EP12-A
Vol. 22 No. 14

Replaces EP12-P
Vol. 20 No. 15

User Protocol for Evaluation of Qualitative Test Performance; Approved Guideline

CLINICAL AND
LABORATORY
STANDARDS
INSTITUTE ™

(Formerly NCCLS)
Providing NCCLS standards and guidelines,
ISO/TC 212 standards, and ISO/TC 76 standards

This document provides a protocol designed to optimize the experimental design for the evaluation of qualitative tests; to better measure performance; and to provide a structured data analysis.

A guideline for global application developed through the NCCLS consensus process.

NCCLS

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 specialties are urged to volunteer for participation in NCCLS projects. Please contact the NCCLS Executive Offices for additional information on committee participation.

User Protocol for Evaluation of Qualitative Test Performance; Approved Guideline

Abstract

NCCLS document EP12-A—*User Protocol for Evaluation of Qualitative Test Performance; Approved Guideline* provides the user with a consistent approach for protocol design and data analysis when evaluating qualitative diagnostic tests. Guidance is provided for both reproducibility and method-comparison studies.

NCCLS. *User Protocol for Evaluation of Qualitative Test Performance; Approved Guideline*. NCCLS document EP12-A (ISBN 1-56238-468-6). NCCLS, 940 West Valley Road, Suite 1400, Wayne, Pennsylvania 19087-1898 USA, 2002.

THE NCCLS consensus process, which is the mechanism for moving a document through two or more levels of review by the healthcare community, is an ongoing process. Users should expect revised editions of any given document. Because rapid changes in technology may affect the procedures, methods, and protocols in a standard or guideline, users should replace outdated editions with the current editions of NCCLS documents. Current editions are listed in the *NCCLS Catalog*, which is distributed to member organizations, and to nonmembers on request. If your organization is not a member and would like to become one, and to request a copy of the *NCCLS Catalog*, contact the NCCLS Executive Offices. Telephone: 610.688.0100; Fax: 610.688.0700; E-Mail: exoffice@nccls.org; Website: www.nccls.org

EP12-A
ISBN 1-56238-468-6
ISSN 0273-3099

User Protocol for Evaluation of Qualitative Test Performance; Approved Guideline

Volume 22 Number 14

Larry W. Clark, M.S., Chairholder
Patricia E. Garrett, Ph.D.
Robert Martin, Dr. P.H.
Kristen L. Meier, Ph.D.

NCCLS

Suggested Citation

(NCCLS. *User Protocol for Evaluation of Qualitative Test Performance; Approved Guideline.* NCCLS document EP12-A [ISBN 1-56238-468-6]. NCCLS, 940 West Valley Road, Suite 1400, Wayne, Pennsylvania 19087-1898 USA, 2002.)

Proposed Guideline
July 2000

Approved Guideline
August 2002

ISBN 1-56238-468-6
ISSN 0273-3099

Committee Membership

Area Committee on Evaluation Protocols

Jan S. Krouwer, Ph.D. **Chairholder**	**Krouwer Consulting** **Sherborn, Massachusetts**
Daniel W. Tholen, M.S. **Vice-Chairholder**	**Dan Tholen Statistical Services** **Traverse City, Michigan**

Subcommittee on Qualitative and Semiquantitative Testing

Larry W. Clark, M.S. **Chairholder**	**Bayer Corporation** **Elkhart, Indiana**
Patricia E. Garrett, Ph.D.	Boston Biomedica, Inc. West Bridgewater, Massachusetts
Henry T. Lee, Jr.	Harpers Ferry, West Virginia
Robert Martin, Dr. P.H.	Centers for Disease Control and Prevention Atlanta, Georgia

Advisors

Richard H. Albert, Ph.D.	Food and Drug Administration Washington, D.C.
Michael Lynch	Bayer Corporation Walpole, Massachusetts
Kristen Meier, Ph.D.	FDA Center for Devices/Rad. Health Rockville, Maryland

Jennifer K. McGeary, M.T.(ASCP), M.S.H.A. *Staff Liaison*	NCCLS Wayne, Pennsylvania
Patrice E. Polgar *Editor*	NCCLS Wayne, Pennsylvania
Donna M. Wilhelm *Assistant Editor*	NCCLS Wayne, Pennsylvania

Active Membership
(as of 1 July 2002)

Sustaining Members

Abbott Laboratories
American Association for Clinical Chemistry
Beckman Coulter, Inc.
BD and Company
bioMérieux, Inc.
CLMA
College of American Pathologists
GlaxoSmithKline
Nippon Becton Dickinson Co., Ltd.
Ortho-Clinical Diagnostics, Inc.
Pfizer Inc
Roche Diagnostics, Inc.

Professional Members

AISAR-Associazione Italiana per lo Studio degli
American Academy of Family Physicians
American Association for Clinical Chemistry
American Association for Respiratory Care
American Chemical Society
American Medical Technologists
American Public Health Association
American Society for Clinical Laboratory Science
American Society of Hematology
American Society for Microbiology
American Type Culture Collection, Inc.
Asociación Española Primera de Socorros (Uruguay)
Asociacion Mexicana de Bioquimica Clinica A.C.
Assn. of Public Health Laboratories
Assoc. Micro. Clinici Italiani-A.M.C.L.I.
British Society for Antimicrobial Chemotherapy
CADIME-Camara De Instituciones De Diagnostico Medico
Canadian Society for Medical Laboratory Science—Société Canadienne de Science de Laboratoire Médical
Clinical Laboratory Management Association
COLA
College of American Pathologists
College of Medical Laboratory Technologists of Ontario
College of Physicians and Surgeons of Saskatchewan
ESCMID
Fundación Bioquímica Argentina
International Association of Medical Laboratory Technologists
International Council for Standardization in Haematology
International Federation of Clinical Chemistry
Italian Society of Clinical Biochemistry and Clinical Molecular Biology
Japan Society of Clinical Chemistry
Japanese Committee for Clinical Laboratory Standards
Joint Commission on Accreditation of Healthcare Organizations
National Academy of Clinical Biochemistry
National Association of Testing Authorities – Australia
National Society for Histotechnology, Inc.
Ontario Medical Association Quality Management Program-Laboratory Service
RCPA Quality Assurance Programs PTY Limited
Sociedade Brasileira de Analises Clinicas
Sociedade Brasileira de Patologia Clinica
Sociedad Espanola de Bioquimica Clinica y Patologia Molecular
Turkish Society of Microbiology

Government Members

Association of Public Health Laboratories
Armed Forces Institute of Pathology
BC Centre for Disease Control
Centers for Disease Control and Prevention
Centers for Medicare & Medicaid Services/CLIA Program
Centers for Medicare & Medicaid Services
Chinese Committee for Clinical Laboratory Standards
Commonwealth of Pennsylvania Bureau of Laboratories
Department of Veterans Affairs
Deutsches Institut für Normung (DIN)
FDA Center for Devices and Radiological Health
FDA Center for Veterinary Medicine
FDA Division of Anti-Infective Drug Products
Iowa State Hygienic Laboratory
Massachusetts Department of Public Health Laboratories
National Center of Infectious and Parasitic Diseases (Bulgaria)
National Health Laboratory Service (South Africa)
National Institute of Standards and Technology
New York State Department of Health
Ohio Department of Health
Ontario Ministry of Health
Pennsylvania Dept. of Health
Saskatchewan Health-Provincial Laboratory
Scientific Institute of Public Health; Belgium Ministry of Social Affairs, Public Health and the Environment
Swedish Institute for Infectious Disease Control
Thailand Department of Medical Sciences

Industry Members

AB Biodisk
Abbott Laboratories
Abbott Laboratories, MediSense Products
Acrometrix Corporation
Ammirati Regulatory Consulting
Anaerobe Systems
Asséssor
AstraZeneca
AstraZeneca R & D Boston
Aventis
Axis-Shield POC AS
Bayer Corporation – Elkhart, IN
Bayer Corporation – Tarrytown, NY
Bayer Corporation – West Haven, CT
Bayer Medical Ltd.
BD

BD Biosciences – San Jose, CA
BD Consumer Products
BD Diagnostic Systems
BD Italia S.P.A.
BD VACUTAINER Systems
Beckman Coulter, Inc.
Beckman Coulter, Inc. Primary Care Diagnostics
Beckman Coulter K.K. (Japan)
Bio-Development SRL
Bio-Inova Life Sciences International
Bio-Inova Life Sciences North America
BioMedia Laboratories Sdn Bhd
BioMérieux (NC)
bioMérieux, Inc. (MO)
Biometrology Consultants
Bio-Rad Laboratories, Inc.
Bio-Rad Laboratories, Inc. - France
Biotest AG
Blaine Healthcare Associates, Inc.
Bristol-Myers Squibb Company
Canadian External Quality Assessment Laboratory
Capital Management Consulting, Inc.
Carl Schaper
Checkpoint Development Inc.
Chiron Corporation
ChromaVision Medical Systems, Inc.
Chronolab Ag
Clinical Design Group Inc.
Clinical Laboratory Improvement Consultants
Cognigen
Community Medical Center (NJ)
Control Lab (Brazil)
Copan Diagnostics Inc.
Cosmetic Ingredient Review
Cubist Pharmaceuticals
Dade Behring Inc. - Deerfield, IL
Dade Behring Inc. - Glasgow, DE
Dade Behring Inc. - Marburg, Germany
Dade Behring Inc. - Sacramento, CA
Dade Behring Inc. - San Jose, CA
Diagnostics Consultancy
Diagnostic Products Corporation
Eiken Chemical Company, Ltd.
Elan Pharmaceuticals
Electa Lab s.r.l.
Enterprise Analysis Corporation
Essential Therapeutics, Inc.
EXPERTech Associates, Inc.
F. Hoffman-La Roche AG
Fort Dodge Animal Health
General Hospital Vienna (Austria)
Gen-Probe
GlaxoSmithKline
Greiner Bio-One Inc.
Helena Laboratories
Home Diagnostics, Inc.
Immunicon Corporation
Instrumentation Laboratory
International Technidyne Corporation
IntraBiotics Pharmaceuticals, Inc.
I-STAT Corporation
Johnson and Johnson Pharmaceutical Research and Development, L.L.C.
Kendall Sherwood-Davis & Geck
LAB-Interlink, Inc.
Laboratory Specialists, Inc.
Labtest Diagnostica S.A.
LifeScan, Inc. (a Johnson & Johnson Company)
Lilly Research Laboratories
Macemon Consultants
Medical Device Consultants, Inc.
Merck & Company, Inc.
Minigrip/Zip-Pak
Molecular Diagnostics, Inc.
mvi Sciences (MA)
Nabi
Nichols Institute Diagnostics (Div. of Quest Diagnostics, Inc.)
NimbleGen Systems, Inc.
Nissui Pharmaceutical Co., Ltd.
Nippon Becton Dickinson Co., Ltd.
Norfolk Associates, Inc.
Novartis Pharmaceuticals Corporation
Ortho-Clinical Diagnostics, Inc. (Raritan, NJ)
Ortho-Clinical Diagnostics, Inc. (Rochester, NY)
Oxoid Inc.
Paratek Pharmaceuticals
Pfizer Inc
Pharmacia Corporation
Philips Medical Systems
Powers Consulting Services
Premier Inc.
Procter & Gamble Pharmaceuticals, Inc.
The Product Development Group
QSE Consulting
Quintiles, Inc.
Radiometer America, Inc.
Radiometer Medical A/S
David G. Rhoads Associates, Inc.
Roche Diagnostics GmbH
Roche Diagnostics, Inc.
Roche Laboratories (Div. Hoffmann-La Roche Inc.)
Sarstedt, Inc.
SARL Laboratoire Carron (France)
Schering Corporation
Schleicher & Schuell, Inc.
Second Opinion
Showa Yakuhin Kako Company, Ltd.
Streck Laboratories, Inc.
SurroMed, Inc.
Synermed Diagnostic Corp.
Sysmex Corporation (Japan)
Sysmex Corporation (Long Grove, IL)
The Clinical Microbiology Institute
The Toledo Hospital (OH)
Theravance Inc.
Transasia Engineers
Trek Diagnostic Systems, Inc.
Versicor, Inc.
Vetoquinol S.A.
Visible Genetics, Inc.
Vysis, Inc.
Wallac Oy
Wyeth-Ayerst
Xyletech Systems, Inc.
YD Consultant
YD Diagnostics (Seoul, Korea)

Trade Associations

AdvaMed
Association of Medical Diagnostic Manufacturers
Japan Association Clinical Reagents Ind. (Tokyo, Japan)
Medical Industry Association of Australia

Associate Active Members

20th Medical Group (SC)
31st Medical Group/SGSL (APO, AE)
67th CSH Wuerzburg, GE (NY)
121st General Hospital (CA)
Academisch Ziekenhuis-VUB (Belgium)
Acadiana Medical Laboratories, LTD (LA)
Adena Regional Medical Center (OH)
Advocate Healthcare Lutheran General (IL)
Akershus Central Hospital and AFA (Norway)
Albemarle Hospital (NC)

Allegheny General Hospital (PA)
Allegheny University of the Health Sciences (PA)
Allina Health System (MN)
Alton Ochsner Medical Foundation (LA)
American Medical Laboratories (VA)
Antwerp University Hospital (Belgium)
Arkansas Department of Health
ARUP at University Hospital (UT)
Armed Forces Research Institute of Medical Science (APO, AP)
Associated Regional & University Pathologists (UT)
Aurora Consolidated Laboratories (WI)
Azienda Ospedale Di Lecco (Italy)
Bay Medical Center (MI)
Baystate Medical Center (MA)
Bbaguas Duzen Laboratories (Turkey)
Bermuda Hospitals Board
Bo Ali Hospital (Iran)
British Columbia Cancer Agency (Vancouver, BC, Canada)
Brooks Air Force Base (TX)
Broward General Medical Center (FL)
Calgary Laboratory Services
Carilion Consolidated Laboratory (VA)
Cathay General Hospital (Taiwan)
CB Healthcare Complex (Sydney, NS, Canada)
Central Peninsula General Hospital (AK)
Central Texas Veterans Health Care System
Centre Hospitalier Regional del la Citadelle (Belgium)
Centro Diagnostico Italiano (Milano, Italy)
Champlain Valley Physicians Hospital (NY)
Chang Gung Memorial Hospital (Taiwan)
Changi General Hospital (Singapore)
Children’s Hospital (NE)
Children’s Hospital & Clinics (MN)
Children’s Hospital Medical Center (Akron, OH)
Children’s Hospital of Philadelphia (PA)
Children’s Medical Center of Dallas (TX)
Clarian Health–Methodist Hospital (IN)
Clendo Lab (Puerto Rico)
Clinical Laboratory Partners, LLC (CT)
CLSI Laboratories (PA)
Columbia Regional Hospital (MO)
Commonwealth of Kentucky
Community Hospital of Lancaster (PA)
CompuNet Clinical Laboratories (OH)
Cook County Hospital (IL)
Cook Children’s Medical Center (TX)
Covance Central Laboratory Services (IN)
Danish Veterinary Laboratory (Denmark)
Danville Regional Medical Center (VA)
Delaware Public Health Laboratory
John F. Kennedy Medical Center (NJ)
John Peter Smith Hospital (TX)
DesPeres Hospital (MO)
DeTar Hospital (TX)
Detroit Health Department (MI)
Diagnosticos da América S/A (Brazil)
Dr. Everett Chalmers Hospital (New Brunswick, Canada)
Doctors Hospital (Bahamas)
Duke University Medical Center (NC)
E.A. Conway Medical Center (LA)
Eastern Maine Medical Center
East Side Clinical Laboratory (RI)
Eastern Health (Vic., Australia)
Elyria Memorial Hospital (OH)
Emory University Hospital (GA)
Esoterix Center for Infectious Disease (TX)
Fairview-University Medical Center (MN)
Federal Medical Center (MN)
Florida Hospital East Orlando
Foothills Hospital (Calgary, AB, Canada)
Fort St. John General Hospital (Fort St. John, BC, Canada)
Fox Chase Cancer Center (PA)
Fresenius Medical Care/Spectra East (NJ)
Fresno Community Hospital and Medical Center
Frye Regional Medical Center (NC)
Gambro Healthcare Laboratory Services (FL)
Gateway Medical Center (TN)
Geisinger Medical Center (PA)
Grady Memorial Hospital (GA)
Guthrie Clinic Laboratories (PA)
Hahnemann University Hospital (PA)
Harris Methodist Erath County (TX)
Harris Methodist Fort Worth (TX)
Hartford Hospital (CT)
Headwaters Health Authority (Alberta, Canada)
Health Network Lab (PA)
Health Partners Laboratories (VA)
Heartland Regional Medical Center (MO)
Highlands Regional Medical Center (FL)
Hoag Memorial Hospital Presbyterian (CA)
Holmes Regional Medical Center (FL)
Holzer Medical Center (OH)
Hopital du Sacre-Coeur de Montreal (Montreal, Quebec, Canada)
Hôpital Maisonneuve – Rosemont (Montreal, Canada)
Hospital for Sick Children (Toronto, ON, Canada)
Hospital Sousa Martins (Portugal)
Hotel Dieu Hospital (Windsor, ON, Canada)
Houston Medical Center (GA)
Huddinge University Hospital (Sweden)
Hurley Medical Center (MI)
Indiana State Board of Health
Indiana University
Institute of Medical and Veterinary Science (Australia)
International Health Management Associates, Inc. (IL)
Jackson Memorial Hospital (FL)
Jersey Shore Medical Center (NJ)
John C. Lincoln Hospital (AZ)
John F. Kennedy Medical Center (NJ)
John Peter Smith Hospital (TX)
Kadlec Medical Center (WA)
Kaiser Permanente Medical Care (CA)
Kaiser Permanente (MD)
Kantonsspital (Switzerland)
Keller Army Community Hospital (NY)

Kenora-Rainy River Regional Laboratory Program (Ontario, Canada)
Kern Medical Center (CA)
Kimball Medical Center (NJ)
King Faisal Specialist Hospital (Saudi Arabia)
King Khalid National Guard Hospital (Saudi Arabia)
King's Daughter Medical Center (KY)
Klinični Center (Slovenia)
Laboratories at Bonfils (CO)
Laboratoire de Santé Publique du Quebec (Canada)
Laboratório Fleury S/C Ltda. (Brazil)
Laboratory Corporation of America (NJ)
Laboratory Corporation of America (MO)
LAC and USC Healthcare Network (CA)
Lakeland Regional Medical Center (FL)
Lancaster General Hospital (PA)
Langley Air Force Base (VA)
LeBonheur Children's Medical Center (TN)
L'Hotel-Dieu de Quebec (Canada)
Libero Instituto Univ. Campus BioMedico (Italy)
Louisiana State University Medical Center
Maccabi Medical Care and Health Fund (Israel)
Magee Womens Hospital (PA)
Malcolm Grow USAF Medical Center (MD)
Manitoba Health (Winnipeg, Canada)
Martin Luther King/Drew Medical Center (CA)
Massachusetts General Hospital (Microbiology Laboratory)
MDS Metro Laboratory Services (Burnaby, BC, Canada)
Medical College of Virginia Hospital
Medicare/Medicaid Certification, State of North Carolina
Memorial Medical Center (IL)
Memorial Medical Center (LA) Jefferson Davis Hwy
Memorial Medical Center (LA) Napoleon Avenue
Methodist Hospital (TX)
Methodist Hospitals of Memphis (TN)
MetroHealth Medical Center (OH)
Michigan Department of Community Health
Mississippi Baptist Medical Center
Monte Tabor – Centro Italo - Brazileiro de Promocao (Brazil)
Montreal Children's Hospital (Canada)
Montreal General Hospital (Canada)
MRL Pharmaceutical Services, Inc. (VA)
MRL Reference Laboratory (CA)
Nassau County Medical Center (NY)
National Institutes of Health (MD)
Naval Hospital – Corpus Christi (TX)
Naval Surface Warfare Center (IN)
Nebraska Health System
New Britain General Hospital (CT)
New England Fertility Institute (CT)
North Carolina State Laboratory of Public Health
North Kansas City Hospital (MO)
North Shore – Long Island Jewish Health System Laboratories (NY)
Northwestern Memorial Hospital (IL)
O.L. Vrouwziekenhuis (Belgium)
Ordre professionnel des technologists médicaux du Québec
Ospedali Riuniti (Italy)
The Ottawa Hospital (Ottawa, ON, Canada)
Our Lady of Lourdes Hospital (NJ)
Our Lady of the Resurrection Medical Center (IL)
Pathology and Cytology Laboratories, Inc. (KY)
The Permanente Medical Group (CA)
Piedmont Hospital (GA)
Pikeville Methodist Hospital (KY)
Pocono Hospital (PA)
Presbyterian Hospital of Dallas (TX)
Queen Elizabeth Hospital (Prince Edward Island, Canada)
Queensland Health Pathology Services (Australia)
Quest Diagnostics Incorporated (CA)
Quintiles Laboratories, Ltd. (GA)
Regions Hospital
Reid Hospital & Health Care Services (IN)
Research Medical Center (MO)
Rex Healthcare (NC)
Rhode Island Department of Health Laboratories
Riyadh Armed Forces Hospital (Saudi Arabia)
Royal Columbian Hospital (New Westminster, BC, Canada)
Sacred Heart Hospital (MD)
Saint Mary's Regional Medical Center (NV)
St. Alexius Medical Center (ND)
St. Anthony Hospital (CO)
St. Anthony's Hospital (FL)
St. Barnabas Medical Center (NJ)
St-Eustache Hospital (Quebec, Canada)
St. Francis Medical Ctr. (CA)
St. John Hospital and Medical Center (MI)
St. John Regional Hospital (St. John, NB, Canada)
St. Joseph Hospital (NE)
St. Joseph's Hospital – Marshfield Clinic (WI)
St. Joseph Mercy Hospital (MI)
St. Jude Children's Research Hospital (TN)
St. Luke's Regional Medical Center (IA)
St. Mary of the Plains Hospital (TX)
St. Mary's Hospital & Medical Center (CO)
St. Paul's Hospital (Vancouver, BC, Montreal)
St. Vincent Medical Center (CA)
Ste. Justine Hospital (Montreal, PQ, Canada)
Salina Regional Health Center (KS)
San Francisco General Hospital (CA)
Santa Clara Valley Medical Center (CA)
Seoul Nat'l University Hospital (Korea)
Shanghai Center for the Clinical Laboratory (China)
South Bend Medical Foundation (IN)
Southwest Texas Methodist Hospital (TX)
South Western Area Pathology Service (Australia)

Southern Maine Medical Center
Specialty Laboratories, Inc. (CA)
Stanford Hospital and Clinics (CA)
State of Washington Department of Health
Stony Brook University Hospital (NY)
Stormont-Vail Regional Medical Center (KS)
Sun Health-Boswell Hospital (AZ)
Sunrise Hospital and Medical Center (NV)
Swedish Medical Center – Providence Campus (WA)
Tampa General Hospital (FL)
Temple University Hospital (PA)
Tenet Odessa Regional Hospital (TX)
The Toledo Hospital (OH)
Touro Infirmary (LA)
Trident Regional Medical Center (SC)
Tripler Army Medical Center (HI)
Truman Medical Center (MO)
UCSF Medical Center (CA)
UNC Hospitals (NC)
University College Hospital (Galway, Ireland)
University Hospital (Gent) (Belgium)
University Hospitals of Cleveland (OH)
The University Hospitals (OK)
University of Alabama-Birmingham Hospital
University of Alberta Hospitals (Canada)
University of Colorado Health Science Center
University of Chicago Hospitals (IL)
University of Illinois Medical Center
University of the Ryukyus (Japan)
University of Texas M.D. Anderson Cancer Center
University of Virginia Medical Center
University of Washington
UZ-KUL Medical Center (Belgium)
VA (Denver) Medical Center (CO)
Virginia Department of Health
VA (Kansas City) Medical Center (MO)
VA (Western NY) Healthcare System
VA (San Diego) Medical Center (CA)
VA (Tuskegee) Medical Center (AL)
VA Outpatient Clinic (OH)
Vejle Hospital (Denmark)
Washington Adventist Hospital (MD)
Washoe Medical Center Laboratory (NV)
West Jefferson Medical Center (LA)
West Shore Medical Center (MI)
Wilford Hall Medical Center (TX)
William Beaumont Army Medical Center (TX)
William Beaumont Hospital (MI)
Williamsburg Community Hospital (VA)
Winn Army Community Hospital (GA)
Winnipeg Regional Health Authority (Winnipeg, Canada)
Wishard Memorial Hospital (IN)
Yonsei University College of Medicine (Korea)
York Hospital (PA)

OFFICERS

BOARD OF DIRECTORS

Contents

Contents (Continued)

Foreword

Qualitative diagnostic tests have been used since the early days of laboratory medicine for the screening, diagnosis, and management of a variety of diseases. These tests are found in many specialties of the clinical laboratory. Method evaluation procedures for such tests are diverse, with each laboratory specialty often emphasizing different issues in both the experimental design and in the data analysis and interpretation of such studies.

There have been two key published efforts to standardize both the experimental details as well as the data analysis of qualitative information.[1,2] The International Federation of Clinical Chemistry published a guideline in 1989 on protocol design and data analysis, featuring examples for urinary glucose and albumin by visually read reagent strips. [1] The European Committee for Clinical Laboratory Standards published a guideline in 1990 that focused on the evaluation of qualitative tests.[2] A very prominent work on the assessment of both quantitative and qualitative laboratory tests was written by Gambino and Galen in 1975, *Beyond Normality: The Predictive Value and Efficiency of Medical Diagnosis*. NCCLS document GP10— *Assessment of the Clinical Accuracy of Laboratory Tests Using Receiver Operating Characteristic (ROC) Plots* describes the assessment of the accuracy of a test compared to the clinical status of the patient.

The latter two references both focus on the relation of the test result to the clinical status of the patient, for either qualitative or quantitative tests. In many laboratories, the clinical information is not readily available, so it is important that protocols for evaluations be established that enable comparison of a new test to other laboratory procedures, in much the same way that most method evaluation studies are performed for quantitative tests. Ideally, the comparison should be made to a "reference" procedure or "gold standard." However, comparison with a method in current use is also of interest. This guideline describes two different situations for these studies: the first is when the laboratory knows the diagnosis of each patient specimen in the study, and the second is when the laboratory does not know the clinical diagnosis of each patient specimen. These are treated separately, to enable appropriate data analysis. Parameters such as specificity, sensitivity, and predictive value for the test method are estimated in the former situation, and agreement measures are estimated in the latter situation.

This guideline is intended to promote uniformity in performance assessment of qualitative testing among

- laboratories of all types that perform qualitative tests;
- manufacturers of qualitative diagnostic kits, for design of the studies they use to demonstrate kit performance, as well as the way kit performance is described; and
- regulatory agencies and laboratory surveyors.

Key Words

Analytical goals, qualitative test, semiquantitative test

The Quality System Approach

NCCLS subscribes to a quality system approach in the development of standards and guidelines, which facilitates project management; defines a document structure via a template; and provides a process to identify needed documents through a gap analysis. The approach is based on the model presented in the most current edition of NCCLS HS1- *A Quality System Model for Health Care.* The quality system approach applies a core set of "quality system essentials (QSEs)," basic to any organization, to all operations in any healthcare service's path of workflow. The QSEs provide the framework for delivery of any type of product or service, serving as a manager's guide. The quality system essentials (QSEs) are:

QSEs

Documents & Records
Organization
Personnel
Equipment
Purchasing & Inventory
Process Control
Information Management
Occurrence Management
Assessment
Process Improvement
Service & Satisfaction
Facilities & Safety

EP12-A Addresses the Following Quality System Essentials (QSEs)

Documents & Records	Organization	Personnel	Equipment	Purchasing & Inventory	Process Control	Information Management	Occurrence Management	Assessment	Process Improvement	Service & Satisfaction	Facilities & Safety
					X						

Adapted from NCCLS document HS1— *A Quality System Model for Health Care*

User Protocol for Evaluation of Qualitative Test Performance; Approved Guideline

1 Introduction

Qualitative diagnostic tests, which are found in many specialties of the clinical laboratory, have been used for the screening, diagnosis, and management of a variety of diseases. Method evaluation procedures for such tests are diverse, with each laboratory specialty often emphasizing different issues in the experimental design as well as the data analysis and interpretation of such studies.

A new qualitative test can be implemented in the clinical laboratory for a number of reasons. The new test might be easier to use, more economical, have enhanced performance, or otherwise better meet the user's needs. Before patient test results are reported with a new qualitative test, users must document the test performance in their clinical laboratories. Documentation of test performance is not limited to comparison with another method. Laboratory staff training and proficiency with the new qualitative test, preparation of proper specimen collection and handling, and documentation of a quality control system are necessary before implementation of any new test.

Although universal evaluation guidelines for all qualitative tests are not feasible or practical, several common features do exist. Before collecting performance evaluation data, proper familiarization, training, and a quality assurance plan should be completed. Any qualitative test must provide the user with consistent and correct results, and reproducibility studies and comparison of methods studies with patient specimens are used to demonstrate the test's performance capabilities.

This guideline is intended to promote uniformity in performance assessment of qualitative testing among laboratories of all types that perform qualitative tests; manufacturers of qualitative diagnostic kits, for design of the studies they use to demonstrate kit performance, as well as the way kit performance is described; and regulatory agencies and laboratory surveyors.

2 Scope

This guideline provides evaluation protocols for the demonstration of qualitative test performance. Here, a qualitative test is restricted to those tests that have only two possible outcomes. Future revisions may be expanded to include qualitative tests that have more than two outcomes. EP12 is written for clinical laboratory personnel who are the end users of such tests. Demonstration of test performance by the user can satisfy internal (as well as external) expectations that the test performs acceptably in meeting the user's clinical and analytical goals. This guideline for test performance may help the user meet documentation and regulatory needs, but it is not intended to meet all of the user's goals and requirements, because regulatory and documentation needs vary.

3 Clinical Utility

Qualitative tests may be used clinically for screening, diagnostic, confirmatory, or monitoring purposes. The test's sensitivity, specificity, predictive values, and efficiency, and the prevalence of the disease or condition in the population being tested, determine the clinical utility of the qualitative test just as for a quantitative test.

3.1 Screening Tests

Clinically, screening methods are used to test entire populations (or subsets of such populations) for the presence of the analyte or agent. Examples may be the detection of blood in feces or the use of the Venereal Disease Research Laboratory (VDRL) syphilis serology test. As a rule, these qualitative tests used for screening purposes should have a high sensitivity to ensure that true-positive results are detected. Generally, screening tests produce more false-positive results than diagnostic or confirmatory tests. This lower specificity can be tolerated if a good confirmatory test exists and if the social/economic consequences of the false-positive results are not too severe.

The need to follow up positive screening results with confirmatory testing to detect false-positive results can be preferable to the occurrence of false-negative test results, because false negatives can result in errors, such as the transfusion of infectious blood, or in the failure to treat a serious, treatable condition.

3.2 Diagnostic Tests

Qualitative tests are often used to diagnose a particular disease or condition based on a clinical suspicion that it might be present. The use of various microbiology culture tests to detect infection is one example of a diagnostic test. Clinically, the requirement for timely and proper treatment demands that diagnostic tests have excellent sensitivity and specificity. If a confirmatory test always follows the diagnostic test, the specificity requirement can be somewhat lower.

3.3 Confirmatory Tests

Confirmatory tests are used to follow up screening or diagnostic test results. The verification or confirmation of the previous test result permits the clinician to establish a diagnosis. Confirmatory tests are designed to be specific (at the expense of sensitivity, if necessary) and have a high positive predictive value. The Fluorescent Treponemal Antibody Absorption test (FTA-ABS) syphilis serology test is an example of a confirmatory test that follows a screening test, such as the VDRL syphilis serology test.

4 Definitions[a]

The following terms are defined for use in this guideline:

Accuracy, *n* **- 1)** Closeness of the agreement between the result of a measurement and a true value of the measurand {/analyte}.

Analyte, *n* **-** A substance or constituent for which the laboratory conducts testing; **NOTE:** This includes any element, ion, compound, substance, factor, infectious agent, cell, organelle, activity (enzymatic, hormonal, or immunological), or property, the presence or absence, concentration, activity, intensity, or other characteristics of which are to be determined. See **Measurand**.

Clinical sensitivity, *n* – The proportion of patients with a well-defined clinical disorder whose test values are positive or exceed a defined decision limit (i.e., a positive result and identification of the patients who have a disease); **NOTE**: The clinical disorder must be defined by criteria independent of the test under consideration.

Clinical specificity, *n* - The proportion of subjects who do not have a specified clinical disorder whose test results are negative or within the defined decision limit.

Control//control material, *n* - A device, solution, or lyophilized preparation intended for use in the quality control process; **NOTES:** a) The expected reaction or concentration of analytes of interest are known within limits ascertained during preparation and confirmed in use; b) Control materials are generally not used for calibration in the same process in which they are used as controls.

[a] Some of these definitions are found in NCCLS document NRSCL8—*Terminology and Definitions for Use in NCCLS Documents*. For complete definitions and detailed source information, please refer to the most current edition of that document.

Cutoff, *n* – **(1)** The test response point below which a qualitative test result is determined to be negative and above which the result is determined to be positive (or vice versa); **NOTE**: For a truly qualitative test, the cutoff is the (only) medical decision point; for a qualitative test derived from dichotomizing a quantitative or ordinal scale, there are many possible choices for a cutoff. **(2)** The analyte concentration at which repeated tests on the same sample yield positive results 50% of the time and negative results for the other 50% (*ECCLS*).

Efficiency, *n* - ***Immunoassay;*** The percentage (number fraction multiplied by 100) of results that are true results, whether positive or negative.

False negative result//False negative (FN), *n* **-** A negative test result for a patient or specimen that is positive for the condition or constituent in question.

False positive result//False positive (FP), *n* **-** A positive test result for a patient or specimen that is negative for the condition or constituent in question.

Gold standard, *n* - A nonspecific term that indicates that a process or material(s) is the best available approximation of the truth; **NOTE**: Its use is deprecated.

Measurand, *n* - A particular quantity subject to measurement*;* **NOTE:** This term and definition encompass all quantities, while the commonly used term “analyte” refers to a tangible entity subject to measurement. For example, “substance” concentration is a quantity that may be related to a particular analyte.

Negative predictive value, *n* - The likelihood that an individual with a negative test does not have the disease, or other characteristic, which the test is designed to detect.

Positive predictive value, *n* - The likelihood that an individual with a positive test result has a particular disease, or characteristic, that the test is designed to detect.

Prevalence, *n* - The extent of occurrence expressed as a fraction of the numbers affected by the disease or condition compared to the total number of members in the specified group.

Qualitative tests, *n* - Those test methods that provide only two categorical responses (i.e., positive/negative or yes/no); **NOTE**: A truly qualitative test is based on a single medical decision point; alternatively, some tests labeled as qualitative are derived from dichotomizing a quantitative or ordinal scale.

Reproducibility, *n* - The closeness of the agreement between the results of measurements of the same measurand, where the measurements are carried out under changed conditions; **NOTES:** a) Changed conditions may include: principle or method of measurement, observer, measuring instrument, location, conditions of use, and time; b) Reproducibility may be expressed quantitatively in terms of dispersion characteristics of the results.

True negative//True negative result (TN), *n* - A negative result of a test for a disease or condition in a subject in whom the disease or condition is absent.

True positive/True positive result (TP), *n* - A positive result of a test for a disease or condition for a subject in whom the disease or condition is present.

5 Device Familiarization and Training

5.1 Purpose

A familiarization and training period with the new test method must be completed before evaluation data is collected. The test method proficiency should include a demonstrated understanding of sample handling and storage, kit reagent handling and storage, proper test protocol, proper interpretation of results, and QC for the system. Each test operator must be proficient with the test method.

5.2 Duration

A training session in which the method is demonstrated for, and practiced by, each operator should be followed on a separate day by another session where each operator demonstrates proficiency with the method. An objective judgment that proficiency was achieved by all operators during the familiarization and training period should be made before continuing the evaluation.

6 Evaluation Materials

6.1 Controls

Appropriate quality control materials are required to ensure consistent test performance. The control(s) provided by the manufacturer should be used as directed. Other stable commercial controls or clinical controls may be used if care is taken to ensure control material free of matrix effects. If possible, the same quality control material should be tested with all methods if a multiple-method comparison is being made. For many qualitative tests, the daily use of a negative and positive control is sufficient, while some qualitative test methods may require more frequent testing of controls. Some qualitative tests produce numerical results that can be monitored like quantitative method controls to assess test performance (please refer to NCCLS document C24—*Statistical Quality Control for Quantitative Measurements: Principles and Definitions* for more information). Corrective action must be taken if the expected results are not obtained for the controls.

6.2 Specimen Collection and Handling

Clinical specimens used for the evaluation must be collected according to the manufacturer's instructions using good clinical laboratory practices. If required for the method, fresh specimens should be collected and tested without delay to reduce concerns about specimen quality. It may be appropriate to process the evaluation specimens with a transport system if the transport system represents typical specimen collection and handling. The timing of the collection of the specimens for some clinical conditions is critical and should be consistent during the evaluation.

7 Reproducibility Studies

7.1 Negative and Positive Controls

During the course of the evaluation, control materials should be tested with each run to document that the assay met expected performance requirements, and thus the data are to be considered valid. The manufacturer's recommended negative and positive control materials are to be tested during each run of the test method over the course of the comparison of methods study (see Section 8). If the comparison of methods study is completed in ten days, then duplicate measurements of each control material should be made in each run, to give a total of 20 replicates. If the comparison of methods study is completed over 20 days, then single measurements of each control material should be made in each run, to also give a total of 20 replicates.

If either of the control materials does not give expected results, the run must be rejected. Patient results from rejected runs must not be accepted for the study. A new run on either the same day or an additional day must be scheduled to replace the rejected run. Obviously, the laboratory must investigate the cause for the unacceptable QC result. No more than one run may be rejected in a ten-day study or two runs in a twenty-day study. If there are more rejected runs than this, the laboratory must discontinue testing and consult with the kit manufacturer to identify the cause and implement corrective action.

7.2 Analyte Concentrations Near the Cutoff

Reproducibility studies for qualitative tests should provide an estimate of the precision of the method at analyte concentrations near the cutoff. It is not appropriate to measure the reproducibility of a qualitative assay with low-negative or high-positive samples, as these are too far away in analyte concentration from the medical decision point.

A useful definition for the cutoff point in a qualitative test method is the analyte concentration at which repeated tests on the same sample yield positive results 50% of the time and negative results for the other 50%.[1] Then, increasing concentrations of the analyte in small increments and performing multiple tests on each sample would be expected to yield correspondingly larger percentages of positive results and smaller percentages of negative results. Likewise, decreasing concentrations by the same increments might be expected to yield an opposite pattern of positive and negative results.

This description of the cutoff point or discrimination point in a qualitative test allows an understanding of the fact that, at concentrations of analyte near the cutoff point, there will be imprecision, and test results (positive or negative, plus or minus, absent or present) will not be completely consistent on multiple observations of the same sample if its analyte concentration is in this range.

The concentrations above and below the cutoff point at which repeated results are 95% positive or 95% negative, respectively, have been called the "95% interval" for the cutoff point for that method.[1] At concentrations of analyte higher or lower than the cutoff concentration and beyond the 95% interval, the ability of a method to consistently produce the same result on repeated observations of the same sample is a characteristic of a good, or robust, method.

The range of concentrations within the 95% interval, and the range of concentrations required to reach the point where consistent results are produced on the same sample, may be different in different tests for the same analyte, and the ability to distinguish this difference can be a useful evaluation tool.

The reproducibility experiment described below cannot define either the 95% interval or the range of concentrations required for a consistent result, but it can indicate whether those ranges are within or outside of a 20% concentration range from the cutoff point. A more extensive experiment is described by the European Committee for Clinical Laboratory Standards (ECCLS).[1] Both experiments assume that the dose response curve (a plot of concentration vs. observed test result) is linear for a diluted sample in the concentration range near the cutoff point. When this assumption does not hold, as in the case of enzyme immunoassays (EIA) screening results for mixtures of HIV antibodies, neither experiment will yield valid results.[3]

7.3 A Qualitative Method-Reproducibility Experiment for Analyte Concentrations Near the Cutoff

(1) The purpose of this section is to establish the analyte's cutoff concentration for the test method under study and determine that the +/- 20% concentration range at the cutoff concentration is within the 95% interval. The package insert for the test method might state the cutoff concentration for the analyte, but often it does not. If the cutoff concentration cannot be estimated by this or other means, a dilution series can be made from a positive sample, and dilutions can be tested in replicate to

determine the dilution that yields 50% positive and 50% negative results. This dilution then contains the analyte concentration at the cutoff point.

(2) Prepare samples at the cutoff concentration and with concentrations 20% above and 20% below the cutoff concentration in sufficient volume to allow up to 20 replicate tests on the same sample.

(3) Test the samples in replicates up to 20, and determine the percentage of positive and negative results for each sample.

If it is not feasible to test each sample 20 times, useful information can be gained from fewer replicates, but the statistical power of such results is less.

(4) Use the percentages derived from step 3 to determine the following:

(a) Was the estimated cutoff concentration accurate? If it was, the replicates on that sample should have yielded 50% positive and 50% negative results. If the test results for the cutoff concentration sample were different from 50% positive/50% negative, the estimate of the cutoff concentration was inaccurate, the number of tests performed was insufficient to yield an accurate result, or the dose-response curve for the method is not linear near the cutoff point.[3]

(b) Is the +20 to –20% concentration range within, at, or outside the 95% interval for the test method?

If the +20% sample yielded positive results $\geq$ 95% of the time, and the –20% sample yielded negative results $\geq$ 95% of the time, then this range is at or outside of the 95% interval for the method. Thus, samples $\geq$ 20% away from the cutoff concentration can be expected to yield consistent results with this method.

If the +20% sample yielded positive results <95% of the time, and/or the -20% sample yielded negative results <95% of the time, then this range is within the 95% interval for the method. Thus, samples 20% away from the cutoff concentration cannot be expected to yield consistent results with this method, and the 95% interval for the method is >20% away from the cutoff concentration.

In the case of a method with a 95% interval >20% away from the cutoff concentration, another experiment or series of experiments is required to determine the actual 95% interval.

As an example, a visual qualitative human chorionic gonadotropin (hCG) test in urine could have a cutoff concentration of 16 mIU/mL. A sample prepared at 16 mIU/mL (i.e., by a quantitative method) yielded 50% positive and 50% negative results with the visual hCG test. The +20% sample (19 mIU/mL) yielded positive results ≥95% of the time for the 20 replicates, and the –20% (13 mIU/mL) sample yielded negative results ≥ 95% of the time.

8 Comparison of Methods

In one common type of comparison of methods study design, the same set of specimens is tested by two or more methods and the results are compared. The particular study design varies depending on the qualitative test being evaluated. The comparative method may be another qualitative method (such as the user's current method), the "gold standard" method, a quantitative method, or the clinical diagnosis. Reference panels and proficiency samples may also be utilized to study test performance. PT materials may suffer from matrix interferences which could generate misleading conclusions regarding method performance.[4,5,6] This limitation would be particularly important when the PT material had analyte content near the threshold of positive or negative, or when the analyte molecular form were different from that of a native specimen such that an immunologic method may not recognize the epitope. The following sections describe what should be considered in the design of the comparison of methods study.

8.1 Test Specimens

Test specimens should be obtained from each patient in large enough amounts to complete testing with the test method and comparative method. The introduction of any bias in obtaining multiple specimens at one time should be resolved. Efforts must be made to ensure that the specimen(s) are typical of those that will be routinely tested and that they remain stable before testing. Some specimens should be fresh, while other specimens should not be as fresh as possible, e.g., blood spots. When possible, testing by both methods should be done at nearly the same time to minimize bias resulting from age differences between the specimens at the time of testing.

8.2 Number of Specimens

The total number of specimens to be tested during the study depends on the evaluator's intended use of the data. If, for instance, a total of 100 routine, incoming specimens is tested and the disease prevalence in the laboratory's population is 5%, there will be about five specimens that have positive results and about 95 specimens that have negative results. This might be a good sample size for evaluating the rate of false-positive results in the negative population (1-specificity), but it is probably not a sufficient number of specimens with positive results for evaluating the number of false-negative results in the positive population (1-sensitivity). The options available to the evaluator are to test specimens until the desired number of positive and negative results with the comparative method are obtained, or to test specimens until a desired number of specimens with positive results are obtained with the comparative method using all specimens with negative results tested along the way in the analysis. As a minimum guideline, testing should continue until at least 50 positive specimens are obtained with the comparative method. At least 50 negative specimens are to be obtained using the comparative method to determine the specificity of the test method.

The evaluator should consult a statistician to determine the number of specimens to test to meet the evaluator’s requirements for acceptable statistical variation. It is also important to test enough (representative) specimens to capture the biological variation in subjects with and without disease. See Section 8.7 for more on this topic.

8.3 Duration

The comparison of methods study with clinical specimens should be conducted daily for 10 to 20 days. Spreading the specimen testing over a number of days allows the user to obtain a representative number of specimens and to evaluate the test method under the typical laboratory usage. If feasible, all specimens tested during the evaluation should be properly stored and saved for additional testing, if necessary, to resolve questionable results. This additional testing with the same comparative method, another comparative method, or use of the clinical diagnosis might provide information to explain differences in test method results.

8.4 Inspection of Data During Collection

All data should be recorded and examined immediately to allow early detection of any sources of analytical system or human errors. If it is determined that some of the results are due to explainable error, the error condition must be noted and the data not included in the data analysis. If a reason for a discrepancy cannot be determined, retain the original results in the data set.

8.5 Discrepant Results

Discrepancies between the test and comparative methods may arise due to errors in the test method or due to errors in a comparative method that is not 100% accurate. When the comparative method is not 100% accurate, specimens with discrepant results between the test and comparative methods can be tested by a "gold standard" (reference method) to provide results that could be useful in resolving the discrepancy. For those tests with numerical values that are converted to qualitative results, a data listing of the test and comparative method results should be examined to investigate whether the discrepant results are near the test or comparative method cutoff point. Numerical values may also be analyzed to determine the difference between the results for test and comparative method specimens. The patient's clinical diagnosis and other clinical information for the specimens should be reviewed to determine if there is a predominant clinical condition among the specimens with discrepant results that should be investigated further.

When the comparative method is not 100% accurate, retesting discrepant results only is generally not sufficient for determining statistically valid estimates of sensitivity and specificity (unless all of the results are discrepant and retested by a 100% accurate method).[7-11] In order to estimate sensitivity and specificity in this situation, at least some concordant specimens need to be retested in addition to the discordant specimens. The number of samples that need to be retested depends on several factors that are unique to each setting. These factors include the desired precision of estimated sensitivity and specificity; the prevalence of the given disease or condition; and the correlation between the test method, the comparative method, and clinical diagnosis.[12,13] These methods require careful statistical planning and data analysis. Approaches for describing test performance that are not based on discrepant resolution are described in Section 9.[1,2,3]

8.6 Reference Specimen Panels

Clinical specimens that were previously tested, or referenced to well-defined methods or the clinical diagnosis of interest, are useful for the evaluation of qualitative methods. These reference panels or challenge panels are desirable in that the specimen's true value is well established and traceable to a well-characterized method or clinical diagnosis. The reference panel may be recognized by government agencies, regulatory agencies, industry, professional societies, or literature citations.

The panel should contain clinical specimens with different concentrations of the analyte of interest. A range of specimens that contains substances that can interfere with the test should be included in the panel if possible. Substances that can interfere can vary depending on the qualitative test being evaluated. False-negative or false-positive results can result from a number of conditions, such as autoimmune disease, spirochetal disease, heterophilic antibodies, rheumatoid arthritis, multiple myeloma, and others.

Although reference panels may be advantageous in the effective and efficient use of the evaluator's resources and time, such panels might not be available. Further, the use of reference panels is limiting in that the evaluator does not evaluate the performance of the test in the laboratory's own clinical population representing the typical prevalence and spectrum of the disease or condition, which precludes the use of predictive test values as performance indicators. Reference panels, as well as proficiency samples, used in conjunction with routine clinical specimen testing can, however, provide significant credibility to the evaluation process.

8.7 Clinical Diagnosis

It is beyond the scope of this document to provide detailed information about the determination of clinical sensitivity and specificity. The reader should follow the most current edition of NCCLS document GP10—*Assessment of the Clinical Accuracy of Laboratory Tests Using Receiver Operating*

Characteristic (ROC) Plots, which describes the assessment of the accuracy of a test compared to the clinical status of the patient.

As an overview, the following clinical parameters should be addressed:

- The patient specimens used for the comparison of method studies should include a representative population of clinical states expected in the clinical practice. A reasonable mix of patients according to age and gender should be obtained. The intent is to obtain patient specimens typical of future usage of the test.

- As stated in NCCLS document GP10—*Assessment of the Clinical Accuracy of Laboratory Tests Using Receiver Operating Characteristic (ROC) Plots,* the establishment of the true clinical state (case definition) of each patient requires that the criteria for each clinical state be accurate.

- Clinical information may be useful in resolving test result differences between the test method and the comparative method.

Although method comparison studies for the evaluation of qualitative tests are commonly found in the literature and product inserts, the clinical diagnosis remains the "gold standard" against which test results must be compared.

9 Data Analysis

9.1 Diagnosis is Known

The performance of truly qualitative tests is most commonly described in terms of "sensitivity" and "specificity." The calculation of these two quantities is most easily done when the true diagnosis of each patient specimen has been confirmed by clinical information independent of the biochemical tests being compared. Table 1 is a 2 x 2 contingency table that compares results of a qualitative test with the known true diagnosis of the specimens. The entry in each cell of the table represents the number of specimens corresponding to the labels in the margins. Following the table is a description of the calculation of estimated sensitivity, specificity, predictive value, and efficiency of the test.

Table 1. 2 x 2 Contingency Table

Method X	**True Diagnosis**		
	Positive	**Negative**	**Total**
Positive	A	B	A + B
Negative	C	D	C + D
Total	A + C	B + D	N

$$\textit{Estimated Sensitivity}\,(\textit{sens}) = 100\%\left[A/\left(A+C\right)\right]$$
$$\textit{Estimated Specificity}\,(\textit{spec}) = 100\%\left[D/\left(B+D\right)\right]$$

$$\textit{Based on evaluation specimens with disease prevalence} = 100\%\left(A+C\right)/N\ ,$$

$$\textit{Study Predictive Value of a Positive Test Result (PVP)} = (100\%)\left[A/(A+B)\right]$$
$$\textit{Study Predictive Value of a Negative Test Result (PVN)} = 100\%\left[D/(C+D)\right]$$

$Estimated\ Efficiency = 100\%[(A+D)/N]$, is the total agreement of the test results with the true diagnosis and is represented by the percent of all results that are true results, whether positive or negative.

The study PVP, study PVN, and estimated efficiency are all functions of estimated sensitivity, specificity, and estimated disease prevalence. Therefore, study PVP, study PVN, and estimated efficiency are meaningful for a particular patient population only when the disease prevalence of the evaluation specimens is the same as the patient population of interest. Even if the prevalence is the same, efficiency is meaningful only when false-positive (1-specificity) and false-negative (1-sensitivity) results are equally undesirable.

When a qualitative test is derived from dichotomizing a quantitative or ordinal scale, the performance of the test is best described through an ROC plot. The reader should follow the most current edition of NCCLS document GP10—*Assessment of the Clinical Accuracy of Laboratory Tests Using Receiver Operating Characteristic (ROC) Plots*.

9.1.1 Data Analysis

Both the test method and the comparative method are independently compared to the clinical diagnosis. The formulas above can be used to estimate the sensitivity, specificity, etc., for each method. These estimated performance measures are generalizable (unbiased) for the laboratory's expected test performance only to the extent that the evaluation specimens are typical of the specimens analyzed in the laboratory. Users should refer to the most current edition of NCCLS document GP10—*Assessment of the Clinical Accuracy of Laboratory Tests Using Receiver Operating Characteristic (ROC) Plots,* which describes how to select representative study subjects. Estimated performance measures are also subject to variability, because a (random) selection of specimens is used in the evaluation. This variability can be quantified by confidence limits, which decrease as the number of specimens evaluated increases.

There are several different methods available for calculating confidence limits for sensitivity and specificity. A common, simple method found in most textbooks is based on the normal approximation to the binomial distribution. However, this method is valid only when the data are normal. In addition, common "rules of thumb" for when this method is valid are not always sufficient.

Exact confidence limits (Clopper-Pearson method) for sensitivity and specificity can be computed from the binomial distribution and can be calculated by many statistical software packages, calculated by hand using F-tables, or obtained from published tables.[14,15,16] Users that have the capability of calculating exact confidence limits may use those. However, exact limits tend to be conservative (i.e., too wide) in some situations. Alternatively, Altman, et al.[17] and Agresti and Coull[16] recommend a direct calculation method called the *score confidence interval*, attributed to Wilson,[18] that can be applied in all cases. Rather than provide the details and statistical arguments for when the above methods are not appropriate, which the subcommittee believes would not be of interest to most readers, the use of score confidence limits is recommended and is described below.

A 95% score confidence interval for sensitivity or specificity is calculated as:

$$[100\%(Q_1 - Q_2)/Q_3, 100\%(Q_1 + Q_2)/Q_3],$$

where the quantities Q_1, Q_2, and Q_3 are computed from the data using the formulas below.

For sensitivity,

$$Q_1 = 2A + 1.96^2 = 2A + 3.84$$

$$Q_2 = 1.96\sqrt{1.96^2 + 4AC/(A+C)} = 1.96\sqrt{3.84 + 4AC/(A+C)}$$
$$Q_3 = 2(A + C + 1.96^2) = 2(A+C) + 7.68$$

For specificity,

$$Q_1 = 2D + 1.96^2 = 2D + 3.84$$
$$Q_2 = 1.96\sqrt{1.96^2 + 4BD/(B+D)} = 1.96\sqrt{3.84 + 4BD/(B+D)}$$
$$Q_3 = 2(B + D + 1.96^2) = 2(B+D) + 7.68$$

In the formulas above, 1.96 is the quantile from the standard normal distribution that corresponds to 95% confidence.

If the estimated sensitivity and specificity (performance) of the qualitative test is acceptable to the user, additional data analysis might not be necessary. However, the user might want to determine whether there is a statistical difference in the performance of the two methods. When a qualitative test is derived from dichotomizing a quantitative or ordinal scale, the evaluation is done by comparing the respective ROC plots. ROC plots can help one distinguish between differences in sensitivity and specificity due to choice of cutoff (the test method and comparative method performance represent two different points on the same ROC plot) versus real differences in diagnostic performance (the test method and comparative method have two different ROC plots). It is beyond the scope of this document to provide information about comparing ROC plots. The reader should follow NCCLS document GP10—*Assessment of the Clinical Accuracy of Laboratory Tests Using Receiver Operating Characteristic (ROC) Plots,* which provides literature references for how to compare the diagnostic ability of two tests.

In general, a comparison of sensitivities (specificities) alone when the corresponding true specificities (sensitivities) differ is an arbitrary one, since changing the cutoff can always increase the sensitivity (specificity) at the expense of lowering specificity (sensitivity). However, a joint comparison of the sensitivity/specificity pairs may still be meaningful if the sensitivity and specificity of one method are *both* larger than the sensitivity and specificity of the other method. When the sensitivity of one method is better than another, but its specificity is worse, it is not obvious which test is better. Biggerstaff[19] provides useful insight into how to compare methods in this situation.

McNemar's test[20] is commonly used to conclude whether there is a statistically significant difference between the sensitivity/specificity pairs of the two methods. This statistical test makes no assumption that one method is superior to the other in the clinical performance; they are both considered subject to the possibility of diagnostic error. However, this test does not provide information about the magnitude of possible differences. Instead, confidence intervals for the difference between sensitivities and for the difference between specificities may be more useful.

For the study design used here, the data are considered "paired," because the same specimens were tested by both the test method and the comparative method. In order to compute the appropriate confidence intervals or correct McNemar test statistic for this type of study design, the data need to be reported as a three-way comparison between the test method, the comparative method, and clinical diagnosis. For example, Table 2 below provides a comparison of the test method results to the comparative results when the true diagnosis is *positive* (for comparison of sensitivities) and when the true diagnosis is *negative* (for comparison of specificities).

Table 2. A Three-Way Comparison Between the Test Method, the Comparative Method, and Clinical Diagnosis

Method Result		Total Specimens	True Diagnosis	
Test Method	**Comparative Method**		**Positive**	**Negative**
Positive	**Positive**	$a = a_1 + a_2$	a_1	a_2
Positive	**Negative**	$b = b_1 + b_2$	b_1	b_2
Negative	**Positive**	$c = c_1 + c_2$	c_1	c_2
Negative	**Negative**	$d = d_1 + d_2$	d_1	d_2
Total		N	n_1	n_2

Note that the data in Table 2 can be used to construct two tables in the form of Table 1 (one for the test method, the other for the comparative method), but the converse is not true. Table 1 (A, B, C, and D) for the test method can be obtained from Table 2 using the following formulas:

$$A = a_1 + b_1$$
$$B = a_2 + b_2$$
$$C = c_1 + d_1$$
$$D = c_2 + d_2$$
$$N = n_1 + n_2$$

From Table 2, the estimated sensitivity of the (new) test method is:

$$sens_{new} = 100\%\left[\left(a_1 + b_1\right)/\, n_1\right],$$

the estimated sensitivity of the (old) comparative method is:

$$sens_{old} = 100\%\left[\left(a_1 + c_1\right)/\, n_1\right],$$

and the estimated difference is:

$$sens_{new} - sens_{old} = 100\%\left[\left(b_1 - c_1\right)/\, n_1\right].$$

Similarly, the respective estimated specificities are:

$$spec_{new} = 100\%\left[\left(c_2 + d_2\right)/\, n_2\right],$$
$$spec_{old} = 100\%\left[\left(b_2 + d_2\right)/\, n_2\right],$$

and the estimated difference is:

$$spec_{new} - spec_{old} = 100\%\left[\left(c_2 - b_2\right)/\, n_2\right].$$

Approximate confidence limits for the underlying difference between sensitivities and specificities are the standard statistical formulas for the difference between paired proportions.[20,21]

However, these confidence limits are approximate and may not be reliable, especially when the total number of specimens where the two tests disagree is small. Therefore, confidence limits described in Altman, et al.[17] (due to Newcombe[21]) are recommended, that can be used in all situations. These limits can be calculated directly and are described below.

Note that if a different study design were used where a different set of specimens had been used to evaluate the old test independently from the new test, then the sensitivities and specificities could be compared using the standard textbook formulas for comparing two independent proportions, and a three-way comparison of results (such as in Table 2) would not be applicable.

The 95% confidence interval for the difference between paired sensitivities, $D = sens_{new} - sens_{old}$ is:

$$\left(D - \sqrt{Q_5}\,, D + \sqrt{Q_6}\right)$$

where Q_5 and Q_6 are computed from the data using the formulas below. First, compute separate score confidence intervals for $sens_{new}$ and for $sens_{old}$, using the formulas at the beginning of Section 9.1.1, and then compute quantities Q_1 through Q_6 below.

l_1=lower limit of 95% score confidence interval for $sens_{new}$
u_1=upper limit of 95% score confidence interval for $sens_{new}$

l_2=lower limit of 95% score confidence interval for $sens_{old}$
u_2=upper limit of 95% score confidence interval for $sens_{old}$

$Q_1 = (a_1+b_1)(c_1+d_1)(a_1+c_1)(b_1+d_1)$ *(If Q_1=0, then Q_4=0. Go to Q_5.)*

$Q_2 = a_1d_1 - b_1c_1$

$$\begin{array}{lll} Q_3 = & Q_2 - n_1/2 & \text{if } Q_2>n_1/2 \\ Q_3 = & 0 & \text{if } 0 \le Q_2 \le 0 \\ Q_3 = & Q_2 & \text{if } Q_2<0 \end{array}$$

$Q_4 = Q_3/\sqrt{Q_1}$ $(Q_4=0 \text{ if } Q_1=0)$

$$Q_5 = (sens_{new} - l_1)^2 - 2Q_4(sens_{new} - l_1)(u_2 - sens_{old}) + (u_2 - sens_{old})^2$$

$$Q_6 = (sens_{old} - l_2)^2 - 2Q_4(sens_{old} - l_2)(u_1 - sens_{new}) + (u_1 - sens_{new})^2$$

The 95% confidence interval for the difference between paired specificities, $D= spec_{new}–spec_{old}$ is similarly computed as:

$$\left(D - \sqrt{Q_5}\,, D + \sqrt{Q_6}\right)$$

where Q_5 and Q_6 are computed from the data using the formulas below. First, compute separate score confidence intervals for $spec_{new}$ and for $spec_{old}$, using the formulas in Section 9.1.1, and then compute quantities Q_1 through Q_6 below.

l_1=lower limit of 95% score confidence interval for $spec_{new}$
u_1=upper limit of 95% score confidence interval for $spec_{new}$

l_2=lower limit of 95% score confidence interval for $spec_{old}$
u_2=upper limit of 95% score confidence interval for $spec_{old}$

$Q_1 = (a_2+b_2)(c_2+d_2)(a_2+c_2)(b_2+d_2)$ *(If Q_1=0, then Q_4=0. Go to Q_5.)*

$$Q_2 = a_2 d_2 - b_2 c_2$$

$$\begin{array}{lll} Q_3 = & Q_2 - n_2/2 & \text{if } Q_2 > n_2/2 \\ Q_3 = & 0 & \text{if } 0 \le Q_2 \le 0 \\ Q_3 = & Q_2 & \text{if } Q_2 < 0 \end{array}$$

$$Q_4 = Q_3/\sqrt{Q_1} \quad (Q_4 = 0 \text{ if } Q_1 = 0)$$

$$Q_5 = (spec_{new} - l_1)^2 - 2Q_4(spec_{new} - l_1)(u_2 - spec_{old}) + (u_2 - spec_{old})^2$$

$$Q_6 = (spec_{old} - l_2)^2 - 2Q_4(spec_{old} - l_2)(u_1 - spec_{new}) + (u_1 - spec_{new})^2$$

9.1.2 Special Comment on Prevalence and Predictive Values

Prevalence is the frequency of a given disease or condition in a specified group or population expressed as a fraction (percent or decimal). These guidelines are designed for evaluation of the test in the user's patient population so that the prevalence for the evaluation and the prevalence expected during the typical usage of the test are similar.

The predictive value of a test combines disease prevalence with test sensitivity and specificity. The predictive value of a positive test result is the proportion of patients testing positive who have the disease. It is calculated as the number of true-positive results divided by the number of all positive test results (true-positive and false-positive results combined). The predictive value of a negative test result is the proportion of patients testing negative do not who have the disease. It is calculated as the number of true-negative test results divided by the number of all negative test results (true-negative and false-negative results combined). The number of the true-positive, true-negative, false-positive, and false-negative results is a function of the prevalence in the population, and the sensitivity and specificity of the test in question. The prevalence for the evaluation is significant in the determination of the test's negative and positive predictive values. The predictive values apply only to those defined populations that have a similar prevalence and spectrum of disease to that used to create the estimate of the predictive value.

9.2 Diagnosis is Not Known

In this common situation, sensitivity and specificity cannot be readily estimated. Many researchers have proposed that the sensitivity and specificity of a test method may still be estimated from fairly simple formulas by assuming that the sensitivity and specificity of the comparative method are known to a close approximation from past experience. However, such methods are typically based on the additional assumption that the test method and the comparative method are "conditionally independent." That is, if the test method's error rate among true diagnosed positives is the same for the comparative method positives and comparative method negatives, and similarly among true diagnosed negatives, then the test method and the comparative method are conditionally independent. In practice, this assumption is not easily verified and is often unrealistic. Alternatively, reporting how often the test method and comparative method agree could be useful.

9.2.1 Sensitivity and Specificity Corrected for Errors in the Comparative Method

Estimates of sensitivity and specificity corrected for errors in the comparative method are typically based on the conditional independence assumption. This assumption can be evaluated from data in the form of Table 2 above. The test method's estimated error rate among true diagnosed positives for the comparative method positives is $c_1/(a_1 + c_1)$, and for the comparative method negatives is $d_1/(b_1 + d_1)$. Conditional independence assumes that the true error rates estimated by these quantities are the same. Similarly, the test method's estimated error rate among true diagnosed negatives for the comparative method positives is $a_2/(a_2 + c_2)$, and for the comparative method negatives is $b_2/(b_2 + d_2)$. Conditional independence also assumes that the true error rates estimated by these quantities are the same. Fisher's exact test for comparing two proportions could be used to test these hypotheses, if such data were available. However, the reason for making an adjustment in the first place is that the true diagnosis is unknown!

In the event that there is data to support the conditional independence assumption, then estimates of sensitivity and specificity of the test method can be obtained for assumed values of the sensitivity and specificity of the comparative method. Formulas are provided in the literature.[22,23]

Alternatively, Thibodeau[24] provides bounds for corrected estimates of sensitivity and specificity that do not depend on the conditional independence assumption. However, this approach assumes that the comparative method sensitivity and specificity are both at least as large as the sensitivity and specificity of the test method. Again, this assumption may not be realistic or easy to verify.

9.2.2 Agreement Between the Test Method and the Comparative Method

When the diagnosis is not known, reporting a 2 x 2 table of results and how often the test method and comparative method agree may be useful. An example of how to present the results is shown in Table 3. Note that summing the results in the last two columns of Table 2 gives the results in Table 3.

Table 3. 2 x 2 Contingency Table When True Diagnosis is Unknown

Test Method	Comparative Method		
	Positive	**Negative**	**Total**
Positive	a	b	a + b
Negative	c	d	c + d
Total	a + c	b + d	n

There are many different statistical measures of agreement. A discussion by M.M. Shoukri, on different types of agreement measures, can be found under "Agreement, Measurement of" in the Encyclopedia of Biostatistics.[25] A simplistic approach is to report the percent agreement, which is given below.

$$\textit{Percent agreement} = 100\%\left[\left(a+d\right)/n\right]$$

Since agreement on absence of disease does not provide direct information about agreement on presence of disease, it may be useful to report two additional measures of agreement.

$$\textit{Agreement of test method with comparative method-positive} = 100\%\,[\,a/(a+c)]$$

$$\textit{Agreement of test method with comparative method-negative} = 100\%\,[\,d/(b+d)]$$

Caution must be used when generalizing agreement measures to any other population, since the disease prevalence for the evaluation specimens (typically unknown in this case) can grossly affect the agreement measures. As a hypothetical example, suppose that the test method and the comparative method agree closely when the true diagnosis is negative, but they do not agree very well when the true diagnosis is positive. The overall agreement between the two methods will be higher in a study where the evaluation specimens have low disease prevalence, and lower in a study where the evaluation specimens have high disease prevalence. If the disease prevalence is unknown, then it is unclear how to generalize the agreement measure to a population with a different prevalence.

An exact confidence interval for percent agreement can be computed as discussed in Section 9.1.1.

A 95% score confidence interval for agreement is calculated as:

$$[100\%(Q_1 - Q_2)/Q_3, 100\%(Q_1 + Q_2)/Q_3],$$

where the quantities Q_1, Q_2, and Q_3 are computed from the data using the formulas below.

$$Q_1 = 2(a+d) + 1.96^2 = 2(a+d) + 3.84$$
$$Q_2 = 1.96\sqrt{1.96^2 + 4(a+d)(b+c)/n} = 1.96\sqrt{3.84 + 4(a+c)(b+d)/n}$$
$$Q_3 = 2(n + 1.96^2) = 2n + 7.68$$

In the formulas above, 1.96 is the quantile from the standard normal distribution that corresponds to 95% confidence.

Confidence intervals for agreement of test method with comparative method-positive and agreement of test method with comparative method-negative are not easily formulated, because the imperfect standard results are subject to variability and the nature of the variability depends on unknown factors.

9.2.3 Special Case: When Sensitivity and Specificity of Comparative Method are Perfect (=100%)

In this case, Table 3 is equivalent to Table 1, and the sensitivity and specificity of the test method can be estimated directly using the formulas described in Section 9.1.

9.3 Examples

9.3.1 Diagnosis is Known

Data are presented that will enable the test method and the comparative method to be compared individually to an independently obtained diagnosis. From the publication of Meijer, et al.[26] an example is shown for eight commercial enzyme-linked immunosorbent assays (ELISA) used to test sera taken from 102 patients in whom *Helicobacter pylori* infection status had been determined.

The performance of the test method and comparative method was determined with the true diagnosis of each patient specimen. A 2 x 2 contingency table of the results from the comparative and test methods are listed below as Examples 1a and 1b. The formulas listed in Section 9.1 are used to estimate the sensitivity, specificity, etc. for each method.

Example 1a. 2 × 2 Contingency Table for Test Method Versus True Diagnosis

		True Diagnosis: *H. pylori*		
		Positive	**Negative**	**Total**
Test Method	**Positive**	57	2	59
	Negative	4	39	43
	Total	61	41	102

$$Estimated\ Sensitivity\ (sens) = 100\%[A/(A+C)] = 100\%[57/61] = 93.4\%$$
$$Estimated\ Specificity\ (spec) = 100\%[D/(B+D)] = 100\%[39/41] = 95.1\%$$

$$\text{Based on evaluation specimens with disease prevalence} = 100\%(A+C)/N = 100\%(61/102) = 59.8\%$$

$$\begin{aligned}\text{Study Predictive Value of a Positive Test Result } (PVP) &= (100\%)[A/(A+B)] \\ &= 100\%(57/59) = 96.6\%\end{aligned}$$

$$\begin{aligned}\text{Study Predictive Value of a Negative Test Result } (PVN) &= 100\%[D/(C+D)] \\ &= 100\%(39/43) = 90.7\%\end{aligned}$$

$$\text{Estimated Efficiency} = 100\%[(A+D)/N] = 100\%(96/102) = 94.1\%$$

Sensitivity for Test Method:

$$sens = 100\%(57/61) = 93.4\%$$

Exact 95% confidence limits are 84.1% to 98.2%.

95% score confidence limits are 84.3% to 97.4%.

Calculations for 95% score confidence limits:

$$Q_1 = 2 \times 57 + 3.84 = 117.84$$

$$Q_2 = 1.96\sqrt{3.84 + 4 \times 57 \times 4/61} = 8.496$$

$$Q_3 = 2 \times 61 + 7.68 = 129.68$$

$$100\%(Q_1 - Q_2)/Q_3 = 100\%(117.84 - 8.496)/129.68 = 84.3\%$$

$$100\%(Q_1 + Q_2)/Q_3 = 100\%(117.84 + 8.496)/129.68 = 97.4\%$$

Specificity for Test Method:

$$spec = 100\%(39/41) = 95.1\%$$

Exact 95% confidence limits are 83.5% to 99.4%.

95% score confidence limits are 84.6% to 98.7%.

Calculations for 95% score confidence limits:

$$Q_1 = 2 \times 39 + 3.84 = 81.84$$

$$Q_2 = 1.96\sqrt{3.84 + 4 \times 39 \times 2/41} = 6.632$$

$$Q_3 = 2 \times 41 + 7.68 = 89.68$$

$$100\%(Q_1 - Q_2)/Q_3 = 100\%(81.84 - 6.632)/89.68 = 83.9\%$$

$$100\%(Q_1 + Q_2)/Q_3 = 100\%(81.84 + 6.632)/89.68 = 98.7\%$$

Example 1b. 2 × 2 Contingency Table for Comparative Method Versus True Diagnosis

		True Diagnosis: *H. pylori*		
		Positive	**Negative**	**Total**
Comparative Method	**Positive**	54	7	61
	Negative	7	34	41
	Total	61	41	102

$$\textit{Estimated Sensitivity (sens)} = 100\%[A/(A+C)] = 100\%[54/61] = 88.5\%$$
$$\textit{Estimated Specificity (spec)} = 100\%[D/(B+D)] = 100\%[34/41] = 82.9\%$$

$$\textit{Based on evaluation specimens with disease prevalence} = 100\%(A+C)/N = 100\%(61/102) = 59.8\%$$

$$\begin{aligned}\textit{Study Predictive Value of a Positive Test Result } (PVP) &= (100\%)[A/(A+B)]\\ &= 100\%(54/61) = 88.5\%\end{aligned}$$

$$\begin{aligned}\textit{Study Predictive Value of a Negative Test Result } (PVN) &= 100\%[D/(C+D)]\\ &= 100\%(34/41) = 82.9\%\end{aligned}$$

$$\textit{Estimated Efficiency} = 100\%[(A+D)/N] = 100\%(88/102) = 86.3\%$$

Note that in Example 1b, sensitivity = PVP and specificity = PVN. This will not always be true. This occurs here, and in general, whenever the total number of patients with a positive true diagnosis equals the total number of positive test results, and whenever the total number of patients with a negative true diagnosis equals the total number of negative test results, respectively.

Confidence limits for the estimated values are computed using the methods described in Section 9.1.1.

Sensitivity for Comparative Method:

$$sens = 100\%(54/61) = 88.5\%$$

Exact 95% confidence limits are 77.8% to 95.3%.

95% score confidence limits are 78.2% to 94.3%.

Calculations for 95% score confidence limits:

$$Q_1 = 2 \times 54 + 3.84 = 111.84$$

$$Q_2 = 1.96\sqrt{3.84 + 4 \times 54 \times 7/61} = 10.487$$

$$Q_3 = 2 \times 61 + 7.68 = 129.68$$

$$100\%\left(Q_1 - Q_2\right)/\,Q_3 = 100\%\left(111.84 - 10.487\right)/\,129.68 = 78.2\%$$

$$100\%\left(Q_1 + Q_2\right)/\,Q_3 = 100\%\left(111.84 + 10.487\right)/\,129.68 = 94.3\%$$

Specificity for Comparative Method:

$$spec = 100\%\left(34\,/\,41\right) = 82.9\%$$

Exact 95% confidence limits are 67.9% to 92.8%.

95% score confidence limits are 68.7% to 91.5%.

Calculations for 95% score confidence limits:

$$Q_1 = 2\times 34 + 3.84 = 71.84$$

$$Q_2 = 1.96\sqrt{3.84 + 4\times 34\times 7\,/\,41} = 10.196$$

$$Q_3 = 2\times 41 + 7.68 = 89.68$$

$$100\%\left(Q_1 - Q_2\right)/\,Q_3 = 100\%\left(71.84 - 10.196\right)/\,89.68 = 68.7\%$$

$$100\%\left(Q_1 + Q_2\right)/\,Q_3 = 100\%\left(71.84 + 10.196\right)/\,89.68 = 91.5\%$$

The publication used ROC analysis to describe the performance of each assay, and could have compared the correlated ROC plots. Alternatively, a joint statistical comparison of the sensitivity/specificity pairs would be meaningful here, since the estimated sensitivity and specificity for the new test method are *both* larger than the estimated sensitivity and specificity for the old comparative method. In order to compare the sensitivities and specificities between the two methods, we need the three-way comparison between old, new, and true diagnosis. These comparative results are indirectly provided in Table 4 of the publication and are redisplayed as Example 1c.

Example 1c. A Three-Way Comparison Between the New Test Method, the Old Comparative Method, and True Diagnosis

Method Result		**Total Specimens**	**True Diagnosis**	
Test Method	**Comparative Method**		**Positive**	**Negative**
Positive	**Positive**	55	53	2
Positive	**Negative**	4	4	0
Negative	**Positive**	6	1	5
Negative	**Negative**	37	3	34
Total		102	61	41

Comparison of sensitivities:

$$D = sens_{new} - sens_{old} = 93.4 - 88.5\,or 100\%\cdot\left(4 - 1\right)/\,61 = 4.9\%$$

l_1= 84.3% (from score confidence limit for $sens_{new}$)
u_1= 97.4%

l_2 = 78.2% *(from score confidence limit for sens$_{old}$)*
u_2 = 94.3%

$$Q_1 = (53+4)(1+3)(4+3) = (57)(4)(54)(7) = 86{,}184$$

$$Q_2 = (53 \cdot 3) - (4 \cdot 1) = 155$$

$$n_1 / 2 = 61/2 = 30.5 < 155 = Q_2$$

$$Q_3 = Q_2 - n_1 / 2 = 155 - 30.5 = 124.5$$

$$Q_4 = Q_3 / \sqrt{Q_1} = 124.5 / \sqrt{86{,}184} = 124.5 / 293.57 = 0.4241$$

$$Q_5 = (93.4 - 84.3)^2 - 2(0.4241)(93.4 - 84.3)(94.3 - 88.5) + (94.3 - 88.5)^2 = 71.68$$

$$Q_6 = (88.5 - 78.2)^2 - (0.4241)(88.5 - 78.2)(97.4 - 93.4) + (97.4 - 93.4)^2 = 87.14$$

$$D - \sqrt{Q_5} = 4.9 - \sqrt{71.68} = -3.6$$

$$D + \sqrt{Q_6} = 4.9 + \sqrt{87.14} = 14.2$$

The 95% confidence interval for $D = sens_{new} - sens_{old}$ is *(–3.6%, 14.2%)*

Comparison of specificities:

$$D = spec_{new} - spec_{old} = 95.1 - 82.9 \, or \, 100\% \cdot (5 - 0) / 41 = 12.2\%$$

l_1 = 83.9% *(from score confidence limit for spec$_{new}$)*
u_1 = 98.7%

l_2 = 68.7% *(from score confidence limit for spec$_{old}$)*
u_2 = 91.5%

$$Q_1 = (2+0)(5+34)(2+5)(0+34) = (2)(39)(7)(34) = 18{,}564$$

$$Q_2 = (2 \cdot 34) - (0 \cdot 5) = 68$$

$$n_2 / 2 = 41/2 = 20.5 < 68 = Q_2$$

$$Q_3 = Q_2 - n_1 / 2 = 68 - 20.5 = 47.5$$

$$Q_4 = Q_3 / \sqrt{Q_1} = 47.5 / \sqrt{18{,}564} = 47.5 / 136.25 = 0.3486$$

$$Q_5 = (95.1 - 83.9)^2 - 2(0.3486)(95.1 - 83.9)(91.5 - 82.9) + (91.5 - 82.9)^2 = 132.25$$

$$Q_6 = (82.9 - 68.7)^2 - 2(0.3486)(82.9 - 68.7)(98.7 - 95.1) + (98.7 - 95.1)^2 = 178.96$$

$$D - \sqrt{Q_5} = 12.2 - \sqrt{132.25} = 0.7$$
$$D + \sqrt{Q_6} = 12.2 + \sqrt{178.96} = 25.6$$

The 95% confidence interval for $D{=}spec_{new}{-}spec_{old}$ is *(0.7%, 25.6%)*

Since the confidence limits for the difference in sensitivities include zero, we cannot conclude that the sensitivities are statistically different. However, the limits for specificity do not include zero, so there is evidence that the specificities are statistically different (the difference could be quite small or as large as 25%).

9.3.2 Diagnosis is Not Known

A typical situation may be that the evaluator does not know the diagnosis but wants to compare the test method results to another method. An example from a publication of Schrier, et al.[27] of evaluation results from two *H. pylori* antibody tests is presented. The test method was an immunochromatographic method, while the comparative method was an ELISA (HM-CAP EIA) method. The 2 x 2 contingency table of the evaluation results is shown below.

HM-CAP EIA Comparative Method

		Positive	**Negative**	**Total**
Immunochromatic Method Test Method	**Positive**	285	15	300
	Negative	14	222	236
	Total	299	237	536

$$Percent\ agreement = 100\% \cdot (a + d) / n = 100\% \cdot 507 / 536 = 94.6\%$$

Agreement of test method with HM-CAP EIA-positive = 100%×285/299 = 95.3%

Agreement of test method with HM-CAP EIA-negative = 100%×222/237 = 93.7%

The confidence limit for percent agreement is calculated using the methods described in Section 9.2.2.

Exact 95% confidence limits are 92.3% to 96.4%.

95% score confidence limits are 92.3% to 96.2%.

Calculations for 95% score confidence limits:

$$Q_1 = 2 \times 507 + 3.84 = 1{,}017.84$$

$$Q_2 = 1.96\sqrt{3.84 + 4 \times 507 \times 29 / 536} = 20.887$$

$$Q_3 = 2 \times 536 + 7.68 = 1{,}079.68$$

$$100\%(Q_1 - Q_2) / Q_3 = 100\%(1017.84 - 20.887) / 1079.68 = 92.3\%$$

$$100\%(Q_1 + Q_2) / Q_3 = 100\%(1017.84 + 20.887) / 1079.68 = 96.2\%$$

References

[1] European Committee for Clinical Laboratory Standards. *Guidelines for the Evaluation of Diagnostic Kits. Part 2. General Principles and Outline Procedures for the Evaluation of Kits for Qualitative Tests.* Lund, Sweden: ECCLS; 1990.

[2] Berg B, Hellsing K, Jagenburg R, Kallner A. Guidelines for evaluation of reagent strips. Exemplified by analysis of urine albumin and glucose concentration using visually read reagent strips. *Scand J Clin Lab Invest.* 1989;49:689-699.

[3] Thorn RM, Braman V, Stella M, Yi A. Assessment of HIV-I screening test sensitivities using serially diluted positive sera can give misleading results. *Transfusion.* 1989;29(1):78-80.

[4] Ross JW, Miller WG, Myers GL, Praestgaard J. The accuracy of laboratory measurements in clinical chemistry. A study of 11 routine chemistry analytes in the College of American Pathologists Chemistry Survey with fresh frozen serum, definitive methods, and reference methods. *Arch Pathol Lab Med.* 1998;122:587-608.

[5] Cattozzo G, Franzini C, d'Eril GV. Myoglobin and creatine kinase isoenzyme MB mass assays: intermethod behaviour of patient sera and commercially available control materials. *Clin Chim Acta*. 2001;303:55-60.

[6] Sokoll LJ, Witte DL, Klee GG, Chan DW. Redesign of proficiency testing materials improve survey outcomes for prostate-specific antigen. *Arch Pathol Lab Med*. 2000;124:1608-1613.

[7] Hagdu A. Bias in the evaluation of DNA-amplification tests for detecting *Chlamydia trachomatis*. *Statistics in Medicine.* 1997;16:1391-1399.

[8] Lipman HB, Astles JR. Quantifying the bias associated with use of discrepant analysis. *Clin Chem.* 1998;44;1:108-115.

[9] Miller WC. Bias in discrepant analysis: When two wrongs don't make a right. *Clin Epidemiol.* 1998;51:219-231.

[10] Hayden CL, Feldstein ML. Dealing with discrepancy analysis. Part 1: The problem of bias. IVD *Technology.* Jan/Feb 2000:37-42.

[11] Hayden CL, Feldstein ML. Dealing with discrepancy analysis. Part 2: Alternative analytical strategies. IVD *Technology.* March/April 2000:51-57.

[12] Alonzo TA, Pepe MS. Using a combination of reference tests to assess the accuracy of a new diagnostic test. *Statistics in Medicine.* 1999;18:2987-3003.

[13] Hawkins DM, Garrett JA, Stephenson B. Some issues in resolution of diagnostic tests using an imperfect gold standard. *Statistics in Medicine.* 2001;20:1987-2001.

[14] Armitage P, Berry G. *Statistical Methods in Medical Research*. 3rd ed. Cambridge: Blackwell Science; 1994.

[15] Zwillinger D, Kokoska S. *CRC Standard Probability and Statistics Tables and Formulae*. Boca Raton: Chapman & Hall/CRC; 2000:205-221.

[16] Agresti A, Coull BA. Approximate is better than “exact” for interval estimation of binomial proportions. *The American Statistician*. 1998;52(2):119-126.

[17] Altman DA, Machin D, Bryant TN, Gardner MJ, eds. *Statistics with Confidence.* 2nd ed. British Medical Journal; 2000.

[18] Wilson EB. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association.* 1927;22:209-212.

[19] Biggerstaff BJ. Comparing diagnostic tests: a simple graphic using likelihood ratios. *Statistics in Medicine.* 2000;19:649-663.

[20] Fleiss JL. *Statistical Methods for Rates and Proportions.* 2nd ed. New York, NY: John Wiley & Sons; 1981.

[21] Newcombe RG. Improved confidence intervals for the difference between binomial proportions based on paired data. *Statistics in Medicine.* 1998;17:2635-2650.

[22] Gart JJ, Buck AA. Comparison of a screening test and a reference test in epidemiologic studies. II: a probabilistic model for the comparison of diagnostic tests. *Am J Epidemiol.* 1966;83:593-602.

[23] Staquet M, Rozencweig M, Lee YJ, Muggia FM. Methodology for the assessment of new dichotomous diagnostic tests. *J Chron Dis.* 1981;34:599-610.

[24] Thibodeau LA. Evaluating diagnostic tests. *Biometrics.* 1981;37:801-804.

[25] Shoukri MM. “Agreement, Measurement of.” In Armitage P, Colton T, eds. *Encyclopedia of Biostatistics*. New York, NY: John Wiley & Sons; 1998:103-117.

[26] Meijer BC, Thijs JC, Kleibeuker JH, van Zwet AA, Berrelkamp RJP. Evaluation of eight enzyme immunoassays for detection of immunoglobin G against *Helicobacter pylori. J Clin Microbiol.* 1997;35(1):292-294.

[27] Schrier WH, et al. Development of FlexSure® HP – an immunochromatographic method to detect antibodies against *Helicobacter pylori*. *Clin Chem.* 1998;44(2):293-298.

Additional References

Dixon WJ, Massey FJ. *Introduction to Statistical Analysis*. 4th ed. New York, NY: McGraw-Hill; 1983.

Galen R, Gambino S. *Beyond Normality: The Predictive Value and Efficiency of Medical Diagnoses*. New York, NY: John Wiley & Sons; 1975.

NCCLS consensus procedures include an appeals process that is described in detail in Section 9 of the Administrative Procedures. For further information contact the Executive Offices or visit our website at www.nccls.org.

Summary of Comments and Subcommittee Responses

EP12-P: *User Protocol for Evaluation of Qualitative Test Performance; Proposed Guideline*

General

1. The guideline is well written and probably will be useful.

- **The subcommittee appreciates the compliment.**

2. The document as it stands seems to be a cross between a protocol that a manufacturer would perform to demonstrate kit performance for an FDA submission and testing that a laboratory would want to perform to verify the kit works in the environment. Since there are few standards available for *in vitro* diagnostics products, we believe this document would be a candidate for recognition by the FDA if the Scope of the document were not limited to the clinical laboratory and certain details were added to the protocols themselves. On the other hand, the studies presented in the document seem too extensive for the typical clinical laboratory to perform in order to verify the test system performs in the environment. This issue is apparent in the "Scope" and "Introduction" to the document. The "Introduction" indicates manufacturers, regulatory agencies, and laboratory surveyors, as well as the clinical laboratory should use this document. The "Scope" indicates that this document is written for "laboratory personnel who are the end users of such tests." This document contains many protocols which can be used by manufacturers during product performance claims testing, but which then should not have to be repeated by the end user. Clarification of this point would add to the usefulness of the document.

- **The Introduction and Foreword state that the purpose of this document is to provide uniform guidance among all users in the performance assessment of qualitative testing. The Scope identifies laboratory personnel as the end users of this protocol. The end users will decide dependent upon their needs if all sections of EP12 are necessary to complete.**

3. Basically, this "guideline" addresses the evaluation of qualitative test performance for dichotomous outcomes only (i.e., discrete binomial random variables). It does not address situations involving multiple outcomes such as PAP smear testing with multiple discrete outcomes (e.g., negative, ASCUS/AGUS, LSIL, HSIL, or carcinoma) or ANA testing where both patterns of staining and dilution titers with multiple discrete outcomes (e.g., negative, 1:40, 1:80, 1:160, 1:320, etc.) must be considered when making a determination for a patient. An addendum to this evaluation protocol should be included, since many diagnostic tests involve more than just two discrete outcomes.

- **EP12 was designed to provide users with a protocol for evaluation of *qualitative* test performance. It is outside the scope of the document to address semiquantitative tests.**

Section 7 (formerly Section 5)

4. Reproducibility Studies: The FDA has on various occasions suggested that the panel identification be blinded to the user and randomized during the testing.

- **Manufacturers' controls do not need to be blinded for this study. Controls are to be used by users as intended.**

Section 7.1 (formerly Section 5.1)

5. Please clarify and explain why the criteria of a control being out of spec on more than one run for the ten-day study (duplicate values) or more than two runs for the twenty-day study (single values) is significant. It is assumed that when using a significance level of 5%, it is expected that due to chance alone 5 out of 100 runs (or 1 out of 20 runs) would yield an incorrect result. This means that one replicate value out of all twenty values would be expected to yield an incorrect result for either the ten-day or twenty-day study. Thus, more than two replicate values indicates a 10% significance level of rejection which would be the basis for the criteria described on the top of page 5 of the document. Please clarify. Also, does this mean for the ten-day study, if both replicate values for a control versus only one replicate value not giving the expected result, that the run should be rejected? Please clarify.

- **As stated in the second paragraph of this section, "If either of the control materials does not give expected results, the run must be rejected." Selection of the criteria of no more than one run rejected in a ten-day study or two runs in a twenty-day study was a decision based on the fact that manufacturers' positive and negative controls are strong positive and negative levels and not near the cutoff of the test. Therefore, the controls not performing as expected is significant, and the laboratory should discontinue testing and consult with the kit manufacturer to identify the cause and implement corrective action.**

Section 7.2 (formerly Section 5.2)

6. These sections are used to define the assay detection limit, i.e., the sensitivity of the assay. In the case of a Strep A assay it was difficult to prepare and quantitate the organism due to the fact of comparing colony-forming units versus "dead" organisms.

- **The subcommittee agrees with the commenter; however, manufacturers do list detection limits.**

Section 7.3 (formerly Section 5.3)

7. What about when controls are spiked swabs? What about when liquid controls are not available?

- **When controls are spiked swabs, the analyte or measurand is the percentage of cells extracted for the assay compared to the number added to the swab. Controls are defined as needed.**

8. The procedure does not consider two important factors that may be critical in a reproducibility experiment with qualitative tests: (1) variability between different lots or batches of manufacturers' reagents, and (2) subjectivity of the user when visually reading results for those qualitative tests that do not use instrumentation to detect test results. Consider providing a note in Section 5.3 that these specific factors may impact test results, and direct the user to refer to Section 6 for more information. The statement in Section 6.4, "All data should be recorded and examined immediately to allow early detection of any sources of analytical system or human errors," is sufficient to alert the user that variability between manufacturers' reagents and subjectivity associated with visual detection of test results may impact test results.

- **These factors may be important but are facts and not included as part of the evaluation.**

9. The explanation about establishing the analyte cutoff concentration is a bit confusing. After re-reading the paragraph several times, I gather that there is some inherent value (signal) that establishes the result, either positive or negative, for the analyte in a sample. If the concentration of the analyte can, somehow, be established in a pool, then the pool could be diluted into a series of related concentrations that could be evaluated against the cutoff signal. However, there still seems to be some circular reasoning in this paragraph. I certainly could not use it, practically, to establish the cutoff

value for an analyte, for example, Rubella antibody in serum. Perhaps the committee could revisit the wording of this paragraph to provide a more enlightening description of how to establish cutoff values.

- **An introductory sentence with the purpose stated has been added.**

Section 7.3 (formerly Section 5.3.1)

10. More utility could be gained if an example justifying this section were included.

- **The subcommittee agrees with the commenter. An example has been added for justification.**

Section 8 (formerly Section 6)

11. For specimens such as swab specimens where it is possible to collect two swabs from a single subject, the two swabs may still have different amounts of bacteria even after diligent efforts to provide consistency in sampling. Consider suggesting some alternatives that would avoid the introduction of variability from the sample and lead to more objective comparisons of methods; e.g., use of a reference material or standard solution that could be used repeatedly.

- **The subcommittee is unaware of an alternative that would always avoid the variability from the sample. Using a standard solution may not work for all comparisons of methods.**

12. This section should allow the use of frozen specimens as a subset in comparison testing, assuming fresh versus frozen samples do not affect results.

- **Use of frozen specimens is acceptable as long as the test results are not affected.**

Section 8.2 (formerly Section 6.2)

13. This could be very cumbersome and expensive in a low prevalence area unless "positive" and "negative" can be determined by another method, and samples stored until ready to use.

- **This practice is acceptable if storage does not degrade the specimens. After storage, specimens should be tested by performing the comparative method and the test method at the same time.**

Section 8.5 (formerly Section 6.5)

14. Please clarify the statement, "Retesting discrepant results only is generally not sufficient for determining statistically valid estimates of sensitivity and specificity (unless all of the results are discrepant!)." Does this statement infer it is necessary to retest the concordant samples as well? How many samples are needed to result in a statistically valid estimate of sensitivity and specificity?

- **This statement does imply that some concordant specimens need to be retested in addition to the discordant specimens in order to estimate sensitivity and specificity. The number of samples that need to be retested depends on several factors that are unique to each setting. These factors include the desired precision of estimated sensitivity and specificity, the correlation between the tests, comparative method and clinical diagnosis, and the prevalence of the given disease or condition. These methods require careful statistical planning and data analysis. Approaches for describing test performance that are not based on discrepant resolution are described in Section 9. Section 8.5 has been expanded to clarify these points.**

15. Please expand this section to incorporate the analytical strategies outlined in the following reference articles:

    Hayden DL, Feldstein ML. Dealing with Discrepancy Analysis. Part 1: The Problem of Bias. IVD *Technology.* Jan/Feb 2000:37-42.

    Hayden DL, Feldstein ML. Dealing with Discrepancy Analysis. Part 2: Alternative Analytical Strategies. IVD *Technology.* March/April 2000:51-57.

- **References have been added throughout the guideline as appropriate.**

Section 9.1.1 (formerly Section 7.1.1)

16. Paragraph 6. The first sentence doesn't make grammatical sense. I believe the committee's intent is to describe the use of "A… McNemar test…." The word "the" immediately following the "(df)" should be deleted to make the sentence read correctly.

- **The text has been modified as suggested.**

17. The third paragraph states, "Exact confidence limits can be computed from the binomial distribution and are the preferred limits." There is no reference attached to this statement even though reference #7 does contain the information needed. I would hope this would be added after the statement to give the reader an idea of where to look for this information.

    Also, I think this statement needs to be a little stronger. In those cases where the estimated sensitivity or specificity is 100%, then the exact formulas must be used, as the approximate formulas give the false impression that there is no confidence interval associated with the estimate.

    As a whole the document is well written and provides a good discussion on how to deal with these types of tests.

- **This section has been modified. A reference has been cited as the source of information for the statement in the second paragraph.**

Section 9.2.2 (formerly Section 7.2.2)

18. Please clarify when the normal approximation calculations for 95% confidence intervals are valid — specifically, when "npq" ≥ 5 where "p" is the agreement proportion and "q" is defined as "1 – p" or "1 – agreement."

- **This section has been revised. Normal approximation calculations are no longer recommended.**

Section 9.3.1 (formerly Section 7.3.1)

19. Please include the application of the method described, as well as the conclusion based on the results of the 95% confidence intervals.

    Specifically (on pages 15 & 16),

    Sensitivity: New Whittaker Test Method = 98.4%
    ["npq" = (61)(0.885)(0.115) = 6.2 → OK to use normal approx.]
    95% CI of Old Comparative Method (Approx.) = 80.3 - 96.7%
    95% CI of Old Comparative Method (Exact) = 77.8 - 95.3%

Specificity: New Whittaker Test Method = 97.6%
[“npq” = (41)(0.829)(0.171) = 5.8 → OK to use normal approx.]
95% CI of Old Comparative Method (Approx.) = 71.1 - 94.7%
95% Cl of Old Comparative Method (Exact) = 67.9 - 92.8%

Since the sensitivity of 98.4% for the new Whittaker test method does not fall in the limits of the approximate or exact 95% confidence interval of the old comparative method, it can be concluded that the sensitivity of the new test method is significantly better than that of the old comparative method. Similarly, since the specificity of 97.6% for the new Whittaker test method does not fall in the limits of the approximate or exact 95% confidence interval of the old comparative method, it can be concluded that the specificity of the new test method is significantly better than that of the old comparative method.

- **The document has been modified to include example calculations and appropriate conclusions.**

20. The second example this section gives a perfect opportunity to illustrate the exact binomial confidence limits. The Whittaker method had a sensitivity of 98.4% and a specificity of 97.6%. In both cases, calculating the confidence intervals using the standard formulas would have produced limits greater than 100%. This could have been pointed out in the text with the appropriate confidence intervals from the binomial distribution also mentioned. This would show the difference between the two and also help to highlight why the binomial is preferred (as stated in the document).

- **See response to Comment 19.**

Section 9.3.2 (formerly Section 7.3.2)

21. Please include the application of the method described, as well as the conclusion based on the results of the 95% confidence intervals, as was done in the previous section.

- **The document has been modified to include example calculations and conclusions for the 95% confidence intervals as suggested.**

Summary of Delegate Comments and Subcommittee Responses

EP12-A: *User Protocol for Evaluation of Qualitative Test Performance; Approved Guideline*

Section 7.2

1. "Dividing the 95% interval (2 SD) by 2" should be replaced with "dividing the 95% interval (2 times 1.645 SD) by 3.290." Note, the 95th percentile of the standard normal is located at z = 1.645.

- **The above-mentioned information, included in a previous draft of EP12, has been deleted.**

Section 9.1

2. The efficiency concept is mentioned without any warning as to whether it is meaningful. It is so only if false-negative and false-positive results are equally undesirable.

- **The following text has been added to address the commenter's concern: "The study PVP, study PVN, and estimated efficiency are all functions of estimated sensitivity, specificity, and estimated disease prevalence. Therefore, study PVP, study PVN, and estimated efficiency are meaningful for a particular patient population only when the disease prevalence of the evaluation specimens is the same as the patient population of interest. Even if the prevalence is the same, efficiency is meaningful only when false-positive (1-specificity) and false-negative (1-sensitivity) results are equally undesirable."**

3. There is a commendable emphasis on precise methods of confidence limit calculation – but there is no warning that these take into account sampling variation only. They do not take into account potential sources of bias. The latter are a greater threat to generalizability in practice.

- **The following text has been added to the first paragraph of Section 9.1.1 to address the commenter's concern: "These estimated performance measures are generalizable (unbiased) for the laboratory's expected test performance only to the extent that the evaluation specimens are typical of the specimens analyzed in the laboratory. Users should refer to the most current edition of NCCLS document GP10—*Assessment of the Clinical Accuracy of Laboratory Tests Using Receiver Operating Characteristic (ROC) Plots,* which describes how to select representative study subjects. Estimated performance measures are also subject to variability, because a (random) selection of specimens is used in the evaluation. This variability can be quantified by confidence limits, which decrease as the number of specimens evaluated increases."**

Related NCCLS Publications[*]

C24-A2 **Statistical Quality Control for Quantitative Measurements: Principles and Definitions; Approved Guideline—Second Edition (1999).** This guideline provides definitions of analytical intervals; plans for quality control procedures; and guidance for quality control applications.

EP5-A **Evaluation of Precision Performance of Clinical Chemistry Devices; Approved Guideline (1999).** EP5-A offers guidelines for designing an experiment to evaluate the precision performance of the 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.

GP10-A **Assessment of the Clinical Accuracy of Laboratory Tests Using Receiver Operating Characteristic (ROC) Plots; Approved Guideline (2001).** This document describes the design of a study to evaluate clinical accuracy of laboratory tests and contains procedures for preparing ROC curves; glossary of terms; and information on computer software programs.

I/LA18-A2 **Specifications for Immunological Testing for Infectious Diseases; Approved Guideline—Second Edition (2001).** This guideline outlines specimen requirements; performance criteria; algorithms for the potential use of sequential or duplicate testing; recommendations for intermethod comparisons of immunological test kits for detecting infectious diseases; and specifications for development of reference materials.

NRSCL8-A **Terminology and Definitions for Use in NCCLS Documents; Approved Standard (1998).** This document provides standard definitions for use in NCCLS standards and guidelines, and for submitting candidate reference methods and materials to the National Reference System for the Clinical Laboratory.

[*] Proposed- and tentative-level documents are being advanced through the NCCLS consensus process; therefore, readers should refer to the most recent editions.

NOTES

NOTES

NCCLS ▼ 940 West Valley Road ▼ Suite 1400 ▼ Wayne, PA 19087 ▼ USA ▼ PHONE 610.688.0100
FAX 610.688.0700 ▼ E-MAIL: exoffice@nccls.org ▼ WEBSITE: www.nccls.org ▼ ISBN 1-56238-468-6

NCCLS