EP09-A2-IR
Vol. 30 No. 17
Replaces EP09-A2
Vol. 22 No. 19

# Method Comparison and Bias Estimation Using Patient Samples; Approved Guideline—Second Edition (Interim Revision)

NOTE: Multiple corrections have been made to the formulae and information in this document. For a listing of all corrections, see page xvii.

This document addresses procedures for determining the bias between two clinical methods, and the design of a method comparison experiment using split patient samples and data analysis.

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

CLINICAL AND LABORATORY STANDARDS INSTITUTE®

# Clinical and Laboratory Standards Institute

*Advancing Quality in Health Care Testing*

Clinical and Laboratory Standards Institute (CLSI, formerly 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 health care 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 health care issues. Our process is based on the principle that consensus is an effective and cost-effective way to improve patient testing and health care services.

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

**PUBLICATIONS**

A 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 CLSI 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 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 consensus level.

**Proposed** A consensus document undergoes the first stage of review by the health care 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.

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

Our 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 CLSI's established consensus procedures. Provisions in CLSI 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 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 addressed by the committee in an appendix to the document. Readers are strongly encouraged to comment in any form and at any time on any document. Address comments to Clinical and Laboratory Standards Institute, 940 West Valley Road, Suite 1400, Wayne, PA 19087, USA.

**VOLUNTEER PARTICIPATION**

Health care professionals in all specialties are urged to volunteer for participation in CLSI projects. Please contact us at customerservice@clsi.org or +610.688.0100 for additional information on committee participation.

# Method Comparison and Bias Estimation Using Patient Samples; Approved Guideline—Second Edition (Interim Revision)

## Abstract

CLSI document EP09-A2-IR—*Method Comparison and Bias Estimation Using Patient Samples; Approved Guideline—Second Edition (Interim Revision)* is written for laboratorians as well as manufacturers. It describes procedures for determining the relative bias between two methods, and it identifies factors to be considered when designing and analyzing a method-comparison experiment using split patient samples. For carrying out method-comparison evaluations, an overview of the experiment, sample data recording and calculation sheets, and an overview flowchart and a detailed flowchart for preliminary data examination are included. As an additional aid, a sample scatter plot and bias plot are introduced for those who are unfamiliar with these procedures. The final section contains recommendations for manufacturers' evaluation of bias and statement format for bias claims.

CLSI. *Method Comparison and Bias Estimation Using Patient Samples; Approved Guideline—Second Edition (Interim Revision)*. CLSI document EP09-A2-IR (ISBN 1-56238-731-6). Clinical and Laboratory Standards Institute, 940 West Valley Road, Suite 1400, Wayne, Pennsylvania 19087-1898 USA, 2010.

EP09-A2-IR
ISBN 1-56238-731-6
ISSN 0273-3099

# Method Comparison and Bias Estimation Using Patient Samples; Approved Guideline—Second Edition (Interim Revision)

Volume 30 Number 17

Jan S. Krouwer, Ph.D.
Daniel W. Tholen, M.S.
Carl C. Garber, Ph.D.
Henk M.J. Goldschmidt, Ph.D.
Martin Harris Kroll, M.D.
Kristian Linnet, M.D., Ph.D.
Kristen Meier, Ph.D.
Max Robinowitz, M.D.
John W. Kennedy

CLINICAL AND LABORATORY STANDARDS INSTITUTE®

**Suggested Citation**

CLSI. *Method Comparison and Bias Estimation Using Patient Samples; Approved Guideline—Second Edition (Interim Revision).* CLSI document EP09-A2-IR. Wayne, PA: Clinical and Laboratory Standards Institute; 2010.

**Proposed Guideline**
January 1986

**Tentative Guideline**
April 1993

**Approved Guideline**
December 1995

**Approved Guideline—Second Edition**
September 2002

**Approved Guideline—Second Edition (Interim Revision)**
July 2010

ISBN 1-56238-731-6
ISSN 0273-3099

## Committee Membership

The changes in this interim revision were approved by the Area Committee on Evaluation Protocols and the Board of Directors as follows:

### Area Committee on Evaluation Protocols

**Greg Cooper, CLS, MHA**
**Chairholder**
**Bio-Rad Laboratories, Inc., QSD Division**
**Plano, Texas, USA**

**R. Neill Carey, PhD, FACB**
**Vice-Chairholder**
**Peninsula Regional Medical Center**
**Salisbury, Maryland, USA**

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

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

George S. Cembrowski, MD, PhD
University of Alberta Hospital
Edmonton, Alberta, Canada

David L. Duewer, PhD
National Institute of Standards and Technology
Gaithersburg, Maryland, USA

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

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

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

Lakshmi Vishnuvajjala, PhD
FDA Ctr. for Devices/Rad. Health
Rockville, Maryland, USA

**Staff**

Clinical and Laboratory Standards Institute
Wayne, Pennsylvania, USA

Lois M. Schmidt, DA
*Vice President, Standards Development*

Robin Levy, MS, MT(ASCP)BB
*Staff Liaison*

Melissa A. Lewis, ELS
*Editorial Manager*

**OFFICERS**

Janet K.A. Nicholson, PhD,
President
Centers for Disease Control and Prevention

Mary Lou Gantzer, PhD, FACB,
President-Elect
Siemens Healthcare Diagnostics, Inc.

Jack Zakowski, PhD, FACB,
Secretary
Beckman Coulter, Inc.

W. Gregory Miller, PhD,
Treasurer
Virginia Commonwealth University

Gerald A. Hoeltge, MD,
Immediate Past President
Cleveland Clinic

Glen Fine, MS, MBA, CAE,
Executive Vice President

**BOARD OF DIRECTORS**

Maria Carballo
Health Canada

Russel K. Enns, PhD
Cepheid

Prof. Naotaka Hamasaki, MD, PhD
Nagasaki International University

Christopher M. Lehman, MD
University of Utah Health Sciences Center

Valerie Ng, PhD, MD
Alameda County Medical Center/
Highland General Hospital

Luann Ochs, MS
BD Diagnostics – TriPath

Robert Rej, PhD
New York State Department of Health

Donald St.Pierre
FDA Center for Devices and Radiological Health

Michael Thein, PhD
Roche Diagnostics GmbH

James A. Thomas
ASTM International

Harriet R. Walsh, MA, MT(ASCP)
Centers for Medicare and Medicaid Services

**Acknowledgment**

CLSI and the Area Committee on Evaluation Protocols gratefully acknowledge James Huntington and Simon Huntington, Co-founders, Analyse-it®, Leeds, United Kingdom, for their unwavering commitment and focused effort on the joint venture partnership with CLSI in the development of software to help laboratories easily implement the CLSI method evaluation protocols.

Special thanks go to Simon Huntington for painstakingly reviewing the statistics in EP09-A2 and applying his expert knowledge of statistical analysis for method validation to identify and offer solutions for the discrepancies and errors that have been corrected in this interim revision.

Jeffrey R. Budd, PhD, Beckman Coulter, Inc., Chaska, Minnesota, USA, serving as Chairholder of the Subcommittee on Method Comparison and Bias Estimation Using Patient Samples, reviewed the reported issues and confirmed the recommended resolutions, which were approved by the Area Committee on Evaluation Protocols.

## Committee Membership

### Area Committee on Evaluation Protocols

| | |
|---|---|
| **Jan S. Krouwer, Ph.D.**<br>**Chairholder** | **Krouwer Consulting**<br>**Sherborn, Massachusetts** |
| **Daniel W. Tholen, M.S.**<br>**Vice-Chairholder** | **Statistical Services**<br>**Traverse City, Michigan** |
| Carl C. Garber, Ph.D. | Quest Diagnostics Assurance<br>Teterboro, New Jersey |
| Henk M.J. Goldschmidt, Ph.D. | Tilburg, The Netherlands |
| Martin Harris Kroll, M.D. | Dallas Veterans Affairs Medical Center<br>Dallas, Texas |
| Kristian Linnet, M.D., Ph.D. | Psychiatric University Hospital<br>Risskov, Denmark |
| Kristen Meier, Ph.D. | FDA Center for Devices/Rad. Health<br>Rockville, Maryland |
| Max Robinowitz, M.D. | FDA Center for Devices/Rad. Health<br>Rockville, Maryland |

**Advisors**

| | |
|---|---|
| R. Neill Carey, Ph.D. | Peninsula Regional Medical Center<br>Salisbury, Maryland |
| Patricia E. Garrett, Ph.D. | BBI Clinical Laboratories<br>New Britain, Connecticut |
| John W. Kennedy | Medstat Consultants<br>Palo Alto, California |
| Jacob (Jack) B. Levine, M.B.A. | Bayer Corporation<br>Tarrytown, New York |
| Jennifer K. McGeary, M.T.(ASCP), M.S.H.A.<br>*Staff Liaison* | NCCLS<br>Wayne, Pennsylvania |
| Patrice E. Polgar<br>*Editor* | NCCLS<br>Wayne, Pennsylvania |
| Donna M. Wilhelm<br>*Assistant Editor* | NCCLS<br>Wayne, Pennsylvania |

**Acknowledgements**

The Area Committee on Evaluation Protocols would also like to recognize the valuable contributions of the members and advisors of the Working Group on Method Comparison and Bias Estimation that developed the first approved edition of this guideline.

John W. Kennedy
R. Neill Carey, Ph.D.
Richard B. Coolen, Ph.D.
Carl C. Garber, Ph.D.
Henry T. Lee, Jr.
Jacob B. Levine
Iris M. Osberg

# 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
David G. Rhoads Associates, Inc.
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
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
Department of Health & Community Services (New Brunswick, Canada)
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 Promacao (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)
New Mexico VA Health Care System
North Carolina State Laboratory of Public Health
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)

## Contents

Abstract ..... i

Committee Membership ..... v

Active Membership ..... ix

Interim Revision Changes to EP09-A2 ..... xvii

Foreword ..... xix

The Quality System Approach ..... xx

1 Introduction and Scope ..... 1

1.1 Overview of the General Comparison Experiment ..... 1
1.2 Symbols Used in the Text ..... 2
1.3 Definitions ..... 3

2 Device-Familiarization Period ..... 4

3 Comparison of Methods Experiment ..... 4

3.1 Test Samples ..... 4
3.2 Comparative Method ..... 4
3.3 Range of Measurement ..... 5
3.4 Number of Samples ..... 5
3.5 Sample Sequence ..... 6
3.6 Time and Duration ..... 6
3.7 Inspection of Data During Collection ..... 6
3.8 Quality Control ..... 7
3.9 Documentation of Rejected Data ..... 7

4 Preliminary Data Examination ..... 7

4.1 Outlier Tests on Within-Method Duplicates ..... 11
4.2 Plotting the Data ..... 12
4.3 Visual Check for Linear Relationship ..... 12
4.4 Visual Check for Between-Method Outliers ..... 12
4.5 Test for Adequate Range of X ..... 13

5 Linear Regression ..... 14

5.1 Computations ..... 14
5.2 Visual Check for Constant Scatter ..... 16

6 Computing Predicted Bias and Its Confidence Interval ..... 16

6.1 Linear Regression Procedure (When Data Pass Adequate Range and Uniform Scatter Checks) ..... 16
6.2 Computing Average Bias Using Partitioned Individual Differences When Data Fail Adequate Range Check (Partitioned Biases Procedure) ..... 18
6.3 Computing Predicted Bias Using Partitioned Residuals When Data Have Nonconstant (Variable) Precision (Partitioned Residuals Procedure) ..... 19

7 Interpreting Results and Comparing to Internal Performance Criteria ..... 19

## Contents (Continued)

8 Manufacturer Modifications ....20
8.1 Experimental Design....20
8.2 Data Analysis....20
8.3 Statement of Bias Performance Claims ....20
References....24
Appendix A. Sample Data Recording Sheet....25
Appendix B. Scatter Plots Derived from Example ....27
Appendix C. Calculation Example....31
Appendix D. Calculation of Deming Slope ....36
Summary of Comments and Working Group Responses ....37
Summary of Delegate Comments and Committee Responses....51
Related NCCLS Publications....54

## Interim Revision Changes to EP09-A2

Section 1.2

- Added definition "$\bar{x}_i$" the average of the $x_i$ replicates
- Added definition "$\bar{y}_i$" the average of the $y_i$ replicates
- Definition "$x_{ij}$ or $y_{ij}$" corrected by changing "run" to " sample"
- Definition "$\bar{x}$ or $\bar{y}$" clarified by addition of the term "overall"
- Definition $s_{y \bullet x}$ clarified by adding (standard deviation of the residuals)

Sections 4.1 and 4.2

- Presentation of symbols standardized for consistency (ie, upper case "X" and "Y" changed to lower case, as appropriate)

Formula Corrections:

| | Summation (Σ) equations clearly defined with the index i or j and their range value | Subscript j changed to i | Intermediate step added for clarity | Subscript ij changed to i | y bar added to denominator | Subscript m changed to i | Bar added to x and/or y in numerator | Change SD to s |
|---|---|---|---|---|---|---|---|---|
| Formula Affected by Correction | 10, 12, 13, 14, 15, 16, 17, 19, 20, 24, 25, 27, 28 (now 29), 29 (now 30), 31 (now 32) | 13, 17, 23, 25 | 14, 15, 19, 20 | 22, 24, 25 | 13 | 28 (now 29), 29 (now 30), 31 (now 32) | 28 (now 29), 29 (now 30), 31 (now 32) | 30 (now 31), 31 (now 32) |

Section 5.1

- First sentence – term $(x_{ij} - y_{ij})$ corrected to $\left(\bar{x}_i, y_{ij}\right)$
- Formula 16 – Equation was mathematically incorrect and has been replaced with the correct formula

Section 6.1

- First paragraph, last sentence rewritten for clarity and $\left(\bar{x}_j, y_{ij}\right)$ changed to $\left(\bar{x}_i, y_{ij}\right)$
- Formula 27 redisplayed to handle using individual replicates
- Formula 28 added to handle using sample averages

Sections 6.2 and 6.3

- Reference to "m" dummy subscript deleted, and x and y changed to $\bar{x}$ and $\bar{y}$
- Note added at end of section regarding dealing with replicates

Appendix B. Scatter Plots Derived from Example

- Corrected: Graph B2. Scatter Plot for All Results From Example

Appendix C. Calculation Example

- C3. Adequate Range Test-Correlation (Section 4.5) – Data regenerated based on corrected formula
- C4. Regression Parameter Estimates (Section 5.1) – Data regenerated based on corrected formula
- C5. Residuals and Standard Error of Estimate ($s_{y.x}$) (Section 6.1) – Data regenerated based on corrected formula

## Foreword

The current literature contains many examples of user and manufacturer product evaluations, with many different experimental and statistical procedures[1] for comparing two methods that measure the same analyte. This methodologic variety has caused confusion, and users have reported that comparisons often lack sufficient data and description to be reproducible.

There has also been an increasing awareness that the scope of evaluation procedures appropriate for manufacturers of diagnostic devices is not always appropriate for their users. The manufacturer is concerned with establishing valid and achievable performance claims for bias when compared with a generally accepted standard or reference method. The user might wish to compare a candidate method with a different one than the manufacturer used in establishing the bias claims. The scope of the experimental and data-handling procedures for these two purposes can often differ.

Therefore, in preparing this document, the working group drew on the experience of users and representatives of industry, statisticians, and laboratory and medical personnel. Because of the many *in vitro* diagnostic methods and kits now available, the working group realizes that a single experimental design is not appropriate for all types of user and manufacturer method comparisons. Therefore, this guideline was developed primarily to give conceptual help in structuring an experiment for comparing two methods. To illustrate representative duration, procedures, materials, methods of quality control, statistical data handling, and interpretation of results, an example experiment is presented.

Throughout the development of this protocol, the working group had to decide which procedural and statistical methods to recommend in the example experiment. To respond to the needs of laboratorians and manufacturers, the working group combined input from users of analytical methods, manufacturers of these methods, and representatives of regulatory agencies. The working group also included the recommendations necessary for a scientifically valid comparison. Compromises were necessary to accommodate both the simplicity of operation protocol and the complexity of design and statistical calculations necessary for valid conclusions. This document is adaptable within a wide range of analytes and device complexity.

The focus of this document is the independent establishment of bias performance characteristics. If appropriate, the user is then free to compare these performance estimates with either the manufacturer's labeled claims or the user's own internal criteria.

The working group believes that standard experimental and statistical procedures in user method comparisons will make such evaluations more reproducible and reflective of actual performance, and the statements of evaluation results considerably more reliable. Also, the misuse and misinterpretation of statistical methods, such as regression and correlation, involved in comparing *in vitro* diagnostic devices can seriously impair the usefulness of such evaluations. Therefore, this document is intended to promote the effective use of statistical analysis and data reporting.

Manufacturers of laboratory devices are encouraged to use this guideline to establish and standardize their bias performance claims. Many different forms have been used for such claims, and they have not always been sufficiently specific to allow user verification.

### Key Words

Bias, evaluation protocol, experimental design, linear regression, method comparison, quality control, residuals

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

**EP09-A2-IR 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.*

# Method Comparison and Bias Estimation Using Patient Samples; Approved Guideline—Second Edition (Interim Revision)

## 1 Introduction and Scope

This document provides both users and manufacturers of clinical laboratory devices with guidance for designing an experiment to evaluate the bias between two methods that measure the same analyte. Ideally, a test (or candidate) method should be compared with a reference method. For users, the comparative method is often the current routine method, however, and the purpose of the evaluation is to determine if the two methods yield equivalent results within the statistical power of the experiment. In this case, determining whether the test method is a suitable replacement for a current method is the primary concern.

This guideline allows the estimation of the bias (expected difference) between two methods at various concentrations. If the comparative method is the same one used by the manufacturer in the statement of claims, it is possible to compare statistically the experimental results to the manufacturer's claims to verify acceptable performance.

### 1.1 Overview of the General Comparison Experiment

Evaluating an analytical method requires the following:

- Sufficient time for the operators to become familiar with the device's operation and maintenance procedures.
- Sufficient time for the operators to become familiar with the evaluation protocol.
- Assurance that both the test and the comparative methods are in proper quality control throughout the evaluation period.
- Sufficient data to ensure representative results for both the test and the comparative methods. (What constitutes sufficient data will depend on the precision and interference effects of the two methods, the amount of bias between the two methods, the range of sample analyte values available, and the medical requirements of the test.)

During the device familiarization period, the operators of the test and comparative methods must become familiar with all aspects of set-up, operation, maintenance, trouble-shooting, and quality control of both methods. This period can precede other parts of the evaluation process or coincide with the manufacturer's training period. Run routine laboratory quality control procedures on both methods.

After the familiarization period, the method-comparison experiment can begin. The working group recommends that at least 40 patient samples be analyzed over at least 5 operating days. The reliability and effectiveness of the experiment increase by analyzing more samples over more time, while following the manufacturer's recommendations for calibration.

Analyze each patient sample in duplicate using both the test method and the comparative method. Analyze the duplicates for each method within the same run for that method. Whenever possible, at least 50% of the samples run should be outside the laboratory's reference interval.

When the experiment is completed, record the data in a logical manner (such as that which is suggested in the Appendix). Plot the data and assess the diagram visually and statistically for relative linearity,

adequate range, and uniform scatter. Based on the results of the data examination, use either simple linear regression or alternative procedures to estimate expected (average) bias and the confidence interval for expected bias at any desired medical decision level. Then, these estimates can be compared with claims or internal criteria to judge the acceptability of the method.

## 1.2 Symbols Used in the Text

The following symbols are used in this document:

| | |
|---|---|
| X | comparative method |
| Y | test method |
| $DX_i$ or $DY_i$ | absolute value of the difference between duplicates for method X or Y |
| i | sample number |
| N | total number of samples |
| 1,2 or j | duplicate number or replicate number (as a subscript) |
| $\overline{DX}$ or $\overline{DY}$ | mean absolute difference of method |
| $DX'_i$ or $DY'_i$ | normalized (relative) absolute difference of method |
| $\overline{DX'}$ or $\overline{DY'}$ | normalized (relative) mean absolute difference of method |
| $E_{ij}$ | absolute difference between methods |
| $\overline{E}$ | mean absolute difference between methods |
| $E'_{ij}$ | relative absolute difference between methods |
| $\overline{E'}$ | relative mean absolute difference between methods |
| $TL_E$ | test limit |
| r | correlation coefficient |
| x | observation from comparative method |
| $\bar{x}_i$ | the average of the $x_i$ replicates |
| y | observation from test method |
| $\bar{y}_i$ | the average of the $y_i$ replicates |
| $x_{ij}$ or $y_{ij}$ | observation (x or y) from sample i, replicate j |

| | |
|---|---|
| $\bar{x}$ or $\bar{y}$ | overall average of x or y |
| b | slope |
| a | y intercept |
| $\hat{Y}$ | predicted value for test method |
| $s_{y \cdot x}$ | standard error of estimate (standard deviation of the residuals) |
| $\hat{B}_c$ | estimate of predicted bias at concentration c |
| $X_c$ | medical decision level |
| $B_c$ | true bias at concentration $X_c$ |
| $N_K$ | number of data points in group K (K = 1, 2, 3) |
| $\sum_{m=1}^{N_K}$ | denotes that the sum is performed on the paired x's and y's in group K (K = 1,2,3) |
| $\bar{B}_K$ | average bias in group K (K = 1,2,3) |
| $s_K$ | standard deviation of biases in group K. |

## 1.3 Definitions[a]

**Analytical measurement range, AMR,** *n* - The range of analyte values that a method can directly measure on the sample without any dilution, concentration, or other pretreatment that is not part of the typical assay process.

**Bias,** *n* - The difference between the expectation of the test results and a true value.

**Clinically reportable range, CCR,** *n* - The range of analyte values that a method can report as a quantitative result, allowing for sample dilution, concentration, or other pretreatment used to extend the direct analytical measurement range.

**Correlation coefficient, *r, n* - *For measured data,*** the ratio of the covariance of two random variables to the product of their standard deviation; **NOTE:** For this document, the correlation coefficient is defined as the square root of the slope of y regressed on x, times the slope of x regressed on y.

**Deming regression,** *n* - A method to estimate slope and intercept parameters from a method comparison experiment with allowance for both methods to have measurement error. The measurement error for each method is used in the estimation procedure.[2]

**Measurand,** *n* - A particular quantity subject to measurement ***(VIM93-2.6);*** **NOTE:** This term and definition encompass all quantities, while the commonly used term "analyte" refers to a tangible entity

[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.

subject to measurement. For example, "substance" concentration is a quantity that may be related to a particular analyte.

**Passing-Bablok,** *n* - A method to estimate slope and intercept parameters from a method comparison experiment using a nonparametric procedure.[3]

**Trueness,** *n* **-** The closeness of agreement between the average value obtained from a large series of test results and an accepted reference value; **NOTE**: The measure of trueness is usually expressed in terms of bias.

# 2 Device-Familiarization Period

The operators of both the test and the comparative methods must be familiar with the following:

- Operation
- Maintenance procedures
- Methods of sample preparation
- Calibration and monitoring functions.

Manufacturers' training programs, when offered, can be a part of the familiarization period. Set up and operate the test device in the laboratory long enough to ensure that the operators understand all procedures and can properly operate the device. The working group recommends five days for the device familiarization period. For extremely simple devices, a shorter period can suffice; for complex, multichannel devices, a longer period can be required.

The operators should practice analyzing actual sample materials to bring to their attention all possible contingencies (such as error flags, error correction, calibration, etc.) that might arise during routine operation of either device. Data should not be collected during this period. The device familiarization period is not complete until the operators can operate the device with confidence. (This procedure may not be necessary for all user evaluations.) Before beginning the method-comparison evaluation, ensure that routine quality control procedures are in place with appropriate control limits.

# 3 Comparison of Methods Experiment

## 3.1 Test Samples

Collect and handle patient samples according to accepted laboratory practice and manufacturer's recommendations.

### 3.1.1 Storage

The duration and conditions of storage depend on the stability of the measurand to be analyzed. Avoid storing samples, if possible.

### 3.1.2 Excluded Samples

If a sample is excluded, record the reason for the exclusion.

## 3.2 Comparative Method

For the comparative method, use the laboratory's current method, the method used by the manufacturer in the labeled claims, or a recognized reference method.

If the comparison method is a reference method, then the difference between the two methods measures the trueness of the new method, measured as bias. If the comparison method is not a reference method, then the trueness of the new method cannot be determined. In this case, one should refer to the difference then simply as a difference, and not bias. Since the preferred approach is to use a reference method as the comparison method, the term "bias" is used in this document.

This experiment gives an estimate of the bias between two methods and the confidence interval for the bias, at any particular concentration. So that differences between the two methods are attributable to errors in the test method, the comparative method should do the following:

- Have better precision than the test method, which can be achieved by replication, if needed.
- Be free from known interferences, whenever possible.
- Use the same units as the test method.
- Have known bias relative (traceable) to standards or reference methods, whenever possible.

This experiment does not segregate the various sources of bias into those coming from each of the methods being compared. (See the most current version of NCCLS document EP14—*Evaluation of Matrix Effects*, for information on detection of matrix interference.) Interference effects may contribute as much as imprecision effects to the differences between methods. (Proper characterization of interference effects on each method can be determined by a separate experiment; see the most current version NCCLS document EP7—*Interference Testing in Clinical Chemistry.*)

## 3.3 Range of Measurement

Evaluate the test method over the clinically meaningful range, i.e., where medical decisions are made. In general, this range extends from below to substantially above the expected reference range. Analyte concentrations should be distributed over the analytical measurement range to the extent possible. The analytical measurement range is the analyte concentration interval claimed by the manufacturer to provide acceptable performance. Tables 1a and 1b show a recommended distribution that takes into account the availability of abnormals for a set of analytes.

### 3.3.1 Analytical Measurement Range

The range of the study is limited by the analytical measurement ranges of the two methods. The range of the comparative method should be at least as wide as the range of the test method so that bias at the limits of the analytical measurement range can be compared.

## 3.4 Number of Samples

Analyze at least 40 samples that meet the criteria stated above. More samples will improve the confidence in the statistical estimates and increase the opportunity to incorporate the effects of unexpected interfering substances (individual idiosyncratic biases). See Figures A1 and A2 for examples of data recording sheets.

### 3.4.1 Duplicate Measurements

For the following reasons, obtain a sufficient quantity of each sample: (1) duplicates can be analyzed by the test method; (2) duplicates can be analyzed by the comparative method; and (3) follow-up studies can be performed, if required.

### 3.4.2 Pooled Samples

If the required volume of a sample cannot be obtained from a single patient, then make "minipools" by mixing samples from two (but not more) patients with approximately the same level of measurand and similar disease histories. Use the "minipools" for the two sets of duplicate analyses. If the samples are whole blood, mixing requires serologic compatibility.

**NOTE:** The process of pooling can mask by averaging out unique or sample-specific biases and thus can lead to an optimistic picture of the comparability of the two methods.

## 3.5 Sample Sequence

Assign the first aliquot of the selected samples sequential positions in the run. Run the second (duplicate) aliquots in reverse order. Reversing the order of the second aliquots minimizes the effects of carryover and drift on the averages of the duplicates within the run. Make every effort to randomize the samples in the sequence. For example, the samples could be run in the following order: 1, 2, 3, 4, 5, 6, 7, 8 and 8, 7, 6, 5, 4, 3, 2, 1. Follow this reverse procedure for both the comparative and test methods, but different initial sequences can be used for each method.

## 3.6 Time and Duration

For a given sample, analysis by the comparative and test methods should occur within a time span consistent with the analyte stability. For all analytes, the time span should not exceed two hours for analysis by each method. If possible, use samples drawn the day of the analysis. If stored samples are used, make sure they were all stored in a manner that ensures their stability and meets the stated requirements of both the test and the comparative methods. Store samples in the same manner for both procedures to avoid introducing storage conditions as a variable.

If the comparison of methods experiment is carried out after the precision experiment (described in the most current version of NCCLS document EP5—*Evaluation of Precision Performance of Clinical Chemistry Devices*), up to eight samples can be selected and analyzed on a single day. If the comparison of methods experiment and precision evaluation experiment are done simultaneously, only four samples a day should be analyzed 10 to 15 days after the protocol familiarization period. Spreading the patient sample data over many days and runs is preferable.

## 3.7 Inspection of Data During Collection

### 3.7.1 Analytical System Errors

Document data collected during a time when the device indicates that an error condition exists, but do not include it in the final data analysis.

### 3.7.2 Human Error

Record any data for which the operator can document that an error was made, but do not include it in the final analysis.

### 3.7.3 Evaluation of Other Discrepant Data

Record the pairs of duplicate data points for which no errors were detected without editing. If a reason for any discrepancy cannot be determined, retain the original results in the data set, subject to the outlier checks in Sections 4.1 and 4.4.

### 3.8 Quality Control

Follow the laboratory's and/or manufacturer's routine quality control procedures during the experiment. Keep control charts, and repeat any run that appears to be out of control on either method until the required number of samples is obtained.

### 3.9 Documentation of Rejected Data

Carefully document and retain a record of any situation that requires the rejection of data along with any discovered acceptable causes and problems.

## 4 Preliminary Data Examination

Figure 1 shows an overview of the examination process described in this section. Figure 2 shows the logic flow chart of the individual steps in the process. Refer to these figures while reading the following sections.

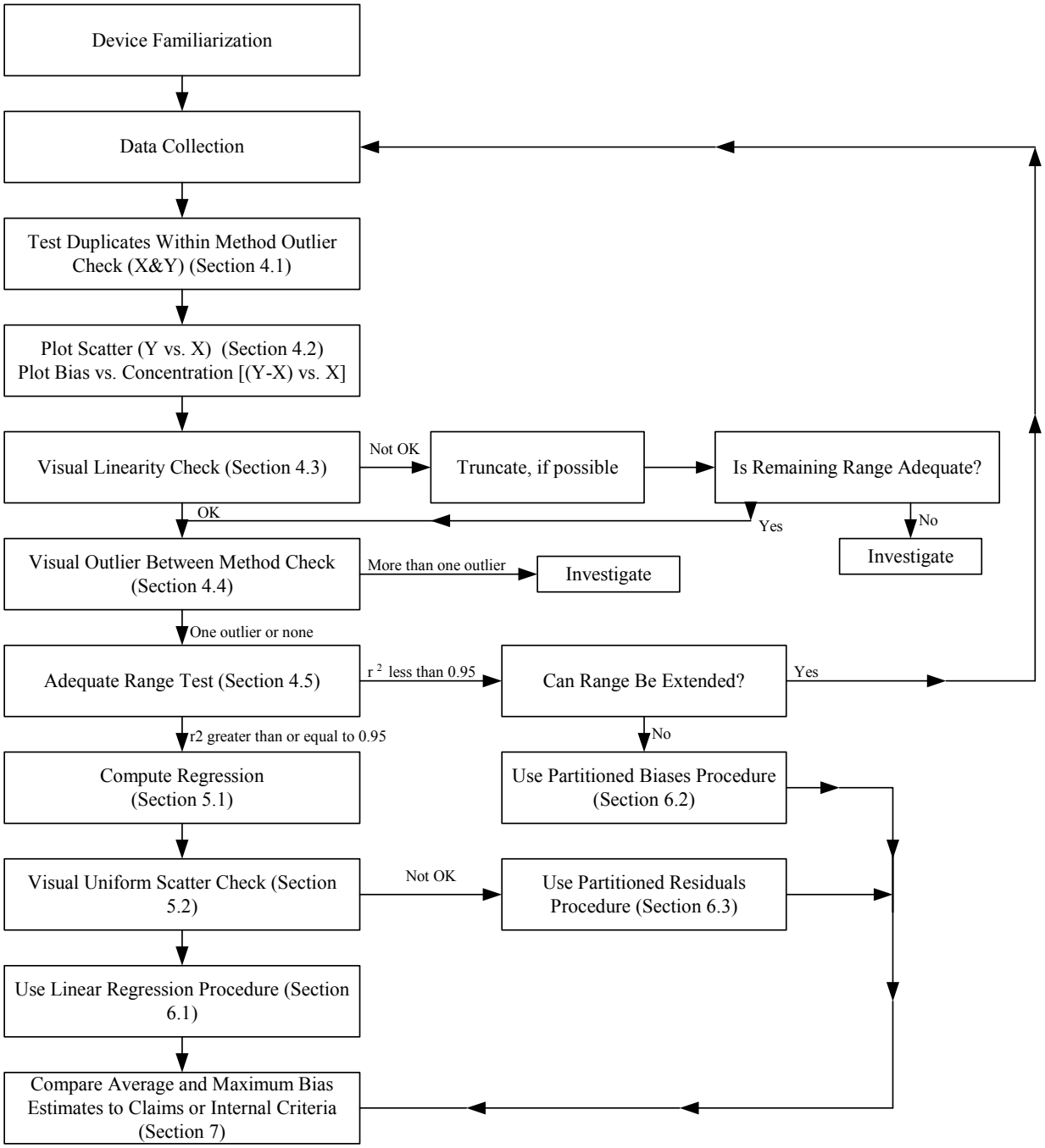



**Figure 1. Overview Flowchart of Protocol**

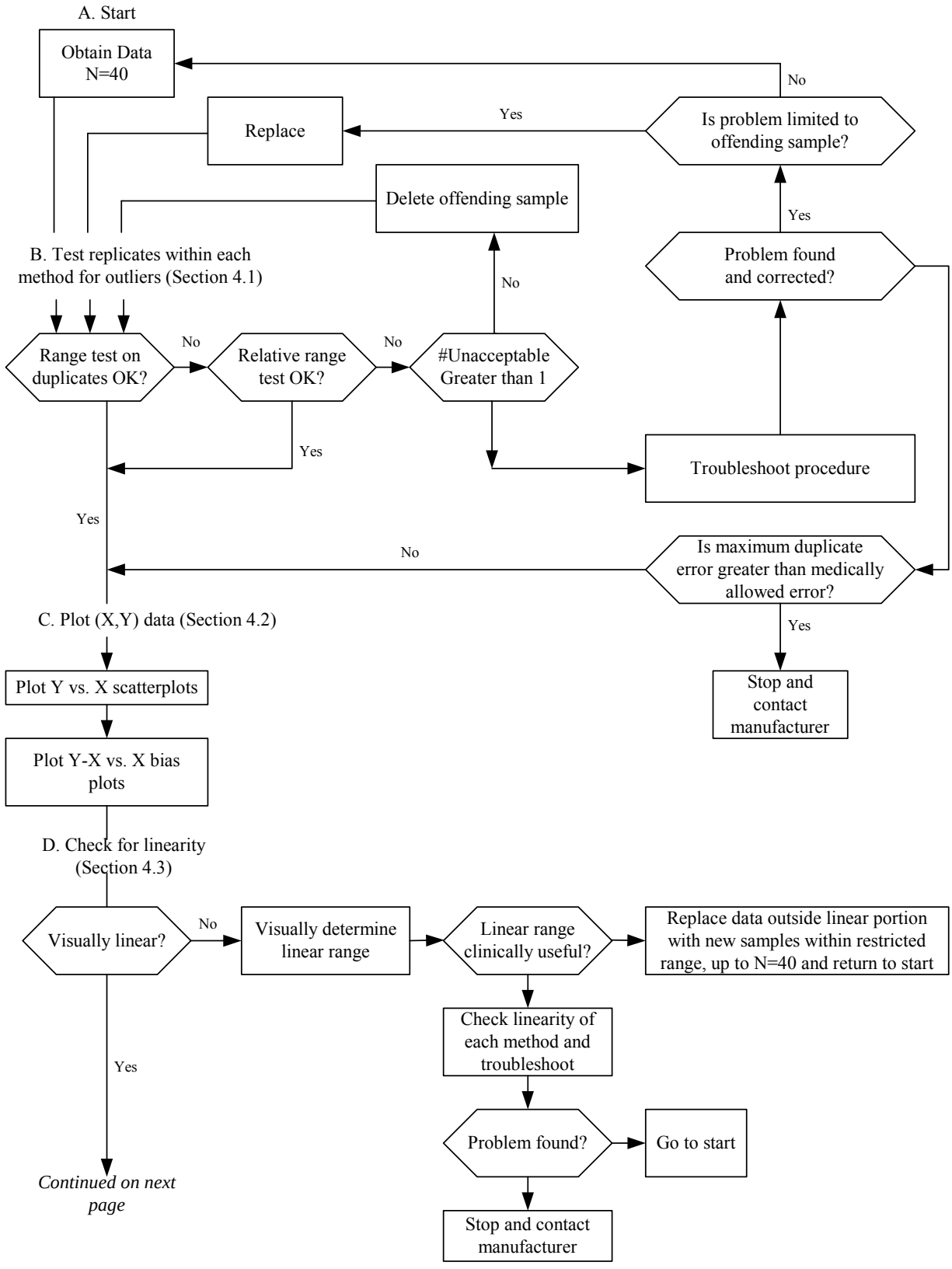



**Figure 2. Detailed Flowchart of Protocol**

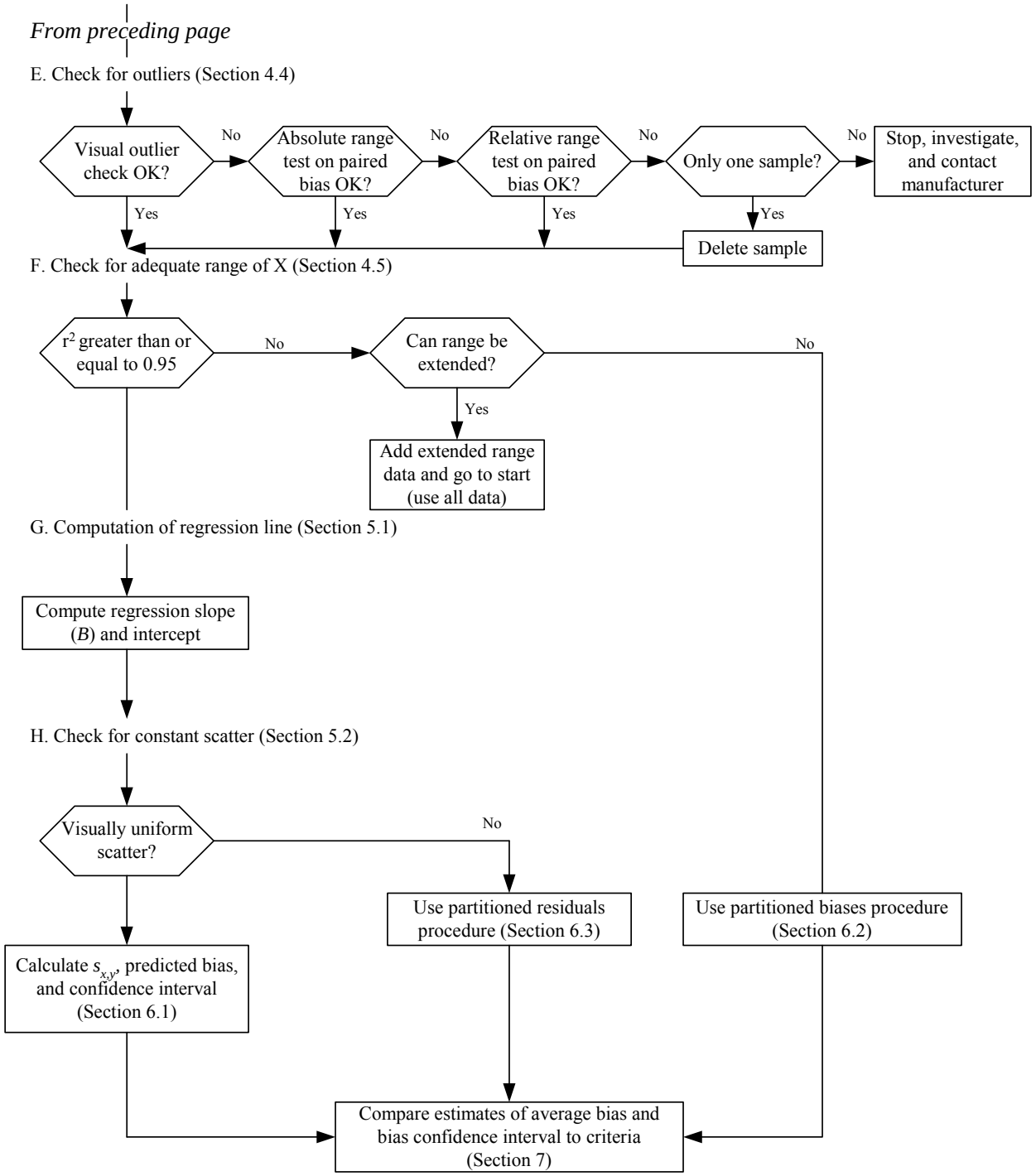



**Figure 2. (Continued)**

### 4.1 Outlier Tests on Within-Method Duplicates

The analysis should be done with all points and with any outliers which have been removed. Apply the following procedure to the duplicates on the Test (Y) and Comparative (X) Methods.[4] The analysis should be done two ways, 1) with all points and 2) with any outliers which have been removed. Compute the absolute values of the differences between the duplicates for each sample as follows:

$$DX_i = | x_{i1} - x_{i2} | \quad (1)$$

$$DY_i = | y_{i1} - y_{i2} | \quad (2)$$

where i = the sample number (which goes from 1 to N, and N = total number of *samples*).

Compute the mean absolute difference between duplicates for each method:

$$\overline{DX} = \frac{\sum DX_i}{N} \quad (3)$$

$$\overline{DY} = \frac{\sum DY_i}{N} \quad (4)$$

Compute "acceptability" limits of four times these mean absolute differences for each method (rounded up to the next higher reportable value). If any individual absolute difference exceeds the appropriate (X or Y) limit value, make an additional calculation for each method using normalized (relative) absolute differences; thus:

$$DX'_i = \frac{| x_{i1} - x_{i2} |}{\overline{x}_i} \quad (5)$$

$$DY'_i = \frac{| y_{i1} - y_{i2} |}{\overline{y}_i} \quad (6)$$

$$\overline{DX'} = \frac{\sum DX'_i}{N} \quad (7)$$

$$\overline{DY'} = \frac{\sum DY'_i}{N} \quad (8)$$

Limits of four times the mean values of the relative differences provide test limits for the normalized values.

If a single data point falls outside the limits for *both* the range and relative range procedures, investigate why it did so, and delete the sample from the data set. Continue analyzing the data after deleting all data (x and y) for that sample.

If more than one sample has to be deleted, carry out an expanded investigation into the cause of the discrepancies. If the source of the problem can be identified and traced to the offending samples alone, replace those samples in the data set. The cause of the problem must be documented. If it can be

corrected but not traced to specific samples, the entire data set must be recollected. If the problem is neither found nor corrected, evaluate the size of the maximum difference between duplicates relative to the allowable medical decision limits for precision of the method. If those limits are not exceeded, return the data and follow the subsequent steps. If these limits are exceeded, stop the experiment and notify the manufacturer. (See Section 3.9 on documentation of rejected data.)

## 4.2 Plotting the Data

Make four plots of the data. The first is the scatter plot of $\bar{y}_i$ (mean of duplicates) versus $\bar{x}_i$ (mean of duplicates), treating the test method as the Y variable and the comparative method as the X variable (see Figure B1). Make the origins and scales of both axes identical, and draw a line with the slope of 1.0 going through the origin. The second should plot each individual $y_{ij}$ against its mean $\bar{x}_i$ in the same way (see Figure B2).

The third is the bias plot for which the X axis variable depends on whether the comparative method is a reference method.[5,6] If this is the case then the third plot is the bias plot where the differences between the mean Y and mean X values $(\bar{y}_i - \bar{x}_i)$ for each assay are plotted against the $\bar{x}_i$ value (see Figure B3). The horizontal centerline of this plot has the value of zero. The fourth plot, as above, plots the individual Y differences from the average X $(y_{ij} - \bar{x}_i)$ against the same $\bar{x}_i$ values (see Figure B4).

If the comparative method is not a reference method or if one is not sure, then the third plot is the bias plot where the differences between the mean Y and mean X values $(\bar{y}_i - \bar{x}_i)$ for each assay are plotted against the $(\bar{y}_i + \bar{x}_i)/2$ value (see Figure B3). The horizontal centerline of this plot has the value of zero. The fourth plot, as above, plots the individual Y differences from the average X $(y_{ij} - \bar{x}_i)$ against the same $(\bar{y}_i + \bar{x}_i)/2$ values (see Figure B4).

Using all four of these plots is helpful because the differences in scale between them can be used to balance decisions on the effect of nonlinear relationship, outliers, and nonconstant variance on the comparison between the test and comparative methods.

## 4.3 Visual Check for Linear Relationship

Check the plots of the data for a linear relationship between X (the comparative method) and Y (the test method) throughout the measured range. If there appears to be a satisfactory linear relationship, examine the data according to the procedures given in Section 4.4. (Please refer to the most current version of NCCLS document EP6—*Evaluation of the Linearity of Quantitative Analytical Methods*, for additional information.)

If there is evidence of a nonlinear relationship, visually determine whether the data contains a linear portion. Often, nonlinearity will occur at the extremes of the concentration values. If this is the case, truncate the data point(s) where they begin to be nonlinear. Examine the remaining linear portion to determine whether it is sufficiently wide to cover the medically useful range. If so, analyze additional samples within that range to replace these excluded samples. Then examine the new data set beginning again at Section 4.

If no linear portion is evident, or if the linear portion is too small, stop the evaluation and notify the manufacturer. If the source of the nonlinearity can be identified and corrected, begin the experiment again with new data.

## 4.4 Visual Check for Between-Method Outliers

Examine data plot A and data plot C for visually obvious outliers. If there are no such points, proceed to Section 4.5. If outliers exist, carry out the following calculations similar to those used for the duplicates in Section 4.1.

Compute the absolute differences between methods and their average; thus:

$$E_{ij} = | y_{ij} - \overline{x}_i | \quad (9)$$

where i = the sample number 1...40 and j = the duplicate number 1 or 2.

$$\overline{E} = \frac{1}{2N} \sum_{i=1}^{N} \sum_{j=1}^{2} E_{ij} \quad (10)$$

Compute the Test Limit ($TL_E$) as $4 \times \overline{E}$, rounded *up* to the next higher reportable value. Compare each $E_{ij}$ with this test limit, and label any point that exceeds this limit.

Compute the *relative* absolute differences between methods and their average; thus:

$$\overline{E'_{ij}} = \frac{\left| y_{ij} - \overline{x}_i \right|}{\overline{x}_i} \quad (11)$$

$$\overline{E'} = \frac{1}{2N} \sum_{i=1}^{N} \sum_{j=1}^{2} E'_{ij} \quad (12)$$

Compute a relative test limit as 4 x $\overline{E'}$, and compare each $E'_j$ with this limit (do not round this limit up). Label any points that exceed this limit.

Any point ($X_{ij}$, $Y_{ij}$) that fails *both* tests is an outlier. A maximum of 2.5% of the data may be deleted from the data set.

If more than 2.5% of the data are identified as outliers by this test, investigate possible interferences, human error, and device malfunctions. If several analytes are being simultaneously evaluated on the same device, examine the results for the offending sample on other analytes. Also, review the quality control results during the runs. If obvious causes cannot be determined, and if the differences resulting between the values exceed the bounds of medical significance, then stop the evaluation or add 40 new samples.

If more than one outlier is detected, but the outliers do not exceed a medically significant difference, retain and use the data. If the expanded investigation shows reasons for the outliers, analyze additional samples and use the data from them to augment the data set.

## 4.5 Test for Adequate Range of X

The results of a regression analysis are valid only if certain assumptions about the data are true. One of these assumptions is that the X variable is known without error. In the clinical laboratory, this is not true because every measurement has intrinsic error. However, if the range of the data is sufficiently wide, the effect of this error on the regression estimates can be considered negligibly small. The correlation coefficient, r, can be used as a rough guide to assess the adequacy of the X range in overcoming this problem. The formula for r is as follows:

(13)

$$r = \frac{\sum_{i=1}^{N}(\bar{x}_i - \bar{x})(\bar{y}_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(\bar{x}_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(\bar{y}_i - \bar{y})^2}}$$

where

$$\bar{x} = \frac{\sum_{i=1}^{N}\bar{x}_i}{N} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{2}x_{ij}}{2N} \quad (14)$$

$$\bar{y} = \frac{\sum_{i=1}^{N}\bar{y}_i}{N} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{2}y_{ij}}{2N} \quad (15)$$

As a general guide, the range of X can be considered adequate if $r \geq 0.975$ (or equivalently, if $r^2 \geq 0.95$). If the data yield an *r* that satisfies this requirement, the error in X is adequately compensated by the range of data, and simple linear regression can be used to estimate the slope and intercept.

If $r^2 < 0.95$, then the range of the data must be extended by assaying additional samples. Then, begin examining the entire data set again. If the range cannot be extended, use the partitioned biases procedure described in Section 6.2 in place of linear regression to estimate average bias.

**NOTE:** This procedure assesses the *range* of the data; it does not measure the *distribution* of the data within the range. One must still obtain an even distribution of data throughout the range.

# 5 Linear Regression

## 5.1 Computations

For the set of paired observations $(\bar{x}_i, y_{ij})$ the slope (b) and the *y*-intercept (a) are calculated according to the following formulas:

The average X value for each pair of X observations is calculated, and = $\bar{x}_i$ .

For individual Ys versus average X,

$$b = \frac{\sum_{i=1}^{N}\sum_{j=1}^{2}[(\bar{x}_i - \bar{x})(y_{ij} - \bar{y})]}{2\sum_{i=1}^{N}(\bar{x}_i - \bar{x})^2} \quad (16)$$

For average Y versus average X,

$$b = \frac{\sum_{i=1}^{N}[(\bar{x}_i - \bar{x})(\bar{y}_i - \bar{y})]}{\sum_{i=1}^{N}(\bar{x}_i - \bar{x})^2} \quad (17)$$

$$a = \bar{y} - b\bar{x} \quad (18)$$

where

$$\bar{y} = \frac{\sum_{i=1}^{N}\bar{y}_i}{N} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{2}y_{ij}}{2N} \quad (19)$$

and

$$\bar{x} = \frac{\sum_{i=1}^{N}\bar{x}_i}{N} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{2}x_{ij}}{2N} \quad (20)$$

Thus, the computed line is described by the following equation:

$$\hat{Y} = bX + a \quad (21)$$

For any given concentration value (X), the equation may be used to produce a *predicted* value ($\hat{Y}$) for the test method. Save the results of this regression for later use. Alternative regression procedures, such as Deming (orthogonal as a special case when $\lambda = 1$) or Passing-Bablok, may be used for estimating the slope and the intercept *only*. After fitting such a model, follow all other steps below. One should not use the orthogonal regression or Deming procedures for calculation of the standard error of estimate because this value will be artificially low unless one computes the standard error based on the vertical and not the orthogonal distance.

### 5.2 Visual Check for Constant Scatter

Examine the scatter and bias plots (Figures B1 through B4) for constant scatter. Although few methods have constant imprecision (which contributes to constant scatter) throughout the analytical measurement range of that test, visual examination determines whether there are dramatic and significant differences (approximately 3:1 or greater) between the standard deviations at the upper and lower ends of the range. If the data appear to exhibit reasonably constant scatter, use the linear regression procedure described in Section 6.1 to compute average bias. In this situation, ordinary least squares regression can still be used to estimate the slope and intercept of the line relating X and Y. Even when scatter is not constant, the estimates of slope and intercept will be unbiased (in the statistical sense). However, in this case, the standard error of estimate ($S_{y.x}$) is not usable for measuring variability around the regression line. Use the partitioned residuals procedure described in Section 6.3 for variability estimates and for making statements about average bias.

Deciding adequate constant scatter is difficult when only 40 samples (80 paired analysis points) are available. Therefore, the working group recommends that more samples be collected if nonconstant scatter is suspected.

Alternatively, standard statistical procedures exist for correcting regression in the presence of non-constant scatter. These techniques include using transformed data (such as logarithms and weighted regressions).

## 6 Computing Predicted Bias and Its Confidence Interval

### 6.1 Linear Regression Procedure (When Data Pass Adequate Range and Uniform Scatter Checks)

The difference, measured in the *Y* direction, between a given data point and the regression line is called the *residual* for that point. The standard error of estimate ($S_{y.x}$) is the standard deviation of these residuals and is thus a measure of the "scatter" of the points around the regression line.

The residual for individual Ys versus average X $\left(\bar{x}_i, y_{ij}\right)$ can be calculated using the following formula:

$$\text{Residual}_{ij} = y_{ij} - \hat{Y}_i = y_{ij} - (a + b\bar{x}_i) \tag{22}$$

and for the average $\left(\bar{x}_i, \bar{y}_i\right)$:

$$\text{Residual}_{i} = \bar{y}_i - \hat{Y}_i = \bar{y}_i - \left(a + b\bar{x}_i\right) \tag{23}$$

The standard error of estimate (standard deviation of the residuals) is given by performing the following calculations for individual $y_{ij}$:

$$s_{y\cdot x} = \sqrt{\frac{\sum_{i=1}^{N}\sum_{j=1}^{2}(y_{ij} - \hat{Y}_i)^2}{2N-2}} \tag{24}$$

and for average $\bar{y}_i$

$$s_{y\cdot x} = \sqrt{\frac{\sum_{i=1}^{N}(\bar{y}_i - \hat{Y}_i)^2}{N-2}} \tag{25}$$

The estimate of the predicted bias ($B_c$) at a given Medical Decision Level $X_c$, is given by:

$$\hat{B}_c = a + (b - 1)\, X_c \tag{26}$$

The 95% confidence interval for $B_c$ (the true bias at $X_c$) is given by:

Formula 27 when using individual replicates and $s_{y\cdot x}$ from **formula 24**

$$[\hat{B}_{c,low'}\ \hat{B}_{c,high}] = \hat{B}_c \pm 2s_{y\cdot x}\sqrt{\frac{1}{2N} + \frac{(X_c - \bar{x})^2}{2\sum_{i=1}^{N}(\bar{x}_i - \bar{x})^2}} \tag{27}$$

Formula 28 when using sample averages and $s_{y\cdot x}$ from **formula 25**

$$[\hat{B}_{c,low'}\ \hat{B}_{c,high}] = \hat{B}_c \pm 2s_{y\cdot x}\sqrt{\frac{1}{N} + \frac{(X_c - \bar{x})^2}{\sum_{i=1}^{N}(\bar{x}_i - \bar{x})^2}} \tag{28}$$

Use the procedures in Section 7 for interpretation of these statistics.

## 6.2 Computing Average Bias Using Partitioned Individual Differences When Data Fail Adequate Range Check (Partitioned Biases Procedure)

Tabulate the data in order of increasing X values, and then divide the data into three groups (low, middle, and high) where each group contains approximately the same number of data points. Base this grouping on the X value of each data pair. One can carry out this process by counting points in from the extremities of the bias plot 2*N*/3 points to identify the boundaries of the three groups. (Allocate points that fall on the boundary such that each group maintains an approximately equal number of points.) Label the data on the recording sheet as to which of the three groups each data point belongs. Then, calculate separately the average bias for each group using the following formulas:

[$N_K$ = number of data points in group K (K = 1,2,3)].

(Note that the sum is performed on the paired $\bar{x}$ and $\bar{y}$ in group K.)

$$\bar{B}_K = \frac{\sum_{i=1}^{N_K} (\bar{y}_i - \bar{x}_i)}{N_K} \tag{29}$$

$$s_K = \sqrt{\frac{\sum_{i=1}^{N_K} [(\bar{y}_i - \bar{x}_i) - \bar{B}_K]^2}{N_K - 1}} \tag{30}$$

This calculation sequence computes the biases (differences) for each point in the group and the standard deviations of those biases. The value of $\bar{B}_K$ is the estimated predicted (average) bias for the appropriate concentration range, and the set of three $\bar{B}_K$'s replaces the $\hat{B}_c$ of Section 6.1. If the set of 3 $\bar{B}_K$'s are approximately the same, then report the average of the three, as $\bar{B}$.

The medical decision levels are chosen for clinical utility and do not depend at all on the way the data is broken up into ranges. If it turns out that an important medical decision level is close to the boundary between two partitions, it is often useful to move the partition to avoid the discontinuity in the bias estimate (or choose the larger one).

The 95% confidence interval for the predicted bias ($\hat{B}_c$) at a medical decision level concentration $X_c$ is given by choosing the appropriate K for $X_c$ and performing the following computations:

$$\left[\hat{B}_{c,low'}\ \hat{B}_{c,high}\right] = \bar{B}_K \pm 2\frac{(s_K)}{\sqrt{N_K}} \tag{31}$$

**NOTE:** In a manner similar to Section 6.1, the calculation can also be performed using the individual replicates.

### 6.3 Computing Predicted Bias Using Partitioned Residuals When Data Have Nonconstant (Variable) Precision (Partitioned Residuals Procedure)

Divide the data into three groups, as in Section 6.2, with approximately equal numbers of data points in each group. Then, calculate the following separately for each group, where $N_K$ = number of data points in group K (K = 1, 2, 3).

$$s_K = \sqrt{\frac{\sum_{i=1}^{N_K} (\bar{y}_i - \hat{Y}_i)^2}{N_K - 1}} \tag{32}$$

(Note that the sum is performed on each paired $\bar{x}_i$ and $\bar{y}_i$ in group K.)

The estimate of the predicted bias ($\hat{B}_c$) at a given medical decision level $X_c$ is:

$$\hat{B}_c = a + (b - 1)\, X_c \tag{33}$$

and a 95% confidence interval for $B_c$ is given by choosing the appropriate K for $X_c$ and performing the following equation:

$$\left[\hat{B}_{c,low'}\ \hat{B}_{c,high}\right] = \hat{B}_c \pm 2\,\frac{(s_K)}{\sqrt{N_K}} \tag{34}$$

**NOTE:** In a manner similar to Section 6.1, the calculation can also be performed using the individual replicates.

## 7 Interpreting Results and Comparing to Internal Performance Criteria

In most instances, the difference between a current method and a candidate replacement method is of interest. In these cases, compare the confidence interval of the predicted bias with the definition of acceptable error at the medical decision point $X_c$. Each laboratory should develop its own criteria (in consultation with its medical staff and/or the technical literature). If the confidence interval for predicted bias includes the defined acceptable bias, then the data do not show that the bias of the candidate method is different from the acceptable bias. However, if the confidence interval for expected bias does not contain the defined acceptable bias, then one of the two following decisions can be made:

- If the acceptable bias is less than the lower limit of the confidence interval of the predicted bias, the following conclusion can be drawn:

  There is a high level of probability (>97.5%) that the predicted bias is greater than the acceptable bias and, therefore, the performance of the candidate method is not equivalent to the current method and may not be acceptable for the defined application.

- If the acceptable bias is greater than the higher limit of the confidence interval of the predicted bias, the following conclusion can be drawn:

There is a high probability (>97.5%) that the predicted bias is less than the acceptable bias and, therefore, the performance of the candidate method is equivalent to the current method and is acceptable for the defined application.

If nonequivalence is observed, and yet it is believed that the candidate replacement method is more specific, rather than reject the new method, obtain new clinical data for it (such as a new reference range) before putting it into routine use. Remember that the criteria developed for the laboratory should define allowable differences between two methods. Criteria for medically allowable errors for precision alone might not apply when comparing allowable error for two methods. Error limit guidelines can be found in the literature for intra-individual biological variation for the test being studied.

Where a manufacturer has provided method comparison data for the test method, an additional assessment of performance can be made. However, remember that in order to make a valid comparison to the manufacturer's data, the comparative method and operating procedures must be identical to the manufacturer's. If the manufacturer's claim for average bias is included in the 95% confidence interval, then it can be concluded that the candidate method has provided equivalent results.

# 8 Manufacturer Modifications

## 8.1 Experimental Design

The manufacturer should obtain a minimum of 100 patient samples, spread throughout the claimed analytical measurement range of the method or device. The manufacturer may choose to employ more than 100 patient samples, particularly if multiple sites are used to collect the samples, or if other factors require study.

Patient samples may be used to assess multiple analytes.

## 8.2 Data Analysis

Follow the basic procedure described in this document for preliminary examination of the collected data. The manufacturer may choose to analyze the data with any valid statistical procedure, but the end point must be the estimation of the bias between the test and comparative methods at relevant medical decision points. To assess the error in the parameters, the manufacturer should compute the standard errors of the regression slope and intercept, as well as the standard error of the predicted value at the points used for the bias claims. If the standard errors are unacceptably large, additional data can be required. Avoid the use of invalid procedures, such as measuring the standard error of estimate in the perpendicular (orthogonal) direction (Deming).

## 8.3 Statement of Bias Performance Claims

The following items should be included in a manufacturer's claim for method comparison bias. Unless the comparative method is an established reference method, the terms "accuracy" and "trueness" should not be used. Items listed as optional may be included at the manufacturer's discretion.

- The slope and intercept of the fitted linear regression line (by any method).
- The total number of points used in the regression.
- The bias calculated from the regression line at stated medical decision points (either at generally recognized decision points or at the extremes of the reference interval).

- The range of data (the highest and lowest value of X included in the regression).
- The comparative method used in the regressions.
- Whether individual observations were used in the regressions or means of replicate determinations and, if so, how many repetitions in each mean. This should be noted for both X and Y.
- The standard error of estimate of the data, *calculated in the vertical* (Y) *direction*, if consistent throughout the claimed analytical measurement range; or the standard error in multiple concentration ranges, if the overall estimate is not appropriate.
- The confidence intervals on the slope and intercept.
- The confidence interval on the bias at each level.
- The correlation coefficient.
- A scatter plot of the observed data, using identical scales and ranges for the *x* and *y* axes, with *all* data indicated, including those data points identified as outliers with a different plot symbol. The scatter plot should include the fitted regression line (if appropriate) and the line of identity (X = Y).
- The method used to fit the linear regression line (ordinary least squares, weighted regression, Deming, orthogonal regression) and a scatter plot illustrating the line of best fit.
- The number of days and calibration cycles used to collect the data on the test (Y) method.

**Table 1a. Suggested Distribution of Data for Comparison of Methods Experiment (Mass Concentration)**

| | Group A | | Group B | | Group C | | Group D | | Group E | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Test** | **Range** | **%** | **Range** | **%** | **Range** | **%** | **Range** | **%** | **Range** | **%** |
| Glucose (mg/dL) | <50 | 10 | 51-110 | 40 | 111-150 | 30 | 151-250 | 10 | 251-SL | 10 |
| BUN (mg/dL) | <15 | 10 | 15-25 | 40 | 26-50 | 20 | 51-100 | 20 | 100-SL | 10 |
| $Na^+$ (mmol/L) | 120-130 | 20 | | | 131-140 | 40 | 141-150 | 30 | 151-160 | 10 |
| $K^+$ (mmol/L) | <3.0 | 20 | 3-4.5 | 35 | 4.5-6.0 | 35 | >6 | 10 | | |
| $Cl^-$ (mmol/L) | 80-95 | 30 | 95-105 | 40 | 105->120 | 30 | | | | |
| $CO_2$ (mmol/L) | <15 | 10 | 15-20 | 30 | 20-30 | 40 | 30-40 | 10 | >40-SL | 10 |
| Uric acid (mg/dL) | <3.0 | 20 | 3-5 | 20 | 5-8 | 20 | 8-10 | 20 | >10-SL | 20 |
| Calcium (mg/dL) | <8.0 | 10 | 8-9 | 20 | 9-11 | 40 | 11-13 | 20 | >13-SL | 10 |
| Inorganic phosphates (mg/dL) | <2.5 | 10 | 2.5-4.5 | 60 | 4.5-6.5 | 20 | >6.5 | 10 | | |
| Alkaline phosphatase (U/dL) | <NL/2 | 30 | NL-2NL | 20 | NL-2NL | 20 | 2NL-4NL | 20 | 4NL-SL | 10 |
| Total protein (g/dL) | <5 | 10 | 5-7 | 40 | 7-9 | 40 | >9 | 10 | | |
| Albumin (g/L) | <3 | 10 | 3-4 | 40 | 4-5 | 40 | >5 | 10 | | |
| Total bilirubin (mg/dL) | 0-1.0 | 30 | 1-2 | 30 | 2-5 | 20 | 5-10 | 10 | 10-SL | 10 |
| Cholesterol (mg/dL) | 120-180 | 20 | 181-220 | 30 | 221-260 | 30 | 261-400 | 20 | | |
| Triglycerides (mg/dL) | <75 | 10 | 75-125 | 30 | 125-200 | 30 | 200-300 | 20 | 300-SL | 10 |
| AST/SGOT (U/L) | NL/2 | 20 | NL/2-NL | 30 | NL/2-NL | 30 | 2NL-4NL | 10 | 4NL-SL | 10 |
| GGT (U/L) | | | 0-NL | 40 | NL/2-NL | 40 | 2NL-4NL | 10 | 4NL-SL | 10 |
| ALT/SGPT (U/L) | NL/2 | 20 | NL/2-NL | 20 | NL/2-NL | 40 | 2NL-4NL | 10 | 4NL-SL | 10 |
| LD (U/L) | NL/2 | 15 | NL/2-NL | 25 | NL/2-NL | 30 | 2NL-5NL | 20 | 5NL-SL | 10 |
| CK (U/L) | NL/2 | 15 | NL/2-NL | 25 | NL/2-NL | 30 | 2NL-5NL | 20 | 5NL-SL | 10 |
| Creatinine (mg/dL) | 0-1.0 | 20 | 1.1-2.5 | 30 | 2.5-5.0 | 20 | 5-10 | 20 | 10-SL | 10 |
| Fe (µg/dL) | <50 | 20 | 50-150 | 50 | 150.300 | 20 | 300-SL | 10 | | |
| Amylase (U/L) | | | 0-NL | 40 | NL/2-NL | 40 | 2NL-4NL | 10 | 4NL-SL | 10 |
| Hemoglobin (g/dL) | <9.0 | 15 | 9.1-12.0 | 25 | 12.1-17.0 | 50 | | | 17.1-SL | 10 |
| RBC ($x10^{12}$/L) | <3.0 | 10 | 3.1-4.0 | 30 | 4.1-6.0 | 50 | 6.1-SL | 10 | | |
| WBC ($x10^9$/L) | <2.0 | 10 | 2.1-5.0 | 20 | 5.1-11.0 | 40 | 11.1-25.0 | 20 | 25.1-SL | 10 |
| Platelets ($x10^9$/L) | <50.0 | 10 | 51.0-150.0 | 20 | 151.0-300.0 | 30 | 301.0-450.0 | 30 | 451.0-SL | 10 |

BUN, blood urea nitrogen
Na, sodium
K, potassium
Cl, chloride
$CO_2$, carbon dioxide
AST, aspartate aminotransferase
SL, scale limit
SGOT, serum glutamic oxaloacetic transaminase
GGT, gamma-glutamyl transferase
ALT, alanine aminotransferase
SGPT, serum glutamic pyruvate transaminase
LD, lactate dehydrogenase
CK, creatine kinase
NL, upper limit of laboratory's normal range
Fe, iron
RBC, red blood cell (count)
WBC, white blood cell (count)

**Table 1b. Suggested Distribution of Data for Comparison of Methods Experiment (Substance Concentration)**

| | Group A | | Group B | | Group C | | Group D | | Group E | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Test** | **Range** | **%** | **Range** | **%** | **Range** | **%** | **Range** | **%** | **Range** | **%** |
| Glucose (mg/dL) | <2.76 | 10 | 2.81-6.06 | 40 | 6.12-8.27 | 30 | 8.32-13.78 | 10 | 13.83-SL | 10 |
| BUN (mg/dL) | <2.50 | 10 | 2.50-4.16 | 40 | 4.33-8.33 | 20 | 8.50-16.65 | 20 | 16.65-SL | 10 |
| $Na^+$ (mmol/L) | 120-130 | 20 | | | 131-140 | 40 | 141-150 | 30 | 151-160 | 10 |
| $K^+$ (mmol/L) | <3.0 | 20 | 3-4.5 | 35 | 4.5-6.0 | 35 | >6 | 10 | | |
| $Cl^-$ (mmol/L) | 80-95 | 30 | 95-105 | 40 | 105->120 | 30 | | | | |
| $CO_2$ (mmol/L) | <15 | 10 | 15-20 | 30 | 20-30 | 40 | 30-40 | 10 | >40-SL | 10 |
| Uric acid (mg/dL) | <178 | 20 | 178-297 | 20 | 297-476 | 20 | 476-595 | 20 | >595-SL | 20 |
| Calcium (mg/dL) | <2.0 | 10 | 2.0-2.25 | 20 | 2.25-2.75 | 40 | 2.75-3.24 | 20 | >3.24-SL | 10 |
| Inorganic phosphates (mg/dL) | <0.8 | 10 | 0.8-1.5 | 60 | 1.5-2.1 | 20 | >2.1 | 10 | | |
| Alkaline phosphatase (U/dL) | <NL/2 | 30 | NL-2NL | 20 | NL-2NL | 20 | 2NL-4NL | 20 | 4NL-SL | 10 |
| Total protein (g/dL) | <50 | 10 | 50-70 | 40 | 70-90 | 40 | >90 | 10 | | |
| Albumin (g/L) | <435 | 10 | 435-580 | 40 | 580-725 | 40 | >725 | 10 | | |
| Total bilirubin (mg/dL) | 0-17.1 | 30 | 17.1-34.2 | 30 | 34.2-85.5 | 20 | 85.5-171 | 10 | 171-SL | 10 |
| Cholesterol (mg/dL) | <3.9 | 10 | 3.9-6.5 | 40 | 6.5-9.1 | 30 | >9.1 | 20 | | |
| *Triglycerides (mg/dL) | <0.086 | 10 | 0.086-0.14 | 30 | 10.14-0.23 | 30 | 0.23-0.34 | 20 | 0.34-SL | 10 |
| AST/SGOT (U/L) | NL/2 | 20 | NL/2-NL | 30 | NL/2-NL | 30 | 2NL-4NL | 10 | 4NL-SL | 10 |
| GGT (U/L) | | | 0-NL | 40 | NL/2-NL | 40 | 2NL-4NL | 10 | 4NL-SL | 10 |
| ALT/SGPT (U/L) | NL/2 | 20 | NL/2-NL | 20 | NL/2-NL | 40 | 2NL-4NL | 10 | 4NL-SL | 10 |
| LD (U/L) | NL/2 | 15 | NL/2-NL | 25 | NL/2-NL | 30 | 2NL-5NL | 20 | 5NL-SL | 10 |
| CK (U/L) | NL/2 | 15 | NL/2-NL | 25 | NL/2-NL | 30 | 2NL-5NL | 20 | 5NL-SL | 10 |
| Creatinine (mg/dL) | 0-88.4 | 20 | 97.2-221 | 30 | 221-442 | 20 | 442-884 | 20 | 884-SL | 10 |
| Fe (μg/dL) | <8.95 | 20 | 8.95-26.9 | 50 | 26.9-53.7 | 20 | 53.7-SL | 10 | | |
| Amylase (U/L) | | | 0-NL | 40 | NL/2-NL | 40 | 2NL-4NL | 10 | 4NL-SL | 10 |
| Hemoglobin (g/dL) | <5.58 | 15 | 5.65-7.45 | 25 | 7.50-10.55 | 50 | | | 10.61-SL | 10 |
| RBC ($x10^{12}$/L) | <3.0 | 10 | 3.1-4.0 | 30 | 4.1-6.0 | 50 | 6.1-SL | 10 | | |
| WBC ($x10^9$/L) | <2.0 | 10 | 2.1-5.0 | 20 | 5.1-11.0 | 40 | 11.1-25.0 | 20 | 25.1-SL | 10 |
| Platelets ($x10^9$/L) | <50.0 | 10 | 51.0-150.0 | 20 | 151.0-300.0 | 30 | 301.0-450.0 | 30 | 451.0-SL | 10 |

*Based on relative molecular mass=875.

**NOTE:** The proposed unit for catalytic amount (enzymatic activity) is the katal (symbol: kat). There is, as yet, no general acceptance of this unit. 1 (conventional) U=16.67 nkat.

BUN, blood urea nitrogen
Na, sodium
K, potassium
Cl, chloride
$CO_2$, carbon dioxide
AST, aspartate aminotransferase
SL, scale limit
SGOT, serum glutamic oxaloacetic transaminase
GGT, gamma-glutamyl transferase
ALT, alanine aminotransferase
SGPT, serum glutamic pyruvate transaminase
LD, lactate dehydrogenase
CK, creatine kinase
NL, upper limit of laboratory's normal range
Fe, iron
RBC, red blood cell (count)
WBC, white blood cell (count)

## References


[1] Linnet K. Evaluation of regression procedures for method comparison studies. *Clin Chem.* 1993;39:424-432.

[2] Mandel J. *The Statistical Analysis of Experimental Data*. Dover, New York: 1964:282-292.

[3] Passing H, Bablok W. A new biometrical procedure for testing the equality of measurements from two different analytical methods. *J Clin Chem Clin Biochem*. 1983;21:709-720.

[4] Beyer WH, Ed. *CRC Standard Probability and Statistics Tables and Formulae*. Boca Raton, Florida: CRC Press. 1999;Table VIII.2:270.

[5] Bland JM, Altman DG Statistical method for assessing agreement between two methods of clinical measurement. *Lancet.* 1986:307-310.

[6] Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. *The Statistician.* 1983;32:307-317.

# Appendix A. Sample Data Recording Sheet

## A1. Example: Blank Worksheet

Sheet #____of____

| Date(s): | Analyte: |
|---|---|
| Test method: | |
| Comparative method: | |

| | Test Method | | Comparative Method | | Test Method (Y) | Comparative Method (X) |
|---|---|---|---|---|---|---|
| Sample # | Result 1 | Result 2 | Result 1 | Result 2 | Mean | Mean |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |

# Appendix A. (Continued)

## A2. Example: Completed Sample Data Recording Sheet

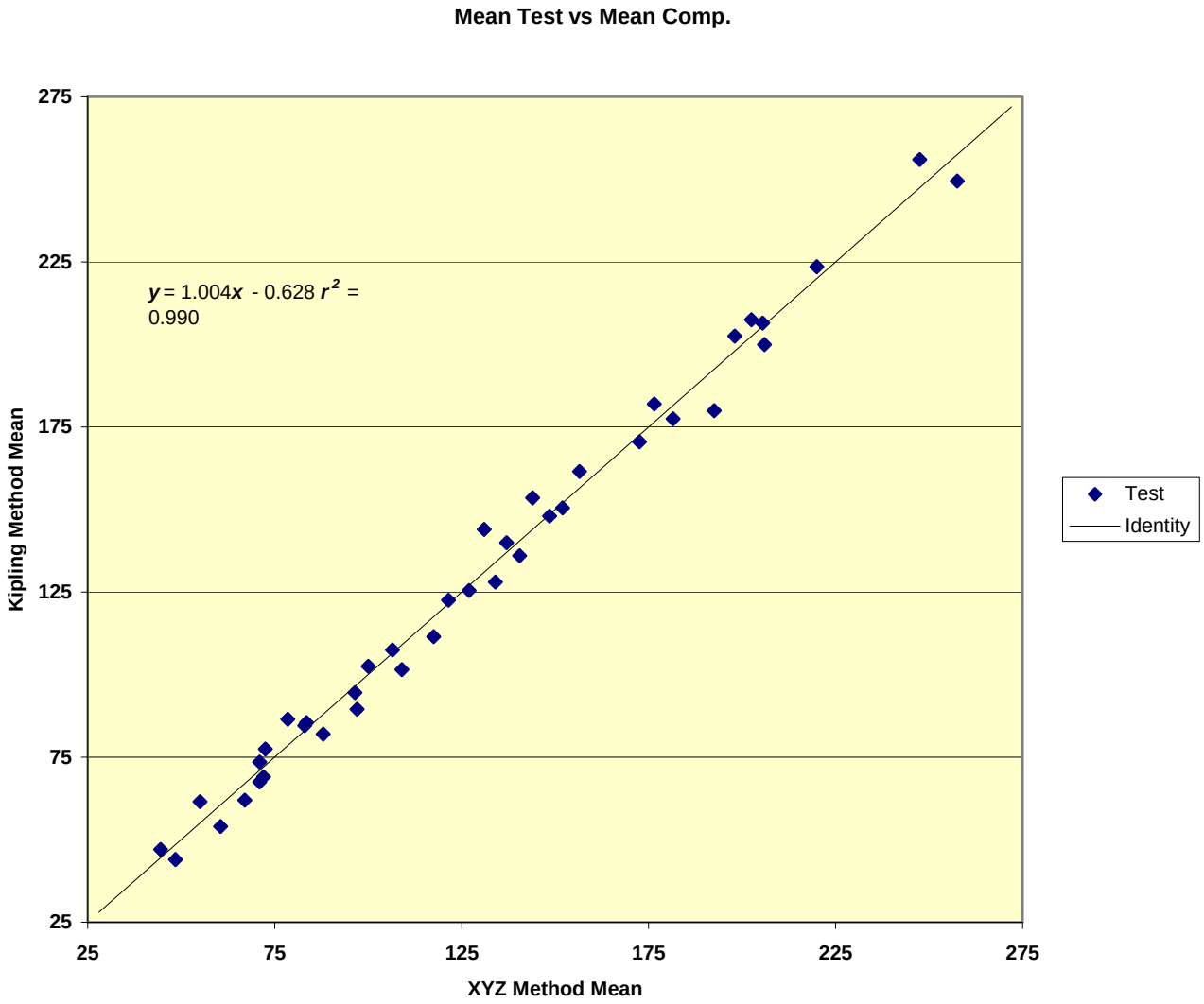
Sheet # 1 of 2

| | |
|---|---|
| **Date(s):** 1/29/93 | **Analyte:** Calculation Example |
| **Test method:** Kipling | |
| **Comparative method:** XYZ | |

| | Test Method | | Comparative Method | | Test Method (Y) Absolute Value Rep1-Rep2 | Comparative Method (X) Absolute Value |
|---|---|---|---|---|---|---|
| **Sample #** | **Result 1** | **Result 2** | **Result 1** | **Result 2** | **Mean** | **Mean** |
| 1 | 87 | 82 | 86 | 80 | 5 | 6 |
| 2 | 165 | 158 | 155 | 158 | 7 | 3 |
| 3 | 197 | 208 | 202 | 194 | 11 | 8 |
| 4 | 43 | 45 | 47 | 50 | 2 | 3 |
| 5 | 68 | 70 | 72 | 72 | 2 | 0 |
| 6 | 184 | 180 | 176 | 177 | 4 | 1 |
| 7 | 227 | 220 | 218 | 222 | 7 | 4 |
| 8 | 140 | 140 | 136 | 138 | 0 | 2 |
| 9 | 168 | 173 | 175 | 170 | 5 | 5 |
| 10 | 87 | 86 | 79 | 78 | 1 | 1 |
| 11 | 144 | 152 | 147 | 150 | 8 | 3 |
| 12 | 264 | 248 | 250 | 245 | 16 | 5 |
| 13 | 45 | 49 | 45 | 44 | 4 | 1 |
| 14 | 92 | 87 | 98 | 96 | 5 | 2 |
| 15 | 74 | 73 | 69 | 73 | 1 | 4 |
| 16 | 63 | 60 | 53 | 57 | 3 | 4 |
| 17 | 147 | 154 | 149 | 155 | 7 | 6 |
| 18 | 204 | 209 | 200 | 211 | 5 | 11 |
| 19 | 106 | 97 | 110 | 108 | 9 | 2 |
| 20 | 125 | 120 | 123 | 120 | 5 | 3 |
| 21 | 132 | 124 | 136 | 132 | 8 | 4 |
| 22 | 101 | 104 | 98 | 102 | 3 | 4 |
| 23 | 211 | 204 | 199 | 206 | 7 | 7 |
| 24 | 67 | 68 | 72 | 70 | 1 | 2 |
| 25 | 184 | 176 | 192 | 193 | 8 | 1 |
| 26 | 97 | 92 | 95 | 98 | 5 | 3 |
| 27 | 143 | 145 | 132 | 130 | 2 | 2 |
| 28 | 106 | 117 | 113 | 122 | 11 | 9 |
| 29 | 84 | 80 | 86 | 90 | 4 | 4 |
| 30 | 201 | 199 | 207 | 205 | 2 | 2 |
| 31 | 154 | 153 | 147 | 141 | 1 | 6 |
| 32 | 76 | 79 | 75 | 70 | 3 | 5 |
| 33 | 55 | 53 | 62 | 59 | 2 | 3 |
| 34 | 181 | 174 | 179 | 184 | 7 | 5 |
| 35 | 243 | 256 | 261 | 254 | 13 | 7 |
| 36 | 127 | 124 | 128 | 126 | 3 | 2 |
| 37 | 84 | 87 | 85 | 82 | 3 | 3 |
| 38 | 62 | 62 | 68 | 66 | 0 | 2 |
| 39 | 137 | 135 | 138 | 143 | 2 | 5 |
| 40 | 104 | 111 | 106 | 107 | 7 | 1 |

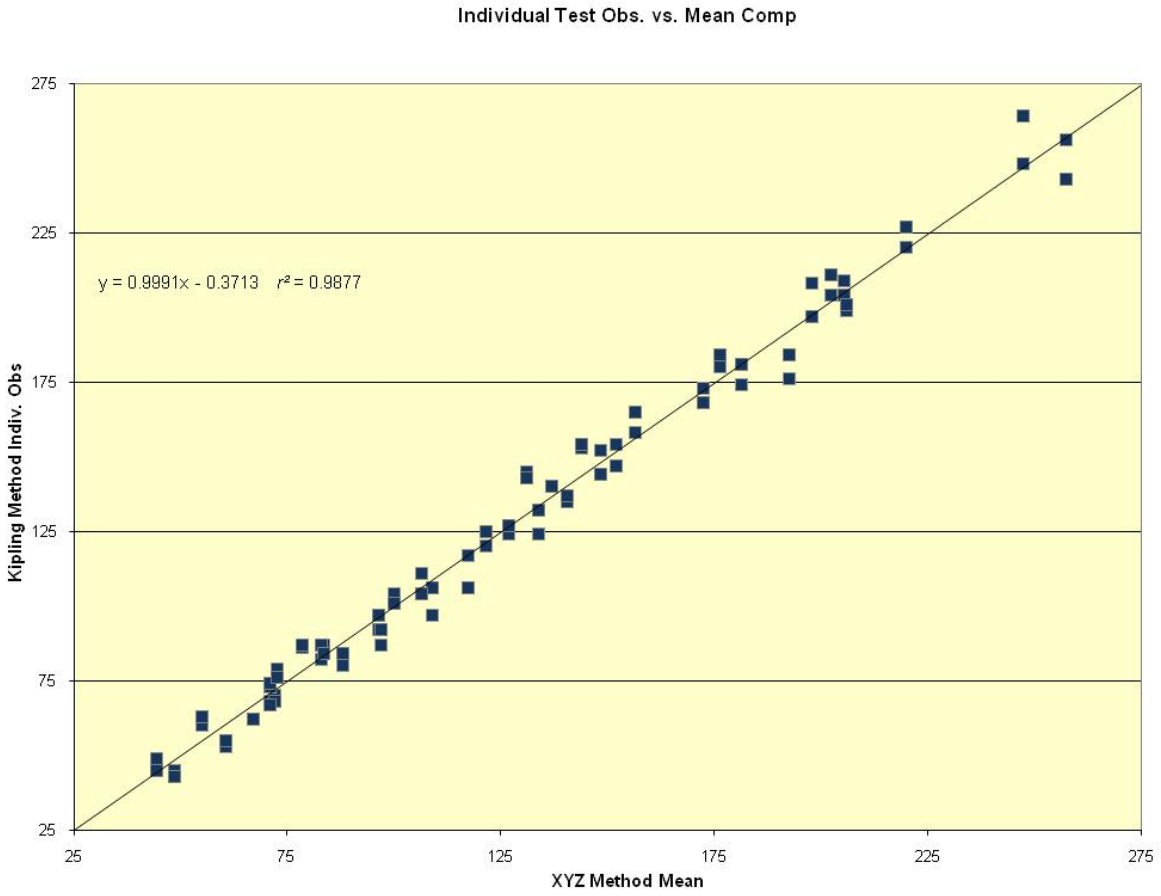
# Appendix B. Scatter Plots Derived from Example

## B1. Scatter Plot for Mean of Replicates From Example



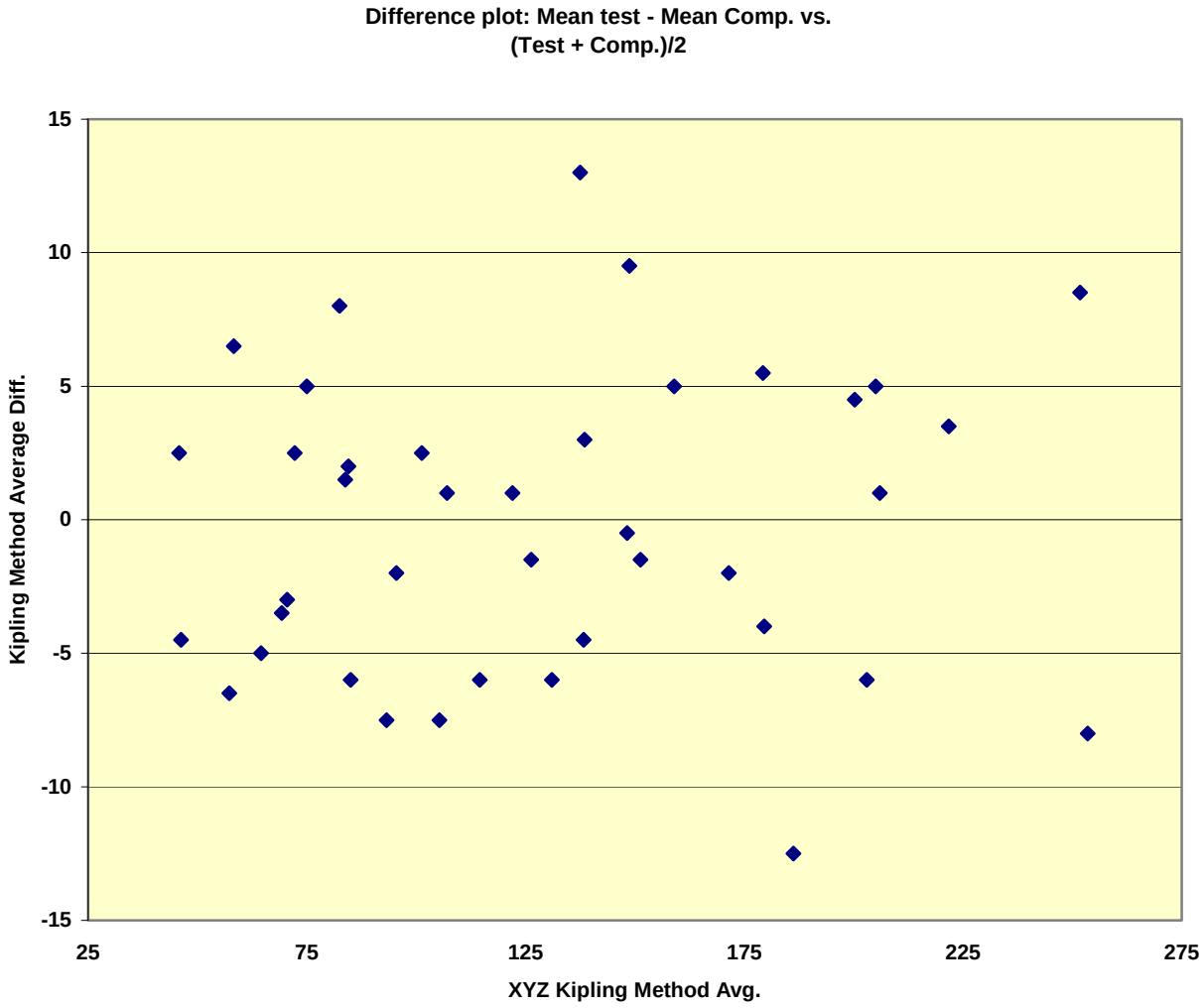
# Appendix B. (Continued)

## B2. Scatter Plot for All Results From Example



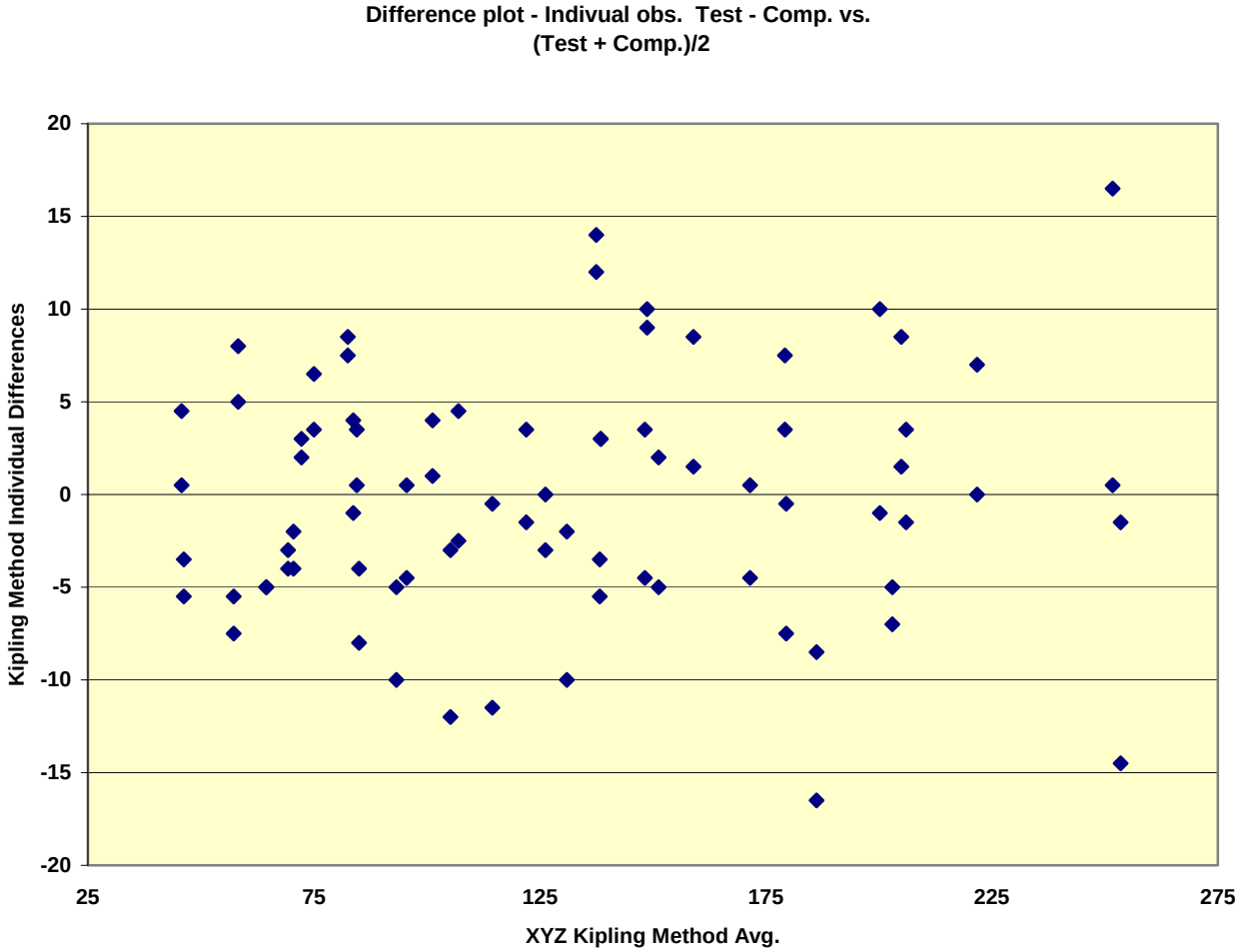
## Appendix B. (Continued)

### B3. Bias Plot Means-of-Replicate Deltas Versus Mean of Test and Comparative Method



## Appendix B. (Continued)

### B4. Bias Plot—Individual Results Deltas Versus Mean of Test and Comparative Method



# Appendix C. Calculation Example

## C1. Within-Method Duplicates Check (Section 4.1)

$x_{11} = 86 \quad x_{12} = 80$

$y_{11} = 87 \quad y_{12} = 82$

$$DX_1 = |x_{11} - x_{12}| = |86 - 80| = 6 \quad \bar{x}_1 = \frac{(x_{11} + x_{12})}{2} = 83$$

$$DY_1 = |y_{11} - y_{12}| = |87 - 82| = 5 \quad \bar{y}_i = \frac{(y_{11} + y_{12})}{2} = 84.5$$

$$DX'_1 = \frac{|x_{11} - x_{12}|}{\bar{x}_i} = \frac{6}{83} = 0.0723$$

$$DY'_1 = \frac{|y_{11} - y_{12}|}{\bar{y}_i} = \frac{5}{84.5} = 0.0592$$

Similarly,

| i | $x_{i1}$ | $x_{i2}$ | $y_{i1}$ | $y_{i2}$ | $DX_i$ | $DY_i$ | $DX_i'$ | $DY_i'$ |
|---|---|---|---|---|---|---|---|---|
| 2 | 155 | 158 | 165 | 158 | 3 | 7 | 0.0192 | 0.0433 |
| 3 | 202 | 194 | 197 | 208 | 8 | 11 | 0.0404 | 0.0543 |
| 4 | 47 | 50 | 43 | 45 | 3 | 2 | 0.0619 | 0.0455 |
| • | • | • | • | • | • | • | • | • |
| • | • | • | • | • | • | • | • | • |
| • | • | • | • | • | • | • | • | • |
| 38 | 68 | 66 | 62 | 62 | 2 | 0 | 0.0299 | 0 |
| 39 | 138 | 143 | 137 | 135 | 5 | 2 | 0.0356 | 0.0147 |
| 40 | 106 | 107 | 104 | 111 | 1 | 7 | 0.0094 | 0.0651 |

$\overline{DX} = 3.775$ Control limit $= 4 \cdot \overline{DX} = 15.1$ Rounded up $= \underline{16}$

$\overline{DY} = 4.975$ Control limit $= 4 \cdot \overline{DY} = 19.9$ Rounded up $= \underline{20}$

$\overline{DX'} = 0.0320$ Control limit $= 4 \cdot \overline{DX'} = \underline{0.1280}$

$\overline{DY'} = 0.0392$ Control limit $= 4 \cdot \overline{DY'} = \underline{0.1567}$

No duplicates exceeded both control limits.

## Appendix C. (Continued)

### C2. Test for Outliers (Section 4.4)

$x_{11} = 86$ $x_{12} = 80$

$y_{11} = 87$ $y_{12} = 82$

$E_{11} = |y_{11} - x_{11}| = |87 - 86| = 1$

$E_{12} = |y_{12} - x_{12}| = |82 - 80| = 2$

$$E_{1'1} = \frac{|y_{11} - x_{11}|}{x_{11}} = \frac{1}{86} = 0.0116$$

$$E_{1'2} = \frac{|y_{12} - x_{12}|}{x_{12}} = \frac{2}{80} = 0.0250$$

Similarly,

| #/i | $E_{i1}$ | $E_{i2}$ | $E_{i1'}$ | $E_{i2'}$ |
|---|---|---|---|---|
| 2 | 10 | 0 | 0.0645 | 0 |
| 3 | 5 | 14 | 0.0248 | 0.0722 |
| 4 | 4 | 5 | 0.0851 | 0.1000 |
| • | • | • | • | • |
| • | • | • | • | • |
| • | • | • | • | • |
| 38 | 6 | 4 | 0.0882 | 0.0606 |
| 39 | 1 | 8 | 0.0072 | 0.0559 |
| 40 | 2 | 4 | 0.0189 | 0.0374 |

$$E = \frac{1}{80} \cdot \sum_{i=1}^{40} \sum_{j=1}^{2} E_{ij} = \frac{1}{80} \cdot 428 = 5.35$$

$$E' = \frac{1}{80} \cdot \sum_{i=1}^{40} \sum_{j=1}^{2} E'_{ij} = \frac{1}{80} \cdot 3.6839 = 0.0473$$

Control limit for E = $4 \cdot E$ = 21.4 Rounded up = <u>22</u>

Control limit for E' = $4 \cdot E'$ = <u>0.1892</u>

No duplicates exceeded both control limits.

## Appendix C. (Continued)

### C3. Adequate Range Test-Correlation (Section 4.5)

$$r = \frac{\sum_{i=1}^{N}(\bar{x}_i - \bar{x})(\bar{y}_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(\bar{x}_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(\bar{y}_i - \bar{y})^2}}$$

| $i$ | $(\bar{x}_i)$ | $(\bar{x}_i - \bar{x})$ | $(\bar{x}_i - \bar{x})^2$ | $(\bar{y}_i)$ | $(\bar{y}_i - \bar{y})$ | $(\bar{y}_i - \bar{y})^2$ | $(\bar{x}_i - \bar{x})(\bar{y}_i - \bar{y})$ |
|---|---|---|---|---|---|---|---|
| 1 | 83 | -46.34 | 2147.3956 | 84.5 | -44.66 | 1994.5156 | 2069.5444 |
| 2 | 156.5 | 27.16 | 737.6656 | 161.5 | 32.34 | 1045.8756 | 878.3544 |
| 3 | 198 | 68.66 | 4714.1956 | 202.5 | 73.34 | 5378.7556 | 5035.5244 |
| 4 | 48.5 | -80.84 | 6535.1056 | 44 | -85.16 | 7252.2256 | 6884.3344 |
| 5 | 72 | -57.34 | 3287.8756 | 69 | -60.16 | 3619.2256 | 3449.5744 |
| 6 | 176.5 | 47.16 | 2224.0656 | 182 | 52.84 | 2792.0656 | 2491.9344 |
| 7 | 220 | 90.66 | 8219.2356 | 223.5 | 94.34 | 8900.0356 | 8552.8644 |
| • | • | • | • | • | • | • | • |
| • | • | • | • | • | • | • | • |
| • | • | • | • | • | • | • | • |
| • | • | • | • | • | • | • | • |
| • | • | • | • | • | • | • | • |
| • | • | • | • | • | • | • | • |
| 34 | 181.5 | 52.16 | 2720.6656 | 177.5 | 48.34 | 2336.7556 | 2521.4144 |
| 35 | 257.5 | 128.16 | 16424.9856 | 249.5 | 120.34 | 14481.7156 | 15422.7744 |
| 36 | 127 | -2.34 | 5.4756 | 125.5 | -3.66 | 13.3956 | 8.5644 |
| 37 | 83.5 | -45.84 | 2101.3056 | 85.5 | -43.66 | 1906.1956 | 2001.3744 |
| 38 | 67 | -62.34 | 3886.2756 | 62 | -67.16 | 4510.4656 | 4186.7544 |
| 39 | 140.5 | 11.16 | 124.5456 | 136 | 6.84 | 46.7856 | 76.3344 |
| 40 | 106.5 | -22.84 | 521.6656 | 107.5 | -21.66 | 469.1556 | 494.7144 |

$\bar{x} = 129.34$

$\bar{y} = 129.16$

$\sum(\bar{x}_i - \bar{x})^2 = 127067.69$ $\sum(\bar{y}_i - \bar{y})^2 = 129204.19$

$\sum(\bar{x}_i - \bar{x})(\bar{y}_i - \bar{y}) = 127513.06$

$$r = \frac{127513.06}{\sqrt{127067.69}\sqrt{129204.19}} = 0.995$$ (Data pass adequate range test)

## Appendix C. (Continued)

### C4. Regression Parameter Estimates (Section 5.1)

Slope (b): Using calculated data from previous page,

$$b = \frac{\sum_{i=1}^{N}[(\bar{x}_i - \bar{x})(\bar{y}_i - \bar{y})]}{\sum_{i=1}^{N}(\bar{x}_i - \bar{x})^2}$$

$$= \frac{127513.056}{127067.694} = 1.003504 = 1.003$$

Intercept (a)

$$a = \bar{y} - b \cdot \bar{x} = 129.1625 - 1.00354 \cdot 129.3375$$

$$= 126.1625 - 126.7954$$

$$= -0.6329$$

$$= -0.63$$

## Appendix C. (Continued)

### C5. Residuals and Standard Error of Estimate $(s_{y\cdot x})$ (Section 6.1)

Predicted Values: $\hat{Y}_i = a + b \cdot \bar{x}_i = -0.63 + 1.003 \cdot \bar{x}_i$

Residual $_i = \bar{y}_i - \hat{Y}_i$

| $i$ | $\bar{y}_i$ | $\hat{Y}_i$ | Residual$_i$ | $\left(\bar{y}_i - \hat{Y}_i\right)^2$ |
|---|---|---|---|---|
| 1 | 84.5 | 82.619 | 1.881 | 3.538161 |
| 2 | 161.5 | 156.3395 | 5.1605 | 26.63076 |
| 3 | 202.5 | 197.964 | 4.536 | 20.5753 |
| 4 | 44 | 48.0155 | −4.0155 | 16.12424 |
| 5 | 69 | 71.586 | −2.586 | 6.687396 |
| • | • | • | • | • |
| • | • | • | • | • |
| • | • | • | • | • |
| • | • | • | • | • |
| • | • | • | • | • |
| 37 | 85.5 | 83.1205 | 2.3795 | 5.66202 |
| 38 | 62 | 66.571 | −4.571 | 20.89404 |
| 39 | 136 | 140.2915 | −4.2915 | 18.41697 |
| 40 | 107.5 | 106.1895 | 1.3105 | 1.71741 |

Sum of the squared residuals = $\sum_{i=1}^{40} (\bar{y} - \hat{Y}_i)^2 = 1244.4259$

Degrees of freedom = $N$−2 = 38

Sample standard deviation from the regression =

$$s_{y\cdot x} = \sqrt{\frac{1244.4259}{38}} = 5.7225 = 5.7$$

Bias at a decision level "$c$" of 150 = $B_c$
$= -0.6329 + 1.003504 \cdot 150 - 150$
= −0.1073
= −0.11

The lower 95% limit of confidence interval of bias estimate

$$B_c - 2 \cdot s_{y\cdot x} \sqrt{(1/N) + (X_c - \bar{x})^2 / \sum_{i=1}^{40} (\bar{x}_i - \bar{x})^2}$$

$$= -0.1073 - 2 \cdot 5.7225 \sqrt{(1/40) + (150 - 129.34)^2 / 127067.694}$$
= −0.1073 – 1.927
= −2.035
= −2.04

The upper 95% limit of confidence interval of bias estimate
= −0.1073 + 1.927
= 1.820
= 1.82

## Appendix D. Calculation of Deming Slope

The following formulas can be used in a spreadsheet to calculate the Deming slope and intercept.

Calculate the ordinary least squares slope with y as the dependent and x as the independent variables = $b_{yx.}$

Calculate the ordinary least squares slope with x as the dependent and y as the independent variables = $b_{xy.}$

Define $f = \frac{1}{b_{xy}} - \frac{\lambda}{b_{yx}}$

where $\lambda$ = the ratio of the error variance (imprecision) in the y variable to the error variance (imprecision) in the x variable. The imprecision can be calculated as the variance between replicate results, pooled across samples.

The Deming slope $b_d = 0.5 \times (f + \sqrt{f^2 + (4 \times \lambda)})$

The Deming intercept $= \bar{y} - (b_d \times \bar{x})$

### Reference for Appendix D

Parvin C. A direct comparison of two slope-estimation techniques used in method-comparison studies. *Clin Chem.* 1984;30:751-754.

**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 Working Group Responses

EP09-A: *Method Comparison and Bias Estimation Using Patient Samples; Approved Guideline*

General

1. In general, the standard is a good attempt to provide a uniform implementation of Comparison and Bias Estimation using patient samples. However, the working group may wish to review the statistical application from the hematologist's viewpoint. Alternative methods, such as binomial, have proved to be more useful that some more standard methods (linear regression).

- **See the response to Comment 71.**

2. The general methodology of "regression" is used extensively throughout the document. It is well known in our industry that these methods are often inappropriate for many measurements used in blood cell analysis. Therefore, it would be helpful if:

   a) the document scope explicitly identified parameters for which the standard is appropriate; and
   b) for all parameters in scope above, some minimal preliminary experimental data should be obtained or references cited supporting the proposed methodology prior to external reviews.

- **The working group has given guidelines for appropriate use of the regressions methods. This is more useful than a list of analytes suitable for regression analysis.**

3. I find that customers and laboratory evaluators find it very difficult to follow documents with calculations but without data examples.

- **There is a data example in the appendix.**

4. Please include a flowchart that shows where "bias evaluation" fits into the overall instrument evaluation (e.g., do it first, last, or in the middle!).

- **This suggestion is beyond the scope of the document; however, general guidance is given in Section 1.**

5. References to the articles in the bibliography should be made in the body of the document.

- **The working group agrees with the commenter. References have been inserted where appropriate.**

Section 1.1

6. To be consistent with other documents, the term "sample" should be replaced with "specimen." The term "sample" denotes an aliquot of a "specimen."

- **The working group disagrees. In keeping with NCCLS's commitment to harmonization, the term "specimen" has replaced the term "sample."**

7. EP9 recommends a minimum of 40 samples where as H20 recommends a total of 200 of which 100 are normal and 100 abnormal. For consistency within the NCCLS organization, these differing recommendations should be reconciled. From our experience, 120 samples representing tertiary care patients generally covers the expected range.

- **The comment for consistency across documents is valid and will be passed on to the Area Committee on Hematology. However, an important issue for any evaluation is always "what is the minimum number of samples required?" A usual answer to this question depends on many other parameters such as specific parameter(s) to assess, confidence level desired, cost, and perhaps most importantly, how representative the sampling is. The minimum of 40 samples is a consensus opinion of the committee.**

8. In general, the "n" required depends upon the objectives of the test. In this case, the objectives may be to determine a particular "precision," "accuracy," "sensitivity," etc. at some statistical level of confidence. From these objectives, the "n" requirement can be determined. Perhaps even more globally, an objective may be driven based upon ...establishing a "Clinical Level of Significance" (required) which ... dictates "precision," "accuracy," etc., requirements which ... dictate sample size "n" at some level of significance.

- **The working group agrees. See the response to Comment 7.**

9. An "n" of 40 is inappropriate for some recommended statistical analyses mentioned later in the draft document. An "n" = 40 would also prove inconsistent for use by manufacturer's for validation and research purposes. An "n" = 40 may be appropriate for some customers depending on the recommended statistical analysis and sample population.

- **The working group agrees. See the response to Comment 7.**

10. The straightaway recommendation of regression analysis may not be appropriate if the spread of the data is not sufficient. Since the paragraph is readable without this sentence, the working group may wish to consider deletion. Alternatively, the working group may cite Section 4.5 of the document that limits the regression application.

- **The working group agrees. The sentence "Analyze the data to detect outliers, and fit a regression line to the data" has been deleted.**

Section 1. 2

11. The definition for $Dx_i$, or $Dy_i$ should read "**absolute** difference between duplicates for method X or Y."

- **The definition has been changed to read "absolute value of the difference between duplicates for methods X or Y."**

Section 2

12. I disagree with the following sentence, "This procedure is not necessary for the user evaluations." Device familiarization is essential for proper instrument operation and understanding of results (numbers, flags, etc.).

- **The working group agrees. The sentence has been changed to read, "This procedure may not be necessary for all user evaluations."**

Section 3.1

13. Patient samples must also be collected and handled in the manner consistent with manufacturer's recommendations.

- **The working group agrees. The sentence "Collect and handle patient samples according to accepted laboratory practice" has been changed to read, "Collect and handle patient samples according to accepted laboratory practice and to the manufacturer's recommendations."**

Section 3.1.2

14. Prequalification of patient samples has the tendency to truncate the experiment design.

- **Section 3.1.2 has been rewritten to address the commenter's concern.**

15. The working group may wish to recommend recording: [1] sample age (time elapsed post phlebotomy), [2] sample condition (e.g., hemolysis) and [3] patient identification (blind code samples to ensure anonymity). If an outlier is detected, this recorded information will provide a database for identification of interfering substances) which may include disease state and therapy. Once the reason of interference has been documented, the sample results may be removed from the database before final analysis is undertaken.

- **The working group agrees. Section 3.1.2 has been rewritten to read, "If a sample is excluded, record the reason for the exclusion."**

16. I would clarify that samples with known interfering substances regarding the "Gold Standard" Comparator should not be used in the evaluation since the "Gold Standard" Comparator would be in error thus negating any meaningful conclusions. It should be understood, however, that samples that may interfere with the "Instrument Under Test (IUT)" results but do not interfere on the "Gold Standard" Comparator may be acceptable, and are probably desirable depending on test objectives.

- **The working group disagrees with this comment. Observed bias may or may not be consistent with true differences in the actual analyte concentration, regardless of the quality of the comparison method. Yet these biases are important to report, especially since one is often assessing a proposed method with an existing one, used to report patient results and contribute to medical decision making.**

17. NCCLS and other consensus bodies have provided detailed procedures for the evaluation of interfering substances. By the time a bias evaluation begins, the interfering substances for both the Comparator and test method should have been defined. Samples with these substances should be excluded from the evaluation altogether.

- **The working group disagrees with this comment as it would be impractical. If 20 drugs interfered, one would either have to assay for all drugs or review patient records.**

18. If either the Comparator or test method lists a substance that is specific to only one of the methods, a procedure should be provided for the laboratory to gain evidence about the performance of those specific samples. Additionally, NCCLS may wish to suggest alternate methods for evaluation of such samples.

- **See the response to Comment 15.**

Section 3.2

19. Define confidence in the following sentence. "This experiment gives an estimate of the bias between two methods and the confidence for the bias, at any particular concentration."

- **The term "confidence" has been changed to "confidence interval" in Section 3.2 to address the commenter's concern.**

20. As the performance of devices improves, the newer generations generally have improved precision; therefore, requiring that the Comparator (previous generation) be more precise may be an inconsistency. Alternatively, the working group may wish to use the Root-Sum-Square (RSS) of Test and Comparator imprecision to predict limits of agreement. Manufacturers continue to seek new technologies with better performance characteristics than those of existing (often Reference) methods. However, the precision should be "accounted" for, since there are many statistical techniques that can compare and measure bias for tests with different precision.

- **The observed precision of any device can be improved with replication. Therefore, the bullet has been changed to read, "Have better precision than the test method, which can be achieved by replication, if needed."**

21. The requirement "to be free from known interference" essentially dictates that the Comparator system shall be a method of high order such as a Reference method which may not be generally available in a clinical laboratory. The working group may wish to rephrase: "[Comparator] Interfering substances should be known and controlled."

- **The working group has modified this bullet to read, "To be free from known interferences, whenever possible."**

22. For evaluations where reporting units differ, the reporting values can be converted to Standard Deviation Index (SDI; College of American Pathologists terminology for "Z" factor; see formula 1). Since the reporting units cancel during the transform process, unitless data is thus made available for analysis.

$$\mathrm{SDI} = \frac{(\mathrm{Patient}_{\mathrm{Assay}}) - (\mathrm{Reference}_{\mathrm{Interval}_{\mathrm{Mean}}})}{\mathrm{Reference}_{\mathrm{Interval}_{\mathrm{SD}}}}$$

- **The working group does not see the benefit of this transformation.**

Section 3.3

23. The working group defines "clinically meaningful range" but fails to define "analytical measurement range."

- **A definition for "analytical measurement range" has been added.**

24. The recommended distribution in Table 1 does not adequately cover abnormalities for reticulocyte and white cell differentials, let alone upcoming flow parameters (e.g., CD4).

- **The area committee will update the tables as suggested in a future revision of the document**.

25. The hemoglobin in table lb should be in g/dL and not mmol/L.

- **The working group disagrees. The units for hemoglobin are mmol/L.**

Section 3.3.2

26. As technology improves, it provides the manufacturers with the means to improve on their design and offer better performance. Extension of the analytical measurement range is a natural corollary. Being restricted by the analytical measurement range of an older Comparative method or device would prove limiting to advancement. Dilution of samples using the Comparative method would be one solution to the issue of limited analytical measurement range. If this is not advocated, please elaborate and provide further instruction of how both manufacturers and users could study these extended ranges.

- **The working group agrees with this comment. Section 3.3.2 has been deleted.**

27. The clinical samples collected should cover or be slightly wider than the narrowest analytical measurement range. If the analytical measurement range recommended by the manufacturer for the test method is wider than the Comparator, alternative methods (e.g., linearity, standard materials) must be found to validate the clinical range.

- **See the response to Comment 26.**

Section 3.4.1

28. It should be mentioned that the samples should be taken from a homogeneous sample.

- **It is the opinion of the working group that this is part of good laboratory practice and does not need to be mentioned.**

Section 3.4.2

29. Since this document may be used by other laboratory disciplines (Hematology) the working group may which to add "serologic compatibility" when whole bloods are pooled.

- **A sentence has been added to the end of Section 3.4.2 to read, "If the samples are whole blood, mixing requires serologic compatibility."**

30. Remove references to pooled samples unless covered in detail (e.g., ABO and Rh Compatibility).

- **See the response to Comment 29.**

Section 3.5

31. We concur that "Reversing" or "Randomizing" the second sample reduces systematic carryover but it may also increase random error. For example, a high sample, independent of sample position, will potentially give carryover interference. Therefore, if the samples are reversed or randomized, this high carryover also becomes randomized and thus more difficult to detect. In our studies, we analyze duplicates sequentially and subsequently review the data for patterns that suggest drift and/or carryover interference.

- **The working group has modified the text for clarity.**

32. Move the last sentence in paragraph 1 to follow the second sentence. (The last sentence would become the third sentence.)

- **The working group agrees and has made the suggested change.**

33. Alternatively, the ordering of the second aliquot could be randomized rather than in the reverse order.

- **The sequence chosen is to minimize linear drift and carryover, which randomization would not achieve.**

Section 4.1

34. For hematology the 'acceptable' limits appear to be less useful because generally the mean difference between pairs is essentially zero. Therefore, the "four rule" would exclude the majority of samples that did not have zero difference. In our experimental designs, we examine the distribution of differences and apply the appropriate statistical method (parametric, nonparametric). Values that exceed the 99% confidence limits are investigated. If the error-source is a human error, the sample is deleted from the database with adequate documentation.

- **A sentence has been added to Section 4.1 to read "The analysis should be done two ways, 1) with all points and 2) with any outliers which have been removed."**

35. The method for computing 'acceptability' limits of four times the mean absolute differences is unfamiliar; a reference for this rule should be given. Alternatively, Tukey (1977, *Exploratory Data Analysis*. Addison-Wesley, Reading) provides a rule of thumb for identifying data points which are "far out." A far out value is one which is larger than 3 times the interquartile range (difference between the 75th and 25th percentile of the data). Such a technique applied to the difference between duplicates for method X or Y (versus the absolute differences, $Dx_i$ and $Dy_i$ may be easier to use than the one presented here.

- **Please see the source of the suggested method in response to Comment 42. The user may employ Tukey's rule as described, Dixon's tests, or any other defensible outlier detection method in the analysis of the data. The method described herein is a suggestion only.**

Section 4.2

36. We concur that plotting Test versus Comparator as reported and as differences are essential evaluation tools. However, we plot Test Sample 1 results versus Comparator Sample 1 results to provide a more realistic assessment of expected use unless the laboratory routinely reports the mean of pairs.

- **Any additional comparison that is helpful should be used. The working group chose the four plots presented as a reasonable view of the data.**

Section 4.3

37. The third paragraph, "if no linear portion is evident. . . the linearity of each method must be verified independently," needs clarification. Is each reference method tested against a third method? We question if the user will know what he/she is supposed to do.

- **The working group has deleted the following text from the third paragraph of Section 4.3 in order to address the commenter's concern: "…the linearity of each method must be verified**

**independently. If this expanded evaluation of the methods does not reveal the source of the problem…"**

Section 4.4

38. Clarify what data plots to examine and what is meant by 'obvious outlier.'

- **The working group has added the word "visually." It is always somewhat difficult to allow for judgment in a guideline, but in the opinion of the working group, that is what is needed.**

39. The procedure should ensure that the bounds of medical significance were determined up front, before experimentation. Choosing them after the experiment has been performed is like writing the specification from the results you obtained.

- **The commenter is correct; however, the working group feels that this concession will not adversely affect the goals of this study.**

40. Give standardized names for statistical techniques and calculations (e.g., "Use Pearson Correlation Coefficient").

- **The working group agrees with the comment and has checked the document for use of standard (statistical) terms.**

41. Manufacturers of hematology devices strive to have essentially zero difference between analyzers. This is especially true when the analyzers represent only a model difference and not an analytical method difference. Therefore, the "4X" rule will identify a majority of paired samples where the difference is not zero. Therefore, we use the 'parametric, nonparametric' methods.

- **Please see the response to Comment 42. Certainly adjustments to the outlier procedures must be made when the data are clearly categorical, of restricted range, or non-Gaussian.**

42. Please provide the reference for the value "four times."

- **The 4x the average absolute difference rules provided in the document are derived from the control limits used for standard statistical quality control charts for the range of subgroups of two. Four times the mean range translates into a three times the mean range difference of an individual pair delta from the average range, and was taken as approximately the upper 99.9% control limit for an R-chart. A more complex statistical derivation of this simple limit can be derived from tables (i.e., Beyer WH, Ed. *CRC Standard Probability and Statistics Tables and Formulae*. Boca Raton, Florida. CRC Press; 1999. Table VIII.2; 270); wherein, the estimate of the SD of the data itself is 0.8862 times the mean range, and the upper 99.9% percentage point is 4.65*0.8862, or a factor of 4.1. This outlier test was designed to catch only the most extreme of outliers in a dataset. The addition of the similar relative range test follows the same logic but adjusts the ranges by the concentration of the analyte in case of nonconstant variance.**

43. The section also needs a discussion regarding the scaling of XY scatter plots and bias plots. Often a trend or shift can be ignored because the axis range was inappropriate. Also, XY scatter plots should have identical scales (x minimum = y minimum; x maximum = y maximum) in order to make the lines of identity and regression easier to view. NCCLS should consider the additional of medically useful clinical decision limits or confidence limits to these plots.

- **The plots do have identical scales. The working group considered adding medical decision limits or confidence lines; however, it believes addition of such would detract from the visual**

**information in the graph. There is nothing wrong with adding either set of lines. See Section 4.2 for more information on plotting data.**

44. We concur that if the observed differences exceed medical significance then stopping the evaluation is an appropriate action. However, we fail to see the benefit of testing addition, samples (n=40) if the already observed differences demonstrate a lack of medical utility.

- **This flexibility is offered to the laboratory. Repeating a study that gives unexpected results is often warranted as some unknown problem may have occurred which invalidated the results.**

45. The matching of $x_{ij}$ and $y_{ij}$ with respect to j does not seem appropriate unless they are actually paired measurements, i.e., measured from the same aliquot. Simply by chance, $y_{ij}$ could be considered an outlier when compared to $x_{i1}$, but not $x_{i2,}$ for example. In addition, since the goal is to assess the deviation of individual results using the test method versus the comparative method, it is desirable to have the best estimate for the comparison method, the average of $x_{ij}$ and $x_{i2}$. Therefore, it is suggested that in Equation (9), $x_{ij}$ is replaced with $\bar{x}_i$ .

- **The working group agrees. The equations have been revised.**

Section 4.5

46. The formula presented as equation (13) is correct; however, the complexity may discourage the average medical technologist. If the working group's purpose is to define a calculation for subsequent software algorithm development, the formula meets that intent. However, if the working group is using the formula to present the general concept used to calculate Correlation Coefficient, then Geigy's presentation may be more appropriate *(Geigy Scientific Tables; Introduction to Statistics, Statistical Tables, Mathematical Formulae;* Volume 2; Ciba-Geigy; USA; 1982; Page 215; Formula 708). Free text could be employed to define the terms.

$$r = \frac{S_{xy}}{\sqrt{(S_x)(S_y)}}$$

- **The formula and notation used in the document conforms to many formulas found in standard statistics textbooks, and was chosen both to be somewhat familiar to those who have been exposed to the concept, and to allow matching of this document to more complete statistics references for the background, interpretation, and computation of the correlation coefficient. It is the opinion of the working group that the equivalent shorthand notation suggested here would be less easily interpreted, and less readily programmed into spreadsheets or other computation aids.**

47. The second paragraph of Section 4.5 is an interesting use of Correlation Coefficient (r). Perhaps, a reference would be a benefit for those wishing to research this application more fully. The working group has used $r^2$ without a definition. This can be solved by including it in Section 1.3 *Symbols Used in the Text.* Perhaps, it may be time to introduce ($r^2$ x 100) as the Predictive Value of Y from X. This use of $r^2$ could be incorporated into the discussion of predicting Y from X using regression analysis (last paragraph of Section 5.1).

I see no sense in using "r" to judge the adequacy of the clinical range. It would be appropriate to give a range of purportedly useful clinical decision levels, and indicate the "n" to be analyzed (or attempted) at the extremes or medical cutoffs for each parameter. Use H20A as an example. Also, be aware that the method for statistical analysis CAN be driven by external forces, such as CAP,

JCAHO. Government bodies in Europe also turn investigations of new technologies over to groups who use existing methods even when inappropriate (linear regression can be one of these).

- **The dependence of the correlation coefficient on the range of the data is a natural byproduct of its defining formula. The basis for this suggestion is the empirical examination of many types of sets of data, and is familiar to anyone who has looked at the correlation coefficient for limited range analytes such as potassium or sodium. Also, the effects of the range of the data on r can be seen by example where a single point at the extremes of the range can dramatically affect the value of r quite out of proportion to the rest of the data. The real basis of this empirical rule comes from Hald 1952 (*Statistical Theory with Engineering Applications*, Wiley Interscience) where it was shown that both r and the regression slope are affected by a factor which he calls lambda, which is the ratio of the error (imprecision) variance in the X variable to the overall distribution variance of the observed Xs. If the range of the data (in X) is sufficiently wide, then the imprecision in X has little effect on the slope. This same ratio appears in the formula for the correlation coefficient, so the purpose of the test indicated in NCCLS document EP9 is to ensure that the range of the data is sufficiently wide to minimize the bias in the slope due to error in X.**

48. For some hematological parameters, the naturally occurring range of the analyte is limited (Mean Corpuscular Hemoglobin Concentration, basophil, monocyte). For these cases, linear regression and partitioning appear to be of limited value. Therefore, the working group may wish to describe bias as a simple mean difference if the bias plots (Section 4.2) do not show a concentration-dependent difference.

- **The working group agrees and has added a section to this effect.**

49. We do not understand what the magnitude of correlation coefficient has to do with adequacy of range. Since this section deals with the problem of X being subject to measurement error, we are of the opinion that the utility of looking for a very high (and positive) correlation coefficient is to determine whether a Deming regression should be used, rather than ordinary least squares. The usual diagnostic indicating a need for the Deming approach is a difference in the Y-on-X and X-on-Y regression slopes, which would correspond to a low value of the correlation coefficient.

    It is not obvious why the sample correlation can be used to assess the adequacy of the range of X. A reference should be given for this. Note that under certain assumptions, there are simple procedures for adjusting regression estimates when there is error in the Xs. It may be helpful to reference these methods (e.g., Fuller, Wayne A. (1987). *Measurement Error Models*. John Wiley and Sons, Inc., New York.)

- **See the response to Comment 47.**

Section 5.1

50. We concur with the statement "Do *not* use the Orthogonal regression or Deming procedures for calculation of the standard error of estimate because this value will be artificially low and is *not* valid." and perhaps to emphasize the warning, the working group may wish to underline the text.

- **The working group appreciates the commenter's suggestion; however, for consistency, the format of the text will be maintained.**

51. Given the use of these techniques, standard error is judged differently. The techniques themselves may not be appropriate. Same comments are also applicable to the last sentence of Section 8, Item 8.2.

- **The working group cannot respond to this comment without further information.**

52. Deming regression is mentioned in this section, but there is no explanation of what it is or when it should be used.

- **The working group has added definitions for "Deming regression" and "Passing-Bablok" to the Definitions section of the guideline.**

Section 5.2

53. The sentence "Although few methods have constant imprecision throughout the analytical measurement range of that test, visual examination determines whether there are dramatic and significant differences (approximately 3:1 or greater) between the biases at the upper and low end of the range" is an interesting concept. Perhaps the working group should reference it for the benefit of those who may wish to research the application in more detail. For parameters that have an extended expected clinical range (WBC: 0.1 to 100 x 109/L), the 3:1 rule may be overly optimistic; however, the principle may be valid if the ratio expands as a function of range. Thoughts and comments would be appreciated.

    **The 3:1 limit covers many cases of routine clinical analytes. For assays with large ranges, the 3.1 limit will often be exceeded. In these cases, transformations or weighted least squares may have to be employed to adjust for visually obvious heteroscedasticity. The reference here is a technical report written by John Tukey from Princeton University. Heteroscedasticity does not introduce a bias, but the consequence is that confidence intervals become erroneous and the statistical analysis is less efficient than an appropriately weighted modification.**

54. Visual examination cannot be used to quantitatively (3:1) measure significance.

- **Although the commenter is correct, the working group needs to compromise between ease of use and rigor. A visual approximation will meet most needs.**

55. The working group may wish to provide a database-size recommendation. For example, if the ratio is 1:6, the working group may suggest that 75 additional samples be collected and that these should be selected to provide a uniform distribution of values throughout the expected clinical range.

- **Again, the working group needs to ensure that the document is not too complicated. In this and other instances, user judgment with document guidance is suggested rather than some algorithmic approach.**

56. It is unclear if the working group means that standard error of the estimate is unusable for all data or that it is unusable for nonuniform distributions.

- **Throughout this section "uniform" has been replaced with "constant" to address the commenter's concern.**

57. Please provide a reference for "usable standard error of the estimate."

- **The working group is using its judgment here. Logically, one cannot use a nonconstant number (such as an increasing standard error) to make a conclusion across the analytical range.**

58. We concur with the use of transforms when it is appropriate; however, the working group may wish to emphasize that drawing conclusions regarding clinical significance is, at best, difficult when transformed data is used. In addition, transformation may mask clinically significant imprecision. The

working group may wish to delete this paragraph since, without a data example, it may tend to confuse the reader.

- **The working group is trying to be complete in its guidance and believes that inclusion of this section is appropriate.**

59. A reference should be given for "nonuniform scatter."

- **The working group has changed the term from "uniform" scatter to "constant" scatter throughout this section.**

60. Has the question of nonindependence of the data been considered and what would the justification be? The recommended regression procedure involves using two replicates on each subject, as if they were two independent observations.

- **We are assuming that duplicate results from the same samples are independent from the sense of the analyzer but they are not totally independent since they are not separate draws from the patient. In Section 3.5, the sample sequence minimizes effects from biases such as drift that would affect independence. The computations recommend both individual Ys versus average X and average Y versus average X.**

61. In the document, a sample size of 40 subjects (two reps each) is recommended. What is this based on? How does one know whether this is enough or not too much? It is preferable to determine each sample size according to the study hypothesis in each study individually. This guarantees that the sample size is right in the given situation.

- **See the response to Comment 7.**

62. An associate works with indwelling measurement devices, and hence, it is not possible to follow NCCLS EP9-A guideline explicitly in terms of taking a duplicate measurement on each patient sample. In his case, his system can take a new sample from the patient eight minutes later for a "second" reading, but now we are dealing with a situation where any change may be due more to physiologic changes than due to imprecision in the measuring device itself.

    So the working group should address this situation in a way that recognizes that certain technologies are not amenable to taking duplicate measurements of patient samples. (In addition, the working group should consider what would be an appropriate sample for the "comparative" method, which presumably would be a standard laboratory assay. When and from where should this be collected, relative to the "indwelling" measurement?) In the future, there will be more technologies developed that are based on "in-flow" or "*in vivo*" measurements, such that the need to clarify this issue will become of greater importance.

    Another situation where duplicate testing may not be amenable is when the cost to perform a single assay is extremely high, such that a laboratory may not be able to afford performing a test in duplicate.

- **The document has no restrictions as to how samples are collected. It is up to the user to ensure that samples are representative. Duplicate samples are ideal but one can proceed without them.**

63. The axes labels indicate that all individual $Y_{ij}$ vs. $X_{ij}$ points are plotted whereas the instructions (Plotting the data) call for a plot of the "$Y_{ij}$s against the mean $X_{ij.}$" Please clarify.

- **The working group agrees and the error has been corrected.**

64. Calculated results presented on the graph are confusing. The $r^2$ presented corresponds to a value calculated in accordance with the instructions ($Y_{ij}$ vs. $X_i$). I would consider this correct. However, the slope and intercept given do not correspond to such a regression (Y = 1.0035X -0.6283, identical to the equation presented). Nor do the values given correspond to a regression analysis of the individual data points ($Y_{ij}$ vs. $X_{ij}$). Please clarify.

- **This is an error and has been corrected. The regression equation is calculated from individual y-observations and mean x-observations.**

65. The objective of NCCLS document EP9-A is to estimate the predicted bias ($B_c$) and its confidence interval at a predefined medical decision level ($X_c$). Therefore, the sentence "If the 95% confidence interval ....=149.9" should read "The 95% confidence interval of $B_c$ at $X_c$ = 150:

$$\hat{B}_c = -0.429 + (1.002 - 1) * 150 = 0.129$$

1. (computation based on formula $23, \hat{B}_c = a + (b-1)^* X_c$)
2. Lower 95%. . . ."

Attached is a copy of "Table 5-2. Confidence Intervals and prediction intervals for straight-line regression analysis" (*Applied Regression Analysis and Other Multivariate Methods*, Kleinbaum, Kupper & Muller, PWS-KENT Publishing Company), the formula 24 cannot be directly derived from the table provided. Could you provide the derivation of formula 24? As software validation is becoming an important issue in the industry, all the calculations from our testing protocols have to be verified. It will be very helpful if you can provide the derivation of formula 24.

- **The formula provided in Equation 24 is centered around the BIAS value, not the predicted value on the regression line like in standard references. Since we are primarily interested in this bias value (expected difference) at $X_c$, the confidence interval has been adjusted to center around this value.**

66. The outlier tests for within-method duplicates seem to have very large limits of acceptability at four times the mean absolute differences and four times the relative differences. I have observed significant differences between duplicates that passed both outlier tests.

- **The commenter shared the data set with the working group. The data set had three outliers in the set of 40 pairs. When the percentage of outliers is so large, no statistical test can distinguish the "outliers" from the rest of the data. The "outliers" may be representative of the precision performance of the method. If the precision performance of either method causes large disagreements between the Test and Comparative methods, these should be detected in Section 4.4, "Visual Check for Between-Method Outliers."**

Section 6.1

67. The formulae present a rather complex mathematical picture that may distance the audience (readers). Perhaps, the working group may wish to try a simpler approach that still retains clinical relevance.[†] For example, the 95% CI of future *average value* of y can be roughly determined using the Standard Error of the Mean as:

[†] Weisbrat IM. *Statistics for the Clinical Laboratory*. J.B Lippincott Company; USA 1985:35; Syx.

$$SEM_{95\%_{Confidence}} = \frac{\pm 2S_{yx}}{\sqrt{n_{data_{pairs}}}}$$

- **The working group agrees that many of the formulas in the document are complex and can be confusing and daunting. However, it was felt as a general principle that the correct and defensible definitions and formulas should be used throughout, rather than simpler, rough approximations such as this, which could just as easily be criticized from the other direction.**

Sections 6.2

68. Please include a data example that will further clarify the partition method.

- **The working group believes that this procedure is sufficiently simple and that the additional volume of an example is not required.**

Section 7

69. We concur that a new method may be less precise but have other benefits, e.g., less interfering substances. In these cases, the working group may wish to recommend that the patient assay result should also contain a statement of imprecision at 95% confidence (see below):

    Glucose: 87 (± 4) mg/dL

    This concept is fairly novel since laboratory test results generally do not include an imprecision statement. However, when one considers the routine use of therapy protocols, the incorporation of the uncertainty of an assay may make medical decision points less razor sharp.

- **The suggestion for uncertainty estimates is beyond the scope of this document.**

70. This section requires additional discussion. It is possible to find a new technology that is biased from the existing Comparator method that can be BETTER. Precision, reference range, specificity and the sensitivity at the medical decision levels must be taken into account.

- **The working group agrees in principle but wishes to limit the scope of the document.**

71. We concur with the statement "If the manufacturer's claim for average bias is included in the 95% confidence interval, then it can be concluded that the candidate method has provided equivalent results" because much time can be lost explaining that the user-verification result is statistically a subset of the manufacturer's validation study.

- **The working group thanks the commenter for the comment.**

Section 8.3

72. The suggested labeling assumes that regression analysis is appropriate for data analysis. In some cases, this is true; however, for other analysis such as the leukocyte differential parameters, other analyses may be more appropriate (Binomial). Therefore, the working group may wish to recognize alternative analysis methods in the requirements for Claims.

- **The scope of this document is limited to tests that have continuous quantitative results.**

73. The slope and intercept of the fitted linear regression line may be an inappropriate technique for the technology.

- **See the response to Comment 72.**

## Summary of Delegate Comments and Committee Responses

EP09-A2: *Method Comparison and Bias Estimation Using Patient Samples; Approved Guideline—Second Edition*

General

1. An important feature of the assay validation problem is that both *x* and *y* are measurements of an analyte, *z*, that is unobserved and changes from subject to subject. The purpose of method comparison is to assess the deviation of individual results using the test method versus the reference method. The task for method comparison is to confirm; hence, linear regression may not be a good method for this evaluation.

- **The committee does not agree with the commenter's general model. The standard regression model, which applies to diagnostics assays when the comparison assay is a reference method, assumes that "X" is known and without error. There is no need to invoke a variable "z."**

Section 1.4

2. Use of analytical measurement range (AMR) is excellent and makes the document conform to definitions used by CAP in its accreditation program. I suggest the term "reportable range" be deleted throughout the document and replaced with the term "AMR." Replacing the single term "reportable range" with two terms "AMR" and "clinically reportable range" increases the clarity of what each term refers to.

- **The committee agrees with the commenter. The suggested change has been incorporated.**

Section 3.1.1

3. Replace "constituent" with "measurand" to be ISO compliant.

- **The committee agrees with the commenter. The suggested change has been incorporated.**

Section 3.2

4. Second bullet; delete "To"; fourth bullet replace "national" with "metrological" to improve global acceptance.

- **The committee believes neither "to" nor "national" is needed. The terms have been deleted.**

Section 3.6

5. "Analysis ... within two hours" is arbitrary and while applicable to many analytes will not be satisfactory for some such as blood gasses, whole blood glucose, ammonia, etc. A suggestion is: "For a given sample, analysis by the comparative and test methods should occur within a time span consistent with the analyte stability. For all analytes, the time span should not exceed two hours for analysis by each method."

- **The committee agrees with the commenter. The suggested change has been incorporated.**

Section 3.8

6. Add in the first line “and/or manufacturer's” after “Follow the laboratory’s”.

- **The committee agrees with the commenter. The suggested change has been incorporated.**

Section 4

7. Correct the spelling of “internal” in the box at the bottom of the page in Figure 1.

- **This editorial correction has been made.**

8. Figure 2 and various places in the text use the term bias to mean the difference between values measured by test and comparison methods, e.g., "bias plot." The definition given for the term bias is "difference between the expectation of the test results and a true value." A true value (in ISO terminology) is only available if the sample has been assayed in replicate by a reference method. I suggest the term "bias" be replaced by the term "difference" unless the true value is actually known.

- **The committee agrees with the comment and has addressed it by adding the terms “accuracy,” “bias,” and “trueness” to Section 1.4 on Definitions and the following paragraph to Section 3.2: "If the comparison method is a reference method, then the difference between the two methods measures the trueness of the new method, measured as bias. If the comparison method is not a reference method, then the trueness of the new method cannot be determined. In this case, one should refer to the difference simply as a difference, and not bias. Since the preferred approach is to use a reference method as the comparison method, the term 'bias' is used in this document."**

Section 4.2

9. Second and third paragraph the word "or" is used when "for" is meant in the phrase "... or each assay."

- **The committee agrees with the commenter. The suggested change has been incorporated.**

Section 4.5

10. Following equation (15) end of first sentence "(or equivalently, if $r^2 = 0.95$)" should be "$r^2 \geq 0.95$."

- **The committee agrees with the commenter. The suggested change has been incorporated.**

11. There are analytes for which extending the range of the data is not possible, e.g. Na, Cl, Ca, etc., and partitioned biases procedure is not suitable. I suggest adding the following sentence, "In cases where physiologically the analyte spans a relatively small range (e.g., sodium, chloride, calcium) extending the range or using a partitioning approach may not be possible. In these cases, additional data may improve the analysis or the scatter in data may need to be acknowledged as a limitation in data interpretation."

- **The committee does not agree with the commenter. The partitioned bias method was designed to cover the situation for analytes that have a limited range.**

Section 8.3

12. In the "test" column of Tables 1a and 1b, the units given (e.g., mg/dl) are the same in both tables, although the numbers in the tables are different. Are these meant to be gravimetric versus molar concentrations?

- **Gravimetric (as opposed to volumetric) is a means to prepare concentrations (such as molar concentrations); therefore, the information presented in the "test" column of Tables 1a and 1b has been maintained.**

13. Replace "accuracy" with "trueness," or with "accuracy or trueness."

- **The committee agrees with the commenter. The sentence has been modified to read: "Unless the comparative method is an established reference method, the terms "accuracy" and "trueness" should not be used.**

14. Appendix B1, B3, and B4 should be redrawn to put the x-axis numbers at the bottom of the graph along the x-axis.

- **The committee agrees with the commenter's concerns relative to the figures in B3 and B4. The suggested changes have been incorporated. The figure in B1 has been maintained.**

## Related NCCLS Publications*

**EP5-A** **Evaluation of Precision Performance of Clinical Chemistry Devices; Approved Guideline (1999).** This document provides guidance for designing an experiment to evaluate the precision performance of clinical chemistry devices; recommendations on comparing the resulting precision estimates with manufacturer's precision performance claims and determining when such comparisons are valid, as well as manufacturer's guidelines for establishing claims.

**EP6-P2** **Evaluation of the Linearity of Quantitative Analytical Methods; Proposed Guideline—Second Edition (2001)**. This document provides guidelines for characterization the linearity of a method during a method evaluation; for checking linearity as part of routine quality assurance; and for determining and stating a manufacturer's claim for linear range.

**EP7-P** **Interference Testing in Clinical Chemistry; Proposed Guideline (1986).** This document provides background information and procedures for characterizing the effects of interfering substances on test results.

**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 (NRSCL).

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

**NOTES**

**NOTES**

940 West Valley Road ▾ Suite 1400 ▾ Wayne, PA 19087 ▾ USA ▾ PHONE 610.688.0100
FAX 610.688.0700 ▾ E-MAIL: customerservice@clsi.org ▾ WEBSITE: www.clsi.org ▾ ISBN 1-56238-731-6