Developing Clinical Evidence for Regulatory and Coverage Assessments in *In Vitro* Diagnostics (IVDs)

A Framework for Developing Credible Evidence of Analytical Validity, Clinical Validity, and Clinical Utility for IVDs

A Report of the IVD Clinical Evidence Working Group
of the Medical Device Innovation Consortium (MDIC)

August 13, 2019

AUTHORS

MDIC has assembled a working group comprised of member organizations and other subject matter experts to guide work on this project.

MDIC Clinical Evidence Working Group and Advisors

Susan Alpert, MD, PhD | Expert Advisor
Vicki Anastasi | ICON (Working Group Chair)
Naomi Aronson, PhD | Expert Advisor
Amy Durtschi, PhD | Abbott
Tremel Faison, MS | BARDA
Rochelle Fink, MD, JD, MBA | FDA - CDRH - OCD
Alberto Gutierrez, PhD | Expert Advisor
Jaime Houghton, MS | Sysmex
Louis Jacques, MD | Expert Advisor
Marina Kondratovich, PhD | FDA-CDRH-OIR
Fred Lasky, PhD | Expert Advisor
Maya Mahue, PhD | Hologic
Greg Payne | BD
Susan Piotrowski | Abbott
Mike Reiner, MBA | Hologic
Sandra Statz, MS | Exact Sciences
Timothy Stenzel, MD, PhD | FDA – CDRH - OIR
Lee Termini | ICON
Songbai Wang, MD, MSPH | Johnson & Johnson
Carolyn Hiller, MBA | MDIC Clinical Diagnostics Program Director

EXECUTIVE SUMMARY

A Framework for Developing Credible Evidence of Analytical Validity, Clinical Validity, and Clinical Utility for* In Vitro *Diagnostics (IVDs)

Background on the MDIC Clinical Diagnostics Program
The Medical Device Innovation Consortium (MDIC) is a 501(c)(3) non-profit organization and the first public-private partnership with a mission to advance regulatory science in the medical device industry. MDIC coordinates the development of methods, tools, and resources used in managing the total product life cycle of a medical device to improve patient access to cutting-edge medical technology.

The MDIC Clinical Diagnostics Program develops new tools and methods that will improve processes to assess safety, effectiveness, and the value proposition of diagnostic tests.

Purpose of the Clinical Evidence Framework
This Framework is intended to help IVD manufacturers make decisions on how to develop credible evidence of analytical and clinical validity, as well as clinical utility.

This Framework should be considered an initial thought piece and not a prescriptive, "how-to" guide.

Reading and following this document neither guarantees FDA approval/clearance nor payment from insurance companies.

Overview of the Framework
The Framework is organized into five sections, outlined below:

Section One: Introduction – introduces the Framework, definitions of analytical validity, clinical validity, and basic concepts of FDA clearance and approval of medical devices, including IVDs.

Section Two: Analytical Validity – provides a list of studies frequently used to demonstrate analytical validity. Useful terms, test considerations and requirements, and related references are provided.

Section Three: Clinical Validity – discusses assay types and measures of clinical validation, as well as design considerations for clinical validation based on the intended use of the IVD.

Section Four: Clinical Utility – explores the general strategy for developing evidence of clinical utility for payers, presents a clinical utility "self-assessment" framework that IVD developers can use for planning, and details additional clinical utility considerations by IVD test type.

Section Five: References – lists regulatory documents not provided in-line with the text and other references cited in this Framework.

TABLE OF CONTENTS

1 Introduction

1.1 Scope of the Framework

For an In Vitro Diagnostic (IVD), analytical and clinical validity are required for the United States Food and Drug Administration (FDA) clearance/approval, and clinical utility is required for payer coverage and provider adoption. Many sponsors collect analytical and clinical validity data first to facilitate approval/clearance through the FDA. But, waiting to collect clinical utility data until after a regulatory approval/clearance can lead to duplicate data collection efforts, changes in the intended use (and a resultant increase in FDA regulatory burden), a delay in market entry, or limited coverage. Therefore, defining pathways to collect clinical utility data in parallel with analytical and clinical validity data can potentially increase the efficiency and reduce costs of IVD development.

The MDIC IVD Clinical Evidence Framework provides a high-level overview that IVD manufacturers can use to make decisions on how to develop credible evidence of analytical and clinical validity, as well as clinical utility. The MDIC IVD Clinical Evidence Working Group hopes that this Framework will help IVD manufacturers develop a strategy to generate clinical evidence for regulatory and reimbursement purposes in parallel.

Audience:
We believe this Framework will be useful for individuals and sponsors new to the diagnostic space and for those with intent to develop a new IVD.

Contents

The Framework is organized into five sections, outlined below:

Section One: Introduction – introduces the Framework, definitions of analytical validity, clinical validity, and basic concepts of FDA clearance and approval of medical devices, including IVDs.

Section Two: Analytical Validity – provides a list of studies frequently used to demonstrate analytical validity. Useful terms, test considerations and requirements, and related references are provided.

Section Three: Clinical Validity – discusses assay types and measures of clinical validation, as well as design considerations for clinical validation based on the intended use of the IVD.

Section Four: Clinical Utility – explores the general strategy for developing evidence of clinical utility for payers, presents a clinical utility "self-assessment" framework that IVD developers can use for planning, and details additional clinical utility considerations by IVD test type.

Section Five: References – lists regulatory documents not provided in-line with the text and other references cited in this Framework.

It is important to note that reading and following this document neither guarantees FDA approval/clearance nor payment from insurance companies.

1.2 Definitions

In this Framework, we have chosen to define analytical validity, clinical validity, and clinical utility as follows:

- Analytical Validity – the ability of a test to accurately and reliably measure or detect the analyte(s) of interest in specimens that are representative of specimens that would be obtained in the intended use population.

  Analytical validity is demonstrated by evidence that the assay results are accurate, repeatable and reproducible (precision), specific (minimal impact of bias-causing substances or conditions), and sensitive (limit of detection and limit of quantification if appropriate).

  References:
  CLSI. A Framework for Using CLSI Documents to Evaluate Clinical Laboratory Measurement Procedures. 2nd ed. CLSI report EP19. Wayne, PA: Clinical and Laboratory Standards Institute; 2015

  *Evaluation of Genomic Applications in Practice and Prevention (**EGAPP™**) Working Group*[1]

- Clinical Validity – the ability of a test to accurately and reliably predict the clinically defined disorder or phenotype of interest.

  For example, for binary qualitative tests, clinical validity encompasses three variables:

  - *Clinical Sensitivity* – The percentage of individuals with the target condition (disease) that will have positive test results.
  - *Clinical Specificity* – The percentage of individuals that do not have the target condition who will have negative test results.
  - *Predictive values of positive and negative test results that combine positive and negative likelihood ratios with target condition prevalence* – The proportion of individuals in the intended use population who have the target condition.

  Reference: FDA Guidance "Design Considerations for Pivotal Clinical Investigations for medical Devices", 2013

- Clinical Utility – the ability of a test to meaningfully improve patient health outcomes, when used to inform and support clinical decisions that increase the likelihood of improved patient outcomes, compared to decisions that would be made without the test results.

  Clinical utility will vary with each new IVD based on its intended use, existing tests, and the payer. Although it is common to separate the terms clinical validity and clinical utility, a clear-cut separation is not always possible since the clinical validity of the tests depends on the intended use or the claims made on the use of the test.

  References:
  Grosse SD, Khoury MJ. What is the clinical utility of genetic testing? Genet Med. 2006;8(7):448-450[2]

Teutsch SM, Bradley LA, Palomaki GE, et al. The Evaluation of Genomic Applications in Practice and Prevention (EGAPP) Initiative: methods of the EGAPP Working Group. Genet Med. 2009;11(1):3-14.[3]

A study designed to assess analytical validity, clinical validity, or clinical utility may not necessarily be named as such, despite having one of those objectives. To clarify this variation in nomenclature, we have created a table that outlines IVD study objectives and then lists the terms used for studies that may be required to support the objectives (Table 1.1). The list of studies necessary is not exhaustive and different tests may require more evaluations.

Table 1.1 IVD Study Types Classified by Study Objective

Study Objective	Terms Used	Examples (Not a Complete List)
Evaluation of how well the test can measure (or detect) a measurand of interest (e.g., target analyte)	**Analytical Validity**	• Evaluation of precision • Evaluation of cross-reactivity • Evaluation of measuring interval • Evaluation against a comparator device to determine accuracy (trueness) and/or analytical sensitivity in detecting a specific biomarker/measurand/analyte
Evaluation of the ability of the test to classify a patient into disease or prognosis category	• **Clinical Validity** • **Clinical Performance**	• Estimation of measures of clinical performance • For a binary qualitative test, these measures are clinical sensitivity and specificity, positive and negative likelihood ratios, positive and negative predictive values for prevalence • For a qualitative test with multiple outputs, these measures are pre-test risk, post-test risks for each output, likelihood ratio for each output, and percent of patient with the output for each output

Evaluation of the ability of the test to direct clinical management and potentially improve patient outcomes	• **Diagnostic Thinking Efficacy** • **Benefit-Risk Analysis** • **Clinical Validity/Clinical Utility**	• Impact on clinician judgment about diagnosis/prognosis • Impact on the choice of management
Evaluation of the ability and magnitude of the test to improve patient outcomes	**Clinical Utility**	Impact on mortality or morbidity
Evaluation of the test to benefit society as a whole	**Societal Efficacy/Clinical Utility**	Cost-effectiveness analysis

1.3 Basic Concepts of FDA Approval/Clearance of Medical Devices

In the United States, the FDA regulates medical devices, which includes IVDs. In order to approve or clear a device for market entry, the FDA must determine that the device is safe and effective relying only upon valid scientific evidence. A device is safe when there is reasonable assurance that the probable benefits to health outweigh the probable risks. It is effective when there is reasonable assurance that in a significant proportion of the intended use population the device will provide clinically significant results [21 CFR 860.7].

1.3.1 Safety and Effectiveness

The United States Food, Drug, and Cosmetic (FD&C) Act section 513(2)[4] states:

> *The safety and effectiveness of a device are to be determined---*
> *(A) with respect to the persons for whose use the device is represented or intended,*
> *(B) with respect to the conditions of use prescribed, recommended, or suggested in the labeling of the device, and*
> *(C) weighing any probable benefit to health from the use of the device against any probable risk of injury or illness from such use.*

In addition, FDA has interpreted the statutory standard (FD&C Act) through regulation as follows:

- *21 CFR 860.7(d)(1).*[5] *There is reasonable assurance that a device is safe when it can be determined, based upon valid scientific evidence, that the probable benefits to health from use of the device for its intended uses and conditions of use, when accompanied by adequate directions and warnings against unsafe use, outweigh any probable risks.*

  The valid scientific evidence used to determine the safety of a device shall adequately demonstrate the absence of unreasonable risk of illness or injury associated with the use of the device for its intended uses and conditions of use.

- *21 CFR 860.7(e)(1).*[5] *There is reasonable assurance that a device is effective when it can be determined, based upon valid scientific evidence, that in a significant portion of the target population, the use of the device for its intended uses and conditions of use, when accompanied by adequate directions for use and warnings against unsafe use, will provide clinically significant results.*

Reasonable assurance of device safety and effectiveness must be supported by valid scientific evidence. The evidence must support that the device, when accompanied by the appropriate device labeling, will provide clinically meaningful results in the target population. Further, a determination is based on balancing the probable benefit to health with any probable risk of injury or illness resulting from device use.

1.3.2 Valid Scientific Evidence

Valid scientific evidence is defined through regulation as follows:

- *21 CFR 860.7(c)(2).*[5] *Valid scientific evidence is evidence from well-controlled investigations, partially controlled studies, studies and objective trials without matched controls, well-documented case histories conducted by qualified experts, and reports of significant human experience with a marketed device, from which it can fairly and responsibly be concluded by qualified experts that there is reasonable assurance of the safety and effectiveness of a device under its conditions of use.*

  The evidence required may vary according to the characteristics of the device, its conditions of use, the existence and adequacy of warnings and other restrictions, and the extent of experience with its use.

  Isolated case reports, random experience, reports lacking sufficient details to permit scientific evaluation, and unsubstantiated opinions are not regarded as valid scientific evidence to show safety or effectiveness. Such information may be considered, however, in identifying a device the safety and effectiveness of which is questionable.

FDA regulations also consider which types of evidence support reasonable assurance of safety and effectiveness:

- *21 CFR 860.7(d)(2).*[5] *Among the types of evidence that may be required, when appropriate, to determine that there is reasonable assurance that a device is safe are investigations using laboratory animals, investigations involving human subjects, and nonclinical investigations including in vitro studies.*

- *21 CFR 860.7(e)(2).*[5] *The valid scientific evidence used to determine the effectiveness of a device shall consist principally of well-controlled investigations, as defined in [21 CFR 860.7(f)], unless [FDA] authorizes reliance upon other valid scientific evidence which [FDA] has determined is sufficient evidence from which to determine the effectiveness of a device, even in the absence of well-controlled investigations. [FDA] may make such a determination where the requirement of well-controlled investigations in [21 CFR 860.7(f)] is not reasonably applicable to the device.*

Evidence of effectiveness of a medical device must generally be obtained from well-controlled studies (as described in 21 CFR 860.7(f)). However, the regulations provide FDA with some flexibility regarding its determination of the type of evidence that may be considered valid scientific evidence to demonstrate the safety of a medical device.

1.3.3 Benefit-Risk Assessment

FDA's mission is to protect the public health by ensuring the safety, efficacy, and security of all regulated products.[6]

21 CFR 860.7(b)(3).[5] states that, in determining the safety and effectiveness of a device, FDA must weigh "the probable benefit to health from the use of the device…against any probable injury or illness from such use." This concept is often referred to as the risk-benefit assessment.

Probable benefit to health refers to the benefit(s) to a subject's health that results from the use of the medical device. Probable injury or illness refers to a characterization of the risks, either objective or subjective, associated with the use of the medical device.

Evaluation of the benefit(s) of the medical device as compared to the risk(s) should account for the factors below, as applicable[6]:

- Type(s) of benefits to overall patient health and clinical management
- Likelihood of patients to experience one or more benefit
- Duration of device effects
- Patient perspective on device benefit(s)/risk(s)
- Benefit(s) and risk(s) potentially affecting healthcare providers or caregivers
- Medical necessity
- Potential severity of harm
- Likelihood of risk(s) associated with patient harm, device malfunctions, incorrect device results, number of patients that may be exposed to harm, etc.

The evidence supported through analytical and clinical validation will be evaluated based on a risk-benefit analysis. See 21 CFR 860.7(b)(3)[5] for more information.

2 IVD Analytical Validity

Analytical validity is demonstrated evidence that the analytical performance of the IVD device that measures or detects a specified measurand (e.g., genetic marker or biological characteristic) is capable of reliably satisfying established criteria for meeting its stated intended purpose. Examples of analytical performance characteristics for IVDs may include but are not limited to, imprecision, bias (e.g., for quantitative tests), analytical specificity (minimal impact of bias-causing substances or conditions), and analytical sensitivity (limit of detection and limit of quantification (for quantitative tests)).

2.1 List of Studies Frequently Used to Support Analytical Validity

In this section, we present a list of studies that FDA frequently considers when evaluating the analytical validity of qualitative, quantitative, and semi-quantitative IVD tests. The sponsor should utilize a risk-based approach when deciding and designing any study to support analytical validity that may be applicable.

For each of the thirteen studies listed, we define key terms and list data sources, requirements, and reference documents. This section is meant to be a starting point for sponsors; thus, we refer the reader to the relevant reference documents in each section. Regulatory and standards bodies and documents mentioned here include:

- *CLSI* – The Clinical & Laboratory Standards Institute. CLSI "provides standards and guidelines for medical professionals through its unique consensus process"[7]
- *ISO* – The International Organization for Standardization. ISO "develops and publishes international standards."[8]
- *JCGM* – The Joint Committee for Guides in Metrology. JCGM "develops and maintains, at the international level, guidance documents addressing the general metrological needs of science and technology, and to consider arrangements for their dissemination."[9]
- *CFR – Title 21:* The Code of Federal Regulations. CFR is the "codification of the general and permanent rules published in the Federal Register by the departments and agencies of the Federal Government."[10] Title 21 is the section of CFR pertaining to rules of the FDA.
- *FDA Guidance Documents* – In general, FDA Guidance documents are written and published by FDA and represent the agency's current thinking on a topic.[11] FDA has issued several Guidances that are specific to various IVDs. These documents should be reviewed thoroughly prior to designing studies to support analytical validity.
- *IUPAC* – The International Union of Pure and Applied Chemistry. IUPAC is "the world authority on chemical nomenclature and terminology, including the naming of new elements in the periodic table; on standardized methods for measurement; and on atomic weights, and many other critically-evaluated data."[12]

For the purpose of this Framework, intended use should include the specimens to be used with the IVD test and the anticipated use of the IVD test results.

2.1.1 Matrix and Specimen Type Comparison

Specimen type and/or matrix-related issues may arise when assessing the performance of the IVD test. Specimen type comparison issues are those related to performance using different types of specimens (i.e., serum, plasma,

urine). Matrix-related comparison issues may include the use of surrogate samples and issues in which calibrators, controls, or proficiency panels are formulated in matrices that are not identical to the matrix from a patient. It is necessary to perform tests to detect the appearance and extent of these specimen types and matrix-related comparison issues.

Definitions

- *Specimen* – Type of biological material, body fluid or tissue, taken from a single human subject for examination (i.e., serum, plasma, urine).
- *Matrix* – The components in a specimen other than the targeted analyte.
- *Matrix Effect* – The combined effect of components in a specimen (excluding the analyte) on the measurement (or detection) of the quantity of interest.
- *Surrogate Sample* - Material or combination of materials used as a substitute for body fluid or tissue taken from a single human subject for examination.
- *Interference* – If a specific component can be identified as affecting the test result, then this is referred to as interference.
- *Commutability* – The equivalence of a given reference material, as demonstrated by the closeness of agreement between (a) the relation among the measurement results obtained according to different measurement procedures for a stated quantity of this material and (b) the relation obtained among the measurement results for other specified materials.
- *Uncertainty* – A situation that involves imperfect and/or unknown information.
- *Standard Document* – *See* CLSI. Evaluation of Commutability of Processed Samples; Approved Guideline — Third Edition. CLSI document EP14-A3. Wayne, PA: Clinical and Laboratory Standards Institute; 2014.

Summary of Requirements

Performance validation for different specimen types and matrices

1. Performance characteristics must be assessed for all acceptable specimen types for which the test will be used. However, specimen equivalency studies may be conducted as an alternative to executing all validation studies in all specimen types. Taking into consideration the specimen type equivalency study conclusions, some tests may require complete performance assessment for each specimen type.

   Using native patient specimens is preferred for all analytical performance characterization studies. In cases in which this is not feasible, surrogate samples can be utilized. This, however, requires that the relationship between analytical performance characteristics of the patient specimens and surrogate samples have been characterized and understood. For more details about basic principles and hierarchy in preparation of surrogate samples, consult the MDIC Surrogate Sample Framework.

2. Finally, differences due to interfering substances in matrices for the same type of specimen should be characterized.

Calibration considerations

1. Commutability across matrices must be assessed to establish proper control limits and assay calibration. A separate calibration may be required for each matrix if a systematic difference exists between specimens with different matrices.

References

- CLSI. Evaluation of Commutability of Processed Samples; Approved Guideline—Third Edition. CLSI document EP14-A3. Wayne, PA: Clinical and Laboratory Standards Institute; 2014.
- ISO 15194:2009 In vitro diagnostic medical devices -- Measurement of quantities in samples of biological origin.

2.1.2 Accuracy (Comparison Study)

A comparison study is typically performed to assess whether the candidate method successfully determines the value (quantitatively or qualitatively) of the measurand.

Definitions

- *Accuracy* – Accuracy (measurement) is the "closeness of agreement between a measured quantity value and a true quantity value of a measurand" (JCGM 200:2012). It reflects the combined contributions of random and systematic error sources in a measurement procedure.
- *Comparator* – Method used to determine the true quantity value of the measurand. May be an FDA cleared device or a standard or reference measurement procedure with low imprecision or bias.

Summary of Requirements

The accuracy of an IVD must be proved through valid scientific evidence and is expressed through the appropriate statistical analysis of data generated from validation testing comparing the results obtained from the candidate method to the true quantity value across the entire analytical measuring interval. The outcome is the quantification of bias between the candidate method and the true quantity. The statistical approach to determine the bias can vary depending on the type of the IVD.

References

- ISO-3534-1:2006 Statistics -- Vocabulary and symbols -- Part 1: General statistical terms and terms used in probability
- JCGM 200:2012 The international vocabulary of metrology — Basic and general concepts and associated terms (VIM), Third Edition
- CLSI. Measurement Procedure Comparison and Bias Estimation Using Patient Samples. 3rd ed. CLSI guideline EP09c. Wayne, PA: Clinical and Laboratory Standards Institute; 2018.
- CLSI. User Protocol for Evaluation of Qualitative Test Performance; Approved Guideline—Second Edition. CLSI document EP12-A2. Wayne, PA: Clinical and Laboratory Standards Institute; 2008.

2.1.3 Precision: Repeatability, Within-Laboratory Precision, and Reproducibility

Multiple factors may contribute to the variability of the device over the applicable range. For qualitative assays, evaluation of variability is particularly important for samples at or near the cutoff(s) or at or near Medical Decision Levels (MDLs) for quantitative assays. Sources of variability may include different days, operators, lots of reagents, equipment, calibration of equipment, and time between measurements.

Definitions

- *Precision* – The closeness of agreement between indications or measured quantity values obtained by replicate measurements on the same or similar objects under specified conditions.

- *Repeatability* – Precision under the same operating conditions over a short interval of time that includes same operators, same operating conditions, and same location.
- *Reproducibility* – Precision under reproducibility conditions that includes different locations, operators, and different instruments.

Summary of Requirements

Refer to CLSI documents EP05-A3 and EP12-A2 (*references listed at the end of this section*) for details on study design and analysis. Additionally, FDA has issued several guidance documents outlining the requirements that pertain to specific IVDs such as HIV, companion diagnostics, HPV, next-generation sequencing, nucleic-acid based testing, etc.

1. Repeatability is estimated in the reproducibility study or in the within-laboratory study. A usual study design that allows one to estimate repeatability should include multiple replicates (at least two) for the same run. The number of replicates may vary based on the selected study design and/or IVD device.
2. A typical design for the reproducibility study is a study that includes 3 testing sites, 5 days, 2 runs per day, and 3 replicates per run. In this study design, the following components of variance are evaluated: repeatability, between-run, between-day, and between-site. Different operators can also be included in the reproducibility study if needed. Different lots can also be included in the study or estimated in a separate study.
3. Within-laboratory precision (intermediate precision) is precision that includes within-laboratory sources of variability: different runs, different days, different operators, different reagent lots, and so on. The requirements may vary based on the selected study design and IVD device.

References

- JCGM 200:2012 The international vocabulary of metrology — Basic and general concepts and associated terms (VIM), Third Edition
- ISO-5725-1:1994 Accuracy (trueness and precision) of measurement methods and results -- Part 1: General principles and definitions
- CLSI. Evaluation of Precision of Quantitative Measurement Procedures; Approved Guideline—Third Edition. CLSI document EP05-A3. Wayne, PA: Clinical and Laboratory Standards Institute; 2014.
- CLSI. User Protocol for Evaluation of Qualitative Test Performance; Approved Guideline—Second Edition. CLSI document EP12-A2. Wayne, PA: Clinical and Laboratory Standards Institute; 2008.

2.1.4 Reference Interval

An established reference interval for a quantitative assay is used as the basis for comparing the results of the new IVD test for an individual patient. This established reference interval may also be called the “normal” range. Frequently, the lower and upper limits of the reference interval are limits for change in the clinical management of the patient, therefore, these limits can also be Medical Decision Levels (MDLs). The term “expected values” is used for qualitative tests and to describe the expected percent of test results in the target population (for example, for a qualitative HPV test with three outputs, expected values are percent of subjects in the target population with “16/18 positive results,” “12 Other HPV positive results,” and “Negative results”).

Definitions

- *Reference interval* – The interval between the lower reference limit and the upper reference limit of the population to be evaluated. The reference interval is often expressed at a level of 95%, meaning that 95% of the reference population is between these limits.
- *Medical Decision Levels (MDLs)* – Limits, often defined by consensus, at which the clinical management of the patient changes.

Summary of Requirements

Typically, a reference interval is established by assessing approximately 120 subjects from reference (usually, apparently healthy) individuals to determine the interval that covers 95% of the values of the reference samples. These reference individuals should adequately represent various age and gender groups from representative geographical locations. Different age groups, gender groups, or race/ethnicity groups may require establishment of different reference intervals. These studies may be from a single location or multi-center or multi-geographic locations.

Different types of samples, such as serum or plasma, may require the establishment of different reference intervals. The analysis of the data can use parametric or non-parametric statistical analysis. Consult CLSI document EP28-A3c for more information (*reference listed at the end of this section*).

These data can often be obtained in a preliminary fashion from early versions of the IVDs but should be confirmed or determined with IVDs representative of the final manufacturing configuration. Between- and within-person variability may be evaluated as part of establishing a reference interval, along with any variants of the analyte.

References

- CLSI. Defining, Establishing, and Verifying Reference Intervals in the Clinical Laboratory; Approved Guideline—Third Edition. CLSI document EP28-A3c. Wayne, PA: Clinical and Laboratory Standards Institute; 2008.

2.1.5 Detection Capability

The measurement boundaries of the IVD device must be assessed. Typically, this assessment requires determining the lowest level of detection capability and can be represented, for example, as the Limit of Blank, Limit of Detection, or Limit of Quantitation.

Definitions

- *Limit of Blank (LoB)* – The highest measurement result that is likely to be observed for a blank (negative) sample.
- *Limit of Detection (LoD)* – The lowest level of analyte that can reliably be detected by a given diagnostic test.
- *Limit of Quantitation (LoQ)* – The upper limit of quantification (ULoQ) and lower limit of quantification (LLoQ) boundaries in a detection range that are quantifiable with acceptable performance characteristics, including accuracy, precision, and linearity.

Summary of Requirements

1. Multiple methods can be used to establish LoB, LoD, and LoQ. The method will be dependent on the analysis system, the test methodology, and the analyte in question. The LoB, LoD, and LoQ should be established for each specimen type/matrix that is being considered unless matrix equivalence was established for samples with low values. Multiple methods for LoB, LoD, and LoQ determination are outlined in CLSI document EP17 (*reference listed at the end of this section*).
2. For establishing a measuring interval, the LLoQ and the ULoQ should be established. To establish the linearity interval, three or more samples with concentrations at or near the upper limit of the assay should be tested along with 9-11 dilutions per sample and a few replicates.

References

- CLSI. Evaluation of Detection Capability for Clinical Laboratory Measurement Procedures. CLSI document EP17; Approved Guideline—Second Edition. Wayne, PA: Clinical and Laboratory Standards Institute; 2012.
- CLSI. Evaluation of the Linearity of Quantitative Measurement Procedures: A Statistical Approach; Approved Guideline. CLSI document EP06-A. Wayne, PA: Clinical and Laboratory Standards Institute; 2003.

2.1.6 Analytical Specificity

The susceptibility of the IVD device to quantify the measurand in the presence of potentially interfering substances must be characterized. Both endogenous and exogenous sources of potentially interfering substances should be considered, and applicable testing completed.

Definitions

- *Analytical specificity* – The ability of the method to assess, unequivocally, the analyte in the presence of other components that are expected to be present (e.g., similar molecules, impurities, degradation products, matrix components, etc.).

  Analytical specificity is distinct from clinical specificity, which refers to the percentage of individuals with assay negative results among all individuals who truly do not have the target condition.

Summary of Requirements

Analytical specificity would include tests such as reactivity/inclusivity testing and cross-reactivity testing. An example of specificity testing for an IVD that detects the H1N1 influenza virus would include reactivity/inclusivity testing of the IVD with a number of influenza virus strains to prove that there is no reactivity with the intended H1N1 test; cross-reactivity testing could be evaluated using common pathogens of respiratory origin.

References

- CLSI. Interference Testing in Clinical Chemistry. 3rd ed. CLSI guideline EP07. Wayne, PA: Clinical and Laboratory Standards Institute; 2018.
- CLSI. Supplemental Tables for Interference Testing in Clinical Chemistry. 1st ed. CLSI supplement EP37. Wayne, PA: Clinical and Laboratory Standards Institute; 2018.

- ISO 17822-1:2014 In vitro diagnostic test systems -- Qualitative nucleic acid-based in vitro examination procedures for detection and identification of microbial pathogens -- Part 1: General requirements, terms and definitions

2.1.7 Reagent Calibrator, & Quality Control Sample Stability

IVD reagents, calibrators, and quality control (QC) samples must maintain their performance characteristics over a given period of time (shelf-life) when stored, transported, and used in the conditions specified by the manufacturer or when prepared, used, and stored according to the manufacturer's instructions for use.

Definitions

- *Stability* – The ability of a reagent, calibrator, or quality control material to maintain its performance characteristics consistently over time.

Summary of Requirements

Various types of stability testing are required, including:

- Establishing IVD reagent and instrument shelf life, including the transport conditions required to maintain product specifications
- Establishing stability of the IVD reagent after the first opening of the primary container (open vial or in-use)
- Monitoring stability of IVD reagents already placed on the market
- Verifying stability specifications after modifications of the IVD reagent that might affect stability

References

- ISO Guide 30: 2015 Reference materials – Selected terms and definitions
- ISO 23640:2011 In vitro diagnostic medical devices -- Evaluation of stability of in vitro diagnostic reagents
- CLSI. Evaluation of Stability of In Vitro Diagnostic Reagents; Approved Guideline. CLSI document EP25-A. Wayne, PA: Clinical and Laboratory Standards Institute; 2009.
- 21 CFR 809 In Vitro Diagnostic Products for Human Use

2.1.8 Flex Studies

Flex studies, also known as Guardband Studies or Assay Characterization, test the limits of the assay design in various scenarios.

Definitions

- *Flex study* – a study that establishes the robustness of an IVD to function correctly under varying conditions of improper use, ensuring that slight variations in the test do not affect results.

Summary of Requirements

The acceptable specifications for an assay's tolerance of improper use should be determined in advance of a flex study, with consideration to the intended users, testing location, and all the potential sources of error.

Two key aspects to consider are user errors and environmental factors. For example, a flex study assessing user error for an IVD that requires the input of three drops of sample should test the effects of inputting 1, 2, 3, 4, 5, and 6 or more drops. A flex study assessing potential errors due to an environmental factor such as temperature for a device specified for use at room temperature (between 15 and 30°C) would also assess device performance at

temperatures lower and higher than the specified temperature range. Other potential sources of error are identified in the FDA Guidance referenced below.

Reference

- Guidance for Industry and FDA Staff: Recommendations for Clinical Laboratory Improvement Amendments of 1988 (CLIA) Waiver Applications for Manufacturers of In Vitro Diagnostic Devices Jan. 30. 2008.

2.1.9 Usability

The user can influence the accuracy, performance characteristics, and interpretation of IVDs. User-related factors that can influence an IVD include who collects the specimen, conducts the test, reads the output, and interprets the results.

For example, tests intended to be performed in a qualified laboratory by highly trained individuals will be evaluated under those conditions. In contrast, tests intended for home use by individuals without other training will be evaluated in the hands of lay users who only receive the labeling and test kit in their home environment or in a structured environment that simulates the home setting.

IVDs intended for near-patient testing in clinics, emergency rooms, or a physician office laboratory may also be influenced by factors not typically at play in a highly controlled laboratory environment. These factors should be considered and evaluated during usability testing.

Definitions

- *User testing* – Testing of the final device under the labeled conditions of use, on patients indicated for use, and performed by individuals with the background, education, and training of those who will perform the test in its intended environment.

Summary of Requirements

User testing should be conducted on the final device configuration and with final draft instructions for use. Importantly, testing should also assess labeling interpretation.

Testing must be objective, without bias, and performed under a clear protocol to ensure the resulting data are reflective of actual use conditions. Individuals to be tested should represent those with the relevant medical conditions of target individuals. The demographics of the test users, as well as the location of testing, should reflect the target users and environment.

References

- Guidance for Industry and FDA Staff: Recommendations for Clinical Laboratory Improvement Amendments of 1988 (CLIA) Waiver Applications for Manufacturers of In Vitro Diagnostic Devices. Issued January 30, 2008.

2.1.10 Specimen Collection and Stability

For the IVD device under investigation, the method of specimen collection needs to be determined. Collection vessels must be stored according to manufacturer's instructions. The method of collection must also be reviewed for any potential problems, such as hemolysis of the sample. Control of sample transport is also required to maintain stability. The transport time interval, storage conditions, and temperatures are critical considerations.

Except for point of care testing**,** testing a patient sample is usually not performed directly after collection but after the sample has been processed, transported, and stored. Therefore, it is important that sample stability be maintained over the relevant stored/ transport conditions.

Stability is not only related to the chemical integrity of a molecule but also to other factors during transport and storage, for example, solvent evaporation, adsorption to containers, and non-homogeneous distribution over a sample. It is recognized that this topic is broad and sometimes complicated (it includes short-term and long-term stability issues).

Sample stability can be assessed by two different approaches. In the first approach, the measurand is considered stable for a stated period and under defined conditions when the average change in the values measured by the IVD device between (a) the results for a stored sample and (b) the results for a corresponding fresh sample (time=0) is less than a value, δ, which depends on the pre-specified risks of decision error.

In the second approach, analyte stability results are obtained by comparing the concentration after storage to a reference value that can be either the theoretical concentration or the concentration experimentally determined in an aliquot of the sample that has not been subjected to a storage (time=0).

Summary of Requirements

Stability at 2-8°C, room temperature (15-30°C), and frozen should be determined and demonstrated, as applicable. Additionally, a series of freeze-thaw studies can assess sample robustness with repeated manipulations.

Sample stability should be obtained by using multiple samples with values close to MDLs for quantitative assays or close to the cutoff for qualitative assays. These should be tested with a single lot of reagent and calibrator. All samples, such as serum, plasma, whole blood, dried blood spots, etc., and all QC materials must be analyzed with the sample material being tested.

References

- "*Stability: Recommendation for Best Practices and Harmonization from the Global Bioanalysis Consortium Harmonization Team*," by Nico van de Merbel, Natasha Savoie, Manish Yadav, Yoshiaki Ohtsu, Joleen White, Maria Francesca Riccio, Kelly Dong, Ronald de Vries, and Julie Diancin in The AAPS Journal, Vol. 16, N. 3, May 2014, page 392-399.
- CLSI. Collection, Transport, and Processing of Blood Specimens for Testing Plasma-Based Coagulation Assays and Molecular Hemostasis Assays; Approved Guideline—Fifth Edition. CLSI document H21-A5. Wayne, PA: Clinical and Laboratory Standards Institute; 2008.

2.1.11 Linearity (for Quantitative and Semi-Quantitative Tests Only)

Manufacturers of quantitative tests must determine or establish the concentration(s) at which the test is linear over the intended measuring interval. The extent of the test's nonlinearity must also be assessed.

Definitions

- *Linearity* – The ability, within a given range, to obtain test results that are directly proportional to the concentration of the analyte in the test sample.

Summary of Requirements

A linearity study panel includes seven to eleven samples. These samples have known concentrations or a known relative relationship to each other that is established by dilution. The samples should span the anticipated linearity interval.

The sample with the highest concentration in the linearity panel should be above the upper limit of the linearity interval and the sample with the lowest concentration in the linearity panel should be below the lower limit of the linearity interval.

At each level, multiple replicates should be tested, depending on the expected imprecision of the assay. Deviations of the mean values of multiple replicates from the best fitted straight line corresponding to a line that assay results are directly proportional to the concentration describe the extent of the assay nonlinearity.

References

- CLSI. Evaluation of the Linearity of Quantitative Measurement Procedures: A Statistical Approach; Approved Guideline. CLSI document EP06-A. Wayne, PA: Clinical and Laboratory Standards Institute; 2003.

2.1.12 Measuring Interval

For quantitative tests, the measuring range or interval needs to be defined. This is typically defined in an interval where the imprecision is acceptable, the test is linear, and biases are acceptable (if applicable). The low value of the range is called the lower limit of quantitation (LLoQ), and the upper value is called the upper limit of quantitation (ULoQ).

Definitions

- *Measuring Interval* – A set of values of the same kind that can be measured by a given measuring instrument or measuring system with specified instrumental uncertainty, under defined conditions.
- *Limit of Quantitation (LoQ)* – The upper limit of quantification (ULoQ) and lower limit of quantification (LLoQ) boundaries in a detection range that are quantifiable with acceptable performance characteristics, including accuracy, precision, and linearity.
- A*nalytical measuring interval* (*AMI*) – The interval in which specimen concentrations are measured within the medical and laboratory needs for accuracy without dilution, concentration, or other pretreatment not part of the standard or routine measurement process. The AMI includes the interval in which linearity, precision and bias (if applicable) have been deemed acceptable and extends from LLoQ to ULoQ.
- *Extended measuring interval (EMI)* – The interval in which concentrations are measured with appropriate accuracy by diluting the specimen before taking a measurement with developed measurement process.

- *Reportable Interval* – The interval that includes the AMI and EMI and also extends to the lower limit of detection.

Summary of Requirements

For establishing a measuring interval, the lower limit of quantitation (LLoQ) and the upper limit of quantitation (ULoQ) should be established. Often, LoB and LoD should also be established. To establish the linearity interval, samples with concentrations at or near the upper limit of the assay should be tested along with nine to 11 dilutions per sample. Precision/reproducibility studies are used for precision profiles and decisions of whether imprecisions are acceptable. The measuring interval is an interval where accuracy, imprecision, and deviation from linearity are acceptable. Notably, different types of samples, such as serum or plasma, may require the establishment of different measuring intervals.

These data can often be obtained in a preliminary fashion from early versions of the assay but should be validated with assays representative of the final manufacturing configuration.

References

- JCGM 200:2012 The international vocabulary of metrology — Basic and general concepts and associated terms (VIM), Third Edition
- ISO 18113-1:2009 In vitro diagnostic medical devices – Information supplied by the manufacturer (labelling) – Part 1: Terms, definitions and general requirements
- CLSI. Evaluation of Detection Capability for Clinical Laboratory Measurement Procedures. CLSI document EP17; Approved Guideline—Second Edition. Wayne, PA: Clinical and Laboratory Standards Institute; 2012.
- CLSI. Evaluation of the Linearity of Quantitative Measurement Procedures: A Statistical Approach; Approved Guideline. CLSI document EP06-A. Wayne, PA: Clinical and Laboratory Standards Institute; 2003.
- CLSI. Establishing and Verifying an Extended Measuring Interval Through Specimen Dilution and Spiking. 1st ed. CLSI guideline EP34. Wayne, PA: Clinical and Laboratory Standards Institute; 2018.

2.1.13. Carry-Over and Cross Contamination Effects

The performance of an IVD can be compromised when undesired materials – for example, diluents, wash solutions, certain parts of a specimen, or even reagents from other IVD products – are carried into a reaction mixture in which they do not belong. These carry-over phenomena occur in manual, automated, and semi-automated IVD devices and are critical to evaluate.

Definitions

Carry-over can be classified either according to:

- *the material that is carried over*, e.g., carry-over of specimen, diluent, reagent, reaction mixture, or wash solution
- *the site where the carry-over occurs*, e.g., carry-over in a specimen cup, sample probe, reagent probe, reaction system, signal detection system, or wash system.

In practice, the following combinations of carry-over have been observed:

- *Specimen-to-specimen* – Carry-over of specimen to specimen in a sample probe. Carry-over from a preceding sample probe into the following specimen may also be referred to as contamination.
- *Diluent-to-specimen* – Carry-over from the diluent to a specimen.
- *Reagent-to-reaction mixture* – Includes carry-over from reagent(s) to the reaction mixture by a reagent probe, carry-over from reaction mixture to reaction mixture in a processing system or signal detection system, and carry-over by insufficient washing and drying.

Summary of Requirements

A typical study design for evaluation of the carry-over effect usually includes two stages that will differ depending on if the IVD test is quantitative or qualitative.

For quantitative tests:

- In the first stage, sample *S* with a concentration close to an MDL (e.g., an upper limit of the reference interval) is measured in one run with a few (at least two) replicates.
- In the second stage, the same sample *S* is tested in a series alternating with a high positive sample *H* in patterns dependent on the operational function of the device. High positive samples in the study should be close to the upper limit of the test's measuring interval. At least five runs with alternating samples should be performed during the carry-over study.
- The carry-over effect can be calculated as the difference between *B* and *A*:

  $$\text{Carry-over} = B - A$$

  where *A* is the mean value of replicates for sample *S* in the first stage and *B* is the mean value of the measurements for sample *S* in the second stage.

  Calculation of the value *B* includes only such measurements of sample *S* for which potential carry-over can occur.

  In addition, the carry-over effect can be calculated as a percentage:

  $$\text{Percent carry-over} = \frac{(B - A)}{H} \times 100\%$$

- The second stage can also be performed for the sample *S* tested in a series alternating with a low concentration (sample *L*) in patterns dependent on the operational function of the device. Low concentration samples in the study should be close to the lowest concentration in the target population (close to the lower limit of the measuring interval or true negative samples).

For qualitative tests:

- In the first stage, negative sample *S* is tested in one run with a few (at least two) replicates.
- In the second stage, the same sample *S* is tested in a series alternating with a high positive sample *H* in patterns dependent on the operational function of the device. High positive samples in the study should be high enough to exceed 95% or more of the results obtained from specimens of diseased patients in the

intended use population. At least five runs with alternating samples should be performed during the carry-over study.

- The carry-over effect can be estimated by a difference between the percent of negative results for the negative samples that are adjacent to high positive samples in the second stage of the carry-over study, *B*, and the percent of negative results in the absence of adjacent high positive samples in the first stage of the study, *A*.
- The difference in the mean signal value of the negative sample *S* in the second stage and the mean value of the negative samples in the first stage should also be evaluated.

References

- IUPAC: Proposals for the Description and Measurement of Carry-Over Effects in Clinical Chemistry (recommendations 1991)

3 IVD Clinical Validity

Clinical validity for an IVD is clinical evidence to demonstrate that the measurements of the IVD medical device (test) are capable of reliably predicting the clinically defined condition, disorder, or health status of interest in a clearly defined target (patient or subject) population. The clinical evidence supports the scientific validity and performance of the device described in the labeling of the IVD medical device and marketing claims.

For many analytes, the clinical utility is well-established and therefore determination of the accuracy on clinical samples may be all that is required for clinical validity. For example, calcium in serum is well-established as being linked to the diagnosis and treatment of parathyroid disease, a variety of bone diseases, chronic renal disease, and tetany. In contrast, for novel analytes or combinations of analytes the clinical validity and/or utility of the analyte (s) may not be established, and this may have an impact on the clinical validity study design.

As scientific and medical knowledge further develops, the initially established clinical validity for an analyte might change and/or expand, e.g., C-reactive protein (CRP). This analyte was initially established as being linked to the detection and evaluation of infection, tissue injury, and inflammatory disorders. However, CRP was later found to be linked to the risk of cardiac disease. Adding additional indications for a device may impact the clinical validity study design.

Examples of clinical validity may include but are not limited to, clinical sensitivity, clinical specificity and, if prevalence is known or assumed, positive predictive value and negative predictive value, as well as positive and negative likelihood ratios. Clinically defined conditions include the intended purpose of the examination, e.g., for diagnosis or population screening.

The design of the clinical validity study will depend on the intended use of the medical device with the following considerations:

- test purpose(s) (e.g., diagnosis, screening, monitoring)
- target population(s) (e.g., age, race, gender, geography, clinical condition)
- specimen type(s) (e.g., serum, plasma, urine)
- intended user(s) (e.g., layperson)
- established analytical validity characteristics (e.g., precision, interference, measuring interval (range), cutoff)
- expected clinical validity characteristics (e.g., clinical sensitivity, specificity)
- novelty of the technology and/or clinical use (e.g., relevant previous experience)
- availability of an appropriate method to establish the true clinical status of the patient

Evaluation of continued clinical validity of an assay should be monitored post-market (e.g., adverse event reports, results from performance studies, published literature) as this information may indicate a need to reassess the benefits and risks of an IVD medical device. Labeling and marketing claims may be required to be adjusted based on this information.

In the rest of this chapter, we discuss the types of assays used to assess clinical validity. Then, we present clinical validity study design considerations based on the intended use of the IVD.

3.1 Assay Types/Measurement Procedures for Clinical Validation

Assays or measurement procedures to assess clinical validity include qualitative, quantitative, semi-quantitative, titered, and Multi-Analyte Assays with Algorithmic Analyses (MAAA), also known as In Vitro Diagnostics Multivariate Index Assays (IVDMIA). The choice of assay type is based on the results that will ultimately be provided to physician/patients, as well as the possible valid interpretations of the results. The different types are further explained below:

3.1.1 Qualitative Assays

Qualitative assays are assays that are designed to determine whether a target condition is present or absent from the intended use population. A qualitative assay provides results to the user in terms of nominal values. Nominal values are not ordered categories and have no magnitude.[11] Examples include the color of a spot test in chemistry, a sequence of amino acids in a polypeptide, and pregnancy tests and ovulation tests.

Types of qualitative assays include:

- **Binary qualitative assays with only two outputs**: Positive and negative or detected and not detected. Qualitative assays help to identify the presence or absence of the analyte (pathogen, toxin, antigen, antibody) or an amount of the analyte above some threshold.
- **Qualitative assays with equivocal/Indeterminate results**: In some cases, the result may be reported as equivocal, indicating a level of response (signal) that falls in a gray zone (equivocal zone) of neither positive nor negative.
- **Qualitative assays with multiple (>2) outputs**: The outputs can be coded with numbers, but the order is arbitrary and any calculations (e.g., computing the average) would be meaningless. Examples include genetic tests with three outputs such as aa, aA, and AA and a genotyping HCV assay with outputs: 1a, 1b, 2, 3, 4, 5, and 6.

Clinical validity study designs may include expected values for comparison. For qualitative tests with binary outputs (positive, negative), quantitative tests with one cutoff, and MAAA tests with one cutoff, the following analysis is applicable: Clinical sensitivity and specificity based on the known clinical/physiological state of the individual, likelihood ratios (positive and negative), and predictive values – Positive Predictive Value (PPV) and Negative Predictive Value (NPV) – based on the prevalence of the disease.

Clinical validity for a binary qualitative test with outputs of either positive and negative results can be assessed following the method indicated below using actual clinical study data. In the tables and calculations below, *N* represents the total number of subjects from the target population and *A1, B1, A2, B2*, etc. are the numbers of patient results in each of the possible categories.

Outputs	Target Condition Present or Comparative Method Positive	Target Condition Absent or Comparative Method Negative	Total
Positive	A1	B1	A1+B1
Negative	A2	B2	A2+B2
Total	A1+A2	B1+B2	N

Clinical sensitivity and specificity are described by

$$Clinical\ sensitivity = \frac{A1}{A1 + A2}$$

$$Clinical\ specificity = \frac{B2}{B1 + B2}$$

Outputs	Likelihood Ratios	Risks	Percent of Subjects with a Given Output
Positive	[A1/(A1+A2)]/[B1/(B1+B2)]	PPV = A1/(A1+B1)	(A1+B1)/N
Negative	[A2/(A1+A2)]/[B2/(B1+B2)]	1-NPV = A2/(A2+B2)	(A2+B2)/N
		Prevalence = (A1+A2)/N	

Clinical validity for qualitative tests with multiple outputs, for quantitative tests with multiple cutoffs, MAAA tests with multiple cutoffs, and semi-quantitative tests can be determined with the following methods: likelihood ratio (LR) for each output, prevalence, risk for each output, and percent of subjects with corresponding output in the population.

For example, for the test with three outputs (O_1, O_2, O_3) and data:

	Target Condition Present	Target Condition Absent	Total
Output O_1	A1	B1	A1+B1
Output O_2	A2	B2	A2+B2
Output O_3	A3	B3	A3+B3
Total	A1+A2+A3	B1+B2+B3	N

Clinical performance is described:

	Likelihood Ratios	Risks	Percent of Subjects with a Given Output
Output O1	LR1 =[A1/(A1+A2+A3)]/[B1/(B1+B2+B3)]	A1/(A1+B1)	(A1+B1)/N
Output O2	LR2 =[A2/(A1+A2+A3)]/[B2/(B1+B2+B3)]	A2/(A2+B2)	(A2+B2)/N
Output O3	LR3 =[A3/(A1+A2+A3)]/[B3/(B1+B2+B3)]	A3/(A3+B3)	(A3+B3)/N
		Prevalence = (A1+A2+A3)/N	

Regarding invalid results where batch controls failed, the percentage of invalid results should be provided separately but not included in the calculation of the test performance with valid results. The same is true for failed runs and failed specimens where a result is not obtained.

3.1.2 Quantitative Assays

Quantitative assays produce results that provide information on the amount of analyte in the sample relative to the international or national reference standards, e.g., World Health Organization (WHO) or National Institute of Standards and Technology (NIST). Examples of quantitative assays include assays providing analyte concentration on a log scale, assays based on a counting process (results can be non-negative integer numbers), and assays producing percentages (values from 0% to 100%).

The results of quantitative assays are reported as numerical values, have properties of linearity, and are either in the form of interval or ratio data. Interval data are numeric values that can be compared by the difference between

values, while ratio data can be compared also by division. Typically, reference intervals should be established for interpretation of the quantitative assay results.

3.1.3 Semi-Quantitative Assays

Semi-quantitative assays provide ordinal numerical values; however, they cannot be considered quantitative because the values of the assay cannot be compared either by difference or by division. Ordinal data are data where the order matters but the difference between values does not. For example, consider a urine dipstick assay with results: negative, trace, 1+, 2+, 3+, and 4+. A result of 4+ means there is more analyte than in a 2+ result because the order of the results determines their interpretation. However, the difference between 4+ and 2+ may not be the same as between 3+ and 1+. A ratio of 4+ and 2+ does not mean that the amount of analyte for 4+ is 2 times larger than for 2+.

Semi-quantitative numeric results from different assays may not be comparable, though the qualitative interpretation of the results will be similar. Semi-quantitative results are compared against an accompanying reference interval to provide a qualitative interpretation.

3.1.4 Titered Assays

A titered assay is a variation of a semi-quantitative assay that reports the relative amount of an analyte within the sample. The assay requires titration of the sample into serial dilutions to detect the point at which the analyte is no longer detected. The dilution standards are set forth by the performing laboratory and can vary significantly between laboratories. Titers are usually reported as a ratio of the highest dilution that still allows detection of the analyte, for example, 1:8.

3.1.5 Multi-Analyte Assays with Algorithmic Analyses (MAAA) or In Vitro Diagnostics Multivariate Index Assays (IVDMIA)

MAAA (or IVDMIA) combine the values of multiple variables using an interpretation function to yield a single, patient-specific result, e.g., a “classification,” “score,” or “index,” that is intended for use in the diagnosis of disease or other conditions, or in the cure, mitigation, treatment or prevention of the disease or condition.

3.2 Design Considerations for Clinical Validation Based on Intended Use of the IVD

The intended use of the IVD helps to determine the clinical study design. The test may be used to determine the current state of the patient, such as a test used to screen for, diagnose, or aid in diagnosing a specific disease state. Alternatively, the test may be used to determine the future state of the patient, such as those for predisposition, prognosis, and treatment response tests. Below we present design considerations for clinical validation of assays, classified by the intended use of the IVD.

3.2.1 Diagnosis

A diagnostic test is one that may determine whether a patient has or does not have a particular disease. In some cases, the diagnosis is defined by the test, e.g., diabetes might be diagnosed based on the levels of HbA1c in blood. Most tests will require a clinical study to support approval or clearance. For example, in infectious disease testing, new tests are often compared to a gold standard for the diagnosis – a test that identifies the microorganism’s presence utilizing a standard laboratory process, or for cardiac markers, a test that detects troponin is compared to clinical outcomes for myocardial infarction.

Another example is breath testing for *Helicobacter pylori* infection in the stomach. The breath tests utilize the urea-splitting activity of the bacteria, an activity that is only found in the stomach when the bacteria are present, and an orally ingested isotope-labeled urea, which, when split by the bacterial urease can be detected as labeled CO_2 in the

breath. The gold standard, in this case, is a combination of demonstrated bacterial presence via histology and bacterial enzymatic activity detected in endoscopically acquired gastric biopsy specimen. New breath tests must establish a dose of the substrate and a cut-off for detection of the labeled CO_2, and then confirm them in a clinical study establishing sensitivity and specificity versus the gold standard.

3.2.2 Aid in Diagnosis

Physicians often order diagnostic testing based on the symptomology of the presenting patient in order to confirm or refine their initial impressions. Therefore, results from the diagnostic tests are used along with other information available to the physician as an aid in the diagnosis of a certain disease or condition. Examples of these types of tests are those for various infectious agents where the patient is presenting with a fever, cough, or other symptoms normally observed with that infection.

3.2.3 Screening

Screening is an early-detection preventive health measure that identifies the probable presence of a disease in individuals deemed to be at risk of developing the targeted disease (average risk of disease) but who are lacking signs or symptoms. Screening IVDs are intended for large-scale population testing and require defined clinical pathways to diagnosis following positive results or results indicative of disease presence. An optimal screening paradigm may include the sequential use of tests. Performance of screening tests can be designed to favor sensitivity or specificity based on the intent to detect or rule out disease, respectively. Disease targets can include but are not limited to genetic markers, epigenetic markers, proteins, DNA, RNA, viruses, bacteria from blood, saliva, stools, and tissue biopsy. Examples of screening IVDs include blood tests for various genetic conditions (prenatal, newborn screening), stool tests for colorectal cancer, liquid biopsy for cervical cancer, the total PSA test for prostate cancer, and a blood test for latent tuberculosis.

3.2.4 Monitoring

Monitoring is defined as a test that is used to monitor the progress of a disease state or the response to a medical treatment. Monitoring can be performed by measuring certain parameters, which is achieved by repeatedly performing a medical test on samples (of various types). For example, blood glucose monitors are a type of monitoring IVD.

Clinical validation for monitoring tests should characterize the subject population and state the decision-making process that will be made based on the IVD test results (for example, calculation of a difference in the test results between patient test results at different visits). Trends in IVD tests used for monitoring include enabling personalized treatment regimes, monitoring disease progression, and helping to guide therapeutic options.

Clinical validation issues related to monitoring IVDs include the ability to obtain true positive and negative samples throughout the progression of the testing/monitoring. Tests with different technologies can be used for monitoring, thereby creating challenges in the evaluation of these tests. An example is the INR test (Prothrombin Time); quantitative values obtained from one methodology are difficult to interpret relative to those obtained from other methodologies, yielding challenges in the evaluation of test accuracy.

3.2.5 Predisposition (Risk Assessment)

Predisposition assays are used to determine the likelihood of disease onset in asymptomatic patients in whom the disease is not usually present. These assays assess the risk of developing the disease in the future. Predisposition assays differ from screening assays, where the disease must be present in order to be detected (although the patient may be asymptomatic). The assay results may warrant preventive interventions for patients at sufficient risk.

These assays are designed to evaluate a patient's future state. Examples include the genetic test for apolipoprotein E to assess the risk of developing Alzheimer's disease and BRCA1/BRCA2 mutation status testing to assess the risk of developing breast cancer.

3.2.6 Prognosis

Prognostic tests are also designed to evaluate a patient's future state in patients already diagnosed with a disease/condition. Such tests may be used to estimate the natural progression of the disease (i.e., outcome in the absence of treatment) or to determine the likelihood of a clinical outcome irrespective of therapeutic intervention. Examples include highly sensitive C-reactive protein measurement for the risk estimation of future cardiac events for patients with acute coronary syndromes and a baseline HIV-1 RNA level to assess prognosis for HIV patients.

3.2.7 Treatment Response (Predication)

Treatment response refers to the ability of a given therapy to affect a specific disease condition, thereby decreasing the harm caused by the disease. Diagnostic tests linked to predictive outcomes are used for diagnosis, screening, and monitoring of a specific disease, and for predicting susceptibility to a disease. Predictive diagnostic tests can directly aid in drug selection (for example, HER2/neu genetic testing to predict treatment response to drug therapy), or they can be used to monitor and assess disease conditions or markers that correlate to therapy response (i.e., viral load tests to monitor antiviral efficacy, glucose tests to assess insulin response, and cholesterol monitoring to predict the utility of statins). Challenges related to the clinical validation of treatment response assays include the need to have well-established therapy cutoffs and decision points, along with specifications for acceptable assay error and precision, particularly at the medical decision level (MDL).

3.2.8 General Issues of Clinical Validity Study Design

A critical component in designing any type of IVD device clinical validity study is selection of specimens/subjects for testing.[13] Issues with regard to prospectively obtained specimens and archived samples are discussed.

- In the clinical validity studies, specimens may be collected and tested immediately, or under certain circumstances, may be collected and stored prior to being tested. Specimens are said to be *prospectively* obtained when a pre-specified protocol is used, and only specimens from subjects meeting the protocol criteria are obtained.

  In a prospectively planned study, a pre-specified protocol is used. Such a protocol would pre-specify the study design, including inclusion/exclusion criteria, method of subject recruitment and selection, testing protocol, and analysis methods to be used. Subjects meeting inclusion/exclusion criteria would be selected over the study duration. Well-executed prospective planning can help ensure that the study population provides an adequate representation of the target population so that the study provides evidence to support the intended use.

- Specimens that are obtained from collections that are assembled without pre-specified use or that were part of a pre-specified protocol for a different study are not considered to be prospectively obtained. These are *archived* samples. In certain situations, it may be acceptable to supplement a prospective study with archived samples, e.g., when the target condition is very rare, and it is very difficult to obtain a sufficient number of subjects with the target condition in a prospective manner. Sometimes, only archived samples can be used to assess the performance of the device, provided that the potential for bias and other concerns can be adequately addressed.

Inclusion of previously archived samples can introduce additional challenges. In general, for specimens selected in a prospective manner, the selection process is under the control of the investigator(s). In contrast, archived samples may be limited to, for example, samples from subjects with a reference method result. The concern is that the archived samples may be non-representative of the target population (e.g., archived samples may represent only extreme cases of the target condition).

The use of archived specimens thus requires consideration of several possible issues, including:

i. The purpose for which the samples were collected (with respect to representativeness to the current target population)
ii. Possible degradation of samples or change of technology used to acquire and store samples over time
iii. Non-random depletion of archival samples

4 IVD Clinical Utility

The real-world viability of an IVD test depends on the willingness of insurers to pay for its use, regardless of how technically innovative the IVD may be.

Practically speaking, this means a diverse group of influential independent organizations must be convinced that a given IVD has clinical utility. Historically, payers define positive clinical utility as the extent to which a given service improves patient outcomes. These include specific, patient-focused endpoints such as hospital readmission, length of stay, and morbidity and mortality. Increasing weight is also given to patient functional outcomes, including cognitive ability and the successful accomplishment of activities of daily living, and patient quality of life. For a service to be covered for a specific population, evidence must show it leads to improved outcomes in that population.

Demonstrating clinical utility for IVDs can be complex, particularly in comparison to other medical devices and treatments.

For example, the central question when evaluating most other medical technologies' clinical utility is: "Does the intervention increase the chances of improved patient outcomes in the target population?" One can generate direct evidence of the link between a therapeutic intervention and patient outcomes through clinical trials and/or observational studies.

For IVDs, however, the test results indirectly affect patient outcomes by influencing clinical decisions about treatment and care, which in turn affect patient outcomes. Therefore, the central question for IVD clinical utility differs: "Do the test results inform and support clinical decisions that increase the likelihood of improved patient outcomes, compared to decisions that would be made without the test results?"

Clinical utility builds on analytical validity and clinical validity. A diagnostic test must be accurate enough in real-world conditions that the clinical benefit of decisions exceeds the consequence of decisions based on false results. Furthermore, the clinical utility of new tests is typically compared with existing tests. If the new test is more accurate, faster, less invasive, or offers other advantages, this may add to its clinical utility. However, it is insufficient for an IVD to merely identify a unique new analyte or be more sensitive or specific than an existing test. The new test results also must be clinically relevant, i.e., the results should enable care decisions that are equivalent to or that meaningfully improve patient outcomes beyond what can be done with existing test options.

Economic Utility

Economic utility is closely related to clinical utility. While the Centers for Medicare & Medicaid Services (CMS) do not consider cost in making Medicare coverage decisions, private insurers often do. As a result, new tests that provide incremental advantages, but cost significantly more than existing options may be less viable in the real world.

In addition, both public and private payers are rapidly moving to value-based reimbursement through mechanisms such as bundled payments and accountable care organizations. These mechanisms put providers at financial risk for both improving patient outcomes and reducing overall costs – which will likely further increase price sensitivity for prospective IVDs. Clinicians may carefully weigh any added clinical utility a new test offers against any added cost of the test itself and its impact, if any, on overall treatment costs.

Under the FDA-CMS Program for Parallel Review of Medical Products, a limited number of tests and devices have undergone a formal concurrent review for FDA approval and Medicare coverage approval from CMS. Clinical measures (endpoints) for devices under FDA review also increasingly consider patient outcomes.

In short, IVDs are increasingly unlikely to be commercially viable without evidence of clinical utility and economic utility. The contents of the following subsections are intended to help IVD manufacturers and sponsors systematically develop credible evidence to assess the clinical and economic viability of specific tests to inform development and acquisition decisions.

4.1 Developing Evidence of Clinical Utility for Payers

For evidence of clinical utility for IVDs to be considered credible by payers, it generally must include three elements that are causally linked. These are (1) specific patient outcomes likely resulting from (2) specific clinical decisions and (3) specifically how the test's results influence those decisions. For the utility of the outcome to be attributed or partially attributed to the test, evidence must be presented for each element along with the chain of logic that connects those elements.

Payers may need additional information to support payment decisions. For example, Medicare requires that a service involving a test must fall into a benefit category defined in the Social Security Act. Similarly, commercial payers and risk-bearing providers generally require that tests relate to services covered under contract terms. Before approving payment, they will often assess the cost-effectiveness of a new test relative to existing tests (or no test at all).

However, while third-party payer reimbursement for clinical patient care services generally is the largest potential revenue source for new IVDs, it is not the only source. Significant revenue can be generated from patient self-paid services, developers of new clinical therapies, and research programs, such as gene mapping and large-scale epidemiological studies. Indeed, many IVDs are developed specifically to identify biomarkers targeted by new therapies still in development, while others may be developed specifically to be marketed directly to patients.

Below is a framework of critical elements to consider in estimating clinical utility and how those elements may relate to new IVDs for purposes of assessing market viability. This framework focuses primarily on services covered by commercial payers because this is by far the largest potential market.

4.1.1 Demonstrate Positive Patient Outcomes

The clinical utility of an IVD ultimately rests on its ability to support improved patient outcomes. In descending order of importance, these include:

- *Survival*. Preventing or delaying death is perhaps the most obvious benefit patients may receive from medical interventions in lethal conditions. Evidence for an intervention's utility in extending life generally comes from randomized clinical trials (RCTs) or observational studies, comparing the outcomes of patients managed one way versus another. Note that mortality sometimes may be deferred or delayed by foregoing or stopping a toxic treatment to which the patient is not responding or by preventing an incorrect treatment. These positive outcomes should be referenced and quantified to the extent possible in building reimbursement cases.
- *Reduced illness or morbidity*. Similarly, curing or reducing the effects of an existing illness, and preventing or delaying the onset of additional complications are highly valued positive patient outcomes. Evidence for an intervention's utility in reducing morbidity generally comes from RCTs or observational studies

comparing the outcomes of patients managed one way versus another. Arresting or slowing the progression of chronic diseases — such as in diabetes, heart failure, COPD, dementia, arthritis, macular degeneration, and cancers — is increasingly important as these conditions become more prevalent. Like mortality, morbidity often is reduced as much by avoiding erroneous treatment as it is by initiating correct therapies. The value of a test's utility in arriving at correct diagnoses and treatment decisions should be emphasized in building reimbursement cases.

- *Reduced pain and suffering or symptoms.* Patients feel physical and psychological pain from disease and other conditions, and alleviating such pain is a major patient benefit. Pain is subjective, though it can be reliably characterized using statistically validated patient-reported outcomes instruments in RCTs and observational studies. Symptoms such as fatigue and inability to concentrate can be assessed with validated patient-reported instruments, while others, such as confusion and dementia, may use both patient responses and behavioral observation scales. Since freedom from pain and other bothersome symptoms significantly improves patient quality of life, the utility of a treatment for reducing or eliminating these should be included in building reimbursement cases.
- *Improved clinical biomarkers and signs*. Clinical signs – such as blood pressure, blood sugar, cholesterol, and blood chemistry, as well as patterns of rashes and other inflammation, and any number of biomarkers – may not be perceived negatively by patients but have in some cases been shown to correlate with disease processes in RCTs and observational studies. As such, they might be considered as intermediate outcomes that reliably predict a poor patient outcome. But this depends on the availability of a strong evidence-based link between the biomarker or sign and the relevant clinical outcome as experienced by the patient. The extent to which improved signs resulting from a therapy can be shown to improve patient outcomes should be considered in building reimbursement cases.
- *Positive impact on daily living activities and other functional outcomes.* Functional outcomes, including self-care, ambulation, and the ability to drive, cook, or live independently, are highly valued by patients and often can be improved by medical treatment of diseases and conditions limiting them, such as arthritis and cataracts. Evidence for these can be generated using validated patient-reported and behavioral observation instruments in RCTs and observational studies. Indeed, CMS values improved functional outcomes, sometimes giving them precedence over improved signs and symptoms. The extent to which therapies can be shown to improve functional outcomes should be considered in building reimbursement cases. Patient-reported outcomes are vulnerable to placebo bias and rigorous attention should be focused on designing appropriate controls to minimize bias.
- *Reduced patient burden, risk, or cost for the same or better outcome.* Some clinical tests and therapies are painful or carry with them some risk of an adverse event. New tests that can be shown to provide similar information for guiding useful clinical decisions as an existing test, but are safer, less intrusive, or less painful can be seen as having increased clinical utility. The value of avoiding these adverse patient outcomes may also justify extra cost.

4.1.2 Link Tests to the Clinical Utility of Care Decisions

Typically, a diagnosis does not rest exclusively on a single test result, even though the result may play a critical role in a particular decision. Characterizing the influence that a test result has on a given clinical decision is critical for establishing the chain of logic linking the test result to the utility realized in the patient outcome.

Various models[14,15] exist for characterizing how a test influences a clinical decision, though most roughly follow the hierarchy of evidence for clinical test utility proposed by Fryback and Thornbury[14] in 1991. Such influence may be

quantified by RCTs comparing outcomes of patients correctly diagnosed and treated using the test results with those who were not. It may also be suggested by observational studies. Such studies can show the benefits of treatment as well as any harm from the treatment itself or the harm of failure to treat, which must be balanced to arrive at an assessment of clinical utility.

How a test may influence the management of a patient's treatment plan is described in a hierarchy in order of ascending influence as follows:

- *Analytical and clinical validity.* This is the most basic level of evidence needed to support arguments for the clinical utility of a test, and it must exist to validate the test before any results can be considered in treatment decisions.
- *Influence on physician decision-making.* This intermediate level of evidence suggests a test result is helpful but may not be compelling for diagnosis. It is a necessary step closer to linking test and outcome, but insufficient to show clinical utility. Payers may look to evidence-based clinical society guidelines to see if a test is accepted as appropriate or acceptable practice by a relevant physician specialty. This level of evidence suggests that a test result directly influences the care a clinician delivers, but it is one step removed.
- *Effect on patient outcomes of a physician acting reasonably upon the test result.* This is a high level of evidence of the clinical utility of a test. It exists when it can be shown that the information a test provides leads to a treatment decision that might not otherwise be likely and that such guided treatment meaningfully increases the likelihood of a better patient outcome.

Note that each step in the hierarchy must be fulfilled to establish the chain of evidence needed to link a test to patient outcomes in support of technology assessment or reimbursement decisions. Tests that do not yield results that support differentiating among treatment options are unlikely to be reimbursable in clinical practice.

4.2 Self-Assessment Framework for IVD Clinical Utility

In this section, we provide tools that IVD manufacturers can use to assess the clinical utility and market viability of their prospective IVDs. First, we broadly discuss adopting a "question first" approach to IVD development. Next, we present a series of staged questions that help guide evidence development for clinical utility. Finally, we highlight methods of evidence generation and recommend parties to be responsible for developing evidence.

4.2.1 "Question First" Approach to IVD Development

We propose that IVD manufacturers adopt a "question first" approach to development. This approach generates its own evidence for clinical utility by starting with a clinically useful question to be answered and then develops a test to answer that question.

A "question first" development pathway begins with considering the pivotal clinical decisions faced by the physician treating the patient. The goal is to identify specific clinical decisions currently being made without enough information to tell whether a chosen option improves the patient's chances for better outcomes compared to the discarded options.

Such inadequately informed clinical management decision points typically are highly specific. They often are limited to specific conditions, or even stages of conditions, within the context of specific treatment pathways for specific patient populations. They often involve a pattern of outcomes suggesting that some patients respond better than others due to a specific treatment, possibly due to an underlying biological pathway that is not currently detectable

or understood. Development of an IVD to accurately predict patient outcomes can assist in choosing personalized treatment plans that benefit patient outcomes, treatment timelines, and potential complications.

For example, patients with clinically similar appearing neovascular age-related macular degeneration (nAMD) may resolve with a single intraocular injection of ranibizumab, or they may need monthly injections for years. Repeated injections expose the patient to risks of additional procedure-related complications, along with additional financial burdens. Ideally, a patient would only receive the number of injections needed to preserve acceptable vision. As of yet, no one can accurately predict a patient's response. Identifying these clinical management decision points may best be done through a systematic needs assessment of physicians practicing within a relevant subspecialty treating a specific condition in a specific patient population. In the case of nAMD, retinal surgeons are treating patients before fibrosis sets in.

Demonstrating clinical utility, however, requires more than finding a predictive diagnostic marker. Finding that marker also must make possible a treatment decision that can make a difference in a patient's outcome. For nAMD, as interesting as predicting responses to ranibizumab may be, the clinical utility of such a test might be limited by the fact that it would not significantly affect the course of treatment. Regardless of the test result, patients would likely be re-treated based on periodic clinical assessments, though the frequency of those assessments might be reduced for good responders. To demonstrate clinical utility, we would ideally see evidence that patients whose treatment was curtailed after a single injection, based on a favorable test result, retained acceptable vision with fewer complications compared to patients who were not tested. Similarly, we would see evidence that patients with unfavorable test results went on to complete long-term therapy that preserved their vision, compared to patients who would otherwise have been unwilling to continue treatment with intraocular injections.

The advent of a new treatment option, such as an extended release medication that could reduce the need for repeated risky intraocular injections for poor responders, might greatly increase the hypothetical test's clinical utility – and its chances for reimbursement by payers.

Cost is another factor. The $2,000-plus cost and high risk of ranibizumab injections might support a high price for a test that could reasonably reduce the need for them. But for low-cost, low-risk treatments, for example, a common generic oral drug or a vaccination costing $10, payers, and particularly private payers, likely would not cover a test that costs more than the treatment it might save.

The permutations of disease states, treatment options, the influence of a test on a treatment decision, and the costs and risk of one treatment option over another that factor into a clinical utility assessment are too varied and numerous to list. Notably, these will change over time as evidence accumulates of treatment outcomes guided by test results. For example, the presumed clinical utility of the PSA test as a guide for prostatectomy declined after studies showed that in general, aggressively treated patients had no better survival than patients who opted for watching and waiting. As a result, some payers may not reimburse regular screening with the PSA test except for specific patient sub-groups in which it has been shown more reliable for predicting aggressive cancers

.4.2.2 Five Questions to Guide Clinical Utility Evidence Development

An IVD test that is clinically useful must answer specific clinical questions and those answers must matter clinically, practically, and economically. In this section, we identify five relevant clinical questions along with accepted best practices for answering them. This section is designed to help developers assess the clinical utility for any type of IVD and guide the development of credible evidence to support clinical utility claims.

4.2.2.1 Question One: Does this test generate information that enables clinicians to make patient care decisions that are likely to improve patient outcomes?

If the answer is "No," or it cannot be answered, it is unlikely payers will reimburse for the test. However, in some cases, patients or physicians may find some value in a test that is diagnostic, prognostic, or predictive, even in the absence of a curative treatment option, because the result may provide additional information regarding the patient's status. Test results may inform the goals of care, i.e., palliation versus seeking cure or enrollment in a clinical trial.

Less complex, but related questions include:

- Does this test provide actionable results for a significant portion of patients?
- Do the results lead to reasonable changes in patient management, such as changes in treatment or hospital release?
- Are these changes causally associated with better patient outcomes?

Answering "No" to any of these suggests the test may not have enough clinical utility to be considered for payment. Should the sponsor believe that the IVD does not meet the requirements of clinical utility, a risk-based approach should be utilized to consider additional design input/outputs to meet the clinical utility requirements.
Regardless of how the question is formulated, it must be answered in a specific clinical context. The following steps help establish this vital context.

Define the specific application.
Precisely defining the specific clinical application of an IVD is essential to develop evidence of clinical utility.

Questions to address are:

- What is the intended use of the new test?
  This should be described in an intended use statement specifying the condition under investigation or treatment, the appropriate patient population, the clinical pathway, the clinical decision within the pathway that the test supports, and how and when the test should be used to support that decision. For example, is the test intended for screening asymptomatic individuals versus establishing a diagnosis in symptomatic patients?

- How clear is the intended use statement?
  Does it accurately reflect the actual use of the test? To be useful, the statement must be understandable and reflect actual or potential clinical practice in real-life settings.

- In what circumstances would offering the new test be clearly inappropriate?
  Such circumstances should include any physical or environmental limitations, such as whether the test requires special facilities, specimen collection, handling, transport or processing; as well as clinical contraindications and identification of patient populations for which the test increases risk or has no utility.

Define the purpose of the test and how it applies to clinical decision-making.
To assess the clinical utility of the test we must ask patient-centered and clinician decision-centered questions, including:

- What is the purpose of the new test?
  The answer should describe the test's potential patient benefits, such as supporting a differential clinical decision that could not otherwise be made or providing similar information to an existing test with reduced risk, e.g., avoid ionizing radiation associated with the current diagnostic paradigm, higher accuracy, lower costs, clinically meaningful greater speed, or some other advantage.

- What clinical question does the new test answer?
  It should be unambiguous and clearly related to advancing along a clinical pathway.

- What is the clinical decision to be made?
  The clinical decision option should be clearly articulated to define the continued diagnostic, treatment, or management options to be decided upon based on test results.

- How does the test inform the clinical decision?
  It should be clearly described how the information the test provides is used in making that clinical decision, and why those test results are necessary for making that decision.

If the test's purpose cannot be defined in terms of potential patient benefit, or if the logic by which the test results influence clinical decisions leading to that benefit cannot be articulated, it may be difficult to demonstrate clinical utility.

Formulate PICOTS.

PICOTS[16,17] is an acronym for a method of formulating clinical questions for evidence-based assessment. It refers to Patients, Interventions, Comparators, Outcomes, Timing, and Setting. The purpose is to identify the characteristics of an intervention or test that allow assessment of its performance relative to existing practice and technology.

PICOTS is useful for both engaging patients and clinicians in developing credible evidence for clinical utility and presenting evidence of clinical utility to payers. The Agency for Healthcare Research and Quality (AHRQ) characterizes the steps of PICOTS as follows[16,17]:

- Patients
  The patient population of interest should be described specifically in terms of the disease, stage or severity, prior treatment, and management options that are being considered. Management options may include treatment, additional testing or surveillance. Conversely, options may include no treatment, or no additional testing, or no surveillance.

- Intervention
  The intervention is the technology that is being studied. For an IVD, the relevant evidence includes consideration of downstream therapies that are furnished or avoided based on the new test result. The new test must be specifically defined. In some cases, the new test may be considered as a class (commodity) of tests that is interchangeable with respect to characteristics and performance.

Consider, for example, certain clinical chemistry tests that share aligned FDA-approved labeling for an analyte within a common clinically relevant range of values for the test result. A physician treating a patient for renal failure might consider serial serum potassium levels (using a standard scale reported in mEq/L) over time to represent a meaningful trend, even though the tests may not have been performed on the same analyzer. A new test falling within a well-established "commodity" will likely be able to leverage the established clinical utility of that commodity and receive limited payer scrutiny. That is unless there are red flags such as significantly higher cost, questions about differential performance in relevant settings of care, or concerns about performance by a different type of operator, such as a patient at home rather than a licensed technician in a certified laboratory setting.

However, one cannot assume interchangeability for new and established tests that fall within a general category but have distinct characteristics and performance. For example, various proprietary tests for breast cancer measure different analytes and clinical inputs and apply unique algorithms to produce a risk score that does not fall into a standard report scale. Differences between tests include targeted nucleic acid sequences, epigenetic factors such as methylation, and the relative contribution of these and other inputs in a unique algorithm. Similar examples are found in prostate cancer.

- Comparator
  In the context of health technology assessment reviews of IVDs, the comparator is the medical management that the patient would receive without the incremental information of the new test result. This "usual care" may include currently covered diagnostic technologies, other IVDs, imaging, and other measures. In some circumstances, appropriate comparators may also include clinical criteria, or no test, or empirical treatment as a kind of test. This use of the term "comparator" differs from other uses of the term, which speak more to the analytic performance of the test.

  The analytic "gold standard" comparator for a new test depends on the availability of performance reference standards and the relevance of any such references to disease outcomes as experienced by the patient. In some cases, the "truth" may be known only by autopsy. More commonly, a reference standard exists, often a test for which performance for disease outcome is well-known.

- Outcomes
  For a new test, the outcomes of interest are largely indirect – they are the consequences of the management decisions taken as a result of the incremental information furnished by the test. The usual comparator would be the outcomes based on management decisions in the status quo; that is, the clinical information and prior test results, such as imaging, that are generally available before the new test is done.

  Beneficial outcomes may accrue in the form of administration of correct beneficial therapy, avoidance of unnecessary or incorrect therapy with its attendant burdens, referral for appropriate additional testing, or avoidance of unnecessary or incorrect follow-up testing. Another beneficial outcome of a new IVD with reasonable costs could replace a test that would be invasive or involve radiation.

  Harmful outcomes arise from the consequences of an incorrect result or the mismanagement of the patient based on a technically correct, but misapplied test result; burdens of incorrect or unnecessary therapy; failure to treat with available beneficial therapy; burdens of incorrect or unnecessary additional testing; and failure to obtain appropriate additional testing.

- Timing
  The timing is that of the assessment of the outcome and/or durability of the outcome. Examples include the duration of follow-up and whether a single assessment or multiple follow-up assessments are required.

- Setting
  The setting is the location of the test assessment, e.g., ambulatory (including primary and specialty care) and inpatient settings. It considers whether a beneficial result depends on receiving care from specialty practitioners working in centers of excellence with specific infrastructure or staffing.

Develop an analytic review framework.
Developing an analytic review framework helps clarify the relationships of test results to clinical decisions and to patient outcomes that are necessary to demonstrate the clinical utility of tests. In the context of health technology assessment, analytic review refers not to the narrow technical process required to demonstrate an IVD's analytic validity, but to a broad examination of the links of indirect evidence necessary to demonstrate that a test or treatment may improve health outcomes.

It is often useful to depict the links of evidence required to demonstrate test clinical graphically as a flowchart, since they are complex and often involve intermediate outcomes. Decision trees also may be helpful for mapping multiple potential downstream consequences of complex decisions.[18]

According to AHRQ, key considerations that should be addressed in a graphical analytic framework are[17]:

- Direct evidence that testing reduces mortality and/or morbidity
- Test accuracy
- Impact of test on patient management
- Impact of management on health outcomes
- Impact of management on intermediate outcomes
- Impact of intermediate outcomes on health outcomes
- Adverse events, acceptability of test procedure
- Adverse events of subsequent treatment and/or other tests

An analytic review framework for an IVD also may include a benefit-risk analysis. The benefit-risk analysis should be based on:

- Overall test accuracy based on its analytical and clinical validity – this may include a matrix of true and false negatives and positives and their outcomes.
- Potential outcomes of false negatives or positives – this may include steps to mitigate the harm of false negatives through ongoing monitoring and re-tests and through confirmatory tests.
- Risk of the test itself – this may include any invasive procedure required or exposure to radiation of toxic agents.
- Statistical uncertainty of test results – these should be presented in terms of confidence interval, as well as uncertainty of positive impact of therapies undertaken because of a test result.

Taken together, these data can help assess whether the potential benefits of a test outweigh the risks and can be used to quantify a favorable risk profile for a new test over an existing alternative. Additional examples are available in task force reports on the US Preventive Services Task Force website.[16]

Note: The clinical utility of a test is suspect, and it may not be commercially viable if:

- the above steps cannot be precisely defined and articulated, or
- if they do not demonstrate that patient outcomes are meaningfully improved, or
- that the test's benefits justify its risks.

4.2.2.2 Question Two: Are there alternative existing sources for the same information the test provides?

If the answer is "No," the clinical utility of a new test likely will need to be established with reference to the better patient outcomes that become possible as a result of clinical decisions based on the test's results, as outlined above.

However, many, if not most, new IVDs do not provide completely new clinical information. Rather, they provide similar information to existing tests with some advantage. In this case, the following questions should be examined to establish whether the new test has sufficient clinical utility to justify its use

- Are the new test results truly comparable to existing alternatives? Does it have the same or better clinical validity as an alternative test?

  To be considered equivalent within an existing clinical pathway, the new test must show similar or more accurate results in a similar patient population with similar disease states. Note that demonstrating comparable clinical validity to a test already recognized as part of a useful clinical pathway may be all that is required to establish clinical utility for payment purposes since the links to a specific clinical decision and its outcomes utility presumably already exist for the comparator test.

- What, if any, advantages does the new test offer over existing tests? Is it safer, more accurate, faster, less expensive, less painful, etc.?

4.2.2.3 Question Three: Is the test usable under existing clinical conditions?

Tests requiring special equipment, handling, or training may be considered to have less clinical utility, particularly if the conditions required are not common in the specialty. Conversely, if a new test eliminates the need for expensive handling, equipment, or training, its clinical utility will likely be perceived as higher by payers and providers.

4.2.2.4 Question Four: Do the potential clinical benefits justify any additional cost?

While CMS does not include cost-effectiveness in its technology assessments or coverage decisions, private payers may not support expensive tests for marginal potential outcome improvement. Indeed, private payers may not perceive that a molecular test has more utility than a microscopic test for the same condition even if the molecular test is faster and easier, particularly if there is no or marginal clinical benefit to a faster diagnosis.

4.2.2.5 Question Five: Are the test and any indicated management covered services?

Medicare benefit categories are strictly defined in law and private payer coverage policies vary. A test that cannot be separately billed may not be viable.

4.2.3 Methods for Evidence Generation

Evidence of clinical utility may aid approval or clearance by regulatory bodies and is essential for payment from most sources, regardless of whether an IVD's prospective market is payers, researchers, or patients directly. Evidence for IVDs and other tests is generated from a variety of sources. These typically include both clinical studies supporting efficacy before and after approval, as well as theoretical and professional analyses supporting their use in specific clinical scenarios.

4.2.3.1 Direct Evidence

A successful RCT that shows benefit(s) from a protocol dependent on the test over alternate treatment choices might be considered direct evidence of the test's clinical utility.

RCTs often are conducted for IVDs that are being developed as companion tests for new therapies, for which a specific test is integral to identifying or staging treatment candidates. In these cases, the role of the IVD is unambiguous and its clinical utility is presumed if the treatment itself improves outcomes.

However, IVDs are more commonly incorporated into complex diagnostic frameworks where they provide only a part of the evidence required for making treatment decisions, and alternative tests often exist for obtaining the same or similar information. RCTs evaluating the entire chain of influence, from IVD to clinical decision to patient outcome, may be much more difficult to organize or may not be practical, or even necessary, for these scenarios. If an IVD can be shown to provide the same information as an existing test with less risk of complications, e.g., avoiding or reducing the need for a biopsy, this by itself may be enough to establish the IVD's clinical utility. This also might be established by an RCT.

4.2.3.2 Indirect Evidence

More often, evidence of an IVD's clinical utility is indirect. Indirect evidence includes observational studies of patient outcomes as well as theoretical constructs accounting for a test's use in clinical decision-making. While such evidence is characterized as indirect because it does not isolate and directly test a hypothesis about an IVD's clinical impact, it is often necessary and can be compelling for guiding clinical decisions.

Observational Studies

RCTs by their nature cannot answer questions about the use and outcomes of IVDs or other tests in the real world or even beyond their often very narrow test populations. Observational studies can and do provide this invaluable supplemental information. They can be conducted before or after approval, prospectively and retrospectively.

Post-approval observational studies based on real-world data from patient registries and electronic medical records are particularly valuable for establishing clinical utility. They can provide crucial evidence about how and whether a test is used under real clinical conditions, how it influences actual clinical decisions and subsequent treatment outcomes, and its effect on patients outside the initial test population that cannot be obtained any other way.

Theoretical Frameworks and Models

Similarly, clinical practice by necessity can run ahead of mature evidence, with off-label use sometimes exceeding the narrow scope of approved indications. Such decisions typically are justified in part by a theoretical framework describing how the test or intervention works and why it may be applicable in a given clinical situation. Results often are verified by clinical observations and observational studies, confirming the clinical utility of the expanded use.

Expert opinion for a given application from key opinion leaders often is viewed as credible indirect evidence of clinical utility. Formal adoption of a practice in the practice guidelines of a major specialty organization is even stronger evidence of clinical utility. Such guidelines typically include a review of existing evidence for their recommendations. These carefully developed opinions can carry considerable weight with clinicians, patients, and payers.

4.2.4 Recommended Parties Responsible for Evidence Development

For evidence of an IVD's clinical utility to be considered credible, both clinicians from the relevant specialties and pathologists should be involved in the development and the demonstration of its accuracy and utility.

Clinical Specialists

Given that the utility of IVD and other test results rests on their usefulness in supporting clinical decisions, it follows that the physician or practitioner specialty group responsible for specific uses should be involved in developing the test, as well as evidence of a test's utility. For example, for a test that distinguishes among various targeted therapies for a particular tumor, oncologists should be involved in the clinical trials and observational studies developing the test. The utility of the test results should be reflected in the contents of high-impact oncology journals and/or practice guidelines generated by relevant oncology specialty or subspecialty groups.

Pathologists

Given that clinical testing is highly technical — involving everything from ensuring the integrity of specimen collection and transport, to ensuring proper lab procedures are followed for accurately detecting the target biomarkers, and that the biomarker is associated with the relevant clinical condition — it follows that clinical pathologists should be involved since they are the medical specialty most involved with these technical issues. Indeed, an IVD or another test is unlikely to be paid by payers or accepted by clinicians, researchers, or patients if its basic analytic and clinical validity cannot be demonstrated.

4.3 Additional Clinical Utility Considerations for Specific IVD Test Types

In addition to the general considerations outlined in the previous sections, IVD manufacturers must also consider their specific test type when planning how to produce evidence of clinical utility for payers. The following list presents unique clinical utility considerations for several types of IVD tests.

4.3.1 Diagnostic Tests

Purpose of test: Detect the presence or absence of a disease or health condition, and/or the extent or severity of a disease or condition.

New diagnostic tests are likely to be considered to have clinical utility if they generate information. Examples include:

- Supporting a specific therapeutic indication in a currently accepted clinical context
- Performing more accurately, more quickly, or more reliably than existing tests
- Being similar to existing tests with less burden or risk to patients from the test itself or equivalent to existing tests
- Providing some combination of those benefits

Diagnostic tests for diseases or conditions that are not considered treatable, or do not support a management decision, may not be considered to have clinical utility by payers. However, these may support a future treatment likely to be developed or support a significant research effort that may result in future clinical decisions.

4.3.2 Prognostic Tests

Purpose of test: Project the likelihood of developing a disease or condition in the future, or the likely future course of an existing disease or condition.

Prognostic tests are considered to have clinical utility by payers if they support clinical decisions to conduct future diagnostic screening tests for high-risk patients, or influence treatment or management decisions of existing diseases, such as determining when to initiate rehabilitative or palliative care.

Payers may not consider prognostic tests to have clinical utility if the tests do not assess risk for a treatable or manageable disease or condition, such as detecting genetic features that have a very low chance of being associated with a disease, or if they do not support management decisions for an existing disease. However, these tests may be market viable if they contribute to fundamental research, such as genetically characterizing diseases or disease risk in a population.

4.3.3 Predictive Tests and Companion Diagnostics

Purpose of test: (In the context of personalized medicine) – Assess the likelihood of a positive response to a specific therapy.

Predictive tests are considered to have clinical utility by payers if they support differential clinical treatment or management decisions, such as companion diagnostic tests for therapies targeting disease variants, identifiable from genetic or protein markers.

Predictive tests that do not support treatment decisions for current therapies are less likely to be seen as having clinical utility by payers, though tests co-developed with future therapies may be market viable.[19-21]

4.3.4 Monitoring Tests

Purpose of test: Assess the progress of a disease, condition, or response to therapy.

Monitoring tests may be diagnostic, prognostic, or predictive. However, the term monitoring may raise payer questions unless there is adequate evidence to support the frequency of testing, and it is clear that the test result drives different management options.

4.3.5 Epidemiological Tests

Purpose of test: Assess the prevalence of a disease or condition in a population.

Epidemiological tests are generally not considered to have clinical utility by payers if they are not used to support a potential clinical decision for an individual patient. They may also be viable for secondary research purposes, though tests for research purposes are generally not included in the scope of health insurance benefits, unless there is an actionable paradigm for individual tested patients, e.g., an influenza test.

4.3.6 Quality Control Tests

Purpose of test: Validate the reliability of lab processes.

Quality control tests are generally not considered to have clinical utility and generally are not clinically reimbursable, even though they may be necessary to ensure quality.

4.3.7 Forensic Tests

Purpose of test: Assess the causes of death or injury.

Forensic tests generally are not considered to have clinical utility because they do not influence patient outcomes. However, they may be necessary for legal or research purposes and may be paid as a result.

References

1. Chapter 23 - The Evaluation of Genomic Applications in Practice and Prevention (EGAPP™) initiative: methods of the EGAPP™ Working Group. https://www.cdc.gov/genomics/resources/books/2010_huge/chap23.htm. Published 2010. Accessed September 17, 2018.
2. Grosse SD, Khoury MJ. What is the clinical utility of genetic testing? *Genet Med.* 2006;8(7):448-450.
3. Teutsch SM, Bradley LA, Palomaki GE, et al. The Evaluation of Genomic Applications in Practice and Prevention (EGAPP) Initiative: methods of the EGAPP Working Group. *Genet Med.* 2009;11(1):3-14.
4. Title 21 - Food And Drugs, Chapter 1 - Food and Drug Administration. Code of Federal Regulations Web site. https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfCFR/CFRSearch.cfm?FR=860.7. Accessed September 17, 2018.
5. FDA: What We Do. https://www.fda.gov/aboutfda/whatwedo/default.htm. Accessed September 17, 2018.
6. *Factors to Consider Regarding Benefit-Risk in Medical Device Product Availability, Compliance, and Enforcement Decisions: Guidance for Industry and Food and Drug Administration Staff.* US Food & Drug Administration;2016.https://www.fda.gov/downloads/medicaldevices/deviceregulationandguidance/guidancedocuments/ucm506679.pdf. Accessed September 17, 2018.
7. *About ISO.*https://www.iso.org/about-us.html. Accessed September 17, 2018.
8. *About the Code of Federal Regulations.*https://www.govinfo.gov/help/cfr. Accessed September 17, 2018.
9. *Charter of the Joint Committee for Guides in Metrology.*https://www.iso.org/sites/JCGM/JCGM-charter.htm. Accessed April 5, 2019.
10. *FDA Basics for Industry: Guidances.*https://www.fda.gov/forindustry/fdabasicsforindustry/ucm234622.htm. Accessed September 17, 2018.
11. *Guidance for Industry and FDA Staff: Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests.* US Food & Drug Administration;2007.https://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071287.pdf. Accessed September 17, 2018.
12. *Who We Are - IUPAC.*https://iupac.org/who-we-are/. Accessed April 5, 2019.
13. *Design Considerations for Pivotal Clinical Investigations for Medical Devices: Guidance for Industry, Clinical Investigators, Institutional Review Boards, and Food and Drug Administration Staff.* US Food & Drug Administration;2013.https://www.fda.gov/media/87363/download. Accessed August 2, 2019.
14. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. *Med Decis Making.* 1991;11(2):88-94.
15. *Evaluation of Clinical Validity and Clinical Utility of Actionable Molecular Diagnostic Tests in Adult Oncology.* Center for Medical Technology Policy;2013.http://www.cmtpnet.org/docs/resources/MDX_EGD.pdf. Accessed September 17, 2018.
16. *Methods Guide for Medical Test Reviews.* Agency for Healthcare Research and Quality;2012.https://effectivehealthcare.ahrq.gov/topics/methods-guidance-tests/overview-2012. Accessed September 17, 2018.
17. Samson D, Schoelles KM. Developing the Topic and Structuring Systematic Reviews of Medical Tests: Utility of PICOTS, Analytics Frameworks, Decision Trees, and Other Frameworks. In: Chang S, Matchar D, Smetana G, Umscheid C, eds. *Methods Guide for Medical Test Reviews.* Rockville, MD: Agency for Healthcare Research and Quality; 2012.
18. Woolf SH. An organized analytic framework for practice guideline development: using the analytic logic as a guide for reviewing evidence, developing recommendations, and explaining the rationale. In: McCormick KA, Moore SR, Siegel RA, eds. *Methodology perspectives: clinical practice guideline development.* Rockville, MD: US Department of Health and Human Services, Public Health Service, Agency for Health Care Policy and Research; 1994:105-113.

19. *Medicare Coverage Center.* US Centers for Medicare & Medicaid Services.https://www.cms.gov/Center/Special-Topic/Medicare-Coverage-Center.html. Accessed September 17, 2018.
20. *National Coverage Analysis (NCA) Tracking Sheet for Next Generation Sequencing for Medicare Beneficiaries with Advanced Cancer (CAG-00450N).* US Centers for Medicare & Medicaid Services.https://www.cms.gov/medicare-coverage-database/details/nca-tracking-sheet.aspx?NCAId=290. Accessed September 17, 2018.
21. *Advanced Diagnostic Laboratory Tests (ADLTs).* US Center for Medicare & Medicaid Services.https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/ClinicalLabFeeSched/Advanced-Diagnostic-Laboratory-Tests.html. Accessed September 17, 2018.

Contact information

For more information, please contact
Carolyn Hiller at chiller@mdic.org