Recommended Principles and Practices for Validating Clinical Molecular Pathology Tests

Lawrence Jennings, MD, PhD; Vivianna M. Van Deerlin, MD, PhD; Margaret L. Gulley, MD;
for the College of American Pathologists Molecular Pathology Resource Committee

• *Context.*—The use of DNA- and RNA-based tests continues to grow for applications as varied as inherited disease, infectious disease, cancer, identity testing, human leukocyte antigen typing, and pharmacogenetics. Progress is driven in part by the huge growth in knowledge about the molecular basis of disease coupled with advancements in technologic capabilities. In addition to requirements for clinical utility, every molecular test also may have limitations that must be carefully considered before clinical implementation. Analytic and clinical performance characteristics as well as test limitations are established and documented through the process of test validation.

***Objective.*—To describe the established principles of test validation, along with relevant regulations in the United States, in order to provide a rational approach to introducing molecular tests into the clinical laboratory.**

***Data Sources.*—PubMed review of published literature, published guidelines, and online information from national and international professional organizations.**

***Conclusions.*—These resources and recommendations provide a framework for validating clinical tests.**

(*Arch Pathol Lab Med.* 2009;133:743–755)

The steps involved in test validation are meant to answer one main question—is a given test ready to be implemented in the clinical laboratory? In order to answer this question, the clinical utility (ie, usefulness) of the test must be established, along with its analytic performance characteristics and limitations. To understand these issues, one needs to define the intended use of the test and identify and quantify the possible sources of error as well as sources of analytic variation and biologic variation. Thus, what seems like a simple question can become quite complex.

The US Food and Drug Administration (FDA) is responsible for the oversight of in vitro diagnostic devices in the United States, but the number of laboratories or manufacturers that have acquired FDA approval for molecular tests is rather limited when one considers the wide range of molecular tests that are presently offered clinically. Furthermore, because of the cost and time commitments involved in gaining FDA approval, those tests are generally limited to high-volume tests (compare 20 heritable conditions listed on www.amp.org/FDATable/FDATable.doc vs 1261 clinical tests on www.genetests.org, accessed on May 23, 2008). Moreover, FDA approval is not required for clinical implementation of molecular tests as long as other regulations are met. As a consequence, individual laboratories have developed most molecular tests currently in use clinically, and these laboratories assume the full responsibility for validating their tests. This manuscript is intended to provide practical guidance for laboratorians with respect to regulatory requirements, established principles, and realistic approaches to test validation for both FDA-approved tests and laboratory-developed tests (LDTs). Analogous requirements exist for some clinical laboratories outside of the United States.[1]

REGULATION OF CLINICAL LABORATORY TESTS IN THE UNITED STATES

Federal regulation of clinical laboratory tests began in 1976 with the Medical Device Amendment to the Food, Drug, and Cosmetic Act. This amendment covered all reagents, tests, and equipment manufactured for the diagnosis or treatment of disease. In 1988, Congress passed the Clinical Laboratory Improvement Amendment (CLIA), which was enacted in 1992 to establish quality standards for all aspects of clinical laboratory testing. The CLIA is overseen by the Centers for Medicare & Medicaid Services (CMS), which publishes regulations and assumes primary responsibility for financial management operations of the CLIA program. The FDA has primary responsibility for categorizing the complexity of laboratory tests into 1 of 3 CLIA regulatory categories: waived for simple tests, and moderate or high complexity for other tests based on their potential risk to public health. The Centers for Disease Control serves an advisory role and provides scientific support.

The Code of Federal Regulations (CFR) is a compilation

Accepted for publication September 12, 2008.

From the Department of Pathology and Laboratory Medicine, Children's Memorial Hospital, Chicago, Illinois (Dr Jennings); the Department of Pathology and Laboratory Medicine, University of Pennsylvania Hospital, Philadelphia (Dr Van Deerlin); and the Department of Pathology, University of North Carolina, Chapel Hill (Dr Gulley).

The authors have no relevant financial interest in the products or companies described in this article.

Reprints: Margaret L. Gulley, MD, Department of Pathology, 913 Brinkhous-Bullitt Bldg, University of North Carolina, Chapel Hill, NC 27599-7525 (e-mail: margaret_gulley@med.unc.edu).

of the general and permanent rules published in the *Federal Register* by the executive departments and agencies of the US federal government. Title 21 of the CFR is reserved for rules of the FDA. In 1997, the FDA published 3 rules that were intended to ensure the quality of and control the risks associated with reagents used in LDTs. These rules (21 CFR 864.4020) defined and classified analyte-specific reagents (ASRs), which the FDA defines 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 specimen, are intended for use in a diagnostic application for identification and quantification of an individual chemical substance or ligand in biological specimens.'' The rules imposed restrictions on the sale, distribution, and use of ASRs (21 CFR 809.30) and established requirements for ASR labeling (21 CFR 809.10(e)). These rules, collectively known as the ''ASR Rule,'' hold that ''clinical laboratories that develop tests are acting as manufacturers of medical devices and are subject to FDA jurisdiction.''[2] To date, the FDA has generally chosen not to enforce the rule as it applies to LDTs (sometimes referred to as ''home brew tests''), at least in part because the FDA is already regulating the primary ingredients of most LDTs (the ASRs), and also because ''CLIA regulated laboratories qualified to perform high complexity testing have demonstrated expertise and ability to use ASRs in test procedures and analyses.''[2]

REQUIREMENTS FOR TEST VALIDATION

The FDA does not specify requirements for test validation, but it provides guidance for commercial manufacturers that intend to submit validation data for FDA approval or clearance. Guidance for pharmacogenetic tests and genetic tests for heritable markers can be found at www.fda.gov/cdrh/oivd/guidance/1549.html (accessed May 23, 2008), and statistical guidance can be found at www.fda.gov/cdrh/osb/guidance/1620.html (accessed May 23, 2008). In addition, the FDA provides a database of internationally recognized consensus standards, which can be found at www.fda.gov/cdrh/stdsprog.html (accessed May 23, 2008). These standards and guidance have largely been developed by the Clinical Laboratory Standards Institute (CLSI, formerly the National Committee for Clinical Laboratory Standards, or NCCLS).

CLIA regulations require certification of all laboratories performing clinical testing of any kind and provide both general guidelines and subspecialty-specific standards. Although there is a cytogenetic subspecialty under CLIA, there is no molecular pathology subspecialty. As a consequence, there are few specific requirements for molecular testing, but the general guidelines and requirements still apply and are considered sufficient. CLIA requirements for test validation vary by test complexity and by whether the test is FDA approved, FDA cleared, or laboratory developed. FDA approval refers to products that are approved for marketing under the premarket approval process for new devices. FDA clearance refers to devices that are cleared for marketing under a 510(k) review, and such clearance is reserved for devices that are conceptually similar to those already on the market for which there is a predicate device. From the perspective of the clinical laboratory, the implications are the same; thus, we will use ''FDA-approved'' to indicate FDA-approved or cleared devices.

CLIA requires that each lab establish or verify the performance specifications of moderate- and high-complexity test systems that are introduced for clinical use.[3] For an FDA-approved test system without any modifications, the system has been validated by the manufacturer; therefore, the laboratory need only verify (ie, confirm) these performance specifications. In contrast, for a modified FDA-approved test or an LDT, the laboratory must establish the test's performance specifications, including accuracy, precision, reportable range, and reference range. The laboratory must also develop and plan procedures for calibration and control of the test system. In addition, the laboratory must establish analytic sensitivity and specificity and any other relevant indicators of test performance. A codified version of CLIA regulations can be found at www.phppo.cdc.gov/clia/regs/toc.aspx (accessed May 23, 2008). How and to what extent performance specifications should be verified or established is ultimately under the purview of medical laboratory professionals and is overseen by the CLIA certification process, including laboratory inspectors.

Professional organizations or state agencies demonstrating that their accreditation requirements meet or exceed those specified by CLIA have ''deemed status'' from CMS to accredit laboratories. The 3 largest organizations with deemed status are the College of American Pathologists (CAP), the Joint Commission on Accreditation of Healthcare Organizations, and the Commission for Office Laboratory Accreditation. Of these, CAP accredits most molecular pathology laboratories, and therefore this manuscript focuses on the requirements of CAP in the context of test validation. CAP requirements largely reflect the performance specifications outlined by CLIA, but in some circumstances CAP requirements exceed those of CLIA. The CAP checklist (checklist 12) for molecular pathology was created in 1991 and is periodically updated to modernize the requirements for test validation in the face of changing technologies and applications (www.cap.org; accessed May 23, 2008).

PRINCIPLES OF TEST VALIDATION

In order to understand the principles of test validation, it is essential to comprehend relevant terminology, yet the regulating organizations generally do not attempt to define these terms. As a result, the interpretation of these terms may vary considerably among laboratories and organizations, and this has been an ongoing source of confusion. The CLSI strives to accurately define the terminology used in its documents. However, the CLSI documents have some inconsistencies because different consensus groups wrote the documents at different times, and the definitions have continued to evolve. Therefore, the CLSI now provides a ''Harmonized Terminology Database,'' which is freely accessible (www.clsi.org; accessed May 23, 2008), to ''encourage broad acceptance and usage of internationally accepted terminology.'' We encourage the use of this terminology, some of which is derived from work by the International Organization for Standardization (ISO) and available at www.iso.org (accessed May 23, 2008). Below, we provide these definitions together with our interpretation and specific examples to help clarify these terms as applied to molecular tests.

VALIDATION VERSUS VERIFICATION

The terms *validation* and *verification* are both used in the context of new test development. Validation is defined as ''confirmation, through the provision of objective evidence, that requirements for a specific intended use or application have been fulfilled'' (ISO 9000).[4] Meanwhile, verification is defined as ''confirmation, through the provision of objective evidence, that specified requirements have been fulfilled'' (ISO 9000).[4] Although the terms are similar, validation goes a step beyond verification. One can interpret this to mean the following: validation determines that we are doing the *correct* test; verification confirms that we are doing the test *correctly.* Thus, validation requires identifying the needs of the user and establishing documented evidence that provides a high degree of assurance that a test will consistently meet those needs (adapted from the *Guideline on General Principles of Process Validation*[5]). In other words, the test is appropriate for its intended use. Verification, on the other hand, requires confirmation that predetermined specifications are consistently met. In other words, the test performs as expected from an analytic perspective. For example, an FDA-approved test has generally been validated for a specific purpose using specified equipment, reagents, and controls that have all been qualified to meet certain specifications; a clinical laboratory wishing to introduce this test must verify the analytic performance characteristics of the test in its own hands before introducing it for clinical use, assuming that the laboratory has the same intended use as the FDA-approved test and that its test is performed using the same qualified equipment, reagents, protocol, and so forth. If the FDA-approved test is altered (including such seemingly minor changes as a different patient population, specimen type, or nucleic acid extraction method), then it may no longer be either appropriate for its intended use (ie, valid) or analytically sound (ie, verifiable). Therefore, the laboratory is required to perform its own validation. Whether a test is FDA approved or not, CLIA requires that analytic performance characteristics initially be defined and that performance is verified at regular intervals by implementing a series of checks to demonstrate reagent, equipment, and technical qualifications (eg, calibration verification). Obviously, for a test to continue to be valid, the specified requirements must continue to be verified.

ANALYTIC VALIDATION VERSUS CLINICAL VALIDATION

Analytic validity focuses on the analyte(s) targeted by the assay, whereas clinical validity focuses on the associated diseases or patient conditions. A goal of analytic validation is to identify and quantify potential sources of technical variation in the analysis of patient samples. The sources of technical variation vary depending on the test, and therefore so do the performance characteristics that need to be addressed in the analytic validation process. Historically, molecular testing, with its origins primarily in hematopathology and genetic testing, was predominantly qualitative (detecting an inherited mutation, for example); however, with newer technologies and applications, quantitative molecular testing has become commonplace. Quantitative molecular tests that measure, for example, gene dosage, tumor burden, or viral load are much more like traditional chemistry tests, both in terms of their performance characteristics and their need for calibration.

A major goal of clinical validation is to identify and quantify potential sources of biologic variation in the analysis of a given sample. This process influences the extent to which the test is medically useful. By defining the ''normal'' and ''abnormal'' ranges of values, a given test provides practical information that can help direct the management of a patient. In this regard, it is important that a physician consultant in every clinical laboratory (mandated by CLIA) be part of the process of vetting a test prior to its clinical implementation (www.cdc.gov/clia/regs/subpart_m.aspx#493.1441; accessed May 23, 2008). Even though the testing laboratory may use published literature as supportive evidence for clinical validity, CLIA requires that every laboratory have a clinical consultant who is ''qualified to consult with and render opinions to the laboratory's clients concerning the diagnosis, treatment and management of patient care'' (CLIA section 493.1455). This implies that the consultant is ethically and medically responsible for assessing clinical performance (either in his or her own hands or by extrapolation from reliable sources, such as published evidence) in order to provide a communication link between the laboratory and clinicians and in order to practice evidence-based laboratory medicine.[6] A test that is analytically sound but has no established clinical utility should not be offered clinically.

ANALYTIC INTERPRETATION VERSUS CLINICAL INTERPRETATION

Interpretation of test results has both an analytic and a clinical component. Developing guidelines for analytic and clinical interpretation of results is part of the analytic and clinical validation process. For example, a polymerase chain reaction amplification may be intended to quantify the CGG repeats in the 5′ untranslated region of *FMR1*, which is associated with fragile X syndrome when >200 repeats are present. Evaluation of the peak pattern on capillary gel electrophoresis, which is usually done in the context of patient sex, comprises the analytic interpretation, and establishing interpretive criteria is part of the analytic validation of the test. Errors in analytic interpretation may occur in a patient with Klinefelter syndrome (male with XXY) when the assay has not been calibrated properly or when factors other than CGG repeat number contribute to peak pattern variation.[7] On the other hand, clinical interpretation depends on the clinical context: For example, an *FMR1* gene premutation is interpreted differently if the patient is a child with developmental delay, a pregnant woman, a woman having trouble becoming pregnant, or an elderly man with ataxia. Medical judgment is often required in individual cases, and it is the role of the laboratory director to ensure that patient reports are appropriately interpreted. In addition, newly validated tests may be less familiar to clinicians and variable across testing laboratories, so an educational component should also be considered for inclusion in patient reports. Guidelines for reporting molecular pathology results have been published.[8]

ANALYTIC PERFORMANCE CHARACTERISTICS

For quick reference, relevant performance characteristics are summarized in Table 1. For quantitative tests, the *accuracy* of an analytic procedure expresses the closeness of agreement between the value found and the value that is accepted either as a conventional true value or an accepted reference value.[9] The closeness of agreement observed is the result of systematic error and random error (ie, total

Table 1. Comparison of Test Parameters That Need to Be Addressed for Quantitative and Qualitative Tests

Parameter	Definition	Comment
Accuracy	Closeness of agreement between a test result and an accepted reference value.	Represents total error, which is the sum of systematic error (bias) and random error. When applied to qualitative tests, ''accuracy'' is equivalent to ''sensitivity and specificity.''
Trueness	Systematic error (bias) resulting in consistent under- and overestimation compared to true value.	A measure of systematic error and can be expressed as a percent deviation from the true value. Applies to quantitative tests.
Precision	Closeness of agreement between independent test results obtained under stipulated conditions.	A measure of random error and usually is expressed as standard deviation or coefficient of variation. Applies to quantitative tests.
Reproducibility	A property of an experiment or process where one tends to receive consistent results from following a specific procedure.	Equivalent to ''precision'' for qualitative and semiquantitative tests. Repeatability is used to indicate within-run reproducibility.
Robustness	Test precision given small, deliberate changes in test conditions (eg, preanalytic delays, variations in storage temperature).	Assessment may be required for laboratory developed tests or modified FDA-approved tests.
Linearity	The ability of the test to return values that are directly proportional to the concentration of the analyte in the sample.	Applies to quantitative tests.
Reportable range	The range of values over which the relationship between the instrument, kit, or system's measurement response is shown to be valid.	Applies to all tests.
Reference range	The range of test values expected for a designated population of individuals.	Some tests, such as HLA genotyping, do not have a ''reference range.''
Interfering substances	An interfering substance in analytical procedures is one that, at the given concentration, causes a systematic error in the analytical result.	May include biologic matrix elements and closely related analytes.
Analytic sensitivity	The ability to obtain positive results in concordance with positive results obtained by the reference method.	Applies to qualitative and semiquantitative tests. A more precise term is ''positive agreement as compared to.''
Analytic specificity	The ability to obtain negative results in concordance with negative results obtained by the reference method.	Applies to qualitative and semiquantitative tests. A more precise term is ''negative agreement as compared to.''
Limit of detection	The lowest amount of analyte that is statistically distinguishable from background or a negative control.	For quantitative and semiquantitative tests, the term should be used instead of ''analytic sensitivity.'' The term has also been applied to qualitative tests to express the lowest concentration of genomic DNA that yields informative results.
Limit of quantification	Lowest and highest concentrations of analyte that can be quantitatively determined with suitable precision and accuracy.	Applies to quantitative tests.
Clinical sensitivity	The proportion of subjects who have a specified disorder and whose test results are interpreted as positive.	If a quantitative test has a cutoff for clinical decision making, the clinical sensitivity is calculated using this cutoff value.
Clinical specificity	The proportion of subjects who do not have a specified disorder and whose test results are interpreted as negative.	If a quantitative test has a cutoff for clinical decision making, the clinical specificity is calculated using this cutoff value.

Abbreviations: FDA, US Food and Drug Administration; HLA, human leukocyte antigen.

error). Systematic error (*bias*, expressed as *trueness*[10]) is the tendency to consistently overestimate or underestimate the true value; in some cases it may be overcome by recalibration. Random error is unpredictable error (such as sampling error) that is expressed by the *precision* of measurement, which is defined as the ''closeness of agreement between independent test results obtained under stipulated conditions'' (ISO 3534-1).[10] Precision of a given measurement procedure is subdivided into *repeatability* that relates to *within-run precision*; *intermediate precision* when one or more of the relevant factors varies; and *reproducibility* for more substantial changes in conditions (eg, different laboratories, calibrations, or measurement systems), which is often termed *interlaboratory precision* (ISO 15195).[11] For practical purposes of validating a new test in one's laboratory, we recommend using the term *reproducibility* to reflect the work done to define and test the variables that one deems likely to have an impact on precision (eg, operators, days, lots of reagents and calibrators, alternative instruments or testing sites). An illustration of these concepts is seen in Figure 1, where a hypothetical example shows multiple measurements of an analyte of known concentration using 4 different procedures[12]; the diagram demonstrates the mean and dispersion about the mean of each method. The deviation of the mean from the true value (percent bias) and the imprecision (random error) need to be considered together to estimate total error and to judge which procedure deserves further consideration. In the example shown, if acceptance limits are set at ±15% total error from the reference value, then procedures 1 and 2 are acceptable because they meet this objective, whereas procedures 3 and 4 are unacceptable.[12]

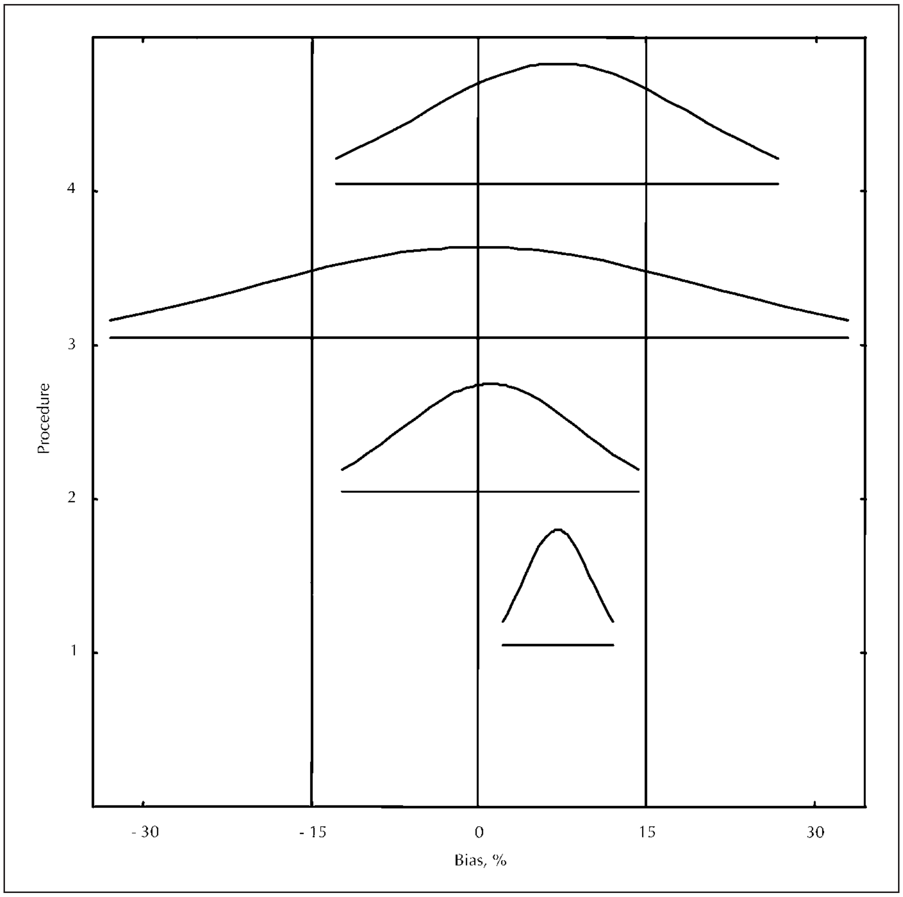

Figure 1. *Diagram demonstrating the association of accuracy, imprecision, and trueness for a hypothetical example of 4 different analytic procedures done repeatedly on the same specimen. Procedure 1 has tightly clustered values with a mean that is greater than the reference standard (systematic bias), implying that adjustments to calibrate this assay might render it even more accurate. Procedures 2 and 3 have negligible systematic bias, although procedure 2 is more precise. Reprinted from Hubert et al.*[12] *Harmonization of strategies for the validation of quantitative analytical procedures. A SFSTP proposal–part I.* J Pharm Biomed Anal. *2004;36:579–586 with permission from Elsevier.*

For qualitative tests, the term *accuracy* is difficult to apply because ''closeness'' to a ''true'' value does not make sense; the result is generally either correct or incorrect when compared to the truth. As a result, several published papers addressing the specific challenges of molecular test validation offer alternative definitions of ''accuracy.''[13–17] The lack of consensus on this issue in the published literature was recognized by the Cochrane Collaboration (an international not-for-profit organization focused on evidence-based health care), and in 1999 this organization began an initiative to define Standards for the Reporting of Diagnostic Accuracy. The Standards for the Reporting of Diagnostic Accuracy steering committee offers a definition of accuracy that includes analytic as well as clinical findings.[18] The committee states, ''the term accuracy refers to the amount of agreement between the information from the test under evaluation, referred to as the index test, and the reference standard.'' The committee defines the reference standard as ''the best available method for establishing the presence or absence of the condition of interest. The reference standard can be a single method, or a combination of methods, to establish the presence of the target condition. It can include laboratory tests, imaging and pathology, but also dedicated clinical follow-up of participants.'' The committee goes on to say, ''diagnostic accuracy can be expressed in many ways, including sensitivity and specificity, likelihood ratios, diagnostic odds ratios and the area under a receiver-operator characteristic curve.'' Ever since the simultaneous publication of the Standards for the Reporting of Diagnostic Accuracy statement in 7 journals in 2003, the Standards for the Reporting of Diagnostic Accuracy statement has been widely adopted.

For qualitative tests, the term *precision* is also difficult to apply, but it is important to demonstrate analytical consistency with a given test. Therefore, within-run repeatability and within-laboratory reproducibility studies are usually performed to show that a test will return the same result regardless of the operator, lot, day, and so forth.

The *reportable range* is defined as ''the range of test values over which the relationship between the instrument, kit, or system's measurement response is shown to be valid'' (US CFR 493; February 28, 1992). CLSI adds the note, ''the range of values (in units appropriate for the analyte [measurand]) over which the acceptability criteria for the method have been met; that is, where errors due to nonlinearity, imprecision, or other sources are within defined limits.''[19] Expressed simply, the reportable range includes all possible results that can be reported, whether it is a range of numbers or else defined qualitative results, such as ''negative'' or ''positive.''

Related terms for quantitative assays include *linearity* and *analytic measurement range* (AMR). Linearity is the ability of the test to return values that are directly proportional to the concentration of the analyte in the sample. Mathematic data transformations to promote linearity may be allowed if there is scientific evidence that the transformation is appropriate for the method. As an example, quantitative polymerase chain reaction uses log base 10 transformation of the concentration of an analyte to generate a linear standard curve.

The AMR is the range of analyte values that a method can directly measure on the sample without any dilution, concentration, or other pretreatment that is not part of the typical test process. Because specimens can be diluted, the reportable range may be greater than the AMR. For example, if the upper end of the AMR for a viral load test is 10 million copies per milliliter, a patient sample of 15 million copies per milliliter is not directly reportable without further evaluation. One can report it simply as greater than 10 million copies per milliliter or make dilutions of the sample into the AMR range after validating the linearity of the dilution procedure. If the dilutions are linear, then these data can be used to extend the reportable range beyond the AMR.

Robustness is defined as test precision given small, deliberate changes in test conditions (eg, prolonged delays, variations in storage temperature). Robustness is thoroughly investigated by manufacturers to develop specimen collection, processing, and storage requirements for FDA-approved tests. Similar investigation of relevant factors is indicated when validating LDTs and when modifying an FDA-approved test, especially as it relates to alternate sample types or specimen handling, such as length of sample storage or storage temperature prior to or after nucleic acid extraction. The results of these evaluations can be used to set preanalytic acceptability criteria.

Interfering substances refer to any substance that, at a given concentration, causes a systematic error in the analytic result. These may include biologic matrix elements that have a negative effect on the test value, such as hemoglobin, high- or low-level DNA, lipid, protein, and glycosaminoglycans. To identify interfering substances, weakly positive specimens could be spiked with plausible matrix materials to test for false-negative results (eg, spike cervical brushings used for human papilloma virus testing with blood, lubricant, or sperm). Studies may also involve closely related analytes (eg, a test that is intended to be specific for herpes simplex virus 1 should not cross-react with herpes simplex virus 2).

The *limit of detection* (LoD) is the lowest amount of analyte that is distinguishable from background (a negative control) with a stated level of confidence. The level of confidence is usually set at 95%, which indicates that at least 95% of the time the given concentration could be distinguished from a negative sample. The LoD obviously applies to quantitative tests, but it also applies to semiquantitative tests and even qualitative tests. For a semiquantitative test in which quantitative data are available and a threshold is set to determine positivity, one might want to identify the lowest concentration of a pathogen or a tumor biomarker that can be detected. In such a case, one may wish to use a receiver-operator characteristic curve analysis to evaluate several cutoff values that could affect both the false-positive and false-negative call rates. For example, van Gils et al[20] demonstrated how a receiver-operator characteristic curve can be used to determine a cutoff value (Figure 2). In this example, they used biopsy as the standard for detecting prostate cancer and determined the ratio of *PCA3/PSA* RNA after digital rectal exam, expressed as a *PCA3* score. They found the optimum cutoffs (calculated as *PCA3* copies/*PSA* copies × 1000) to be 66 in prostatic fluid and 43 in urine. Although the test is reported in a qualitative fashion, it is actually semiquantitative in that quantitative raw data are evaluated, and therefore periodic calibration verification at or near the cutoff value is essential. As another example, an interphase fluorescence in situ hybridization test may involve counting the percentage of scorable nuclei with the expected signal pattern; the LoD might be characterized as the lowest percentage of cells with the expected pattern that is distinguishable from a negative control at least 95% of the time. Obviously, a person who is counting nuclei cannot be calibrated, but demonstrating technologist-to-technologist consistency becomes critical. Limit of detection also has been applied to qualitative tests. In this regard, it corresponds to the lowest concentration of analyte that gives an informative result. For a genetic mutation, this may correspond to the concentration of DNA required to give an accurate result more than 95% of the time. For example, the FDA summary of the Nanosphere Verigene system, which detects genetic polymorphisms of cytochrome P450 2C9 (*CYP450 2C9*) and vitamin K epoxide reductase complex subunit 1 (*VKORC1*), describes LoD as a concentration of genomic DNA of 40 ng/μL, above which the call rate should be 92% or better (www.fda.gov/cdrh/reviews/K070804.pdf, accessed June 30, 2008).

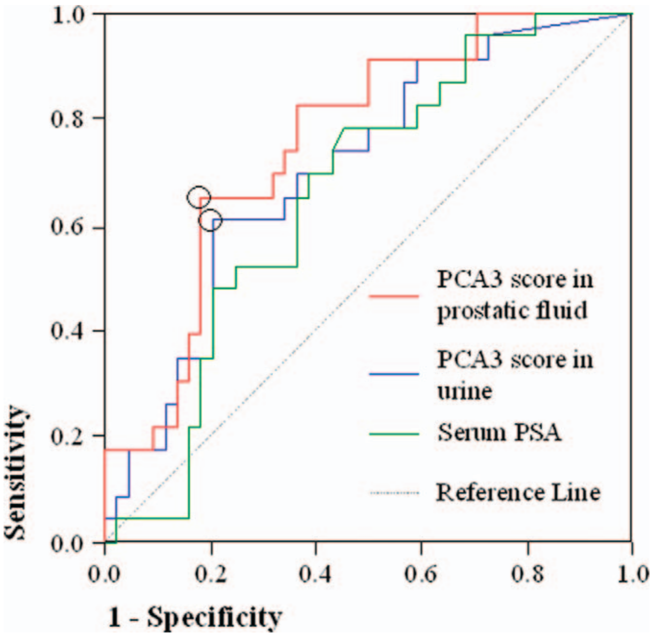

Figure 2. *Demonstration of effective utilization of a receiver-operator characteristic curve. This diagram shows performance characteristics of three molecular tests: serum prostate-specific antigen (PSA) and the PCA3 score in prostatic fluid and in urinary sediments. The open circles indicate the optimal cutoff values for calling each molecular test positive in relation to the reference standard test, which is the histopathologic diagnosis of cancer. Reprinted from van Gils et al.[20] Molecular PCA3 diagnostics on prostatic fluid.* Prostate. *2007;67:881–887 with permission of Wiley-Liss Inc, a subsidiary of John Wiley & Sons Inc.*

The *limits of quantification* (LoQ) are the lowest and highest concentrations of analyte that can be quantitatively determined with acceptable accuracy (total error). Because every test is different with regard to how accurate test values need to be, LoQ is determined at the laboratory director's discretion based on estimates of total error combined with analysis of the clinical impact of inaccurate results. A hypothetical example is illustrated in Figure 3. In this example, λ is the total error that is considered acceptable for quantitation, the horizontal dashed line is the mean of multiple measurements, and the shaded area represents the dispersion about the mean (SD).[12] C_1 through

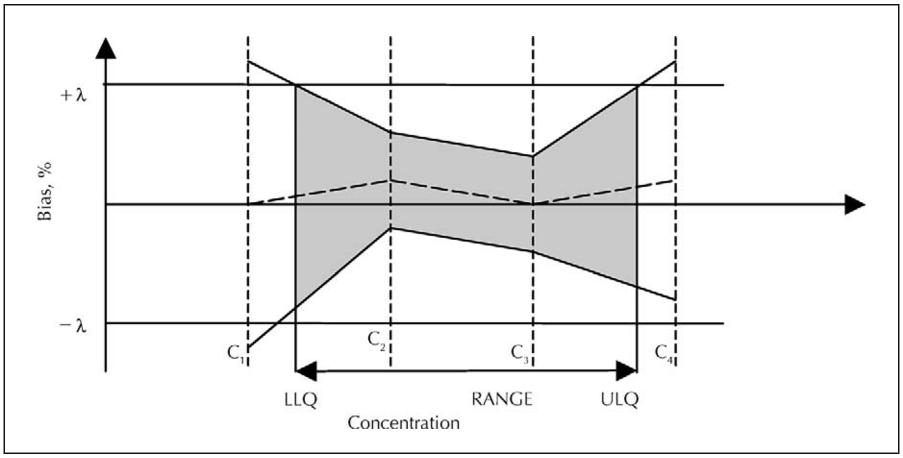

Figure 3. *Illustration of accuracy profile as a tool to determine the lower limit of quantitation (LLQ) and upper limit of quantitation (ULQ). See text for further clarification. Reprinted from Boulanger et al.*[12] *Harmonization of strategies for the validation of quantitative analytical procedures. A SFSTP proposal–part I.* J Pharm Biomed Anal. *2004;36: 579–586 with permission from Elsevier.*

C_4 represent 4 different concentrations of analyte in biologically appropriate matrix. The limits of quantification correspond to the lower and upper concentrations where the means ± 2SD exceed the acceptance limit, so the lower limit of quantification falls between the C_1 and C_2 concentrations and the upper limit of quantification falls between the C_3 and C_4 concentrations. It is important to note that LoD and LoQ determinations should be done in biologically appropriate matrices.

Analytic sensitivity has been defined differently by different organizations, and the meaning has become ambiguous and test dependent. In its Standards and Guidelines, the American College of Medical Genetics defines analytic sensitivity as the proportion of biologic samples that have a positive test result or known mutation and that are correctly classified as positive (www.acmg.net; accessed May 23, 2008). The group differentiates this from *clinical sensitivity*, which is defined as the proportion of individuals who have (or will develop) the phenotype of interest and who have a positive test result (this distinction is discussed further below). These interpretations seem reasonable for a qualitative test or for a quantitative test with decision limits. However, the meaning of the term ''analytic sensitivity'' can be quite different when applied to quantitative tests. In quantitative testing, analytic sensitivity has been defined as the lowest amount or concentration of an analyte that is detectable by a test, and it is therefore equivalent to LoD. The FDA does not attempt to define these terms, but it is evident from the FDA decision summaries of cleared or approved tests that it has uniformly interpreted the term as LoD for all tests, even qualitative genetic tests (www.fda.gov/cdrh/oivd/; accessed May 23, 2008). To avoid confusion, consider avoiding the ambiguous term ''analytic sensitivity'' and instead use the terms ''LoD'' or ''positive agreement to a comparative method.''

Likewise, *analytic specificity* has been defined by the American College of Medical Genetics as the proportion of biologic samples that have a negative test result or no identified mutation (being tested for) and that are correctly classified as negative (www.acmg.net; accessed May 23, 2008). In contrast, CLSI defines it as the ability of a test or procedure to correctly identify or quantify an entity in the presence of interfering substances that might be expected to be present. However, CLSI does add the note that for qualitative or semiquantitative tests, it is the method's ability to obtain negative results in concordance with negative results obtained by the reference method. It seems the FDA has adopted the interpretation that analytic specificity is equivalent to interfering substances and/or cross-reactivity (www.fda.gov/cdrh/oivd/; accessed May 23, 2008). To avoid confusion, consider avoiding the ambiguous term ''analytic specificity'' and instead use the terms ''interfering substances,'' ''cross-reactivity,'' or ''negative agreement to a comparative method.''

CLINICAL PERFORMANCE CHARACTERISTICS

The *reference range* is the range of test values expected for a designated population of individuals (US CFR 493; February 28, 1992). Equivalent terms include *reference interval or normal range* and are determined by testing a series of specimens from a given population who are known to be free of the disease of concern. For most tests, the ''normal'' and ''abnormal'' ranges of values are established from previous studies, and the goal is to confirm the normal range in a particular patient population. For a qualitative test, the reference range may be ''negative,'' ''normal,'' ''no clonal population,'' or some other term that tends to clarify whether the result is normal or abnormal. It should be noted that some tests do not have a reference range. For example, hepatitis C virus genotyping in a known hepatitis C virus–positive sample does not have a ''normal'' or expected genotype among all known genotypes. Similarly, in identity testing and human leukocyte antigen typing, the results are determined for a sample, but their significance may be evaluated only after comparison to one or more additional samples.

Clinical sensitivity is defined as the proportion of individuals who have (or will develop) the phenotype of interest and who have a positive test result, whereas *clinical specificity* is the proportion of subjects who do not have a specified disorder and whose test results are interpreted as negative (www.acmg.net; accessed May 23, 2008). Note that for clinical performance characteristics, clinical diagnosis remains the ''gold standard,'' recognizing that clinical diagnosis is often heavily influenced by laboratory test results; indeed, it is estimated that up to 80% of medical decision making is guided by the use of laboratory tests.[21] For some conditions, diagnosis is so dependent on a particular laboratory result that the test virtually defines the disease, and the distinction between analytic and clinical performance become blurred. Nevertheless, the combination of family and medical history, physical findings, various laboratory tests, and other clinicopathologic findings

Table 2. Clinical Performance Characteristics and Their Relationship to One Another[a]

2 × 2 Contingency Table, Diagnostic Test	Disease		
	Present	Absent	Totals
Positive	a	b	a + b
Negative	c	d	c + d
Totals	a + c	b + d	a + b + c + d

[a] Sensitivity = a/(a + c). Specificity = d/(b + d). Positive predictive value = a/(a + b). Negative predictive value = d/(c + d). Likelihood ratio positive test = a/(a + c)/b/(b + d). Likelihood ratio positive test = sensitivity/(1 − specificity). Likelihood ratio negative test = c/(a + c)/d/(b + d). Likelihood ratio negative test = (1 − sensitivity)/specificity.

makes clinical diagnosis the best standard by which the clinical performance characteristics of novel laboratory tests can be estimated.[22–24]

The distinction between analytic and clinical performance characteristics should be maintained. For example, consider that an allele-specific polymerase chain reaction may consistently detect a particular *CFTR* gene mutation (compared with a reference method such as DNA sequencing), but this polymerase chain reaction only detects one among the more than 1000 known disease-causing mutations. Thus, its clinical sensitivity for diagnosis of cystic fibrosis (as well as for detecting carrier status) would be poor, even though the positive agreement to the comparative test (sequencing) is good for detecting the single mutation. As another example, the presence of nucleic acid from a given human pathogen may be defined as abnormal by the laboratory, but this result does not necessarily indicate disease if that pathogen is nonviable or if it is considered normal flora in some anatomic sites or in some clinical circumstances. Therefore, determining clinical performance requires correlation of laboratory findings with other data from, for example, case-control studies, retrospective chart reviews, or clinical trials.

Clinical trials are often designed to assess the clinical validity of a novel molecular target that can be assessed by a laboratory test. If the test result is used to alter patient management (eg, to meet criteria for enrollment or to assign a treatment arm), then such patient-specific reporting of a laboratory test result requires that the test be reliable, and therefore that it be done in a CLIA-certified clinical laboratory. In this circumstance, the test must be analytically valid, but it may not yet have established clinical validity. Such clinical research requires approval by an institutional review board. For individual laboratories, it may not be possible to identify large numbers of positive cases to establish the clinical performance of a test. In such instances, the laboratory should obtain a reasonable number of positive cases and also cite relevant publications.

Other clinical performance terms and their associations with clinical sensitivity and clinical specificity are shown in Table 2. The *positive predictive value* is the proportion of patients with positive test results who truly have the disease, whereas the *negative predictive value* is the proportion of patients with negative test results who truly are free of the disease. Because diagnosis is not the only use of molecular tests, the concept should be extended to include whatever clinical condition the test has an impact on (eg, prognosis, detecting residual disease). It is important to note that positive and negative predictive values depend on the prevalence of the disease in the target population. For example, a test with a very good sensitivity may have a poor positive predictive value if the prevalence of disease is low. To address this issue and compare tests across different populations, it may be more useful to use a *likelihood ratio*, which is the probability that a specific test result is obtained in patients with the disease divided by the probability of obtaining the same test result in patients without the disease.[25] Likelihood ratios do not vary with changes in disease prevalence.[26] In addition, Bayesian analysis allows one to use likelihood ratios to convert a pretest probability into a posttest probability.[27] The posttest probability of one test can become the pretest probability of the next, so likelihood ratios from multiple tests can be used to generate a combined posttest probability, which provides a rational approach to evidence-based laboratory medicine, although one must be careful to identify tests that are independent of one another.[25,28] For example, venous thrombosis is associated with polymorphism in the *F5* and *F2* genes, as well as other risk factors, such as activated protein C resistance, which is a downstream effect of *F5* polymorphism. One can estimate the risk of future thrombotic events given data on one or more of these risk factors. However, it would be inappropriate to generate a combined posttest probability using data from both *F5* mutation testing and activated protein C resistance testing, because these 2 tests are closely related.

RECOMMENDED APPROACH TO TEST VALIDATION

In the text below, validation work has been artificially divided into 3 steps. Each successive step should be addressed in sequence and carefully documented.

Step 1: Planning Phase to Define the Requirements of the Test

The first step is to determine the intended use of the test, and then use this information to settle on the most appropriate test design and the expected performance characteristics. This step is often overlooked but is essential to designing the validation plan.

For an FDA-approved test, this step is straightforward. The intended use and the expected performance characteristics are stated. The goal of the validation plan would be to verify these, as discussed below in step 2, provided no aspects of the FDA-approved method are altered. For an LDT or modified FDA-approved test, this step can be quite complex. The first items to be addressed are the intended clinical utility (the benefits and risks associated with the use of the test in clinical practice) and the clinical indications (a list of the clinical setting[s] in which the test is appropriately ordered). Some aspects of clinical utility may be previously established if, for example, one desires to replace a traditional method, such as culture or antigen-based testing, with a more sensitive or specific molecular method. If the test is new to the laboratory and perhaps new to the practice of medicine (eg, *JAK2* mutation detection in myeloproliferative disorders), determining clinical utility requires careful evaluation of the literature. An approach to systematic review of the medical literature was recently published along with a 44-item checklist of items to consider when validating a new test.[29] Together, the clinical utility and the indications can be said to comprise the intended use of the test in clinical practice, which in turn requires addressing preanalytic and postanalytic factors, such as: (1) purpose of the test (eg, screening, diagnosis, prognosis, prediction, monitoring, confirmatory), (2) specimen types (eg, frozen tissue, blood, etc), (3) collection and handling procedures, (4) criteria for rejection, (5) target population, and (6) use of test results in patient management. At this stage it is also very important to communicate with clinicians who will be likely users of the test to (1) determine whether they are interested and ready to use the test if it were available, (2) determine how test results would alter patient care, and (3) ensure that the test method selected is likely to meet the performance criteria that are appropriate for the intended use of the test.

The intended medical use of the test, together with certain practical issues, is considered when setting the minimal acceptable performance criteria. Medical issues include such features as the risk to the patient of an incorrect result, the need for rapid turnaround time when timeliness is critical for patient management, and the availability of confirmatory tests. It is recommended that a pathologist or other laboratory physician be consulted regarding the medical aspects of assay design and validation. Practical issues may include: equipment and reagent needs and costs, total duration and hands-on technologist time, prior experience with a method, economy of scale and adaptability of the method for other applications, ease of obtaining reagents and support, and license requirements.

Another important component of validation study planning is determining what controls are necessary for initial and ongoing verification of test performance. These controls are included in each run of the validation study to accumulate data that will help set acceptable criteria for control performance. Controls act as surrogates for patient specimens in the quality control process. They exist in the appropriate matrix and are processed like a patient sample to monitor the ongoing performance of the entire analytical process (eg, specimen preparation, reverse transcription, amplification, interpretation, etc). Quantitative molecular tests typically include negative, low-positive, and high-positive controls in each run. Qualitative tests typically include a negative and a positive control in each run. For a qualitative test in which true positives may be close to the LoD or cutoff, a low-positive control should be included. To detect either poor extraction or the presence of an inhibitor that may interfere with hybridization or amplification in assays for which an undetectable result is a possibility, an endogenous or spiked exogenous control should be designed.

Once a putative test design is in place, pilot testing is done to gather data on the feasibility of the test to meet the minimal acceptable performance characteristics. For example, a molecular test may be intended to replace culture as a screening test for a pathogen. Such a test would likely need to detect a low level of pathogen, and pilot data can be used to estimate the LoD. Ideally, the acceptance criteria should be set before beginning the validation work so that goals are not compromised by what is initially achievable. Troubleshooting and optimization processes are used to refine the test procedure as needed. It should be noted that data collected during the design and optimization phase are helpful to estimate performance characteristics, but they are not generally used as validation data, because methods are continually being adjusted during this time. In fact, the operators who will collect the validation data may include the test developer, although it is wise to include at least one other operator who can comment on whether the protocol is clearly written and who runs samples in a blinded fashion. Before collecting validation data, new operators usually require a familiarization phase to become comfortable with the chosen instrument and methods.

Step 2: Generate Validation Data

The second step is to write the validation plan, generate the data, and summarize the data in a validation summary report. CLSI has published a number of guidelines for designing and carrying out test validation protocols. Many of these have been recognized by the FDA as consensus standards, and they are referenced at the FDA Web site http://www.fda.gov/cdrh/stdsprog.html (accessed May 23, 2008). The validation protocol provides details of the experiments that are performed to define the performance characteristics. The performance characteristics that need to be addressed are different for quantitative and qualitative tests and are addressed below.

For quantitative tests, the performance characteristics include accuracy, trueness, precision, linearity, analytic measurement range, reportable range, LoD, and LoQ. In addition, for LDTs or modified FDA-approved tests, interfering substances and clinical performance characteristics need to be assessed. Although these each can be addressed individually, they are interrelated and often addressed together.

Accuracy can be assessed in a number of ways. First, one can analyze samples that span the analytic measurement range and compare the measured values to the reference method. If a reference method is not available, accuracy may be assessed using a comparative method. Whereas a ''reference method'' is accepted as the ''gold standard'' method and therefore considered to be the best available approximation of the truth, a ''comparative method'' is potentially less accurate than the method being validated. In practice, methods that are considered the ''gold standard'' or reference method are often shown to be less accurate than the new method, and it is therefore recommended to avoid this term and instead use ''comparative method.'' Also, because comparative methods are potentially less accurate, it may be beneficial to use a combination of comparative methods.[30,31] Another approach is known as a ''recovery study,'' in which an analyte of known concentration is spiked into blank matrices (eg, plasma, whole blood), and recovery is determined. By performing multiple analyses of the same samples that cover a broad range of concentrations, trueness and precision can be determined (EP15-A2).[32] In addition, these data can be used to assess linearity (EP6-A).[19]

The LoD, LoQ and analytic measurement range can be estimated by extrapolation from the linearity data.[33] However, further studies would be required to establish these values. The ''classic approach'' to LoD determination is discussed in considerable detail in EP17-A.[33] Basically, the operator determines the distribution of values obtained from multiple measurements of blank samples (normal or negative controls). Values that are greater than the 95th percentile of the distribution are considered significantly different than the blank measurements. This cutoff value is the limit of the blanks. The objective then is to establish the concentration of analyte (LoD) that gives a value above the limit of the blanks 95% of the time.

The LoQ and the analytic measurement range depend on the acceptance criteria set by the laboratory director and reflect the clinical impact of erroneous results. Briefly, the director would set the limit for total error that is acceptable (eg, 15% error for 95% of measurements). Multiple replicates at a given concentration would be used to determine bias and SD to estimate total error at that concentration. If the total error (bias ± 2SD) was less than the acceptance limit, the criteria would be met, as discussed above in relation to Figure 3. It is important to note that accuracy, linearity, LoD, and LoQ determinations should be done in a biologically appropriate matrix that simulates clinical samples. In addition, certain matrix elements could be a source of interfering substances and should be assessed by comparing test results with and without these interfering substances (EP7-A2).[34] Evaluating cross-reactivity would be test specific and might involve spiking closely related analytes into normal or negative controls.

For qualitative tests, the performance characteristics that need to be assessed include positive and negative agreements compared with an established method, as well as reproducibility. In addition, for laboratory-developed and modified FDA-approved tests, interfering substances and clinical performance characteristics need to be assessed. If these performance characteristics are assessed using patient samples in which the diagnoses are known, the results demonstrate clinical performance, and this terminology should be used. If the diagnosis is unknown, results should be expressed as positive and negative agreements to a comparative test, as described (EP12-A).[35] Assessment of reproducibility involves multiple analyses using the same samples to demonstrate consistency. When applicable, LoD can only be determined by trial and error (ie, testing a series of concentrations), and it is usually set at the concentration at which informative results are obtained at least 95% of the time. Interfering substances could be assessed by spiking positive and negative samples with possible confounding matrix elements or analytes. It should be noted that genomic DNA itself is a potential interfering substance that can negatively affect the limit of detection or cause false-positive detection.

For semiquantitative tests in which quantitative data are used to produce a qualitative result, the performance characteristics that need to be addressed are essentially the same as for qualitative tests. However, LoD is typically determined in a manner similar to a quantitative test. In addition, the comparison to another method or clinical diagnosis can be performed by setting multiple thresholds and generating a receiver-operator characteristic curve, as demonstrated in Figure 2.

The perennial question is, ''How many samples do I need to run to validate a given test?'' Unfortunately, the answer is always the same—it depends. It depends on whether the test is intended as a new test or to replace a comparative test, how the test is to be used, which performance criteria are most critical for the intended use, and the confidence level that is required for good medical practice, implying that medical judgment is required. For a new test, the goal is to establish the parameters together with their confidence intervals. Jones et al[36] provide a nice example of how to calculate the number of samples needed given predetermined lowest acceptable sensitivity, specificity, and confidence intervals. If the test is to be compared to an established method to show equivalence, one must determine the allowable analytic error as well as the performance difference that is considered clinically significant.[13,36] Jones et al[36] provide examples to compare one study group to another for both continuous normally distributed data as well as categorical data, and these types of calculations can be used to estimate sample size for comparing test parameters. In addition, statistical tools for power and sample size determinations are freely available (www.statpages.org/#Power; accessed May 28, 2008), and the CLSI guidelines provide estimates on the number of specimens and days of evaluation that are necessary to obtain statistically meaningful data. However, these are only estimates, and the reader is cautioned that much depends on how the test will be used and how much confidence is required before moving forward. Consultation with a biostatistician and a pathologist or other laboratory physician may be helpful.

Naturally, the extent of work is more for manufacturers (including those of LDTs) than it is for verifying the performance characteristics of an established test, as shown in Table 3. CLSI guideline EP15-A2[32] provides advice on how to verify accuracy, whereas EP9-A2[37] and EP5-A2[38] advise how to establish accuracy and precision for a novel test. The basic evaluation protocol is the same, but the statistical analysis for test verification is different (as described in detail in EP15-A2[32]). To verify precision, the guidance advises measuring 2 concentration levels in triplicate over 5 days, and a variance is then determined. If this variance is less than that stated by the manufacturer, this performance characteristic is considered verified. If it is greater, then one would determine whether there is a statistically significant difference using χ^2 distribution for a given α error. Bias is similarly verified by determining the average bias relative to a comparative test, setting an α error and using the *t* distribution. On the other hand, to establish an estimate of precision, one would increase the number of analyses and report the obtained standard deviation and coefficient of variance (EP5-A2).[38] To provide an estimate of trueness, one would report the bias together with the confidence intervals (EP9-A2).[37]

It should be noted that these protocols can be used separately or combined to derive a more efficient evaluation protocol. For example, EP9-A2 can be combined with EP5-A2 to simultaneously evaluate bias and imprecision. Indeed, EP10-A2[39] is intended to evaluate several parameters simultaneously to arrive at estimates of precision, bias, and linearity. However, this protocol is not intended to be used to establish these parameters or even to verify these parameters, but only as a quick cursory check that there are no major issues that need to be addressed before proceeding with validation. Still, some laboratory directors may choose to expand this protocol to acquire more data and increase the statistical power so that it may be suitable to verify or validate performance characteristics. The advantage of such an approach is that it is more efficient than addressing these parameters individually. The disadvantage is that it requires a clear understanding of the statistics involved, and observed deviations from acceptance limits may be due to imprecision, bias, or lack of linearity. Further evaluation would be required to quantify the error associated with each parameter. As another example, a laboratory director may effectively argue that within-laboratory precision (ie, between technologists, machines, lots, days) is within acceptance limits, so there is little need to evaluate within-run precision. However, if the acceptable limits are exceeded, then studies of within-run precision may shed light on the sources and extent of error.

Table 3. Relevant Clinical and Laboratory Standards Institute (CLSI) Evaluation Protocols (EPs) From Among the Many Consensus Standards That Have Been Recognized by the US Food and Drug Administration[a]

Title	CLSI Document	Scope	Estimated No. of Specimens and Days
Evaluation of Precision Performance of Quantitative Measurement Methods; Approved Guideline—Second Edition	EP5-A2	Includes protocols for a developer of test methods or devices, and protocols for users of these methods who wish to determine their own performance capabilities.	Twenty operating days are recommended for the precision evaluation experiment. Because day-to-day imprecision may be large, performance must be evaluated long enough to ensure that the separate components of error are adequately covered by the experiment.
Method Comparison and Bias Estimation Using Patient Samples; Approved Guideline—Second Edition	EP9-A2	Includes protocols for a developer of test methods or devices, and protocols for users of these methods who wish to determine their own performance capabilities.	At least 40 patient samples, preferably spread over many days and runs.
Preliminary Evaluation of Quantitative Clinical Laboratory Methods; Approved Guidelines—Second Edition	EP10-A2	Preliminary evaluation of linearity, bias, linear drift, carry-over, and precision.	Low- and high-range pools of analyte (controls or samples) and a mixture of these to make a midrange pool. At least one run per day for 5 days.
User Protocol for Evaluation of Qualitative Test Performance; Approved Guideline	EP12-A	Written for end users and includes protocols for the demonstration of qualitative test performance.	At least 50 positive and 50 negative specimens, daily for 10 to 20 days.
User Demonstration of Performance for Precision and Accuracy; Approved Guideline	EP15-A2	Intended for verification of precision and accuracy.	For precision, 5 days of 2 concentrations in triplicate. For bias, 20 patient samples in duplicate over several days.

[a] Additional information is found at www.fda.gov/cdrh/stdsprog.html, and the CLSI documents can be obtained through the CLSI Web site www.clsi.org/.

As an example of verification of an FDA-approved quantitative test, one could make a series of 4 or 5 controls using an appropriate matrix to span the analytic measurement range and include the LoD/LoQ. On the first day, one should do 2 runs of a group of 4 samples: a blank (in appropriate matrix), an LoD control and another control run as an unknown, and a clinical sample. On days 2 through 10, one should repeat but use 2 other clinical samples and other controls on each day so that at the end of 10 days, one would have data on 20 blanks and 20 LoDs to verify your LoD (EP17-A),[33] one would have 20 patient results to be judged against one's comparative test and to verify one's reference range (EP15-A2),[32] and one would have 5 values for each control to verify within-lab precision and trueness and to verify one's linearity and analytic measurement range (EP6-A and EP17-A).[19,33] Furthermore, a comparison of expected to actual control results may reveal any unanticipated trends related to how the calibrators were stored between runs (like freeze-thaw effects; EP14-A2).[40] For an LDT, one could take essentially the same approach, except one could also include other patient samples to identify interfering substances (including interference due to a spiked control), and the study would be extended to 20 days so that twice the data would be accumulated (to meet the minimum requirements of EP5-A2[38] and EP09-A2[37]). Also note that some parameters are not addressed with this protocol, like clinical sensitivity and clinical specificity, which would require a review of the literature and perhaps more extensive studies, chart reviews, etc.

For a qualitative FDA-approved test (eg, Roche's Factor V Leiden Kit), one could analyze a series of clinical samples that would sufficiently cover the reportable range (eg, homozygotes, heterozygotes, normal) and evaluate these against a comparative method (eg, sequencing) or reference samples obtained from a laboratory with a validated procedure. The CLSI evaluation protocol for evaluation of qualitative test performance (EP12-A[35]) is intended for end users of such tests and recommends a minimum of 50 positive and 50 negative specimens over 10 to 20 days. In addition, some samples would need to be repeated to demonstrate reproducibility. For an LDT, additional samples should be run, and the number depends on the degree of confidence that is required, which can be estimated as described in EP12-A.[35] Interfering substances and clinical performance characteristics also need to be assessed. If the diagnosis is known, these data could be used to determine clinical sensitivity and specificity as well as other parameters, like the positive and negative likelihood ratios. Positive and negative predictive values depend on the prevalence in the population, which could be extracted from the medical literature when appropriate. If the diagnosis is not known, a careful review of the literature or chart reviews would be required to estimate clinical performance characteristics.

A well-designed validation protocol includes consideration of possible outcomes and calculates the expected confidence intervals around each performance characteristic. Note that it is possible to extend the validation data acquisition to obtain the necessary confidence interval as long as the test is not modified in the process. If the experimental data fail to meet the minimal acceptance criteria, consider whether the protocol should be modified followed by further iterations of the evaluation protocol.

Generation of statistically significant validation data can seem like a daunting task, especially when dealing with rare diseases for which few clinical samples are available.[16] Therefore, the laboratory director (as well as the clinical consultant, if they are different people) must combine medical judgment and technical knowledge to resolve discrepancies and to determine when sufficient validation work has been done. The physician(s) in every CLIA-approved laboratory should take medicolegal responsibility for the patient care done in their facility, and they are likewise commissioned to determine what is appropriate for their patients, including when to introduce a new labo-

Table 4. Components of a Validation Report

Specimens tested: Represent each of the possible reportable results (eg, normal and mutant) or span the range of results (eg, low to high values), plus a reasonable number and representative distribution of specimen types (eg, blood, marrow).
Method comparison: Compare with another valid test, such as another test method, or sample exchange with a laboratory performing a similar test.
Analytic performance characteristics: Sensitivity, specificity, and (for quantitative tests) accuracy, precision and reproducibility, linearity, and analytic measurement range.
Define: Reference values (normal range), reportable range, acceptable sample types and criteria for rejection, recommended reagents/instruments, controls, standards, calibrators, step-by-step procedure, interpretation, and reporting guidelines.
This validation study has been reviewed, and the test is approved for clinical use (director signature and date).

Table 5. Example of an Implementation Checklist

Validation summary completed and approved by director
MSDS book updated
Assay procedure written and reviewed
Necessary worksheets written and reviewed
Training checklist developed
Staff trained on performing assay
Competency of staff documented
Equipment maintenance scheduled (if necessary)
New equipment approved by engineering (if necessary)
Billing codes set up
Test codes set up and requisition created
Laboratory directory updated
Specimen receiving notified
Clinicians notified
Proficiency testing scheduled

Abbreviation: MSDS, Material Safety Data Sheet.

ratory test. It should be recognized that molecular pathology is a rapidly evolving discipline and that full understanding of any new test may take years (or even decades) of work by multiple clinical and basic investigators, especially when long-term outcome studies are involved. Molecular tests are therefore introduced into clinical practice when there is adequate but perhaps not a full understanding of the clinical utility of the test. Incremental progress is certainly allowable, and indeed desirable. As a general guide, a new laboratory test is ready to be introduced when it is demonstrably equivalent to or better than what was already available for supporting the care of a patient. The word ''better'' includes not only assay performance characteristics but also such factors as turnaround time, cost, and less invasive specimen collection. Although supporting evidence is essential to scientific understanding, it must be recognized that statistically significant data may not be available for every aspect of every validation study. There is simply not enough patient material, financial support, or scientist/technologist time to support such rigorous data collection. Therefore, medical judgment in the context of the in-laboratory data and the larger health care system is essential to deciding when a test is ready to be introduced for patient care.

After collecting all of the necessary validation data, a validation summary report is written (Table 4). This report recaps the details of the validation work, especially the aspects that would be of interest to inspectors and clinicians, such as evaluation of analytic performance characteristics, clinical sensitivity and specificity, and positive and negative predictive values in the target population(s). Work done to determine acceptable specimen types, collection and handling optimization, interfering substances, and analytic interpretation are also described. Literature references are provided to complement the local validation study findings and to fill certain gaps for which direct evidence is lacking. At the end of this report should be a statement signed by the laboratory director that this test is (or is not) ready for clinical implementation.

Step 3: Implement the Test

Implementation requires incorporating the test into the workflow of the laboratory and the completion of several documents. The number of details required makes an implementation checklist quite useful (Table 5). A complete procedure must be written and signed by the laboratory director, and it should include such elements as indications, intended use, principle of the test, specimen handling and storage, reagents, controls, equipment, step-by-step procedure, guidance for analytic and clinical interpretation, and references. Other documents might include bench worksheets and reporting templates.

Before implementing a test, there needs to be a procedure for quality control and quality assurance. This procedure should address preanalytic, analytic, and postanalytic issues, many of which may be general (eg, procedure for dealing with an inappropriate specimen), but some of which may be specific to the test system (eg, equipment maintenance). Included in this procedure should be important quality control and quality assurance indicators, such as turnaround time, documentation of control failures or test failures, and trends in test volumes and results. For example, the number of factor V Leiden–positive patients may be tracked over time, and if the patient population remains constant, the proportion of positive results should be rather constant. A significant change in volumes or positive results would prompt an investigation.

Designing the calibration scheme is an aspect of the quality control and quality assurance procedure for quantitative or semiquantitative tests. A calibrator is defined by CLSI as a ''substance, material, or article intended by its manufacturer to be used to establish the measurement relationships of an in vitro diagnostic medical device.'' A related term is *reference material*, which is a substance whose properties are sufficiently homogeneous and well established to be used for the calibration of an apparatus, the assessment of a measurement method, or for assigning values to materials (as paraphrased from CLSI), and CLSI adds the note that a given reference material may be used either as a calibration material to calibrate a measuring system and assign values to materials, or as a control material in assessing the performance of a measurement procedure, but it must not be used for both purposes in a given situation in a particular laboratory.[41] This is because the calibrators are considered part of the test system, and the controls are meant to be external to the test system so as to evaluate the test system as a whole (including the calibrators). A control that is external to the test system may thus be used simultaneously as a control and for calibration verification. Ideally, calibrators and materials for calibration verification are in a matrix appropriate for the clinical specimens tested. Calibration verification must be performed at least every 6 months and upon service events or reagent lot changes.

Several other steps are essential and should not be overlooked, such as enrollment in a proficiency testing program or identification of an alternate form of external quality control, and a training checklist with competency documentation for personnel who will perform the test. Other steps tend to be institution dependent but are also necessary and include coding, billing, and making the test known to and orderable by clinicians.

Once a test is introduced into clinical practice, it behooves the laboratorian to keep abreast of new technologies and ongoing research related to the test and the medical condition(s), so that the test may be withdrawn if necessary or improved when appropriate (with an associated validation study for the altered procedure).

SUMMARY

Test validation is critical to the practice of evidence-based laboratory medicine. Appropriate effort is needed to successfully verify or validate a test prior to clinical laboratory implementation. Although the task can seem overwhelming at times, it is encouraging to know that a laboratory's efforts assure patients and the public that we in the health care system are acting in their best interests. Laboratory professionals should take pride in their work to improve accessibility and to advance progress in laboratory medicine.

References

1. Rabenau HF, Kessler HH, Kortenbusch M, Steinhorst A, Raggam RB, Berger A. Verification and validation of diagnostic laboratory tests in clinical virology. *J Clin Virol.* 2007;40:93–98.

2. Medical devices; classification/reclassification; restricted devices; analyte specific reagents—FDA. Final rule. *Fed Regist.* 1997;62:62243–62260.

3. Medicare, Medicaid and CLIA programs; regulations implementing the Clinical Laboratory Improvement Amendments of 1988 (CLIA)—HCFA. Final rule with comment period. *Fed Regist.* 1992;57:7002–7186.

4. International Organization for Standardization. *Quality Management Systems—Fundamentals and Vocabulary.* Geneva, Switzerland: International Organization for Standardization; 2005. ISO 9000.

5. Food and Drug Administration. *Guideline on General Principles of Process Validation.* Rockville, MD: Food and Drug Administration; 1987.

6. Christenson RH. Committee on Evidence Based Laboratory Medicine of the International Federation for Clinical Chemistry Laboratory Medicine. Evidence-based laboratory medicine—a guide for critical evaluation of in vitro laboratory testing. *Ann Clin Biochem.* 2007;44:111–130.

7. Cecconi M, Forzano F, Rinaldi R, et al. A single nucleotide variant in the FMR1 CGG repeat results in a "Pseudodeletion" and is not associated with the fragile X syndrome phenotype. *J Mol Diagn.* 2008;10:272–275.

8. Gulley ML, Braziel RM, Halling KC, et al. Clinical laboratory reports in molecular pathology. *Arch Pathol Lab Med.* 2007;131:852–863.

9. Rozet E, Ceccato A, Hubert C, et al. Analysis of recent pharmaceutical regulatory documents on analytical method validation. *J Chromatogr A.* 2007; 1158:111–125.

10. International Organization for Standardization. *Statistics—Vocabulary and Symbols—Part 1: General Statistical Terms and Terms Used in Probability.* Geneva, Switzerland: International Organization for Standardization; 2006. ISO 3534-1.

11. International Organization for Standardization. *Laboratory Medicine—Requirements for Reference Measurement Laboratories.* Geneva, Switzerland: International Organization for Standardization; 2003. ISO 15195.

12. Hubert P, Nguyen-Huu JJ, Boulanger B, et al. Harmonization of strategies for the validation of quantitative analytical procedures. A SFSTP proposal—part I. *J Pharm Biomed Anal.* 2004;36:579–586.

13. Association for Molecular Pathology statement. Recommendations for in-house development and operation of molecular diagnostic tests. *Am J Clin Pathol.* 1999;111:449–463.

14. Forbes BA. Introducing a molecular test into the clinical microbiology laboratory: development, evaluation, and validation. *Arch Pathol Lab Med.* 2003; 127:1106–1111.

15. Dimech W, Bowden DS, Brestovac B, et al. Validation of assembled nucleic acid-based tests in diagnostic microbiology laboratories. *Pathology.* 2004; 36:45–50.

16. Maddalena A, Bale S, Das S, Grody W, Richards S. Technical standards and guidelines: molecular genetic testing for ultra-rare disorders. *Genet Med.* 2005;7:571–583.

17. Wiktor AE, Van Dyke DL, Stupca PJ, et al. Preclinical validation of fluorescence in situ hybridization assays for clinical practice. *Genet Med.* 2006;8: 16–23.

18. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Standards for Reporting of Diagnostic Accuracy. *Clin Chem.* 2003;49:1–6.

19. National Committee on Clinical Laboratory Standards. *Evaluation of Linearity of Quantitative Measurement Procedures: A Statistical Approach; Approved Guideline.* Wayne, PA: National Committee on Clinical Laboratory Standards; 2003. NCCLS document EP6-A.

20. van Gils MP, Cornel EB, Hessels D, et al. Molecular PCA3 diagnostics on prostatic fluid. *Prostate.* 2007;67:881–887.

21. IVD manufacturers, labs should be subject to same standards—AdvaMed. *The Gray Sheet: Medical Devices, Diagnostics & Instrumentation.* 2004;30:5.

22. Hadgu A, Dendukuri N, Hilden J. Evaluation of nucleic acid amplification tests in the absence of a perfect gold-standard test: a review of the statistical and epidemiologic issues. *Epidemiology.* 2005;16:604–612.

23. Martin DH, Nsuami M, Schachter J, et al. Use of multiple nucleic acid amplification tests to define the infected-patient "gold standard" in clinical trials of new diagnostic tests for *Chlamydia trachomatis* infections. *J Clin Microbiol.* 2004;42:4749–4758.

24. Baughman AL, Bisgard KM, Cortese MM, Thompson WW, Sanden GN, Strebel PM. Utility of composite reference standards and latent class analysis in evaluating the clinical accuracy of diagnostic tests for pertussis. *Clin Vaccine Immunol.* 2008;15:106–114.

25. Fischer JE, Bachmann LM, Jaeschke R. A readers' guide to the interpretation of diagnostic test properties: clinical example of sepsis. *Intensive Care Med.* 2003;29:1043–1051.

26. Simel DL, Samsa GP, Matchar DB. Likelihood ratios with confidence: sample size estimation for diagnostic test studies. *J Clin Epidemiol.* 1991;44:763–770.

27. Ogino S, Wilson RB, Gold B, Flodman P. Bayesian risk assessment in genetic testing for autosomal dominant disorders with age-dependent penetrance. *J Genet Couns.* 2007;16:29–39.

28. Chu K, Brown AF. Likelihood ratios increase diagnostic certainty in pulmonary embolism. *Emerg Med Australas.* 2005;17:322–329.

29. Gudgeon JM, McClain MR, Palomaki GE, Williams MS. Rapid ACCE: experience with a rapid and structured approach for evaluating gene-based testing. *Genet Med.* 2007;9:473–478.

30. Alonzo TA, Pepe MS. Using a combination of reference tests to assess the accuracy of a new diagnostic test. *Stat Med.* 1999;18:2987–3003.

31. Hawkins DM, Garrett JA, Stephenson B. Some issues in resolution of diagnostic tests using an imperfect gold standard. *Stat Med.* 2001;20:1987–2001.

32. National Committee on Clinical Laboratory Standards. *User Verificaton of Performance for Precision and Trueness; Approved Guideline—Second Edition.* Wayne, PA: National Committee on Clinical Laboratory Standards; 2005. NCCLS document EP15-A2.

33. National Committee on Clinical Laboratory Standards. *Protocols for Determination of Limits of Detection and Limits of Quantitation; Approved Guideline.* Wayne, PA: National Committee on Clinical Laboratory Standards; 2004. NCCLS document EP17-A.

34. Clinical and Laboratory Standards Institute. *Interference Testing in Clinical Chemistry; Approved Guideline—Second Edition.* Wayne, PA: Clinical and Laboratory Standards Institute; 2007. CLSI document EP7-A2.

35. National Committee on Clinical Laboratory Standards. *User Protocol for Evaluation of Qualitative Test Performance; Approved Guideline.* Wayne, PA: National Committee on Clinical Laboratory Standards; 2002. NCCLS document EP12-A.

36. Jones SR, Carley S, Harrison M. An introduction to power and sample size estimation. *Emerg Med J.* 2003;20:453–458.

37. National Committee on Clinical Laboratory Standards. *Method Comparison and Bias Estimation Using Patient Samples; Approved Guideline—Second Edition.* Wayne, PA: National Committee on Clinical Laboratory Standards; 2002. NCCLS document EP9-A2.

38. National Committee on Clinical Laboratory Standards. *Evaluation of Precision Performance of Quantitative Measurement Methods; Approved Guideline—Second Edition.* Wayne, PA: National Committee on Clinical Laboratory Standards; 2004. NCCLS document EP5-A2.

39. National Committee on Clinical Laboratory Standards. *Preliminary Evaluation of Quantitative Clinical Laboratory Methods; Approved Guidelines—Second Edition.* Wayne, PA: National Committee on Clinical Laboratory Standards; 2002. NCCLS document EP10-A2.

40. National Committee on Clinical Laboratory Standards. *Evaluation of Matrix Effects; Approved Guideline.* Wayne, PA: National Committee on Clinical Laboratory Standards; 2003. NCCLS document EP14-A.

41. National Committee on Clinical Laboratory Standards. *Preparation and Validation of Commutable Frozen Human Serum Pools as Secondary Reference Materials for Cholesterol Measurement Procedures; Approved Guideline.* Wayne, PA: National Committee on Clinical Laboratory Standards; 1999. NCCLS document C37-A.