Points to Consider Document:

Scientific and Regulatory Considerations for the Analytical Validation of Assays Used in the Qualification of Biomarkers in Biological Matrices

Biomarker Assay Collaborative Evidentiary Considerations Writing Group, Critical Path Institute (C-Path)

Table of Contents

Introduction

The characterization and analytical validation of biomarker assay performance, used to define its capability and limitations, is a fundamentally important aspect of biomarker qualification. Biases introduced in the conduct or interpretation of the assay results will affect the biomarker's predictive accuracy and thus its evaluation as a useful Drug Development Tool (DDT). Inherent in the measurement of biomarkers, unlike the measurement of xenobiotics (drugs), is that biomarkers are endogenous entities or molecules. Therefore, biomarker assays typically measure an increase or decrease in the endogenous level of the molecule which often fluctuates based on biological factors, pathological (disease) factors, treatment administered and environmental factors. Given this, the requirements and expectations for assays used in the qualification of biomarkers must not only take into consideration the type of molecules being measured, but also the context in which the biomarker is being applied in drug development and in regulatory decision making.

The key criteria for the validation of drug concentration (pharmacokinetic assays) and *in vitro* diagnostic (IVD) methods used in clinical practice have been well defined, but are not universally transferable or applicable to biomarker assays as the utility and expectations for the biomarker qualification assays are distinct from these defined criteria. While the criteria used in the validation of drug concentration assays and IVD methods can be applied as a framework for the development of criteria for biomarker assay validation, they cannot be adopted unequivocally. Thus, analytical validation of biomarker assays used during DDT qualification must be refined to fit the proposed context for which they are ultimately going to be utilized.

The goal of this document is to define the scientific and regulatory considerations for the analytical validation of soluble biomarker assays used for the qualification of biomarkers to be employed as DDTs. The topics to be discussed include the optimization of pre-analytical factors, core assay performance expectations, and setting minimally acceptable assay performance criteria. Technology areas covered include single-plex ligand and immune binding assays, mass spectrometry, and enzyme based assays. Out of scope of this document are pre-qualification activities, immunohistochemistry (IHC), flow cytometry, genetics, genomics, imaging biomarkers, and multiplex assays. Likewise the development and analytical validation of assays used in clinical practice, as well as the use of exploratory biomarkers in clinical drug development, is outside the scope of this document.

The two primary areas that require consensus and agreement are 1) the experimental characterization of the biomarker assays used in qualification (Assay Consideration), and 2) the approach to defining the requisite assay performance and acceptance criteria (Assay Validation Acceptance Criteria).

Biomarker Qualification and the Context of Use

The US Food and Drug Administration's (FDA) Biomarker Qualification Program (BQP) is designed to provide a mechanism for external stakeholders to work with the Center for Drug Evaluation and Research (CDER) to develop biomarkers for use as tools in the drug

development process (FDA 2016). The goals of the BQP are to provide a platform to 1) qualify biomarkers and make supporting information publicly available, 2) facilitate uptake of qualified biomarkers in the regulatory review process, and 3) encourage the identification of new biomarkers to be used in drug development and regulatory decision-making (Amur et al. 2015). Terms used in biomarker qualification have been defined by the FDA-NIH Biomarker Working Group and can be found in the BEST (Biomarkers, EndpointS, and other Tools) Resource (BEST Resource 2016).

> A biomarker is a "defined characteristic that is measured as an indicator of normal biological processes, pathogenic processes, or responses to an exposure or intervention, including therapeutic interventions. Molecular, histologic, radiographic, or physiologic characteristics are types of biomarkers. A biomarker is not an assessment of how an individual feels, functions, or survives" (BEST resource 2016).
>
> Qualification is defined as "a conclusion, based on a formal regulatory process, that within the stated context of use (COU), a medical product development tool can be relied upon to have a specific interpretation and application in medical product development and regulatory review" (BEST resource 2016).

Once a biomarker is qualified, it can be used for the qualified COU in drug development programs without the need for CDER to re-review the supporting information.

> The Context of Use (COU) is "A statement that fully and clearly describes the way the medical product development tool is to be used and the medical product development-related purpose of the use" (BEST resource 2016).

The biomarker COU is commonly defined early, as it is the basis of the level of evidence that is needed for qualification and may be modified as needed based on acquired data. The COU consists of a concise 'Use Statement' containing the biomarker's name, identity and proposed use in drug development, as well as the 'Conditions for Qualified Use', a comprehensive description of how the biomarker will be used in the qualified setting (FDA 2014). This may include assessment of susceptibility or risk, diagnosis of disease or disease subtype, assessment of prognostic outcome of interest, prediction or assessment of patient response or toxicity, and monitoring of therapy response or toxicity. It should be noted that the aims of the COU do not directly overlap with the indications for use of an FDA Premarket Approval Application (PMA) or Premarket Notification (510(k)) for IVD methods used in clinical practice.

The COU also helps to define the *fit-for-purpose* expectations for the validation of the biomarker assay. It sets forth the specific information and the quality of that information that must be provided by the biomarker measurement, and thus by the assay used to measure it. Since decisions will be made based upon the data generated, the assay must be rigorous and specific enough to support those decisions around the scientific validity of the biomarker's performance by heath authorities. In terms of regulatory requirements, and in keeping with the learning from previous guidance documents, method validation for biomarker assays should assess assay capability metrics that are similar to drug concentration assays.

Biomarker assays are required to measure changes in endogenous levels against a variable background, and to be able to accurately and precisely measure relevant changes in those concentrations to enable investigators to make informed decisions. As a result, the magnitude of the biomarker change from baseline to reach a medically actionable level will have a direct effect on the amount of acceptable variability in an assay. For example, if a biomarker has a baseline of 5 units and a medically relevant change in that biomarker is an increase of 2 units, a fit-for-purpose assay will have to be very precise and have only a small amount of total analytical error. However, if a medically relevant change is an increase of 200 units in that biomarker, then a lower level of precision and a higher amount of total analytical error may be acceptable to yield medically useful results. If the assay yields a result of 10 ± 6 in the first example, the data are not useful due to the variability associated with the result; in the second example, the result can be interpreted as a change in the biomarker that is not medically important. This topic is further discussed in the Assay Validation Acceptance Criteria section of this document, and put into the context of a Performance Standard for a biomarker assay and Allowable Error for the biomarker.

The COU will also define the expected reference interval for the assay. The reference interval is the range of values that can be interpreted by the assay. The concentration measurements generated by the assay are compared against a population based reference value (population reference interval) and are interpreted as being in range or out of range of normal. This can be influenced by endogenous factors such as age and sex, and exogenous factors such as exercise or fasting. Genetics, geographical location, different laboratories, and different statistical methods can also impact the proposed reference interval. The COU will help to determine if the assay is appropriate for the population being tested, be that a normal healthy or diseased population or both, and each population will have an appropriate defined reference range.

Analytical Validation vs Clinical Validation

In the qualification of biomarkers, both analytical and clinical factors must be considered. Thus, both analytical validation, as it relates to the correct measurement of the biomarker, and clinical validation, as it relates to the predictive ability of the biomarker, are important. However, these concepts are easily confused and mistakenly combined into one concept.

> Analytical validation is the process of "Establishing that the performance characteristics of a test, tool, or instrument are acceptable in terms of its sensitivity, specificity, accuracy, precision, and other relevant performance characteristics using a specified technical protocol (which may include specimen collection, handling and storage procedures). This is validation of the test, tools, or instrument's technical performance, but is not validation of the item's usefulness." (BEST resource 2016).
>
> Clinical validation is the process of "Establishing that the test, tool, or instrument acceptably identifies, measures, or predicts the concept of interest." (BEST resource 2016)

Therefore, analytical validation includes all factors that are part of the assay and is dependent only upon the critical reagents and the test system, while clinical validation relates to the consistency and predictive accuracy of the biomarker in predicting the outcome claimed. This should not be confused with clinical utility, which expresses to what extent diagnostic testing improves health outcomes relative to the current best alternative (Bossuy et al, 2012).

Biomarker Analytical Assay Validation and Fit-for-Purpose

As stated in the Biomarker Qualification and the Context of Use section of this document, the COU helps to define the fit-for-purpose expectations for the validation of the assay. Fundamentally, all valid bioanalytical assays are fit-for-purpose based on their defined application. The remainder of this document is dedicated to defining the appropriate level of characterization and validation that should be expected for assays used for biomarker qualification.

The goal of biomarker assay development is to construct an assay that is not too simplistic, nor too rigorous, for the goals of the investigation. The term fit-for-purpose is often used in this context. Properly defined, fit-for-purpose is "A conclusion that the level of validation associated with a medical product development tool [assay] is sufficient to support its context of use" (BEST resource 2016). However, too often the term is used inappropriately and without sufficient rationale, labeling assays as such without correlating the level of validation with the assay's purpose.

Assays that measure biomarkers seeking qualification are used to produce the evidence required to establish and confirm decision points, and therefore should undergo full validation to ensure that assay performance and application match (Table 1). A fully validated assay would be required in all confirmatory biomarker qualification studies including the establishment of reference intervals and/or decision points. The fit-for-purpose process can be used to develop an assay that is accomplishing clinically what is necessary and relevant. The concept and proper implementation of fit-for-purpose has been thoroughly summarized by Lee et al. (2006) and Lee (2009). This is an iterative process, where data informs the further development of the assay (Table 1). The fit-for-purpose process involves four iterative, continuous steps with the intended use of the biomarker data as the driving force for the analytical validation. These steps include method development, exploratory method qualification, method validation, and in-study method validation (Lee et al. 2006).

Table 1: Fit-For-Purpose Approaches for Biomarker Assays Validation

	Exploratory Validation	Partial Validation	Full Validation*
Decision level	Rank ordering, screening	Candidate selection	High risk actionable data
State of development	Discovery	Translational Research	Clinical
Reference Standard	• When available, or surrogate	• When available, or surrogate	• Requires reference standard or surrogate
Matrix	• Authentic or surrogate	• Authentic or surrogate matrix • Spiked reference calibrator • Consider disease state, multiple donors • Test parallelism	• Authentic or surrogate matrix • Spiked reference calibrator • Consider disease state, multiple donors • Test parallelism
Standard and Quality Control Accuracy and Precision criteria	• Acceptance criteria not needed • Established based on evaluation results	• Acceptance criteria based on evaluation results and technology-based analytical considerations • Native animal/human samples as quality control samples	• Acceptance criteria based on evaluation results and technology-based analytical considerations • Native animal/human samples as quality control samples
Accuracy and Precision qualification	• Not required	• Minimum one runs	• Six runs
Stability evaluation	• Bench top • Scientific judgment	• Collection, room temperature, freeze/thaw, and long term stability • Reference standard or matrix stability test with acquired animal/human samples	• Collection, room temperature, freeze/thaw, and long term stability • Reference standard or matrix stability test with acquired animal/human samples
Data output	• Qualitative • Semi-quantitative	• Relative quantitative • Semi-quantitative • Absolute quantitative	• Absolute quantitation • Relative quantitative • Semi-quantitative • Qualitative

* - Assays that measure biomarkers seeking qualification are used to produce the evidence required to establish and confirm decision points, and therefore should undergo full validation to ensure that assay performance and application match

History of Guidance Documents and White Papers

Multiple guidance documents have been published for PK/bioequivalence and IVD assay development and validation. These documents contain nearly all of the fundamental concepts necessary for the development and validation of biomarker assays for use in the qualification of DDTs. However, the direct application of these concepts for biomarker assay validation has not been codified. The lessons learned and knowledge gained in the development of these guidance documents can be used to build a new, more relevant document that is directly applicable to biomarker qualification. Below is an overview of the currently available

documents related to establishing guidelines for the validation of biomarker qualification assays.

In 2013 the FDA published a Draft Guidance for Industry on Bioanalytical Method Validation (FDA, 2013). Originally issued in 2001, this guidance was revised to reflect advances in science and technology related to validating bioanalytical methods. The draft guidance was then open to public review and comment. The joint FDA/American Association of Pharmaceutical Scientists (AAPS) Crystal City V Meeting was held in Baltimore from December 3-5, 2013, to continue the feedback/comment process. Although both FDA and industry representatives presented, and consensus was reached on a number of issues (Booth et al., 2015), a final guidance has yet to be issued. In fact, the Crystal City VI Workshop in 2015 was prompted to clarify residual concerns pertaining to validations for Ligand Binding Assays (LBA) and Liquid Chromatography-Mass Spectrometry (LC-MS) assays (Lowes and Ackerman, 2016; King, Arnold et al., *in press*).

Currently, specified criteria for PK and biomarker assay performance outlined in the 2013 Bioanalytical Method Validation Draft Guidance (FDA 2013) are being considered for assays to support biomarker measurement. In the draft guidance document it states that "*Method validation for biomarker assays should address the same questions as method validation for PK assays. The accuracy, precision, selectivity, range, reproducibility, and stability of a biomarker assay are important characteristics that define the method. The approach used for PK assays should be the starting point for validation of biomarker assays, although FDA realizes that some characteristics may not apply or that different considerations may need to be addressed.*" These considerations include expectations around assay accuracy, precision, selectivity, interferences, reproducibility, acceptable total allowable error, sample/matrix stability, etc. However, other factors, such as the nearly ubiquitous lack of certified reference materials, inability to utilize *in vivo* samples as controls, and the necessity to provide most measurements as relative quantitative rather than absolute quantitative, may limit assay characterization and potentially impacts clinical decision making. With the goal of ensuring reliable clinical conclusions, the level of analytical rigor and industry standard practices for validation of biomarker assays have been debated and promulgated for some time (Lee et al. 2006).

To date, industry engaged in biomarker qualification efforts have been referred to regulatory guidance documents (Table 2) for analytical assay performance for biomarkers in biological matrices, which originated within Center for Devices and Radiological Health (CDRH) as part of the 510(k) and PMA submission processes. However, these guidance documents have an entirely different purpose from the qualification of biomarkers as DDTs. A 510(k) is a premarketing submission made to FDA to demonstrate that the device to be marketed is as safe and effective, or substantially equivalent (SE), to a legally marketed device that is not subject to premarket approval (PMA). Premarket Approval (PMA) is the most stringent type of device marketing application required by FDA. Unlike premarket notification, PMA is to be based on a determination by FDA that the application contains sufficient valid scientific evidence that provides reasonable assurance that the device is safe and effective for its intended use or uses. This represents a higher level of rigor than is required for

assays/biomarkers that are to be used in a limited fashion under the well-controlled environment of drug development.

Table 2: Listing of CLSI Guidelines Related to the Validation of Biomarker Assays

CLSI EP05-A3	Evaluation of Precision of Quantitative Measurement Procedures; Approved Guideline – Third Edition
CLSI EP06-A	Evaluation of Linearity of Quantitative Measurement Procedures: A Statistical Approach; Approved Guideline
CLSI EP07A2	Interference Testing in Clinical Chemistry; Approved Guideline – Second Edition
CLSI EP09-A3	Measurement Procedure Comparison and Bias Estimation Using Patient Samples; Approved Guideline – Third Edition
CLSI EP17-A2	Evaluation of Detection Capability for Clinical Laboratory Measurement Procedures; Approved Guideline – Second Edition
CLSI EP21-Ed2	Evaluation of Total Analytical Error for Quantitative Medical Laboratory Measurement Procedures – Second Edition
CLSI EP28-A3c	Defining, Establishing, and Verifying Reference Intervals in the Clinical Laboratory; Approved Guideline – Third Edition

Table 3 is a comparison of assay characteristics across biomarker, PK, and IVD methods. Although conceptually there is significant overlap among these assays, there are also meaningful differences among them. The most obvious similarity is between biomarker and IVD methods with respect to trueness, accuracy, and precision. It is also clear that PK methods are different from biomarker and IVD methods. However, this table fails to point out one important similarity between PK and biomarker assays and that is their exclusive application in drug development. While IVD assays can be used in drug development, their primary application is in clinical practice.

In conclusion, although accepted guidelines that outline the key aspects of biomarker assay development and validation already exist (e.g. draft guidance documents, and CLSI), these guidelines are most relevant for PK assays and IVD methods, and not for biomarker qualification specifically. Furthermore, there is a need for specific guidance for the validation of assays used in the clinical evaluation of biomarkers to enable qualification of the biomarker as a DDT.

Table 3: Comparison of Selected Biomarker, Drug Concentration, and IVD Assay Characteristics

	Biomarker Assay	Drug concentration Assay	In Vitro Diagnostic Assay
Trueness	A link between biomarker (target) and biological effect of interest is established for routine measurement	Drug (target) concentration is quantitated	The measured biomarker (target) must be linked to the biological effect being characterized
Accuracy or Trueness?	➢ Well characterized reference standards are rare ➢ Target is not always homogeneous ➢ Blank matrix is not always available ➢ Endogenous biomarker molecule may have unique modifications that has significant influence on biological effect and therefore may not be directly simulated by recombinant reference material.	➢ Well characterized reference standard always available ➢ Target is homogeneous ➢ Blank matrix is always available ➢ In vivo modifications has little influence on assay quantitation	➢ Well characterized reference standard often available ➢ Target is not always homogeneous ➢ Blank matrix is not always available ➢ In vivo modifications may have significant influence on biological effect and may need to be separately quantitated
Parallelism	To confirm that calibrators react in the assay in the same way as the endogenous molecule	Calibrators are prepared with the exact xenobiotic (drug) that is being measured.	To confirm that calibrators act in the assay in the same way as the endogenous molecule
Precision	Heightened requirement to ensure DDT accuracy or clinical decision making is effective	Relative requirement to ensure analytical accuracy	Heightened requirement to ensure clinical accuracy

Assay Development and Validation

In order to develop this document, several key assumptions were made and are outlined below regarding the nature and use of assays for qualification of soluble biomarkers measured in biological matrices.

1. The validation expectations for biomarker qualification assays are not identical to the expectations outlined for drug concentration or toxicokinetic assays.
2. The performance characteristics of biomarker qualification assays are in line with the COU and ultimately the application of the biomarker as a DDT.
3. Assays used to generate data for biomarker qualification are not sufficient for, and are not intended to be used as, *de facto* substitutes for an IVD or for approvals or clearances by CDRH.
4. Although an FDA approved or cleared assay is not required to support a biomarker qualification effort, adequate assay performance and validation is essential for a biomarker's qualification.
5. Qualified biomarkers and the performance expectations for the associated assays are suitable for use in drug development and regulatory submissions but are not assumed to be directly acceptable in, or transferrable to, clinical practice.

Assay Considerations

Assay Design, Development, and Validation

Assay design usually begins with determining the critical success factors; however, this is not the case for novel biomarker assays as historical information in normal and diseased populations is rarely available. Setting definitive acceptance criteria for the desired analytical precision and total error *a priori* may not be possible. However, working criteria can be used *a priori* to track and develop assay performance, as arbitrary criteria should be avoided. Once these parameters have been determined, the assay reagents can be acquired and the instrument testing can be performed. Proof of Concept (POC) experiments are then conducted to establish preliminary assay parameters. Further development of the assay proceeds with optimization, defining pre-analytical factors, followed by validation, in preparation for implementation of the developed methodology.

Analytical validation involves documenting that the performance characteristics of a method are sufficiently suitable and reliable for the intended analytical applications through the use of specific laboratory investigations. The acceptability of analytical data corresponds directly to the criteria used to validate the methodology. The validation of biomarker qualification assays requires determination of four specific characteristics.

1. The selectivity and specificity of the assay - Is the proper analyte being measured?

2. The limits of the measurement - What is the upper and lower limit (sensitivity) of quantitation or detection and what actions are performed when biomarker results exceed these limits?
3. The variability of the measurement - What is the accuracy and precision/reproducibility associated with the measurement?
4. Stability – How do sample handling conditions (i.e. parameters) affect the sample before measurement?

As stated previously, biomarker qualification assays are by definition being used to make confirmatory or clinical decisions, and therefore should undergo a full validation to ensure that assay performance and application match. As defined by the fit-for-purpose paradigm, a full validation requires a reference standard or in most cases a surrogate. An authentic or surrogate matrix, with a spiked reference standard or surrogate, is used for the determination of accuracy, precision, limit of detection, limit of quantitation, specificity, linearity and reportable range, ruggedness, and robustness. Absolute quantitation, not qualitative data, is required for the output.

Pre-Analytical

Pre-analytical evaluations include assessment of the influence of the anticipated providence of a specimen prior to analysis. Pre-analytical variables include specimen collection, processing, storage, shipment, and handling that affect the integrity of the biological specimens, and later the results of analyses. Pre-analytical variables can introduce inconsistency into assay results, either systematically or randomly, resulting in lack of reproducibility. Not only must these factors be taken into consideration early in the assay development phase, well prior to the full validation of the assay, but they must be established and remain consistent between the validation samples and the qualification samples. Table 4 lists examples of the pre-analytical factors that can affect quantitation of biomarkers.

Table 4: Pre-Analytical Factors to be Considered

Sample Type	Serum, plasma with platelets, platelet-free plasma, neat urine, centrifuged urine, CSF, exosomes as source of protein markers
Overall Collection Parameters	Collection method, volume, reproducibility, replicates, future use in other platforms, timing of sample, draw order
Collection Tube	Type of tube; minimize protein adherence, plastic leaching; breakage;
Collection Variables	Proper mixing; use of additive, preservative, and/or anticoagulant (e.g. clot activator, EDTA, heparin, thrombin, sodium citrate, acid citrate)
Centrifuging	Fixed rotor or bucket, refrigeration, maximum g force necessary, adjustable brake settings, size of tubes, availability at all sites
Post Collection Variables	Collection and immediate storage temperature, minimization of time not stabilized, requirements for protection from light
Identification of abnormal samples	Hemolysis, icterus, lilpemia, etc. annotated if observed and appropriately triaged
Logistics of transport	Temperature (shipping on wet ice, dry ice), permits for human or primate blood, manifests, upright shipping, light exposure
Storage considerations	Long term aims, micro-aliquots (<500 ml), desiccation, oxidation, sublimation, temperature (-4°C, -20°C, -70°C, -80°C, LN2)

Assay Performance

In this section, fundamental assay parameters are discussed as they relate to determining the analytical validity of a biomarker assay. As outlined in the draft PK bioanalytical guidance (FDA, 2013), basic bioanalytical parameters already exist that should be considered when developing an assay for the qualification of biomarkers. It should be noted that not all parameters will be applicable for every biomarker assay, but each should be considered based on the biomarker COU. Good scientific judgment must be used, while keeping the COU in mind at all times. Different platforms will have different requirements for the assessments of performance criteria and may have other considerations beyond this minimal list, or may not include some members of this list. If a parameter is not addressed, a justification should be formulated for why it was excluded.

When considering the performance and rigor of criteria required for biomarker assay validation, it is essential to understand the purpose and clinical requirement of that assay as they relate to the biomarker's COU. Early in the exploration of a biomarker's utility, a simple and minimally validated assay may be sufficient to generate informative data. However, when qualifying a biomarker, a fully validated assay will be needed to provide sufficiently robust data for confirmatory and clinical study sample analysis.

Validation is the confirmation via extensive laboratory investigations that the performance characteristics of an assay are suitable and reliable for its intended analytical use. It describes in mathematical and quantifiable terms the performance characteristics of an assay. At a fundamental level, the validation of a biomarker assay used for qualification should include the assessment of **Precision, Accuracy, Upper and Lower Limits of Detection, Limit of Quantitation, Specificity, Linearity and Measureable Range, Ruggedness and Robustness**. A refined list of analytical parameters that should be evaluated is included in Table 5, and is defined in greater detail in Appendix 1.

Table 5: Analytical Parameters to be Considered during Biomarker Assay Validation

Sensitivity	**Robustness**
Limit of Detection	**Ruggedness**
Lower Limit of Quantitation	**Stability**
Upper Limit of Quantitation	Bench top
Working range/Reportable range	Short term
Specificity /Selectivity	Long term
Accuracy/Trueness	Freeze-thaw
Bias	**Reproducibility**
Drift	Quality Control/
Qualification matrix	Characterization of reference
Precision	materials/commutability
Within sample	Spike Recovery
Within run	Linearity/Dilutional verification
Between days	Parallelism
Between operators	Interference
Between sites	**Standard/calibration curve range and model**
Between lots	

As with validation of all bioanalytical methods, a primary consideration is the number of replicates that will be required during the validation of a biomarker assay. Table 6 gives a range of expectations based upon the guidance documents and standards for PK and IVD method validations. This table has been derived from information condensed from guidance documents and pivotal scientific publications; individual reference documents should be consulted for additional detail and justification. For assays being used to support biomarker qualification, an approach similar to that outlined for the CDER Bioanalytical Full Method Validation in Table 6 is appropriate.

Table 6: Comparison of Regulatory Validation Expectations

	Crystal City White Papers Partial Method Validation [a]	CDER Bioanalytical Full Method Validation [b,c]	CDRH 510(k) [d]	CDRH PMA [d]
Controls	3	6	2	3
Duplicates	2	2	2	2
Replicates	5	5	-	-
Sites	1	1	2	3
Operators	1 [e]	1 [e]	2	3
Reagent Lots	1	1	2	3
Calibration Cycles	0	0	5	5
Runs	6	6	2 [f]	2 [f]
Days	3 [g]	3 [g]	20	20
Runs/Day	1	1	2	2
Min. Obs./ Sample	60	120	640	2160

[a] White Papers – DeSilva (2003), Viswanathan (2007), Lee (2007), Lee (2009); [b] FDA Bioanalytical Method Validation Final 2001; [c] FDA Bioanalytical Method Validation Draft 2013; [d] Harmonized w/ CLSI Approved Guideline Method Evaluation Protocol EP05-A3; [e] DeSilva (2003), Viswanathan (2007), Lee (2006), Lee (2009) recommend two (2); [f] Two runs per day (AM & PM) for 20 days yielding a total of 40 runs; [g] Not per day, but over three days, ergo a total of 6 runs

Method precision and accuracy are performance characteristics that describe the magnitude of random errors (variation) and systematic error (bias) associated with repeated measurements of the same homogeneous sample (pooled with or without biomarker spiked in) under specified conditions. Method accuracy, within-run precision, and between-run precision should be initially established during method development, followed by confirmation during pre-study validation. However, biomarkers rarely have well-characterized reference standards. Therefore, precision and accuracy parameters are established from patient samples or spiked control material. When biomarker samples are being analyzed across multiple laboratories, inter-laboratory reproducibility should also be considered. Table 7 provides a guide for evaluating inter-laboratory reproducibility. However, in cases where only a single laboratory is utilized to conduct biomarker qualification analysis, there is no need to demonstrate inter-laboratory reproducibility.

Table 7: Considerations for Evaluating Inter-laboratory Reproducibility

	Multiple laboratory	Single laboratory
	Validation Sample Replicate Expectations	
Controls	6	6
Duplicates	2	2
Replicates	5	5
Sites	2-3	1
Operators	2-3	1
Reagent Lots	2-3	1
Calibration Cycles	5	0
Runs	40	6
Runs/Day	1	1
Min. Obs./Sample	≥640	120

System Suitability, Assay Format and Detection System

Before beginning assay development, decisions on assay format and the detection system should be based on the characteristics of the analyte. These decisions can be influenced by factors such as the necessary assay sensitivity, the available reagents, and the volume of sample that the study will provide. The system/equipment check is commonly measured by injecting replicate standards on a GC, HPLC, or MS, or detecting known positives with a kit assay.

In-Study Validation and Sample Analysis Acceptance Criteria

During study sample analysis, precision and accuracy should be continuously monitored in order to ensure that the assay continues to perform as per predefined specifications in each study run. As described by Lee *et al*., (2006), this entails the use of quality control (QC) samples, typically at three levels (low, mid, and high concentration) of the analyte, with at least two replicates at each level. Ideally, the QC samples used in the in-study sample analysis phase should be prepared identically to the validation samples used during the assay's validation, although this is not an absolute necessity. It is important that the approaches for assessment of method performance during the generation of qualification data are suitable for the intended purpose. Similar to assay validation, the acceptance criteria for biomarker assay performance will depend heavily on the intended use of the assay and should be based on physiological variability as well.

Assay Validation Acceptance Criteria

Determining assay acceptance criteria for biomarker assays is likely the most challenging exercise for a biomarker assay validation. Unlike the predefined acceptance criteria established for small and large molecule PK assays, the acceptance criteria for biomarker assays are

dependent upon each biomarker's physiological behavior, similar to the validation approach used for IVD methods.

As discussed by Lee *et al.* (2006), the fit-for-purpose status of a biomarker method is deemed acceptable if the assay is capable of discriminating changes that are statistically significant from the intra- and inter-subject variation associated with the biomarker. If the assay is not capable of such discrimination, either the assay lacks the appropriate analytical attributes or the biomarker is not suitable for the proposed purpose. For example, an assay with 40% total error determined during validation may be adequate for statistically detecting a desired treatment effect in a clinical trial for a certain acceptable sample size, but this same assay may not be suitable for a clinical trial involving a different study population that has much greater physiological variability.

An assay's performance characteristics are considered to be acceptable if (1) appropriate assay characterization practices are applied (evaluation of assay precision, accuracy, limit of detection, limit of quantitation, specificity, linearity and range, ruggedness and robustness), and (2) the assay can distinguish biomarker changes that are outside of the normal variability. Of course, it is desirable to have a well-performing, fully validated assay so that if additional analytical error is introduced into the assay, the biomarker's performance will not suffer.

In order to further understand an assay's tolerance in the event of additional bias, the concept of **Performance Standard** (PS) (from the CLSI guidance documents) has been applied (CLSI EP21-Ed2). As both the assay and the biomarker's intrinsic physiological behavior are the primary sources of variability in demonstrating the utility of a biomarker and its qualification, both of these sources of error must be taken into account. This approach is outlined below by defining a minimal **Performance Standard** (PS) for the biomarker.

Performance Standard is defined by the amount of **Allowable Error** (E_A) for the biomarker at the **Decision Level** (X_C).

$$PS = E_A \text{ at } X_C$$

Allowable Error is the amount of error that can be tolerated without invalidating the medical usefulness of the result.

Decision Level is any concentration of the analyte that is critical for medical interpretation (e.g. diagnosis, monitoring and therapeutic decisions).

For biomarkers, **Allowable Error** can be derived from intra-individual biological variation of the biomarker itself, and the magnitude of the biomarker's change from baseline in response to a valid biological stimulus or medically significant event. Thus, the biomarker's minimal **Performance Standard** can be used as a guide to set criteria for the acceptability of the **Total Error** associated with the assay.

Total Error (E_T) is the sum of all systematic bias and variance components that affect a result (i.e., the sum of the absolute value of the **Bias** (B) and **Intermediate Precision** (P_I) of the biomarker assay). This reflects the closeness of the test results obtained by the biomarker assay to the true value (concentration) of the biomarker.

$$E_T = B + P_I$$

Bias is any systematic error that contributes to the difference between the mean of a large number of test results and an accepted reference value.

Intermediate Precision is the within-laboratories variation based on different days, different analysts, different equipment, etc.

Finally, performance criteria can be formulated to judge the acceptability of an assay's performance by comparing the observed analytical **Total Error** to the specification for the **Performance Standard**.

> Performance is acceptable when observed analytical **Total Error** is less than the **Performance Standard** ($E_T < PS$).
>
> Performance is not acceptable when observed analytical **Total Error** is greater than the **Performance Standard** ($E_T > PS$).

Using this approach, biomarkers with a high degree of biological variability and lower amplitude of response to stimulus would require an assay with relatively low **Total Error**. While higher **Total Error** would be acceptable for assays with biomarkers that have low biological variability and higher amplitude of response to stimulus.

The concept of a **Performance Standard** for a biomarker in conjunction with an assay's **Total Error** also allows for the establishment of stability and interference thresholds. Both lack of stability and assay interference introduce bias into an assay and directly contribute to **Total Error**. As described above, if either of these factors result in the **Total Error** exceeding the **Performance Standard,** the performance of the assay would be considered unacceptable.

Case study: Kidney Safety Biomarker Clinical Validation

A collaboration between the Foundation for the National Institutes of Health (FNIH) Biomarkers Consortium Kidney Safety Biomarker Project Team and the Critical Path Institute Predictive Safety Testing Consortium Nephrotoxicity Working Group (FNIH BC/PSTC) resulted in the first successful qualification of safety biomarkers for nephrotoxicity. Partial results have been presented to the FDA, European Medicines Agency (EMA) and Japan's Pharmaceuticals and Medical Devices Agency (PMDA). The initial briefing package was submitted to the FDA in April 2011. The project was titled "Qualification of Translational Safety Biomarkers for Monitoring Renal Safety in Clinical Drug Development Research Trials." This work was designed to extend support for the translational utility of five urinary kidney safety biomarkers: albumin, total

protein, kidney injury molecule-1 (KIM-1), cystatin C (CysC) and clusterin. Each biomarker was qualified by the FDA, EMA and PMDA for use in rat studies during drug development. This work was also intended to provide support for the clinical utility of three additional urinary kidney safety biomarkers: N-acetyl-β-D-glucosaminidase (NAG), neutrophil gelatinase-associated lipocalin (NGAL) and osteopontin (OPN).

The COU for the clinical kidney safety project is as follows: *Qualified renal safety biomarkers are proposed to be used together with conventional kidney biomarker monitoring (e.g., sCr, BUN) in early clinical drug development research (under an IND or CTA) to support conclusions as to whether a drug is likely or unlikely to have caused a mild injury response in the renal tubule at the tested dose and duration. The study population was healthy volunteers and patients with normal renal function, taking into account age and gender. Proposed biomarkers are a Composite measure (CM) of urine CLU, CysC, KIM-1, NAG, NGAL, and OPN*.

Assay parameters and critical success factors for the bioassays kits were defined. In Table 8, Table 9 and Table 10 the assay parameters and critical success factors for the NGAL bioassay are summarized. For the calibration (standard) curve assessment, the calibrators were prepared according to the kit manufacturer's instructions in each case. Each standard curve contained a minimum of six non-zero calibrators, analyzed in duplicate, covering the entire reportable range (including LLOQ), excluding blanks (FDA, 2013). The standard curve was then fit to the simplest regression model providing an appropriate or best statistical fit (FDA, 2013). A minimum of six runs were conducted over at least two days (FDA, 2013). Acceptance criteria for the standard curve were set for ±25% of the nominal value of the standard calibrator concentration at the LLOQ and ±20% of the nominal value at all other concentrations on the curve (FDA, 2013 as starting point, fit-for-purpose for final criteria). ≥75% of non-zero standards must meet the criteria, including LLOQ (FDA, 2013). The total analytical error (accuracy and precision) must be ≤30% (fit-for-purpose).

QC samples were prepared by collecting normal donor urines (six total), prepared by a standard protocol. These were collected, centrifuged, aliquoted and frozen at -80°C. The endogenous analyte concentration was determined for each donor sample individually prior to pooling. To create the Low QC pool (LQC), urine from two donors within three times the LLOQ was pooled. To create the Middle QC pool (MQC), urine from two donors in the assay midrange was pooled. To create the High QC pool (HQC), urine from two donors testing at approximately 70-75% of the high range of the expected study sample concentrations (if available) was pooled. If high range samples were not available, recombinant protein for each biomarker was spiked in to the urine to reach the needed range.

For the precision assessment, a minimum of three (≥3) QC concentrations (LQC, MQC and HQC) in the range of expected study sample concentrations was tested. The precision determined at LQC, MQC and HQC could not exceed ±20% CV, and the precision determined at the LLOQ could not exceed ±25% CV.

Quality control samples were included in each run. A minimum of three (≥3) concentrations of QCs were measured in duplicate per run. The minimum number of QCs required to be analyzed

was the greater of ≥5% of the number of test samples, or six total QCs. The run was accepted if ≥2/3 of QC results were within 20% of respective nominal (theoretical) values (total ≥ four out of six pass) and ≥50% of QCs at each level were within 20% of their respective nominal values, i.e., no QC may fail both replicates (FDA, 2013 as starting point, fit-for-purpose for final criteria).

Spike Recovery (Relative Accuracy) was measured using a minimum of five determinations per concentration, and a minimum of three concentrations of known spiked materials (low, mid, and high) in the range of expected study sample concentrations. Mean values were accepted if within 20% of actual values, except at the LLOQ, where mean values were accepted within 25% of actual values.

The LLOQ (sometimes referred to as Functional Sensitivity) was established by a minimum of five samples generated by dilution of QCs or calibrators. When possible, the appropriate matrix was used for the dilutions, otherwise PBS was used as the diluent. A minimum of five analyses over a minimum of six analytical runs was used to generate the data. The mean, SD and % CV were calculated, and the LLOQ defined as back-calculated concentration of lowest calibrator that did not exceed a 20% CV [recovery ±25%] (FDA, 2013 as starting point, fit-for-purpose for final criteria).

The ULOQ was established by a minimum of five assay runs of highest standard curve calibrator. Mean, SD and % CV were calculated, and the ULOQ defined as back-calculated concentration of lowest calibrator that did not exceed a 20% CV [recovery ±20%] (FDA, 2013 as starting point, fit-for-purpose for final criteria).

The dilutional linearity was determined using a minimum of two urine samples diluted with the appropriate assay diluent to create 7 to 11 evenly distributed samples covering the assay range. Samples were measured in duplicate. To be acceptable, recovery must be within 80-120% of the expected concentration.

Sample Stability was determined using at least two samples (low and high in assay range). Samples were stored for at least 24 hours at -80°C per cycle. The acceptability for change from baseline was ≤20%. Bench-top stability was designed to mimic intended laboratory sample handling conditions (time and ambient temperature) used during sample analysis. For freeze and thaw stability, a minimum of three freeze-thaw cycles were conducted, designed to mimic intended sample handling conditions used during sample analysis. Long term storage stability at -80°C has been carried out past one year and is still ongoing (fit-for-purpose criteria).

Interference Studies were also conducted in accordance with CLSI EP07-A2. Clinically significant differences are difficult to assess for novel urine biomarkers. Thus, an empirical number of five replicates were tested with acceptance criteria set at ±20% of expected value. A minimum of five normal urine samples were pooled and analyzed for each biomarker. In addition, two sub-pools were created by spiking with exogenous analyte (as needed) to create low, normal and high ranges. These sub-pools were split into control pools and test pools. Testing was conducted by addition of drug interferences at highest expected concentration in urine. Five aliquots each of the two test sub-pools, and five aliquots of the control pool were analyzed,

with test and control samples analyzed in duplicate in alternating order. The observed interference was calculated as the difference of test and control samples. Acceptance was within 20% of controls. Interfering substances tested were appropriate for urine specimens in general (erythrocyte, hemoglobin and total protein), as well as disease-specific or treatment related compounds.

With respect to the validity of the assays for use in qualification, each of the assays were appropriately characterized (as described above) and for each of the assays their respective biomarker changes were determined to be outside of the normal variability in response to nephrotoxicity. Thus, these assays are deemed acceptable for use in the qualification of the proposed panel of kidney safety biomarkers. Although the assay clearly distinguished biomarker changes that are outside of the normal variability, in most cases there is little separation between upper limit of normal and the decision point. Thus, the assay Total Error represents the maximal Total Error acceptable for any assay used to measure the biomarkers and there is little tolerance for the additional of more variability into the method.

Table 8: Pre-Analytical Factors Considered during the Validation of Neutrophil Gelatinase-Associated Lipocalin (NGAL)

Sample Type	Neat, centrifuged urine
Interference	Erythrocytes, hemoglobin, lysed leukocytes. Exercise, high protein meals, dehydration and other factors that may elevate urine creatinine used for normalization could bias results.
Overall Collection Parameters	Spot, clean catch, mid-stream.
Collection Tube	Sterile collection cup with no preservatives.
Collection Variables	Maintain sample at room temperature; process and freeze within 4 hours of collection.
Centrifuging	2000xg for 10 minutes, discard pellet
Post Collection Variables	Document processing steps and time between collection and time in freezer.
Identification of abnormal samples	Microscopy of an aliquot of sample to rule out contamination with red or white blood cells is recommended. If samples are visibly colored, strip test for esterase and hemoglobin must be performed.
Logistics of transport	Transport on dry ice.
Storage considerations	Freeze at -70 to -80◦C. Avoid temporary storage at -20◦C.

Table 9: Analytical Parameters Evaluated during the Validation of Neutrophil Gelatinase-Associated Lipocalin (NGAL)

Sensitivity	**Robustness**
Lower Limit of Quantitation	**Ruggedness**
Upper Limit of Quantitation	**Stability**
Working range/Reportable range	Short term
Specificity /Selectivity	Long term
Accuracy/Trueness	Freeze-thaw
Bias	**Reproducibility**
Drift	Quality Control
Precision	Spike Recovery
Within sample	Linearity
Within run	Dilutional verification
Between lot	Interference
	Standard/calibration curve range and model

Table 10: Summary of the Neutrophil Gelatinase-Associated Lipocalin (NGAL) Validation

	Bioanalytical Full Method Validation
Controls	3
Replicates	2
Sites	1
Operators	1
Reagent Lots	1
Runs	6
Days	2
Runs/Day	1

Finally, considerations for inter-laboratory reproducibility were not addressed in the validation of the kidney safety biomarker assay as all confirmatory analyses (samples for evaluation of reference ranges, decision points and confirmatory studies) were conducted at a single laboratory.

Conclusions

The validation of biomarker assay performance is integral to the biomarker qualification process for DDTs. While guidance documents for assay validation exist, they cannot all be broadly generalized to the validation of assays used in the qualification of biomarkers. Biomarkers are by nature endogenous compounds analyzed in the context of fluctuating

background concentrations. *In vivo* samples can rarely be used as controls. Currently, certified reference materials are scarce to nonexistent, depending on the biomarker. Therefore, multiple analytical factors must be taken into account when designing the assays, given that the consequences by definition impact clinical decisions. To ensure reliable clinical conclusions, the level of analytical rigor and quantity of generated data must be based primarily on the biomarker-specific COU. A fully validated assay, as defined by fit-for-purpose criteria, is required for assays used in the qualification of biomarkers. This includes the definition of reference ranges, establishment of decision points, and confirmation of the biomarker's predictive accuracy. An assay's performance characteristics are considered to be acceptable if: (1) appropriate assay characterization practices are applied (evaluation of assay precision, accuracy, lower and upper limits of quantitation, specificity, linearity and range, ruggedness and robustness); and (2) the assay can accurately distinguish biomarker changes that are outside of the range of normal variability.

References

Amur SG, Sanyal S, Chakravarty AG, Noone MH, Kaiser J, McCune S, Buckman-Garner SY. 2015. Building a roadmap to biomarker qualification: challenges and opportunities. Biomark Med. 2015;9(11):1095-105. doi: 10.2217/bmm.15.90.

Arnold, M.E., Booth, B., King, L. et al. AAPS J (2016). Workshop Report: Crystal City VI—Bioanalytical Method Validation for Biomarkers. doi:10.1208/s12248-016-9946-6

BEST (Biomarkers, EndpointS, and other Tools) Resource [Internet] Glossary. FDA-NIH Biomarker Working Group, Available at: http://www.ncbi.nlm.nih.gov/books/NBK338448/ Accessed on August 23, 2016. Published January 28, 2016, last updated April 28, 2016.

Booth B, et al. (2015). *Workshop Report: Crystal City V—Quantitative Bioanalytical Method Validation and Implementation: The 2013 Revised FDA Guidance.* The AAPS Journal, Vol. 17(2). DOI: 10.1208/s12248-014-9696-2.

Bossuyt PM, Reitsma JB, Linnet K, Moons KG. 2012. Beyond diagnostic accuracy: the clinical utility of diagnostic tests. Clinical Chemistry 58:12 1636-1643.

CLSI EP05-A3: *Evaluation of Precision of Quantitative Measurement Procedures; Approved Guideline – Third Edition*. ISBN (1-56238-967-X). CLSI, 940 West Valley Road, Suite 140, Wayne, PA 19087-1898 USA, 2014.

CLSI EP06-A: *Evaluation of Linearity of Quantitative Measurement Procedures: A Statistical Approach; Approved Guideline.* ISBN (1-56238-498-8). CLSI, 940 West Valley Road, Suite 140, Wayne, PA 19087-1898 USA, 2003.

CLSI EP07A2: *Interference Testing in Clinical Chemistry; Approved Guideline – Second Edition.* ISBN (1-56238-584-4). CLSI, 940 West Valley Road, Suite 140, Wayne, PA 19087-1898 USA, 2005.

CLSI EP09-A3: *Measurement Procedure Comparison and Bias Estimation Using Patient Samples; Approved Guideline – Third Edition.* ISBN (1-56238-888-6). CLSI, 940 West Valley Road, Suite 140, Wayne, PA 19087-1898 USA, 2013.

CLSI EP17-A2: *Evaluation of Detection Capability for Clinical Laboratory Measurement Procedures; Approved Guideline -- Second Edition.* ISBN (1-56238-795-2). CLSI, 940 West Valley Road, Suite 140, Wayne, PA 19087-1898 USA, 2012.

CLSI EP21-Ed2: *Evaluation of Total Analytical Error for Quantitative Medical Laboratory Measurement Procedures, 2nd Edition.* ISBN (1-56238-940-8). CLSI, 940 West Valley Road, Suite 140, Wayne, PA 19087-1898 USA, 2016.

CLSI EP28-A3c: *Defining, Establishing, and Verifying Reference Intervals in the Clinical Laboratory; Approved Guideline – Third Edition.* ISBN (1-56238-682-4). CLSI, 940 West Valley Road, Suite 140, Wayne, PA 19087-1898 USA, 2010.

Food and Drug Administration. 2001. Guidance for industry bioanalytical method validation. Available at: http://www.fda.gov/downloads/Drugs/.../ucm070107.pdf. Last Updated May 2001. Accessed on August 23, 2016

Food and Drug Administration. 2013. Guidance for industry bioanalytical method validation. Available at: http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm368107.pdf. Last Updated September 2013. Accessed on August 23, 2016.

Food and Drug Administration. 2014. Biomarker Qualification Context of Use. Available at: http://www.fda.gov/Drugs/DevelopmentApprovalProcess/DrugDevelopmentToolsQualificationProgram/ucm284620.htm Last Updated 9/15/14. Accessed on August 23, 2016.

Food and Drug Administration. Biomarker Qualification Program. 2016. Available at: http://www.fda.gov/Drugs/DevelopmentApprovalProcess/DrugDevelopmentToolsQualificationProgram/ucm284076.htm Last Updated on July 11, 2016. Accessed on August 23, 2016.

Lee J, et al. 2006. *Fit-for-Purpose Method Development and Validation for Successful Biomarker Measurement.* Pharmaceutical Research. DOI: 10.1007/s11095-005-9045-3.

Lee J. 2009. Method validation and application of protein biomarkers: basic similarities and differences from therapeutics. Bioanalysis 1(8), 1461-1474. 10.4155/BIO.09.130. ISSN 1757-6180.

Lowes and Ackerman. 2016. *AAPS and US FDA Crystal City VI workshop on bioanalytical method validation for biomarkers.* Bioanalysis (2016) 8(3), 163–167.

U.S. Department of Health and Human Services Food and Drug Administration Center for Drug Evaluation and Research (CDER). 2016. Considerations for Use of Histopathology and Its Associated Methodologies to Support Biomarker Qualification. Guidance for Industry. Available at: http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm285297.pdf Accessed on August 23, 2016.

Viswanathan CT, Bansal S, Booth B, DeStefano AJ, Rose MJ, Sailstad J, et al. Workshop/conference report—quantitative bioanalytical methods validation and implementation: best practices for chromatographic and ligand binding assays. AAPS J. 2007; 9(1):E30–42.

Appendix 1. Assay Performance Characteristics Definitions

Sensitivity

Sensitivity is the ability to detect the target analyte within the matrix of interest, and practically speaking is the limit of quantitation (see below) of the calibration/standard curve. This can be influenced by interferences in the matrix, affinity of antibodies, etc. Sensitivity is commonly measured by determining the limit of quantitation.

Specificity

Specificity is the ability to assess unequivocally the target analyte in the presence of components or homologs which might be expected to be present. The specificity of an assay is the capability of the assay to differentiate similar analytes or organisms or other interfering compounds from matrix elements that could have a positive or negative effect on the assay value. Antibody Specificity (Interference) is a related concept. For antibody assays, the specificity of the antibody to the epitope adds another layer of specificity to consider. For example, does the detecting antibody pick up epitopes on related molecules other than the analyte of interest? Specificity can be influenced by the similarity of the analyte to other compounds in the matrix or assay materials and can be method/platform dependent. Specificity is commonly measured by sample controls at various concentrations spanning the expected range, with and without the potential interfering substance.

Spike recovery

Spike recovery is the process of comparing the amount of analyte present in a sample after a standard has been added to and extracted from the sample, as compared to the true concentration of the standard added. This can be influenced by the sample type, the means of collection, the preparation and extraction procedure, the chemical properties of the analyte, and the stability of the analyte. Spike recovery is commonly measured by measuring the extraction efficiency of the analyte using an internal standard and showing that it is consistent, precise, and reproducible at more than one concentration.

Accuracy (Relative)

Accuracy is the closeness of the agreement between the result of a measurement and true value of the measure. In practice, an accepted reference value is substituted for the true value. Accuracy can also be expressed as %bias, and is also called Trueness or Bias. This requires a "gold" standard or method but in the absence of a gold standard or method, comparison to established reference laboratory's results may substitute. Accuracy is influenced by the number of measurements (i.e., fewer measurements are usually less accurate than more measurements that can then be averaged). Relative accuracy is commonly measured by comparing the value found for an unknown, to that of a known value of reference material, in replicate samples, preferably in the expected range of concentrations.

Accuracy = (Actual value - (Actual value - Measurement)) / Actual value

Bias

Bias is any systematic error that contributes to the difference between the mean of a large number of test results and an accepted reference value. Thus, it refers to the degree of trueness between an average of a large series of measurements and the true value of the measurement.

Precision

Precision is the closeness of agreement between independent test results obtained under stipulated assay conditions. Precision is usually expressed as imprecision using the standard deviation (SD) or % coefficient of variation (CV) of the results of a replicate set of experiments. Precision includes within assay variability, repeatability (within-day variability), and reproducibility (day-to-day variability). Precision may be established without the availability of a "gold" standard as it represents the scatter of the data rather than the exactness (accuracy) of the reported result.

- Repeatability (of results of measurements) or within sample, measuring closeness of the agreement between results of successive measurements of the same measure, carried out under the same conditions of measurement
- Intra-assay Precision (within assay) and within a single run
- Repeatability (within run precision) measuring precision same method on identical test material in the same laboratory by the same operator using the same equipment within a short interval of time
- Inter-assay Precision (between assay) measuring precision with time, and including different analysts, labs, reagents, equipment, etc.

Precision is influenced by differences in assay conduct and equipment. Precision is commonly measured by measuring multiple replicates of several known concentrations. This can be done between different assays, different days, different laboratories, different analysts, etc.

Intermediate Precision (also called within-laboratory precision)

A measure of precision under a defined set of conditions: same measurement procedure, same measuring system, same location, and replicate measurements on the same or similar objects over an extended period of time. It may include changes to other conditions such as new calibrations, operators, or reagent lots.

Standard/calibration curve range and model

Multiple concentrations of the analyte in the matrix of interest are measured and the simplest mathematical model that can be used to fit a straight line is used to create the standard or calibration curve. This provides a means to determine the concentration of unknown samples

that fall within this range of concentrations. This can be influenced by the affinity of the detection antibodies, the signal to noise ratio of an instrument. A calibration curve is commonly measured by using at least 5 or 6 concentrations of the analyte covering the expected range of the assay, in the matrix that is going to be used, including a blank (no analyte).

Detection Limit or limit of detection (LOD)

Detection Limit or limit of detection (LOD) is the lowest amount of analyte which can be detected, but not necessarily quantitated as an exact value. The detection limit is a low concentration that is statistically distinguishable from background or negative control, but is not sufficiently precise or accurate to be quantitated. This can be influenced by interference of other compounds in the matrix or limitations of the detection methods being used. LOD is commonly measured by determining a minimum signal to noise ratio based on blank samples and samples with known but low concentrations of analyte.

Lower Limit of Quantitation (LLOQ) and Upper Limit of Quantitation (ULOQ)

Limits of Quantitation are the lowest (LLOQ) and highest (ULOQ) concentrations of an analyte in a sample that can be quantitatively determined with suitable precision and accuracy. The LLOQ is often defined by an arbitrary cut-off such as a ratio of signal-to-noise, equal to 1:10, or a value equal to the mean of the negative control plus 5 times the standard deviation of the negative control values. More precise experimental determinations of an assay LLOQ include repeated measurements of samples with low and very low analyte concentrations in several independent experiments and the determination of the LLOQ value using predefined criteria based on precision and recovery of the sample measurement.

Stability

- Bench top

  Samples should be checked for stability for at least the length of time they are anticipated to be at room temperature after thawing or before freezing while being prepared for analysis.

- Freeze-thaw stability

  Generally repeated freeze/thaw cycles should be avoided whenever possible and samples should only be thawed if directly used for measurements or if required for production of aliquots. The stability of an analyte needs to be shown for repeated freeze-thaw cycles if it is expected that samples will be repeatedly frozen and re-measured.

- Short-term stability

  Conditions used in stability experiments should reflect situations likely to be encountered during actual sample handling and analysis of a biomarker. These include usual handling and

processes, and assay processing time to simulate the time samples will be maintained at a certain temperature for analysis.

- Long-term stability

Long-term analyte stability testing can be a complex task due to the need to define biomarker stability under storage conditions and to judge the adequacy of the assay method to monitor stability changes. Procedures need to include an evaluation of analyte stability in the stock solution. Ideally, the storage time in long-term stability evaluations should exceed the time between the date of first sample collection and the date of the last sample analysis. Sufficient samples should be banked to allow longer time points and bridging to cross validate assays as the need might arise.

Stability under all conditions can be influenced by time, temperature, humidity, the presence of degrading enzymes, the natural half-life of the biomarker, storage conditions, the matrix, and the container system. Stability is commonly measured by comparing stored samples under realistic conditions to a set of samples prepared fresh from a stock solution of standard at known concentrations in an interference free matrix. The stored samples should be sufficiently similar in concentration to the freshly prepared stock standards to generate meaningful and decision-worthy data.

Working range/reportable range

Range is the concentrations of analyte or assay values between the low and high limits of quantitation. Within the assay range, linearity, accuracy and precision are acceptable and shown to be valid. This can be influenced by the factors measured above, as well as the overall performance of the assay. Range is commonly measured by examining the low and high limits of quantitation of the assay.

Selectivity/interference

Selectivity is the ability of the assay to determine the identity of the analyte definitively in the presence of the other materials present in the matrix. Usually signal suppression is more common than enhancement, but in both cases the source of the interference is the concentration of cross-reacting, interfering substances. If the lack of selectivity comes from a known source, it is referred to as interference; if it comes from an unknown source, it is referred to as matrix effect (Lee and Hall, 2009). This can be influenced by other endogenous substances, metabolites, decomposition substances, or other xenobiotics or proteins concomitantly administered. Selectivity is commonly measured by analyzing multiple blank samples of matrix and attempting to find the analyte of interest. If the analyte cannot be detected, the assay is selective.

Quality Control/Reproducibility

Method precision and accuracy are performance characteristics that describe the magnitude of random errors (variation) and systematic error (mean bias) associated with repeated measurements of the same homogeneous (spiked) sample under specified conditions. Method accuracy, intra batch (within-run) precision, and inter batch (between-run) precision should be established preliminarily during method development and confirmed in pre-study validation. However, biomarkers rarely have fully characterized reference standards so these parameters are established from patient samples or spiked control material.

Method robustness/ruggedness is part of this reproducibility. Ruggedness is the reproducibility of the assay under a variety of normal, but variable, test conditions. Variable conditions might include different machines, operators, and reagent lots. Ruggedness provides an estimate of experimental reproducibility with unavoidable error. Robustness is a measure of the assay capacity to remain unaffected by small but deliberate changes in test conditions (e.g., incubation time, temperature, sample preparation, buffer pH, and potential interfering substances). Robustness provides an indication of the ability of the assay to perform under normal usage. Reproducibility conditions are conditions where test results are obtained with the same method on identical test items in different laboratories with different operators using different equipment (ISO 5725- 1). Reproducibility (of results of measurements) is the closeness of the agreement between the results of measurements of the same measure, carried out under changed conditions of measurement (VIM93).

Characterization of reference materials (and stability)

If available, WHO reference material can be used for calibration of an assay. However, reference materials are rarely available and a surrogate, such as patient samples or spiked control material, must be used.

Linearity/ Dilution verification/Parallelism

Linearity is the ability of the assay to return values that are directly proportional to the concentration of the target analyte or pathogen in the sample. The linear assay range is considered the most responsive and provides the most reliable quantification. Mathematical data transformations, to promote linearity, may be allowed if there is scientific evidence that the transformation is appropriate for the method. It is acknowledged that the dose response curve of a large number of immunoassays reflects a sigmoidal response characteristic and not a strict linear analyte-signal response behavior, but can still allow determination of analyte concentrations. Related to this is parallelism, or a condition in which dilution does not result in biased measurements (trending up or down) of the analyte concentration. Linearity can be influenced by matrix effects, protein binding, or metabolism of the biomarker. Parallelism likewise can be caused by metabolism or protein/serum binding and both can be tested for by assessing incurred samples against a number of dilutions of standard (if available) over the same range. If a standard is not available, serial dilutions of several high concentration samples over several concentrations could be used.