CLINICAL MICROBIOLOGY REVIEWS, July 2010, p. 550–576 Vol. 23, No. 3
0893-8512/10/$12.00 doi:10.1128/CMR.00074-09

Validation of Laboratory-Developed Molecular Assays for Infectious Diseases

Eileen M. Burd*

Emory University Hospital, Atlanta, Georgia

* Mailing address: Emory University Hospital, 1364 Clifton Rd., N.E., Atlanta, GA 30322. Phone: (404) 712-7297. Fax: (404) 712-4632. E-mail: eburd@emory.edu.

INTRODUCTION

Molecular tests to detect infectious agents are now widely used in many clinical laboratories. The technological advantages of molecular tests make them very powerful diagnostic tools, and they have become particularly valuable for the detection of infectious agents that cannot be grown or are difficult to grow in culture. The field of molecular testing for infectious diseases has expanded greatly and now includes qualitative assays that detect a single target, quantitative assays that are used as a part of monitoring the response to therapy for some viral infections, and multiplexed assays that detect two or more analytes in the same specimen. Most molecular tests used in clinical laboratories are commercially produced, FDA-approved tests. Sometimes, however, tests are developed, evaluated, and validated within one particular laboratory. These "laboratory-developed tests" are used only by the developing laboratory and are not distributed or sold to any other laboratories. Laboratory-developed tests are used in many sections of the laboratory, including chemistry, coagulation, microbiology, hematology, and molecular diagnostics. Molecular assays are often created by a clinical laboratory because a commercial test for the analyte(s) of interest is not currently available. Tests may not be commercially available because the analyte is rare, and the market for such a product would be too small to be profitable. Clinical Laboratory Improvement Amendments (CLIA) regulations recognize that clinical laboratories can run three types of "test systems": (i) test systems that are cleared or approved by FDA and run by the laboratory without modification, (ii) test systems that are cleared or approved by FDA and run after modification by the laboratory, and (iii) test systems that are not subject to FDA clearance or approval (27). In spite of the widespread use of molecular tests, there is still confusion surrounding the requirements that need to be met when bringing a molecular test, whether FDA approved/cleared or laboratory developed, into a clinical laboratory. Part of the confusion has arisen because, even with attempts to align terminology, the terms "validation" and "verification" have been used interchangeably to describe the same process.

All laboratories in the United States that perform clinical testing on humans, excluding clinical trials and basic science research, are regulated by the Clinical Laboratory Improvement Amendments (CLIA) of 1988. The CLIA federal regulatory standards (Public Law 100-578) were passed by the 100th Congress in 1988, published in the *Federal Register* in 1992, and extensively revised in January 2003 (71, 72). The regulatory standards are codified in the Code of Federal Regulations (CFR). The main objective of the CLIA regulations is to ensure the accuracy, reliability, and appropriateness of clinical test results, regardless of where the test is performed. As such, CLIA sets the minimum standards that must be met in validating performance of clinical tests. The Centers for Medicare and Medicaid Services (CMS) has the primary responsibility for operation of the CLIA program. Laboratories are recognized as meeting the requirements of CLIA if they are accredited by professional organizations such as the Joint Commission (JC), the College of American Pathologists (CAP), COLA (formerly the Commission on Office Laboratory Accreditation), or another agency officially approved by CMS. Some states (e.g., New York and Washington) also have state health laboratory organizations that are approved by the government and impose specific requirements that are comparable to or more stringent than CLIA regulations. These states are considered to be CLIA exempt, and the state requirements for evaluating a test or test system must be met. Molecular tests are considered to be nonwaived (formerly called moderate and high complexity) and are subject to all CLIA requirements for nonwaived tests. All method validation requirements must be met before results can be used for decisions regarding patient care.

CLIA defines differences between implementation of FDA-approved tests and implementation of laboratory-developed tests (Table 1). Prior to the 2003 final rule, laboratories could accept the performance characteristics provided by the manufacturer of FDA-approved nonwaived tests instead of performing method validation studies themselves. The 2003 final rule now requires that laboratories do studies for FDA-approved nonwaived tests to verify that the performance specifications established by the manufacturer can be reproduced by the testing laboratory for the population of patients that the laboratory serves. The performance characteristics that must be verified include accuracy, precision, reportable range, and reference interval (28). Laboratories are not required to verify analytical sensitivity or analytical specificity for FDA-approved

TABLE 1. Required performance characteristics with suggested studies needed before implementation of FDA-approved/cleared tests and laboratory-developed tests[a]

Performance characteristic (reference[s]) and suggested study	Requirement(s) for:	
	FDA-approved/cleared test	Laboratory-developed test
Reportable range (8), linearity study (for quantitative assays)	5-7 concentrations across stated linear range, 2 replicates at each concn	7-9 concentrations across anticipated measuring range (or 20-30% beyond to ascertain widest possible range); 2-3 replicates at each concn; polynomial regression analysis
Analytical sensitivity (14, 28, 33), limit-of-detection study	Not required by CLIA, but CAP requires LOD verification for quantitative assays; use 20 data points collected over 5 days	60 data points (e.g., 12 replicates from 5 samples in the range of the expected detection limit); conduct the study over 5 days; probit regression analysis (or SD with confidence limits if LOB studies are used)
Precision (7, 13, 15, 40), replication experiment	For qualitative test, test 1 control/day for 20 days or duplicate controls for 10 days; for quantitative test, test 2 samples at each of 2 concentrations (4 samples) plus one control over 20 days or test 2 concentrations in triplicate over 5 days	For qualitative test, minimum of 3 concentrations (LOD, 20% above LOD, 20% below LOD) and obtain 40 data points; for quantitative test, minimum of 3 concentrations (high, low, LOD) and test in duplicate 1-2 times/day over 20 days; calculate SD and/or CV within run, between run, day to day, total variation
Analytical specificity (28), interference study	Not required by CLIA	No minimum no. of samples recommended; test sample-related interfering substances (hemolysis, lipemia, icterus, etc.) and genetically similar organisms or organisms found in same sample sites with same clinical presentation; spike with low concentration of analyte; paired-difference (*t* test) statistics
Accuracy (trueness) (13), comparison-of-methods study	20 patient specimens within the measuring interval or reference materials at 2 concentrations (low and high) in duplicate over 2-5 runs	Test in duplicate by both the comparative and test procedures over at least 5 operating days; typically 40 or more specimens; *xy* scatter plot with regression statistics; Bland-Altman difference plot with determination of bias; % agreement with kappa statistics
Reference interval (6)	The reference interval stated by the manufacturer may be "transferred" if the stated reference interval is applicable to the population served by the clinical laboratory; if exptl verification is desired, test 20 specimens representative of the population; if the population is different, establish the reference interval by testing 60 (minimum, 40) specimens	If a nucleic acid target is always absent in a healthy individual and the tests is a qualitative test, the reference range is typically "negative" or "not detected" and reference interval studies do not need to be performed; for quantitative assays, the reference interval will be reported as below the LOD or LLOQ; for some analytes, the reference interval may be a clinical decision limit; if the intended use of the test is limited to patients known to be positive for the analyte being assayed, a reference interval may not be applicable

[a] All validation and verification studies must use samples prepared in the appropriate matrix. Each sample type that will be tested in the clinical laboratory must also be evaluated, as well as each genotype and each analyte in multiplex assays.

tests but should verify limit of detection (LOD) for quantitative assays (28, 33). For laboratory-developed tests, however, more extensive studies are required, and the laboratory must establish the performance specifications of the test at the time of test development. The performance characteristics that must be established include accuracy, precision, reportable range, reference interval, analytical sensitivity, and analytical specificity.

Validation and verification studies are not required for tests used by the laboratory before 24 April 2003 but must be done for tests introduced after that date (28). Documentation of all validation and verification experiments must be kept by the laboratory for as long as the test is in use but for no less than 2 years (28). Calibration and control procedures must also be determined based on performance characteristics whether the test is cleared/approved by the FDA or laboratory developed. While CLIA requires analytic validation of an assay described here, CLIA does not require clinical validation of an assay prior to its use in a clinical laboratory. Clinical accuracy is not a property of the test per se but is a property of the clinical application of the test. Establishing clinical accuracy requires clinical trials that may go beyond the purview of an individual

laboratory (18, 34). Studies that document clinical relevance are frequently provided in peer-reviewed literature. CLIA requires laboratories to have a director who is responsible for ensuring, using studies performed by the laboratory or reported in published or other reliable sources, the clinical utility of the tests performed in his or her laboratory (32, 49).

The following discussion will be limited to the processes involved in validation of laboratory-developed tests and will also address the ongoing postvalidation calibration and quality control procedures required to ensure that the expected performance is maintained throughout the life of the test (29, 30). While CLIA lists the performance specifications that must be established, CLIA does not specify the scientific methodology or data analysis tools to be used. Guidelines to assist in establishing performance specifications have been published by the Clinical and Laboratory Standards Institute (CLSI) and International Organization for Standardization (ISO) in several documents. CLSI consensus documents are developed by subcommittees or working groups with representatives from clinical laboratories, manufacturers of products for medical testing, and regulatory and scientific government agencies. These guidelines are reviewed and approved by an official vote of its members. There are guidelines that address various types of molecular diagnostic assays as well as guidelines that address evaluation of specific assay performance characteristics. Some of the CLSI protocols are intended for test developers, and others are intended for laboratory users of FDA-approved tests. Developers of test methods will generally follow protocols intended for manufacturers, although performance characterization studies will not usually need to include a between-laboratory component unless the test will be performed at multiple sites. Use of CLSI protocols is not mandatory, but they are frequently referred to by accrediting agencies and are regarded as good laboratory practices. ISO is a nongovernmental organization that is similarly structured, with technical committees that draft standards which are then submitted to representatives from the 162 member countries for review and approval by vote.

CLIA regulations stipulate that it is the responsibility of clinical laboratory directors to establish performance characteristics for laboratory-developed tests used in their laboratories. Laboratories face many challenges in trying to accomplish this. Laboratories must determine the type of experiments that are required, include an acceptable number and type of specimens, and choose the statistical methods to evaluate the data. Laboratories may follow relevant guidelines from CLSI, ISO, or other sources. Molecular test methods have advanced rapidly and are continuing to change, making existing guidelines often difficult to apply. Clinical laboratories are subject to inspection from a variety of accrediting agencies (e.g., CMS, COLA, CAP, and JC) that also have standards that must be met. Accrediting agency standards must include the minimum standards set by CLIA, but accrediting agencies may have additional, more stringent requirements. A single set of comprehensive guidelines that would help laboratories manage validation studies and that is acceptable to all accrediting organizations is not available. Laboratories today are also under great pressure to control costs and must carry out method validation studies by performing the minimum necessary to satisfy regulatory requirements and ensure robust performance of an assay. Well-designed experiments are essential to accomplish this. One of the major challenges in validating laboratory-developed infectious disease assays is the absence of standards for many analytes. An additional element that must be considered is that, although not technically research under the U.S. Department of Health and Human Services (DHHS) definition, the use of patient specimens for validation studies may require prior approval by an Institutional Review Board (IRB). DHHS defines research as "a systematic investigation, including research development, testing and evaluation, designed to develop or contribute to generalizable knowledge." If the activity is not considered research or if patient specimens are used in such a way that subjects cannot be identified either directly or through identifiers linking the specimens to the subject, the activity may be exempt from IRB approval. An IRB, not the investigator, must determine if a project is exempt. If the activity is not considered research under the DHHS definition, it may still meet the FDA definition of research. Some activities involving FDA-regulated products, including *in vitro* diagnostic tests, will not be exempt, even if specimens are deidentified and the activity seems to fit under the DHHS definition of exempt research. FDA regulations generally require IRB review and approval of activities using FDA-regulated products (19). Current definitions and the regulatory status of laboratory-developed tests have created uncertainty regarding the need for IRB approval for validation studies. Many IRBs have a specific human subject protection program that governs the use of specimens. Since validation studies may or may not meet the DHHS or FDA definitions of human subjects or research, laboratories should contact their IRB to obtain a written determination.

Definitions

FDA-approved assays. Clinical laboratory tests are *in vitro* diagnostic devices (IVDs) that are defined in the *Federal Food, Drug, and Cosmetic Act* as an "instrument, apparatus, implement, machine, contrivance, implant, *in vitro* reagent, or other similar or related article … intended for use in the diagnosis of disease or other conditions, or in the cure, treatment or prevention of disease, in man" (41). Facilities that manufacture, repackage, relabel, and/or import IVDs sold in the United States are regulated by the FDA's Center for Devices and Radiological Health (CDRH) under authority granted by laws in the Code of Federal Regulations passed by Congress. These facilities undergo periodic inspections by the FDA to ensure that they are in compliance with quality system (QS)/good manufacturing practices (GMP) requirements (25).

IVDs that are commercially distributed for diagnostic use in the United States require prior approval or clearance by the FDA. A manufacturer can currently place an IVD into the market in two main ways. One way is by premarket notification 510(k), in which a manufacturer provides documentation and data demonstrating that the new device is substantially equivalent to an existing marketed device in terms of both safety and effectiveness under the conditions of intended use. The FDA review of a 510(k) is entirely a scientific evaluation of data. If the FDA assessment indicates that the new device is substantially equivalent to a legally marketed device, the device is cleared and the manufacturer is free to market it in the United

States. If there is no similar preexisting marketed device, the manufacturer must submit a premarket approval (PMA) application rather than a 510(k). A PMA may be required if the test is for a novel agent, if it is a new method for which clinical relevance and clinical use must be established, or if the analyte poses a health threat (such as *Mycobacterium tuberculosis*) and a false-positive or false-negative result would be of significant risk to the patient or general public. The manufacturer must provide data that demonstrate a reasonable assurance of safety and effectiveness of the device. FDA review of a PMA application includes an in-depth scientific evaluation of the data and a comprehensive good manufacturing practice inspection at the manufacturing facility. Before approval, the PMA may also be reviewed by an FDA advisory panel of outside experts who provide recommendations. If the completed assessment is favorable, the IVD is approved and can be marketed in the United States. An assay that is submitted via the 510K route is listed as "FDA cleared," since clinical relevance has already been established and predicate devices to which results can be compared are available. An assay that is submitted via the PMA route is listed as "FDA approved" for specific clinical applications such as diagnosis, monitoring, etc.

Assays that have been approved or cleared by the FDA are labeled "for in vitro diagnostic use." Labeling regulations also require these assays to have a package insert that indicates the intended use of the test; instructions for specimen collection, transport, and preparation for analysis; storage recommendations; a summary and explanation of the test; step-by-step instructions for performing the test; specific performance characteristics such as accuracy, precision, specificity, and sensitivity; cutoff criteria and interpretation of results; limitations of the assay; quality control recommendations; bibliography; and the name and place of business of the manufacturer, packer, or distributor (21).

Laboratories that use FDA-approved or -cleared assays are required to verify the performance of these assays in their laboratories before reporting patient results.

Laboratory-developed tests. Laboratory-developed tests are considered to be tests that are used for patient management but have been developed within CLIA-certified laboratories for use by those laboratories. Laboratory-developed tests may be (i) FDA-cleared or -approved tests that have been modified by the laboratory, (ii) tests not subject to FDA clearance or approval, or (iii) test systems in which performance specifications are not provided by the manufacturer (28).

For FDA-cleared or -approved tests that have been modified by the laboratory, CLIA does define the term "modified," but modifications are generally considered to include changes in test components (extraction, amplification, and/or detection), procedural parameters, assay cutoff values, specimen types or collection devices, etc. The CAP allows results to be reported with a disclaimer for alternative specimen types while validation studies are in process or if the specimen type is rare and validation studies cannot be done due to insufficient numbers (33). The disclaimer should state that the specimen type has not been validated.

Tests not subject to FDA clearance or approval include standardized textbook procedures or tests developed in the laboratory that performs the assay and reports the results. FDA has historically taken the position that it has the authority to regulate laboratory-developed tests but has exercised "enforcement discretion" and has chosen not to, in part because CLIA and comparable state laws regulate the practice of clinical laboratory testing.

Laboratory-developed tests are accepted as being scientifically valid and are relied on routinely in the delivery of health care in the United States. Laboratory-developed tests are extensively regulated by CMS under CLIA. Clinical laboratories must determine performance specifications for all laboratory-developed tests as required by CLIA and are responsible for both the quality and interpretation of results generated from those tests. Studies to determine performance specifications for laboratory-developed tests are not transferable to other clinical laboratories, and each laboratory must conduct its own studies. Although laboratory-developed tests are not regulated by FDA, some components, such as reagents (general-purpose reagents and/or analyte-specific reagents [ASRs]), controls, or equipment, used in these tests may be purchased from third-party biological or chemical suppliers and may be FDA approved.

For several years, FDA officials have indicated that, because of the increase in the number and complexity of laboratory-developed tests, they are reconsidering the current enforcement discretion exemption from FDA oversight. Numerous advisory committees have made recommendations for creation of a system of oversight. Some recommendations have suggested that FDA should regulate laboratories as manufacturers of medical devices and that all laboratory-developed tests should be reviewed by the FDA in some manner before being offered clinically. Others suggest that manufacturers and laboratories are different entities and that FDA regulation of all laboratory-developed tests would not allow patients to have access to innovative clinical tests. Any regulatory changes could affect the process of validation.

CAP has proposed a three-tier model for regulatory oversight based on potential risk to patients and the extent to which test results influence diagnosis or treatment decisions. Under the CAP plan, laboratory-developed tests would be classified as low, moderate, or high risk. High-risk tests would require FDA review before being placed into clinical use. Moderate-risk tests would be reviewed by the accrediting agency used by the laboratory. Low-risk tests would be validated in the clinical laboratory, and the accrediting agency would review validation procedures and compliance with accreditation standards during regular inspections. Tests in the proposed low-risk classification include those that may affect diagnosis or treatment but are not used independently or directly. Also considered low risk would be tests for rare diseases as well as FDA-approved/cleared tests that are modified by the laboratory. The distinction between the proposed moderate- and high-risk classifications largely concerns the transparency of the methodology by which the test result is obtained and interpreted. Tests that use a proprietary algorithm or calculation that is not accessible to the end user would be considered high risk. The high-risk category would include *in vitro* diagnostic multivariate index assays (IVDMIAs), which combine findings from multiple individual analyses into a single, patient-specific test result using an interpretation function that cannot be independently derived or verified by the end user. The CAP plan also calls for stronger CLIA accreditation standards for laborato-

ries using low- and moderate-risk laboratory-developed tests. A requirement for clinical validation of laboratory-developed tests has been included in the CAP proposal to ensure that tests are accurately correlated to a clinical condition. The CAP Laboratory Accreditation Program currently has the requirement that laboratories demonstrate clinical validity as an item in the Molecular Pathology Checklist (34) but not in the Microbiology Checklist (33).

(i) ASR. The "ASR rule" was published in the Code of Federal Regulations in November 1997 to clarify the role of the FDA in the regulation of laboratory-developed tests and to ensure that the components of those tests were made consistently over time. The "ASR rule" has three major parts: (i) analyte-specific reagents (ASRs) are defined and classified in a rule codified in 21 CFR 864.4020; (ii) restrictions on the sale, distribution, and use of ASRs are imposed in 21 CFR 809.30; and (iii) requirements for ASR labeling are established in 21 CFR 809.10(e) (21, 23, 26). These statutory requirements were enforced by the FDA on 15 September 2008.

ASRs are not diagnostic tests. They are key components of diagnostic tests and are defined as "antibodies, both polyclonal and monoclonal, specific receptor proteins, ligands, nucleic acid sequences, and similar reagents which, through specific binding or chemical reaction with substances in a sample, are intended for use in a diagnostic application for identification and quantification of an individual chemical substance or ligand in biological specimens" (26). ASRs can be manufactured anywhere in the world. ASRs must be manufactured in compliance with current GMPs to ensure that they are manufactured under controlled conditions that ensure that the devices meet consistent specifications across lots and over time (22). ASRs are subject to regulation as medical devices when they are purchased by clinical laboratories for use in laboratory-developed tests or certain IVD tests. The FDA classifies medical devices, including ASRs, into class I, II, or III according to the level of risk associated with the device and the regulatory control necessary to provide reasonable assurance of safety and effectiveness. Most ASRs are classified as class I and are exempt from FDA premarket notification requirements in part 807, subpart E, of 21 CFR (20). ASRs are designated class II and require premarket notification (510k) when they are intended for use in blood banking (e.g., reagents used in cytomegalovirus [CMV] and syphilis serologic tests). Class III ASRs include reagents used in donor screening tests and certain high-risk infectious disease tests (e.g., those for human immunodeficiency virus [HIV] or tuberculosis). FDA regulations require premarket approval (PMA) for all class III ASRs before they can be sold in the United States.

ASRs may be sold only to (i) diagnostic device manufacturers; (ii) clinical laboratories that are CLIA certified to perform high-complexity testing under 42 CFR part 493 or clinical laboratories regulated under the Veteran's Health Administration Directive 1106; or (iii) organizations such as forensic, academic, research and other nonclinical laboratories that use the reagents to make tests for purposes other than diagnostic information for patients and practitioners (23).

ASR manufacturers must provide on the label the proprietary name, common name, quantity, or concentration of the reagent; the source and measure of its activity; and the name and place of the manufacturer (21). Class 1 exempt ASRs must be labeled "Analytic Specific Reagent. Analytical and performance characteristics are not established." Class II or III ASRs must be labeled "Analytic Specific Reagent. Except as a component of the approved/cleared test [name of approved/cleared test], analytical and performance characteristics are not established" (21). ASR manufacturers are not permitted to include any statements regarding the clinical or analytical performance of the ASR or information on methods or techniques. Manufacturers may not assist with optimization of tests or provide technical support. The responsibility for test development is clearly assigned to the laboratory. ASRs may not be promoted for use on designated instruments or in specific tests. Laboratories may combine individual ASRs and other components in the development of their own tests. ASRs cannot be sold as kits. ASRs cannot be sold with validation information or medical or performance claims.

Federal regulations require laboratories to append the following disclaimer to the laboratory-developed test result report: "This test was developed and its performance characteristics determined by [laboratory name]. It has not been cleared or approved by the U.S. Food and Drug Administration" (21, 26). This statement is not required if all of the ASRs used in an assay are created in the laboratory, rather than purchased. The CAP notes that it may be problematic to require clinical laboratories to warn physicians that tests were developed without FDA review and that it would be more accurate to acknowledge that the FDA does not require clearance or approval of laboratory-developed tests (34). CAP recommends adding language to the required disclaimer such as "The FDA has determined that such clearance or approval is not necessary. This test is used for clinical purposes. It should not be regarded as investigational or for research. This laboratory is certified under the Clinical Laboratory Improvement Amendments of 1988 (CLIA-88) as qualified to perform high complexity clinical laboratory testing" (34).

RUO. Research-use-only (RUO) products are intended solely for research purposes, not for diagnostic purposes. RUO products are often discussed as if they are medical devices, but since the intended use is research only, they do not fit the definition of a device and are essentially unregulated. RUO products are addressed very briefly by FDA regulations. FDA regulations provide only a definition and labeling requirements. RUO products are defined by FDA regulations as "in the laboratory research phase of development and not represented as an effective *in vitro* diagnostic product" (21). An RUO product cannot be represented as an effective *in vitro* diagnostic product, and manufacturers cannot make performance claims or give reference values. RUO products must be labeled "For research use only. Not for use in diagnostic procedures" (21). Labeling a product RUO allows it to be available to researchers who can then evaluate whether the product may be potentially useful for some specific diagnostic purpose. RUO products may be used in preclinical or nonclinical research settings, and they may be used with either clinical or nonclinical materials; however, the research cannot have intended clinical use. The research is limited to either basic science research unrelated to product development or the investigation of potential clinical utility of a product in the initial phase of development. RUO products cannot be used for investigational purposes (clin-

ical studies) or for research to establish safety and effectiveness of the product. Results cannot be reported to the patient's physician or medical record and cannot be used to assess the patient's condition or for any diagnostic, prognostic, monitoring, or therapeutic purposes. The sale of RUO products is not restricted to high-complexity, CLIA-certified laboratories. Manufacturers and distributors of RUO products are encouraged to have a certification program that documents the researcher's agreement that the product will not be used for clinical purposes. RUO products are not registered with the FDA, and the FDA does not expect these products to be manufactured in compliance with GMP because they cannot be used for clinical purposes (21). Laboratories should be aware that it is illegal to bill CMS for RUO tests (66). This may or may not apply to nonfederal payers, depending on their specific coverage policies for RUO tests.

CAP allows use of RUO reagents as components of laboratory-developed tests when FDA-approved/cleared or ASR products are not available, and it states that "Antibodies, nucleic acid sequences, etc., labeled 'Research Use Only' (RUO) purchased from commercial sources may be used in home brew tests if the laboratory has made reasonable effort to search for FDA-approved/cleared kits, and ASR class reagents. The results of that failed search should be documented by the laboratory director" (34). In addition, the laboratory director has the option not to select an FDA-cleared/approved test over a laboratory-developed test if the FDA-approved/cleared test is found in documented studies to be inferior to the laboratory-developed test.

IUO. FDA considers investigational-use only (IUO) products to be in the phase of development that requires clinical investigation before a manufacturer can submit an application for product clearance. It is important to note that the federal regulations governing IUO medical devices, including *in vitro* diagnostic products, differ from those governing IUO drugs and biologics (24). IUO products may be distributed only for use in well-controlled clinical trials to establish performance characteristics. Testing may include, but is not limited to, gathering data required to support a 510(k) submission or PMA application to FDA, establishing safety and effectiveness of a product, establishing clinical performance characteristics and expected values in the intended patient population, comparing the usefulness of the product to that of other products or procedures currently recognized as useful, or clinical evaluation of certain modifications or new intended uses of legally marketed devices (21).

If the purpose of the clinical investigation is to establish the safety and effectiveness of a product and the product is of significant risk, the product is regulated under part 812, investigational device exemption (IDE), of the Code of Federal Regulations (24). An approved IDE must be issued by the FDA in order for a device to be shipped lawfully for purposes of conducting the investigation without complying with other FDA requirements that would apply to devices that are in full commercial distribution. All clinical evaluations of IUO devices, unless exempt, must have an approved IDE before the study is initiated. An IDE requires approval by an IRB. An IDE also requires informed consent from all patients, labeling as "CAUTION-Investigational Device. Limited by Federal (or U.S.) law to investigational use," monitoring of the study, and documentation of required records and reports.

IUO devices may be exempt from the IDE requirements of part 812 under certain conditions. For a device to be exempt, testing must be noninvasive, cannot require invasive sampling presenting significant risk, cannot introduce energy into a subject, cannot be used for determining safety and effectiveness, and cannot be used for human clinical diagnosis unless the diagnosis is being confirmed by another, medically established diagnostic product or procedure (24). Simple venipuncture to obtain blood or the use of residual specimens (body fluids or tissues left over from specimens taken for noninvestigational purposes) is considered noninvasive (24). IVDs that are exempt from IDE requirements must be labeled, "For Investigational Use Only. The performance characteristics of this product have not been established" (21). Studies exempt from IDE requirements may or may not require IRB review and approval and may or may not be exempt from informed consent requirements. Compliance with IRB and informed consent regulations depends on the nature and purpose of the study and should be evaluated accordingly.

CMS may reimburse providers for certain costs associated with some IDE-approved clinical trials. The charge cannot exceed the costs of development, manufacture, handling, and research and must be justified in the IDE application. This charge may be passed on to participants in a study only if an IRB-approved informed consent document fully describes any additional costs that may result from participation in the research. Providers seeking CMS reimbursement for clinical trial device and patient care costs must first consult their local Medicare contractor for determination of Medicare reimbursement. There are also special billing instructions for hospitals and physicians who are submitting claims for reimbursement for clinical trials involving an investigational device under the investigational device regulation. Medicare contractors are also responsible for making coverage determinations on IDE-exempt-nonsignificant risk devices that are the responsibility of the hospital's IRB. By law, CMS can pay only for services that are considered reasonable and necessary, and determination of coverage is made by Medicare contractors. Nonfederal payers may have different specific policies regarding coverage of tests that are considered investigational.

Verification/validation. The principles of verification and validation exist to ensure standards of laboratory practice and accuracy of test results generated by clinical laboratories. As discussed above, laboratories in the United States are required to follow federal regulations established by CLIA. Guidelines used by many laboratories to help accomplish this are available in a variety of documents published by CLSI and laboratory accrediting agencies. Efforts to provide worldwide standards are changing some of the current practices and concepts. In 2003, ISO published document 15189, a global quality standard for use in clinical laboratories (47). Many of the concepts and terminologies used in document 15189 were imported from ISO document 9000, which was intended for general applications but used by some laboratories before the laboratory-specific document was published (77). Variations in the terms used as these guidance documents have evolved and undergone revision have caused some confusion.

(i) Verification. Both the FDA and ISO define the term "verification" broadly as "confirmation through the provision of objective evidence, that specified requirements have been fulfilled" (23, 47). CLIA uses the term "verification" specifically to relate to confirmation that the laboratory using a test can replicate the manufacturer's performance claims when the test is used according to the package insert. Initial FDA clearance or approval of a product does not predict how that product will perform in the end user's laboratory under actual testing conditions and with the specimen mix encountered in a particular patient population. Verification applies to unmodified nonwaived (moderate- and high-complexity) tests that have been cleared or approved by the FDA and are labeled "for *in vitro* diagnostic use." An updated list of FDA-cleared/approved molecular diagnostic tests is maintained on the Association for Molecular Pathology website (www.amp.org/) under the "Resources" tab. Laboratories must ensure that devices previously cleared or approved by the FDA are performing as expected. The specific CLIA definition refers to the requirement that each laboratory that introduces an unmodified FDA-cleared or -approved test must demonstrate and document "that it can obtain performance specifications comparable to those established by the manufacturer for the following performance characteristics: accuracy, precision, reportable range of test results for the test system, verify that reference intervals (normal values) are appropriate for the laboratory's patient population" (28).

(ii) Validation. The FDA and ISO both define the term "validation" as "confirmation by examination and provision of objective evidence that the particular requirements for a specific intended use can be consistently fulfilled" (25, 47). The FDA and ISO terms for verification and validation are similar, with the distinction having to do with intended use. The term "intended use" in these documents is established by the manufacturer or developing laboratory at the time of assay development and has to do with the purpose and population for which the test was intended (e.g., diagnosis, following treatment, etc.). Relevant performance characteristics are then determined based on the intended use of the test. Some people interpret "intended use" as referring to the relevant clinical laboratory in which the test is performed, rather than the manufacturer or developing laboratory, thus causing confusion. The World Health Organization (WHO) defines validation as "the action (or process) of proving that a procedure, process, system equipment or method used works as expected and achieves the intended result" (18, 79).

There is also confusion because CLIA does not specifically use the term "validation" but refers to "establishment of performance specifications" (28). Performance specifications for nonwaived (formerly called moderate- or high-complexity) laboratory-developed tests must be established by the laboratory at the time of assay development. Establishing performance specifications is a one-time process that must be completed prior to reporting patient results. The process includes determination of accuracy, precision, reportable range, reference interval, analytical sensitivity, and analytical specificity. The laboratory then performs ongoing quality control and quality assessment as required in the CLIA final rule (29, 30, 31, 32).

The ISO definition of validation includes establishing performance specifications as well as quality assessment measures. According to the ISO definition, specific information should be documented as part of the validation process (47, 77). The information that should be recorded includes identity of the analyte, purpose of the examination and goals for performance, performance requirements (detection limit; quantification limit; linearity; sensitivity; measurement precision, including measurement repeatability and reproducibility; and selectivity/specificity, including interfering substances and robustness), specimen type, required equipment and materials, calibration procedures, step-by-step instructions for performing the examinations, quality control procedures, interferences, calculations, measurement uncertainty, reference intervals, reportable interval, alert/critical values, test interpretation, safety precautions, and potential sources of variation. In addition, for each validation study, acceptance/rejection criteria, results obtained, control and calibration procedures, data analysis, performance characteristics determined, comparison of results with other methods, factors influencing results, carryover when applicable, and interferences or cross-reactivity should be reported (47, 77).

In the context of this review, the term "validation" will be used to refer to the analytic performance characteristics that need to be initially established at the time of assay development. Whether tests are verified (providing confirmation that an FDA-approved test is performing as expected) or validated (establishing performance specifications), there must be ongoing quality control and quality assessment.

ESTABLISHMENT OF PERFORMANCE SPECIFICATIONS FOR LABORATORY-DEVELOPED TESTS

Performance specifications for laboratory-developed tests must be established for the following characteristics: accuracy, precision, analytical sensitivity, analytical specificity to include interfering substances, reportable range, reference intervals (normal values), and any other characteristics required for test performance (28). These performance specifications can be established by doing the following experiments (77): a comparison-of-methods experiment to estimate inaccuracy/bias (may include a recovery experiment), a replication experiment to estimate imprecision, a linearity experiment to determine reportable range and lower limit of quantification (LLOQ) (for quantitative assays), a limit-of-detection experiment to estimate the lowest concentration that can be detected, an interference experiment to determine constant interferences, and a reference value study to determine reference range(s).

It is helpful to organize these experiments into a general plan in which the basic experiments that define the assay characteristics are done first and are followed by the more extensive required studies (77). When using a general plan in this manner, basic data are collected in a minimum amount of time and can reveal analytic errors early in the process. It is important to analyze the data as they are being collected so that problems can be easily identified as the experiments are being performed. If an organized, stratified experimental plan is not followed, unsuitable performance characteristics may be revealed only late in the process, necessitating assay modification and repetition of experiments.

The following order of experiments is suggested as an effective general plan (the plan may be modified to accommodate any unique characteristics of a method or any special re-

TABLE 2. Possible protocol for determining reportable range, analytical sensitivity, and precision in combined experiments based on current CLSI guidelines

Performance characteristic	Analyte concentration tested[a]											Comment(s)
	Low					Medium			High			
	1	2	3	4	5	6	7	8	9	10	11	
Reportable range (for quantitative assays)	×	×	×	×	×	×	×	×	×	×	×	7-11 concentrations across anticipated measuring range; 2-4 replicates on same day
Analytical sensitivity (LOD)	×	×	×	×	×							8-12 replicates of 4-5 samples at the low concentration end over 5 days
Precision												
Qualitative assay	×	×	×									Use concentrations at LOD, 20% above LOD, and 20% below LOD; test in duplicate over 15 days (include data from analytical sensitivity runs to provide data over 20 days)
Quantitative assay		×			×					×		Use high, low, and LOD concentrations; test in duplicate over 19 days (include data from reportable range study as day 1 to provide data over 20 days)

[a] ×, the concentration is tested. The reportable range is from concentration 2 to concentration 10; the LOD, LLOQ, and upper limit of linearity are at concentrations 2, 4, and 10, respectively.

quirements of the laboratory or the patient population it serves): (i) reportable range (linearity study), (ii) analytical sensitivity (limit-of-detection study), (iii) precision (replication study), (iv) analytical specificity (interference studies), (v) accuracy (comparison-of-methods study), and (vi) reference interval (reference value study).

Establishing performance characteristics for molecular assays can be quite costly for clinical laboratories. The most economical approach may be to design an experiment to determine reportable range and include enough replicates at lower concentrations of analyte to simultaneously determine analytical sensitivity (limit of detection) and lower limit of quantification (for quantitative studies) (Table 2). By testing multiple replicates across the dynamic range of the assay in the same run and in different runs and applying suitable statistical tests, precision can also be established using some of the data points from reportable range and analytical sensitivity experiments.

Reportable Range

Definition. CLIA uses the term "reportable range" to refer to the span of test result values over which the laboratory can establish or verify the accuracy of the instrument or test system measurement response (27). Laboratories may quantitatively report only results that fall within the reportable range. Reportable range does not apply to qualitative clinical tests (8). Reportable range and other terms, such as measuring interval, analytical measurement range (AMR), and linear range, are used interchangeably to refer to the same performance characteristic of quantitative assays (8, 77). In practice, it is common to refer to the "linear range," and laboratories generally use a linearity experiment to determine the reportable range for a test (8, 77). It is not mandatory that a method provide a linear response, but a linear response is desirable since test results that are in the linear range are considered to be directly proportional to the concentration of the analyte in the test samples (8, 77). The boundaries of the reportable range are the lowest and highest analyte concentrations that generate results that are reliably produced by a test method without dilution of the specimen (16). The lower limit must also be clinically relevant and acceptable for clinical use (14). The lower limit of linearity is frequently referred to as the lower limit of quantification (LLOQ) and the upper limit of linearity as the upper limit of quantification (ULOQ). The upper limit of linearity may be restricted by the highest available concentration in a sample or by the saturation of the signal generated by the instrument.

Study suggestions. A linearity experiment involves testing a series of samples of known concentrations or a series of known dilutions of a highly elevated patient specimen or standard with concentrations across the anticipated measuring range. Equally spaced intermediate concentrations are recommended but not strictly required (8).The measured values are compared to the assigned values, typically by plotting the measured values on the *y* axis and the assigned values on the *x* axis. The reportable range is assessed by fitting a regression line through the points in the linear range (8). Deming regression may be more relevant than simple linear regression if there is uncertainty in the assigned value of quantification. More complicated statistical calculations may be required if

a wide range of concentrations are tested and the scatter around the regression line does not appear to be constant across the range (14).

It is understood that poor precision will affect the linearity of an assay. Therefore, it is recommended to check for poor repeatability as part of the linearity study by testing two to four replicates at each concentration, depending on the expected imprecision of the assay (8). More replicates may need to be tested at the low end to adequately determine the LLOQ. The use of as many as 25 to 40 replicates has been recommended (14).

If values beyond the upper limit of linearity are of clinical interest, it may be desired to allow dilution of patient specimens to extend reporting of test values beyond the upper limit of the linear range. A dilution protocol to allow retesting of out-of-range patient specimens must be validated. A dilution of 1:2, 1:5, or 1:10 is generally used. It is important to maintain an appropriate matrix when preparing dilutions (12). Water, saline, or base matrix (for plasma specimens) may be acceptable. The appropriate diluents must be documented.

Sample sources. For linearity experiments it is required that samples of known concentrations or known relationships established by dilution be used. For some analytes, reference panels containing samples at a variety of concentrations may be commercially available. When panels are not available, it may be necessary to prepare dilutions of standard solutions or patient specimens containing high concentrations of the analyte of interest. Samples containing high concentrations of analyte may include quality control materials, proficiency testing samples, or patient specimens tested by an acceptable method. Equally spaced concentrations are not required.

If multiple matrix specimen types (urine, serum, spinal fluid, etc.) are to be assayed, a linearity study should be carried out for each specimen type, as the matrix background can significantly affect results.

Number of samples. To establish a linear range for laboratory-developed tests, CLSI recommends using 7 to 11 concentrations across the anticipated measuring range (8). To ascertain the widest possible linear range, additional samples at concentrations 20 to 30% wider than the anticipated measuring range can be used. Samples should be tested in replicates of two to four, depending on the anticipated imprecision of the assay (8). Ideally, all results for a single analyte will be obtained on the same day (8).

Data analysis. The first step in analysis of the data is to prepare an *xy* plot with measurand results on the *y* axis versus the expected or known values on the *x* axis. Individual data points or mean values can be plotted for each set of replicates. Plotting individual results will allow visual detection of outliers that do not fit the pattern represented by the rest of the data. A single outlier in a data set can be removed and does not need to be replaced. Two or more unexplained outliers cast doubt on the precision of the test system. The line can be drawn manually or with the aid of a computer program. A visual examination of the plot will show whether there is obvious nonlinearity or whether the range of testing should be narrowed or expanded. It will also give insight into the most appropriate procedures for the subsequent statistical analysis (8).

There are a wide variety of statistical analysis techniques to evaluate linearity, and there is not consensus on the optimal analytical approach. Different approaches have been advocated by CAP, CLSI, and manufacturers of diagnostic methods and control materials, including visual review, least-squares regression, comparison of slopes for line segments, comparing observed and expected values with allowances for error, and polynomial regression (8, 77). Quantitative and objective approaches are preferred over subjective visual evaluation to describe linearity.

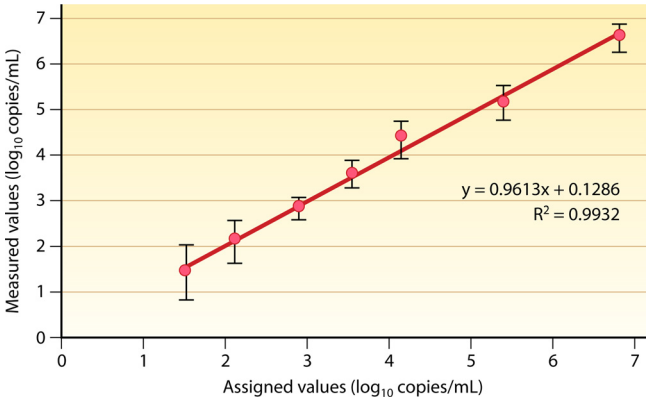

FIG. 1. Plot of results from a linearity experiment to determine reportable range. Seven concentrations of analyte prepared by dilution of a high-concentration standard were tested in triplicate. Assigned values, (converted to $\log_{10}$) were plotted on the *x* axis versus measured values (converted to $\log_{10}$) on the *y* axis using Microsoft Excel. Linear regression analysis gave the equation $y = 0.9613x + 0.1286$ ($r^2 = 0.9932$). A second-order polynomial trendline gave the equation $y = -0.028x^2 + 1.1937x - 0.2667$ ($r^2 = 0.9954$). A third-order polynomial trendline gave the equation $y = 0.009x^3 + 0.1388x^2 + 1.5994x - 0.6948$ ($r^2 = 0.9958$). The second- and third-order polynomials are not significantly better ($P > 0.05$) than the linear equation, which indicates that the linear equation is the best fit for the data. The fitted regression line shows the slope to be significantly different from zero and the intercept to be not significantly different from zero. The regression coefficient of 0.9973 verifies the linearity throughout the range tested. The reportable range in this example translates to 30 copies/ml (LLOQ) through 3,000,000 copies/ml (ULOQ). Because of imprecision at the low end, more replicates in a precision experiment may need to be tested to adequately determine the LLOQ before accepting the reportable range.

(i) Linear regression analysis. It is commonly accepted that linear regression analysis can be used if the relationship between expected and observed values is a line without noticeable curvature and if the scatter in the *y* direction around the regression line appears constant across the concentration range. If the data appear to have a curved relationship, as is often the case when testing analytes that cover a wide concentration range, log transformation of the data points may straighten the line. Log transformation consists of taking the log (generally base 10) of each observed value. All PCR-based data should be plotted after $\log_{10}$ transformation.

Linear regression describes the straight line that best predicts *y* from *x*. Linear regression does this by finding the line that minimizes the sum of the squares of the vertical distances of the points from the line. It is important to realize that linear regression does not test whether the data are linear. It assumes that the data are linear. Linear regression finds the slope and intercept to create the equation for the best-fitting line. In a linear equation, the slope should be significantly different from

zero and the y intercept should not be significantly different from zero. The regression coefficient, r^2, a measure of goodness-of-fit of linear regression, is also calculated. The value r^2 is a fraction between 0.0 and 1.0 and has no units. An r^2 value close to 0.0 means that there is no linear relationship between x and y, and knowing x does not help predict y. When r^2 equals 1.0, all points lie exactly on a straight line with no scatter and knowing x predicts y perfectly.

(ii) Polynomial regression analysis. Because linear regression is based on the perhaps faulty assumption that the data are linear, the preferred statistical approach is to use polynomial regression analysis (Fig. 1) (8). Polynomial regression analysis assumes that the data set is not linear and uses tests for nonlinearity. Polynomial regression analysis can be accomplished with most commercial statistical spreadsheet software programs. The goal of polynomial regression analysis is to find the polynomial function that properly fits the data points and to assess whether there is a significant difference between linear and quadratic fits. First-, second-, and third-order polynomial regression analyses should be performed. It is usually not necessary to carry regression analysis beyond the third-order polynomial, since in most cases it is hard to attach a biological meaning to exponents greater than 3 (8).

To perform polynomial regression analysis, a linear regression is done first, fitting an equation of the form $y = a + bx$ (linear) to the data. Then, a second-order polynomial (quadratic; $y = a + b_{1x} + b_{2x}{}^2$), which produces a parabola (either up or down), is fit to the data. The r^2 will always increase when a higher-order term is added but needs to be tested to see whether the increase in r^2 is significantly greater than that expected due to chance. Next, a third-order polynomial (cubic; $y = a + b_1x + b_2x^2 + b_3x^3$), which produces an S-shaped line where nonlinearity occurs at the ends of the measuring range, is fit and the effect on r^2 is tested. The degrees of freedom are calculated for the second and third (nonlinear) polynomials, and a t test is performed to see whether the equation fits the data significantly better than an equation of a horizontal line. If the second- or third-order polynomials are not significant ($P > 0.05$), the data are considered linear and the analysis is complete, except to check for high imprecision. If either the second- or third-order polynomial is significant ($P < 0.05$), then the data are nonlinear. The deviation (often expressed as a percentage) from linearity should then be calculated for each concentration level. It is possible that the degree of deviation may not be large enough to affect patient results. A predefined acceptable allowance is often difficult to determine and may be based on clinical decision levels, literature review, or discussion with colleagues. If the nonlinear concentration is at either or both ends of the range of concentrations, one option is to reduce the linear range by removing the nonlinear point(s) and rerunning the statistical analysis.

Analysis of linearity experiments must consider the contribution made by random error (precision) (8). The best estimate of precision is the standard deviation (SD) or coefficient of variation (CV). If the estimates of SD or CV are equal across levels and within an acceptable range, the precision is adequate for a reliable determination of linearity. If the SD or CV is larger than acceptable, the precision may not be adequate for a reliable determination of linearity. If repeatability is severely different across concentrations, more sophisticated weighted regression analysis may be required. If precision is unacceptably high at either end of the concentration range tested, the linear range may need to be reduced.

Analytical Sensitivity

Definition. The "analytical sensitivity" of an assay is defined as the ability of the assay to detect very low concentrations of a given substance in a biological specimen (63). Analytical sensitivity is often referred to as the "limit of detection" (LOD). LOD is the lowest actual concentration of analyte in a specimen that can be consistently detected (e.g., in ≥95% of specimens tested) with acceptable precision, but not necessarily quantified, under routine laboratory conditions and in a defined type of specimen (14, 15).

LOD is an important performance characteristic that must be determined for both quantitative and qualitative tests. LOD is expressed as a concentration (usually copies/ml; copies/μg DNA for molecular assays) such that the lower the detectable concentration of analyte, the greater the analytical sensitivity of the assay. Analytical performance at the low concentration limit is often of great interest in molecular infectious disease tests because it defines the ability of the test to diagnose disease and determine treatment endpoints. Knowing the limit of detection is also important to determine the concentration to be used as a low positive control that will be monitored to ensure consistency of performance between runs at levels near the cutoff and to ensure that the LOD does not change when new reagent lots are used. For quantitative tests, it is possible for the LOD to equal the LLOQ if the observed bias and imprecision at the LOD meet the requirements for total error for the analyte (14). Most often, however, the LOD resides below the linear range of an assay and is lower that the LLOQ (14). The LOD cannot be higher than the LLOQ (14).

Analytical sensitivity is an inherent characteristic of an assay and is very different from the diagnostic or clinical sensitivity of the assay. Diagnostic sensitivity becomes relevant only when an assay is used to detect a condition or disease in a population. Diagnostic sensitivity is defined as the proportion or percentage of individuals with a given disorder who are identified by the assay as positive for the disorder (63). An assay with high analytical sensitivity does not necessarily have acceptable diagnostic sensitivity. An assay with perfect analytical sensitivity may fail to give a positive result if the target substance is not present in the processed specimen because of vagaries in collecting or processing the specimen. Conversely, there may not be a clinical need for or interest in very low concentrations approaching zero (primarily with latent viruses).

Study suggestions. A variety of procedures are available for establishing the LOD for laboratory assays. The LOD is generally determined in one of two ways: (i) statistically, by calculating the point at which a signal can be distinguished from background, or (ii) empirically, by testing serial dilutions of samples with a known concentration of the target substance in the analytical range of the expected detection limit (14). For medical applications of molecular assays, it is generally more useful to use the empirical method to estimate the detection limit.

Statistically determining the LOD based on blank specimens (specimens not containing the analyte of interest) is commonly used in analytical chemistry. The rationale behind calculating

the LOD as a function of the limit of blank (LOB) is to ensure that a blank sample does not generate an analytical signal that might overlap with a signal indicating a low concentration of analyte. Adding a population variance measurement from replicate testing of a low-level sample(s) to the calculation ensures that the signal from a specimen containing a low level of analyte can reliably be distinguished from background noise or signals produced in the absence of analyte (77). The challenge of determining the LOD as a function of the LOB in molecular assays is that the raw data associated with blank specimens may not be easily accessible, since the actual signal for results that fall below the threshold is often automatically assigned a value of zero. When attempted, the recommended number of blank and low-level samples to be used to establish the LOD is 60 (14). It may be necessary to spike several samples whose concentrations are in the analytical range of the expected detection limit. Means and pooled standard deviations are calculated from the resulting values. The LOD is estimated as a 95% one-sided confidence limit by calculating the mean of the blank plus 1.65 times the standard deviation of the blank plus 1.65 times the standard deviation of a low-concentration sample(s) ($LOB = mean_{blank} + 1.645SD_{blank}$; $LOD = LOB + 1.645SD_{low\text{-}concentration\ sample}$).

When empirically determining the LOD, repeated measurements are obtained from samples with relevant low concentrations. Experiments to determine analytical sensitivity should be spread over several (e.g., five) days so that the standard deviations reflect performance of the assay over a range of typical laboratory conditions but without a change in reagent lots (14).

Analytical sensitivity should be determined for each type of specimen matrix that will be tested in the clinical laboratory.

Genetic variants (e.g., hepatitis C virus [HCV] genotypes) may not have the same analytical sensitivity in an assay due to differences in primer binding and amplification efficiency. Analytical sensitivity validation per genotype, when possible, should be done.

For multiplex assays, the LOD must be determined individually for each analyte tested by the multiplex assay (18). In addition, analytical sensitivity testing needs to incorporate the use of combinations of target at high and low concentrations to show that a high concentration target will not outcompete a low-concentration target present in the specimen. Similarly, when consensus assays are used for generic amplification (e.g., to generically detect members of the herpesvirus group), the assay needs to be validated for each specific virus.

Sample sources. When determining the limit of detection, it is essential that well-characterized samples with true or accepted values of analyte are used. Characterized samples include cell lines containing known quantities of target, standards, quality control materials, proficiency testing samples, or patient specimens tested by an acceptable method.

Several spiked samples should be prepared in the appropriate matrix using known analyte concentrations within the range of the expected detection limit. The matrix is often previously tested samples that were found to be negative for the target analyte. Since matrix differences exist from specimen to specimen, it is best to spike a set of specimens from different subjects (five or more), rather than using just one patient specimen or pool as the matrix (14).

Number of samples. It is understood that more samples and more measurements provide better estimates with less uncertainty (14). However, when validating molecular tests, the number of samples may be limited by sample availability and budget concerns. The literature often recommends 20 measurements at, above, and below the probable LOD as determined by preliminary dilution studies (77). CLSI also recommends 20 measurements to verify a manufacturer's claim but suggests a minimum of 60 data points (e.g., 12 measurements from each of five samples) for a manufacturer to establish an LOD claim (14).

Data analysis. Calculations to determine LOD based on LOB are presented above. Probit analysis is a commonly used type of regression analysis when empirically determining the lowest concentration of analyte that can be reliably detected by molecular assays (1, 4, 4a, 35, 36, 37, 38, 44, 51, 53, 54, 57, 58, 59, 60, 61, 62, 64, 67, 68, 74, 75, 78, 80, 81). Probit analysis is used for studies that have binomial response variables and reliably estimates biological endpoints when the number of replicates tested at each concentration is small. A binomial response variable refers to a response with only two possible outcomes. For limit-of-detection studies, the response of the assay using various concentrations of analyte has only two outcomes: detected or not detected. Each concentration tested must have a minimum of three replicates, but large sample sizes (10 or more replicates at each concentration) are always better than small sample sizes. The same number of replicates is not needed for each concentration. Probit analysis transforms the concentration-response curve to a straight line that can then be analyzed by regression using either least squares or maximum likelihood. Maximum likelihood is used by most computerized statistical packages and is considered to be more precise. Probit analysis is available using the Analyze-it statistical analysis software addendum for Microsoft Excel (Analyze-it Software Ltd., Leeds, United Kingdom), SPSS (Statistical Package for Social Sciences; IBM, Chicago, IL), Statgraphics (Statpoint Technologies, Warrenton, VA), MiniTab (MiniTab Inc., State College, PA), and other commercially available statistical software packages.

Various endpoints can be used to compare the different concentrations of analyte, but for limit-of-detection studies, typically the 95% endpoint (C_{95}) is the most widely used. The C_{95} represents the concentration at which 95% of the samples containing that concentration of analyte test positive.

Probit analysis can be done by using Finney's table to estimate the probits (short for probability unit) and fitting the relationship by eye; by calculating the probits, regression coefficient, and confidence intervals by hand; or, most easily, by using a computerized statistical package to perform the analysis. Probit analysis involves first listing the concentrations tested, the number of samples per concentration that responded, and the total number of samples tested per concentration. The percentage responding at each concentration is then converted to probits. The $\log_{10}$ of the each concentration is calculated, and a graph of the probits versus the log of the concentrations is constructed. A line of regression is fit, and the C_{95} is determined by searching the probit list for a probit of 9.50 and then taking the inverse log of the concentration with which it is associated (Table 3). Although LOD determinations produce a specific number above which an

TABLE 3. Limit of detection using probit regression analysis[a]

Copies/ml	Log_{10} copies/ml	No. of replicates	No. positive	% Positive	Probit value
1,000	3	8	8	100	NA[b]
500	2.69897	8	8	100	NA
200	2.30103	8	8	100	NA
100	2	8	8	100	NA
50	1.69897	8	7	88	6.13
10	1	8	2	25	4.33

[a] A series of six samples was prepared by diluting a high-concentration standard. Each sample was tested eight times. Probit regression analysis gave a probit value of 6.65, which converts to a C_{95} value (concentration detectable 95% of the time) of 79.60 (log_{10} C_{95} = 1.90), indicating that the limit of detection is about 80 copies/ml and that samples containing that concentration would be detected 95% of the time.

[b] NA, not applicable.

analyte is considered detectable, this is an oversimplification. There is a continuum of uncertainty when dealing with low-level data, and distinct cutoffs do not exist. Most computer programs also automatically calculate the 95% confidence interval, which indicates that there is 95% certainty that the true value lies within the boundaries of the confidence interval as it might be estimated from a much larger study.

The analytical sensitivity of a quantitative assay technically refers to the ability of the assay to detect a change in concentration of the analyte (16). Assays that detect smaller changes have greater analytical sensitivity. The International Union of Pure and Applied Chemistry (IUPAC) definition of analytical sensitivity relates to the slope of the calibration curve (16). Assays that have steeper slopes are more sensitive to slight changes in amount of analyte. This is not useful in validation, however, and in practice, the limit of detection or the limit of quantification is determined more frequently (16).

Precision

Definition. The term "precision" refers to how well a given measurement can be reproduced when a test is applied repeatedly to multiple aliquots of a single homogeneous sample (43). Precision is defined as the "closeness of agreement between independent test/measurement results obtained under stipulated conditions" (45). When considering precision, it is important to remember that a measurement may be very precise (replicates have the same result) but not very accurate (the real value is much different). The ideal assay is both precise and accurate. Precision (also referred to as random analytical error) is related entirely to random error caused by factors that vary during normal operation of the assay. While portions of many laboratory-developed molecular tests are automated, most still have nonautomated steps involving timing, temperature, etc., that may be subject to significant variation, even if the same pipettes, instruments, etc., are used. Differences in the techniques of individual operators may also introduce considerable random variation. Reaction conditions in automated systems may also be the source of (albeit usually smaller) variation. Precision does not have a numerical value but may be expressed qualitatively as high, medium, or low (7). For numerical expression, the term "imprecision" is used. Imprecision is defined as the "dispersion of results of measurements obtained under specified conditions" (7). Different terms have been used to describe different components of precision (or imprecision). "Repeatability" and "reproducibility" are considered to be the extreme measures of precision, with repeatability (or within-run imprecision) being the smallest measure of precision and involving measurements carried out under the same conditions (same operator, reagent lots, instrument, laboratory, time, etc.) and reproducibility (run-to-run imprecision, day-to-day imprecision, etc.) being the largest measure of precision and involving results of measurements under changed conditions (different operators, reagent lots, time, laboratory, etc.) (45, 46). All other measures of precision are considered to be "intermediate" measures, and conditions must be explicitly specified (7).

Study suggestions. Precision is commonly evaluated by performing a replication experiment to observe the variability in results generated under the normal operating conditions in the laboratory. There are many sources of variability to consider when designing a replication experiment. The best design will include sufficient conditions so that all the sources of variability in the setting in which the test is being performed are reflected in the estimate of precision. It is not necessary to separately estimate the relative contribution of each source or component.

The length of time over which the experiment is conducted is an important factor to consider. When samples are analyzed in a single run (within-run imprecision), the variability is expected to be low because results are affected only by the operating conditions present at the time of the run. A single-run study errs in that it underestimates imprecision by failing to include significant sources of variation that occur when an assay is run over time. Single-run studies are most reflective of the best performance of the assay rather than a realistic estimate of precision that is seen over time. To ensure a robust estimate of precision that better reflects the range of results seen over time and best represents the expected future performance of the assay, replication experiments are usually designed in a between-day format. Performing the experiment over a minimum of 20 operating days is generally recommended, because day-to-day imprecision may be large and the experiment must ensure that the separate components of error are adequately covered (7). FDA method validation guidance for some bioanalytical method tests allows precision testing with a minimum of five determinations at each concentration (43).

Laboratories may wish to incorporate other possibly significant sources of variation, such as different operators, multiple reagent lots, multiple instruments, etc., into the study design. The data analysis for these more complex designs becomes more complicated, since it must reflect the influence of all of these factors. A single lot of reagents and a single operator are often used for the entire study, but the protocol should state this fact, and it must be understood that results may underestimate the true long-term precision of the assay. Molecular tests with run times of longer than 2 to 3 h are generally done once per shift in most laboratories. If normal laboratory operations would include performing the assay more than once per day, it would be important to include a between-run evaluation in the experimental design. Likewise, if the same assay will be performed at multiple locations, it is important to include a between-laboratory component in the replication study, since

TABLE 4. Replication experiment to evaluate precision[a]

Concentration (copies/ml)	Log_{10} concentration	% CV (total)	SD	95% CI (log viral load)	Log change 95% CI	Fold change 95% CI
250,000	5.39	1.93	0.104	5.11–5.67	0.56	3.6
5,000	3.69	1.75	0.130	3.43–3.95	0.53	3.3
300	2.48	6.91	0.164	2.15–2.81	0.66	4.6

[a] Three concentrations of analyte spanning the reportable range were used to evaluate the precision of a quantitative assay for human cytomegalovirus. Samples at each concentration were tested in duplicate in one run per day over 20 days using two operators. Between-day, operator-to-operator, and total imprecision were evaluated using analysis of variance of log_{10}-transformed data. The coefficient of variation (CV) and standard deviation (SD) were calculated for each concentration. Total imprecision (% CV) values for the three concentrations ranged from 1.75% to 6.91%, indicating that the assay is less precise at lower concentrations. Relatively little variance was attributable to between-day or operator-to-operator components except for the low-concentration sample, for which the day-to-day imprecision was much greater than with the other concentrations (data not shown). Calculation of 95% confidence intervals (CI) revealed that a 3- to 4-fold change at mid- to high viral loads and about a 5-fold change at viral loads near the limit of detection may represent imprecision of the assay rather than a true biological difference.

variability between laboratories is often the largest single contributor to imprecision (7).

(i) Qualitative tests. Precision studies for qualitative tests should provide an estimate of the imprecision of the method at analyte concentrations near the limit of detection (11). It is not appropriate to measure the imprecision of a qualitative assay with high-positive samples, since they are too far away from the medical decision point (11). CLSI document EP12-A2 describes a protocol for performing a precision experiment at analyte concentrations near the limit of detection (11). The protocol describes a repeatability experiment that could be modified to be performed over several days in order to better incorporate elements of reproducibility, as discussed below for quantitative tests. The protocol suggests preparing three samples: one with an analyte concentration at the limit of detection, one with a concentration 20% above the limit of detection, and one with a concentration 20% below the limit of detection. The three samples would be tested in replicates up to 40. The document acknowledges that it may not be feasible or cost-effective to test 40 replicates and provides confidence limits achievable with fewer replicates, but it also cautions that statistical power is less with smaller numbers of replicates.

(ii) Quantitative tests. Precision studies for quantitative tests would ideally generate data for concentrations covering a large portion of the measuring range. Since cost concerns and time limitations must be considered, it is suggested that the replication study be done with a minimum number of samples and then additional samples be tested if necessary (7). CLSI document EP5-A2 suggests that a high-level sample, a low-level sample, and a sample as close as possible to a medical decision level (usually the limit of detection) should be tested (7). Higher levels of variability will generally be at the low and high ends of the measuring range. If there are large differences in the precision estimates at the three levels, then it may be necessary to test additional concentrations to fully describe the performance of the assay. It is also suggested that the study design include testing the samples in duplicate twice a day over 20 working days (7). This type of study design will allow calculation of within-run, between-run, and between-day variances, which can then be combined to determine the total variance of the assay. The study may be performed using one run per day if that is reflective of how the assay will be used in standard practice (7).

Sample sources. Repeatability studies to determine precision are done using samples with known concentrations of analyte. Test materials could include standards, quality control materials, proficiency testing samples, or patient specimens in sufficient quantity to complete the study. The samples should be selected or prepared in a matrix as close as possible to the appropriate clinical specimens. Since repeatability studies are generally done over a period of several weeks, samples must be adequately stored, generally at −20°C or −70°C or lower, to ensure stability over time (15, 17).

Number of samples. Generally, the estimate of imprecision improves with greater numbers of available observations (7). The estimates of precision using only a few samples might be expected to scatter around the "true" value and the estimates obtained from more observations to cluster more closely around the "true" precision. In general, a larger number of observations leads to more confidence in an estimate.

Data analysis. Precision is usually expressed on the basis of statistical measurements of imprecision, such as the standard deviation (SD) or coefficient of variation (CV) (Table 4). The specific analysis-of-variance (ANOVA) formulae used to calculate precision depend on the number of replicates per run, the number of runs per day, and the number of days over which the experiment is conducted, as well as the number of instruments used in the evaluation, number of reagent lots, number of operators, etc., used in the evaluation. Many statistical software packages correctly calculate the components of variance, but not all of them do (7). It is not possible to generalize the equations needed for all experimental designs, and the statistics can become relatively complex when it is necessary to include the influence of many factors in the estimation of precision (7). CLSI document EP5-A2 provides worksheets and equations to determine the precision at each concentration tested using the protocol described above in which samples are tested in duplicate twice a day over 20 days. Repeatability, between-run/within-day, and between-day variances for each concentration are calculated and then combined to obtain the total variance. Modified equations for use when the experimental design includes only one run per day are also given in CLSI document EP5-A2 (7). These formulae allow for calculation of repeatability and between-day variances but do not contain a between-run/within-day component.

While the calculations for estimating precision are generally available, the criteria for determining whether the calculated variation indicates acceptable performance for molecular assays are not. When validating the precision of a modified FDA-approved test or changing an extraction method or other key component of an already-established test, acceptable ranges of error may be available from the existing valida-

tion studies or CAP or other proficiency testing summaries. When criteria are not available, imprecision is often expressed as the target value plus or minus two or three SDs or the target value plus or minus a percentage (e.g., target ± 10%) (7). FDA guidance is available for some bioanalytical procedures, such as gas chromatography, high-pressure liquid chromatography, combined gas chromatography/mass spectrometry, and some ligand-binding immunological and microbiological procedures (43). The current guidance suggests that the precision around the mean value should not exceed 15% of the CV, except at the LLOQ, where precision should not exceed 20% (43). General guidelines for chemistry assays suggest that analyte levels that vary more than about 30% from their central value are significant (69). Although not specifically formulated for molecular assays, these numbers may be a reasonable general indicator of acceptable imprecision. In determining acceptable performance, it may also be useful to construct precision profiles in which the SD or CV is plotted as a function of analyte concentration, since precision often varies with the concentration of analyte being considered (56). If the precision estimates are the same at all concentrations, then there is evidence of constant precision. If the estimates are not similar, knowing that an assay is less precise at the upper or lower part of the analytical measurement range may be helpful in specifically describing the performance characteristics of the assay.

Analytical Specificity

Definition. The analytical specificity of an assay is different from the diagnostic specificity of an assay (63). "Diagnostic specificity" refers to the percentage of individuals who do not have a given condition and are identified by the assay as negative for the condition. In some situations, the diagnostic specificity of a molecular assay can be diminished without loss of analytical specificity. Situations that contribute to diminished diagnostic specificity of infectious disease molecular assays include false-positive reactions that occur because of sample contamination or detection of nucleic acid fragments from organisms that are not viable and are therefore not capable of causing disease. False-positive results can also be caused by interfering substances or organisms that are genetically similar.

"Analytical specificity" is the parameter that needs to be determined for validation of molecular assays. Analytical specificity refers to the ability of an assay to detect only the intended target and that quantification of the target is not affected by cross-reactivity from related or potentially interfering nucleic acids or specimen-related conditions. The two aspects of analytical specificity are cross-reactivity and interference. Both are determined by performing interference studies.

Cross-reactivity. Organisms that should be tested to rule out potential cross-reactivity include organisms with similar genetic structure, normal flora organisms that could concurrently be present in the specimen, and organisms that cause similar disease states or clinically relevant coinfections.

Other potential cross-reacting nucleic acids may also be revealed by comparing the sequence of the nucleic acid target to other known sequences in publically accessible nucleic acid sequence databases (65). GenBank is the National Institutes of Health (NIH) genetic sequence database and is part of the International Nucleotide Sequence Database Collaboration (www.ncbi/nlm/nih.gov/GenBank). Basic Local Alignment Search Tool (BLAST) sequence analysis tool (http://www.ncbi.nlm.nih.gov) can be used to search the National Center for Biotechnology Information (NCBI) database and will display homologous sequences. If it is revealed that sequence similarities exist in other compounds that could potentially be present in the patient specimen, those compounds would need to be evaluated in interference studies.

Interfering substances. The term "interfering substances" refers to the effect that a compound other than the analyte in question has on the accuracy of measurement of an analyte. Specimens from any source may contain unpredictable amounts of interfering substances. Some substances present in specimens have the potential to affect polymerase activity and interfere with or inhibit amplification of nucleic acid. These substances may originate from a variety of endogenous and exogenous sources (9). Some endogenous substances such as hemoglobin, bilirubin, or triglycerides may be readily visible in the specimen. Other potentially interfering endogenous substances are not visible and include metabolites produced in pathological conditions such as diabetes mellitus, multiple myeloma, cholestatic hepatitis, etc., or compounds introduced during treatment of a patient, such as medications, parenteral nutrition, plasma expanders, anticoagulants, and others. Alcohol, drugs of abuse, and other substances ingested by the patient may also interfere. Exogenous contaminants can be inadvertently introduced during specimen collection from sources such as hand cream, powdered gloves, serum separators, collection tube stoppers, etc. It is critical that specimens be collected in proper collection containers to prevent the presence of known interfering anticoagulants, preservatives, stabilization reagents, etc. Nucleic acid extraction procedures are often helpful in inactivating or removing interfering substances. It is important to ensure that residual reagents, such as alcohol, from nucleic acid extraction procedures are not retained in the extracted sample.

Because of the complexity of specimen matrices and the abundance and variety of potential inhibitors in clinical specimens and because interference may be subtle, the influence of all potential inhibiting substances cannot be easily assessed. Therefore, it is important to demonstrate that each specimen, or nucleic acids extracted from it, will allow amplification. This is commonly accomplished by adding an amplification control nucleic acid sequence to the specimen (further discussed below). An amplification control that fails to amplify or is outside acceptable limits indicates the presence of an inhibitor.

Study suggestions. Interference studies begin by compiling a list of cross-reacting or interfering substances that have the potential to affect the assay being evaluated (9). It is generally accepted that a specified concentration of interfering substance causes a constant amount of systematic error regardless of the concentration of the analyte of interest (77). Recommended concentrations for many common drugs and some common endogenous constituents can be found in Appendix C of CLSI Document EP-7A2 (9).

Two approaches for conducting interference studies are (i) analyzing the effect of potentially interfering substances added to specimens containing the analyte of interest using the test method (interference screen) and (ii) evaluating the bias of representative patient specimens containing the potential in-

terfering substance using both the test method and a comparative reference method if one is available and particularly if it is a method that is in routine use (comparative measurement procedure) (9, 77). The comparative measurement procedure requires that specimens known to contain both the potential interfering substance and the analyte of interest are available. Such studies also require control specimens that do not contain the potential interfering substance but span the same range of analyte concentrations in the test specimens. Appropriate specimens may not be available to establish interference for infectious disease molecular assays using the comparative measurement procedure, and more often the interference screen procedure will be more practical.

An interference screen involves testing samples containing the interfering material, with and without the analyte of interest, in the same analytical run to see if acceptable amplification can occur in the presence of potential interfering material. The analyte of interest should be present in a concentration at the medical decision limit of the assay (9), which is usually at the low end of the reportable range for infectious disease molecular assays. Significant interference is most likely to be revealed in the presence of high concentrations of interfering substance; therefore, to adequately evaluate interference and cross-reactivity, it is suggested that interference studies be designed using the highest concentration of organism or interfering substance anticipated to be found in an actual specimen. It is recommended that each sample be tested in duplicate to properly reveal systematic error with less effect imparted by the random error of the method (9, 77). If interference is found at high concentrations of interfering substance, it may be desired to construct a series of paired samples to determine a concentration, if any, which permits amplification in spite of the presence of the interfering substance (9).

Interference/cross-reactivity should be established for each specimen type used in the test system, using potentially interfering material appropriate for the specimen matrix.

For multiplexed assays, samples containing each target should be used not only to establish assay characteristics for that particular pathogen but also to rule out cross-reactivity or interference between the pathogen and primers/probes designed to detect the remaining pathogens in the assay (18).

Sample sources. Specimens containing the analyte of interest may be available in the testing laboratory and can be used to construct paired samples. The specimen is spiked with diluent (control), and a second aliquot of the specimen is spiked with the potentially interfering substance, cross-reacting organism, or nucleic acid from the organism (test). If specimens containing the target analyte are not available, paired control and test specimens may need to be constructed by adding low concentrations of analyte with and without the potential interfering substance to negative specimens. A limitation of preparing any type of spiked sample is that the substance(s) added to the negative specimen may not have the same properties as found naturally occurring *in vivo* (9).

Number of samples. There is no recommended minimum number of samples that should be tested. Many statistical software packages contain power analysis tools that use the standard deviation of the differences between the means of the paired samples to estimate the sample sizes needed to detect a significant difference.

Data analysis and criteria for acceptance. The data analysis most commonly applied is equivalent to a paired *t* test, repeated-measures test, or paired-difference test and is easily calculated using most standard statistical software packages (9, 77). Analysis is based on the difference between the means of the test and control samples and the allowable error that is clinically significant for the test. The difference between the values is calculated for each sample pair, and the mean and standard error of these differences are determined. The mean is divided by the standard error of the mean to generate a test statistic that follows a *t* distribution with degrees of freedom equal to one less than the number of pairs. Once a *t* value is determined, a *P* value can be found. If the calculated *P* value is below the threshold chosen for statistical significance (e.g., $P < 0.05$), then the influence attributed to the presence of the interfering substance is significantly different from zero.

Accuracy (Trueness)

Definition. Accuracy is a broad term that has generally been used to describe the extent to which a new test method is in agreement with a comparative or reference method. The terminology has changed somewhat to align with that of the ISO. The current metrological use of the term "accuracy" refers to the closeness of the agreement between the results of a single measurement and the true value of the analyte (48). What was previously considered to be "accuracy" is now termed "trueness." "Trueness" has replaced the term "accuracy" when referring to the closeness of the agreement between the average value obtained from a large series of measurements and the true value (if there is an international standard) or accepted reference value (if there is not an international standard) of a measurand (45). Trueness is expressed numerically as bias (lack of agreement). Bias is inversely related to trueness and refers to the average deviation from the true value due to nonrandom effects caused by a factor(s) unrelated by the independent variable (45).

Study suggestions. Trueness studies are the cornerstone of method validation. The intended use of the new test must be defined as part of the test design (15). Intended uses may be to diagnose disease, confirm a serologic diagnosis, evaluate the effect of therapy, etc. The use of a test, once validated, must be restricted to the stated purpose(s). The intended patient population and specimen types should also be fully described.

Two primary approaches have been described to evaluate trueness (10, 11, 13, 15). A comparison-of-methods study is a split-sample experiment in which results from the method under evaluation and a comparative method are assessed. A recovery study uses proficiency testing samples or other assayed materials and compares results from the method under evaluation to the expected reference value. Laboratories may choose to use either or both approaches, depending on the availability of samples.

Ideally, testing should be spread over a minimum of 5 days so that the comparison reflects performance over a range of typical laboratory conditions but does not become dependent on performance of the methods in one particular analytical run (10, 50). Reagent lots should not be changed.

Assessing trueness for multiplex assays has some unique challenges. Several comparative methods may have to be used

if none of the methods in routine use cover all of the analytes in the multiplex assay (18). Another challenge is choosing the appropriate method for data analysis to establish trueness at the level of each individual analyte as well as for the overall multiplex system (18).

(i) Comparison-of-methods study. In a comparison of methods study, specimens are tested in parallel with both the new test and a valid comparative method (13). For the comparative method, the laboratory's existing method or a recognized reference method may be used. If the comparative method is a reference method, then theoretically the exact true value of the analyte is known. If the comparison method is not a reference method, then only an accepted reference value is known, which is considered the true value that can be determined in practice (13). Testing the samples in duplicate by both the comparative and test procedures is recommended so that the data generated represent well-characterized samples (10, 13). If the laboratory expects close agreement between the comparative and test methods, then each sample may be tested singly (13). Routine quality controls should be followed during the experiment, and any run with failed quality control samples should be repeated.

(ii) Recovery study. For new tests with no comparison method, a recovery experiment may be the only practical way to assess trueness (52, 77). Recovery studies test whether the assay can measure the analyte of interest when a known amount is present in the intended specimen matrix. Samples for recovery studies need to be well characterized and may include standards, quality control materials, proficiency testing materials, or patient specimens with known or consensus values. Samples may need to be constructed for testing by adding known amounts of analyte to negative patient specimens. The amount of analyte recovered is then compared to the amount added to the specimen. The difference between the average recovery and 100% recovery can then be used to judge acceptability. This type of experiment may also be helpful to provide an estimate of proportional error that can occur as the concentration of analyte increases (77).

Sample sources. The samples used in a comparison-of-methods study are generally residual patient specimens that have been previously tested by the existing method in the laboratory or in the reference laboratory used by the clinical laboratory. These specimens may be somewhat limiting in that they have no known value other than that obtained using an existing assay, which may, in itself, have weaknesses. Specimens to be evaluated should be representative of the population and clinical conditions expected in the future use of the test and should be reasonably distributed according to age and gender. Specimens should also be distributed over the clinically meaningful range of the test and should include both positive and negative specimens. When multiple genotypes exist for the analyte being evaluated, the performance of each genotype should be assessed in validation of trueness. Also, if multiple specimen types are accepted for the assay, each specimen type should be independently evaluated.

For rare targets or to evaluate samples of known value, samples may need to be obtained from sources external to the routine clinical laboratory. In these cases, it may be possible to obtain samples from other clinical laboratories or public health laboratories. Characterized samples may also be obtained from commercial sources and may include standards, quality control materials, reference panels, or proficiency testing samples. These samples are limiting in that they may not be representative of the laboratory's clinical population or of the typical prevalence and spectrum of the clinical condition of interest. In addition, matrix effects may lead to inaccurate conclusions (12, 50). Costs may also be significant (12). Because of these limitations, it is recommended that as many patient specimens as possible be obtained and supplemented with samples from other sources, if needed (11).

Specimens should be collected and handled according to accepted laboratory practice. Testing by the comparative method and the test method should occur within a time span consistent with the stability of the analyte. It is recommended to avoid storing specimens, if possible. If archived specimens are used, it is essential that they be stored under conditions that ensure their stability and represent the routine specimen handling and processing for the assay (10, 77).

Number of samples. The number of specimens to be tested for a method comparison study is determined by the laboratory performing the study. The appropriate number of specimens depends on many factors, including the precision and complexity of the assay, the prevalence of the target(s) in the indicated population, the established accuracy of the reference method, cost and feasibility, the scheme used for data analysis and the level of statistical confidence that the user is willing to accept (11, 18). It is recommended that no fewer than 20 and typically 40 to 50 or more specimens be tested (10, 52). A minimum of 100 specimens is suggested for manufacturers (10). For qualitative assays, 50 positive and 50 negative specimens are suggested for initial study (11). If the resultant confidence interval is unacceptably wide, more specimens can be tested to obtain a narrower confidence interval. For laboratory-developed molecular tests, 50 positive and 100 negative specimens have been recommended (52).

Studies to evaluate trueness need to be comprehensive enough to describe performance of the assay in detail. It may not be possible to accurately establish the trueness of a test unless a large number of samples are evaluated (55). It is most important that the concentration range represents the variety of diseases and medical conditions relevant for the assay. It is recommended that samples be distributed with one-third in the low to low-normal range, one-third in the normal range, and one-third in the high abnormal range (77). A larger number of samples will improve confidence in the statistical estimates and will allow the potential influence of random errors and/or bias related to performance of the test in subgroups within a population to be revealed (10). It is acknowledged that it may be difficult for laboratories to obtain an ideal number of positive samples for rare pathogens that are observed with low frequency in the indicated population. For these rare pathogens, it may be practical to test only a small number of samples (18).

Data analysis and criteria for acceptance. Molecular methods often prove to be more sensitive than current "gold standard" methods. Assessment of trueness is challenging when the new method has lower detection limits than the old method (73). It is important to use well-characterized specimens or reference material with known target values. International reference standards produced by the World Health Organization (WHO) are currently available for only a few analytes (e.g.,

HIV, hepatitis A virus [HAV], hepatitis B virus [HBV], hepatitis C virus [HCV], and parvovirus B19), with others (cytomegalovirus [CMV] and Epstein-Barr virus [EBV]) being in development. Test and reference methods may yield different values if nominal standards that have been quantified by alternate methods are used in trueness experiments (73). Methods to resolve discrepant results (e.g., sending specimens to another laboratory for testing or using clinical data) need to be established prior to conducting trueness experiments. Sending specimens to another laboratory, particularly for quantitative assays with no international standards, may be problematic if different methods, targets, primers, calibrators, etc., are used. To ensure objectivity, acceptability criteria also need to be decided upon in advance.

(i) Assays with a gold standard. Data obtained in studies that compare a new assay to an established assay or gold standard must be analyzed to see whether the assays agree sufficiently for the new assay to replace the old assay. In method comparison analysis, assays may be found to be equivalent, commutable, or incompatible. Methods are equivalent if they yield identical results for all individuals. Equivalent methods can be interchanged without loss of analytical accuracy. It is unlikely, however, that different assays will be exactly equivalent. Therefore, the best criterion for determining whether two assays can be interchanged is commutability and not equivalence. The magnitude of quantitative differences should be determined to establish whether two assays are commutable. If the difference is not enough to be clinically important, the new assay is considered to be within the medical tolerance interval and can replace the old assay. Commutable assays can be interchanged without loss of diagnostic power for patients. Methods that are incompatible have differences that are greater than the medical tolerance interval and cannot be interchanged.

It is useful to interpret the data graphically as well as statistically. Current guidelines recommend the use of both a scatter diagram with correlation and regression analysis and a difference plot with calculation of the 95% limits of agreement for evaluation of method comparison data (10, 50, 56). The scatter diagram will reveal constant or proportional bias, and the difference plot will reveal whether the bias is constant over the whole range of values or whether the bias is influenced by between-method differences (Fig. 2).

(*a*) *Scatter diagram*. Scatter diagrams are constructed by plotting the mean of duplicates of the test method on the *y* axis versus the corresponding mean of duplicates of the comparative method on the *x* axis. When comparing two methods that measure the same characteristic, the plot should be constructed so that the origins and scales of both axes are identical. Visual examination of the pattern made by the points on the scatter diagram gives an initial impression of the basic nature and strength of the relationship between the two variables. If the variables are correlated, the points will fall along a line or curve. The better the correlation, the closer the points will be to the line. The range of linear agreement and any obvious outliers can also be visually identified. If an outlier is found, every effort should be made to determine the cause, since outliers may have a significant impact on the conclusions derived from the analysis. A single run that has an outlier may be replaced with another run. If outliers are observed in more than one run, possible sources of error should be investigated.

Scatter diagram data should be further analyzed using regression analysis to obtain the equation of the line that fits the data most closely. Regression analysis can also give an assessment of adequate range and relative scatter of the data. The endpoint for examination of data from trueness experiments is an estimation of bias between the test and comparison methods (10). Although it is useful for visual examination, it is not necessary to plot the data in order to calculate the equation of the regression line. Regression analysis can be done using calculator statistical tools, a general statistics computer program, a specialized method validation program, or an electronic spreadsheet. Regression analysis generally includes determination of the slope, *y* intercept, and correlation coefficient (*r*). A variety of linear regression models are available. Ordinary least-squares regression is often used, including computation of limits on the slope and *y* intercept. The least-squares approach is generally valid as long as the line is well defined (50). An argument may be that least-squares linear regression considers the error of the test method only. Weighted Deming regression or Passing-Bablock nonparametric regression models take into account variability in both the test and comparison methods and may be used to estimate the slope and *y* intercept, but they should not be used for calculating the standard error of the estimate because it may be artificially low (10). If the data cover the whole analytical range of the assay, differences between estimates of regression parameters using different models are generally insignificant compared to tolerance limits. The slope is the angle of the line that fits the data most closely. The *y* intercept describes the point where the regression line crosses the *y* axis. Regression line characteristics of comparative data that are in perfect agreement are slope = 1.0 and *y* intercept = 0. A regression line with a slope that is not statistically different from 1.0 but with a *y* intercept not equal to zero indicates constant, or systematic, bias. Proportional bias is indicated by a regression line with a slope that is statistically different from 1.0, regardless of the *y*-intercept value. Correlation coefficients measure the strength of the relationship of the observed data between the two methods. While correlation coefficients have values of between −1.00 and 1.00, values between 0 and 1 are most relevant for microbiologic assays. A correlation coefficient of +1.00 indicates perfect correlation. Values less than 1.00 indicate that there is scatter in the data around the regression line. The lower the value, the more scatter there is in the data. Since correlation coefficients measure the scatter about the line and do not assess the agreement between two variables, correlation coefficients should never be used as an indicator of method acceptability. High correlation does not necessarily mean that there is good agreement between the two assays being compared. For example, if one method gives values that are twice those obtained by the other method, a plot of the data will give a perfect straight line with a slope of 2.0. The correlation coefficient will be 1.0, but the two measurements do not agree (16). Correlation coefficients are affected by the range of samples studied. Two assays that are designed to measure the same parameter will have good correlation when the set of samples is chosen so that they are widely distributed in the reportable range of the assay.

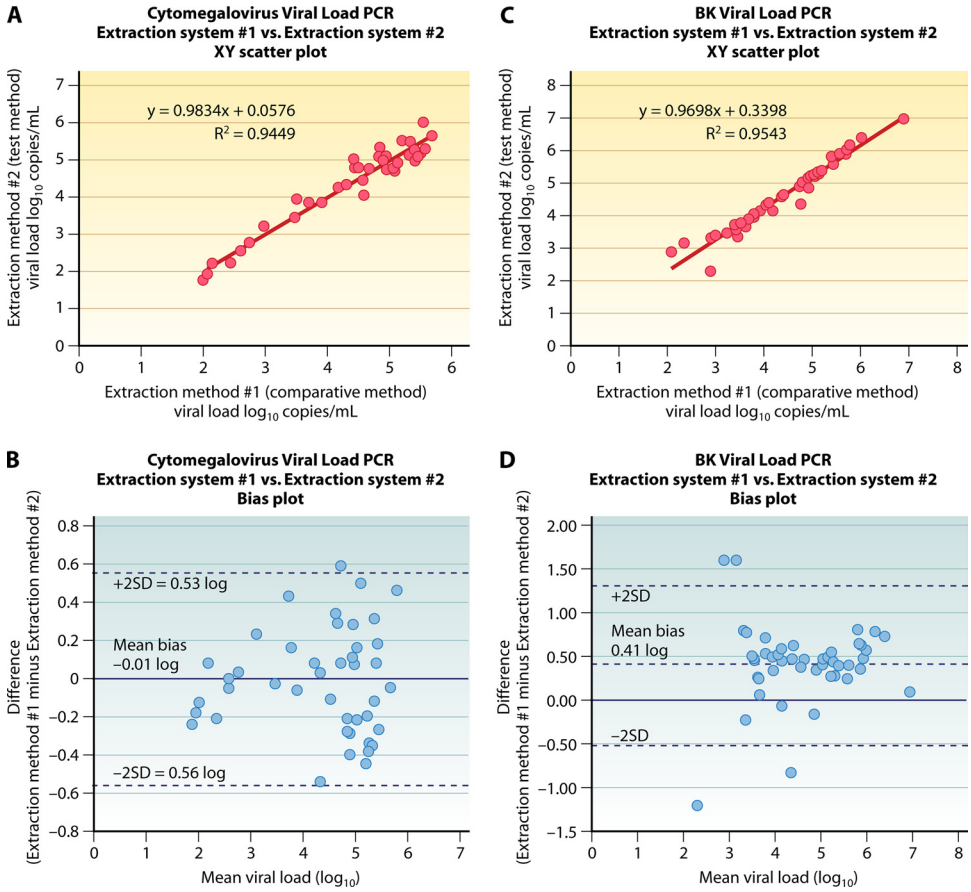

FIG. 2. Bland-Altman bias plots and *xy* scatter plots for two different quantitative real-time PCR assays for which a new extraction method was being evaluated. (Data courtesy of Charles Hill, Emory University, Atlanta, GA; used with permission.) (A) Seventy-two patient specimens were tested using both a new (test) and old (comparison) extraction method for a cytomegalovirus viral load assay. Forty-three specimens that had numerical results by both methods are plotted. Visual inspection of the *xy* scatter plot showed no obvious outliers. Linear regression analysis gives the equation $y = 0.9834x + 0.0576$ with a correlation coefficient (r^2) of 0.9449 to describe the line that fits the data most closely. The correlation coefficient of 0.9449 indicates good correlation. The slope of 0.9834 is very close to 1.00, indicating no proportional bias, and the *y* intercept at 0.0576 is very close to the origin (0.00), indicating no constant systematic bias. (B) Bland-Altman bias plot of the data in panel A. The mean bias was determined to be −0.01 log unit, which indicates no systematic bias. The plot shows that bias is greater at higher viral loads, but the bias is in both directions and within acceptable limits. Both extraction methods were determined to be equivalent, and the new extraction method could be used without need to rebaseline patients. (C) In this comparison study, 46 patient specimens positive for BK virus in a real-time quantitative PCR assay were tested using both a new extraction method and an old extraction method. No obvious outliers are seen in the *xy* scatter plot. Linear regression analysis gives the equation $y = 0.9698x + 0.3398$ with a correlation coefficient of 0.9543. The correlation coefficient of 0.9543 indicates good correlation, and the slope of 0.9698 is very close to 1.00, indicating no proportional bias. However, the *y* intercept at 0.3398 is significantly away from the origin (0.00), indicating the presence of some constant systematic bias. (D) Bland-Altman bias plot showing that the mean bias was determined to be 0.41 log unit, indicating a systematic bias of about 2.5-fold. The bias is not considered to be statistically significant since the 95% confidence interval contains zero (no difference), but it needs to be decided if the bias is clinically significant before the new extraction system is put into use for clinical testing.

Since the correlation coefficient is affected by the range of values, the only appropriate use of correlation coefficients in method comparison studies is to determine whether the sample data are in an adequate range (10, 77). A correlation coefficient of greater than 0.975 (or equivalently, $r^2 \geq 0.95$) is generally accepted for analytical chemistry comparisons to indicate that the range of data is sufficiently wide (10). Acceptable correlation coefficients have not been defined specifically for molecular assays.

(*b*) *Difference plot*. Difference plots can be constructed in a variety of forms, each of which has advantages and disadvantages. A widely used method for appropriately assessing and measuring agreement is the mean-difference plot described by Bland and Altman (3, 39). The main objective of the Bland-Altman approach is to compare the experimentally observed deviations with a preset clinical acceptance limit. Software providing Bland-Altman plots, such as the Analyze-it statistical analysis software addendum for Microsoft Excel (Analyze-it Software Ltd., Leeds, United Kingdom) and MedCalc statistical software (MedCalc Software, Mariakerke, Belgium), as well as others, is commercially available. A Bland-Altman plot is identical to a Tukey mean-difference plot and is a method of plotting data to analyze the agreement between two different assays.

To construct a standard Bland-Altman plot, both the new and old assays are performed on the same sample. Each of the

samples is represented on the graph by plotting the mean of the measurements on the *x* axis and the difference of the two values on the *y* axis. In method comparison studies that cover a wide concentration range, as is often the case for infectious disease molecular tests, log transformation of the data points is recommended. If the *x* axis data cover a narrower concentration range, it may be more useful to plot percent difference as the *y* axis value. Graphing the data in this way allows analysis of the relationship between the differences and the averages of the measurements and will reveal any bias. Analysis of the data will also reveal any variation in differences over the range of measurement and can identify possible outliers. In Bland-Altman analysis, it is common to summarize agreement by calculating the bias as well as estimating the mean difference and the standard deviation of the differences. It is also common to determine the limits of agreement, which are by convention set at the 95% confidence interval of the difference between the methods. This is usually specified as bias ± 1.96SD (average difference ± 1.96 standard deviations of the difference). How far apart measurements can be without causing difficulties will be a question of judgment and ideally should be defined in advance. If the 95% confidence interval for the mean difference includes zero, there is statistically no evidence of bias, but when establishing acceptability criteria, clinical significance as well as statistical significance must be considered (50). If the bias is not clinically important, the two methods may be used interchangeably. If the bias exceeds an acceptable limit and the new assay is a modification of an assay currently in use, the reference interval should be reviewed and clinicians notified that results using the new assay may be different from those previously issued (50). A potential limitation of mean difference analysis is that the error pertains exclusively to the new assay and not to the comparison assay, which may be an acceptable standard for comparative studies but, in itself, is also subject to error. If nonequivalence is observed but the test method is believed to be better than the comparison method, the test method does not necessarily have to be rejected, but clinical data need to be obtained before the assay is put into routine use (10).

(ii) Qualitative or quantitative assays without a suitable comparator. For analytes that do not have an independent test to measure, the gold standard may be a clinical diagnosis determined by definitive clinical methods (10). The results of such a comparison study are often presented in the form of a two-by-two table (56). The usual measure of agreement is the overall fraction or percentage of subjects that have the same test result (i.e., both positive or both negative). Because agreement may depend on the disease prevalence in the population studies, it may be useful to separate the overall agreement into agreement concerning positive and negative results. One concern with these simple agreement measures is that they do not take agreement by chance into account (56). Verifying that the observed agreement exceeds chance levels for categorical data is typically done using kappa statistics. Kappa statistics assess the proportion of times that two or more raters (or in this case, laboratory assays) examining the same data (or specimen) agree about assigning the result to categories other than by chance alone. Kappa statistics are easily calculated, and software is readily available in many standard statistical packages (e.g., SAS [SAS Institute, Inc., Cary, NC]). A kappa value of 1.00 indicates perfect agreement, a kappa value of 0.00 indicates no agreement above that expected by chance, and a kappa value of −1.00 indicates complete disagreement. Not everyone agrees as to what constitutes good agreement, but generally kappa statistics above 0.80 are considered almost perfect.

Reference Interval

The reference interval is usually the last performance characteristic to be studied, since it is not used to decide whether a method is acceptable or not (77). The reference interval ("normal range") of a test is simply defined as the range of values typically found in individuals who do not have the disease or condition that is being assayed by the test (the "normal" population) (6). Defining the reference interval for a test gives clinicians practical information about what is "normal" and "abnormal" that can be used to guide management of a patient. To be clinically useful, the reference interval must be appropriate for the population being served.

If a nucleic acid target is always absent in a healthy individual and the test is a qualitative test, the reference range is typically "negative" or "not detected." In this case, evaluation may not be necessary, and the reference interval study plan may be to state that, based on literature review or other pertinent information, no reference interval study will be performed.

For quantitative assays, the reference interval will be reported as below a particular quantitative measurement, usually either the limit of detection or the limit of quantification. For some analytes, the reference interval may be different from the limit of detection or quantification, and a clinical decision limit may be used instead (6). This is especially true for latent viruses such as cytomegalovirus (CMV), where a cutoff that distinguishes asymptomatic infection from active CMV disease needs to be established (42).

For some infectious disease molecular tests, such as HCV genotyping, the intended use of the test stipulates that the patient should be known to be positive for the analyte being assayed. A reference interval may not be applicable in these situations.

Reference interval study. If it is decided that a reference interval needs to be determined, a reference interval study is performed. Experiments required to establish the reference interval for a particular test may be different from experiments to determine the limit of detection or limit of quantification. To determine the reference interval, specimens from healthy subjects in the intended population are tested and the resultant values are used to determine the normal range (6). Reference interval studies are often done using residual specimens from tests done for other purposes as long as the specimens are representative of the population being served. An appropriate number of specimens must be tested, taking into consideration the desired confidence limits. To establish a reference interval, it is recommended that 120 samples be tested (6). Samples with known interfering substances should be excluded. Criteria should be established for excluding samples for other reasons. Samples should be preserved and tested according to routine practice for patient specimens. The distribution of resulting data should be evaluated and possible outliers identified. A simple nonparametric method is sufficient to analyze the data

in most situations (6). If fewer than 120 samples are tested, more sophisticated methods of estimation using traditional parametric methods, bootstrap-based procedures, or Horn and Pesce methods may be used (6). The reference interval is usually computed as the interval between the lower reference limit (the 2.5th percentile) and the upper reference limit (the 97.5th percentile). In most infectious disease molecular assays, usually only the upper reference limit is of medical importance (6).

If reference intervals of the assay being used have already been determined based on an adequate reference interval study, it may be possible to simply "transfer" the reference interval (6, 77). The laboratory director can make this assessment after documentation of careful review of information from the original study. This type of transfer can be done if the manufacturer provides adequate demographic information for the population studied when the original reference interval was determined and it is determined that the reference interval is applicable to the population served by the clinical laboratory. Even if details of an adequate reference interval study are provided, the laboratory may prefer to experimentally verify the reference interval. In this case, it is recommended that 20 specimens from individuals who represent the laboratory's reference population be analyzed. If two or fewer results fall outside the stated reference interval, the reference interval is considered verified (77). If there are concerns that reference interval information from the manufacturer is not adequate, if the new test is being applied to a different population, or if the new test is based on a different measurement principle, then it is recommended to verify the claimed reference interval by testing 60 (minimum, 40) specimens. The results are evaluated by statistically estimating the reference interval and comparing it to the claimed reference interval (77).

Revalidation

Validation of an analytical method is a one-time process unless the conditions under which the method was developed have changed. Revalidation is required if the existing method is modified (e.g., by addition of new sample type or changes in a critical component or reagent that may affect the assay) (65).

CONTROLS

CLIA regulations state that laboratories must determine and document calibrator and control procedures based on assay performance specifications as applicable (30). CLIA allows laboratories to establish their own protocols for control samples as long as for each run (defined as up to 24 h of stable operation) there are the following (30) (Table 5): (i) amplification controls (for qualitative assays, positive and negative; for quantitative assays, negative and at least two controls of different values [generally low positive and high positive]), (ii) extraction controls for assays with an extraction phase (i.e., a control that is capable of detecting errors in the extraction process), and (iii) inhibition controls if reaction inhibition is a significant source of false-negative results (i.e., a control capable of detecting the inhibition).

TABLE 5. Quality control schedule for infectious disease molecular assays

Test type	Procedure
Qualitative	Controls[a]
	Amplification controls (positive for each anylate, negative, sensitivity, every run, establish target/range values and monitor)
	Internal control (or demonstrate no inhibition)
	Extraction control
	Calibration verification (verify cutoff)[b]
Quantitative	Controls[a]
	Amplification controls (at least 2 levels of positive control at relevant decision points to verify that calibration status is maintained, negative, every run, establish target/range values and monitor)
	Internal control (or demonstrate no inhibition)
	Extraction control
	Calibration verification (verify cutoff)[b]
	Analytical measurement range validation[b]

[a] A positive control that is taken through the extraction process may dually serve as both an extraction control and an amplification control. For multiplex assays, a pooled control that contains all analytes can be used or individual controls can be rotated after lot/shipment validation of all targets.

[b] Every 6 months, after changes of major system components, after lot changes of all reagents, and after failure of quality control, major maintenance, etc., as appropriate.

Amplification Controls

Qualitative assays. A positive control and a negative control must be included in each molecular amplification run. Control materials may be obtained commercially, prepared in-house, or obtained from other sources. Positive-control material may be purified target nucleic acid, patient specimens containing the target nucleic acid, or controls produced by spiking the organism of interest, preferably inactivated, into specimens known not to contain the target. Calibration materials should not generally be used as controls, although CLIA allows the use of calibrators as controls for assays for which control materials are not available (30). If calibrators are used as controls, a different lot number should be used (30). Nuclease-free water is often used as a diluent; however, it is preferred that quality control samples have a matrix that matches that of the specimens in the analytical run as closely as possible. This is important because the matrix confers turbidity and surface tension that can affect pipetting. There also may be interactions between the analyte and the proteins in the sample matrix that may affect the detectable concentration of the analyte. Ideally, a control matrix is tested and found to be negative for HIV, HCV, and HBV.

The best practice is to construct the positive control so that it is at a concentration near the lower limit of detection of the assay. The concentration should be just high enough to provide

consistent positive results but low enough to challenge the detection system near the limit of detection (e.g., 5 or 10 times the LOD) (15). Frequent analysis of a positive control near the limit of detection maximizes the opportunity for the laboratory to detect problems with the test system.

For multiplex systems, positive controls for each analyte should be included in each run or rotated so that all analytes are tested periodically (33).

A blank nontemplate control such as water or buffer is often used in molecular assays as a form of negative control. Blank controls can rule out contamination of reagents with target nucleic acid and are often interspersed throughout the run (15, 40). Blank controls can also be used to compensate for background signal generated by the reagents (15, 40). Blank controls should be taken through the extraction process and should contain all of the reaction reagents (15).

An optimal negative control, however, is a specimen containing known nontarget nucleic acid (15, 40). Nontarget nucleic acid negative controls rule out contamination of reagents with target nucleic acid but also rule out nonspecific PCR amplification or nonspecific detection of amplified product. Patient specimens from noninfected individuals or specimens containing known nontarget organisms or nucleic acids are often used as negative controls. Negative controls can also be made by adding nontarget organisms or nucleic acids to negative specimens (15). Negative controls should be taken through the entire assay, beginning with extraction.

Preparing a 1- to 2-year supply of control material provides long-term stability in the practice of quality control and is recommended whenever possible (3). In controls for molecular tests, the target nucleic acid may degrade over time, whether the control is a patient specimen or created artificially (15). The laboratory must document the stability of control material made in the laboratory. This can be ongoing based on expected performance. Dilute nucleic acids are particularly prone to degradation and may need to be replaced more frequently. Control samples should be stored in tightly capped polyallomer or specially designed polypropylene tubes that do not bind DNA (17). Stability at room and refrigeration temperatures for some samples may be improved by using Tris-EDTA buffer. Control samples in which the target is DNA and may contain DNases should ideally be stored in a nondefrosting freezer at −20°C or colder (17). Similarly, controls that may contain RNases are best stored at −70°C (17).

Quantitative assays. For a quantitative assay, a negative control and two positive controls containing different concentrations of target nucleic acid are required in every run. The concentrations should be chosen to verify assay performance at relevant analytic and clinical decision points (3). This usually means including a high-positive control near the upper limit of the reportable range as well as a low-positive control near the lower limit of detection as discussed above for qualitative assays (16).

Extraction Controls

Nucleic acids are prepared from specimens for molecular testing using any of a variety of manual or automated methods, including preparation of crude lysates, sequence capture to remove the target from specimen matrix, and lysis/extraction procedures. CLIA requires that assays that include a nucleic acid extraction procedure include a control that is capable of detecting errors in the extraction process. CAP more broadly addresses specimen processing for molecular assays and requires that all nucleic acid isolation/preparation processes, not just extraction, be evaluated (33). To properly test for effective nucleic acid isolation, preparation, or extraction of a bacterial or viral target, it is best to use whole bacteria or virus as it would be present in a patient specimen as the control material. If whole organism is not available, purified nucleic acid can be used. To prepare the control, the organism or nucleic acid should be seeded into the appropriate matrix at a low level and run in parallel with patient specimens.

If the positive amplification control is taken through all steps of the assay, it can dually serve as an amplification control and as an extraction control.

Internal Controls

In assays in which reaction inhibition is a significant source of false-negative results, CLIA requires that indicators of inhibition be used. Inhibition in molecular assays can result from alterations in pH, ion concentration, or viscosity or direct inhibition of the polymerase enzyme (2). Inhibitory substances can be endogenous to the specimen or introduced exogenously into the assay and are preliminarily assessed as part of analytical validation of the assay as discussed above. Continual monitoring of test result trends will also reveal problems due to inhibition. If a laboratory gathers sufficient data (100 to 500 specimens) and inhibition rates are found to be within acceptable limits considering the medical implications of a false-negative result, testing for inhibition may be relaxed or discontinued (33, 34).

Inhibitors of amplification can be detected by using an internal control. The rationale is that if a specimen does not allow amplification of an internal control, the amplification of an intended target sequence may also be inhibited. Internal controls cannot differentiate between inhibition and amplification failure due to any number of variables, such as thermocycler well failure, failure to add enzyme or other reagents, etc. Various constructs of internal controls are used in molecular assays (15). Internal controls can be homologous extrinsic, heterologous extrinsic, or heterologous intrinsic. To prevent competition and avoid adverse reductions in sensitivity, the internal control should be set at the lowest concentration that permits consistent detection of the control (15).

Homologous extrinsic controls. If unmodified intended target is used as the internal control, a small amount of intended target is added to a second aliquot of specimen and run in parallel with the original aliquot. Unmodified-target controls are easy to use but have the disadvantage of the cost and space occupied by a second reaction. This type of internal control can detect the influence of chemical constituents and cycling profiles of the reaction but does not control for PCR efficiency or for the loss of DNA/RNA during extraction in the original aliquot from which that patient result is generated (2).

Modified-target controls are constructed so that they can be amplified with the same primers that amplify the intended target, but they contain non-target-derived sequence inserts that are distinguished from the intended target by size (usually

100 to 200 bp longer) or presence of unique internal sequences (2, 15). Modified-target controls can be coamplified with the intended target in the same reaction without the need to run a second "spiked" aliquot. If an unacceptable reduction in sensitivity occurs when the reaction is carried out in a single reaction vessel, there may be competition for PCR reagents and the control reaction may need to be performed with a second specimen aliquot. Unmodified- or modified-target controls can be added to the sample either before or after the sample is prepared. If added before sample preparation, the internal control can also serve as an extraction control (15).

Heterologous extrinsic controls. Heterologous extrinsic controls are non-target-derived controls that require primers and probes different from the target. As with homologous extrinsic controls, heterologous extrinsic controls can be added to the sample either before or after the sample is prepared and can also serve as an extraction control if added before sample preparation (15). Heterologous extrinsic controls must also not be competitive for PCR reagents.

Heterologous intrinsic controls. Heterologous intrinsic controls are often referred to as "housekeeping genes" and are conserved fragments of the host's genome that are present naturally in patient specimens in low copy number. Heterologous intrinsic controls are amplified with a different set of primers and can be amplified in the same or a separate reaction vessel. Commonly used intrinsic controls include the genes encoding β-globin, β-actin, gamma interferon, glyceraldehyde-3-phosphate dehydrogenase (GAPDH), or U1 small nuclear ribonucleoprotein (snRNP)-specific A protein. Depending on the marker chosen and the specimen type, intrinsic controls can be used to establish the presence of cellular material in a clinical specimen. A concern when using intrinsic controls is that the number of human gene copies may be much higher than the target infectious organism copy number and thus have an amplification advantage and not accurately test for inhibition.

Regardless of the type of control used, the control target sequence must be detected for a negative result to be considered valid. In specimens that test positive for the intended target, the internal control may be outcompeted by the target and therefore not detected. In most cases, a reasonably strong positive result with a negative internal control is considered valid (15).

Frequency of Controls

CLIA regulations indicate that controls must be run as specified by the manufacturer or established by the laboratory with a minimum frequency of at least each day (30). In addition, controls must also be run to qualify each shipment or lot-to-lot change of reagents and each time there is major preventive maintenance or replacement of a critical part that may influence test performance. Control results must meet expected results before patient testing is continued.

For those assays that contain electronic/procedural/built-in internal controls and are FDA approved and not modified by the laboratory, external controls may be run with each new lot or shipment only and do not need to be run daily long as the assay validation protocol has documented the adequacy of control frequency (33). The laboratory director has the discretion to run external controls more frequently for these assays, but it is not required.

TABLE 6. Estimating limits for controls[a]

Time point	Avg C_T value	SD
Initial	38.06	2.15
Month 1	38.16	2.60
Month 2	37.62	0.75
Month 3	37.75	1.03
Avg	37.90	1.60

[a] One approach for establishing the target value and limits for controls involves determining an initial temporary target value by assaying the control material a minimum of 20 times. The temporary value is used for the next 3 to 6 months (or for x number of runs, e.g., 20 to 30), during which time testing is presumed to be stable. The final target value and standard deviation are determined from the consecutive monthly values. The data presented are for a low-positive control in a real-time PCR assay to detect herpes simplex virus type 1 (HSV-1). The C_T value refers to the number of cycles required for the fluorescence signal to cross threshold. The final target value (C_T) ± SD is 37.90 ± 1.60.

Location in Run

To best serve as an indicator of the cumulative effects of handling during the assay process, negative controls should be placed at the end of a run as the last sample to which reagents are added (5, 15). If negative controls are placed at the beginning of a run, they may generate falsely low assessments of contamination. Especially in very large batch runs, it may be beneficial to space negative controls evenly throughout the batch to monitor drift or distribute them randomly in the run to detect random errors (5). For many FDA-cleared/approved assays, the position controls in a run is established by the manufacturer and cannot be changed by the user.

Statistical Parameters

Tolerance and acceptability limits must be defined and monitored for all control procedures (30, 33, 34). This involves determining the target value and then setting limits based on the analytical variation of the assay as well as the implications of rejecting a test on that basis (5). Limits should be set to accurately detect random and systematic analytical errors accompanied by an appropriately low false rejection rate. Many different decision criteria can be used to set control limits, but the most common approach is to set control limits based on multiples of the standard deviation on both sides of the observed mean. This involves obtaining repeated measurements on the control samples by the methods used in the laboratory. If there is no history of quality control data, as for newly developed assays or new lots of control material, the following approach is recommended to estimate the limits for the control (5) (Table 6). (i) Determine a temporary target value by assaying the control material using a minimum of 20 separate determinations, ideally on 20 separate days. If fewer days are necessary, no more than four control measurements per day for at least five different days are recommended. (ii) Calculate the mean and standard deviation. Set the range of allowable control values around the temporary target using the standard deviation multiplied by the laboratory's control limit. (iii) Use this value for the next 3 to 6 months or x number of runs (e.g.,

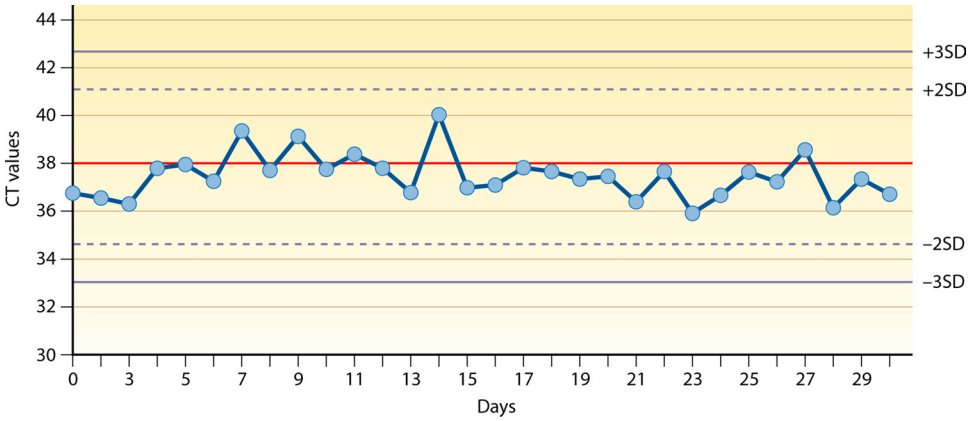

FIG. 3. Continuous monitoring of controls using Levey-Jennings plots. The low-positive control presented in Table 6 was monitored on each day of use using a Levey-Jennings plot. The plot shows the target value (threshold cycle value [C_T] = 37.90, solid red line) as well as expected limits (hatched lines) (two standard deviations = 34.7 to 41.1; three standard deviations = 33.1 to 42.7). All control results are within two standard deviations of the mean target value and are therefore in range using Westgard's multirule analysis.

20 to 30 runs), during which time the testing is assumed to be stable. (iv) Calculate the final target and usual standard deviation from the additional values.

Visual analysis of control data is often helpful for general observation of trends and shifts. The most common visual analysis is the Levey-Jennings plot, in which the established target value and allowable error are plotted on the *y* axis and the days of the month on the *x* axis (Fig. 3). Some laboratory information systems have quality control modules that will allow preparation of Levey-Jennings plots. Levey-Jennings plots use rejection criteria that involve taking action when a single result is beyond the control limit determined by the laboratory, generally two or three standard deviations in either direction from the mean. When control performance is based on ±2 standard deviations, it can be expected that there will be a run failure rate of 5%.

Westgard's multirule analysis uses multiple control rules to judge the acceptability of an analytical run and generally improves detection of quality control failure with concurrent low probability of falsely rejecting a run (76). Multirule analysis is recommended, especially when there are two or more control measurements per run. Multirule analysis involves preparation of standard Levey-Jennings plots designating the mean plus or minus one, two, and three standard deviations. When a single control measurement exceeds two standard deviations from the mean, the control data are inspected using the additional rules. If a single point is beyond three standard deviations, the run is rejected. If results from two consecutive control samples across runs exceed two standard deviations in the same direction, the run is rejected. If one point is greater than plus two standard deviations and another is greater than minus two standard deviations, the run is rejected. Many other rejection rules for assays that use more complex control schemes have been described (76).

CLIA regulations require review of quality control results using the rules established by the laboratory before patient results are reported and documentation of corrective actions when controls exceed action limits (31). For those laboratories inspected by CAP, quality control data must be reviewed at least monthly by the laboratory director or designee (33, 34).

Suggested sequential steps that might be taken when controls exceed action limits include (34) the following: (i) use a fresh aliquot of control, (ii) use a separate or newly prepared control, (iii) look for other obvious problems (reagent levels, mechanical fault, etc.), (iv) use new reagents (one or all), (v) perform maintenance on the instrument, and (vi) recalibrate the instrument. After each step is taken, controls are reanalyzed, and if results are now within acceptable limits, it is assumed that the problem has been resolved.

CALIBRATION VERIFICATION

CLIA regulations require laboratories to determine calibration procedures for each assay used in that laboratory (28). Calibration is defined as the process of testing and adjusting an instrument, kit, or test system readout to establish a correlation between the instrument's measurement of the substance being tested and the actual concentration or amount of the substance (27). Once calibrated, all assays need to be able to "hold calibration" and stay within tolerated limits over a reasonable period of time. Calibration is verified as stable by the process of calibration verification.

Calibration verification is defined as "the assaying of calibration materials in the same manner as patient specimens to confirm that the calibration of the instrument, kit, or test system has remained stable throughout the laboratory's reportable range for patient test results" (27). In calibration verification, known quantities of a material are measured by the assay and results are compared to the known true value of the material (34). The purpose of calibration verification is to ensure that the reported values are set accurately. According to the CAP Laboratory Accreditation Program, calibration verification, as used in CLIA regulations, actually includes two different processes: verification of previously established calibration status (i.e., set point) and verification of the analytical measurement range (AMR). To satisfy the CLIA requirement, CAP requires both calibration verification (as defined by CAP) and validation of the AMR (29, 33, 34).

Calibration verification, as considered in the CAP checklist as a single process, is used to check the correlation between the

instrument's measurement of the substance being tested and the actual concentration of the substance. Calibration verification, when used in this way, verifies only the set point of the test system at the declared cutoff value. For qualitative tests that establish a cutoff value to distinguish positive from negative, the cutoff value is verified.

CLIA requires clinical laboratories to verify calibration once every 6 months. Calibration verification also needs to be performed before resumption of patient testing and reporting results whenever any of the following occur (Table 5) (29). (i) All reagents used for a test procedure are changed to new lot numbers, unless the laboratory can demonstrate that changing reagent lot numbers does not adversely affect the linear range or control values. (ii) There is a major preventive instrument maintenance or replacement of a critical part(s) that may influence the performance of the assay. This includes sending an instrument to the manufacturer for repair. (iii) There is a change in major assay components. (iv) Control values are outside acceptable limits or show an unusual trend or shift, and the usual means used to assess and correct unacceptable control values fail to identify and correct the problem. (v) The laboratory has determined that the test system's reportable range for patient test results should be checked more frequently.

Materials used to verify calibration must have assigned concentration values. Appropriate materials include proficiency testing samples or patient specimens with known values or commercially available standards, calibrators, or reference materials with appropriate matrix characteristics and known target values.

Calibration verification should be documented each time it is performed. If calibration verification results are unacceptable, the calibration procedure for the assay must be repeated. After repeating the calibration procedure, calibration verification must be performed again, and acceptable results must be obtained before patient testing is resumed.

AMR VALIDATION

Quantitative molecular assays are generally performed using three to five standards of known concentration of target nucleic acid to create a standard curve. The amount of target nucleic acid in a test specimen is determined by comparing the result of the specimen with the standard curve. Analytical measurement range (AMR) is the term used by CAP to refer to the range of values that a quantitative method can measure for a specimen without dilution, concentration, or other pretreatment of the specimen prior to testing it. AMR is the same as reportable range in CLIA terminology. The AMR is established by the laboratory for in-house-developed assays using linearity studies as described above. Validation of the AMR corresponds to the CLIA requirement for validation of the reportable range. The purpose of AMR validation is to ensure that the test system is providing accurate results throughout the measurement range.

Materials used for AMR validation should have a matrix that matches the clinical specimens assayed by that method as closely as possible. Materials for validation may include samples used for linearity studies; proficiency testing survey samples; previously tested unaltered patient specimens; previously tested patient specimens altered by admixture, dilution, or spiking negative specimens with known amounts of analyte; standards or reference materials with appropriate matrix characteristics and target values; and calibrators or control materials if they adequately span the AMR (33, 34).

CLIA regulations require validation of the AMR using at least three different concentrations (a minimal value, a midrange value, and a maximum value near the upper limit of the range) of appropriate material within the established measurement limits of the assay (26). Materials for validation are run as patient specimens would be.

AMR validation must be performed at least once every 6 months, whenever there is a change in major system components or reagent lot changes (unless the laboratory can demonstrate no effect of lot number changes), and at other times as appropriate (e.g., after major instrument service, failure of quality control, etc.) (Table 5).

AMR validation should be documented each time it is performed. The laboratory must define acceptance limits for the difference between the measured values and the actual concentrations of the material. Acceptance limits are often defined by the slope of the line obtained in the AMR validation process. If results are unacceptable, corrective action must be taken before patient testing is resumed.

REFERENCES

1. **Antwerpen, M. H., P. Zimmermann, K. Bewley, D. Frangoulidis, and H. Meyer.** 2008. Real-time PCR system targeting a chromosomal marker specific for *Bacillus anthracis*. Mol. Cell Probes **22:**313–315.
2. **Ballagi-Pordany, A., and S. Belak.** 1996. The use of mimics as internal standards to avoid false negatives in diagnostic PCR. Mol. Cell Probes **10:**159–164.
3. **Bland, J. M., and D. G. Altman.** 1986. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet **i:**307–310.
4. **Bolotin, S., C. De Lima, K. W. Choi, E. Lombos, L. Burton, T. Mazzulli, and S. J. Drews.** 2009. Validation of the TaqMan Influenza A Detection Kit and a rapid automated total nucleic acid extraction method to detect influenza A virus in nasopharyngeal specimens. Ann. Clin. Lab. Sci. **39:**155–159.

4a. **Cai, T., G. Lou, J. Yang, D. Xu, and Z. Meng.** 2008. Development and evaluation of real-time loop-mediated isothermal amplification for hepatitis B virus DNA quantification: a new tool for HBV management. J. Clin. Virol. **41:**270–276.

5. **CLSI/NCCLS.** 2006. Statistical quality control for quantitative measurement procedures: principles and definitions. Approved guideline, 3rd ed. CLSI document C24-A3. Clinical and Laboratory Standards Institute, Wayne, PA.
6. **CLSI/NCCLS.** 2008. Defining, establishing and verifying reference intervals in the clinical laboratory. Approved guideline, 3rd ed. CLSI document C28-A3. Clinical and Laboratory Standards Institute, Wayne, PA.
7. **CLSI/NCCLS.** 2004. Evaluation of precision performance of quantitative measurement methods. Approved guideline, 2nd ed. CLSI document EP5-A2. Clinical and Laboratory Standards Institute, Wayne, PA.
8. **CLSI/NCCLS.** 2003. Evaluation of the linearity of quantitative measurement procedures: a statistical approach. Approved guideline. CLSI document EP6-A. Clinical and Laboratory Standards Institute, Wayne, PA.
9. **CLSI/NCCLS.** 2005. Interference testing in clinical chemistry. Approved guideline. CLSI document EP7-A2. Clinical and Laboratory Standards Institute, Wayne, PA.
10. **CLSI/NCCLS.** 2002. Method comparison and bias estimation using patient samples. Approved guideline, 2nd ed. CLSI document EP9-A2. Clinical and Laboratory Standards Institute, Wayne, PA.
11. **CLSI/NCCLS.** 2008. User protocol for evaluation of qualitative test performance. Approved guideline, 2nd ed. CLSI document EP12-A2. Clinical and Laboratory Standards Institute, Wayne, PA.
12. **CLSI/NCCLS.** 2005. Evaluation of matrix effects. Approved guideline, 2nd ed. CLSI document EP14-A2. Clinical and Laboratory Standards Institute, Wayne, PA.
13. **CLSI/NCCLS.** 2005. User verification of performance for precision and trueness. Approved guideline, 2nd ed. CLSI document EP15-A2. Clinical and Laboratory Standards Institute, Wayne, PA.
14. **CLSI/NCCLS.** 2004. Protocols for determination for limits of detection and limits of quantitation. Approved guideline. CLSI document EP17-A. Clinical and Laboratory Standards Institute, Wayne, PA.
15. **CLSI/NCCLS.** 2006. Molecular diagnostic methods for infectious diseases.

Approved guideline. CLSI document MM3-A2. Clinical and Laboratory Standards Institute, Wayne, PA.
16. **CLSI/NCCLS.** 2003. Quantitative molecular methods for infectious diseases. Approved guideline. CLSI document MM06-A. Clinical and Laboratory Standards Institute, Wayne, PA.
17. **CLSI/NCCLS.** 2005. Collection, transport, preparation, and storage of specimens for molecular methods. Approved guideline. CLSI document MM13-A. Clinical and Laboratory Standards Institute, Wayne, PA.
18. **CLSI/NCCLS.** 2008. Verification and validation of multiplex nucleic acid assays. Approved guideline. CLSI document MM17-A. Clinical and Laboratory Standards Institute, Wayne, PA.
19. **Code of Federal Regulations.** 2010. Title 21. Food and drugs, vol. 8, chapter 1. Food and Drug Administration, Department of Health and Human Services, part 56. Institutional review boards. U.S. Government Printing Office, Washington, DC.
20. **Code of Federal Regulations.** 2010. Title 21. Food and drugs, vol. 8, chapter 1. Food and Drug Administration, Department of Health and Human Services, part 807. Establishment registration and device listing for manufacturers and initial importers of devices. U.S. Government Printing Office, Washington, DC.
21. **Code of Federal Regulations.** 2010. Title 21. Food and drugs, vol. 8, chapter 1. Food and Drug Administration, Department of Health and Human Services, part 809. In vitro diagnostic products for human use, section 809.10. Labeling for in vitro diagnostic products. U.S. Government Printing Office, Washington, DC.
22. **Code of Federal Regulations.** 2010. Title 21. Food and drugs, vol. 8, chapter 1. Food and Drug Administration, Department of Health and Human Services, part 809. In vitro diagnostic products for human use, section 809.20. General requirements for manufacturers and producers of in vitro diagnostic products. U.S. Government Printing Office, Washington, DC.
23. **Code of Federal Regulations.** 2010. Title 21. Food and drugs, vol. 8, chapter 1. Food and Drug Administration, Department of Health and Human Services, part 809. In vitro diagnostic products for human use, section 809.30. Restrictions on the sale, distribution and use of analyte specific reagents. U.S. Government Printing Office, Washington, DC.
24. **Code of Federal Regulations.** 2010. Title 21. Food and drugs, vol. 8, chapter 1. Food and Drug Administration, Department of Health and Human Services, part 812. Investigational device exemptions. U.S. Government Printing Office, Washington, DC.
25. **Code of Federal Regulations.** 2010. Title 21. Food and drugs, vol. 8, chapter 1. Food and Drug Administration, Department of Health and Human Services, part 820. Quality system regulation. U.S. Government Printing Office, Washington, DC.
26. **Code of Federal Regulations.** 2010. Title 21. Food and drugs, vol. 8, chapter 1. Food and Drug Administration, Department of Health and Human Services, part 864. Hematology and pathology devices, section 864.4020. Analyte specific reagents. U.S. Government Printing Office, Washington, DC.
27. **Code of Federal Regulations.** 2009. Title 42. Public health, vol. 4, chapter V. Health Care Financing Administration, Department of Health and Human Services, part 493. Laboratory requirements, section 493.2. Definitions. U.S. Government Printing Office, Washington, DC.
28. **Code of Federal Regulations.** 2009. Title 42. Public health, vol. 4, chapter V. Health Care Financing Administration, Department of Health and Human Services, part 493. Laboratory requirements, section 493.1253. Standard: establishment and verification of performance specifications. U.S. Government Printing Office, Washington, DC.
29. **Code of Federal Regulations.** 2009. Title 42. Public health, vol. 4, chapter V. Health Care Financing Administration, Department of Health and Human Services, part 493. Laboratory requirements, section 493.1255. Standard: calibration and calibration verification procedures. U.S. Government Printing Office, Washington, DC.
30. **Code of Federal Regulations.** 2009. Title 42. Public health, vol. 4, chapter V. Health Care Financing Administration, Department of Health and Human Services, part 493. Laboratory requirements, section 493.1256. Standard: control Procedures. U.S. Government Printing Office, Washington, DC.
31. **Code of Federal Regulations.** 2009. Title 42. Public health, vol. 4, chapter V. Health Care Financing Administration, Department of Health and Human Services, part 493. Laboratory requirements, section 493.1282. Standard: corrective actions. U.S. Government Printing Office, Washington, DC.
32. **Code of Federal Regulations.** 2009. Title 42. Public health, vol. 4, chapter V. Health Care Financing Administration, Department of Health and Human Services, part 493. Laboratory requirements, section 493.1455. Standard: clinical consultant qualifications. U.S. Government Printing Office, Washington, DC.
33. **College of American Pathologists, Commission on Laboratory Accreditation.** 2009. Microbiology checklist. College of American Pathologists, Northfield, IL.
34. **College of American Pathologists, Commission on Laboratory Accreditation.** 2009. Molecular pathology checklist. College of American Pathologists, Northfield, IL.
35. **Dreier, J., M. Störmer, and K. Kleesiek.** 2005. Use of bacteriophage MS2 as an internal control in viral reverse transcription-PCR assays. J. Clin. Microbiol. **43:**4551–4557.
36. **Dreier, J., M. Störmer, D. Mäde, S. Burkhardt, and K. Kleesiek.** 2006. Enhanced reverse transcription-PCR assay for detection of norovirus genogroup I. J. Clin. Microbiol. **44:**2714–2720.
37. **Drosten, C., E. Seifried, and W. K. Roth.** 2001. TaqMan 5′-nuclease human immunodeficiency virus type 1 PCR assay with phage-packaged competitive internal control for high-throughput blood donor screening. J. Clin. Microbiol. **39:**4302–4308.
38. **Drosten, C., M. Panning, J. F. Drexler, F. Hänsel, C. Pedroso, J. Yeats, L. K. de Souza Luna, M. Samuel, B. Liedigk, U. Lippert, M. Stürmer, H. W. Doerr, C. Brites, and W. Preiser.** 2006. Ultrasensitive monitoring of HIV-1 viral load by a low-cost real-time reverse transcription-PCR assay with internal control for the 5′ long terminal repeat domain. Clin. Chem. **52:**1258–1266.
39. **Eis, P. S., M. C. Olson, T. Takova, M. L. Curtis, S. M. Olson, T. I. Vener, H. S. Ip, K. L. Vedvik, C. T. Bartholomay, H. T. Allawi, W. P. Ma, J. G. Hall, M. D. Morin, T. H. Rushmore, V. I. Lyamichev, and R. W. Kwiatkowski.** 2001. An invasive cleavage assay for direct quantitation of specific RNAs. Nat. Biotechnol. **19:**673–676.
40. **Espy, M. J., J. R. Uhl, L. M. Sloan, S. P. Buckwalter, M. F. Jones, E. A. Vetter, J. D. C. Yao, N. L. Wengenack, J. E. Rosenblatt, F. R. Cockerill III, and T. F. Smith.** 2006. Real-time PCR in clinical microbiology: applications for routine laboratory testing. Clin. Microbiol. Rev. **19:**165–256.
41. **Federal Food, Drug and Cosmetic Act.** Section 201(h), now codified as Title 21, food and drugs, chapter 9, Federal Food, Drug, and Cosmetic Act, subchapter II definitions, section 321(h), U.S. Code. U.S. House of Representatives, Washington, DC.
42. **Ferreira-Gonzalez, A., and A. M. Caliendo.** 2007. Viral infections in transplant patients. *In* D. G. B. Leonard (ed.), Molecular pathology in clinical practice, Springer Science+Business Media, LLC, New York, N.
43. **Food and Drug Administration.** 2001. FDA guidance for industry: bioanalytical method validation. Center for Drug Evaluation and Research, U.S. Department of Health and Human Services, Rockville, MD.
44. **Garson, J. A., P. R. Grant, U. Ayliffe, R. B. Ferns, and R. S. Tedder.** 2005. Real-time PCR quantitation of hepatitis B virus DNA using automated sample preparation and murine cytomegalovirus internal control. J. Virol. Methods **126:**207–213.
45. **International Organization for Standardization.** 2006. Statistics—vocabulary and symbols, part 1. Probability and general statistical terms. ISO 3534-1. International Organization for Standardization, Geneva, Switzerland.
46. **International Organization for Standardization.** 1993. International vocabulary of basic and general terms in metrology. International Organization for Standardization, Geneva, Switzerland.
47. **International Organization for Standardization.** 2007. Medical laboratories—particular requirements for quality and competence. ISO 15189, 2nd ed. International Organization for Standardization, Geneva, Switzerland.
48. **International Organization for Standardization.** 2003. In vitro diagnostic medical devices—measurement of quantities in biological samples. Metrological traceability of values assigned to calibrators and control materials. ISO 17511. International Organization for Standardization, Geneva, Switzerland.
49. **Jennings, L., V. M. Van Deerlin, and M. L. Gulley.** 2009. Recommended principles and practices for validating clinical molecular pathology tests. Arch. Pathol. Lab. Med. **133:**743–755.
50. **Johnson, R.** 2008. Assessment of bias with emphasis on medical comparison. Clin. Biochem. Rev. **29:**S37–S42.
51. **Klee, S. R., J. Tyczka, H. Ellerbrok, T. Franz, S. Linke, G. Baljer, and B. Appel.** 2006. Highly sensitive real-time PCR for specific detection and quantification of *Coxiella burnetii*. BMC Microbiol. **6:**2.
52. **Koch, D. D., and T. Peters.** 1999. Selection and evaluation of methods *In* C. A. Burtis and E. R. Ashwood (ed.), Tietz textbook of clinical chemistry, 3rd ed. W. B. Saunders Co., Philadelphia, PA.
53. **Kocjan, B. J., K. Seme, and M. Poljak.** 2008. Detection and differentiation of human papillomavirus genotypes HPV-6 and HPV-11 by FRET-based real-time PCR. J. Virol. Methods **153:**245–249.
54. **Koidl, C., M. Bozic, A. Burmeister, M. Hess, E. Marth, and H. H. Kessler.** 2007. Detection and differentiation of *Bordetella* spp. by real-time PCR. J. Clin. Microbiol. **45:**347–350.
55. **Linnet, K.** 1999. Necessary sample size for method comparison studies based on regression analysis. Clin. Chem. **45:**882–894.
56. **Linnet, K., and J. C. Boyd.** 2006. Selection and analytical evaluation of methods—with statistical techniques. *In* C. A. Burtis, E. R. Ashwood, and D. E. Bruns (ed.), Tietz textbook of clinical chemistry, 4th ed. Elsevier Saunders, Philadelphia, PA.
57. **Maaroufi, Y., J. M. de Bruyne, V. Duchateau, R. Scheen, and F. Crokaert.** 2006. Development of a multiple internal control for clinical diagnostic real-time amplification assays. FEMS Immunol. Med. Microbiol. **48:**183–191.
58. **Mahony, J. B., A. Petrich, L. Louie, X. Song, S. Chong, M. Smieja, M. Chernesky, M. Loeb, and S. Richardson.** 2004. Performance and cost eval-

uation of one commercial and six in-house conventional and real-time reverse transcription-PCR assays for detection of severe acute respiratory syndrome coronavirus. J. Clin. Microbiol. **42:**1471–1476.
59. **Müller, J., A. M. Eis-Hübinger, M. Däumer, R. Kaiser, J. M. Rox, L. Gürtler, L. P. Hanfland, and B. Pötzsch.** 2007. A novel internally controlled real-time reverse transcription-PCR assay for HIV-1 RNA targeting the pol integrase genomic region. J. Virol. Methods **142:**127–135.
60. **Nitsche, A., M. Büttner, S. Wilhelm, G. Pauli, and H. Meyer.** 2006. Real-time PCR detection of parapoxvirus DNA. Clin. Chem. **52:**316–319.
61. **Olson, V. A., T. Laue, M. T. Laker, I. V. Babkin, C. Drosten, S. N. Shchelkunov, M. Niedrig, I. K. Damon, and H. Meyer.** 2004. Real-time PCR system for detection of orthopoxviruses and simultaneous identification of smallpox virus. J. Clin. Microbiol. **42:**1940–1946.
62. **Panning, M., J. Kilwinski, S. Greiner-Fischer, M. Peters, S. Kramme, D. Frangoulidis, H. Meyer, K. Henning, and C. Drosten.** 2008. High throughput detection of Coxiella burnetii by real-time PCR with internal control system and automated DNA preparation. BMC Microbiol. **8:**77.
63. **Saah, A. J., and D. R. Hoover.** 1997. "Sensitivity" and "specificity" reconsidered: the meaning of these terms in analytical and diagnostic settings. Ann. Intern. Med. **126:**91–94.
64. **Saldanha, J.** 1993. Assays for viral sequences and their value in validation of viral elimination. Dev. Biol. Stand. **81:**231–236.
65. **Sloan, L. M.** 2007. Real-time PCR in clinical microbiology: verification, validation, and contamination control. Clin. Microbiol. Newsl. **29:**87–95.
66. **Social Security Act.** Section 1862(a) (1), exclusions from coverage and Medicare as a second payer, now codified as Title 42, chapter 7, exclusions from coverage and Medicare as a second payer, U.S. Code. U.S. House of Representatives, Washington, DC.
67. **Störmer, M., K. Kleesiek, and J. Dreier.** 2007. High-volume extraction of nucleic acids by magnetic bead technology for ultrasensitive detection of bacteria in blood components. Clin. Chem. **53:**104–110.
68. **Tedder, R. S., U. Ayliffe, W. Preiser, N. S. Brink, P. R. Grant, K. S. Peggs, S. Mackinnon, F. Kreig-Schneider, S. Kirk, and J. A. Garson.** 2002. Development and evaluation of an internally controlled semiautomated PCR assay for quantification of cell-free cytomegalovirus. J. Med. Virol. **66:**518–523.
69. **Thompson, R., S. L. R. Ellison, and R. Wood.** 2002. Harmonized guidelines for single-laboratory validation of methods of analysis. Pure Appl. Chem. **74:**835–855.
70. Reference deleted.
71. **U.S. Department of Health and Human Services, Centers for Medicare and Medicaid Services.** 1992. Clinical laboratory improvement amendments of 1988 (CLIA). Fed. Regist. **57:**7002–7186.
72. **U.S. Department of Health and Human Services, Centers for Medicare and Medicaid Services.** 2003. Medicare, Medicaid and CLIA programs: laboratory requirements relating to quality systems and certain personnel qualifications. Final rule. Fed. Regist. **16:**3640–3714.
73. **Valenstein, P. N.** 1990. Evaluating diagnostic tests with imperfect standards. Am. J. Clin. Pathol. **93:**252–258.
74. **Valle, L., D. Amicizia, S. Bacilieri, F. Banfi, R. Riente, P. Durando, L. Sticchi, R. Gasparini, C. Esposito, G. Icardi, and F. Ansaldi.** 2006. Performance testing of two new one-step real time PCR assays for detection of human influenza and avian influenza viruses isolated in humans and respiratory syncytial virus. J. Prev. Med. Hyg. **47:**127–133.
75. **Welzel, T. M., W. J. Miley, T. L. Parks, J. J. Goedert, D. Whitby, and B. A. Ortiz-Conde.** 2006. Real-time PCR assay for detection and quantification of hepatitis B virus genotypes A to G. J. Clin. Microbiol. **44:**3325–3333.
76. **Westgard, J. O.** 2002. Basic QC practices, 2nd ed. Westgard QC, Inc., Madison, WI.
77. **Westgard, J. O.** 2008. Basic method validation, 3rd ed. Westgard QC, Inc., Madison, WI.
78. **Wölfel, R., J. T. Paweska, N. Petersen, A. A. Grobbelaar, P. A. Leman, R. Hewson, M. C. Georges-Courbot, A. Papa, S. Günther, and C. Drosten.** 2007. Virus detection and monitoring of viral load in Crimean-Congo hemorrhagic fever virus patients. Emerg. Infect. Dis. **13:**1097–1100.
79. **World Health Organization Expert Committee on Biological Standardization.** 1995. Glossary of terms for biological substances used for texts of the requirements. WHO unpublished document BS/95.1793. World Health Organization, Geneva, Switzerland.
80. **Zimmermann, B., A. El-Sheikhah, K. Nicolaides, W. Holzgreve, and S. Hahn.** 2005. Optimized real-time quantitative PCR measurement of male fetal DNA in maternal plasma. Clin. Chem. **51:**1598–1604.
81. **Zimmermann, P., I. Thordsen, D. Frangoulidis, and H. Meyer.** 2005. Real-time PCR assay for the detection of tanapox virus and yaba-like disease virus. J. Virol. Methods **130:**149–153.

Eileen M. Burd, Ph.D., D(ABMM), is the Director of Clinical Microbiology in the Department of Pathology and Laboratory Medicine at Emory University Hospital in Atlanta, GA. She also holds a faculty appointment as Associate Professor at Emory University School of Medicine. She received her doctoral degree from the Medical College of Wisconsin in Milwaukee and completed postdoctoral and clinical infectious disease fellowships, also at the Medical College of Wisconsin. She was Division Head, Microbiology, at Henry Ford Hospital in Detroit, MI, for 12 years prior to joining the faculty at Emory University in 2007. Her primary focus is the laboratory diagnosis and management of patients with infectious diseases. She is also actively involved in medical education and has varied research interests surrounding the nature of etiologic agents and the diagnosis and epidemiology of infectious diseases.