinical Accuracy of Laboratory Tests Using Receiver Operating Characteristic (ROC) Plots; Approved Guideline

Abstract

Assessment of the Clinical Accuracy of Laboratory Tests Using Receiver Operating Characteristic (ROC) Plots; Approved Guideline (NCCLS document GP10-A) provides guidance for laboratorians 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 plots, the importance of defining the question, selecting a sample group, and determining the "true" clinical state are emphasized. The statistical data generated can be useful whether one is considering replacing an existing test, adding a new test, or eliminating a current test.

[NCCLS. Assessment of the Clinical Accuracy of Laboratory Tests Using Receiver Operating Characteristic (ROC) Plots; Approved Guideline. NCCLS Document GP10-A (ISBN 1-56238-285-3). NCCLS, 940 West Valley Road, Suite 1400, Wayne, Pennsylvania 19087, 1995.]

GP10-A
ISBN 1-56238-285-3
December 1995
ISSN 0273-3099

Assessment of the Clinical Accuracy of Laboratory Tests Using Receiver Operating Characteristics (ROC) Plots; Approved Guideline

Volume 15 Number 19

Mark H. Zweig, M.D.
Edward R. Ashwood, M.D.
Robert S. Galen, M.D., M.P.H.
Ronley H. Plous, M.D., FCAP
Max Robinowitz, M.D.

NCCLS

Suggested Citation

NCCLS. Assessment of the Clinical Accuracy of Laboratory Tests Using Receiver Operating Characteristic (ROC) Plots; Approved Guideline. NCCLS Document GP10-A (ISBN 1-56238-285-3). NCCLS, 940 West Valley Road, Suite 1400, Wayne, Pennsylvania 19087, USA.

Proposed Guideline

March 1987

Tentative Guideline

December 1993

Approved Guideline

Approved by Board of Directors
August 1995

Approved by Membership
November 1995

Published
December 1995

ISBN 1-56238-285-3
ISSN 0273-3099

Contents

Committee Membership

Area Committee on General Laboratory Practices

Gerald A. Hoeltge, M.D. **Chairholder**	**The Cleveland Clinic Foundation** **Cleveland, Ohio**
Donald A. Dynek, M.D. **Vice Chairholder**	**Pathology Medical Services, P.C.** **Lincoln, Nebraska**

Subcommittee on Clinical Evaluation of Tests

Mark H. Zweig, M.D. **Chairholder**	**National Institutes of Health** **Bethesda, Maryland**
Edward R. Ashwood, M.D.	University of Utah School of Medicine Salt Lake City, Utah
Robert S. Galen, M.D., M.P.H.	Case Western Reserve University Cleveland, Ohio
Ronley H. Plous, M.D., FCAP	LabOne, Inc. Shawnee Mission, Kansas
Max Robinowitz, M.D.	FDA Center for Devices and Radiological Health Rockville, Maryland

Advisors

George S. Cembrowski, M.D., Ph.D.	Park Nicollet Medical Center St. Louis Park, Minnesota
William Lee Collinsworth, Ph.D.	Boehringer Mannheim Diagnostics, Inc. Indianapolis, Indiana
William C. Dierksheide, Ph.D.	FDA Center for Devices and Radiological Health Rockville, Maryland
Jerome A. Donlon, M.D., Ph.D.	FDA Center for Biologics Evaluation and Research Rockville, Maryland
Marlene E. Haffner, M.D.	Food and Drug Administration Rockville, Maryland
Marianne C. Watters, M.T.(ASCP) *Board Liaison*	Parkland Memorial Hospital Dallas, Texas
Denise M. Lynch, M.T.(ASCP), M.S. *Staff Liaison*	NCCLS Wayne, Pennsylvania

Foreword

As laboratorians, we are often interested in how well a test performs clinically. This is true whether we are considering replacing an existing test with a newer one, adding a new test to our laboratory's menu, eliminating tests where possible, or just because we want to know something about the value of what we are doing. This project was originally intended to make recommendations about assessing the clinical performance of diagnostic tests. We elected to adopt the concepts of Swets and Pickett,[1] whereby clinical performance is divided into (1) a discrimination or diagnostic accuracy element 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, whether it is a clinical chemistry test, an electroencephalogram, an electrocardiogram, a nuclide scan, an x-ray, a biopsy, a view through an orifice, a pulmonary function test, or a sonogram, is meant to make some definable discrimination. It is important to know just how inherently accurate each tool (test) is as a diagnostic discriminator. *Note that assessing clinical accuracy, without engaging in comprehensive clinical decision analysis, is a valid and useful activity for the clinical laboratory*. Clinical 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, between patients responding to therapy and those not responding, or predicting who will get sick versus who will not. This ability to distinguish or discriminate between two states among patients who could be in either of the two states 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 (i.e., clinical accuracy) is the most basic property of a laboratory test as a device to help in decision making. This property is the place to start when assessing what value a test has in contributing to the patient-management process. If the test cannot provide the relevant distinction, it will not be valuable for patient care. On the other hand, once we establish that a test does discriminate well, then we can explore its role in the process of patient management to determine the practical usefulness of the information in a management strategy. This exploration is clinical decision analysis, and measures of test accuracy provide part of the data used to carry out that analysis.

Usefulness or efficacy refers to the practical value of the information in managing patients. A test can have considerable ability to discriminate, yet not be of practical value for patient care. This could happen for several reasons. For instance, the cost or undesirability of false results can be so high that there is no decision threshold for the test where the trade-off between sensitivity and specificity is acceptable. Perhaps there are less invasive or less expensive means to obtain comparable information. The test may be so expensive or technically demanding that its availability is limited. It could be so uncomfortable or invasive that the subjects do not want to submit to it.

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 or device; rather they are properties of the circumstances of the clinical application. These include the probability of disease (prevalence), the possible outcomes and the relative values of those 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 are characteristics or properties of the context in which test information is used, but they are not properties of the tests themselves. These factors interact with test

Foreword (Continued)

results to affect the usefulness of the test. Thus, it is helpful to conceptually separate the characteristic that is fundamental and inherent to the tests themselves, discrimination ability, from the interaction that results when this discrimination ability is mixed with external factors in the course of patient management.

In summary, we define clinical accuracy as the basic ability to discriminate between two subclasses of subjects where there is some clinically relevant reason to separate them. This concept of clinical accuracy refers to the quality of the information (classification) provided by the test and it should be distinguished from the practical usefulness of the information.[1] Both are aspects of test performance. Second, we suggest that the assessment of clinical 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 go any further in exploring a possible clinical role. If, on the other hand, a test does exhibit 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, we can decide whether to delve into a more complex assessment of its role in patient-care management (decision analysis). This document addresses the assessment of diagnostic accuracy but not the analysis of usefulness, or the role of the test in patient-care strategy.

The subcommittee believes that 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 other devices for performing tests who are interested in assessing or validating test performance in terms of clinical accuracy
- Regulatory agencies interested in establishing requirements for claims related to diagnostic accuracy
- Clinical laboratories that are reviewing data, literature, and/or generating their own data to make decisions about which tests to employ in their laboratory
- Health care/scientific workers interested in critical evaluation of data being presented on clinical test performance.

Key Words

Clinical accuracy, sensitivity, specificity, true-positive fraction, false-positive fraction, false-negative fraction, receiver operating characteristic (ROC) plot, performance evaluation, medical decision analysis, true-negative fraction.

Acknowledgment

The subcommittee thanks Dr. Gregory Campbell (Director, Division of Biostatistics, Office of Surveillance and Biometrics, Center for Devices/Radiological Health, Food and Drug Administration, Rockville, MD) for his invaluable expert statistical consultation on this document.

Assessment of the Clinical Accuracy of Laboratory Tests Using Receiver Operating Characteristic (ROC) Plots; Approved Guideline

1 Scope

This guideline outlines the steps and principles for designing a prospective study to evaluate the intrinsic diagnostic accuracy of a clinical laboratory test, i.e., its fundamental ability to discriminate correctly among alternative states of health expressed in terms of sensitivity and specificity. Each of the steps is discussed in detail, along with its rationale and suggestions for its execution. These same concepts can be used in critical evaluations of data already generated.

2 Glossary

Clinical accuracy (diagnostic accuracy): The ability of a diagnostic test to discriminate between two or more clinical states, for example, discrimination between rheumatoid arthritis and systemic lupus erythematosus, between rheumatoid arthritis and "no joint disease," between chronic hepatitis and "no liver disease," and between rheumatoid arthritis and a "mixture" of other joint diseases.

Clinical state: A state of health or disease that has been defined either by a clinical definition or some other independent reference standard. 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 threshold (also decision level, cutoff): A test score used as the criterion for a "positive test." All test scores at or beyond this test score are considered to be "positive"; those not at or beyond the score are considered to be "negative." In some cases, a low test score is considered to be "abnormal," e.g., L/S ratio or hemoglobin. In other cases, a high test score is considered to be "abnormal," e.g., cardiac enzyme or uric acid concentration.

Diagnostic test: A measurement or examination used to classify patients into a particular class or clinical state.

Efficacy: Actual practical value of the data, i.e., usefulness for clinical purposes.

False-negative result (FN): Negative test result in a subject in whom the disease or condition is present.

False-positive result (FP): Positive test result in a subject in whom the disease or condition is absent.

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 + TP); same as (1-sensitivity).

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 + TN); same as (1-specificity).

Prevalence: The pretest probability of a particular clinical state in a specified population; the frequency of a disease in the population of interest at a given point in time.

Receiver operating characteristic (ROC) plot: A graphical description of test performance representing the relationship between the true-positive fraction (sensitivity) and the false-positive fraction (1-specificity). Customarily, the true-positive fraction is plotted on the vertical axis and the false-positive rate (or, alternatively, the true-negative fraction) is plotted on the horizontal axis. Clinical accuracy, in terms of sensitivity and specificity, is displayed for the entire spectrum of decision levels.

Sensitivity (clinical sensitivity): Test positivity in disease; true positive fraction; ability of a test to correctly identify disease at a particular decision threshold.

Specificity (clinical specificity): Test negativity in health; true-negative fraction; ability of a test to correctly identify the absence of disease at a particular decision threshold.

Study group: A group of persons representing a sample of a clinically defined population of interest. The population of interest is the target group to which the test being evaluated will be applied in practice. Subgroups of the study group will be designated as belonging to particular clinical states by applying the standard criteria (see text).

True-negative result (TN): Negative test result in a subject in whom the disease is absent.

True-positive result (TP): Positive test result in a subject in whom the disease is present.

True-negative fraction (TNF): Ratio of subjects who do not have the disease and have a negative test to all subjects who do not have the disease; TN/(TN + FP); specificity.

True-positive fraction (TPF): Ratio of subjects who have the disease and a positive test to all subjects who have the disease; TP/(TP + FN); sensitivity.

3 Outline of the Evaluation Procedure

3.1 Define the Clinical Question (See Section 4.1)

Use the following procedure to define the clinical question:

(1) Characterize the subject population.

(2) State the management decision to be made.

(3) Identify the role of the test in making the decision.

3.2 Select a Representative Study Sample (See Section 4.2)

Use the following procedure to select a representative study sample:

(1) Select, prospectively, a statistically valid sample that consists of subjects who are representative of the population identified in Section 3.1 above.

(2) Select the sample independent of test results.

(3) Account for patients for whom data are incomplete.

3.3 Establish the "True" Clinical State of Each Subject (See Section 4.3)

Use the following procedure to establish the true clinical state of each subject:

(1) Adopt independent external standards or criteria of diagnostic truth for each relevant clinical state so as to classify each subject as accurately as possible. This may be based on a rigorous diagnostic workup or, alternatively, an assessment of clinical course or outcome.

(2) Classify subjects independent of the test being evaluated, i.e., without knowing the test results and without including the test results in the criteria.

3.4 Test the Study Subjects (See Section 4.4)

Use the following procedure to test the study subjects:

(1) Perform the test without knowing the clinical classification of the subjects.

(2) When comparing multiple tests, perform all tests on all subjects, preferably in a batch mode, and at the same point in their clinical course.

3.5 Assess the Clinical Accuracy of the Test (See Section 4.5)

Use the following procedure to assess the clinical accuracy of the test:

(1) Construct and analyze receiver operating characteristic (ROC) plots to evaluate test accuracy.

(2) Compare alternative tests on the basis of their ROC plots and analysis.

4 Designing the Basic Evaluation Study

4.1 Define the Clinical Question

Laboratory tests are requested to provide information that can be helpful in managing patients. There is always a relevant clinical question. Defining the clinical question is fundamental, then, because it establishes the particular patient-care issue being addressed by the evaluation. Can CK-2 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? Which, among several tests, is the best to use in discriminating between those subjects with breast cancer who will respond to a particular chemotherapy and those who will not? Which, among several tests, is most accurate in distinguishing between iron deficiency and other causes of anemia in elderly patients who present with previously undiscovered anemia?

A given test can perform differently in different clinical settings. A test can perform well in helping to discriminate between young, apparently healthy men with no prostatic disease and middle-aged men with prostatic cancer, but it might not do so well in helping to discriminate between middle-aged men with benign prostatic disease and middle-aged men with malignant prostatic disease. The latter distinction addresses a relevant clinical question applied to symptomatic middle-aged men, whereas the former distinction addresses a different issue that might not be clinically relevant at all.

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 patients with hypercalcemia, which ones have sarcoidosis?" The apparently similar patients share the common characteristic of hypercalcemia. The test helps in the attempt 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.

For the previously mentioned cases, the target population must be defined carefully, 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 (for example, chest x-ray) that are required for including and excluding subjects from the population.

4.2 Select a Representative Study Sample

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

The meaningfulness of the results depends on the care with which the relevant population is identified and sampled. The conclusions that can be drawn follow from the definition of the question and the nature of the subjects selected for study.

It is commonplace in routine laboratory practice to adopt or establish reference intervals, which are usually available with patient results to aid in their interpretation. These intervals are frequently derived from test-result data gathered from blood donors, laboratory workers, students, or other ambulatory, "healthy" volunteers. *Note that such groups might not be relevant for the evaluations of diagnostic accuracy described in this guideline*. When the accuracy of a test as a screening tool is being assessed, then a sample representative of the population to be screened should be used. Consider, for example, fecal occult blood testing for colon cancer. If the goal is to evaluate the accuracy of the test in discovering occult 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,

healthy volunteers and a group already known to have carcinoma of the colon is not appropriate.

The same principles apply when a test is being used, not for screening, but for differentiating between disease states in symptomatic patients. If a test is to be used to identify acute pancreatitis in patients 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 healthy volunteers and patients with well-defined pancreatitis, a study sample composed of such subjects is not appropriate. Conclusions based on such a sample would not serve the purpose of the study.

4.2.1 Selection Bias

To avoid selection biases that could compromise the study's validity or relevance to the question being posed, choose subjects carefully. Using only patients with well-established or clinically apparent disease, for example, can exclude the more typical patients, especially those with occult or early disease. Likewise, using young healthy volunteers can be inappropriate to the presumptive application of the test. The measures of accuracy used here are influenced by the spectrum of disease in the target population and, therefore, in the sample. The importance of the proper spectrum of subjects is discussed in detail in the literature.[2-6]

4.2.2 Retrospective Study

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

4.2.3 Selection Before Testing

Choosing subjects before testing begins acts as a precaution against the biases introduced when the test result directly or indirectly influences the selection of subjects. To avoid any biases, include in the test all patients who meet the definition of the clinical group of interest until a predetermined number of subjects is obtained. Once chosen, subjects should not be dropped from the study. If some patients do not complete the study (because of technical errors, analytical interferences, death, or loss to follow-up), they should be accounted for in the final analysis of the data. The uncertainty and possible biases that the lost subjects cause in the study's conclusions must be considered and reported.

4.2.4 Prevalence of Disease

The approach described here is independent of prevalence of disease, so it is not necessary to have a sample that reflects actual prevalence. It is desirable to have approximately equal numbers of subjects who are truly affected and truly unaffected by the disease.

4.2.5 Consult a Statistician

Consultation with a professional statistician is recommended when planning the definition, size, and selection of study populations that will be used for critical evaluation of test performance. The sample size should be appropriate to the goals of the evaluation and provide valid estimates of ROC plots and comparisons among tests. When this is not possible, the criteria for selection should be clearly described.

4.3 Establish the "True" Clinical State of Each Subject

An objective assessment of clinical accuracy requires comparing the results provided by the test with some independent, external definition of truth. The clinical question, defined above, establishes what the categories of "truth" (states of health) are, relevant to the evaluation. Criteria or standards are applied to place individual persons in their respective categories of truth. The standards may include biopsy data, surgical or autopsy findings, imaging data, and long-term follow-up. Unfortunately, classifying individual persons into distinct categories can be an imperfect operation. The standards can be unreliable and/or can produce bias.[6] Some of them might not fit clearly into one of the defined states of health. Metz suggests that "truth is ultimately a philosophical concept, of course, and standards of truth are adequate for practical purposes if they are

substantially more reliable than the diagnostic system [test] undergoing evaluation."[6(p. 723)]

4.3.1 Validity of Evaluation

When evaluating the clinical 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.

4.3.2 True Clinical Subgroup

Routine clinical diagnoses are likely to be inadequate for evaluation studies. Determining a patient's true clinical subgroup can require such procedures 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.

4.3.3 Approaches to Classification

In many clinical situations, obtaining an independent, accurate classification of the patient's true clinical condition is difficult. Several strategies have been developed to deal with the difficulties in identifying true states of health. One strategy is to define the diagnostic problem in terms of measurable clinical outcomes.[7] A second approach is to employ some sort of consensus, majority rule, or expert review to arrive at a less error-prone identification process.[8] A third solution is to assume for the comparison of several accurate tests that there is some unknown mixture of diseased and nondiseased persons in the subject population and then to estimate this mixture parameter, as well as the other parameters.[9] A fourth approach is to, rather than definitively assign each such patient to one of the groups, say, "diseased" or "nondiseased," assign to each a value between 0 and 1 that corresponds to the (subjective) assessment of how likely it is that this patient belongs to the diseased group (this could be accomplished by logistic regression). Then there is no need to discard the data from these gray, fuzzy cases where group assignment is equivocal.[10-12, 13]

Although diagnostic categories often do predict complications and therapeutic responses, the best evaluation of a test can be in terms of its ability to indicate clinical course or outcome, rather than its ability to assign a diagnosis. For example, it might be possible to classify patients with suspected prostatic disease into those who have cancer and those who do not have cancer based on biopsy results; however, it might be more useful to classify them in terms of which patients progress to overt disease. If the goal of the evaluation is to assess the accuracy of a serum marker in discriminating between those patients who need intervention and those who do not, then it is more relevant to know which patients will progress than to know which have histologic evidence of disease at that moment. This issue is actually one that is properly confronted earlier in formulating the original clinical management task to be addressed by the test under evaluation. Thus, lack of an immediate definitive diagnostic category does not necessarily prevent a valid assessment of the clinical accuracy of a test. In fact, even when the correct diagnosis can be easily established, a study correlating test results with the clinical course can provide a more usefulclinical evaluation than a study that merely correlates test results with patient diagnoses.

4.3.4 Independent Classification

To avoid bias in evaluating the clinical accuracy of a test, the true clinical state should also be determined independent of the test(s) under investigation or used for comparison. Obviously, the new test should not be included in the criteria used to classify the subjects. Neither should a closely related test be included in the criteria for classifying subjects. For example, if an RIA for CK-MB is being evaluated for the diagnosis of AMI, neither CK-MB by electrophoresis or by immuno-inhibition 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 LD-1/LD-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 "truth standard."

4.3.5 Masked Evaluation

To ensure that the classification is not influenced by the result of the test under evaluation, it should be done masked, 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 consensus of experts who each interpret the material masked and independent of each other.

4.4 Test the Study Subjects

4.4.1 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 done before the clinical question is answered. Knowing the answer to the clinical question can introduce subtle biases. Results that do not fit the clinical status might be selectively repeated or rejected on the basis of supposed technical difficulties or interfering factors.

4.4.2 Identical Specimens

When comparing two or more tests, it is important that the subjects and specimens be identical for all tests. Failure to use the identical subjects for evaluating each test can result in misleading conclusions based on sampling errors. Furthermore, subtle biases can affect the selection of subjects for the different groups. Thus, apparent differences in test performance can simply be reflections of 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 ensures that differences in sensitivity and specificity are not simply due to inconsistent application of the diagnostic criteria.

Similarly, if two or more tests are applied to the same subject at different times during the course of his illness, an apparent superiority of one of the tests might simply reflect that it was done when the disease was more easily detected. Therefore, all tests should be performed at the same point in the course of each subject's illness. Using identical specimens for all tests obviates all of the above pitfalls.

4.4.3 Testing Mode

Assaying all samples in one batch, when possible, to minimize the influence of between-run analytical variance, is suggested. However, attention should be given to maintaining analyte stability through proper storage conditions.

4.5 Assess the Clinical Accuracy of the Test

Assessing the performance of a test by examining its clinical accuracy, that is, its ability to correctly classify individual persons into two subgroups, for example, a subgroup of persons affected by some disease (and therefore needing treatment) and a second subgroup of unaffected persons, is suggested. 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 power to discriminate (accuracy).

4.5.1 Diagnostic or Clinical Sensitivity and Specificity

The ability of a test to identify or recognize the presence of disease is its diagnostic sensitivity; its ability to recognize the absence of disease is its diagnostic specificity. Both are measures of

accuracy and can be expressed as percentages, rates, or decimal fractions. A perfect test achieves a sensitivity and specificity of 100% or 1.0. However, tests are rarely perfect, and, usually, they usually do not achieve a sensitivity and a specificity of 100% at the same time.

Diagnostic sensitivity (true-positive rate or fraction) is defined as follows:

$$\frac{\text{Number of True-Positive Results}}{\text{Number of True-Positive + Number of False-Negative Results}}$$

$$\text{or } \frac{TP}{TP + FN} \qquad (1)$$

This is the fraction of persons who are truly affected by a disease who have positive test results.

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

$$\frac{\text{Number of True-Negative Results}}{\text{Number of True-Negative + Number of False-Positive Results}}$$

$$\text{or } \frac{TN}{TN + FP} \qquad (2)$$

This is the fraction of persons who are truly unaffected by a disease who have negative test results.

Often, a test is said to have a particular sensitivity and specificity. However, there is not a single sensitivity or specificity for a test; rather there is a continuum of sensitivities and specificities. By varying the decision threshold (or decision level, upper-limit-of-normal, cut-off value, or reference value), any sensitivity from 0 to 100% can be obtained, and each one will have a corresponding specificity. For each decision threshold used to classify the subjects as "positive" or "negative" based on test results, there is a single combination of sensitivity and specificity. These parameters occur, then, in pairs, and the accuracy of a test is reflected in the spectrum of pairs that can occur (not all pairs being possible for a particular test). For any test in which the distributions of results from the two categories of subjects overlap, there are inevitable "trade-offs" between sensitivity and specificity. As the decision threshold is varied over the range of observed results, the sensitivity and specificity will move in opposite directions. As one increases, the other decreases. For each decision threshold, then, there is a corresponding sensitivity and specificity pair. Which one(s) describe(s) the accuracy of the test? All of them do. Only the entire spectrum of sensitivity/specificity pairs provides a complete picture of test accuracy.

In Figure 1 (p. 13), at a threshold of 6 μg/L, CK-BB exhibits a sensitivity of 100% or 1.0. All 50 subjects with acute mycocardial infarction (AMI) are correctly classified as "positive" or "affected." Likewise, at this same threshold, 9 of the 20 subjects without AMI are incorrectly classified as positive, so the specificity is only 55% (55% true negatives, 45% false positives). However, when the decision threshold is 12 μg/L instead of 6, the sensitivity decreases to 96% (0.96) because only 48 of the 50 subjects with AMI are correctly classified as "positive." Furthermore, because all non-AMI subjects are now correctly classified as unaffected, specifity has increased to 100% (100% true negatives, 0% false positives). Thus the shift in the threshold from 6 to 12μg/L results in a decrease in sensitivity and an increase in specificity. Note that sensitivity is calculated entirely from the affected (AMI) subjects, while specificity is calculated from the unaffected subgroup.

Furthermore, a test can have one set of sensitivity–specificity pairs in one clinical situation but a different set in another clinical situation with a different group of subjects. If CK-BB had been measured in postoperative patients suspected of having an AMI, instead of in emergency department patients (as in Figure 1 p. 13), the sensitivity–specificity pairs could be quite different. The spectrum of pairs contained in the test characterizes its basic accuracy for a particular clinical setting.

4.5.2 Receiver Operating Characteristic Plots

4.5.2.1 General

The spectrum of trade-offs between sensitivity and specificity is conveniently represented by the ROC plot.[14] ROC methodology is based on statistical decision theory and 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.[6] An ROC-type plot was used in the 1950s to

characterize the ability of an automated Pap smear analyzer to discriminate between smears with and without malignant cells.[15]

The ROC plot graphically displays this entire spectrum 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 clinical accuracy of the test by plotting all the sensitivity–specificity pairs resulting from continuously varying the decision threshold over the entire range of results observed. The important part of the plot is generated when the decision threshold is varying within the region where results from the affected and unaffected subjects overlap. Outside of the overlap region, either sensitivity or specificity is 1.0 and not varying; within the overlap region, neither is 1.0 and both are varying as the decision threshold varies. On the *Y* axis, sensitivity, or the true-positive fraction (TPF), is plotted. On the *X* axis, false-positive fraction (FPF) (or 1-specificity) is plotted. This is the fraction of truly unaffected subjects who nevertheless have positive test results; therefore, it is a measure of specificity.

Another option is to plot specificity directly (false-negative fraction) on the *X* axis. This results in a left-to-right "flip," giving a mirror-image of the plot described above. However, if the *X* axis is labeled from 0 to 1.0 from right to left (instead of left to right), then the plot is not flipped over.

As mentioned above for sensitivity and specificity, TP and FP fractions vary continuously with the decision threshold within the region of overlapping results. Each decision threshold has a corresponding pair of TP (sensitivity) and FP (1-specificity) fractions. The rates observed also depend on the clinical setting, as reflected by the study group chosen. The FP fraction is influenced by the type of unaffected subjects included in the study group. If, for example, the unaffected subjects are all healthy blood donors who are free of any signs or symptoms, a test can appear to have much lower FP fractions than if the unaffected subjects are persons who clinically resemble those who actually have the disease.

Likewise, the TP fraction also depends on the study group. A test used to detect cancer can have higher TP fractions when applied to patients with active or advanced disease than to patients with stable or limited disease. This dependence of TP and FP fractions on the study population is the reason that an ROC plot must be generated for each clinical situation.

In the ROC plot, the various combinations of sensitivity and specificity possible for the test in a given setting are readily apparent. Also apparent, then, are the "trade-offs" 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 appreciated 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 appears.

Because true- and false-positive fractions are calculated entirely separately, using the test results from two different subgroups of persons (affected, unaffected), the ROC plot 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 plot, are still influenced by the type (spectrum) of subjects included in the sample.

The ROC plot provides a general, global assessment of performance that is not provided when only one or a few sensitivity–specificity pairs are known. The test performance data obtained to derive ROC plots may also be used to select decision thresholds for particular clinical applications of the test. Several elements besides test performance determine which of the possible sensitivity–specificity pairs (and thus the corresponding decision threshold) is most appropriate for a given patient-care application: (a) the relative cost or undesirability of errors, i.e., false-positive and false-negative classifications (the benefits of correct classifications may also be considered); (b) the value (or "utility") of various outcomes (death, cure, prolongation of life, or change in the quality of life); and (c) the relative proportions of the two states of health that the test is intended to discriminate between (prevalence of the conditions or diseases). While the selection of a decision threshold is usually required for using a test for patient management, this important step is beyond the scope of this guideline. Discussion of this complex issue can be found elsewhere.[3,16-19]

4.5.2.2 Generating the ROC Plot; 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. 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 (ratings) data with rating categories such as "definitely abnormal," "probably abnormal," "equivocal," "probably normal," and "definitely normal."

A tie in laboratory data is of interest when a member of the diseased group has the same result as does a member of the nondiseased group. Such ties are more likely to occur when there are few data categories (i.e., few different results), such as with coarse discrete data (dipstick data, for example) rather than when the number of different results is large, as with continuous data. This results from grouping or "binning" the data into ordered categories. In clinical laboratories, when observations are made on a continuous scale, ties are much less likely (unless intentional grouping into "bins" has occurred). Theoretically, if measurements are exact enough, no two persons would have the same result on a continuous scale. However, the resolution of results in the clinical laboratory is often not so fine as to prevent this, and some ties will occur even with continuous data. Furthermore, intentional binning of continuous data also increases the chance for ties. This occurs when, for example, gonadotrophin results are expressed as whole numbers even though the assay provides concentrations to 0.1 of a unit. It also occurs when all results within intervals, such as 0–50, 51–100, etc., are grouped together. Ties can be caused, then, either by the intentional binning of data or by the degree of analytical resolution of the method of observation.

For both tied and untied data, one merely plots the calculated (1-specificity, sensitivity) points at all the possible decision thresholds (observed values) of the test. (This can be limited to the decision thresholds in the region of overlapping results; see Section 4.5.2.1.) It is the graph of these points that is the ROC plot. For data with no ties, adjacent points can be connected with horizontal and vertical lines in a unique manner to give a staircase figure (Figure 2, p. 14). As the threshold changes, inclusion of a true-positive result in the decision rule produces a vertical line; inclusion of a false-positive result produces a horizontal line. As the numbers of persons in the two groups increase, the steps in the staircase become smaller and the plot usually appears less jagged. Because this ROC plot uses all the information in the data directly through the ranks of the test results in the combined sample, it can also be called the nonparametric ROC plot. The term "nonparametric" here refers to the lack of parameters needed to model the behavior of the plot, in contrast to parametric approaches that rely on models with parameters to be estimated.

When there are ties in continuous data, both the true-positive and false-positive fractions change simultaneously, resulting in a point displaced both horizontally and vertically from the last point. Connecting such adjacent points produces diagonal (nonhorizontal and nonvertical) lines on the plot. Diagonal segments in the ROC plot, then, indicate ties.

As mentioned above, ties may be intentionally introduced in the display of the test results by grouping the results into intervals. A common approach often adopted in the literature is to plot the ROC at only a few points by using only a few decision thresholds and connecting adjacent points with straight line segments. All data falling in an interval between thresholds are treated as tied. Although this bin approach has the advantage of plotting ease, it discards much of the data and introduces many ties in the data. If the points are few and far between, this approximation can be poor and it can misrepresent the actual plot.

4.5.2.3 Qualitative Interpretation of the ROC Plot

A test with good clinical performance achieves high TPFs (sensitivity), while having low FPFs (corresponding to high specificity). Tests with high diagnostic accuracy, then, have ROC plots with points close to the upper left corner where TPFs are high and FPFs are low. A test with perfect accuracy, giving perfect discrimination between affected and unaffected groups, achieves a TPF of 1.0 (100% sensitivity) and an

FPF of 0.0 (100% specificity) at one or more decision thresholds. This ROC plot, then, goes through the point (0, 1.0) in the upper left corner. A simple rule of thumb is that the closer the plot is to this point, the more clinically accurate the test usually is. A test that does not discriminate between truly affected and truly unaffected subgroups has an ROC plot that runs at a 45° angle from the point (0,0) to (1.0, 1.0). Along this line, TPF equals FPF at all points, regardless of the decision threshold. (See "X" in Figure 2, p. 14.) All tests have plots between the 45° diagonal and the ideal upper left corner. The closer the plot is to the upper left corner, the higher the discriminating ability of the test. Visual inspection of the plot, then, provides a direct qualitative assessment of accuracy.

Figure 2 (p. 14) has an ROC plot for a test with modest accuracy. Here the plot is in an intermediate position between the 45° diagonal and the ideal upper left corner. Figure 3 (p. 15) has an ROC plot for a test with high accuracy. Note how closely the plot passes to the upper left corner where sensitivity is highest and the FPF (1-specificity) is lowest. Figure 4 (p. 16) shows ROC plots of results of three tests, all derived from the same sample of persons. This provides a convenient comparison of accuracies. The plot for amylase lies above and to the left of the plot for phospholipase A (PLA). Thus, at most sensitivities (TPF), amylase has a lower FPF (higher specificity) than PLA. Conversely, at most FPFs, amylase has a higher TPF (better sensitivity) than does PLA. Amylase and lipase have nearly identical ROC plots, indicating virtually the same ability to discriminate. Both appear to be more accurate than PLA.

4.5.2.4 Area Under a Single ROC Plot

One convenient way to quantify the diagnostic accuracy of a laboratory test is to express its performance by a single number. The most common measure is the area under the ROC plot. By convention, this area is always ≥ 0.5 (if it is not, one can reverse the decision rule to make it so). Values range between 1.0 (perfect separation of the test values of the two groups) and 0.5 (no apparent distributional difference between the two groups of test values). The area does not depend only on a particular portion of the plot, such as the point closest to the upper left corner or the sensitivity at some chosen specificity, but on the entire plot. This is a quantitative, descriptive expression of how close the ROC plot is to the perfect one (area = 1.0). The statistician readily recognizes the ROC area as the Mann–Whitney version of the nonparametric two-sample statistic[20,21] introduced by the chemist Frank Wilcoxon. An area of 0.8, for example, means that a randomly selected person from the diseased group has a laboratory test value larger than that for a randomly chosen person from the nondiseased group 80% of the time. It does *not* mean that a positive result occurs with probability 0.80 or that a positive result is associated with disease 80% of the time.

When there are no ties between the diseased and nondiseased groups, this area is easily computed from the plot as the sum of the rectangles under this graph. Analytical formulas to calculate the area appear in reports by Bamber[20] and Hanley and McNeil.[21] Alternatively, the area can be obtained indirectly from the Wilcoxon rank-sum statistic.[22]

Parametric approaches to calculating area, employing some model for fitting a curve, have also been described. Both parametric and nonparametric methods are discussed and compared in published reviews.[13,23]

In using global indices such as area under the ROC plot, there is a loss of information. Therefore, it is undesirable to consider area without visual examination of the ROC plot itself as well.

4.5.2.5 Statistical Comparison of Multiple Tests

Direct statistical comparison of multiple diagnostic tests is frequent in clinical laboratories. Usually, two (or more) tests are performed on the same subjects, as in a split-sample comparison.

Tests can be compared to one another at a single observed or theoretical sensitivity or specificity.[24–26] Alternatively, a portion of the ROC plot can be used to compare tests.[27]

A global approach is to compare entire ROC plots by using an overall measure, such as area under the plot; this can be performed either nonparametrically or parametrically.[13] This is especially attractive to laboratories because the comparison does not rely on the selection of a particular decision threshold (which should

consider prevalence and cost trade-off information). However, the user should always inspect the ROC plot visually when comparing tests, rather than relying on the area that condenses all the information into a single number.

4.5.2.6 Other ROC Statistics

Confidence intervals around a point or points on the ROC plot can be estimated both parametrically and nonparametrically for those who so desire such estimates.[13] When "true" diagnoses are not well known for the subjects being studied ("fuzzy" cases), the probability that a given patient belongs to a particular diagnostic category can be assigned, and a "fuzzy" ROC plot can be constructed.[13]

4.5.2.7 Advantages of ROC Plots[13]

The ROC plot has the following advantages: It is simple, graphical, and easily appreciated visually. It is a comprehensive representation of pure clinical accuracy, i.e., discriminating ability, over the entire range of the test. It does not require selection of a particular decision threshold because the whole spectrum of possible decision thresholds is included. It is independent of prevalence: No care need be taken to obtain samples with representative prevalence; in fact, it is usually preferable to have equal numbers of subjects with both conditions. It provides a direct visual comparison between tests on a common scale. It requires no grouping or binning of data, and both specificity and sensitivity are readily accessible.

4.5.2.8 Disadvantages of ROC Plots

The ROC plot has several drawbacks: Actual decision thresholds usually do not appear on the plot (though they are known and are used to generate the graph). The number of subjects is also not part of the plot. Without computer assistance, the generation of plots and analysis is cumbersome. (See the Appendix for available software packages).

5 The Use of ROC Plots: Examples From the Clinical Laboratory Literature

Van Steirteghem et al[28] compared the accuracies of myoglobin, CK-BB, CK-MB, and total CK in discriminating among persons with and without acute mycocardial infarction, who presented to an emergency department with typical chest pain. ROC plots could be constructed for any sampling time by using measurements on multiple, closely sequential serum samples timed from the onset of pain. Leung et al[29] performed a similarly detailed evaluation of total CK and CK-2 in 310 patients admitted to a cardiac care unit with chest pain. These authors also used ROC plots to describe the changing clinical accuracy at various time intervals after the onset of pain.

Several other authors also used ROC plots in various ways. Carson et al[30] investigated the abilities of four different assays of prostatic acid phosphatase to discriminate between those subjects with prostatic cancer and those subjects with either some other urologic abnormality or no known urologic abnormality. Hermann et al[31] compared the clinical accuracies of two versions of a commercial assay for thyrotropin to test a claim that the newer one was superior for discriminating between euthyroidism and hyperthyroidism. Kazmierczak et al[32] used ROC plots in a study of the clinical accuracies of lipase, amylase, and phospholipase A in discriminating acute pancreatitis from other diseases in 151 consecutive patients seen with abdominal pain. Flack et al[33] used ROC plots and areas to compare the abilities of urinary free cortisol and 17-hydroxysteroid suppression tests to discriminate between Cushing disease and other causes of Cushing syndrome. Guyatt et al[34] studied the ability of seven tests, including ferritin, transferrin, saturation, mean cell volume, and erythrocyte protoporphyrin, to discriminate between iron-deficiency anemia and other causes of anemia in subjects older than 65 years who were admitted to the hospital with anemia. Beck,[35] while studying iron-deficiency anemia, also used ROC plots to compare four tests.

6 Summary

The first step in designing a study to evaluate the clinical accuracy of a test is to establish the clinical goal clearly and explicitly. It is essential

to identify what issue of consequence to patient management is to be addressed by the test. The following guidelines are suggested for a clinical test evaluation or diagnostic trial.

- Carefully define the clinical question or goal.

- Choose study subjects who are representative of the clinical population to which the test is ultimately to be applied. Advance consultation with a statistician is recommended.

- Perform all tests being evaluated on the same specimens from the same subjects; perform all tests on individual subjects at the same point in their clinical course.

- Classify the subjects as either "affected" or "unaffected," or into other relevant management subgroups, by rigorous and complete means so that the true diagnoses or outcomes are approached closely. Diagnostic procedures that go beyond routine clinical practice or the use of extensive follow-up, may be required for the purpose of the evaluation. All classification criteria should be independent of the test or tests being studied.

- Evaluate and compare clinical accuracy in terms of sensitivity and specificity at all decision thresholds using ROC plots.

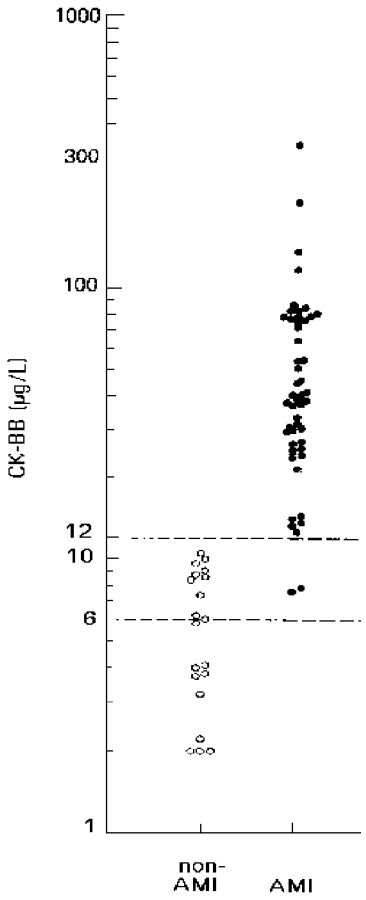


Figure 1. Dot diagram of serum CK-BB concentrations 16 hours after the onset of symptoms in 70 subjects presenting to an emergency department with typical chest pain. Fifty were eventually considered to have had acute mycocardial infarction (AMI); 20 were not. (Data from Van Steirteghem AC, Zweig MH, Robertson EA, et al. Comparison of the effectiveness of four clinical chemical assays in classifying patients with chest pain. *Clin Chem* 1982;28:1319–1324.)

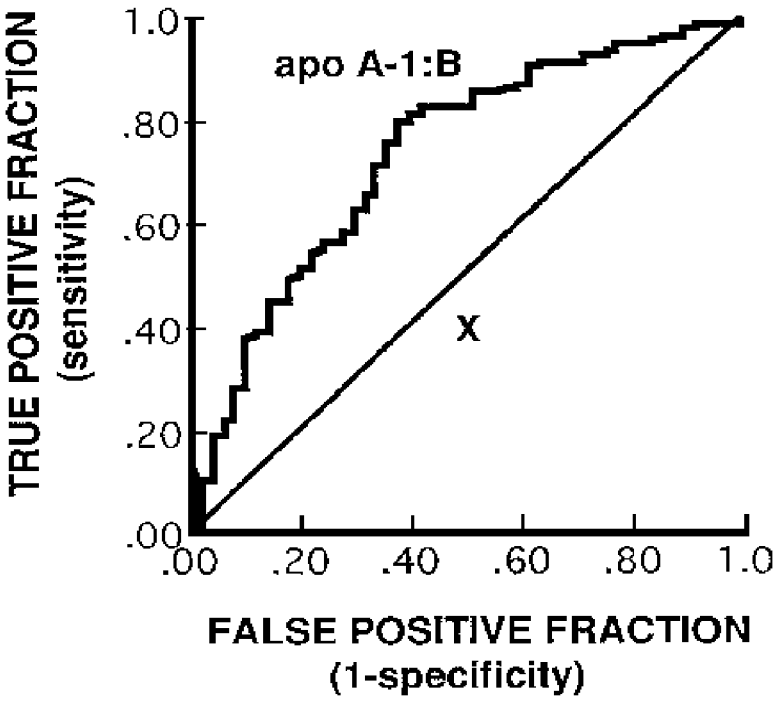


Figure 2. Nonparametric ROC plot of serum apolipoprotein A-I/B ratios used in identifying clinically significant coronary artery disease (CAD) in 304 men suspected of having CAD. Presence or absence of CAD was established by coronary angiography. Area under the ROC plot is 0.75. The line labeled "X" represents the theoretical plot of a test with no ability to discriminate (area = 0.5). (From Zweig MH. Apolipoproteins and lipids in coronary artery disease: Analysis of diagnostic accuracy using receiver operating characteristic plots and areas. *Arch Pathol Lab Med* 1994; 118:141–144.)

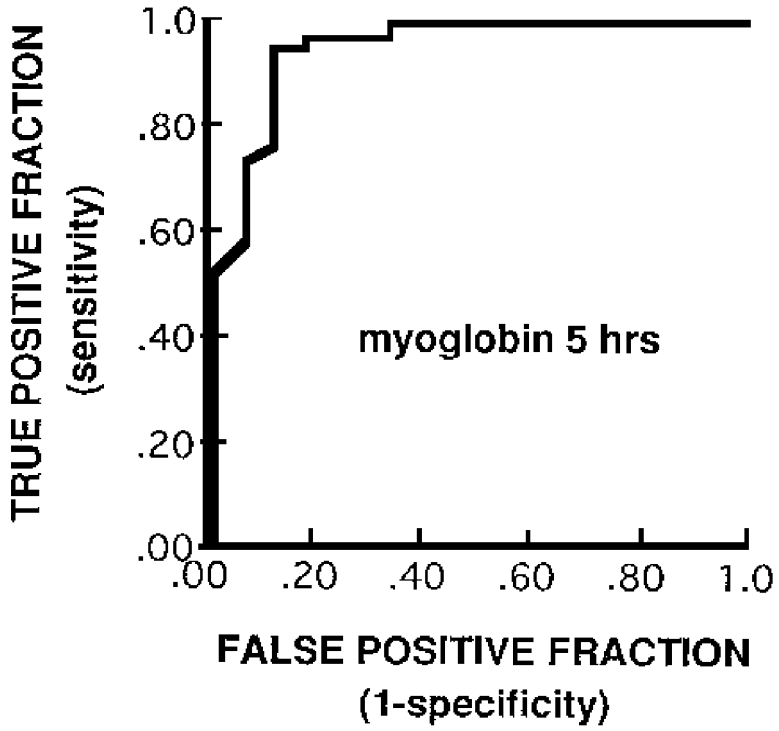


Figure 3. Nonparametric ROC plot of serum myoglobin concentrations, 5 hours after the onset of symptoms in 55 emergency department patients suspected of having acute myocardial infarction. The area under the plot is 0.953. Thirty-seven subjects had an AMI; 18 did not. (Data from Van Steirteghem AC, Zweig MH, Robertson EA, et al. Comparison of the effectiveness of four clinical chemical assays in classifying patients with chest pain. *Clin Chem* 1982;28:1319–1324.)

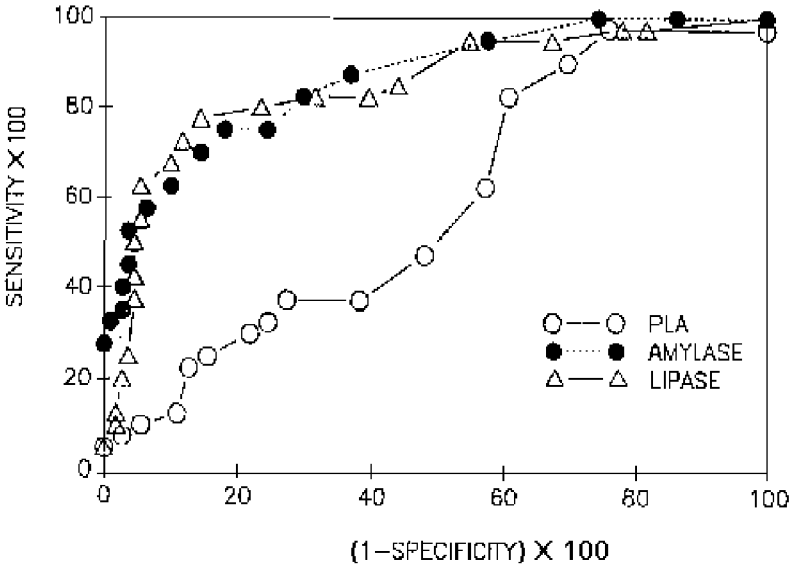


Figure 4. ROC plots for peak serum amylase, lipase, and phospholipase A (PLA) concentrations in identifying acute pancreatitis in 151 consecutive patients with abdominal pain. (From Kazmierczak SC, Van Lente F, Hodges ED. Diagnostic and prognostic utility of phospholipase A activity in patients with acute pancreatitis: comparison with amylase and lipase. *Clin Chem* 1991;37:356–360.)

Appendix: Computer Software for ROC Plotting and Analysis

Commercial, public domain, and shareware is available to calculate sensitivities and specificities; generate ROC plots; calculate areas under the plots; to generate other statistics, such as standard deviation and confidence intervals, and analyses such as comparing the areas for multiple tests. This software has not been evaluated by NCCLS. The list was generated as a starting point for users who might wish to evaluate and purchase a package.

Some programs were developed primarily to deal with discrete or ratings-type data typically used by radiologists, for example, where the number of different results is small. Laboratories almost always generate continuous data with a virtually infinite number of possible results. Some programs are designed to use all the raw continuous data without grouping or compressing it into fewer categories or bins. Whether the data is continuous (not grouped or "binned") or originally discrete (or made discrete by binning), the ROC plot can be generated parametrically or nonparametrically. The nonparametric approach does not use models to fit the curve but rather simply plots the calculated (l-specificity, sensitivity) points at all the possible observed values of the tests. The points are connected to produce the plot. Parametric approaches rely on models with parameters to be estimated in order to fit a curve to the data. These approaches, the underlying assumptions, and a discussion of the characteristics, advantages, and disadvantages for laboratory data are published.[13]

Several commercial and public domain software products for ROC plotting and analysis are listed below. Note that only three of the programs are designed to treat the continuous data directly, without binning (forcing into discrete intervals) the data (numbers 4, 5, 7). A comparison of features is published.[13]

CLINROC. Henry T. Sugiura and George A. Hermann, R. Phillip Custer Laboratories, Presbyterian University of Pennsylvania Medical Center, 39th & Market Streets, Philadelphia, PA 19104. CLINROC does not produce its parametric analysis of likelihood ratios through maximum likelihood methods but rather based on the assumption of normality in the original or log-transformed scale.

Metz programs: LABROC1, CLINROC, ROCFIT, CORROC. Charles E. Metz, Department of Radiology, MC2026, The University of Chicago Medical Center, 5841 South Maryland Avenue, Chicago, IL 60637-1470; [FAX (312) 702-6779; Internet address: c-metz[@] uchicago.edu]. The Metz programs are, for a single diagnostic test, LABROC1 for continuous data and ROCFIT for discrete data, and, for two correlated tests, CLABROC and CORROC, respectively. Program requesters are asked to specify the platform and to include, for microcomputer requests, two appropriate floppy disks. A version for the Macintosh is available.

ROC ANALYZER. Robert M. Centor, 10806 Stoneycreek Drive, Richmond, VA 23233. [E mail address: rcentor@gim.meb.vab.edu]. This program is described by Centor and Keightley.[37]

ROCLAB. James M. DeLeo, Bldg. 12A, Room 2013, Division of Computer Research and Technology, National Institutes of Health, Bethesda, MD 20892. [E mail address: deleoj@6100.dcrt.nih.gov]. ROCLAB provides maximal, as well as trapezoidal, areas for ties. It also has the ability to do ROC plots for fuzzy data.

RULEMAKER. Digital Medicine, Inc., Hanover, NH 03755; [(603) 643-3686]. Rulemaker, which will run on a Macintosh, is still being developed; a release version is anticipated in 1996.

SIGNAL. SYSTAT, Inc., 1800 Sherman Avenue, Evanston, IL 60201. SIGNAL is a module of a much larger commercial package SYSTAT.

TEP-UH (*Test Evaluation Program -University Hospital*). Thomas G. Pellar, Department of Clinical Biochemistry, University Hospital, P.O. Box 5339, 339 Windemere Road, London, Ontario, Canada N6A 5A5. Running TEP-UH requires the parent program MUMPS (Micronetics Design Corp., Rockville, MD).[38]

Appendix (Continued)

Two other packages are available:

TESTIMATE, idv-Data Analysis and Study Planning, Wessobrunner Str. 6, 82131 Gauting/ Munich, Germany. Fax: +49.89.8503666.

SmarTest, idv-Data Analysis and Study Planning, Wessobrunner Str. 6, 82131 Gauting/ Munich, Germany. Fax: +49.89.8503666.

Product/Vendor List in NCCLS Standards

This list includes products known to NCCLS at the time this guideline was published, but it is not all inclusive. NCCLS has not evaluated the listed products. Inclusion of products and/or vendors on the list does not constitute endorsement by NCCLS.

References

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

2. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Eng J Med* 1978;299:926-930.

3. Robertson EA, Zweig MH, Van Steirteghem AC. Evaluating the clinical efficiency of laboratory tests. *Am J Clin Pathol* 1983;79:78-86.

4. Zweig MH, Robertson EA. Why we need better test evaluations. *Clin Chem* 1982;28:1272-1276.

5. Lachs MS, Nachamkin I, Edelstein PH, et al. Spectrum bias in the evaluation of diagnostic tests: lessons from the rapid dipstick test for urinary tract infection. *Ann Intern Med* 1992;117:135-140.

6. Metz CE. ROC methodology in radiologic imagining. *Invest Radiol* 1986;21:720-733.

7. Valenstein PN. Evaluating diagnostic tests with imperfect standards. *Am J Clin Pathol* 1990;93:252-258.

8. Revesz G, Kundel HL, Bonitatibus M. The effect of verification on the assessment of imaging techniques. *Invest Radiol* 1983;18:194-198.

9. Henkelman RM, Kay I, Bronskill M . Receiver operator characteristic (ROC) analysis without truth. *Med Decis Making* 1990;10:24-29.

10. Campbell G, DeLeo JM. Fundamentals of fuzzy receiver operating characteristic (ROC) functions. In: Malone L, Beck K, eds. *Computing Science and Statistics: Proceedings of the Twenty-First Symposium on the Interface*. Alexandria, VA: American Statistical Association, 1989:543-548.

11. DeLeo JM, Campbell G. The fuzzy receiver operating characteristic function and medical decisions with uncertainty. *Proceedings of the First International Symposium on Uncertainty Modeling and Analysis*. College Park, NJ: IEEE Computer Society Press, 1990:694-699.

12. Campbell G, Levy D, Bailey JJ. Bootstrap comparison of fuzzy R.O.C. curves for ECG-LVH algorithms using data from the Framingham heart study. *J Electrocardiol* 1990;23 (suppl):132-137.

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

*14. Metz CE. Basic principles of ROC analysis. *Semin Nucl Med* 1978;8:283-298.

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

16. Lusted LB. Decision making studies in patient management. *N Engl J Med* 1971; 284:416-424.

*Note that these articles give detailed reviews of procedures. Review of these articles is especially recommended.

References (Continued)

17. Lusted LB. Signal detectability and medical decision-making. *Science* 1971;171:1217-1219.

18. McNeil BJ, Keeler E, Adelstein SJ. Primer on certain elements of medical decision making. *N Engl J Med* 1975: 293:211-215.

19. Weinstein MC, Fineberg HV. *Clinical Decision Analysis*. Philadelphia: WB Saunders, 1980.

20. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic curve. *J Math Psychol* 1975;12:387-415.

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

22. Hollander M, Wolfe DA. Nonparametric statistical methods. New York: John Wiley, 1973:67-78.

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

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

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

26. Greenhouse SW, Mantel N. The evaluation of diagnostic tests. *Biometrics* 1950; 6:399–412.

27. Wieand S, Gail MH, James BR, James KL. A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika* 1989; 76:585–592.

28. Van Steirteghem AC, Zweig MH, Robertson EA, et al. Comparison of the effectiveness of four clinical chemical assays in classifying patients with chest pain. *Clin Chem* 1982;28:1319-1324.

29. Leung FY, Galbraith LV, Jablonsky G, et al. Re-evaluation of the diagnostic utility of serum total creatine kinase and creatine kinase-2 in myocardial infarction. *Clin Chem* 1989;35:1435-1440.

30. Carson JL, Eisenberg JM, Shaw LM, et al. Diagnostic accuracy of four assays of prostatic acid phosphatase. Comparison using receiver operating characteristic curve analysis. *J Am Med Assoc* 1985;253:665-669.

31. Hermann GA, Sugiura HT, Krumm RP. Comparison of thyrotropin assays by relative operating characteristics analysis. *Arch Pathol Lab Med* 1986;110:21-25.

32. Kazmierczak SC, Van Lente F, Hodges ED. Diagnostic and prognostic utility of phospholipase A activity in patients with acute pancreatitis: comparison with amylase and lipase. *Clin Chem* 1991;37:356-360.

33. Flack MR, Oldfield EH, Cutler GB, et al. Urine free cortisol in the high-dose dexamethasone suppression test for the differential diagnosis of the Cushing syndrome. *Ann Intern Med* 1992;116:211-217.

*Note that these articles give detailed reviews of procedures. Review of these articles is especially recommended.

References (Continued)

34. Guyatt GH, Patterson C, Ali M, et al. Diagnosis of iron-deficiency anemia in the elderly. *Am J Med* 1990;88:205-209.

35. Beck JR. The role of new laboratory tests in clinical-decision making. *Clin Lab Med* 1982;2:51-77.

36. Zweig MH. Apolipoproteins and lipids in coronary artery disease: Analysis of diagnostic accuracy using receiver operating characteristic plots and areas. *Arch Pathol Lab Med* 1994; 118:141–144.

37. Centor RM, Keightley GE. Receiver operating characteristic (ROC) curve area analysis using The ROC ANALYZER. *Proceedings of the Symposium for Computer Applications to Medical Care*, 1989: 222-226.

38. Pellar TG, Leung FY, Henderson AR. A computer program for rapid generation of receiver operating characteristic curves and likelihood ratios in the evaluation of diagnostic tests. *Ann Clin Biochem* 1988;25: 411-416.

Summary of Comments and Subcommittee Responses

GP10-T: *Assessment of Clinical Sensitivity and Specificity of Laboratory Tests; Tentative Guideline*

General

1. We were very impressed with the document and believe it will be of value to the clinical laboratory. Although most laboratories may not do studies that lead to ROC plots, they certainly need to understand how they are developed and what they mean. This document will be a good start.

- **The subcommittee is pleased to receive this praise. No changes were requested.**

2. The document is a summary of the relevant issues written at an introductory primer level. It will therefore be of use to clinical "laboratorians" who will (one hopes) be guided by senior investigators responsible for experimental design and analysis. In fact, perhaps the most telling line of the document is this (page 5): "Consultation with a professional statistician is recommended..."

   In particular, none of the subtle issues involved in the data analysis are mentioned in the document; there is no display (or explanation) of the results on the double-probability scale that is most frequently used to fit the results with a straight line. Finally, there is a good list of available software and one can find technical guidance by working through the references at the back.

   In a few words, this is an OK introductory primer on the subject. Nevertheless, it is historic and important.

- **The subcommittee is pleased to receive these comments. No changes were requested.**

3. Our group, which routinely determines diagnostic efficiency, prefers cumulative distribution analysis graphs (see for example, BI Bluestein et. al. *Cancer Research* 1984;44:4131–4136) rather than ROC curves.

   Cumulative distribution analysis graphs are more readily understood. Sensitivity and specificity are immediately known for any concentration cutoff. ROC curves do not show concentration at all and specificity only indirectly.

- **Regarding cumulative distribution graphs, the subcommittee recognizes that these have desirable features including the display of decision thresholds. An important limitation is that multiple tests cannot be plotted together and compared directly to one another because the abscissa depends on the concentration scale peculiar to each test. This is the feature that allows for the display of decision thresholds but interferes with comparison of tests. ROC plots, because the axes are normalized, permit all tests to be evaluated, either singly or in multiples, on the very same scale, regardless of the original scale. The subcommittee did not intend to review all graphical or statistical approaches to evaluating test performance, nor did it intend to select one as the best or only approach. As ROC plots have finally received fairly widespread recognition, we feel it is appropriate to recommend them without contending that they are necessarily the only useful approach.**

Summary of Comments and Subcommittee Responses (Continued)

Specificity can be directly shown on an ROC plot simply by employing the variation using an abscissa on which the scale runs right to left instead of left to right. This is already mentioned in the document in the third paragraph of Section 4.5.2.1.

Foreword

4. In the sentence before "*Note that assessing*..." the "a" should be removed from the sentence to read: "It is important to know just how inherently accurate each tool (test) is as a diagnostic discriminator."

- **The subcommittee removed "a" from this sentence.**

Section 4.2.5

5. In this section, the authors recommend consulting a statistician. In our view, this should be emphasized very strongly because, as the authors point out in the response to Comment #46, the statistical techniques and issues are not simple. This is evident even in the subcommittee's own recommendation of McNemar's test or Fisher's exact test to compare ROC plots, which really are not appropriate. Greenhouse and Mantel (*Biometrics* 1950;12:399) derived appropriate non-parametric test statistics to use in this context. This class of statistics was generalized by Wieand, Gail, James, and James (*Biometrika* 1989;76:3:585–592), who provided a useful general nonparametric approach. In addition, parametric binormal models which are discussed by Metz, Hanley, and others, are computationally more manageable than nonparametric approaches, but they require careful assessment of the appropriateness of the statistical assumptions on which the tests and estimators are based.

   As a corollary to the above comment, more emphasis should be given to the importance of adequate sample size to provide a sufficiently precise estimate of the ROC curve and use of confidence intervals to assess the precision of the estimates. Because of the special nature of the test statistics, power/sample size computations are not possible with any currently available packages of which we are aware.

   The standard method of comparing ROC curves using area under the curve, although well accepted, is a blunt instrument, which receives much more emphasis than it deserves. This measure averages in ranges of sensitivity/specificity, which would be of little clinical usefulness and therefore are irrelevant to deciding between two competing technologies. Comparisons of ROC curves at a definite specificity, or over a limited range of relevant specifications, as proposed by Wieand et al above, is better.

- **The subcommittee recognizes the points made here and acknowledges the statistical complexities involved. Because we do not feel it is appropriate to deal with these extensive statistical issues in the document, we have revised Section 4.5.2.5, second paragraph, to be more general and refer the reader to more primary sources, including Greenhouse and Mantel, 1950, and Wieand et al, 1989. Also, we revised Section 4.5.2.4 by adding the caveat that global quantitative indices, such as area under the curve, can mask important information and that visual inspection of the plot is necessary to fully appreciate test accuracy. Likewise, Section 4.5.2.5 is revised to recommend visual inspection when comparing multiple tests. A sentence was added to Section 4.2.5 that emphasizes the need for appropriate sample size.**

Summary of Comments and Subcommittee Responses (Continued)

Section 4.3

6. In the special case when comparison is being done between different implementations of the same test (for instance, comparison of CKMB on different analyzers), it may not be necessary to go through the rigor of establishing the "true" clinical state of the patients. While I believe this is essential for a new test, when the clinical laboratory is assessing equivalence, it may be sufficient to simply compare the current implementation of the test with the new one using the final diagnosis on the chart. Although this diagnosis is biased, because it was determined using the laboratory's current test, the study should be valid because the question being asked is, "Are the two implementations of the test equivalent?" If the ROC plots show that the tests being compared are equivalent, then no additional studies would need to be done. However, if the ROC plots were substantially different, then additional work would need to be done to understand the difference. If the committee agrees with this and could include this type of information in the current guideline without much delay, I believe it would be of value.

- **While we recognize the logic in the approach used for the particular circumstances described in the comment, we prefer not to encourage users of the document to compromise on the rigor of their classification (diagnosis). Those users who are well acquainted with the principles will know when it may not be necessary to seek definitive classifications. Even in the situation described in the comment it is still advisable to establish the "true" clinical state if this had not been done originally when the "current" (old) test was studied.**

Section 4.3.5 & 4.4.1

7. It is our understanding that the terms "blind/blinded/blinding" are now politically incorrect. Contemporary terms are "masked/masking."

- **The subcommittee changed "blind" and "blindly" to "masked" in Sections 4.3.5 and 4.4.1.**

Section 4.5.2.1

8. In the second paragraph you write about plotting sensitivity/specificity pairs "over the entire range of results observed." That is unnecessary. Results only have to be plotted for the overlap region (the range in which sensitivity and specificity are both less than 1.0). This same error occurs in the fourth paragraph with the statement that "TP and FP fractions vary continuously with the decision threshold." No, they only vary when the cut-off point yields true positive fractions >0 and <1.

   An easy way to decide what range to plot is to look at the extremes for each group (disease vs. non-disease). For a test that increases with disease, the range of values to plot on the ROC curve is between the lowest value for the disease group and the highest value for the non-disease group.

   In the last paragraph of this section, the next to last sentence should read: "While the *selection* of a decision...."

- **The subcommittee agrees that results only have to be plotted for the overlap region. Sections 4.5.2.1 and 4.5.2.2 have been revised to add a statement to that effect.**

Summary of Comments and Subcommittee Responses (Continued)

9. On page 11, line 5, the word "section" should be *selection*—"While the selection of a decision threshold..."

- **"Section" has been changed to "selection" as suggested.**

Section 4.5.2.3

10. The word "accuracy" is used where I believe the word "sensitivity" should be used.

- **The term "accuracy," not "sensitivity," was indeed intended. Accuracy is used here to refer to the overall ability of the diagnostic device to discriminate between alternative states of health (see the Foreword). Sensitivity and specificity are components of accuracy. No change is indicated.**

11. The statistical discussion is a little bit difficult to understand. However, the author's recommendation to use commercially available programs is a good one.

- **No change was requested.**

Section 4.5.2.5

12. I believe this section should be expanded. McNemar's statistic for paired data and Fisher's exact test for unpaired data should be thoroughly described. Sample calculations would also be useful. Comparing two tests using their areas under the plot has significant weaknesses. For example, for tests where the ROC plots cross at some point, one test may be significantly better than the other at a certain decision point. This may not be reflected by comparing areas under each plot.

    There is no discussion in this section on test efficiency (TP + TN)/(Total subjects). Efficiency should be defined in the glossary and explained in this section. It is a commonly used method to describe a test's usefulness at a particular decision point. Tests can also be compared by their maximum efficiencies.

- **See Comment 5. The subcommittee recognizes the statistical complexity and notes that Comment 13 also addresses Fisher's exact test. As mentioned in Comment 5, we have added some primary references and simplified the discussion in the document in the belief that a thorough description of all of these approaches is beyond the scope of the document.**

  The term "efficiency" was removed previously in response to an earlier comment (#58). Because efficiency is very dependent on prevalence, it is not actually a characteristic of the test itself but of the interaction of the test with the setting.

13. On page 14, 2nd paragraph, the Fisher's exact test seems vague to us. An idea of the intended audience can be obtained from the "Summary of Comments."

- **See Comment 12.**

Summary of Comments and Subcommittee Responses (Continued)

Appendix

14. Rulemaker is not available. It never completed beta testing, and Digital Medicine, Inc. has not made it available. I am not sure why, since it progressed far enough to be used in studies and mentioned in publications.

- **Rulemaker is still under development and the date of availability is projected to be 1996. GP10 has been revised accordingly.**

Summary of Comments and Subcommittee Responses; Comment 50

15. I agree with item 2 in Comment 50 (page 38). An expansion of this document to discuss selection of decision limits and predictive values would be a significant value. Possibly discussions of "gray zones" could also be included. My experience is that there is a significant lack of understanding of the concepts, how they are determined, and how they should be used. While it may be beyond the scope of NCCLS to address what is essentially an educational issue, I believe guidelines similar in scope to those in the ROC document would help the educational process. However, I would not want to see the ROC document delayed to incorporate this information. I believe it has value and is and should be approved.

- **The subcommittee is pleased to receive the recommendation for approval.**

Related NCCLS Publications

EP5-T2 **Precision Performance of Clinical Chemistry Devices—Second Edition; Tentative Guideline (1992).** EP5-T2 contains guidelines for designing an experiment to evaluate the precision performance of clinical chemistry devices; recommendations on comparing the resulting precision estimates with manufacturer's precision performance claims and determining when such comparisons are valid; and manufacturer's guidelines for establishing claims.

EP6-P **Evaluation of the Linearity of Quantitative Analytical Methods; Proposed Guideline (1986).** EP6-P discusses the verification of the analytical range (or linearity) of a clinical chemistry device.

EP7-P **Interference Testing in Clinical Chemistry; Proposed Guideline (1986).** EP7-P discusses interference testing during characterization and evaluation of a clinical laboratory method or device.

EP9-T **Method Comparison and Bias Estimation Using Patient Samples; Tentative Guideline (1993).** EP9-T discusses procedures for determining the relative bias between two clinical chemistry methods or devices. It also discusses the design of a method comparison experiment using split patient samples and analysis of the data.

EP10-T2 **Preliminary Evaluation of Quantitative Clinical Laboratory Methods—Second Edition; Tentative Guideline (1993).** EP10-T2 addresses experimental design and data analysis for preliminary evaluation of the performance of an analytical method or device.