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...

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.