Points to Consider Document: Scientific and Regulatory Considerations for the Analytical Validation of Assays Used in the Qualification of Biomarkers in Biological Matrices

June 11, 2019

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

Steven P. Piccoli, GlaxoSmithKline and John Michael Sauer, Critical Path Institute

Contributing Authors:

Brad Ackermann, Eli Lilly; John Allinson, Immunologix Laboratories; Mark Arnold, Covance;Shashi Amur, U.S. FDA; Jiri Aubrecht, Takeda; Amanda Baker, Ventana Medical Systems, Inc.; Robert Becker, U.S. FDA; ShaAvrée Buckman-Garner, U.S. FDA; Jennifer Burkey, Critical Path Institute; Martha Donoghue, U.S. FDA; Carmen Fernandez-Metzler, PharmaCadence; Fabio Garofolo, Angelini; Russ Grant, LabCorp; Huidong Gu, Bristol-Myers Squibb; Vinita Gupta, Exelixis; Steve Gutman, Kylie Haskins, U.S. FDA; Illumina; John Kadavil, U.S. FDA; Yan Mao, Boehringer Ingelheim; Nicholas King, Critical Path Institute; Omar Laterza, Merck; Jean Lee, BioQualQuan; Steve Lowes, Q2 Solutions; Vasum Peiris, U.S. FDA; Steven P. Piccoli, GlaxoSmithKline; Mark Rose, CHDI; Afshin Safavi, BioAgilytix; John Michael Sauer, Critical Path Institute; Shelli Schomaker, (retired); Rick Steenwyk, Pfizer (retired); Lauren Stevenson, Immunologix Laboratories; Meena Subramanyam, Takeda; Matt Szapacs, GlaxoSmithKline; Faye Vazvei, Merck; Sue Jane Wang, U.S. FDA; Jianing Zeng, Bristol-Myers Squibb;

Commenters:

Candace Adamo, Pacific Biomarkers, Inc.; Abbas Bandukwala, U.S. FDA; Montserrat Carrasco-Triguero, Genentech; Lou Christodoulou, UCB; Anonymous, European Bioanalytical Forum; Swati Gupta, Allergan; Paula Katavolos, Genentech; Kellie Kelm, U.S. FDA; Dan Krainak, U.S. FDA; Adora Ndu, Biomarin; Corinne Petit-Frère, Roche; Hugo Vanderstichle, ADx NeuroSciences; Parya Nouri, Pfizer; Phillip Turfle, U.S. FDA; Yoshiro Saito, National Institute of Health Sciences; Sofia Stinchi, Merck; Yan Zhang, Bristol-Myers Squibb

Table of Contents

List of Tables

List of Figures

Foreword

Points to Consider Document: Scientific and Regulatory Considerations for the Analytical Validation of Assays Used in the Qualification of Biomarkers in Biological Matrices

This Points to Consider document was originally designed to establish consensus on the expectations for the validation of assays used in the regulatory qualification of fluid biomarkers. The scope of this document was to define the scientific and regulatory considerations for the analytical validation of assays for fluid-based (any protein, peptide, lipid or other chemical entity soluble in plasma, urine, saliva, etc.) biomarkers used in the regulatory qualification of drug development tools (DDT's). However, the scope quickly expanded, and the document is now a comprehensive user guide to the analysis of biomarkers in drug development. At its core, this document contains a complete description of necessary approaches that can be applied to nearly every analytical situation that will be encountered in fluid-based biomarker qualification.

It is important to note that this document should not be thought of as a check list where all points listed must be experimentally evaluated for a given biomarker assay. Instead, only the analytical elements directly relevant to the biomarker of interest and its Context of Use (COU) in drug development should be considered. And based on these considerations, the analytical elements should either be experimentally evaluated, or a rationale should be developed for their lack of evaluation. Consideration of the individual analytical elements based on the intended use of the biomarker truly embraces the fit-for-purpose approach to assay validation. Fit-for-purpose, as defined in biomarker qualification, is a conclusion that the level of assay validation associated with a biomarker is sufficient to support its context of use (i.e., the way the biomarker will be used for regulatory decision making in drug development).

Four major areas must always be considered for the validation of assays to be used in the qualification of biomarkers regardless of the biomarker assay or platform used for analysis.

1. Defining pre-analytical conditions
2. Setting analytical performance requirements for assay
3. Characterizing and documenting assay performance
4. Establishing assay validation acceptance criteria

Furthermore, during the development of the assay validation plan, seven key analytical parameters must also be considered.

1. Accuracy (Relative)
2. Analytical Measurement Range
3. Parallelism
4. Precision
5. Selectivity
6. Specificity

7. Stability (sample)

The robustness of the experimental evaluation of these key analytical parameters may be varied based on the characteristics of the biomarker and its intended clinical application. However, each of these parameters should be well understood in terms of their individual impact on the reliability of decision making associated with use of the biomarker. This Points to Consider document provides in depth information on the four major areas and each of these seven key analytical parameters, including figures and examples to aid in assessing their impact on assays that will be utilized for biomarker qualification.

Finally, it should be stated that this document was created at the behest of many biomarker stakeholders including scientists from the U.S. Food Drug Administration. It represents the efforts of a diverse, dedicated, and expert working group that fully utilized current scientific peer-reviewed literature, input from discussion sections at scientific meetings, and public expert opinion. The result of this exercise has been the development of a valuable consensus document that outlines best practice approaches that can be applied to the development, characterization, and validation of assays to support fluid biomarker qualification. It is our intent to update this Points to Consider document routinely as the science in the field evolves. It is our goal that this document will serve as a resource for analytical and biomarker scientists to aid in biomarker assay validation.

Steven P. Piccoli and John Michael Sauer

Introduction

The evolutionary process for the eventual use of a decision-driving biomarker in drug development must utilize a biomarker assay which has undergone complete analytical validation and whose performance characteristics are therefore sufficiently well understood to precisely specify the capability of the assay to establish the value of the target biomarker as a qualified Drug Development Tool (DDT) as defined by the FDA. To ensure reliability and reproducibility of the data generated to support biomarker qualification, assays should be analytically validated before confirmatory clinical validation studies for the biomarker are performed, and analytically re-validated during confirmatory clinical validation studies whenever changes to the assay may significantly impact assay performance. It is important that the assay procedure and resulting measurements are suitable for the assay's intended purpose. Measurement errors that could result in biases and negatively affect the biomarker's predictive accuracy would thus limit its utility.

Inherent in the measurement of biomarkers is that biomarkers are endogenous entities or molecules unlike the measurement of xenobiotics (drugs which do not have an endogenous counterpart). Therefore, biomarker assays typically measure an increase or decrease in the endogenous level of the molecule which often fluctuates because of individual variability in physiology, disease biology, pathology, comorbidities, treatment administered, and environmental factors. In addition, both accessibility of biomarker samples (not always easily obtainable or fluid-based), as well as potential challenges in assessibility of the biomarker (molecular isoforms, e.g., proteolytic cleavages, glycosylation, lipidation, phosphorylation, nitration, oxidation; and the omnipresent scourge of inadequate reference standards), must be given due deliberation. It is therefore key to understand and describe which specific isoform is to be quantified and its exact relationship to the pathological form in subjects. Given these factors, the acceptance criteria and expectations for assays used in the qualification of biomarkers must take into consideration 1) the type of molecule being measured and 2) the context in which the biomarker is being applied in drug development and in regulatory decision making. Some principles have been proffered in draft form for the specific intent of companion diagnostics (FDA 2016a).

The key acceptance criteria for the analytical validation (*hereafter, unless otherwise specified, the term validation refers exclusively to analytical validation; for exact definition, see* Analytical Validation vs Clinical Validation) of pharmacokinetic (PK) assays (i.e., drug concentration), and for *in vitro* diagnostic devices (IVD) used in clinical practice, have been well defined but are not universally transferable or applicable to biomarker assays as DDTs. This is because the expectations (both clinical and analytical) for assays used to support biomarker qualification are distinct. While the criteria used in the validation of drug concentration assays used in clinical practice can be applied as a limited framework for the development of criteria for biomarker assay validation, they cannot be adopted unequivocally due to the broad spectrum of technologies utilized in biomarker measurements, each with disparate technical requirements. Thus, the analytical validation of assays used to generate data for biomarker qualification must be refined to fit the proposed drug development context of use.

The goal of this document is to define the scientific and regulatory considerations for the analytical validation of assays for fluid-based (any protein, peptide, lipid, or other chemical entity soluble in plasma, urine, saliva, etc.) biomarkers used in the qualification of DDTs. It does not address exploratory activities, i.e., those activities not designed for regulatory submission but for internal decision making only, nor does it address the use of biomarkers as primary or secondary endpoints in clinical trials. The topics to be discussed include considerations for assay design and technology selection, optimization of pre-analytical factors, core assay performance expectations, and setting minimally acceptable assay performance criteria. Technology areas covered include singleplex ligand binding assays (LBA) and immunometric assays, singleplex and certain[1] multiplex mass spectrometry assays, and enzyme-based (kinetic rate) assays. Out of scope of this document are immunohistochemistry (IHC), flow cytometry, imaging biomarkers, multiplex LBA, as well as nucleic acids, genetics, genomics, and transcriptomics. Multiplex LBAs should follow the same general principles but will require additional characterization of parameters not detailed here. Criteria for determining that analytical performance must be re-validated based on changes to the assay or determining that one assay has sufficient analytical performance to be substituted for another in the context of confirmatory clinical validation studies are also not within the scope of this document. Likewise, the development and analytical validation of assays to be used in commercial clinical practice (i.e., those regulated solely by Clinical Laboratory Improvement Amendments of 1988 [CLIA]), as well as assays used for measuring exploratory biomarkers in clinical drug development, are outside the scope of this document. However, the general analytical validation principles outlined in this document for biomarker assays may also be applicable to biomarker methods used in clinical development of biopharmaceutics.

The three primary areas of this document that require consensus and agreement by the contributing authors and supporting groups are 1) the experimental characterization of the assay used during qualification of a biomarker (Assay Design, Development and Validation), 2) the approach to defining the requisite assay performance and acceptance criteria (Assay Validation Acceptance Criteria), and 3) Total Analytical Error. It is recognized that multiple iterations of each may be necessary to achieve the final objectives.

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 2016b). 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 for use in drug development and regulatory decision-making. This program is described by the FDA both as publicly available information (FDA 2016b), as well as refereed literature (Amur et al. 2015). Terms used in

[1] The only time a multiplexed MS assay is within scope is when the analytical measurement range of all analytes fits within the dynamic range of the MS instrument, and all analytes follow identical extraction and sample handling procedures.

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) under BEST glossary.

> 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).
>
> 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).
>
> 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).
>
> Further clarification of qualification as a DDT has been provided in the 21st Century Cures Act:
>
> "… a drug development tool qualified under this section may be used for— ''(A) supporting or obtaining approval or licensure (as applicable) of a drug or biological product (including in accordance with section 506(c)) under section 505 of this Act or section 351 of the Public Health Service Act; or ''(B) supporting the investigational use of a drug or biological product under section 505(i) of this Act or section 351(a)(3) of the Public Health Service Act." (U.S. Congress 2016)

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

A biomarker's COU should be proposed early in the biomarker qualification process, at the Letter of Intent stage (FDA 2014a) (*note that this guidance document, though active, was written before the 21st Century Cures Act, and is in the process of being updated*), as it is the basis of the level of evidence that needs to be considered for qualification (FDA 2014b). The COU consists of a concise 'Use Statement' containing the BEST biomarker category and proposed use in drug development. The BEST biomarker categories in drug development include susceptibility/risk, diagnostic, monitoring, prognostic, predictive, pharmacodynamic/response and safety. Example of drug development uses include defining inclusion/exclusion criteria, supporting clinical dose selection, or defining treatment allocation arms. Examples of COUs are: a predictive biomarker to enrich for enrollment of a sub group of asthma patients who are more likely to respond to a novel therapeutic in Phase 2/3 clinical trials; a prognostic biomarker to enrich the likelihood of hospitalizations during the timeframe of a clinical trial in phase 3 asthma clinical trials; or a safety biomarker for the detection of acute drug-induced renal tubule alterations in male rats. It should be noted that the

aims of the COUs are specific to drug development and do not overlap with the indications for use of an FDA Premarket Approval Application (PMA) or Premarket Notification (510(k)) for IVD devices used in clinical practice. As such, there is a continuum of validation requirements to address the needs of qualifications resultant from disparate COUs. Biomarker qualification is not a regulatory decision on the assay(s) used in the qualification process. However, future use of the qualified biomarker must demonstrate assay performance characteristics sufficient to support the COU.

The COU determines the assay rigor by defining the use of the biomarker measurement in drug development. Since drug development decisions will be made based upon qualified biomarkers, the assay used to measure the biomarker must be robust, sensitive, specific and selective enough to support the specific decisions defined by the COU.

The intended use population defined by the COU will also determine the expected reference interval for the biomarker. The reference interval, or commonly, reference range, is the central 95% of the range of values present within the distribution (mean ± 1.96SD) of healthy subjects. If the distribution of the measurement of interest is not normally distributed (not Gaussian), transformation of the measurement may be explored until the distribution is at least approximately symmetric. The measurements generated by the assay are described as in range or out of range bounded by the upper and lower limits of the distribution of healthy subjects. The reference interval can be influenced by endogenous factors such as age and sex, comorbidities, and exogenous factors such as exercise or fasting. Genetics, geographical location, different laboratories, and different statistical analysis methods can also impact the reference interval. This may result in the reference range for the intended use population being different than the reference range for healthy subjects.

Assays for soluble biomarkers are required to measure changes in response to disease or treatment in endogenous concentrations or activities of biomolecules against a variable background found in the intended use population defined by the COU. It is important that the relevant changes in biomarker concentrations are measured as accurately and precisely as necessary to enable investigators and health authorities to make informed decisions. Therefore, the magnitude of the biomarker change from baseline to reach a relevant level (such as a cut-off value) will have a direct effect on the amount of acceptable analytical 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, an assay capable of appropriate discrimination must be very precise with only a small amount of Total Analytical Error (TAE). However, if a medically relevant change is an increase of 200 units in that biomarker, then a lower level of assay precision and a higher amount of TAE 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, this result is useful and can be interpreted as an important change in the biomarker that is not medically relevant. This determination is further compounded by intra- and inter-individual variation for normal and diseased states for the biomarker. This topic is further discussed in the Assay Validation Acceptance Criteria section and Appendix 3 of this document, and put into the context of a Performance Standard (PS) for a biomarker assay and allowable Total Analytical Error (aTAE) for the biomarker.

The COU will help to determine the performance characteristics for the assay based in part on the medical decision point for the population being tested, be that a normal or diseased population, or both, and each population will have an appropriately defined reference interval or cut-off value. Both a reference interval and a cut-off may separately be needed to make an informed decision when using a biomarker assay. Sometimes, the reference interval determines the cut-off, but not always. Rarely are reference intervals generated for disease populations, and receiver operating characteristic (ROC) curves are usually generated to define clinical sensitivity and specificity with appropriate cut-off values, reflecting the Positive and Negative Predictive Values (PPV and NPV) of the assay.

Analytical Validation vs Clinical Validation

In the qualification of biomarkers, both analytical and clinical factors must be considered. Thus, for biomarker qualification, demonstration of both analytical validation (as it relates to the accurate and precise measurement of the biomarker) and clinical validation (as it relates to the correct interpretation of the biomarker measurement for a specific COU) are necessary. 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)

Analytical validation supports the biomarker measurement and includes all factors that are part of the assay and is dependent only upon the acceptability of the samples, critical reagents, and the performance characteristics of the test system. Clinical validation supports the interpretation of the biomarker measurement and is dependent on the clinical performance (clinical sensitivity, clinical specificity, clinical accuracy) of the biomarker in predicting the outcome claimed. Clinical validation should not be confused with clinical utility, which expresses to what extent diagnostic testing improves health outcomes relative to the current best alternative (Bossuyt et al. 2012), or "The conclusion that a given use of a medical product will lead to a net improvement in health outcome or provide useful information about diagnosis, treatment, management, or prevention of a disease. Clinical utility includes the range of possible benefits or risks to individuals and populations." (BEST Resource 2016) This document focuses solely on the analytical validation of fluid-based biomarker assays used to generate data for biomarker qualification.

Biomarker Assay Validation and the Fit-for-Purpose Paradigm

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 of the assessments needed for the validation of the assay.

> Fit-for-purpose is a conclusion that the level of [assay] validation associated with a medical product development tool is sufficient to support its context of use (BEST Resource 2016).

Fundamentally, all valid bioanalytical assays are fit-for-purpose based on their defined application. The remainder of this document is dedicated to providing guidance to define 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 adequately meets the goals of the investigation. The term fit-for-purpose is often used in this context. 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. It may cover significant differences in study sample testing results, such as the magnitude of expected change which will determine assay precision requirements, but not establish a 95% CI for baseline/heathy patients or encompass biological variation within or amongst individuals.

Assays that measure biomarkers seeking qualification are used to produce the evidence required to establish and confirm decision points, and therefore should undergo sufficiently extensive and rigorous 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 ranges and biomarker response decision points.

The fit-for-purpose process can be used to develop an assay that is accomplishes what is necessary and relevant for the context of use. The concept and proper implementation of fit-for-purpose has been thoroughly summarized by Lee et al. (2006) and Lee et al. (2009). This is an iterative process, where data informs further development and refinement of the assay (Table 1). The fit-for-purpose process involves four continuous steps including method development, exploratory method validation, "full" or extensive method validation, and in-study method validation, in an iterative progression with the intended use of the biomarker data as the driving force for the analytical validation (Lee et al. 2006). This process must be directly related to and support the COU.

Table 1: Approaches for Biomarker Assay Validation

	Discovery/Exploratory Validation	Translational/Partial Validation	Full Validation *
Decision level (examples)	Screening (internal)	Candidate selection (internal)	Actionable data (external)
Drug development stage	Discovery	Translational Research	Clinical trials
Reference Standard	• When available, or surrogate	• When available, or surrogate	• Requires calibrator or reference standard or surrogate
Matrix	• Authentic or surrogate • Test parallelism if samples available	• Authentic or surrogate matrix • Spiked reference standard • 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 two runs	• Six runs for LBA and minimum three runs for MS assays (based on aTAE)
Stability evaluation* **	• Bench top • Scientific judgment	• Collection, room temperature, freeze/thaw, and long-term stability as needed • Matrix stability test with acquired animal/human samples	• Collection, room temperature, freeze/thaw, and long-term stability • Matrix stability test with acquired animal/human samples
Data output (Lee et al. 2006)	• Qualitative • Quasi-quantitative	• Qualitative • Quasi-quantitative • Relative quantitative	• Qualitative • Quasi-quantitative • Relative quantitative • (Absolute) quantitative**

*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

**For heterogeneous (i.e., large molecule) biomarkers, the calibrators are generally prepared with recombinant reference material in a surrogate matrix. The assay cannot provide absolute quantification; only relative accuracy can be evaluated. Thus, the term relative accuracy (rather than accuracy) is appropriate for nearly all biomarkers where the calibration material differs from the

endogenous biomarker. However, as the concept of accuracy is based on a comparison of test values to "true value", clinical outcome may serve as the true value.

*** For heterogeneous (i.e., large molecule) biomarkers, if spiked reference standards are used, the assay cannot provide insight on endogenous biomarker stability, only on the stability of the recombinant molecule.

History of Guidance Documents Relevant to Assay Validation

Multiple draft and finalized guidance documents have been published for pharmacokinetic (PK)/bioequivalence and IVD assay development and validation. These documents directly and indirectly recommend fundamental concepts necessary for the development and validation of biomarker assays for use in the qualification of DDTs. Although the 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 more comprehensive and relevant document that is directly applicable to biomarker qualification. Below is an overview of the currently available documents that should be considered regarding guidelines for the validation of biomarker qualification assays.

In 2001 CDER and the Center for Veterinary Medicine (CVM) at the FDA jointly published the "Guidance for Industry, Bioanalytical Method Validation" (FDA 2001). This document addressed the validation of methods for use in human clinical pharmacology, bioavailability, and bioequivalence studies requiring a PK evaluation. It described three types of validation (full, partial, and cross-validation) and identified key parameters recommended for validation: selectivity, accuracy, precision, recovery, calibration curve, and stability of analyte in spiked samples.

In September 2013, the FDA published a revised draft of the 2001 guidance "Guidance for Industry, Bioanalytical Method Validation" (FDA 2013). This draft guidance was intended to address recent advances in science and technology related to bioanalytical method validation, while still identifying a familiar list of fundamental parameters for method validation including accuracy, analytical measurement range, parallelism, precision, selectivity, specificity, and stability (sample). After the draft guidance was opened to public review and comment, the joint FDA/American Association of Pharmaceutical Scientists (AAPS) Crystal City V Meeting took place in Baltimore from December 3-5, 2013, to continue the feedback/comment process. A consensus was reached on several issues at this meeting (Booth et al. 2015). As a follow-up to Crystal City V, the AAPS Workshop Crystal City VI: Bioanalytical Methods Validation on Biomarkers was held in September 2015 in Baltimore to clarify residual concerns pertaining to validation of Ligand Binding Assays (LBA) and Liquid Chromatography-Mass Spectrometry (LC-MS) assays (Lowes and Ackermann 2016; Arnold et al. 2016) and to distinguish biomarker assay validation from these principles. The final FDA Bioanalytical Method Validation guidance was issued in May 2018 (FDA 2018).

On August 21, 2015, the University of Maryland's Center of Excellence in Regulatory Science and Innovation (M-CERSI), the U.S Food and Drug Administration (FDA), and the Critical Path Institute co-sponsored a symposium and proceedings paper titled "Evidentiary Considerations for Integration of Biomarkers in Drug Development" at the University of Maryland School of Pharmacy. The

objective of the symposium was to begin to define and ultimately codify the scientific and regulatory expectations for the qualification of safety and trial enrichment biomarkers. It was at this meeting that the concept of forming the analytical writing group responsible for this document was first discussed.

Currently, specified criteria for PK assay performance outlined in the 2018 Bioanalytical Method Validation Guidance (FDA 2018) are being considered for assays to support biomarker measurement. The final guidance states that "*Method validation for biomarker assays should address the same questions as method validation for drug assays. The accuracy, precision, sensitivity, selectivity, parallelism, range, reproducibility, and stability of a biomarker assay are important characteristics that define the method. The approach used for drug 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.*"

To date, the most comprehensive collection of guidance documents addressing analytical validation of biomarker assays cleared or approved as IVDs is that provided by the Clinical and Laboratory Standards Institute (CLSI) (www.clsi.org.) CLSI is a recognized standards development organization and has a well-defined process for issuing standards and other guidance documents, many of which are recognized internationally. CLSI has published dozens of documents addressing issues related to laboratory testing and the development of laboratory testing products for commercial distribution, some of which (Table 2) are directly relevant to this document. The Center for Devices and Radiological Health (CDRH) also has a formal process for standards recognition and has granted formal or informal, full or partial recognition, to a number of the CLSI published standards.

CLSI documents, although directed at parameters and metrics akin to those described in this document, have a different purpose but can still be a valuable resource during study design and data analysis by laboratories or industry to ensure that analytical validation experiments are robust and appropriately demonstrate the performance of the assay in the proposed COU.

This Points to Consider document is intended to address the performance of an assay used to generate data to support biomarker qualification. Testing in this setting is likely to be limited and performed under the well-controlled environment of drug development. Exploratory biomarkers are most often expected to be used as part of early feasibility testing to make decisions during development and identify likely drug candidates for further testing. However, qualified biomarkers are necessary to monitor safety of individual patients or help in the planning of more definitive clinical trials. These uses may allow for a validation that is suitable for its COU but is insufficient for a commercial diagnostic product being sold to multiple laboratories for potential use in multiple different settings.

Table 2: 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 EP07-A2	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
CLSI C62-A	Liquid Chromatography-Mass Spectrometry Methods; Approved Guideline

In addition to use of CLSI standards, the Office of In Vitro Diagnostics and Radiological Health (OIR) in CDRH has published over 100 guidance documents addressing a wide variety of products (OIVD Guidance Documents).

Although there are no free-standing documents addressing analytical validity per se, many of the product specific documents have sections outlining current thinking on best practices for establishing the analytical validity of different types of new tests and test technology.

Finally, if an unapproved or uncleared biomarker test is used for clinical decision making in the context of medical practice outside of a clinical trial, or under some circumstances in a controlled clinical study, it is termed a Laboratory Developed Test (LDT) (FDA 2014c) and becomes subject to oversight by the Clinical Laboratory Improvement Amendments of 1988 (CLIA) administered by the Centers for Medicare and Medicaid (CMS) and may be subject to FDA regulation if it is distributed. Oversight by CMS covers specific requirements for analytical performance, calibration, and quality control (CLIA manual).

In conclusion, there are already a number of guidance documents in place published by FDA, CLSI, and CMS to aid in establishing the analytical validity of biomarker assays. These have varying relevance for biomarker assays intended to support biomarker qualification, depending on the COU,

testing objectives, analytes of interest, and types of regulatory control that may be dictated by current government requirements. This document is intended to develop a practical and pragmatic approach to establishing analytical performance, specifically for use in biomarker qualification as a DDT.

Assay Design, Development and Validation

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

1. Assay design, technology selection considerations and the expectations for the performance characteristics of assays used in biomarker qualification are dependent on the COU and ultimately the application in drug development.
2. The analytical validation parameters for assays used in biomarker qualification are not necessarily identical to the expectations outlined for pharmacokinetic (drug concentration) or toxicokinetic assays.
3. Qualification of a biomarker does not indicate that assays used to generate the qualification data are approved or cleared by CDRH.
4. An assay to support a biomarker qualification effort is not required to be FDA approved or cleared. Assays which are approved or cleared will have a strictly defined intended use statement which may not match the desired COU as a DDT and may be considered an "off-label" use of the biomarker assay. Thus, the final validated method should meet acceptable performance characteristics to support qualification of a biomarker.
5. The performance characteristics of the assays used for qualifying biomarkers are considered suitable for use in drug development and regulatory submissions but are not assumed to be directly acceptable in, or transferrable to, regulated clinical practice, without clearance or approval by CDRH.

Assay Design and Technology Selection

The most important pre-requisite for assay design and technology selection is the definition of the COU of the biomarker. Full consideration of the COU will focus attention on practical considerations for the assay's design. For example, consideration of risks and standard of care practices for sample acquisition might affect assay design. Likewise, intended use under highly controlled or field-like conditions, for high volume or low volume testing, and with professionally trained or lay operators can drive design decisions enabling practical use of the assay. For some COUs, planning for iterative changes in the assay is needed (i.e., for testing volume, single vs. multiple sites, or for different/improved performance over the course of a drug development program).

Even though limited historical data for novel biomarkers may be available regarding endogenous levels and prevalence in normal and diseased populations, establishing the working criteria for the assay is foundational for the selection of appropriate detection technology, and for designing the assay format and selection of optimal reagents. Table 3 highlights points to consider in assay design and development.

Table 3. Points to Consider in Assay Design and Development

Biomarker	Stability of biomarker if known in disease conditions Biology, structure and isoforms
Context of Use	Application requirements
	Test population (e.g., human (healthy, disease), animal)
	Patient population comorbidities
	Sample acquisition
Use Environment	Lab vs. field
	Ruggedness / Robustness
	User training
	Maintenance
	Sample collection timing, methods, transport and storage
	Immediate vs stored analyses
	Sample preservation
	Single/multiple sites
	Single/multiple use (aliquot size; storage stability and freeze/thaw)
	Contamination effects (e.g., blood in CSF)
Assay features	Analyte(s) selection (measurands)
	Qualitative/Semi-Quantitative/Quantitative
	Calibrators/ reference material
	Controls (external, internal)
	Reportable range
	Reference interval
	Specimen volume/quantity requirements
	Desired analytical precision and aTAE, desired detection sensitivity – upper and lower limits and putative detection range
	Selectivity and specificity considerations including probable interference factors in endogenous matrix
	Results turn-around time
	Batch mode vs random access performance in automated clinical analyzers
	Automation
	Process software
	Analyte or reagent carry-over
	Analytical software, user interface
	Waste/hazard containment
	Cost
	Technical support requirements
Iteration (versioning, migration, convergence)	Platform/technology
	Interim data evaluation
	Assay refinement

Selection of a technology platform for biomarker detection will be primarily driven by the nature of the biomarker being measured (protein, lipid, etc.), the sensitivity and selectivity requirements, and the availability of the platform. The biophysical nature of the assay technology and the quality of the assay reagents will impact the absolute and relative measurements of the intended biomarker. Typically, plate- and bead-based assay formats and a variety of detection modalities including fluorescence, chemiluminescence, electrochemiluminescence, chromogenic detection, mass spectrometer-based assessments, and relatively new acoustic detection systems can be considered for the evaluation. Since most of the current automated technologies demonstrate acceptable precision, comparing various assay parameters using available reagents for biomarker detection with a given technology becomes a critical consideration for technology selection. A method comparison of performance between technology platforms may be assessed using a fixed set of assay reagents and normal and QC samples to estimate the reproducibility and relative error of back fit concentrations of the biomarker (spiked or preferably endogenous), preferably in specimens, else in relevant buffer matrix. Then the comparison should be extended to disease samples of interest to measure endogenous biomarker detection. Use of parallelism criteria will enable the identification of potential interference factors in the desired matrix. Another important consideration for technology selection includes scalability. Manual methodologies requiring high technical expertise may not be suitable for a biomarker method that requires global implementation. Likewise, the ease of use and the validation of the data processing software are important considerations.

Once the technology platform is selected, the assay can be optimized prior to finalizing the assay format using a checkerboard or design of experiment (DOE) approach (fractional factorial experiments, central composite designs, etc., as appropriate) to simultaneously evaluate multiple parameters (where applicable) such as minimum required dilution (MRD) of samples, assay reagent concentrations, calibrator levels, incubation periods, blocking and washing parameters, etc. Another important consideration in the assay design finalization phase is the selection of the most appropriate regression model for quantitative assays for the calibration curve (e.g., polynomial [linear, quadratic]; nonlinear models [e.g., four or five parameter logistic models, power model]) to assess the performance characteristics of the prototype method and to show acceptability of system suitability criteria. Further development of the assay may then proceed, defining pre-analytical factors, followed by validation, in preparation for implementation of the developed methodology.

Pre-Analytical

The following discussion is meant to provide points to consider but will not necessarily apply to each qualification submission. It is important to evaluate early within each project which pre-analytical factors are relevant and strive to find the appropriate balance of rigor necessary in a fit-for-purpose approach. Pre-analytical factors refer to all procedures that occur prior to sample analysis including sample collection, processing, transportation, and storage (See Table 4 for some examples). The physiology and/or patient specific characteristics of the human research participants are largely outside the control of the laboratory but can also have a significant impact on laboratory results. These include, but are not limited to, such factors as age, gender, ethnicity and ongoing diseases. Factors such as exercise, eating, drinking, and medication can also affect patient results. These factors should be thought of as part of the sample history and inclusion/exclusion criteria and should be documented as completely as possible.

Table 4 lists some examples of the pre-analytical factors in sample handling and processing that can affect quantitation of biomarkers. These variables can introduce inconsistency to assay results. Not only must these factors be taken into consideration with regard to the COU in specific populations early in the assay development phase prior to the full validation of the assay, but they must be established and remain consistent across assay validation, qualification, and post-qualification use. To ensure consistency, standard operating procedures, quality control indices, and criteria for sample acceptance or exclusion must be developed.

It should be appreciated that the pre-analytical factors may change across multiple assays for the same biomarker and need to be established for each assay. The pre-analytical factors for the same biomarker may also be different depending upon the biological matrix being analyzed. Not all biomarkers will be impacted by all factors, but as learning increases, documentation of earlier studies will make previously collected data interpretable. Table 4 is not meant to be exhaustive, but to provide common examples of factors to be considered, but not all are required for every application; sound scientific expertise and understanding must be utilized for each assay developed for a specific COU.

Table 4: Examples of Pre-Analytical Factors to be Considered

Pre-Analytical Factor	**Examples (not all inclusive)**
Sample Type	Whole blood (venous vs. capillary), cord blood, serum, plasma, platelet-poor plasma (PPP), platelet-rich plasma (PRP), peripheral blood mononuclear cells (PBMC), neat urine, centrifuged urine, saliva, ocular fluid, cerebrospinal fluid (CSF)
Interference	Endogenous: lipids (lipemia), hemoglobin (hemolysis), icterus (bilirubin), glucose, rheumatoid factors (immunoglobulins, C-reactive protein (CRP) Exogenous: drug interferences, OTC medications, skin disinfectants, collection tube additives/preservatives, bacterial contamination
Collection Procedure	Collection method (catheter vs. venipuncture vs fingerstick), type of needle, time of venous occlusion, collection site, volume, draw order, patient posture, adherence of staff to SOPs, timing of sample to pretreatment (protein inhibitors)
Collection Tube	Anticoagulant or preservative type and concentration (e.g. clot activator, EDTA, heparin, thrombin, sodium citrate, acid citrate, sodium fluoride, protein inhibitors), tube composition (low protein adherence, plastic leaching); breakage, proper tube labelling

Sample Collection Time	Time of day, frequency, fasting status
Collection Variables	Proper mixing; use of additive, preservative, and/or anticoagulant, temperature, light exposure, timing between collection and processing
Sample processing	Centrifugation (relative centrifugal force, angle/pelleting factor, time, braking, temperature), aliquoting (e.g. micro-aliquots < 500 µL relative to tube volume), storage tube material, closure, type of aliquot tube, de-salting, solid-phase extraction, adherence of staff to SOPs
Post Collection Variables	Collection and immediate storage temperature, minimization of time not stabilized, requirements for protection from light
Logistics of transport	Temperature (shipping on wet ice, dry ice), permits for human or primate blood, manifests, upright shipping, light exposure
Storage Considerations and Stability	Desired short- and long-term stability goals (timeframe), desiccation, oxidation, sublimation, temperature (-4°C, -20°C, -70°C, -80°C, -120°C, -196°C (liquid nitrogen)), freeze/thaw cycles
Freezing/Thawing Considerations	Rate of freezing (dry ice bath, air at freezer temperature, snap freezing in liquid nitrogen) Thawing temperature and rate (room temperature, 37° water bath, etc.), addition of stabilizers

Several resources and references have been developed to help identify and control sources of pre-analytic variation including CLSI guidelines which are listed in Appendix 2. The NCI Biospecimen Research Database (http://brd.nci.nih.gov) provides a compilation of primary literature that addresses biospecimen science. The Biospecimen Reporting for Improved Study Quality recommendations outline and prioritize elements for biospecimen studies (Moore et al. 2011). The International Society for Biological and Environmental Repositories Biospecimen Science Working Group developed a "Standard PREanalytical Code" (SPREC) that provides a common list of preanalytical variables for fluid samples and corresponding sample labeling system (code) that is intended to provide a generic format for specimen comparison (Betsou et al. 2010). Many of the measures implanted in clinical diagnostics as quality indicators for the preanalytical phase may also apply to the qualification setting. For example, the International Federation of Clinical Chemistry Working Group on Laboratory Errors and Patient Safety has defined quality indicators for the preanalytical stage (West et al. 2017). These are some of the many references available to aid the assay developer in controlling preanalytical variability.

The importance of documenting patient characteristics and understanding the influence of pre-analytical factors cannot be overemphasized. Standardized techniques for sample collection and handling need to be employed, quality control procedures developed, and personnel adequately trained to ensure sample integrity. It is important to realize that some retrospective/banked samples may not have been collected in a manner consistent with the pre-analytical conditions defined during assay validation. Indeed, some banked samples may have incomplete documentation which makes it difficult or impossible to establish the full history of the patient or status of the sample. Samples with incomplete documentation should be used with caution, and consideration given to not using the sample in question or flagging the results as questionable. In addition, storage stability may not be known for that duration of time.

Analytical Performance Requirements for Biomarker Assays

When considering the performance needs of a biomarker assay, it is expected that efforts are made to understand the biological variables which affect the biomarker. While it may be difficult to execute this to complete satisfaction, a process will be described to provide preliminary assessment for comparison of required precision and bias (relative accuracy) goals based upon biological variation of the biomarker levels in the intended population. When this approach is not technically feasible, a consideration for the "confidence" in measuring effect sizes is described (such as treatment over time without a control group or treatment versus control group), derived from analytical parameters determined during assay validation. These will be discussed in detail in the Appendix 1.

Assay Performance

Parameters for Validating Analytical Assay Performance Characteristics

In this section, the seven key analytical assay parameters needed to validate a biomarker assay performance are discussed. As outlined in the final PK bioanalytical guidance (FDA 2018), basic parameters have already been identified that should be considered when developing an assay for the qualification of biomarkers. It should be noted that not all parameters suitable for PK will be applicable or sufficient for a biomarker assay, but each should be considered based on the biomarker COU. If a parameter is to be included, a scientific rationale should be provided. If a parameter is not addressed, a justification should be formulated for why it was excluded at that time. The parameter may be added back as needed, such as if the COU changes. Different platforms will have different requirements for the assessment of performance criteria and may have other considerations beyond this list or may not include some parameters.

When considering the performance and rigor of criteria required for biomarker assay analytical validation, it is essential to understand the purpose and clinical requirements of that assay as they relate to the biomarker's COU. Early in the exploration of a biomarker's usefulness, a simple and minimally validated assay (see Table 1) may be sufficient to generate informative data. However, when qualifying a biomarker, a fully analytically validated assay will be needed to provide robust data for confirmatory and clinical study sample analysis.

Analytical validation is the confirmation via extensive laboratory investigations that the analytical performance characteristics of an assay are suitable and reliable for its intended use. At a fundamental level, analytical validation of a biomarker assay used for qualification should include the assessment of seven parameters: accuracy (relative), analytical measurement range (including LLOQ and ULOQ), parallelism (and dilutional linearity where appropriate), precision (inter-laboratory precision where appropriate, reproducibility), selectivity, specificity, and stability (Table 5). In some cases, information on additional analytical performance parameters may be needed, including trueness/accuracy, robustness, ruggedness, and occasionally drug interference assessment (Table 6). Detailed definitions of these measurements can be found in Appendix 1. Information on how to evaluate assays using these parameters can be found in the next section titled Assay Validation Acceptance Criteria and also in Appendix 1.

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

- Accuracy (Relative)
- Analytical Measurement Range
    - Lower limit of quantitation
    - Upper limit of quantitation
- Parallelism:
    - Minimum Required Dilution
    - Dilutional linearity
- Precision (Imprecision; intermediate precision, reproducibility)
    - Within run
    - Between runs
    - Between days
    - Between operators (if applicable)
    - Between lots (if applicable)
- Selectivity
- Specificity
- Stability (sample)
    - Bench top
    - Short term
    - Long term
    - Freeze-thaw

Table 6: Additional Analytical Parameters to be Considered during Biomarker Assay Validation

- Accuracy/Trueness
- Robustness
- Ruggedness
- Drug Interference

As with validation of all bioanalytical methods, a primary consideration is the number of samples that will be required during the validation of the biomarker assay. Table 7 gives a range of

expectations for evaluations of precision derived from information condensed from guidance documents and pivotal scientific publications; individual reference documents should be consulted for additional detail and justification. **Note that most of the values given are not expressly for biomarker determination and have a different COU defined and therefore are not directly comparable.** It is imperative that the source documentation be consulted before making assumptions, as the comparisons shown are for informational purposes only. However, the trend holds true that method validation for use in exploratory/feasibility studies requires the least amount of performance data; review of class III (high risk) medical devices requires the most.

Additional samples may be needed depending on the number of analytical parameters being characterized and the clinical context. For assays being used to support biomarker qualification, the approach outlined for the CDER Bioanalytical Full Method Validation in Table 7 is most appropriate but will vary according to the COU.

Table 7: Comparison of Regulatory Expectations for Precision Validation Studies

	Crystal City White Papers Partial Method Validation Exploratory / Feasibility Phase of Testing [a]		CDER Bioanalytical Full Method Validation For Use in Biomarker Qualification [b, c]		CDRH 510(k) For Clinical Use (Class II Medical Devices) [d]	CDRH PMA For Clinical Use (Class III Medial Devices) [d]
	LBA	LC-MS	LBA	LC-MS	LBA	LBA
Controls, analytical (validation QC)	3	6 (Lo, Mid, High in duplicate)	6 (3 levels in duplicate)	20 (LLOQ, Lo, Mid, High in 5 replicates)	2	3
Duplicates, analytical (Std)	2	1	2	1	2	2
Replicates, sample (for precision)	5	Det'd with QC's	5	Det'd with QC's	-	-
Sites	1	1	1	1	2	3
Operators	1 [e]	1	1 [e]	1	2	3
Reagent Lots	1	1 [h]	1	1 [h]	2	3
Runs	6	3	6	3/6 [b,c]	2 [f]	2 [f]
Days	3	3	3	3	20	20
Runs/Day	2	1	2	1	2	2

[a] White Papers – DeSilva (2003), Viswanathan (2007a,b), 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 (2007a,b), 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; [h] For hybrid LBA/LC-MS assays

Method precision and relative 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 (native or spiked pools or spiked non-physiological matrices, in decreasing order of preference) under specified conditions (see Appendix 3). Within-run precision, between-run precision, and relative accuracy should be initially established during method development, followed by confirmation during pre-study validation. However, unlike small molecules where absolute quantification may be possible, protein biomarkers rarely have well-characterized reference standards or calibrators. Therefore, precision and relative accuracy parameters are often established either using patient or subject samples (preferably), or a surrogate such as the most appropriate recombinant control material spiked into a blank or appropriate surrogate matrix (normal or synthetic). In the absence of reference methods or materials, sponsors developing assays to support biomarker qualification should have well-defined and well-characterized surrogates for reference standards according to the needs of the COU to ensure performance of the assay of interest remains consistent over time. However, it is expected that there will be experience with, and data generated from, samples containing adequate endogenous biomarker levels before proceeding to qualification.

It should be noted that if testing is being performed for the purpose of individual clinical decision-making within clinical trials, such as patient dose selection, assay performance will be subjected to oversight by the CLIA administered by the CMS. CLIA requires laboratories to establish and test analytical performance and to assure constant test performance by carrying out calibration verification using samples with known values at 6-month intervals. This repeat testing is possible only if there is a reliable source of a reference or surrogate standard for assessment of calibration drift.

When biomarker samples are being analyzed across multiple laboratories, both intra- and inter-laboratory reproducibility should be evaluated. Table 8 provides a guide for comparing sample requirements for inter-laboratory versus intra-laboratory reproducibility based on CLSI EP9. In cases where only a single laboratory is utilized to conduct biomarker validation or qualification, there may be no need to demonstrate inter-laboratory reproducibility. However, inter-instrument reproducibility may be applicable. These numbers reflect diagnostic standards and are shown only for comparative purposes.

Table 8: Considerations for Evaluating Inter-laboratory vs. Intra-laboratory Reproducibility (CLSI EP9)

	Multiple laboratories	Single laboratory
	Validation Sample Replicate Expectations	
Controls	6	3
Duplicates	2	1-2
Replicates	5	1-5
Sites	2-3	1
Operators	2-3	1
Reagent Lots	2-3	1
Runs	40	6
Days	20	2-3
Runs/Day	2	1

System Suitability, Assay Format and Detection System

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

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. However, a more difficult question is the nature of the appropriate validation samples.

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, the biomarker is not suitable for the proposed purpose, or the study size / subject selection is inappropriate. For example, an assay with 40% aTAE (determined during validation) may be adequate for statistically detecting a desired treatment effect in a clinical trial for a certain acceptable sample size (See Appendix 3), but this same assay may not be suitable for a clinical trial involving a different study population that has much greater physiological variability. In this example, bias would be ~10% and imprecision would be 18% (Bias + 1.65 x imprecision). The incorrect assumption that bias is zero mistakenly allows for apparently increased imprecision. However, if bias goes up, the required

imprecision goes down; i.e., if bias = 20%, then imprecision must equal 12% to meet this TAE criterion.

To be considered acceptably validated: (1) appropriate assay characterization practices must be applied (the Seven Key Analytical Parameters defined in Table 5, plus relative accuracy), and (2) the assay must be able to distinguish biomarker changes that are outside of the normal biological variability.

Accuracy (Relative)

Accuracy is the closeness of agreement between the result of a measurement and the true value of the measure. In practice, an accepted reference value where available is substituted for the true value. Accuracy can be expressed as %bias and is also called Trueness or Bias (Information Technology Laboratory 2013). Ideally this requires a "gold" standard material or reference method procedure which is frequently not available for biomarkers. In the absence of these metrological anchors, a comparison to an appropriately validated method or an established reference laboratory's results may substitute. Accuracy is influenced by the number of measurements (i.e., fewer measurements are usually less precise than more). Relative accuracy is commonly measured by comparing the measured value of a specimen to that of a known value of an imperfect reference material (e.g., recombinant, non-glycosylated, etc.) in replicate samples, preferably in the expected range of concentrations.

%Accuracy = ((Actual value – Measured value) / Actual value) x 100%

Analytical Measurement Range (AMR)

The Analytical Measurement Range (AMR) is the range of analyte values that a method can directly measure on the specimen. AMR validation is the process of confirming that the assay system will correctly recover the concentration or activity of the analyte over the AMR. As an example, for assays that can measure a specimen without dilution (for instance, externally calibrated MS assays with isotope dilution), the AMR is determined using the maximum validated dilution and calculated as:

Analytical Measurement Range = LLOQ up to the (ULOQ * maximum validated dilution)

For assays which require specimen dilution prior to measurement (Immunoassays using specimen dilution which differs from calibrator dilution), the AMR is calculated using the minimum required dilution and the maximum validated dilution as:

Analytical Measurement Range = (LLOQ * Minimum required dilution) up to the (ULOQ * maximum validated dilution)

Sensitivity has also been formally defined as the slope of a linear calibration curve in an analyzer, but is often practically defined by the assay LLOQ during parallelism studies. However, the AMR is still bounded by the LLOQ, the lowest concentration of analyte that has been demonstrated to be

measurable with acceptable levels of total error, and, for most immunometric assays, the calibration curve is non-linear and the second definition for sensitivity above is inappropriate. Total error may be initially evaluated (as described using ANOVA (See Appendix 3) for replicate specimens from control or diseased/pre-treatment specimens) or determined from pooled authentic matrices (repetition for precision) or spiking test samples in surrogate matrix (ULOQ and LLOQ back-fit accuracy and precision specimens).

For commercial diagnostic kits the analytical sensitivity is usually defined by the limit of detection (LOD), which is determined via extrapolation of concentrations from a response signal of + 3SD of the mean background signal determined using blank matrix samples (n > 10, usually assay diluent). It must be noted that the variability at the LOD to LLOQ range is much higher than that in the working range. Therefore, data below the LLOQ should be applied with caution (CLSI EP17-A2).

A minimum of five samples with known concentrations spaced evenly across the range (previously assigned via higher order methods or less desirably prepared through spiking into authentic or surrogate matrix), including samples with concentrations that exceed the limits by 10-20% are used to validate the AMR. Samples should be measured in duplicate for assays without internal standardization (in singlicate for assays with internal standardization). Regression analysis by an appropriate linear or non-linear method should be performed comparing the measured to the expected analyte across the quantification range. The results should be plotted, a best fitting line determined, and the y-intercept should be close to zero. For ligand binding assays, the acceptance criteria for the correlation coefficient (r) should be predetermined based on the COU. The general considerations for the following r value ranges are: 0-0.19 very weak, 0.2-0.39 weak, 0.40-0.59 moderate, 0.6-0.79 strong, and 0.8-1 very strong correlation. Correlation coefficient criteria are not commonly applied to MS methods. Calibration systems with non-linear response functions (such as sigmoidal curves for immunoassays) may require consideration be given to the range of concentrations used in this regression analysis due to increased imprecision at asymptotic regions of measurement.

To further understand an assay's tolerance in the event of additional bias, the concept of Performance Standard (PS) 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 sources of error must be taken into account. This approach is outlined below by defining a minimal PS for the biomarker.

PS is defined by the amount of aTAE for the biomarker at the Decision Level (X_C).

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

aTAE is the amount of error that can be tolerated without invalidating the clinical utility of the result.

Decision Level is any concentration of the analyte that is critical for clinical utility (i.e., diagnosis and monitoring).

For biomarkers, acceptable imprecision 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. The bias needs to be calculated as the sum of squares from both the individual and group variances. Thus, the biomarker's minimal PS can be used as a guide to set criteria for the acceptability of the TAE associated with the assay.

TAE 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.

$$TAE = B + P_I$$

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-laboratory 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 TAE to the specification for the final Performance Standard. This is generally not possible for exploratory or partially validated methods.

Performance is acceptable when observed TAE is less than the PS (TAE < PS).

Performance is not acceptable when observed TAE is greater than the PS (TAE > 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 aTAE, while higher aTAE would be acceptable for assays with biomarkers that have low biological variability and higher amplitude of response to stimulus.

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

Parallelism

Parallelism is the extent to which the dose-response relationship between two materials (i.e., calibrator versus unknown specimens) is constant for the examined range of concentrations. It is now regarded as required for validation of LBAs, not optional.

Owing to the presence of endogenous analyte in control matrix samples, a vast majority of quantitative biomarker assays performed by LBA or LC-MS involve the use of a surrogate matrix,

which ranges in composition from biological matrix depleted of the target analyte to synthetically prepared mixtures designed to mimic the chemical composition of the biological control matrix. An ideal surrogate matrix behaves identically to the study sample matrix without presenting target analyte-specific interference. During method development and validation, it is essential that parallelism be established between the surrogate matrix and authentic biological matrix. Parallelism is the assurance that observed changes in response per given change in analyte concentration are equivalent for the surrogate and authentic biological matrix across the range of the assay.

Parallelism has historically been associated with LBAs and the term is often incorrectly used synonymously with 'dilutional linearity,'. Both, however, do use a dilutional approach. Dilutional linearity is performed with spiked (with reference or calibrator) control samples to demonstrate that the measured concentration versus the expected concentration of the diluted samples yields a linear response with slope = 1. Parallelism is performed with samples containing endogenous analyte to demonstrate whether the sample dilution-response curve is parallel to the standard concentration-response curve. For LBA, parallelism is largely a function of preserving binding conditions between the antibody reagents and the analyte. Because this binding is influenced by competing substances in the matrix and may be disproportionate at various analyte concentrations, parallelism is often obtained only after diluting the sample in the surrogate matrix several-fold to limit such interactions, which defines the MRD.

LC-MS biomarker assays also employ surrogate matrices; however, a fundamental difference between the methods is that LC-MS methods use extraction, chromatographic separation, as well as internal standards to compensate for sample to sample variation in matrix composition including differences between the control and surrogate matrix. Stable isotope labelled (SIL) internal standards are able to compensate for differences in analyte extraction recovery, as well as ion suppression /enhancement (which refers to the competition for ionization which occurs for co-eluting substances when introduced to the mass spectrometer interface). Because of the importance of assessing such 'matrix effects' in the development of LC-MS assays, this subject is discussed in the subsequent section on Selectivity.

Not surprisingly, experimental differences between LBA and LC-MS methods have also led to differences in experiments and criteria to assess parallelism (selectivity) with both sets of approaches being viewed as legitimate. In recognition of this situation, detailed practices and recommendations for parallelism assessment are treated separately and given in Appendix 4.

Precision

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, in the case of protein biomarkers which rarely have fully characterized reference standards, these parameters should be performed on patient samples with endogenous analyte whenever possible but are sometimes established from spiked control material (Table 10).

Selectivity

Selectivity is the ability of the assay to accurately measure the analyte unequivocally in the presence of interferences or structurally unrelated components that may be expected to be present in the intended matrix. Samples from multiple individuals of normal and target subject populations (such as ten from each population) should be tested for the endogenous value of the target biomarker in each individual sample. A rigorous assessment of selectivity can be undertaken as in Table 11b, using the common concentration method (Stevenson and Purushothama 2014) since parallelism across multiple individuals effectively demonstrates that the endogenous analyte is being selectively measured in the context of complex matrix components (Valentin et al. 2011). Although it may not be needed or only limited assessment of the effect of interferents is performed during the exploratory biomarker analysis, as the program matures and moves toward full bioanalytical validation for biomarker qualification, the effect on sample analysis of appropriate potential matrix and drug interferents, should be evaluated if appropriate samples are reasonably available (as per CLSI EP07-A2). Recovery of the analyte reference standard spiked into each at high and low levels is used as an approximation of selectivity if no other option is available and is calculated by subtraction of the basal value. The assay TAE may be used as acceptance criteria of spike recovery. A pre-specified sufficient proportion of the test samples should be found to be acceptable based on the TAE.

Specificity

Specificity is the ability of a measurement procedure to determine only the component (measurand) it purports to measure or the extent to which the assay responds only to all subsets of a specified measurand and not to other substances present in the sample.

For small molecule biomarkers, the exact structure of the target analyte is known, as well as its metabolites and structurally similar moieties in the intended matrices. If these compounds are available, various amounts can be spiked into pooled matrix samples to test for interference. On the other hand, protein biomarkers may have multiple endogenous forms, with unknown isoforms and/or catabolites. Therefore, specificity evaluation may not be feasible for the large molecule biomarkers using ligand binding methods. For LC-MS/MS-based methods, multiple ion ratios may be used to check selectivity from both known knowns and unknown unknowns (CLSI C62-A); however, care should be taken to ensure that changes in the ion used for quantitation is reflective of the analyte.

The acceptance criteria for small molecule biomarkers should be generally similar to acceptance criteria for PK analysis because they follow similar experimental designs. However, they are not identical and must depend strictly on the COU. For example, pro-peptides and catabolites are tested by spiking them into validation samples with consideration given to their endogenous levels in the intended matrices (such as normal volunteers, untreated patients, and treated patients if the drug dosing is expected to modulate their levels (CLSI EP07-A2 and O'Hara 2012).

Matrix Effects in LC-MS/MS

Components in the sample matrix may suppress or enhance the ion current response of the analyte and/or the internal standard when applying LC-MS/MS assays and especially during the use of electrospray ionization. These effects are not uncommon and may be disproportionate from sample to sample resulting in increased assay variability and a negative impact on overall assay sensitivity, accuracy and precision. Matrix effects impacting ionization efficiency should be investigated for all LC-MS assays regardless of the ionization technique utilized.

Determination of the Matrix Factor (MF) can be one useful technique when applicable. MF is the determination of the absolute and internal standard normalized peak responses by post extraction spiking of analyte (low and high concentrations) and internal standard into a minimum of 6 lots of blank matrix (if it is not an endogenous analyte) that have been processed through the full sample preparation defined in the method (Jenkins 2015). It is recommended that the internal standard normalized matrix effects should not exceed 20% CV, but ultimately this should be determined by the COU for the biomarker assay. The MF approach may be limited by the biomarker chemotype (small vs large molecule) and the ability to identify true blank matrix samples for these experiments (that have not been mechanically depleted of the relevant biomarker), and thus alternative approaches may be necessary and are reviewed elsewhere (Yang 2016).

Ionization effects should also include additional evaluations with appropriate matrix samples from disease subjects where potential interferences may be anticipated. Additional evaluations should include assessment of matrix effects caused by hemolyzed samples (described above) (CLSI EP07-A2) and evaluation of the effects of any co-administered agents which co-extract and chromatographically co-elute with the analyte or internal standard and, therefore, have the potential to differentially impact MS ionization efficiency. A variety of approaches to evaluate matrix effects have been applied to qualify the impact of endogenous materials on assay quality including evaluating recovery and precision of out-of-range spiked samples or preparation of QC's with a variety of matrix lots (representing disease subjects, etc.) (Panuwet 2016, CLSI EP07-A2).

When immunoaffinity capture is used for large molecule biomarker assay sample clean up, the impact from interference or structurally unrelated components in the sample matrix on the binding capability of the capture reagent and analyte and their impact on the ionization of the analyte and internal standard need to be evaluated. The evaluation can follow the recommendation outlined in the Selectivity section.

Selection of appropriate sample extraction techniques, elimination of non-specific or specific binding (including from anti-drug antibody) and the use of stable-labelled internal standards frequently for small molecules but also for discrete fragments of large molecules is critical to help manage assay matrix variability. In the absence of appropriate internal standards (such as when using analog small molecules or non-surrogate peptide internal standards [Song 2016]) the impact of matrix effects can be significantly different from sample to sample resulting in erroneous results. This situation can be further exacerbated if the internal standard does not chromatographically co-elute with the analyte.

Stability (Sample)

Stability under all conditions can be influenced by time, temperature, humidity, presence of degrading enzymes, the natural half-life of the biomarker, storage conditions, the matrix, exposure to light and the container system. Stability samples should be prepared using native matrix and endogenous analyte(s) whenever feasible, as recombinant protein may give false results. Appropriate surrogate systems may be considered based on scientific justification. Stability samples should be as close as possible in composition to clinical samples at the time of collection and should be prepared from individuals that are relevant to a study population (e.g., same disease, age). Once collected, samples should be immediately frozen and stored under the study sample storage conditions (typically ≤ -70°C). It is recommended that stability samples span the calibrated range of the assay or the anticipated clinical range of the analytes. Multiple pools from discrete individuals may be required when native analyte concentration ranges are narrow or inter-individual sample stability differences are suspected. Stability is then commonly measured by comparing the stored subject samples under realistic conditions to a set of freshly prepared samples (time zero/baseline results) from a stock solution of standard at known concentration in an interference free matrix or samples drawn freshly (time zero/baseline) and sub-aliquoted for stress testing (time, temperature, and storage condition). Sample stability is thus determined by measurement of observed bias to baseline specimens. It is also helpful to monitor the trends of stability evaluations over time and apply control chart methods to identify any out of control behaviors that could potentially be related to sample stability. Calibrators and QC must also meet the method-specific performance criteria as specified by aTAE. As time zero (t_0) samples are frequently difficult to achieve for biomarkers, one may instead consider the trend of degradation measured at a series of times, e.g. t_1, t_2, t_3.

Processed Sample Stability

In chromatography-mass spectrometry methods, samples are processed in batches/analytical runs. Usually samples are analyzed shortly after the preparation without a significant delay in the start of the analysis. In such cases, the acceptability of the analytical run, as displayed by the acceptability of the calibrators and QC, is indicative of samples being stable in the injection solution form for the analysis period. However, there may be circumstances in which there is a delay in injection after preparation or in which samples have to be reinjected due to an instrument failure. In such cases, the stability of processed samples at the appropriate storage conditions needs to be established. This is usually performed by reinjecting QC samples after storage at the appropriate storage conditions against a freshly prepared calibration curve (CLSI C62-A). Jenkins et.al. (2015) have cautioned that for large molecule bioanalysis by mass spectrometry, this approach may pose a challenge due to day to day variability in recovery resulting from procedures that are not well controlled such as digestion or immune capture steps. In such cases, the stored quality control samples or individual samples may be reinjected, and their concentrations calculated against the original calibration curve they were initially analyzed against. It should be noted that a loss of signal does not always indicate a lack of stability as it may be due to non-specific binding to the injection vial.

Case Study: Analytical Validation Approach for Kidney Safety Biomarkers

This case study describes the analytical validation approach for the proposed qualification of kidney safety biomarkers for use in clinical drug development. 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, and it is important to remember that many of the points to consider in this paper were reached and agreed upon substantially after the submission of the briefing package (5-6 years), resulting in the fact that some of the data (spike recovery, LLOQ, ULOQ) were established by methods that are not the preferred methods now described in this paper. 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 proposed COU for the clinical kidney safety project was 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.*

Note that FDA has now qualified the biomarker panel as interpreted via a composite measure of the following six urinary biomarkers, CLU, CysC, KIM-1, NAG, NGAL and OPN, as a composite safety biomarker panel to be used in conjunction with traditional measures to aid in the detection of kidney tubular injury in phase 1 trials in healthy volunteers when there is an *a priori* concern that a drug may cause renal tubular injury in humans. See FDA qualification letter dated August 15, 2018.

Assay parameters and critical success factors for all of the bioassay kits were defined. In Table 9, Table 10 and Table 11, 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 2018). The standard curve was then fit to the simplest regression model providing an appropriate or best statistical fit (FDA 2018). A minimum of six runs

were conducted over at least two days (FDA 2018). 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 2018) as a starting point, using fit-for-purpose for final criteria. ≥ 75% of non-zero standards were required to meet the criteria, including LLOQ (FDA 2018). The aTAE (accuracy and precision) was chosen to be ≤ 30%.

QC samples were prepared by collecting normal donor urines (six total), prepared by a standard protocol. After collection, the samples were 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 (≥ four out of six) were within 20% of respective nominal (measured) values 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 a starting point, using 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 in the range of expected study sample concentrations (low, mid, and high). 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 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 phosphate buffered saline (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 a starting point, using 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 highest concentration calibrator that did not exceed a 20% CV [recovery ± 20%] (FDA 2013) as a starting point, using fit-for-purpose for final criteria.

Parallelism was determined using a minimum of two urine samples (native where possible) diluted with the appropriate assay diluent to create 7 to 11 evenly distributed samples covering the assay range (CLSI EP06-A). Samples were measured in duplicate. Acceptable recovery was required to 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 of 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 (erythrocytes, 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 each of the assays distinguished their respective biomarker changes 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 TE represents the maximal TE acceptable for any assay used to measure the biomarkers and there is little tolerance for the addition of more variability into the method.

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.

Table 9: Pre-Analytical Factors Considered during the Validation of Neutrophil Gelatinase-Associated Lipocalin (NGAL) (specific to the BioPorto assay)

Pre-Analytical Factor	Process
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 the 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.
Sample processing	
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 and stability	Freeze at -70 to -80°C. Avoid temporary storage at -20°C.

Table 10: Analytical Parameters Evaluated during the Validation of Neutrophil Gelatinase-Associated Lipocalin (NGAL)

Accuracy (Relative)	**Selectivity**
Bias	**Specificity**
Drift	**Stability**
Spike Recovery	Bench top
Analytical Measurement Range	Short term
Lower Limit of Quantitation	Long term
Upper Limit of Quantitation	Freeze-thaw
Parallelism	**Standard/calibration curve range and model**
Reproducibility	
Quality Control	
Linearity	
Dilutional verification	
Interference	
Within sample	
Within run	
Between lot	

Table 11: 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

Case Study: Analytical Validation Approach for Glutamate Dehydrogenase (GLDH) as a Liver Specific Biomarker of Hepatocellular Injury

As part of the Critical Path Institute (C-Path) Predictive Safety Testing Consortium's (PSTC) ongoing efforts to augment translational biomarker tools for drug induced liver injury (DILI), the Hepatotoxicity Working Group (HWG) is proposing to qualify GLDH activity as a marker of liver injury in human subjects with ALT elevations from suspected extrahepatic sources such as muscle, i.e., as a biomarker to confer specificity to the liver. GLDH activity is proposed to be utilized as a complement to the existing guidance and standard methods for assessing DILI.

A joint FDA/EMA biomarker qualification consultation meeting for GLDH as a liver specific biomarker of hepatocyte injury in humans was held in March 2017. At this meeting, the FDA and EMA supported the novel qualification approach for using organ injury induced by diseases with a wide range of etiologies as approximation of chemical-induced organ injury for the evaluation of performance of novel biomarkers. On November 27, 2017, the EMA issued a Letter of Support (LOS) for GLDH demonstrating further support of the qualification effort and providing feedback regarding additional data needed to potentially enable formal qualification of GLDH as a "Drug Development Tool" in the future. Based on scientific feedback from both agencies, the validation and statistical plans have been revised and confirmatory studies have been designed and included in the qualification plan submitted to the agencies.

The context of use (COU) for GLDH as a liver specific biomarker of hepatocellular injury is as follows: Serum glutamate dehydrogenase (GLDH) is a safety biomarker capable of detecting hepatocellular injury that can be used as a safety biomarker to evaluate drug-induced liver injury (DILI) in conjunction with standard hepatic injury monitoring in Phase I through Phase III clinical trials for subjects and patients with elevated serum transaminases due to muscle degeneration or hemolysis.

For this validation, GLDH was measured in serum on Siemens ADVIA Automated Chemistry Systems with a commercially available kit (Randox Labs Ltd, Roche). The Randox GLDH assay utilizes the conversion of α-oxoglutarate to glutamate for detection of GLDH enzymatic activity. In this reaction, the kinetics of NADH oxidization are proportional to the GLDH activity and is measured spectrophotometrically as a decrease in absorbance per minute at 340 nm. The Randox GLDH assay kit was manufactured in the United Kingdom with ISO13485 certification as evidence of Good Manufacturing Practice and is an approved IVD assay in Europe, Canada, and China.

The Randox GLDH assay was validated according to Centers of Medicare and Medicaid Services' Clinical Laboratory Improvement Amendments (CLIA) guidelines for Laboratory Developed Tests (LDT). With the exception of the method to method comparison, the analysis was performed at a single site. During assay validation the following parameters were tested: precision, relative accuracy, method to method comparison, analytical sensitivity (Limit of blank (LOB)), analytical measurement range, freeze/thaw stability, short- and long-term stability, analytical specificity to include interfering substances and reference interval. Appropriate quality control samples were

applied during the validation procedure and were used throughout the subsequent sample analysis to ensure data reliability and data comparability over the different clinical studies included in the qualification package. Commercially available products from Randox were utilized as quality control samples including Acusera Human Assay Control 2 and 3 and included lot specific acceptance criteria equivalent to the assigned mean ± 2 standard deviations.

Blood samples were collected from healthy subjects and subjects with clinically demonstrable liver injuries for validation samples. Blood was centrifuged at 3000 x g for 10 minutes at room temperature. Serum samples were recovered from serum-separator tubes and kept at 4°C for up to 72 hours before aliquots were frozen and stored at -80°C until analysis. A stability assessment of GLDH confirmed acceptable stability at 4°C for up to 96 hours. Randox Acusera Calibration 2 and 3 were also utilized as validation samples.

A summary of pre-analytical factors relevant to the validation of the GLDH assay are listed in Table 12. A list of the GLDH assay validation parameters and a summary of the precision requisites for the validation are shown in Table 13 and Table 14 respectively.

Precision testing was performed using human serum samples at four concentrations (Near Detection Level, Low, Mid, and High) and three levels of quality control material run in duplicate 2 times per day over 20 days. Precision testing samples were selected to span the range of expected study sample concentrations tested including samples bridging the medically relevant cutoffs established during the qualification. The precision determined for all samples could not exceed ± 10% CV with the exception of the samples Near Detection Level that could not exceed ± 15% CV.

Spike Recovery (Relative Accuracy) was performed using a pooled serum sample with a low analyte concentration, < 3 U/L GLDH. Samples were spiked with 4 concentrations of kit calibrator and 2 concentrations of a high patient sample in order to span the range of expected study sample concentrations tested including samples bridging the medically relevant cutoffs established during the qualification. Acceptable performance was based on percent recovery of each spiked sample being within ± 20% of the expected calculated concentration.

In the absence of an FDA cleared assay, a method comparison study was performed to evaluate “accuracy” as a measure of the closeness of agreement between a split-sample experiment performed at 2 separate CLIA certified laboratories using the same assay. Forty human serum samples spanning the range of expected study sample concentrations were split and analyzed in duplicate over 5 operating days at 2 sites utilizing the Randox GLDH method. A linear regression analysis was performed and correlation coefficient (R), slope and %bias calculated. Acceptable performance was based on obtaining an R value of ≥ 0.90.

Limit of Blank (LOB) was performed using a blank, deionized water. Twenty replicates of the blank were analyzed in a single run to verify the LOB. The mean and SD of the blank was calculated, and the LOB established according to the following:

$$LOB = \text{mean blank} + 1.645(SD_{blank}) = 0.1 + 1.645\ (0.31) = 0.610 = 1$$

Limit of Detection (LOD) was performed using the blank, deionized water, and a low concentration sample. The low concentration sample was created using a 10% solution of the lowest available calibrator (Calibrator 1 = 28 U/L). The mean and SD of the blank was calculated and the LOD established according to the following formula:

$$LOD = LOB + 1.645(SD_{low\ concentration\ sample}) = 0.1 + 1.645\ (0.92) = 1.61 = 2$$

Both formulas were adapted from Armbruster and Pry (2008), Burd (2010), and CLSI EP17-A2.

The analytical measurement range was established by measuring several samples with GLDH concentrations across the anticipated measuring range. Dilutions of a human sample with elevated GLDH levels were made in the appropriate assay diluent to create 8-10 evenly distributed samples. Samples were run in triplicate. Assay linearity was assessed based on percent recovery at each dilution (80% to 120%) and visual assessment of the linearity using the slope and correlation coefficient as guides. The analytical range was established based on the lowest and highest value that recovered within ± 20% of the expected concentration.

Sample freeze/thaw stability was performed with 3 human serum samples with varied GLDH concentrations designed to mimic intended sample handling conditions used during sample analysis. The samples were assayed after 1, 2, 3, and 4 freeze/thaw cycles from -80°C. Freeze-thaw stability was evaluated as recovered analyte concentration relative to the sample undergoing one (i.e. initial) freeze/thaw. Acceptable freeze/thaw stability was based on percent recovery of each sample being within ± 20% of the initial freeze/thaw value.

Sample stability was performed with 3 human serum samples with low, mid, or high GLDH concentrations. The samples were assayed at baseline, 4, 24, 48, 72 and 96 hours for both room and refrigerated temperatures. Three sample sets with low, mid, or high analyte concentrations were assayed at baseline and after storage of approximately: 1 week, 2 weeks, 1 month, 3 months, 6, 12 and 18 months at-80°C. The percent recovery for each storage timepoint was calculated relative to the baseline value. Acceptable stability was based on percent recovery of each sample being within ± 20% of the baseline value.

Analytical specificity to variable matrix-related interferences including hemolysis, lipemia, and icterus was evaluated. This was achieved by spiking a high GLDH pooled human sample with interferent. A pair of test interference samples was prepared at 6 different interferent concentrations. The pooled serum sample spiked with interferent at 1 of 6 different concentrations was run in parallel with the same high GLDH pooled serum sample spiked with a serum sample with a GLDH value < 3 U/L at the same volume as the interferent. Both samples were analyzed and compared to each other. Results were deemed acceptable if ≥ 80% of the samples tested resulted in %CV of ≤ 30% and were within ± 30% of the respective nominal concentrations. All samples met the acceptance criteria with the exception of lipemia which is in agreement with the manufacturer's reagent package insert. The Siemens Advia analyzer has flagging capabilities when interfering substances are detected. Any sample positive for high triglycerides is rejected by the instrument and the result not reported.

Samples from three populations were obtained to establish a reference range for GLDH. Data for GLDH was generated on the Siemens Advia Chemistry Analyzer using samples from 552 human samples and a reference interval was generated from 274 males and 278 females. Data also enabled the examination of the impact of age, gender, ethnicity, and intra- and inter-individual variability. A reference range that includes 97.5% of the population was established as < 3-10 U/L GLDH.

Table 12: Pre-Analytical Factors Relevant for the Validation of the Glutamate Dehydrogenase (GLDH) Assay

Sample Type	Serum
Interference	Hemoglobin (hemolysis), lipids (lipemia), and bilirubin (icterus)
Collection Tube	Serum separator
Sample processing	Maintain sample at room temperature for at least 20 minutes; centrifuge at 3000xg for 10 minutes; process and freeze within 4 hours of collection.
Identification of abnormal samples	Advia instrument identifies and flags abnormal samples.
Storage considerations	Freeze at -80°C. Avoid temporary storage at -20°C.
Thawing considerations	Thaw samples on the bench.
Logistics of transport	Transport frozen on dry ice.

Table 13: Analytical Parameters Evaluated during the Validation of the Glutamate Dehydrogenase (GLDH) Assay

Precision	**Stability**
Within run	Bench Top
Between run	Short term
Day to day	Long term
Accuracy (Relative)	Freeze-thaw
Spike Recovery	**Specificity (Interference)**
Method to Method Comparison	**Reference Interval**
Analytical Measurement Range/Linearity	

Table 14: Summary of the Precision Requisites for the Validation of the Glutamate Dehydrogenase (GLDH) Assay as a Laboratory Developed Test

QC Samples	**3**
Precision Samples (Human)	**3**
Replicates	**2**
Sites	**2**
Operators	**2+**
Reagent Lots	**2+**
Precision Runs	**40**
Days	**20**

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. Though unaltered endogenous biological samples should always be used as part of a quality control system, they are frequently difficult to obtain in appropriate quantity. Currently, certified reference materials are scarce to nonexistent, depending on the biomarker, though as individual assays develop, better standards may evolve as well. Therefore, multiple analytical factors must be considered when designing the assays, given that the results impact drug development and patient management decisions. To ensure reliable 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 and analytical and clinical performance. 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, parallelism, ruggedness, and robustness); and (2) the assay can accurately distinguish biomarker changes that are outside of the range of normal analytical 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.

Armbruster DA, Pry T. Limit of blank, limit of detection and limit of quantitation. (2008) Clin Biochem Rev 29 Suppl 1:S49-52.

Arnold, M.E., Booth, B., King, L. et al. (2016) AAPS J 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.

Betsou, Fotini, Sylvain Lehmann, Garry Ashton, Michael Barnes, Erica E. Benson, Domenico Coppola, Yvonne DeSouza, et al. (2010) "Standard Preanalytical Coding for Biospecimens: Defining the Sample PREanalytical Code." *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology* 19 (4): 1004–11. https://doi.org/10.1158/1055-9965.EPI-09-1268.

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.

Burd EM. (2010) Validation of laboratory-developed molecular assays for infectious diseases. Clin Microbiol Rev 23:550–76. doi:10.1128/CMR.00074-09.

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 EP07-A2: 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.

CLSI C62-A: Liquid Chromatography-Mass Spectrometry Methods, 1st Edition. ISBN Number: 1-56238-977-7. CLSI 940 West Valley Road, Suite 140, Wayne, PA 19087-1898 USA, 2014.

Cox JM, Butler JP, Lutzke BS, Jones BA, Buckholz JE, Biondolillo R, et al. (2015) A validated LC-MS/MS method for neurotransmitter metabolite analysis in human cerebrospinal fluid using benzoyl chloride derivatization. Bioanalysis 7:2461–75. doi:10.4155/bio.15.170

DeSilva B, Smith W, Weiner R, Kelley M, Smolec J, Lee B, Khan M, Tacey R, Hill H, Celniker A. (2003) Recommendations for the bioanalytical method validation of ligand-binding assays to support pharmacokinetic assessments of macromolecules. Pharm Res. Nov;20(11):1885–900.

European Medicines Agency. (2011) Guideline on bioanalytical method validation. Available at: https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-bioanalytical-method-validation_en.pdf Accessed on April 8, 2019.

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) Draft 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. (2014a) Qualification Process for Drug Development Tools. Available at: https://www.fda.gov/downloads/drugs/guidances/ucm230597.pdf. Accessed on February 25, 2019.

Food and Drug Administration. (2014b) Biomarker Qualification Context of Use. Available at: https://www.fda.gov/Drugs/DevelopmentApprovalProcess/DrugDevelopmentToolsQualificationPro

gram/BiomarkerQualificationProgram/ucm535072.htm Last Updated 9/15/14. Accessed on August 23, 2016.

Food and Drug Administration. (2014c) Draft Guidance for Industry, Food and Drug Administration Staff, and Clinical Laboratories: Framework for Regulatory Oversight of Laboratory Developed Tests (LDTs) Available at: https://www.fda.gov/downloads/medicaldevices/deviceregulationandguidance/guidancedocuments/ucm416685.pdf Accessed on October 10, 2018.

Food and Drug Administration. (2016a) Principles for Codevelopment of an In Vitro Companion Diagnostic Device with a Therapeutic Product: Draft Guidance for Industry and Food and Drug Administration Staff. Available at: https://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/UCM510824.pdf Accessed on October 10, 2018.

Food and Drug Administration. (2016b) Biomarker Qualification Program. Available at: https://www.fda.gov/Drugs/DevelopmentApprovalProcess/DrugDevelopmentToolsQualificationProgram/BiomarkerQualificationProgram/default.htm Last Updated on April 2, 2018. Accessed on October 10, 2018.

Food and Drug Administration. (2018) Bioanalytical Method Validation: Guidance for Industry. Available at: https://www.fda.gov/downloads/drugs/guidances/ucm070107.Pdf Accessed on October 16, 2018.

Fraser CG, P Hyltoft Petersen, JC Libeer and C Ricos. (1997) Proposals for setting generally applicable quality goals solely based on biology Ann Clin Biochem. 34: 8-12.

Fraser CG. (2001) Biological Variation: From Principles to Practice, C. G. Fraser, Publisher: Amer. Assoc. for Clinical Chemistry.

Global Bioanalysis Consortium. Team L2 Large Molecule Specific Assay Operation. Available at: http://www.globalbioanalysisconsortium.org/site/gbc/assets/documents/HT%20slide%20decks/Team%20L2_final.pdf Accessed on March 29, 2019.

Hougton R, Gouty D, Allinson J et al. (2012) Recommendations on biomarker bioanalytical method validation by GCC. Bioanalysis. (20):2439-46. doi: 10.4155/bio.12.197.

ICH Harmonised Tripartite Guideline. Validation of Analytical Procedures: Text and Methodology Q2(R1). Current Step 4 version. Parent Guideline dated 27 October 1994. (Complementary Guideline on Methodology dated 6 November 1996 incorporated in November 2005). Available at: https://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Quality/Q2_R1/Step4/Q2_R1__Guideline.pdf Accessed on February 25, 2019.

Ichihara K, Boyd JC, IFCC Committee on Reference Intervals and Decision Limits (C-RIDL). (2010) An appraisal of statistical procedures used in derivation of reference intervals. Clin Chem Lab Med. Nov;48(11):1537–51.

Information Technology Laboratory. Engineering Statistics Handbook. 2.1.1.3 Bias and Accuracy. Available at: https://www.itl.nist.gov/div898/handbook/mpc/section1/mpc113.htm Accessed on March 29, 2019.

Jani D, Allinson J, Berisha F, et al. (2016) Recommendations for Use and Fit-for-Purpose Validation of Biomarker Multiplex Ligand Binding Assays in Drug Development. AAPS J. (1):1-14. doi: 10.1208/s12248-015-9820-y. Epub 2015 Sep 16.

Jemal M, Schuster A, Whigan DB. (2003) Liquid chromatography/tandem mass spectrometry methods for quantitation of mevalonic acid in human plasma and urine: method validation, demonstration of using a surrogate analyte, and demonstration of unacceptable matrix effect in spite of use of a stable isotope analog internal standard. Rapid Commun Mass Spectrom 17:1723–34. doi:10.1002/rcm.1112.

Jenkins R, Duggan JX, Aubry A-F, Zeng J, Lee JW, Cojocaru L, et al. (2015) Recommendations for validation of LC-MS/MS bioanalytical methods for protein biotherapeutics. AAPS J 17:1–16. doi:10.1208/s12248-014-9685-5.

Jenkins RG. (2016) Accuracy: a potential quandary in regulated bioanalysis of “endogenous” analytes. Bioanalysis 8:2393–7. doi:10.4155/bio-2016-0247.

Jian W, Edom R, Weng N, Zannikos P, Zhang Z, Wang H. (2010) Validation and application of an LC-MS/MS method for quantitation of three fatty acid ethanolamides as biomarkers for fatty acid hydrolase inhibition in human plasma. J Chromatogr B Analyt Technol Biomed Life Sci 878:1687–99. doi:10.1016/j.jchromb.2010.04.024

Jones BR, Schultz GA, Eckstein JA, Ackermann BL. (2012) Surrogate matrix and surrogate analyte approaches for definitive quantitation of endogenous biomolecules. Bioanalysis 4:2343–56. doi:10.4155/bio.12.200.

Klee GG. (2010) Establishment of Outcome-Related Analytic Performance Goals, Clinical Chemistry 56:5 714–722.

Krouwer JS. (2002) Setting Performance Goals and Evaluating Total Analytical Error for Diagnostic Assays. Clinical Chemistry 48:6, 919–927.

Lee JW, Devanarayan V, Barrett YC, Weiner R, Allinson J, Fountain S, Keller S, Weinryb I, Green M, Duan L, Rogers JA, Millham R, O'Brien PJ, Sailstad J, Khan M, Ray C, Wagner JA. (2006) Fit-for-purpose method development and validation for successful biomarker measurement. Pharm Res. 23:312-28.

Lee JW, Figeys D, Vasilescu J. (2007) Biomarker assay translation from discovery to clinical studies in cancer drug development: quantification of emerging protein biomarkers. Adv Cancer Res. 96:269-98.

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.

Lee JW, Hall M. (2009) Method validation of protein biomarkers in support of drug development or clinical diagnosis/prognosis. J Chromatogr B Analyt Technol Biomed Life Sci. 877:1259-71.

Li W, Cohen LH. (2003) Quantitation of endogenous analytes in biofluid without a true blank matrix. Anal Chem 75:5854–9. doi:10.1021/ac034505u.

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

Moore HM, Kelly AB, Jewell SD, McShane LM, Clark DP, Greenspan R, et al. (2015) Guidelines for the standardization of preanalytic variables for blood-based biomarker studies in Alzheimer's disease research. Alzheimers Dement. May;11(5):549–60.

O'Hara DM, Theobald V, Egan AC, Usansky J, Krishna M, TerWee J, Maia M, Spriggs FP, Kenney J, Safavi A, Keefe J. (2012) Ligand Binding Assays in the 21st Century Laboratory: Recommendations for Characterization and Supply of Critical Reagents. AAPS J. Mar 14;14(2):316–28.

Ongay S, Hendriks G, Hermans J, van den Berge M, ten Hacken NHT, van de Merbel NC, et al. (2014) Quantification of free and total desmosine and isodesmosine in human urine by liquid chromatography tandem mass spectrometry: a comparison of the surrogate-analyte and the surrogate-matrix approach for quantitation. J Chromatogr A 1326:13–9. doi:10.1016/j.chroma.2013.12.035.

Panuwet P, Hunter RE, D'Souza PE, Chen X, Radford SA, Cohen JR, et al. (2016) Biological Matrix Effects in Quantitative Tandem Mass Spectrometry-Based Analytical Methods: Advancing Biomonitoring. Crit Rev Anal Chem 46:93–105. doi:10.1080/10408347.2014.980775.

Song A, Lee A, Garofolo F, Kaur S, Duggan J, Evans C, et al. (2016) White Paper on recent issues in bioanalysis: focus on biomarker assay validation (BAV): (Part 2 - Hybrid LBA/LCMS and input from regulatory agencies). Bioanalysis 8:2457–74. doi:10.4155/bio-2016-4988.

Stevenson LF, Purushothama S. (2014) Parallelism: considerations for the development, validation and implementation of PK and biomarker ligand-binding assays. Bioanalysis. 6(2):185-98. doi: 10.4155/bio.13.292.

United States Congress. (2016) PUBLIC LAW 114–255—DEC. 13, 2016; TITLE III—DEVELOPMENT, Subtitle B—Advancing New Drug Therapies, Sec. 3011. Qualification of drug development tools., Amending Chapter V of the Federal Food, Drug, and Cosmetic Act (21 U.S.C. 351 et seq.) by addition of ''SEC. 507. QUALIFICATION OF DRUG DEVELOPMENT TOOLS." Available at: https://www.gpo.gov/fdsys/pkg/PLAW-114publ255/pdf/PLAW-114publ255.pdf Accessed on October 10, 2018.

United States Pharmacopoeia, The National Formulary, (USP-NF) United States Pharmacopeial Convention Inc. Available at: https://www.uspnf.com/ Accessed on February 25, 2019.

Valentin MA, Ma S, Zhao A, Legay F, Avrameas A. (2011) Validation of immunoassay for protein biomarkers: bioanalytical study plan implementation to support pre-clinical and clinical studies. J. Pharm. Biomed. Anal. 55(5), 869–877.

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

Viswanathan CT, Bansal S, Booth B, DeStefano AJ, Rose MJ, Sailstad J, Shah VP, Skelly JP, Swann PG, Weiner R. (2007b) Quantitative bioanalytical methods validation and implementation: best practices for chromatographic and ligand binding assays. Pharm Res. 24:1962-73.

Welink J, Yang E, Hughes N, Rago B, Woolf E, Sydor J, et al. (2017) White Paper on recent issues in bioanalysis: aren't BMV guidance/guidelines "Scientific"? (Part 1 - LCMS: small molecules, peptides and small molecule biomarkers). Bioanalysis 9:1807–25. doi:10.4155/bio-2017-4975

West J, Atherton J, Costelloe SJ, Pourmahram G, Stretton A, Cornes M. (2017) Preanalytical errors in medical laboratories: a review of the available methodologies of data collection and analysis. Ann Clin Biochem. Jan;54(1):14–9.

Westgard JO, Carey RN, Wold S. (1974) Criteria for judging precision and accuracy in method development and evaluation. Clin Chem. Jul;20(7):825–33.

Yang E, Welink J, Cape S, Woolf E, Sydor J, James C, et al. (2016) White Paper on recent issues in bioanalysis: focus on biomarker assay validation (BAV) (Part 1 – small molecules, peptides and small molecule biomarkers by LCMS). Bioanalysis 8:2363–78. doi:10.4155/bio-2016-4992.

Appendix 1. Assay Performance Characteristic Definitions

Accuracy (Relative)

Accuracy is the closeness of agreement between the result of a measurement and the true value of the measure. In practice, an accepted reference value is substituted for the true value. Accuracy can be expressed as %bias and is also called Trueness or Bias. Ideally this requires a "gold" standard or method. In the absence of a gold standard or method, a comparison to an 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). 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 - Measurement)) / Actual value

Note that evolving terminology reflects a shift from "accuracy" to "relative accuracy" in almost all cases for large molecule analytes, due to the nature of the reference standard (recombinant vs. endogenous analyte).

Analytical Measurement Range (AMR)

The analytical measurement range is the range of analyte values that a method can directly measure on the specimen without any dilution, concentration, or other pretreatment not part of the usual assay process.

Analytical Validation

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, tool, or instrument technical performance, but is not validation of the item's usefulness.

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.

Characterization of Reference Materials (and Stability)

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

Clinically Reportable Range (CRR)

Clinically reportable range is defined as the range of analyte values a method can measure, allowing for specimen dilution, concentration or other pretreatment used to extend the direct analytical

measurement range. Within the assay range, linearity, accuracy and precision are acceptable and shown to be valid. This can be influenced by affinity of the detection antibodies (if used), and the signal to noise ratio of an instrument, as well as the overall performance of the assay. See also Analytical Measurement Range.

Context of Use

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.

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.

Intended Use

The specific clinical circumstance or purpose for which a medical product or test is being developed. In the regulatory context, “intended use” refers to the objective intent of the persons legally responsible for the labeling of medical products.

Linearity

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 ligand binding assays reflects a sigmoidal response characteristic and not a strict linear analyte-signal response behavior but can still allow determination of analyte concentrations.

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 specified precision and accuracy. For chromatographic methods of small molecules, 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. However, for large molecule biomarkers, more relevant and precise experimental determinations of an assay LLOQ include repeated measurements of samples with low and very low analyte concentrations in several independent experiments, with the final determination of the LLOQ value by predefined criteria based on the precision and accuracy of the sample measurement.

Parallelism

Parallelism is the extent to which the dose-response relationship between two materials (i.e., calibrator versus unknown specimens) is constant for the examined range of concentrations. It is performed with samples containing endogenous analyte to demonstrate whether the sample dilution-response curve is parallel to the standard concentration-response curve. It is thus different from Dilutional Linearity, which is linearity performed with spiked control samples to demonstrate that the measured concentration vs the expected concentration of the diluted samples yields a linear response with slope = 1.

Parallelism is a condition in which dilution does not result in biased measurements (trending up or down) of the analyte concentration. It is related, but not identical, to linearity. Linearity can be influenced by matrix effects, protein binding, or metabolism of the biomarker. Parallelism likewise can be affected by matrix effects, protein or serum component binding, or metabolism of the biomarker. 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.

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 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. Proper design of lot bridging experiments that evaluate the effect of different lots of assay kits is critical to ensure consistent measurement of the analyte over the course of the study.

- Repeatability (within-series, within-run, or intra-assay) – precision determined under unchanged conditions, measured using the same method on identical test material in the same laboratory by the same operator using the same equipment within a short interval of time
- Intermediate (within-laboratory) – precision under a set of conditions that includes the same measurement procedures, same location, and replicate measurements over an extended period of time; that may also include other conditions involving changes such as new calibrators, equipment, operators, or reagent lots. Also known as Ruggedness.
- Reproducibility (inter-assay) – precision measured over time under changed conditions, measured using the same method on identical test material in different laboratories with different operators using different equipment

Quality Control/Reproducibility

Method precision and relative 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 often established from patient samples or spiked control material. See also Robustness and Ruggedness.

Reportable Range

Reportable range is the functional range of an assay over which the concentrations of an analyte can be measured with acceptable (specified) accuracy and precision. Reportable range should not be confused with reference range. Reportable range includes analytical measurement range (AMR) and clinically reportable range (CRR).

Robustness and Ruggedness

Method robustness is part of quality control and reproducibility. Robustness is defined as: “The robustness of an analytical procedure is a measure of its capacity to remain unaffected by small, but deliberate variations in method parameters and provides an indication of its reliability during normal usage" (ICH Guideline 1994). This 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. It is a measure of the assay capacity to remain unaffected by small but deliberate changes in test conditions. Ruggedness provides an indication of the ability of the assay to perform under normal usage and is defined as: "The ruggedness of an analytical method is the degree to of reproducibility of test results obtained by the analysis of the same samples under a variety of conditions such as different laboratories, different analysts, different instruments, different lots of reagents, different elapsed assay times, different assay temperatures, different days, etc. Ruggedness is normally expressed as the lack of influence of operational and environmental factors of the analytical method. Ruggedness is a measure of reproducibility of test results under the variation in conditions normally expected from laboratory to laboratory and analyst to analyst" (USP-NF 2019).

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 in PK 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. This is not applicable to most biomarkers, which have a measurable endogenous value in most samples. Interference is thus generally determined by running studies using patient samples spiked with the potential interferents.

Sensitivity (Analytical)

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

Specificity (Analytical)

Specificity is the ability to unequivocally assess 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 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 evaluating sample controls at various concentrations spanning the expected range, with and without the potential interfering substance.

Spike Recovery

For an analytical method that includes an extraction process (such as LC-MS methods), 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 measurement 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 determined 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. For an analytical method that does not include an extraction process (such as most LBA), the analyte reference standard is spiked into individual samples and the spike recovery is determined against the concentration of the unspiked sample. However, there is little to no utility for spike recovery in the context of LBA.

Stability

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. It must be demonstrated with endogenous, rather than spiked, samples.

Bench top

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

Freeze-thaw stability

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. 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.

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 the data 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 that can be reliably measured. This can be influenced by the affinity of the detection antibodies, and 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, including a blank (no analyte), covering the expected range of the assay, in the matrix that is going to be used. Curves for LBAs are rarely linear, and the most appropriate model may be exponential or four- or five-parameter logistic fit.

Appendix 2. Pre-analytical Resources

Table 2A. CLSI guidelines for Pre-analytical Variables

C9-A	Analysis of Body Fluids in Clinical Chemistry; Approved Guideline (Vol 27, No. 14) April 2007
GP16-A3	Urinalysis; Approved Guideline – Third Edition (Vol. 29, No. 4) February 2009
H04-A6	Procedures and Devices for the Collection of Diagnostic Capillary Blood Specimens; Approved Standard – Sixth Edition (Vol. 28, No.25 September 208
H18-A4	Procedures for Handling and Processing of Blood Specimens for Common Laboratory Tests; Approved Guideline – Fourth Edition
H21-A5	Collection, Transport, and Processing of Blood Specimens for Testing Plasma Based Coagulation Assays and Molecular Hemostasis Assays; Approved Guideline – Fifth Edition (Vol. 28, No. 5) January 2008
H3-A6	Procedures for the Collection of Diagnostic Blood Specimens by Venipuncture Approved Standard - Sixth Edition
H56-A	Body Fluid Analysis for Cellular Composition; Approved Guideline (Vol.26, No. 26) June 2006

Websites

National Cancer Institute Best Practices for Biospecimen Resources. http://biospecimens.cancer.gov/practices/

NCI Biospecimen Research Network. http://biospecimens.cancer.gov/researchnetwork/.

National Institute on Aging, Biospecimens best practice guidelines for the Alzheimer's Disease Centers V 30 (24 June 2014). https://www.alz.washington.edu

National Institute of Diabetes and Digestive and Kidney Diseases Best Practices for Sample Storage: Urine as a Paradigm. niddkweb.niddk.nih.gov/urine/Best_Practices_for_Sample_Storage

Standardization and improvement of generic pre-analytical tools and procedures for in-vitro diagnostics (SPIDIA). http://www.spidia.eu/.

Case Studies of Existing Human Tissue Repositories—"Best Practices" for a Biospecimen Resource for the Genomic and Proteomic Era http://www.rand.org/pubs/monographs/MG120/index.html

Further Peer-Reviewed Resources

Bernini P, Bertini I, Luchinat C, Nincheri P, Staderini S, Turano P. Standard operating procedures for pre-analytical handling of blood and urine for metabolomic studies and biobanks. J Biomol NMR. 2011 Apr;49(3-4):231-43.

Betsou F, Lehmann S, Ashton G, Barnes M, Benson EE, Coppola D, DeSouza Y, Eliason J, Glazer B, Guadagni F, Harding K, Horsfall DJ, Kleeberger C, Nanni U, Prasad A, Shea K, Skubitz A, Somiari S,

Gunter E, Science IS for B and ER (ISBER) WG on B. Standard Preanalytical Coding for Biospecimens: Defining the Sample PREanalytical Code. Cancer Epidemiol Biomarkers Prev. 2010 Apr 1;19(4):1004–11.

Coppens A, Speeckaert M, Delanghe J. The pre-analytical challenges of routine urinalysis. Acta Clin Belg. 2010 Jun;65(3):182–9.

Cornes MP, Church S, van Dongen-Lases E, Grankvist K, Guimarães JT, Ibarz M, Kovalevskaya S, Kristensen GB, Lippi G, Nybo M, Sprongl L, Sumarac Z, Simundic A-M, Working Group for Preanalytical Phase (WG-PRE) and European Federation of Clinical Chemistry and Laboratory Medicine (EFLM). The role of European Federation of Clinical Chemistry and Laboratory Medicine Working Group for Preanalytical Phase in standardization and harmonization of the preanalytical phase in Europe. Ann Clin Biochem. 2016 Sep;53(Pt 5):539–47.

Cornes MP. Exogenous sample contamination. Sources and interference. Clin Biochem. 2016 Dec;49(18):1340-1345.

del Campo M, Mollenhauer B, Bertolotto A, Engelborghs S, Hampel H, Simonsen AH, Kapaki E, Kruse N, Le Bastard N, Lehmann S, Molinuevo JL, Parnetti L, Perret-Liaudet A, Sáez-Valero J, Saka E, Urbani A, Vanmechelen E, Verbeek M, Visser PJ, Teunissen C. Recommendations to standardize preanalytical confounding factors in Alzheimer's and Parkinson's disease cerebrospinal fluid biomarkers: an update. Biomark Med. 2012 Aug;6(4):419–30.

Delanghe J, Speeckaert M. Preanalytical requirements of urinalysis. Biochem Med (Zagreb). 2014 Feb 15;24(1):89–104.

Delanghe JR, Speeckaert MM. Preanalytics in urinalysis. Clin Biochem. 2016 Dec;49(18):1346–50.

Engel KB, Vaught J, Moore HM. National Cancer Institute Biospecimen Evidence-Based Practices: A Novel Approach to Pre-analytical Standardization. Biopreserv Biobank. 2014 Apr 1;12(2):148–50.

Hassis ME, Niles RK, Braten MN, Albertolle ME, Witkowska HE, Hubel CA, Fisher SJ, Williams KE. Evaluating the effects of preanalytical variables on the stability of the human plasma proteome. Anal Biochem. 2015 Jun 1;478:14–22.

Lehmann S, Guadagni F, Moore H, Ashton G, Barnes M, Benson E, Clements J, Koppandi I, Coppola D, Demiroglu SY, DeSouza Y, De Wilde A, Duker J, Eliason J, Glazer B, Harding K, Jeon JP, Kessler J, Kokkat T, Nanni U, Shea K, Skubitz A, Somiari S, Tybring G, Gunter E, Betsou F, International Society for Biological and Environmental Repositories (ISBER) Working Group on Biospecimen Science. Standard preanalytical coding for biospecimens: review and implementation of the Sample PREanalytical Code (SPREC). Biopreserv Biobank. 2012 Aug;10(4):366–74.

Moore HM, Kelly AB, Jewell SD, McShane LM, Clark DP, Greenspan R, Hayes DF, Hainaut P, Kim P, Mansfield EA, Potapova O, Riegman P, Rubinstein Y, Seijo E, Somiari S, Watson P, Weier H-U, Zhu C,

O'Bryant SE, Gupta V, Henriksen K, Edwards M, Jeromin A, Lista S, Bazenet C, Soares H, Lovestone S, Hampel H, Montine T, Blennow K, Foroud T, Carrillo M, Graff-Radford N, Laske C, Breteler M, Shaw L, Trojanowski JQ, Schupf N, Rissman RA, Fagan AM, Oberoi P, Umek R, Weiner MW, Grammas P, Posner H, Martins R. Guidelines for the standardization of preanalytic variables for blood-based biomarker studies in Alzheimer's disease research. Alzheimers Dement. 2015 May;11(5):549–60.

Stankovic AK, DiLauri E. Quality improvements in the preanalytical phase: focus on urine specimen workflow. Clin Lab Med. 2008 Jun;28(2):339–350, viii.

Vanderstichele H, Bibl M, Engelborghs S, Le Bastard N, Lewczuk P, Molinuevo JL, Parnetti L, Perret-Liaudet A, Shaw LM, Teunissen C, Wouters D, Blennow K. Standardization of preanalytical aspects of cerebrospinal fluid biomarker testing for Alzheimer's disease diagnosis: a consensus paper from the Alzheimer's Biomarkers Standardization Initiative. Alzheimers Dement. 2012 Jan;8(1):65–73.

Vaught J. Biospecimen reporting for improved study quality (BRISQ). Cancer Cytopathol. 2011 Apr 25;119(2):92–101.

Appendix 3. Performance Specification and Total Analytical Error

Performance Specification

To further understand an assay's tolerance in the event of additional bias, the concept of Performance Standard (PS) 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 PS for the biomarker.

PS is defined by the amount of aTAE for the biomarker at the Decision Level (X_C).

PS = aTAE at X_C

aTAE 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 (i.e. diagnosis and monitoring).

For biomarkers, acceptable imprecision 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. The Bias needs to be calculated as the sum of squares from both the individual and group variances. Thus, the biomarker's minimal PS can be used as a guide to set criteria for the acceptability of the TAE associated with the assay.

TAE 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.

$$TAE = B + P_I$$

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-laboratory 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 TE to the specification for the final Performance Standard. This is generally not possible for exploratory or partially validated methods.

Performance is acceptable when observed analytical TAE is less than the PS (TAE < PS).

Performance is not acceptable when observed analytical TAE is greater than the PS (TAE > 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 TE, while higher TE would be acceptable for assays with biomarkers that have low biological variability and higher amplitude of response to stimulus.

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

Total Analytical Error (TAE) and Allowable Total Analytical Error (aTAE)

Total analytical error is often estimated by combining imprecision (SD) and average bias in the equation: total analytical error = bias + 1.65 × imprecision. This indirect estimation model (referred to as the simple combination model) often leads to an underestimation of total analytical error compared to that of a direct estimation method (referred to as the distribution-of-differences method) or of simulation (Krouwer 2002). A number of other approaches are proposed based on the use of a biomarker for decision making (Klee 2010). Many of these approaches require a clinical context for assignment of appropriate performance characteristics. A more tractable determination of appropriate quality specifications is derived from two components of biological variation, namely, within-subject (Coefficient of Variation [CV_I]) and between-subject (CV_G) variation (CV = standard deviation/mean, expressed as a percentage [Fraser et al. 1997]). These are base determinants specifying the minimum meaningful change in biomarker concentration which can be used to support or demonstrate a significant clinical change. As analytical variation (CV_A) will add variability to the "true" test result, three levels of CV_A are proposed. The optimal specification is $CV_A < 0.25*CV_I$, where CV_A comprises ~3% of CV_I. A more appropriate and widely accepted quality standard is a "desirable" specification of $CV_A < 0.5*CV_I$, where CV_A comprises ~12% of CV_I. In the situation where the desired performance is outside of the performance capability of the current technology or methodology, a minimal $CV_A < 0.75*CV_I$, where CV_A is ~25% of CV_I is proposed, with desirable specifications set as an improvement goal.

Furthermore, analytical bias (B_A) may be considered in a similar context, that is, the acceptable error associated with a measurement that would incorrectly assign a change from a group as a function of analytical performance (i.e., the error in accuracy that an effect of treatment is assigned when compared to the group of subjects receiving treatment or the subject result pre-treatment). Three tiers are proposed. Optimal bias is defined as $B_A < 0.125*(CV_I^2 + CV_G^2)^{1/2}$, (falsely assigning a maximum of 3.3% and minimum of 1.8% of subjects outside the group at a 90% confidence interval of the reference limits [mean $\pm$ 1.645 times the SE, where $SE = SD/N^{1/2}$, with N being the relevant sample size]). Desirable bias is defined as $B_A < 0.25*(CV_I^2 + CV_G^2)^{1/2}$, (falsely assigning a maximum of 4.4% and minimum of 1.4% outside the group at a 90% confidence interval). Minimal acceptable bias is thus defined as $B_A < 0.375*(CV_I^2 + CV_G^2)^{1/2}$, (5.74% and 1.4% above and below the group at a

90% confidence interval). As noted above, minimal bias specifications should only be used when the performance capability of the current technology or methodology does not facilitate achievement of desirable bias goals, the latter being a goal for enhancement of method performance.

Preliminary Determination of CV_I and CV_G

A streamlined and simplified proposal for provisional determination of CV_I and CV_G is described below (Ichihara and Boyd 2010) and is an excellent approach to define assay acceptance criteria. The example overcomes the confounding variables effect of univariate analysis by way of nested analysis of variance (ANOVA) allowing simultaneous comparison of multiple sources of variance within a single experiment. The study comprises drawing samples from a minimum of three subjects (pure component of between-individual variance, CV_G) over a minimum of three days (pure component of within-individual variance, CV_I) and measuring each specimen twice (singlicate measure on two separate days to derive in part the pure determination of analytical variance, CV_A). The subjects required for this study should either be normal (control arm of study), diseased (testing arm of the study) or replicated as both a control arm and testing arm independently (two sets of 3 subjects). The goal of the study is to determine CV_I and CV_G in one or the other of the subject groups to define performance needs. Therefore, the 3 subjects must not be a mixture of normal and diseased groups. More subjects are naturally optimal if the expected change in the biomarker is small (refer to "power" and statistician engagement comments below, where small is arbitrarily assigned as < 20%, to reflect the methodological constraints often associated with immunometric and MS based assay performance).

Table 3A: Example data and two-level Nested ANOVA for Preliminary CV_I and CV_G determination

	Sample Draw		
Subject/Assay run	Day 1	Day 2	Day 3
Subject 1 run 1	23	25	27
Subject 1 run 2	25	24	25
Subject 2 run 1	28	35	39
Subject 2 run 2	28	34	40
Subject 3 run 1	52	48	37
Subject 3 run 2	50	48	36

Two-level Nested ANOVA				Alpha	0.05	
	SS	*df*	*MS*	*F*	*p-value*	*sig*
Between individual variance (CV_G)	1244.3	2	622.17	10.04	0.0122	yes
Within individual variance (CV_I)	371.7	6	61.94	69.69	0.0000	yes
Residual	8.0	9	0.89			
Total Variance	1624	17	95.53			

Analysis of variance component (VC)

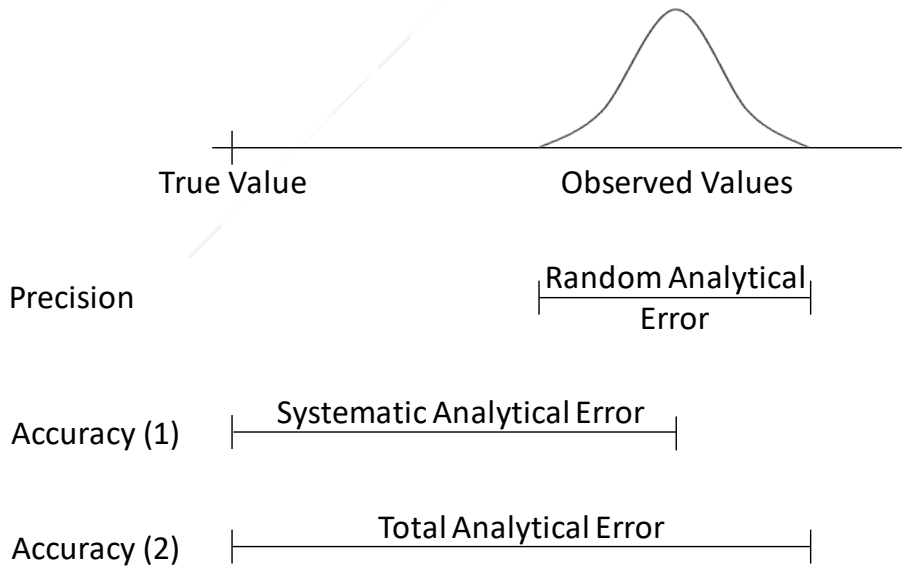
	VC	*VC, %*	*SD*	*CV (VC)*
Between individual variance (CV_G)	93.370	74.824	9.663	27.874
Within individual variance (CV_I)	30.528	24.464	5.525	15.938
Residual	0.889	0.712	0.943	2.720
Sum of variance	124.787			
Grand mean	34.67			

Using data derived from Table 5 of Ichihara and Boyd 2010, the preliminary desirable determination of analytical precision (CV_A) would be < 7.969% ($CV_A < 0.5*15.938$) and preliminary desirable analytical bias (B_A) would be < 8.027% ($B_A < 0.25*((15.938)^2 + (27.874)^2)^{1/2}$). This experiment enables determination of analytical specifications *a priori*. It is advisable that in-study analytical performance is also evaluated *a posteriori* following analysis of subjects (re-assessing CV_I and CV_G) to further refine desirable specifications or determine whether the analytical assay met the required purpose.

Determination of Total Analytical Error (TAE) and impact on confidence

To minimize TE when CV_A is large, B_A should be minimized, and conversely, when B_A is large, CV_A should be minimized, realizing that analytical precision (CV_A) and bias (B_A) are intrinsically related in the determination of TE. Quality specifications for TE may be computed a number of ways, the most usual way being the addition of bias (as an absolute value, no consideration to the positive or negative direction of bias) and precision in a linear manner (Westgard et al. 1974; Fraser 2001). Figure 3A demonstrates the influence of precision and accuracy (Bias (%) = 100 – Accuracy (%)) in a figure recreated from the literature (Westgard et al. 1974).

Figure 3A: Definitions of Precision and Accuracy in terms of Random, Systematic and Total Analytical Errors



One recommendation for the determination of TAE is TAE = B_A + 2*CV_A (Six Sigma processes may utilize TAE = B_A + 5*CV_A or B_A + 6*CV_A for characterization of test quality). In practice, TAE is routinely used (Fraser 2001; Krouwer 2002) and derived with 95% probability (confidence) of a one-sided distribution, thus allowing for a 5% error rate. When including both the upper and lower ends of the distribution, 10% of results are excluded in total. As 90% of the distribution is included in the estimation of TAE, a multiplier of 1.65 is used (Z-score, 5% excluded at both ends of a distribution). Consequently, the formula becomes TAE = B_A + 1.65*CV_A.

Analytical determination of precision is generally derived from inter-assay precision studies; however, determination of bias requires some consideration to how to define absolute truth, something not generally feasible for relative quantitative methods. For established biomarkers, higher order reference methods (with materials for testing) or comparison to existing assays are used to determine bias of new methods (Klee 2010; Westgard et al. 1974; Fraser 2001). Determination of bias in the absence of these comparators may require consideration of analytical parameters that enable calculation such as spike and recovery (Bias (%) = 100-recovery (%)), or via back-calculated bias samples of known concentration such as Lower Limit of Quantification (LLOQ) and Upper Limit of Quantification (ULOQ) replicates from inter-assay accuracy studies (Bias (%) = 100 – accuracy (%)). The influence of precision on measurement is reduced by assaying replicates in multiple runs to reduce imprecision by a factor of $n^{1/2,}$ (n= number of replicates [Fraser 2001]).

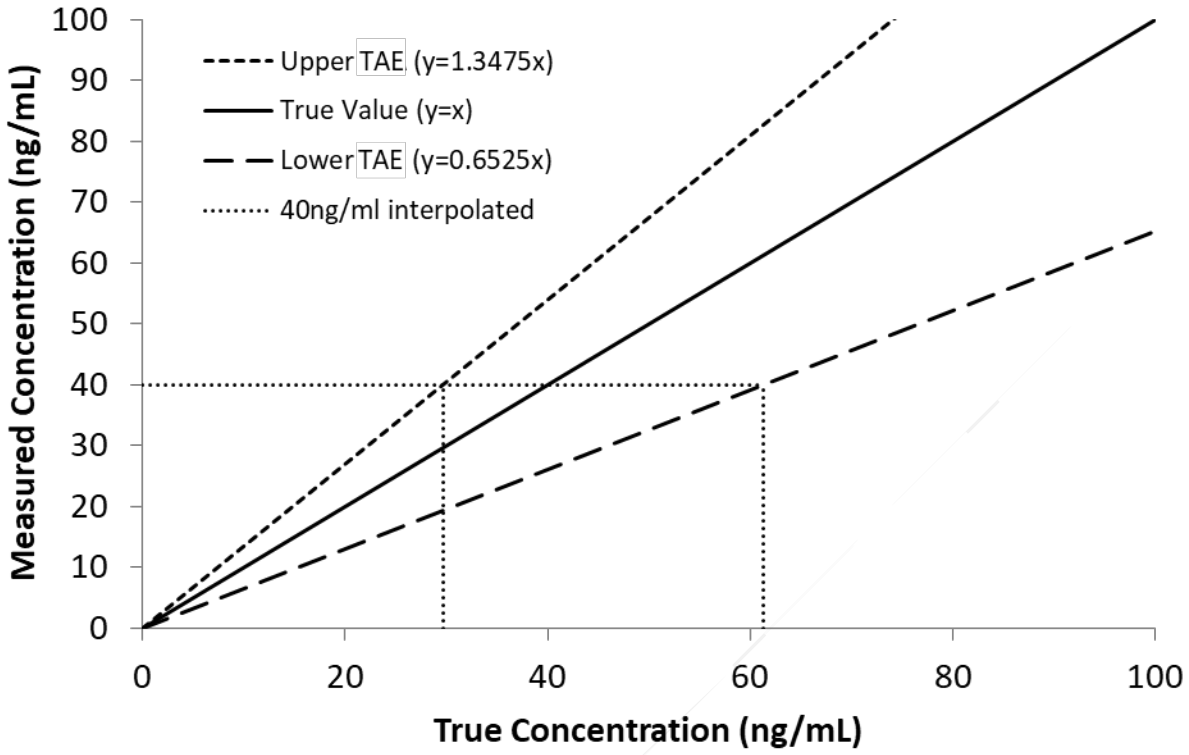
An example of the generation of TAE for an assay with a bias of 10% and precision of 15% (assumed homoscedasticity) is shown in Table 3B. The calculation of TAE = 10 + 1.65*15 = 34.75%. When analyzing samples of true concentrations (10, 30 and 50 ng/mL), the measurable concentration range incorporating TAE is calculated as upper (true result * (100+TAE)/100)) and lower boundaries (true result * (100-TAE)/100)).

Table 3B: Calculating TAE from Bias and Precision and Determining Measurement Ranges (Uncertainty)

True Concentration (ng/mL)	Bias (%, B_A)	Precision (%, CV_A)	Total Allowable Error (%, TAE)	Measured Concentration Range within TAE (ng/mL)	
				Lower	Upper
10	10	15	34.75	6.525	13.475
30	10	15	34.75	19.575	40.425
50	10	15	34.75	32.625	67.375

The results from Table 3B are graphically displayed in Figure 3B. The line of unity (solid) is bracketed with divergent TAE boundaries for upper (short dash) and lower (long dash) lines with slopes of y = 1.3475x and y = 0.6525x respectively. For analysis of a sample with a measured result of 40 ng/mL (dotted line), the true result can be interpolated from these TAE boundary conditions; lower range of true result = 29.685 ng/mL (40 ng/mL/1.3475) and upper range of true result = 61.303 ng/mL (40 ng/mL/0.6525). The range of these results represents, in part, the measurement uncertainty.

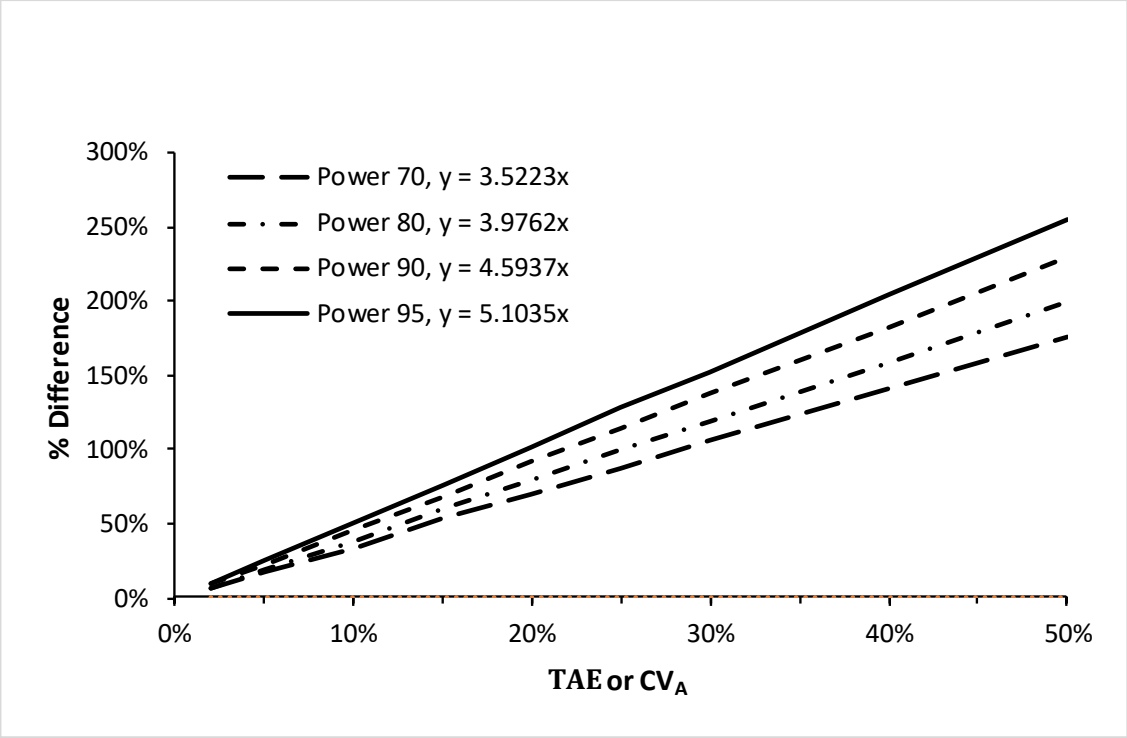
Figure 3B: Extrapolation of Measurement Uncertainty from TAE



The penultimate component of the process determines the change that may be observed in a subject following treatment (pretreatment versus post treatment measurement), where the impact of treatment upon biology is demonstrated following de-convolution of measurement uncertainty from the two measurements. Considering whether an observed difference may be assigned to biological changes requires consideration to the degree of false positivity that is acceptable (incorrectly assigning measurement error to biological change, type 1 error, in Figure 3C) together with the degree of false negativity that is acceptable (incorrectly missing biological change due to the results falling within measurement error, type 2 error). Using a 95% power as in Figure 3C, biological change would be inferred with a 5% false negative rate.

Consequently, for a TAE of 34.75%, the difference between two measures of the same subject that could be attributed to biological changes with 5% false positivity and 5% false negativity is calculated as 177.34% (biological change threshold = 5.1035*34.75%). In other words, an almost three-fold difference between two results would be necessary before there is confidence that a biological change is being observed. Lower power results in less confidence that observed biological changes are true. For example, using an 80% power (20% type 2 error) measured differences > 138.17% (greater than two-fold changes) are attributable to biological change, however, the false negative rate is 4-fold higher than at a 95% power.

Figure 3C: Influence of power analysis on measurement differences as a function of TAE or CV_A



The final step (or perhaps the first step when one considers anticipated effects sizes that are small (< 20%)) when considering applicability of the analytical method (from analytical validation studies, not *a priori* CV_I and CV_G assessment) involves the "effect size" that needs to be measured with confidence (prescribed false negative and false positive rate). While the above example demonstrates the implications of TAE when measuring one subject at two discrete time points in singlicate, smaller "effect sizes" may be discernible by incorporation of a larger number of subjects in cohorts or repeat analysis of subject specimens from all time points. We recommend that the details described thus far are used as a framework for discussion with an appropriate statistician. Ideally, the study design incorporates these criteria to discern significant biological changes from analytical limitations *a priori*, ensuring that appropriately powered studies are carried out to support the COU. A scientific justification for the selection of clinically acceptable TAE used to set analytical validation acceptance criteria should be included in biomarker qualification submissions.

Appendix 4. Parallelism

Parallelism for LBA

Parallelism is most simply understood as "recovery in dilution" where experiments attempt to prove that when an endogenous biomarker present in the biological matrix of interest is serially diluted, the dilution-corrected results report back similar values. This is in contradistinction to dilution linearity, where a similar assessment is performed, but with artificial molecules spiked into matrix.

Parallelism evaluation is an extremely relevant and necessary assessment for the clinical qualification of an endogenous fluid-based biomarker using immunometric methods (dilution linearity is not), since in a LBA, a binding interaction is being measured rather than an intrinsic physico-chemical property of the endogenous analyte being measured in the designated matrix (measurand). Consequently, it is necessary to establish that the interaction of the critical assay reagents with the calibrator material is similar to their interaction with the measurand in patient samples, resulting in parallel lines from the dilution series. Thus, there is no apparent trend or bias toward increasing or decreasing estimates of analyte concentrations over the range of dilutions when a test sample is serially diluted to produce a set of samples having analyte concentrations that fall within the calibration range of the assay and the assay is appropriate for quantification of the measurand. For this parallelism to be perfect, all dilutions will show a recovery of 100%. This, of course never occurs, as there is intrinsic analytical variability in the method itself which is best described by the inter-assay precision (CV%). Different methods of acceptance strategies are discussed below.

Parallelism assessment should be initiated early during the assay development stage, utilizing appropriate normal or disease state samples, as available. In some cases, there may be a need to wait for incurred samples (study samples) to be available to appropriately evaluate parallelism. This is especially the situation for biomarkers that are not usually present in normal subjects, or biomarkers found only in rare disease states and where samples acquired from biobanks may be inadequate depending on their provenance and storage history.

For this initial assessment, one can screen a series of samples in the proper matrix (disease-state and/or normal) to find several suitable samples, i.e., those with high endogenous concentrations of the biomarker that allow several dilutions to be made that remain within the analytical range of the method. Typically, a minimum of four serial dilutions of each sample is performed, although the actual number of dilutions performed may be more or less depending upon the specific biological constraints of the measurand, and 6-10 (when available) may be considered ideal. Here the same diluent as that used for the calibration reference material is used to dilute the biological samples. The goal is to cover the entire analytical measurement range, and/or expected concentration range of samples if the biomarker exists in a narrow physiological range (even when abnormally elevated or reduced).

There has been some consensus across industry (Lee et al. 2006, Hougton et al. 2012, Jani et al. 2016) on methodologies to evaluate parallelism. These and other methods of acceptance strategies are discussed below.

Recommended Approach: Inter-assay Precision Method

This method uses the analytical performance of the assay to determine which dilution results are within statistically relevant limits.

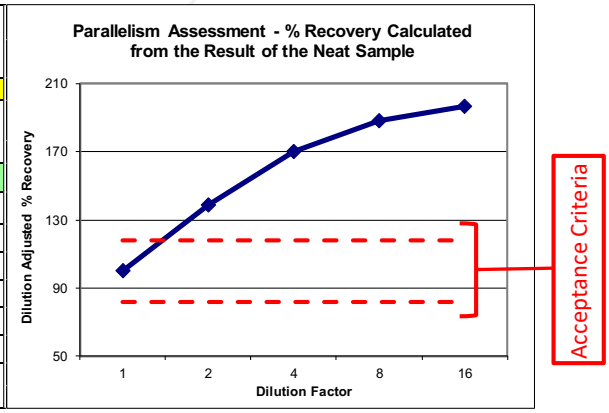
Using this method, acceptance criteria are set at ≤3 x inter-assay CV% (as calculated from the mean CV% of the Validation samples, which are ideally endogenous analyte samples spanning the range of the assay). The neat sample and relevant dilutions are analyzed on one run and results documented in a table and graphed (Steps 1 to 3 below).

Recovery results are determined using the neat result as the target (true) value of the analyte, or if this result is above the ULOQ, from the least diluted sample to generate a result within the analytical range. The % recovery of each subsequent dilution of the sample is then calculated after adjusting each result for the dilution factor.

In the example below, the method has a mean inter-assay CV of 6%, therefore acceptance limits are ± 18% so recovery must be within 82 to 118% of target. This is equivalent to ± 3SD or 99.7% CL.

Step 1 –Neat result used as the "true" value to calculate % recovery of other dilutions:

Neat Sample Result used as the "true" value for the calculaton of % Recovery in other dilutions			
Matrix Sample # 1			
Dilution Factor	Observed Result	Observed Result x Dilution Factor	% Recovery
1	413	413	100
2	287.0	574	139.0
4	175.8	703	170.2
8	97.3	778	188.4
16	50.8	813	196.9
	= Results that pass acceptance criteria *		
The method inter-assay CV% is 6.0% in this example			
Acceptance criteria is therefore +/- 18% = 82 to 118% Recovery			

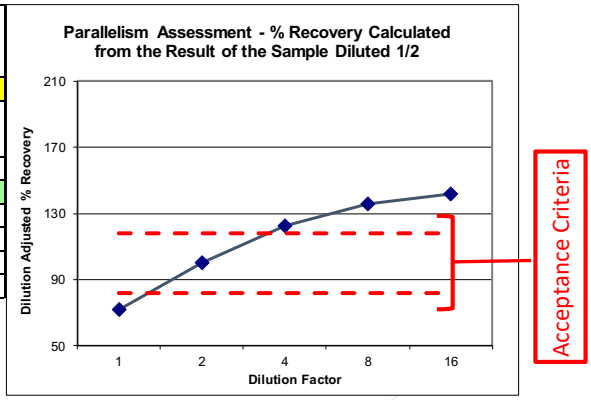


Step 1 Conclusions:

1. All results for dilutions calculated against the observed result for the Neat sample are outside acceptance limits.
2. Move to next step.

Step 2 - Dilution factor = 2 result is used as the "true" value to calculate % recovery of other dilutions:

Result from sample diluted 1/2 used as the "true" value for the calculaton of % Recovery			
Matrix Sample # 1			
Dilution Factor	Observed Result	Observed Result x Dilution Factor	% Recovery
1	413	413	72.0
2	287.0	574	100.0
4	175.8	703	122.5
8	97.3	778	135.5
16	50.8	813	141.6
	= Results that pass acceptance criteria *		

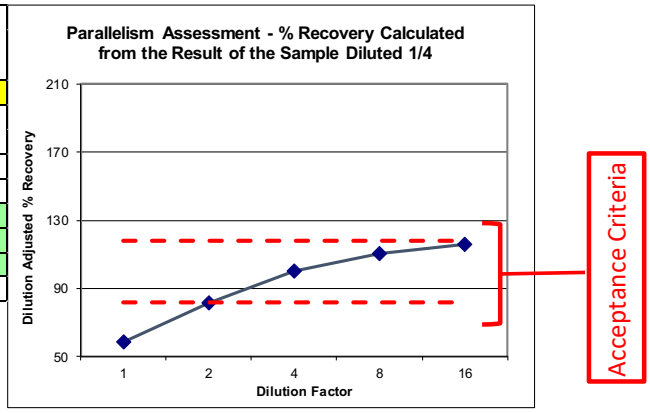


Step 2 Conclusions:

1. All results for dilutions calculated against the observed result for the sample Diluted 1/2 are outside acceptance limits.
2. Move to next step.

Step 3 - Dilution factor = 4 result is used as the "true" value to calculate % recovery of other dilutions:

Result from sample diluted 1/4 used as the "true" value for the calculaton of % Recovery			
Matrix Sample # 1			
Dilution Factor	Observed Result	Observed Result x Dilution Factor	% Recovery
1	413.0	413	58.7
2	287.0	574	81.7
4	175.8	703	100.0
8	97.3	778	110.7
16	50.8	813	115.6
	= Results that pass acceptance criteria *		



Step 3 Conclusions:

1. Results for dilutions calculated against the observed result for the sample Diluted 1/4 are within acceptance limits at dilutions of 1/8 and 1/16.
2. The data support using a dilution of test samples between1/4 and 1/16.
3. The method minimum required dilution = 4

Data for multiple samples can be plotted on the same chart to provide a visual picture of how many samples pass or fail the parallelism assessment. Note that a high number of parallelism failures with a given disease/subject population may be indicative that the MRD needs to be adjusted for that population. In some instances, in practice it may be necessary to test samples at multiple dilutions to ensure that reliable data are obtained. As more data are generated, and a better understanding of sample interferences is understood a final method MRD may then be established.

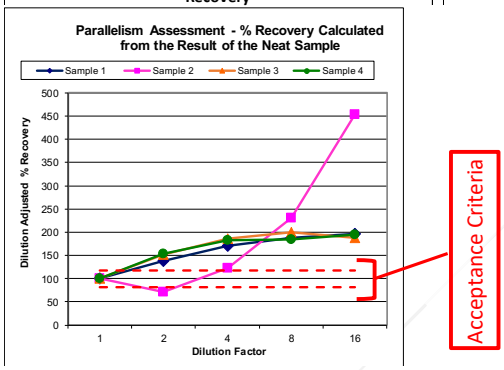
It is recommended that data be graphed to easily identify samples that pass versus fail the assessment as data tables can easily be misinterpreted. As an example, Steps 1A – 3A below show data from 3 additional samples.

Step 1A –Neat result used as the "true" value to calculate % recovery of other dilutions:

Neat Sample Result used as the "true" value for the calculaton of % Recovery in other dilutions			
Matrix Sample # 2			
Dilution Factor	Observed Result	Observed Result x Dilution Factor	% Recovery
1	75.1	75.1	100
2	26.5	53	70.6
4	23.0	92.1	122.6
8	21.6	173	230.4
16	21.3	341	454.1
	= Results that pass acceptance criteria *		
The method inter-assay CV% is 6.0% in this example			
*Acceptance criteria is therefore +/- 18% = 82 to 118% Recovery			

Neat Sample Result used as the "true" value for the calculaton of % Recovery in other dilutions			
Matrix Sample # 3			
Dilution Factor	Observed Result	Observed Result x Dilution Factor	% Recovery
1	185	185	100
2	140.5	281	151.9
4	86.5	346	187.0
8	46.3	370	200.0
16	21.8	348	188.1
	= Results that pass acceptance criteria *		
The method inter-assay CV% is 6.0% in this example			
*Acceptance criteria is therefore +/- 18% = 82 to 118% Recovery			

Neat Sample Result used as the "true" value for the calculaton of % Recovery in other dilutions			
Matrix Sample # 4			
Dilution Factor	Observed Result	Observed Result x Dilution Factor	% Recovery
1	219	219	100
2	167.5	335	153.0
4	99.8	399	182.2
8	50.4	403	184.0
16	26.75	428	195.4
	= Results that pass acceptance criteria *		
The method inter-assay CV% is 6.0% in this example			
*Acceptance criteria is therefore +/- 18% = 82 to 118% Recovery			



Step 1A Conclusions (sample #1 results added to graph):

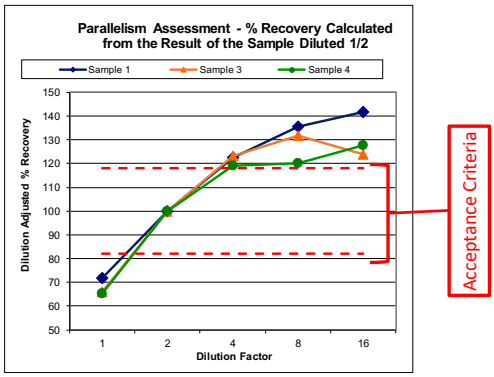
1. All results for dilutions calculated against the observed result for the Neat sample are outside acceptance limits.
2. Move on to the next step.

Step 2A - Dilution factor of 2X result is used as the “true” value to calculate % recovery of other dilutions:

Result from sample diluted 1/2 used as the "true" value for the calculaton of % Recovery			
Matrix Sample # 2			
Dilution Factor	Observed Result	Observed Result x Dilution Factor	% Recovery
1	75	75	141.5
2	26.5	53	100.0
4	23.0	92.1	173.8
8	21.6	173	326.4
16	21.3	341	643.4
	= Results that pass acceptance criteria *		

Result from sample diluted 1/2 used as the "true" value for the calculaton of % Recovery			
Matrix Sample # 3			
Dilution Factor	Observed Result	Observed Result x Dilution Factor	% Recovery
1	185	185	65.8
2	140.5	281	100.0
4	86.5	346	123.1
8	46.3	370	131.7
16	21.8	348	123.8
	= Results that pass acceptance criteria *		

Result from sample diluted 1/2 used as the "true" value for the calculaton of % Recovery			
Matrix Sample # 4			
Dilution Factor	Observed Result	Observed Result x Dilution Factor	% Recovery
1	219	219	65.4
2	167.5	335	100.0
4	99.8	399	119.1
8	50.4	403	120.3
16	26.75	428	127.8
	= Results that pass acceptance criteria *		



Step 2A Conclusions (sample #2 results removed from graph for clarity):

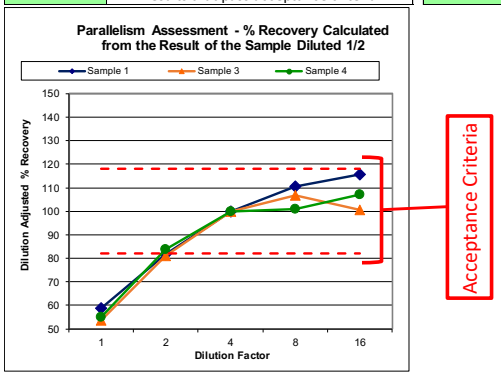
1. All results for dilutions calculated against the observed result for the sample diluted 1/2 are outside acceptance limits.
2. Move on to the next step.

Step 3A - Dilution factor of 4X result is used as the "true" value to calculate % recovery of other dilutions:

Result from sample diluted 1/4 used as the "true" value for the calculaton of % Recovery			
Matrix Sample # 2			
Dilution Factor	Observed Result	Observed Result x Dilution Factor	% Recovery
1	75	75	81.4
2	26.5	53	57.5
4	23.0	92.1	100.0
8	21.6	173	187.8
16	21.3	341	370.2
	= Results that pass acceptance criteria *		

Result from sample diluted 1/4 used as the "true" value for the calculaton of % Recovery			
Matrix Sample # 3			
Dilution Factor	Observed Result	Observed Result x Dilution Factor	% Recovery
1	185	185	53.5
2	140.5	281	81.2
4	86.5	346	100.0
8	46.3	370	106.9
16	21.8	348	100.6
	= Results that pass acceptance criteria *		

Result from sample diluted 1/4 used as the "true" value for the calculaton of % Recovery			
Matrix Sample # 4			
Dilution Factor	Observed Result	Observed Result x Dilution Factor	% Recovery
1	219	219	54.9
2	167.5	335	84.0
4	99.8	399	100.0
8	50.4	403	101.0
16	26.75	428	107.3
	= Results that pass acceptance criteria *		



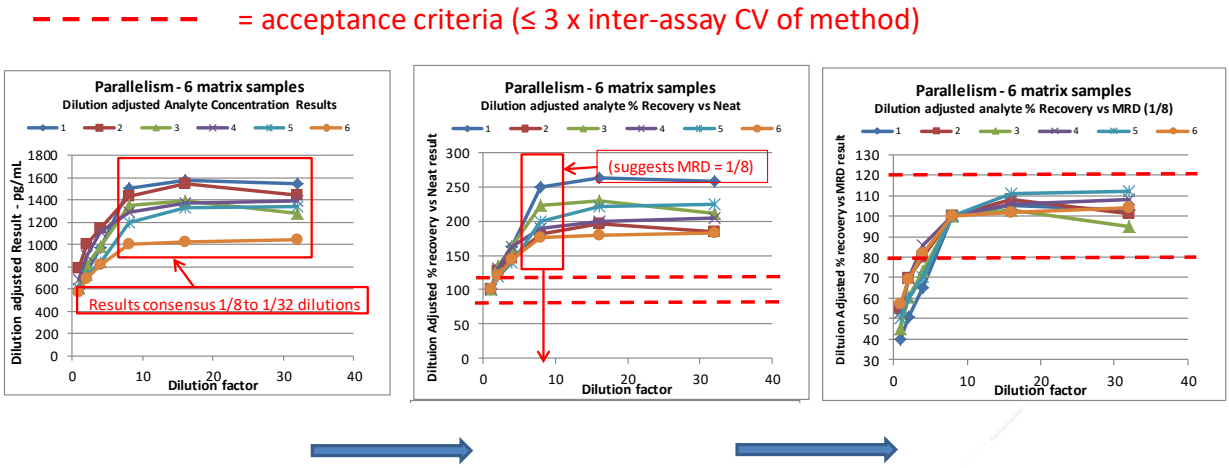
Step 3A conclusions:

1. Results for dilutions calculated against the observed result for samples diluted 1/4 are within acceptance limits at dilutions of 1/8 and 1/16.
2. The data support using a dilution of test samples between 1/4 and 1/16.
3. The method MRD = 4.

It should be noted that Sample #2 has a different profile to the other three samples since a significant positive bias is observed regardless of dilution. It is clear that this sample contains a substance(s) that is interfering with the method which may be subject specific. Seeing this data in such a small number of samples indicates that testing parallelism on a larger number of samples will be necessary to understand the incidence of such interferences in the patient population. Assessments may need to continue during the sample analysis phase to ensure that samples that have high concentrations are not the result of similar interferences as in sample #2. However, a real study sample demonstrating this profile may need to be reported as "No Result due to unidentified assay interference", if further investigation with the sample is not possible due to informed consent restrictions.

This method of data interpretation has previously been recommended in multiple white papers (Hougton et al. 2012, Jani et al. 2016, Lee et al. 2006). An additional example of application of this methodology is reproduced below (Figure 4A) with permission of the authors.

Figure 4A: Graphical Display and Confirmation of Parallelism



By comparison, the method commonly described for PK assay parallelism assessments involves calculating the CV of the dilution adjusted results from a dilution series to demonstrate whether it is within predetermined limits (e.g., %CV ≤ 15% for LC-MS, 20-25% for hybrid LC-MS and ≤ 30% for LBA) such as the example in Table 4A. although with the caution that data sets should be examined carefully as the CV criterion can be met even when significant bias remains (Global Bioanalysis Consortium). In the case of biomarker assays, it is not possible to apply global criteria as appropriate criteria will be inextricably dependent upon the specified context of use.

Table 4A: Example of biomarker assay with pre-specified CV criterion of 25%.

		Dilution Corrected Biomarker Concentration (pg/mL)			
Dilution Fold	1/Dilution	Sample 1	Sample 2	Sample 3	Sample 4
1	1.000	413*	75.1*	185*	219*
2	0.500	574	53.0	281	335
4	0.250	703	92.1	346	399
8	0.125	778	173	370	403
16	0.0625	813	341	348	428
	CV (%):	14.8	77.5	11.4	10.2

*Measured values for neat samples not included in calculation for %CV as they would have resulted in CVs higher than the 25% maximum specified for this assay by the TAE.

Using this method, the TAE for the COU is used to drive acceptance criteria in proving parallelism. However, using this approach would indicate that a two-fold minimum dilution is adequate (excluding sample 2 as in prior methods). However, the possible %Bias seen even after eliminating sample #2 as above, is from 100 to 142% (see tables and charts from Steps 1-3). Using the statistically based approach, for this data to pass a 99.7% CL, the inter-assay CV% of the method being used is required to be ≥ 14%. If the actual Inter-assay CV% of the method is < 14.0% then these data would be outside 99.7% CL and therefore unacceptable.

If we use the Inter-assay Precision method which demonstrates that the correct MRD is a 4-fold dilution the degree of bias seen in the same three samples is reduced from 100 - 142% to 100 – 116%. Whilst the TAE may allow large variability, the assay performance in this case (Inter-assay CV%) dictates that data results outside the 3 x CV% are highly unreliable. It is recommended that either the correct MRD be used to remain statistically aligned with the method performance or a method that suffers less from matrix interference should be used. It is acknowledged that cases may exist where a need for greater sensitivity may lead to consideration of a lesser MRD that is not optimal but meets assumed needs of the COU. However, this is expected to be an extremely rare occurrence and should be pursued with caution.

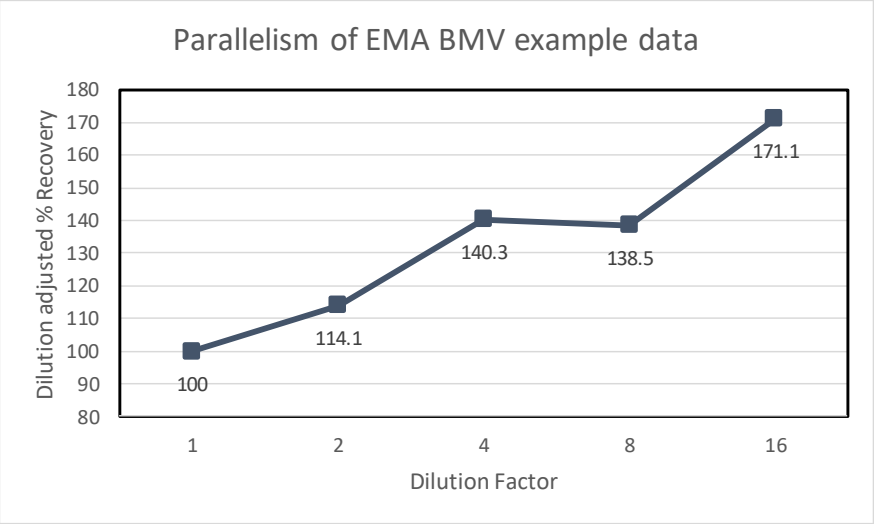
To further demonstrate the benefit of using a statistically based approach to define acceptance criteria for proof of parallelism, an additional example is provided (Table 4B, Figure 4B) that applies classical PK acceptance criteria. Although the data set in Table 4B clearly meets the ≤30% CV LBA acceptance criterion (EMA 2011; FDA 2018), Figure 4B demonstrates the presence of a significant positive bias (up to 171%). Such data contradict the definition of parallelism whereby dilution does not result in biased measurements (trending up or down) of the analyte concentration.

Table 4B: Parallelism in an LBA

Parallelism in an LBA: BMV guidelines			
Dilution Factor	Mean conc. (ng/mL)	Dilution adjusted mean	Precision of series (%)
1	279.6	279.6	
2	159.5	319	
4	98.1	392.4	20.6
8	48.4	387.2	
16	29.9	478.4	

LBA <25% passes BMV criteria

Figure 4B:



Conclusion

Once the validation data have been correctly interpreted, the performance results can be assessed against the COU to ensure the method can deliver data within the set TAE limits. However, while tolerance for inherent variability in overall method performance (precision and relative accuracy) may vary with COU and acceptancy criteria tightened or loosened accordingly, even when variability is within the TAE of the COU, it is important that sample results are reliable within those defined limits. The Inter-assay Precision Method of performing parallelism ensures this is the case.

Beyond demonstration of parallelism itself using the recommended methodology one can assess multiple assay parameters by leveraging a single well-designed experiment to inform assay MRD, selectivity and provisional lower limit of quantification with respect to the endogenous analyte. (Stevenson and Purushothama 2014).

Dilutional Linearity:

Dilutional Linearity in the context that we are using it here is NOT parallelism. The difference is that whilst both are conducted using real matrix dilutional linearity is conducted using recombinant or other “artificial” molecules spiked into the matrix before constructing a dilution series whereas parallelism uses a dilution series constructed using endogenous molecules present in the matrix.

Testing of dilutional linearity will give some insight into potential matrix effects and an estimation of an MRD if required. However, it does not replace parallelism testing which must still be conducted when suitable samples become available to verify any matrix effect findings from dilutional linearity.

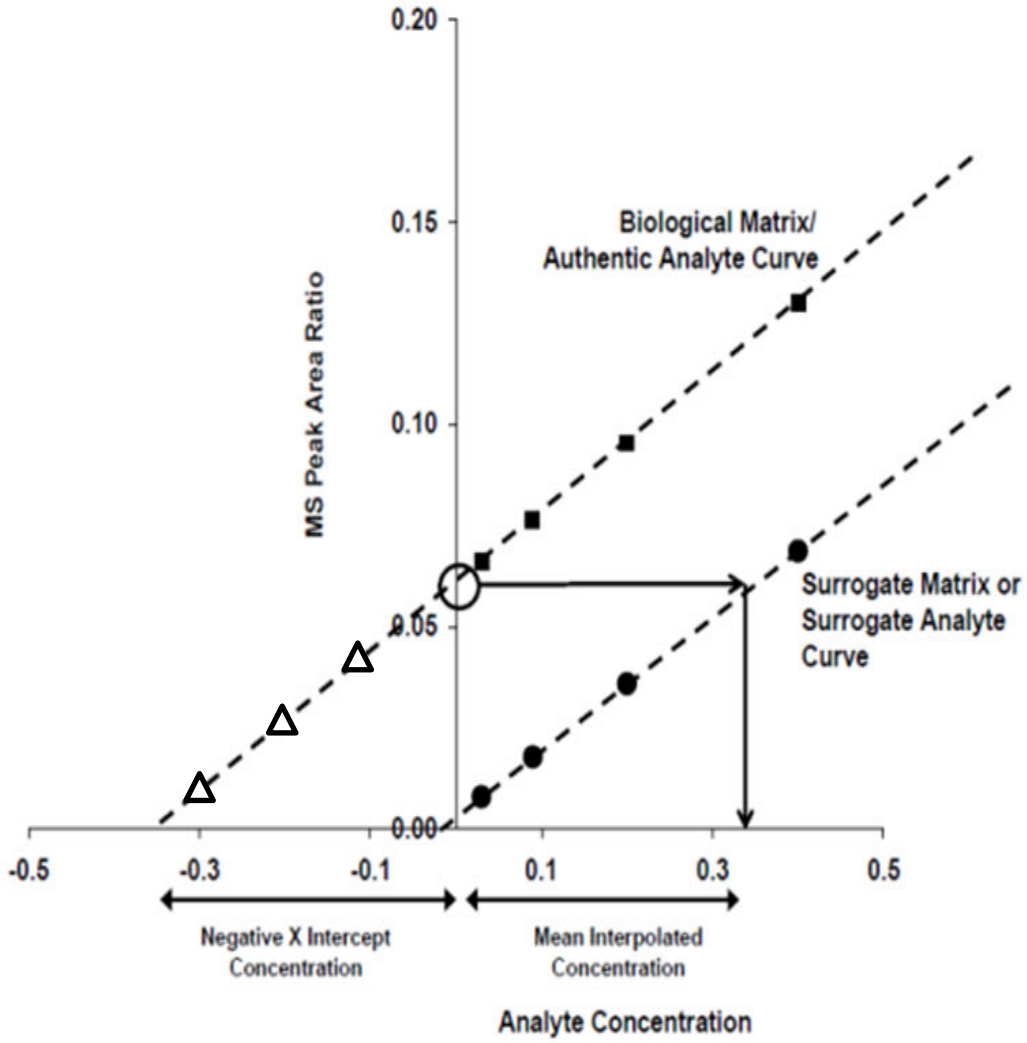
Prozone effect:

Prozone effects are primarily limited to homogeneous assays (without washes between binding and detection steps) and to analytes which have a physiological range greater than the AMR.

Parallelism for Small Molecules by LC-MS

Parallelism assessment for small molecule biomarkers using the surrogate matrix approach is relatively straightforward owing to several properties unique to LC-MS. In contrast to proteins, high quality reference standards are available with an identical chemical structure as the endogenous small molecule. In addition, small molecules exhibit far less matrix binding interference due to disruption of these interactions during sample preparation; making it easier to use synthetically prepared surrogate matrices (e.g. bovine serum albumin/phosphate buffered saline as a substitute for plasma or serum). A further advantage is that the linear calibration curves associated with MS are well-suited to parallelism assessment. This characteristic is illustrated in Figure 4C which displays a graphical representation of standard curves prepared in either surrogate or authentic sample matrix using a common set of analyte spiking solutions. This figure, adapted from a paper by Jones, et. al. (2012), illustrates three methods for parallelism assessment to be performed during method development to qualify a surrogate matrix for subsequent use in validation: spike-recovery, dilutional linearity and standard addition. The back-calculated error of each point on the upper, authentic matrix curve provides an indication of spike-recovery when measured relative to the lower surrogate matrix calibration curve. Dilutional linearity is used to assess parallelism for samples less concentrated than the authentic matrix pool and is shown conceptually by the open triangle symbols to the left of the y-axis. A third assessment involves a comparison of calculated concentrations of the unspiked matrix pool determined by two methods: 1) extrapolation of the spiked authentic matrix curve through the negative x-axis using the method of standard addition and 2) calculation by direct measurement using the surrogate matrix calibration curve (interpolation). Agreement between these values serves as a demonstration of parallelism. The merit of each of these approaches was recently discussed in a commentary on measuring the accuracy for endogenous analytes by Jenkins (2016).

Figure 4C: Parallelism Assessment in LC-MS assays. Adapted from Jones et al. (2012) with permission of Future Science Ltd.



Despite the information gained in the surrogate matrix qualification experiments described above, it is important to note that true confirmation of parallelism comes through the precision and accuracy results from a multi-day, 3-run assay validation experiments. Similar to the FDA Bioanalytical Method Validation (BMV) guidance (FDA 2018), five levels of validation QCs are prepared across the analytical range. Levels higher than the endogenous pool are obtained by spiking whereas those below are prepared by dilution with surrogate matrix. Typically, the grand mean from the 3 validation experiments is used to assign the endogenous pool concentration, although

determination prior to validation through multiple analyses with the intended method is also an accepted practice (Welink 2017).

An example of a multi-analyte small molecule biomarker assay validated toward full regulatory expectations was recently published by Cox et al. (2015) This assay, which analyzed concentrations of four neurotransmitter metabolites in human cerebrospinal fluid (CSF), incorporated benzoyl chloride derivatization to promote LC retention and MS sensitivity. Validation data from this assay, which employed artificial CSF containing 0.2% BSA as the surrogate matrix, are found in the reference.

LC-MS Proteins

Given the various methodologies for LC-MS protein quantification and the rapidly evolving nature of this field, it is not surprising that there are inconsistent practices for assessing parallelism across the biomarker field. Three categories for analysis of proteins by LC-MS were identified in the AAPS white paper on biotherapeutic protein quantification: conventional extraction, immunoaffinity extraction/protein target, and immunoaffinity extraction/peptide target (Jenkins 2015). In addition to this diversity, proteins may be measured intact (top-down) or after enzymatic digestion to yield a surrogate peptide (bottom-up). Despite these differences, the use of protein standards for calibration is strongly recommended along with adherence, wherever possible, to the practices described above for surrogate matrix qualification and parallelism assessment for small molecules. Because of acknowledged issues associated with the use of recombinant proteins as standards, protein quantification by MS methods are generally regarded as having relative accuracy. Finally, surrogate matrix selection for proteins is more difficult owing to the increased presence of binding partners for proteins in biological matrices, as well as the ability of proteins or surrogate peptides to non-specifically bind to labware (although these binding issues may also occur for small analyte biomarkers as well). Nonetheless, successful application of surrogate matrices for protein biomarkers in accessible circulating fluids is possible using a variety of methods including immunodepletion or the use of matrix from a different species. It is further recommended that immunoaffinity-MS methods consider the MRD approach used with LBA to minimize matrix effects.

LC-MS Surrogate Analyte

Surrogate analyte methods take advantage of the closeness in physiochemical properties of SIL analogs and their native analyte counterparts, while retaining distinguishability based on mass (Li and Cohen 2003; Jemal 2003). This feature avoids the issue of signal superposition when spiking a calibrator analyte into a biological control matrix. While notable examples have been reported for biomarkers using the surrogate analyte method (Jian 2010; Ongay 2014), insufficient consensus currently exists on the experimental protocol for validation of these methods. Because the key assumption behind this approach is the analytical equivalence of the two forms, it is essential that any empirical difference in response for equimolar mixtures be normalized using a correction factor. Tuning the MS instrument to achieve response balance is another way to address this issue (Jones 2012). In either case, it is important that appropriate SOPs or concise validation plans exists for these procedures, as well as the methods used to establish isotopic purity of the reagents used.

Although surrogate analyte methods use authentic biological matrix for calibration, a demonstration of parallelism between calibrant and endogenous analyte is important. To assess parallelism, we recommend that a QC sample spiked at the ULOQ be serially diluted over the range of analysis using control matrix, and dilution-corrected recovery be calculated. The ability to successfully quantitate QC samples prepared by spiking native analyte is offered as a further measure of assessment and it should be noted that the method of standard addition is also possible (Jones 2012). Despite the common use of SIL peptides as standards for targeted proteomics (reverse curves), implementation of this practice for fully validated biomarker methods is difficult owing to the mandate for protein standards as calibrants.