Pharmaceutical Research, Volume 23, No. 2, February 2006 (© 2006)
DOI: 10.1007/s11095-005-9045-3

Research Paper

Fit-for-Purpose Method Development and Validation for Successful Biomarker Measurement

Jean W. Lee,[1,16,17] Viswanath Devanarayan,[2] Yu Chen Barrett,[3] Russell Weiner,[3] John Allinson,[4] Scott Fountain,[5] Stephen Keller,[6] Ira Weinryb,[7] Marie Green,[8] Larry Duan,[9] James A. Rogers,[10] Robert Millham,[10] Peter J. O'Brien,[11] Jeff Sailstad,[12] Masood Khan,[13] Chad Ray,[14] and John A. Wagner[15]

Received July 28, 2005; accepted October 7, 2005

Abstract. Despite major advances in modern drug discovery and development, the number of new drug approvals has not kept pace with the increased cost of their development. Increasingly, innovative uses of biomarkers are employed in an attempt to speed new drugs to market. Still, widespread adoption of biomarkers is impeded by limited experience interpreting biomarker data and an unclear regulatory climate. Key differences preclude the direct application of existing validation paradigms for drug analysis to biomarker research. Following the AAPS 2003 Biomarker Workshop (J. W. Lee, R. S. Weiner, J. M. Sailstad, *et al.* Method validation and measurement of biomarkers in nonclinical and clinical samples in drug development. A conference report. *Pharm Res* **22**:499–511, 2005), these and other critical issues were addressed. A practical, iterative, "fit-for-purpose" approach to biomarker method development and validation is proposed, keeping in mind the intended use of the data and the attendant regulatory requirements associated with that use. Sample analysis within this context of fit-for-purpose method development and validation are well suited for successful biomarker implementation, allowing increased use of biomarkers in drug development.

KEY WORDS: assay validation; biomarkers; drug development; fit-for-purpose method development and validation.

[1] Formerly MDS Pharma Services, Lincoln, Nebraska, USA.
[2] Merck and Company, Inc., Blue Bell, Pennsylvania, USA.
[3] Bristol-Myers Squibb, Princeton, New Jersey, USA.
[4] BAS Analytics Ltd., Kenilworth, UK.
[5] Pfizer Global Research and Development, Ann Arbor, Michigan, USA.
[6] Protein Design Labs, Inc., Fremont, California, USA.
[7] Wyeth Research, Collegeville, Pennsylvania, USA.
[8] Millenium Pharmaceuticals, Cambridge, Massachusetts, USA.
[9] Quest Pharmaceutical Services, Newark, Delaware, USA.
[10] Pfizer Global Research and Development, Groton–New London, Connecticut, USA.
[11] Therakos, Inc., Exton, Pennsylvania, USA.
[12] Trimeris Inc., Morrisville, North Carolina, USA.
[13] Covance Laboratories, Inc., Chantilly, Virginia, USA.
[14] Eli Lilly and Company, Indianapolis, Indiana, USA.
[15] Merck and Company, Inc., Rahway, New Jersey, USA.
[16] Present Address: 1 Amgen Center Drive, Mailstop 30E-3-B, Thousand Oaks, California 91320, USA.
[17] To whom correspondence should be addressed. (e-mail: jwlee@amgen.com)

ABBREVIATIONS: AAPS, American Association of Pharmaceutical Sciences; BQL, below quantifiable limit; CDER, Center for Drug Evaluation and Research; CMS, Centers for Medicare and Medicaid Services; CLAS, Clinical Ligand Assay Society; CLIA, Clinical Laboratory Improvement Amendments; CLSI, Clinical and Laboratory Standards Institute; GLP, Good Laboratory Practices; LBABFG, Ligand Binding Assay Bioanalytical Focus Group; LOD, lower limit of detection; LLOQ, lower limit of quantification; MRD, minimum required dilution; NCCLS, National Committee for Clinical Laboratory Standards; PD, pharmacodynamic; PK, pharmacokinetic; QC, Quality Controls; QL, quantification limits; ULOQ, upper limit of quantification; VEGF, vascular endothelial growth factor; VS, validation sample.

INTRODUCTION

Historical Perspectives and Scope

Major advances in the basic science of drug discovery and development have led to an enormous increase in the number of new drug targets; however, despite increasing commitments of time and money to the effort, these advances have not culminated in an increase in new drug approvals (2,3). Consequently, efforts to improve the efficiency of this process are being implemented across the continuum of drug development activities (4–10). The use of biomarkers to identify the most promising drug candidates may ultimately allow a more economical and timely application of developmental resources.

Clinicians have traditionally used biomarkers, typically laboratory and other measures, to monitor therapeutic progress, disease progression, and the efficacy of interventions. Only recently has this use become formalized in drug development. Such increased use has been accompanied by extensive and at times confusing application of the terms "biomarker" and "validation" to variously related activities, highlighting the need for harmonization of terminology and validation approaches. The confusion is compounded by a general absence of official guidelines for the validation of laboratory biomarker assays, leading to inconsistent adaptations of related regulations (11,12) in both bioanalytical and clinical laboratories. In May 2001, the FDA issued guidance for industry for bioanalytical method validation, addressing validation of assays to support pharmacokinetic (PK) assessments of conventional small molecule drugs (11). Meanwhile, laboratories that perform testing on human specimens for diagnosis, prevention, or treatment of any disease or impairment, or for the assessment of the health of individual patients, are certified under the Clinical Laboratory Improvement Amendments (CLIA) of 1988 or have similar accreditation in countries outside the US (13). The standard practices most frequently required for CLIA certification were developed and published by the Clinical and Laboratory Standards Institute [CLSI, formerly the National Committee for Clinical Laboratory Standards (NCCLS)] (14). Therefore, because of the diverse nature of biomarker analysis and its varied applications in drug development, neither the FDA bioanalytical drug assay guidance nor the CLSI guidelines fully meet the needs of drug development and diagnostic applications of biomarker assays (1). Table I compares and contrasts these two validation paradigms, and illustrates some of the unique validation challenges of biomarker assays.

To further define and address these challenges, the American Association of Pharmaceutical Scientists (AAPS) and Clinical Ligand Assay Society (CLAS) cosponsored a Biomarker Method Validation Workshop in October 2003 (1). Members of the AAPS Ligand Binding Assay Bioanalytical Focus Group Biomarker Subcommittee subsequently

Table I. Comparison of Pharmacokinetic (PK) Drug, Biomarker, and Diagnostic Assays

	PK assay	Biomarker assay for drug development	Biomarker assay for diagnostic
Intended use	Bioequivalence, PK	Safety, mechanism of action, PD	Distinguish diseased from healthy
Method category	Most assays are definitive quantitative	Most assays are relative or quasiquantitative	
Nature of analyte	Exogenous in most cases	Endogenous	
Calibrators/Standards	Well characterized. Standards prepared in study matrix	Typically not well characterized, may change from vendor to vendor, lot to lot. Standards/calibrators are made in matrix different than study samples	
Validation samples (VS) and quality control (QC)	Made in study matrix. 4–5 VS levels and 3 QC levels	Made in study matrix. 5 VS levels and 3 QC levels. If study matrix is limited, (e.g., tissue samples) may use surrogate matrix	QC often in lyophilized form, supplied by the vendors, commonly 2 or 3 levels
Assay sensitivity	LLOQ defined by acceptance criteria	LLOQ and LOD	LOD is often used
Validation of accuracy	True accuracy can be achieved by testing spike recovery	In majority of cases only relative accuracy can be achieved. Endogenous background needs to be considered if spike recovery is used	Measured result compared to an accepted reference value obtained by an accepted method
Validation of precision	2–6 replicate samples per run, 3–6 runs	2–6 replicate samples per run, 3–6 runs	3 replicate samples per run, one run per day for 5 days. Samples ran in random order
Stability testing	Freeze/thaw, bench top, and long-term measured by spiking biological matrix with drug	Freeze/thaw, bench top, and storage stability with study samples, when available. If not, with spiked samples	Focus on stability of reagents rather than analytes. Long-term analyte stability not routinely tested
Assay acceptance criteria	4–6–20/30 rule	Establish confidence interval or 4–6–X rule	2 SD ranges, Westgard Rules, Levy–Jennings Chart
Regulatory requirements	GLP compliant	No specific guidelines	Methods are FDA-approved, result generation follows CLIA and CLSI guidelines in US

collaborated to develop an approach to validate laboratory biomarker assays in support of drug development. Although the focus is on ligand-binding methods to gauge biomarkers measured *ex vivo* from body fluids and tissues, many of the recommendations can generally be applied across other platforms and types of biomarkers. Here we summarize the results of these discussions, and describe a "fit-for-purpose" approach for biomarker method development and validation. The key component of this approach is the notion that assay validation should be tailored to meet the intended purpose of the biomarker study, with a level of rigor commensurate with the intended use of the data.

Nomenclature

Numerous publications have described the validation and use of biomarkers in clinical and nonclinical drug development. The nomenclature associated with this field, however, is not standardized. It was agreed that the NIH working group's definition of a biomarker would be used. A *biomarker* is thus "a characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacologic response to a therapeutic intervention" (15). A *clinical endpoint* is defined as a "characteristic that reflects how a patient feels, functions or survives"; a *surrogate endpoint* is a "biomarker intended to substitute for a clinical endpoint." The many contexts in which the term "surrogate marker" is used can generate substantial confusion. Therefore we recommend the term not be used interchangeably with "biomarker." Furthermore, an important distinction should be made between biomarker analytical method validation and clinical qualification. Analytical method *validation* is the process of assessing the performance characteristics of a given assay, whereas clinical *qualification* is the evidentiary and statistical process linking biologic, pathologic, and clinical endpoints to the drug effect, or linking a biomarker to biologic and clinical endpoints.

The NIH working group recommended that the term "validation" be used for analytical methods, and "evaluation" be used for the determining surrogate endpoint candidacy of biomarkers (15). More recently, the term "qualification" has been suggested for biomarker clinical evaluation to avoid confusion with the process of method validation (16,17). Biomarker method validation and biomarker qualification can be intertwined, because the statistical linkages of disease regression, biomarker modulation, and drug effects can depend upon the analytical soundness of a biomarker assay. The fit-for-purpose validation approach addresses the extent to which a biomarker assay should be validated vis-à-vis the intended purpose for which the data are being generated. For example, the validation of a surrogate endpoint assay would require the most rigorous validation and assay performance on the continuum of validations described below (17).

Categories of Biomarker Assay Data

Understanding exactly what is being measured and its biological relevance is crucial to the utility of biomarker data, as is an understanding of the limits of data produced in a given assay format. Lee and colleagues (18) defined categories of biomarker data that reflect the type of assay employed.

A *definitive quantitative assay* uses calibrators fit to a regression model to calculate the absolute quantitative values for unknown samples. Such assays are only possible when the reference standard is well defined and *fully* representative of the endogenous biomarker, such as in the case of small molecule bioanalytes (for example, steroids). Definitive quantitative assays using either physicochemical or biochemical methods can be validated to be accurate and precise. To confirm that these assays are reflective of the biological activities of the biomarker, orthogonal assays are sometimes performed.

A *relative quantitative* assay depends upon a response–concentration calibration function. However, as is the case for many cytokine immunoassays, reference standards may not be available in a purified form fully characterized, or fully representative of an endogenous biomarker. In such cases, precision performance can be validated but accuracy can only be estimated.

A *quasiquantitative assay* (quasi: "possesses certain attributes") does not employ the use of a calibration standard, but has a continuous response and the analytical result is expressed in terms of a characteristic of the test sample. For example, antibody titers of antidrug antibody assays can demonstrate assay precision, but not accuracy.

A *qualitative assay* generates categorical data that lack proportionality to the amount of analyte in a sample. Such data may be nominal, such as the presence or absence of a gene or gene product, or ordinal with discrete scoring scales like those often used for immunohistochemical assays. In general, qualitative methods are more applicable for differentiating marked effects such as the all-or-none effect of gene expression, or effects on relatively homogenous cell populations.

In all but definitive quantitative assays, the use of experimental designs that provide appropriate comparison controls, such as placebo or normal control subjects, are necessary. A full discussion of quasiquantitative and qualitative assays and related statistical considerations is beyond the scope of this paper. Readers may refer to the report of Mire-Sluis *et al.* (19), which describes recommendations for the design and optimization of immunogenicity immunoassays. In this work we build on recommendations relating to ligand-binding assays for macromolecule drugs made by DeSilva and colleagues (20), and focus on issues relevant to definitive and relative quantitative biomarker assays.

Fit-for-Purpose Biomarker Method Validation

Generally, validation should demonstrate that a method is "reliable for the intended application" (21,22). Accordingly, the rigor of biomarker method validation increases as the biomarker data are used for increasingly advanced clinical or otherwise business-critical decision making. For biomarker assays, we propose the adoption of a continuous and evolving *fit-for-purpose* strategy. Fit-for-purpose method validation provides for efficient drug development by conserving resources in the exploratory stages of biomarker characterization. For example, a biomarker under exploratory development in an early phase clinical trial would be less

rigorously validated than an already well-qualified biomarker in the same trial. By definition, exploratory biomarker data would be used for less critical decisions than data describing a well-qualified biomarker. In the latter case, an advanced assay validation would be required to ensure adequate confidence in the measurements. The 2003 Biomarker Conference Report provides a snapshot of trends in biomarker applications across the drug development continuum, and offers examples illustrating the concept of fit-for-purpose, stage-appropriate method validation (1). Figure 1 and Table II further illustrate that biomarker method validation is a graded, cyclical process of assay refinement with validation criteria that are appropriate for the intended use of the resulting data. Table II was designed as a ready guide for researchers to consult once the objectives of a given study have been established and along with recommendations for the level of rigor to apply to various levels of assay validation, incorporates previously described assay validation parameters (19,20).

Three major factors influence the establishment of assay acceptance criteria, which should be predefined and appropriate for their intended application. First and foremost, acceptance criteria should meet the predefined needs of the study rather than simply reflecting the performance capabilities of the assay. The second factor is the nature of the assay methodology and the data generated using that assay, and third is the biological variability of the biomarker within and between the study populations. Below, we discuss key elements of biomarker method validation, contrasting assay-related issues unique to biomarkers with "small molecule" drug assays.

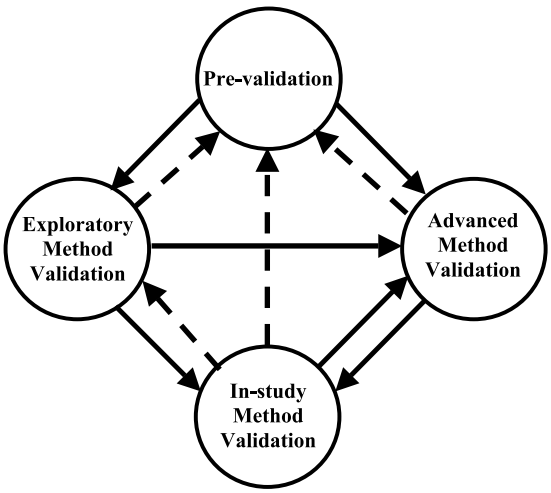

Fig. 1. Conceptual diagram of fit-for-purpose method validation. The method validation processes include four activity circles of prevalidation (preanalytical consideration and method development), exploratory method validation, in-study method validation, and advanced method validation. The processes are continuous and iterative, and driven by the intended purpose of the biomarker data. The solid arrows depict the normal flow of biomarker development (prevalidation), method validation (exploratory or advanced), and application (in-study method validation). The process could be moving the chosen biomarkers from mechanism exploration to pilot in-study and to advanced validation for confirmatory studies; or from exploratory validation to advanced validation due to a critical business decision change. The broken arrows represent scenarios that validation data do not satisfy the study purpose requirements, and backtrack processes for refinement or modification are required.

ISSUES AND RECOMMENDATIONS

Preanalytical Considerations

Biomarker Work Plan and Target Population

Study objectives should be defined in a biomarker work plan prior to the commencement of assay development to aid the timely identification of reagents, controls, and experimental samples. In early-phase clinical trials, preclinical studies and literature reviews can provide background information for a given population, allowing the establishment of appropriate precision requirements for the assay. This plan also defines the level of rigor to be applied to the assay validation and summarizes study objectives and the intended use of assay data. For example, an assay measuring concentrations of a bone resorption biomarker may be well characterized for osteoporotic or arthritic patients. For those populations, data on the inter- and intra-donor variability of the biomarker in those populations may exist, allowing one to predict minimally detectable changes in a biomarker (23). In certain populations, additional analytical measurements may be necessary, as changes in the initial biomarker may be affected by measurable morbidity- or therapy-induced changes to a sample.

Sample Collection

Results from biomarker assays are valid only if sample integrity is maintained from sample collection through analysis. Early, consistent application of predefined sample collection and handling techniques is especially important when such manipulations might affect sample and/or biomarker integrity. The life cycle of a study sample by necessity typically includes freeze/thaw cycles, so a definitive assessment of short-term, freeze/thaw, bench-top, and long-term stability is necessary (see Table II). A complete set of separate investigations should evaluate the most appropriate conditions for collecting and treating study samples to ensure that sample integrity is maintained. Patient-related factors including diurnal, disease-related, and behavioral effects (i.e., emotional state, posture, food intake, etc.) may necessitate alterations in sampling methods and careful interpretation of biomarker data. Provision of a detailed sample collection and storage protocol and adequate training of clinical trial site personnel are especially important when extraordinary measures are necessary to assure analyte integrity.

Biological Fluids. The collection of human and animal biological fluids can seem straightforward, but certain analytes and collection methods can necessitate extra attention in collection to assure that the variability arising from sample collection is minimized. Regardless of potential limitations encountered in the collection process, it is important to apply the collection procedure in a consistent manner for continuity of biomarker measurement.

(1) Type of needle, duration of the blood draw, type of collection tube container (see below), and the type and

Table II. Fit-for-Purpose Elements of Biomarker Assay Development and Method Validation

Parametersassay elements	Preanalytical and method development[a]	Exploratory method validation[b]	Advanced method validation[c]
Reagents and reference material	Consistent and accessible source—do due diligence)	Initial characterization Stablity initiated	Well characterized Inventoried Establish stability Establish change control
Target range	Estimate in biomarker work plan Define expectation of LLOQ and ULOQ	Acquiring data	Establish from incurred samples
Dynamic range (lower and upper quantitation limits)	Determine preliminary assay range with precision profile over target range	Use 3 validation runs	Use at least 6 runs (in-study validation data can be utilized) Establish LLOQ and ULOQ
Sensitivity	Define minimum detectable range Define requirements of sensitivity (LOD) and LLOQ	Estimate sensitivity Consider LOD *vs.* LLOQ	Establish sensitivity
Curve fitting	Choose appropriate calibration model fitting method and tools	Confirm choice of calibration model from 3 validation runs	Use 6 validation runs to confirm calibration model
Selectivity and specificity	Reagent specificity from supplier or literature Assess matrix effects and minimize if possible. Determine minimum required dilution (MRD) Sample and substitute matrix	Investigate likely sources of interference, including the therapeutic agent	Extensive testing of interference and risk recommendation Assessment of biomarker heterogeneity and isoforms
Parallelism	N/A	Use incurred samples, if available	Investigate in targeted population Determine maximum tolerable dilution
Dilution linearity	Determine if applicable, as defined in the biomarker plan (test range)	Use spiked samples	Use spiked samples and dilution VS if applicable
Precision and accuracy (analytical)	Establish expectations early on in biomarker work plan Consider heterogeneity	Use 3 validation runs	Use of total of at least 6 runs (in study validation data can be utilized)
Relative accuracy/recovery (biological)	Establish expectations early on in biomarker work plan	Use spiked incurred samples at multiple concentrations Addition recovery	Use multiple donors
Robustness (reagent and change control)	Determine need Consider availability of biological matrix	NA	Establish tolerability on crucial elements
Sample handling, collection, processing, and storage	Establish feasible conditions	Establish short-term and bench top stability Optimize conditions and effects on assay	Establish freeze/thaw and long-term sample stability
Documentation	Biomarker work plan Draft procedures Assess outsourcing options	Appropriate documentation to support the intended use of the data	Appropriate documentation to support the intended use of the data

In-study validation criteria are not listed in this table; they are defined through empirical testing of the Pre-, Exploratory, and Advanced Validation. Refer to section In-Study Validation and Sample Analysis Acceptance Criteria for elements of In-study Validation.
The recommendation is an example for typical immunoassay to obtain adequate statistical data. For assays with less variability, such as LC-MS/MS, less validation runs may be used.
[a] See sections Preanalytical Considerations and Method Development.
[b] See section Exploratory Method Validation.
[c] See section Advanced Method Validation.

concentration of anticoagulant may be very important determinants of biomarker stability. For example, some assays may be affected by the activation of endothelial cells and platelets, or other tissue damage from venipuncture. In these instances, it may be advisable to discard the first few milliliters of blood, to consider the use of an indwelling port rather than standard venipuncture, or to examine the use of stabilizing additives and anticoagulants. Cellular components from platelets and red blood cells may be generated by sample mishandling, and be problematic. Deliberate "stressing" of samples may reveal such artifacts, and provide vital sample collection tolerance information.

(2) Sample collection tubing and tubes, transfer pipettes, and storage containers should be evaluated to ensure that no significant adsorption of the biomarker to the contact surfaces occurs.

(3) Although serum is preferred over plasma for some analytical formats, some biomarkers such as those involved in the coagulation pathway in platelet activation or susceptible to proteolysis can only be accurately quantified in plasma. The initial coagulation of serum may not be suitable for temperature-labile analytes. In such cases, procedures may need to be developed for lower temperature coagulation. The effect of hemolysis should be evaluated for serum and plasma collection procedures.

(4) Differences in the handling of cells and whole blood can affect subsequent biomarker assays, particularly those that assess cellular functions. Cells can be inadvertently activated through incorrect handling of the blood thus complicating the quantification of cellular antigens, functional responses, and secreted products. In the latter case, the stability of both whole-blood samples and the serum or plasma samples under expected environmental conditions should be investigated. Moreover, the conditions of the *ex vivo* processing should be optimized.

Tissues. The above-mentioned principles often apply for tissue specimens, but the inherent heterogeneity of most tissue samples typically necessitates customized sampling techniques. As with biological fluids, it is important to devise standard protocols for tissue processing and storage to achieve uniformity for multiple sites over the period of the study (1). These protocols are often conducted by a certified medical professional, such as a pathologist or a histologist, with an established sampling protocol. For example, the sampling of live tissues for imaging or subsequent culture needs to be conducted under sterile conditions, and often in the presence of antibiotics, whereas tissues for histology studies typically require immediate fixation or freezing. Assays of tissue homogenates can benefit from appropriate normalization (e.g., to tissue mass, protein content, or the presence of a relevant tissue marker). These conditions must be characterized and defined prior to study sample collection to maintain the integrity of both the biomarker and the normalization factor. The identification of a representative and reproducibly collected section of the tissue to serve as positive and negative controls in assay validation and sample analysis is often crucial to the success of a tissue-based assay. For example, tissue samples from healthy normal subjects, and biopsies of patients at identified cancer stages can be used as negative control and various stages of positive controls; and maintained by a sample repository-coordinating agency.

Method Development

As illustrated in Fig. 1 and described briefly above, method validation is an iterative, "fit-for-purpose" process that does not end with the completion of exploratory validation. Rather, method validation is a process that requires continuous reassessment of data and optimization of the assay method. The extent and rigor of method development depends on whether the validation is exploratory or advanced. This section reviews the key concepts and recommended practices.

Method Feasibility

Method feasibility studies address the likelihood that an assay will be able to achieve its intended purpose. Numerous factors influence the answer to this question, including the availability of reagents of sufficient quality, and whether the performance characteristics of the assay are appropriate for the needs of the study. Accordingly, an early objective assessment of the assay working range is essential for initial assay development. For immunoassays and other ligand-binding methods, the most commonly used method for macromolecule biomarkers, the *precision profile*, is a useful tool that provides preliminary evidence if an assay is capable of measuring the "analyte of interest" at some predetermined concentration range. It is simply a plot of the coefficient of variation (CV) of the backcalculated calibrator concentrations *vs.* the concentration in log scale that serves two main purposes. First, during initial assay development and optimization, where only calibration curve data may be available, it affords preliminary estimates of the lower and upper quantification limits (QL). Second, the precision profile is a valuable tool in the decision to proceed to exploratory validation. Figure 2 in Appendix A contains an illustration of the use of a precision profile of a typical immunoassay. The illustration also demonstrates an important and often misunderstood concept for immunoassays that the apparently "linear" portion of the calibration curve does not always define the optimal working range for an assay.

Reagents, sample matrix (from patients or other target population), and substitute matrix for calibrators, if appropriate, should be acquired and evaluated for their initial behavior with regard to assay selectivity, linearity, and range of quantitation with calibrator matrix. Sample matrix often causes substantial interference with suppressed and variable analyte signals. Despite the likely adverse effect on the assay's lower limit of quantification (LLOQ), the common corrective strategy is to perform sample cleanup (usually not an option for macromolecules) or to dilute out the matrix interference until there is no longer a major effect on analyte recovery [the minimum required dilution (MRD)].

The concept of "specificity" refers to an assay's ability to unequivocally distinguish the "analyte of interest" from structurally similar substances. The degree to which unrelated matrix components cause analytical interference is a measure of assay selectivity. The presence of the "drug of interest" in a test sample can pose a challenge to both an assay's accuracy and selectivity, particularly when that drug targets directly interacts with the "biomarker of interest." For immunoassays, the drug binding to the biomarker can substantially alter the antibody–antigen reaction and influence detection. Assessment of the drug interference should be conducted to provide appropriate interpretation of the biomarker data. For example, a protein drug with a long half-life might affect the detection of the antidrug antibody results, which would require special interpretations (19).

If a commercial reagent or kit is used, it is acceptable, during feasibility studies, to use a manufacturer's quoted

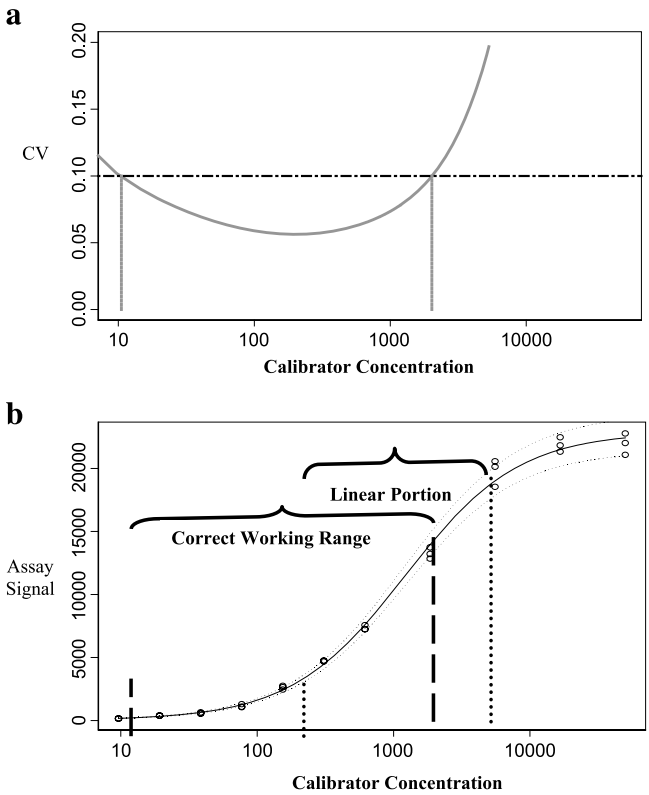

Fig. 2. (a) Precision profile. Data were from an immunoassay calibration curve. The dashed lines represent the "preliminary" working range based on a stringent threshold of 10% CV. As the CV based only on the calibration curve data typically underestimate the overall imprecision of the assay, a more stringent threshold than that of validation samples should be used. As discussed in the section Method Feasibility, the precision profile serves as a useful tool during the assay development and optimization, and as a screening tool before proceeding to the formal validation phase. (b) The calibration curve used for the precision profile in panel (a). The dashed lines represent the working range derived from the precision profile, and the dotted lines represent the linear portion of the curve. Note that the working range does not correspond to the linear portion of the curve. This is true for most assay formats such as immunoassays (see section Method Feasibility).

reagent stability and antibody specificity information. Generally, it is prudent to confirm such data with in-house experiments. Care needs to be taken to distinguish between the stability of packaged components and that of stored reconstituted lyophilized components. Generally, reagent expiry dates may be extended if stability has been proven in-house using the same acceptance criteria to monitor performance as those obtained in validation.

Early method feasibility studies that choose a commonly used analytical platform, such as chromatographic methods with mass spectroscopic detection, or immunoassay, allow flexibility and reliability in outsourcing biomarker assays. Fee-for-service analytical laboratories supporting these common analytical methods generally deliver reliable and expeditious, high capacity services. This is particularly relevant when contemplating large numbers of biomarker assays in late-stage clinical trials. For novel or exploratory biomarker assays or techniques, consideration of outsourcing early in assay development can help facilitate technology transfer.

Calibrators, Validation Samples, and Quality Controls

Calibrators. The selection and preparation of the calibration curve are central in the design, validation, and application of all quantitative assay methods. Calibration curves prepared from heterogeneous, impure, and/or poorly characterized analytes are more likely to yield greater uncertainty and higher variability. For low molecular weight, homogeneous analytes (i.e., <1,000 Da), it should be possible to construct an assay format that yields consistent data with a high degree of accuracy and precision. In such cases, the measurements can be as accurate and precise as seen with many definitive quantitative assays. Higher molecular weight biomarkers and those of unknown or ill-defined heterogeneity often necessitate analytical compromises, because assumptions must be made regarding assay accuracy. The assumptions underlying such relative quantitative assays can be tested in "parallelism" experiments (see section Dilution Linearity *vs.* Parallelism). Should these assumptions prove invalid, the method would be considered to be quasiquantitative. It is desirable to use a reference standard material from a single lot over the duration of the validation program. Alternatively, provision should be made for a "bridging" standard that can be used for comparisons across studies using different reference standard lots.

Unlike most drug/metabolite assays, biomarker assay matrices are often complicated by the presence of endogenous analyte. Endogenous analyte may be removed by a variety of procedures, such as charcoal stripping, high-temperature incubation, acid or alkaline hydrolysis, or affinity chromatography. Alternatives to matrix processing indclude the use of "surrogate" protein-containing buffers (which offer better stability, convenience, and long-term consistency) or a heterologous matrix (e.g., another species) that lacks the analyte or contains a less reactive homolog. These strategies result in a substitute calibrator matrix that is different from the test sample matrix. Nonetheless, the approach is suitable as long as the agreement in concentration–response relationship between the calibrators and test sample analyte is acceptable. However, for multiplexed biomarker assays, removal of endogenous matrix analytes or selection of a synthetic buffered substitute that is suitable for all analytes may be impractical or impossible; thus nonlinearity and bias could be inevitable for some of the analytes (24).

Validation Samples and Quality Controls. It is useful to distinguish validation samples (VS), used in assay validation experiments to estimate intra- and inter-run accuracy/precision and stability, from quality control (QC) samples that are used during study sample analysis to judge the acceptability of assay runs. VS from prestudy validation are usually suitable for subsequent use as QC samples during in-study validation. VS and QC are used to assess the ability of the assay to measure the biomarker of interest for its intended use, allowing one to distinguish assay variability from inherent differences between samples. Thus, VS and QC should be as closely related to the study samples as possible. The availability of rare matrices and the presence of multiple analytes at various concentrations may require the use of substitute reagents including cell lines, or tissues from related species, necessitating additional validation.

If alternative or less optimal matrices and controls allow the measurement of a biomarker with adequate precision and

accuracy for the intended use of the data, clinical samples may be analyzed with QC samples and calibrators made up in the substitute matrix. Ultimately, retrospective validation of that matrix may be possible if sufficient amounts of the clinical specimens remain following sample testing. One approach uses VS/QC spiked with known amounts of the biomarker and prepared after pooling clinical test samples containing low concentrations of a biomarker. Basic assay performance data can thus be generated throughout exploratory and advanced validation. Cross-validation between the substitute and rare matrices is thus possible because the endogenous biomarker concentration can be subtracted from all results to yield amore accurate assessment of biomarker concentrations.

Another approach, similar to that discussed above, is to screen naive or incurred study specimens with a preliminary assay to identify those with low or undetectable analyte concentrations. Such samples are then pooled and supplemented with varying amounts of reference material to create VS and QC samples across the anticipated range of quantitation. Depending on its endogenous level, the biomarker concentration may be ignored or added to the spiked amounts to give the nominal VS and QC sample concentrations.

For quasiquantitative or qualitative assays, positive and negative controls from healthy normal subjects and patients should be set up as VS/QCs with the expected responses (or scores) and the toleration limits.

Calibration Curve Model Selection

During the method development phase, an appropriate "working regression model" for the calibration curve should be chosen to optimize the method protocol and to derive the preliminary performance characteristics of the method. The regression model can be chosen from calibration data if VS are not available. The optimization of the method protocol and the preliminary assessment of the method performance should be based on this working model of the calibration curve. This model should be confirmed with data from VS and routinely applied during the sample analysis, unless there is a major change in the assay protocol or reagents.

An inappropriate choice of the statistical model for the calibration curve can contribute to the total variability in biomarker determination and limit the useful assay quantitation range. Owing to the wide variety of assay formats and analytes in biomarker applications, the choice of a curve-fitting model for calibration curves should be tailored to each analytical method. Some of the commonly used models include a variety of polynomial (linear, quadratic, etc.) and nonlinear models (four or five parameter logistic model, power model, etc.). The practical importance of using the right model is discussed and illustrated with an example in Appendix B.

Weighting for Calibration Curves

The default position that the standard deviation (SD) of the assay response is constant for the entire range of the calibration curve, i.e., every data point is weighted equally, is usually not valid for ligand-binding assays. Curve fitting without appropriate weighting for unequal SDs can produce suboptimal results and incorrect assessment of the true assay performance [for example, inferior sensitivity and limited assay range in sandwich enzyme-linked immunosorbent assay (ELISA)]. Thus, a curve-fitting method that assigns weights to the data proportionate to the amount of variability (SD) should be considered for calibration curves. It should be noted that the weighting is done to reflect and adjust for unequal SDs, not unequal CVs. Appendix C illustrates the importance of weighting in the bias, precision, and quantification limits of the analytical method.

Outliers

Outliers may negatively affect the quality of a calibration curve and so must be either removed or accommodated using outlier-resistant statistical methods. Setting *a priori* criteria for outlier identification and removal is an acceptable, although possibly problematic practice (20). Given the limited degree of replication and the sample sizes in validation experiments, outlier tests lack power to detect all but extreme values and are somewhat arbitrary in the criteria used to declare a point as an outlier. A less subjective and more consistent approach is to use outlier-resistant statistical techniques. For example, weighting factors may be estimated with triplicate calibrators, and the curve fit using median values for each triplicate set. In cases where only duplicates are available, an alternative is to collaborate with a statistician to use Tukey's biweight function. In the evaluation of reported results (as opposed to the context of curve fitting, which can be considered part of the assay method), it becomes important not only to remove the influence of outliers, but also to record their presence. Apparent outliers may provide valuable insight into anomalous model or method performance, which has significant analytical implications. Evidence for a prozone phenomenon (hook effect) in an immunoassay is a relatively commonplace example. Thus, although it is desirable that summary statistics (e.g., medians and median absolute deviations) be uninfluenced by outliers and reflect what is happening to *most* of the data, it is also necessary to augment such summary statistics with graphical analyses or tabulation that facilitates recognition of possible outliers.

Exploratory Method Validation

General Comments

Exploratory Validation (Table II) quantitatively characterizes the assay method with respect to its basic analytical elements. Such procedures need to be geared to meet the common challenge of method validation for numerous and diverse biomarkers using limited resources.

A biomarker plan is highly recommended to guide the investigation of method performance. The plan should include the requirements of the key elements of the study and the level of documentation of data. At least three evaluation runs should be carried out to provide the basic assay performance including accuracy, precision, sensitivity, parallelism (using incurred samples if possible), relative selectivity (through investigation of likely sources of interference); initial biomarker concentration range in normal

individuals and in the target population, assay dynamic range, short-term biomarker stability in the expected biological matrix, and dilutional linearity using spiked samples. The data should be statistically evaluated to estimate if the method would meet study requirements with respect to these parameters. Such initial characterization should constitute sufficient method validation to support exploratory study sample testing.

Assay Dynamic Range and Limits of Quantitation/Detection

Assay dynamic range, also known as the "reportable range," extends from the lower to upper limits of quantification (LLOQ to ULOQ). Within these limits, the analyte is measurable with acceptable levels of accuracy, precision, and total error (see section Advanced Method Validation, Accuracy and Precision for details and illustration). For ligand-binding assays, sample dilution is often validated to extend the assay range at the higher end for test samples. The utility of a biomarker assay may be evaluated *via* an initial assessment of baseline samples. This initial assessment may provide information as to whether an assay will have sufficient sensitivity and *dynamic range* to cover the target range of the potential (inhibitory or stimulatory) impact of a drug. In addition, assessments of biomarker levels in the presence of high concentrations of drug may provide additional guidance on suitable assay ranges for PD markers. A common practice is to generate values from individual humans or animals with and without the targeted disease state, whenever possible.

Biomarkers have been used in the clinical diagnostic arena for many years, where the diagnostic kits/reagents are often "borrowed" for clinical study analyses. These include "for research only" kits, FDA-approved tests, and other commercial reagents. Because the intended use of a biomarker in drug development usually differs from its diagnostic utility, each laboratory must define the intended biomarker application and carry out the assay validation accordingly. Typically, initial assay performance is assessed using the manufacturers' recommendations to verify the kit specifications. However, caution should be exercised when assay sensitivity is evaluated, because the term "limit of detection" (LOD) is often incorrectly used to describe assay "sensitivity" (see Table I). LOD is the concentration resulting in a signal that is significantly higher than that of background (usually mean signal at background + 2 or 3 SD), whereas LLOQ is often the lowest concentration that has been demonstrated to be measurable with acceptable levels of bias, precision, and total error (20).

The use of the calibration curve beyond the lowest quantifiable limit to report the LOD should be undertaken with caution. It is a common practice in pharmacokinetic (PK) analyses to assign a zero or "below quantifiable limits" value to results lower than the LLOQ. For biomarkers, concentrations below LLOQ but above LOD may be useable for quantitative estimates (e.g., % of control). Even though the variability of these values is high, reporting of the numeric results can be justified because it provides actual estimates rather than the assignment of "no result," <LLOQ, or zero (1). This can have a major impact on the study conclusions if a study group yields many values near the lower end of the curve. It is also worth noting that the actual study sample LLOQ is often different from that of the lowest calibration curve point, because calibrators may not exhibit parallelism or acceptable analytical precision as the LLOQ concentration (see section Dilution Linearity *vs*. Parallelism).

Accuracy and Precision

Even though it may not be possible to establish absolute accuracy for a biomarker, relative accuracy data can be informative. An *addition recovery* experiment is used for accuracy evaluation over the anticipated analyte concentration range (target range). This test should be conducted in the biological matrix from a few donors or distinct pools at several concentrations over the target range. As most biomarkers are present in a given biological matrix, the amount of analyte to be added to the sample should be high enough to minimize the contributory effect from the endogenous component. It is critical to include multiple measurements on unmodified samples in these evaluations.

In general, evaluation of the precision of a biomarker assay will provide information on the statistical significance of biomarker results in a study. Similar to drug/metabolite bioanalysis, precision is normally evaluated as intermediate precision (see Glossary) and intraassay variation (repeatability) with at least three levels of VS, with divergent concentrations of analyte, over at least three analytical runs. The optimum number of replicates required per level of VS may vary but, in general, will be higher than suggested for drug/metabolite validations. In many cases, three reportable results per VS level are adequate and three assay runs are a starting point; however, additional assay runs will likely be needed to draw statistically sound conclusions. Between-run variance usually exceeds the within-run variance in ligand-binding assays. It is thus usually preferable to increase the number of runs as opposed to increasing replication within a run to statistically improve the estimates of interassay precision.

Dilution Linearity vs. Parallelism

The absence of suitable blank matrices means that many biomarker immunoassays use calibrators prepared in a substitute matrix that differs from the test sample matrix. *Parallelism* documents that the concentration–response relationship of the analyte in the sample matrix from the study population is sufficiently similar to that in the substitute matrix. Parallelism between dilution curves, where dilution of test samples in the range of the calibration curve does not result in significantly different extrapolated analyte concentrations, validates the use of the substitute matrix for calibrator preparation. Results of these experiments may also define suitable dilution ranges should dilution be necessary to alleviate matrix effects.

A substantial departure from parallelism would invalidate the use of reference standards and would indicate the need for a quasiquantitative interpretation, for which uncalibrated assay signal is reported. It is important to emphasize that method validity can be achieved in the face of a tolerable degree of nonparallelism, as opposed to a complete lack of nonparallelism. Experimental approaches are pre-

sented in Appendix D. Figure 6a and b shows cases of proven and failed parallelism plotted on two commonly used charts.

When available samples contain too little analyte to perform parallelism assessments, method selectivity may be assessed using spiked recovery and dilution linearity of spiked samples. For PK assays, dilutional linearity demonstrates accuracy for samples with concentrations above the ULOQ after dilution into the assay range, and further demonstrates the lack of a high dose hook effect. If the reference standard is reflective of the endogenous biomarker and the test conducted with multiple dilutions including samples from multiple donors, dilutional linearity supports parallelism.

Stability Testing—Short Term and Benchtop

Biomarker storage stability can be a complex issue owing to the difficulties in defining biomarker stability under storage conditions and in judging the adequacy of the assay method to monitor stability changes. A case-by-case evaluation must be made to address whether chromatographic, immunoreactive, or biological activity assays are most appropriate to monitor stability of a given biomarker or biomarker family. One should beware that, unlike small molecules, stability measures of endogenous macromolecules can be method-dependent. For example, aggregation resulting from freezing samples might change the binding characteristics for an immunoassay to show instability, whereas an LC-MS or HPLC assay might not render the same result. For immunoassays, antibodies that recognize epitopes distal to the site of the biological activity could also lead to misleading stability data. Stability data should be interpreted with the understanding of the method in relevance to the biology of the biomarker. Besides degradation, reactions producing the analyte can occur to artificially increase the analyte concentration in the biological matrix. Therefore, careful consideration should be given to optimal, practical sample storage conditions to minimize these possibilities.

Some commonly used conditions to evaluate biomarker short-term and benchtop stability include: whole blood at ambient temperature for most flow cytometric assays, cell medium or human peripheral blood mononuclear cells for intracellular proteins, and serum/plasma samples for most extracellular proteins at ambient, refrigerator, or ice-water bath, and −20°C (as many clinical sites do not have −80°C freezers). If a biomarker is unstable in a processed matrix after 24 h storage, it will be of limited value to support drug development.

Analogous to the approach to evaluation of sample collection integrity (see section Sample Collection), analyte short-term storage stability should be evaluated under conditions mimicking those expected at a typical clinical site and in the intended matrix from the study population. The use of incompletely characterized reference standards to prepare VS/QC samples for stability evaluation may not reflect that of the clinical samples. Incurred samples or healthy donor samples of sufficiently high biomarker concentrations from multiple donors could be used in the stability test.

For an exploratory validation, it is advised that samples from at least three individual sources be evaluated for short-term storage. Pools are best avoided where possible unless required to obtain sufficient volumes. Furthermore, it may be appropriate during an exploratory validation to cite literature for stability information for certain collection conditions.

Advanced Method Validation

General Comments

Because method validation is an iterative, evolving process (Fig. 1), all of the performance characteristics listed in the exploratory validation should also be included in the advanced validation, with additional characterization as described in the following sections. The increased rigor of advanced validation is undertaken in a scaled, fit-for-purpose approach as the impact of the biomarker data on decisions around critical safety, efficacy, pharmacodynamic, differentiation, or surrogate information increases. Methods undergoing advanced validation may have been validated and used for sample testing. Alternatively, advanced validation may be undertaken as the initial phase of formal performance characterization for the method. In the former case, in-study validation (QC sample recovery) represents an ideal source of data because it reflects accuracy/precision performance from actual assay use. The advanced validation biomarker work plan for such a method would include experiments to supplement, rather than replace, existing validation data. If the existing validation data do not fully meet the work plan requirements, more experiments in the prevalidation should be conducted as depicted in the broken arrows in Fig. 1. The biomarker work plan for advanced validation should consider the level of document control as discussed in the conference report (1).

Selectivity and Matrix Effect

During the advanced validation, more rigorous testing of potential interfering endogenous components should be implemented as later-stage clinical trials typically include more diverse populations (including patients) with less control of diet and sample collection, and more concomitant medications. In addition, hemolysis, lipidemia, elevated bilirubin, and other matrix environments would present more commonly. Methods should not be considered suitable for the analysis of samples containing elevated levels of these substances until proven by matrix interference experiments. The selectivity of the assay in the presence of endogenous substances can be evaluated by using blank matrices from multiple individuals (and both genders). In addition, a known concentration of the biomarker can be spiked into the various individual lots of blank matrix to assess the recovery.

Accuracy and Precision

The advanced validation entails a sufficient number of validation batches (such as six independent runs) to provide increasingly rigorous statistical data for the confidence of the assay performance. The calibration curves should include a minimum of six nonzero concentration calibrators, and the VS should span at least five independently (if possible) prepared concentrations in duplicate or more. Two of the VS

levels should be at or near the target LLOQ, one near the middle of the range, and two at or near the target ULOQ. The VS levels (except at the LLOQ and ULOQ) should not be identical to the calibrator concentrations, and should not be used as part of the calibration curve.

Analysis of the prestudy validation data described above can be used to estimate the bias, intermediate precision, and total error (20). Using these results, the sensitivity and working range of the assay can be established. The LLOQ is defined as the lowest concentration at which the total error is within a prespecified threshold (e.g., 30%). The ULOQ is defined similarly.

Figure 4 shows the Total Error profile (solid line) and the absolute value of the percent bias (dashed line) for VS in a validation experiment of a biomarker assay. As the two lowest concentration samples have total error greater than the 30% threshold used as the criteria in this example, the LLOQ is established at the next highest VS concentration tested (~14 pg/mL). In this experiment, the ULOQ is set at the highest VS tested concentration tested because all the high samples are within 30% total error.

Using these validation data, the model for the calibration curve and the weighting method may be finalized and confirmed. Various mathematical models may be fit to the calibration curve data from each run (e.g., linear, four-parameter logistic, and five-parameter logistic). Then, the total error, bias, and precision may be estimated with respect to each of these models. The model that results in the lowest total error throughout the entire range of the curve is then chosen for the production (in-study) phase of the assay. This is illustrated in Fig. 4, section Calibration Curve Model Selection, and Appendix B (calibration curve model selection). Similarly, the optimal weighting method can also be chosen based on these data.

If test/study samples need to be diluted to accommodate limits of the method range, then dilutional linearity should be established within the context of parallelism assessment as described in section Assay Dynamic Range and Limits of Quantitation/Detection, and the maximum tolerable dilution should be determined based on the validation sample data. Dilutional linearity can be assessed with respect to the prespecified limit on the CV of the difference between the dilution-adjusted results and the expected results. For example, if the specified limit on the CV is 20%, then the maximum tolerable dilution is the highest dilution where the CV is within 20%.

Where VS are in the same matrix as test samples, and there is endogenous biomarker present in the matrix used, then actual values need to take the endogenous concentration into account. This will help assign "nominal" target values for the VS and QC samples in prestudy and in-study validation. The endogenous level can be estimated via direct interpolation from the calibration curve, and if that is not possible, it can be estimated via the "standard additions" method. Here, linear regression (in log scale) of observed results (y) *vs.* expected results (x) of VS/QC samples is recommended. The negative X-intercept approximately quantifies the endogenous component and when multiplied by -1 and added to the expected result value will yield the "estimated" nominal target value to be used when calculating bias.

Parallelism

In addition to the basic considerations described in Dilution Linearity *vs.* Parallelism, the advanced validation should ensure that the issue of parallelism has been explored in the targeted population, and that if a substantial deviation from ideal behavior exists, the implications of the significant nonparallelism for the reliable estimation of analyte concentrations are understood. One aspect of the interpretation of such results will be the relationship of the reference standards used to prepare VS and calibrators to the endogenous analyte, and indeed (when appropriate) the variability of reference standards from commercial kit to commercial kit from different vendors, and between lots of commercial kits from the same vendor.

In the case that a calibration curve has a different curve shape than that in study matrix, decisions should be made on what the assay reporting ranges will be, whether a different type of matrix should be used for the calibration curve, or whether dilutions to the study samples will alleviate this concern. It may be prudent in the longer term to expend considerable time and effort to obtain a matrix as close as possible to that of the study samples. The key issue here is to ensure the reliability of an assay and to seek consistency in the concentration–response behavior of the reference standard from lot to lot. It must be recognized that in some cases it may not be possible to assess parallelism until incurred (or other) samples are available that contain high enough concentrations of the biomarker of interest. In these cases, other experiments may be used to give confidence that interferences from the matrix of interest are not present. However, these do not investigate the tolerability of nonparallelism, and therefore the specific experiments outlined in Dilution Linearity *vs.* Parallelism and Appendix D should still be conducted retrospectively. If nonparallelism exists, its extent, tolerability, and effect on data interpretation should be considered.

Stability Testing—Freeze/Thaw and Long Term

Advanced validation should include assessments of freeze/thaw and long-term stability sufficient to cover the range of conditions to which samples are likely expected to be exposed. Practical limitations to the assessment of long-term stability of incurred sample storage may occur, particularly for extended longitudinal studies in which specific biomarker measures are not identified at study initiation. Interpretation of biomarker modulation from extended in-life studies should consider any uncertainties in long-term stability assessments.

Validation for Reagent or Change Control

As a drug development program matures, consistency in the performance of a biomarker method over a long period is an important consideration. Data commenting on such consistency, which is an aspect of the robustness (see Glossary) of the assay, may be generated over time by the use and evaluation of VS pool results, or by a series of experiments designed to test the response of the method to

anticipate perturbations of the assay environment. Furthermore, additional validation and systematic comparison of method revisions may be necessary to ensure the reliability of biomarker measurements. The need for such additional validation work (e.g., qualification, cross-validation, partial validation, or revalidation, depending on circumstances) may arise as a result of a change in critical assay reagent(s) (e.g., antibody lot or kit vendor), assay laboratory (e.g., transfer from a pharmaceutical company to contract research organization), process instrumentation (e.g., plate reader or robotic pipette upgrade), or methodology (e.g., plate-based ELISA to point-of-care devices). For successful biomarker applications, the laboratory should consider having defined procedures (e.g., SOPs) for qualification/requalification of key analytical reagents.

In-Study Validation and Sample Analysis Acceptance Criteria

The in-study validation phase seeks to ensure that the assay continues to perform as per predefined specifications in each study run (i.e., to ensure the assay remains "in control"). This entails the use of QC samples, typically at three levels—at low, mid, and high concentration of the analyte with at least two replicates at each level.

Ideally, the QC samples used in the in-study sample analysis phase should be prepared identically to the VS used in the prestudy validation phase, although this is not an absolute necessity. Where this is possible, it avoids the need for reassessment (e.g., lot-to-lot changes) and assignment of "nominal" target values for the purpose of calculating bias. Such continuity of samples for the purpose of assay control ensures that assay performance has not changed from the prestudy validation phase.

The similarity of the reference standard to the analyte in study samples should be investigated as described in Dilution Linearity *vs*. Parallelism if such samples were not evaluated during the prestudy validation phase. Stability of the biomarker may also need to be assessed at this stage if samples with high enough concentrations of the biomarker of interest have not been available. If possible, incurred samples should be used to perform stability testing instead of spiked samples.

It is important that the approaches for assessment of method performance are suitable for the intended purpose. QC methods used in PK assays [e.g., the 4–6–*X* scheme; (21), see Glossary)] and clinical diagnostics (control charts with confidence limits) may both be applicable. The laboratory performing the analysis should choose the most relevant method to use and justify it based on relevant scientific and statistical grounds and on clinical need. These judgments will be pivotal in order to assign an appropriate value to *X* (the total error, or bias + precision) in the 4–6–*X* scheme.

It should be emphasized that the acceptance criteria for biomarker assays will depend heavily on the intended use of the assay and should be based on physiological variability as well. The appropriate value of *X* in 4–6–*X* can be determined based on the variability of the total error estimates in prestudy validation. When it is feasible to use more QC samples in each run, 8–12–*X* or 10–15–*X* will have much better statistical properties than the 4–6–*X* criterion. Alternatively, use of control charts (25) or tolerance limits (26,27) provides better control of relevant error rates, and thus may be preferable to the approach of fixed criteria.

As indicated above, an important consideration for defining the performance criteria of most biomarker methods is the physiological/biological variability in the study population of interest and other variables such as diurnal effect, operator, instrument, etc. (28). That is, to determine whether a biomarker method is fit-for-purpose, we should determine whether it is capable of distinguishing changes that are statistically significant based on the intra- and intersubject variation. For example, an assay with 40% total error determined during validation may be adequate for statistically detecting a desired treatment effect in a clinical trial for a certain acceptable sample size, but this same assay may not be suitable for a clinical trial involving a different study population that has much greater physiological variability. Thus whenever possible, physiological variation should be considered when evaluating the suitability of biomarker methods for specific applications. When the relevant physiological data (e.g., treated patients of interest) are not available during the assay validation phase, then healthy donor samples should be used to estimate the intra- and inter-subject variation. If healthy donor samples are not available, then other biological rationale should be considered and periodically updated as more information become available. In the absence of physiological data or other biological rationale, only the assay performance characteristics determined from validation experiment such as the bias, precision, and total error should be reported. The sensitivity and dynamic range of the assay can be defined based on a "working criteria" of say, 20 or 30%, on bias, precision and/or total error. However, any decision regarding the suitability of the assay should be based on the availability of adequate information related to the physiological data.

Setting a definitive acceptance criteria on the desired analytical precision and total error *a priori* may not be appropriate (or even possible) when taking into account all possible outcomes in the analytical phase—especially as the values seen in the incurred samples may not be what is expected or predicted. This is especially the case for new or novel biomarkers as opposed to those where historical information in normal and diseased populations is available. However, the "working criteria" can be used as *a priori* criteria to track assay performance.

CONCLUSIONS

Biomarker data can be extremely valuable as early predictors of drug effects and can yield important efficacy and safety information regarding the dose–response relationship. Thus, biomarkers are potentially useful for successful and efficient drug development. The intended diverse arrays of biomarker applications present an analytical challenge when one attempts to adopt regulatory guidelines for either PK assays or diagnostics development. This paper is the result of intense and ongoing discussions by the authors following the AAPS 2003 Biomarker Workshop (1). Here we propose the conceptual strategy for a fit-for-purpose approach

for biomarker method development and validation (depicted in Fig. 1) with four activities: prevalidation (preanalytical consideration and method development), exploratory validation, in-study validation, and advanced validation. The recommended processes of these activities are summarized in Table II. A biomarker work plan should be prepared to define the study purpose and requirements. The recommended basic approaches are conceived and designed to avoid major pitfalls without stifling research efficiency. The key elements for fundamental validation include sample stability from the time of collection, preparation of calibrators, VS and QCs, setting target and dynamic ranges with appropriate calibration curve-fitting, selectivity, precision, and accuracy. The process from exploratory to advanced validation is continuous and iterative with increasing rigor for all the validation elements, and with additional needs focused on method robustness, cross-validation, and documentation control. Critical differences between biomarker assays and those of drug bioanalysis and diagnostics have been discussed to provide clarification to readers more familiar with the latter disciplines. We hope that this position paper will stimulate more discussion and foster consensus building on best practices in the relatively young field of biomarker development.

Acknowledgments

The authors are grateful to Ronald R. Bowsher, PhD (Linco Diagnostic Services), for providing encouragement and critical review on the manuscript. We also thank Wesley Tanaka, PhD, and Omar Laterza, PhD (Merck & Co.) for providing input and suggestions.

Glossary

The following definitions are meant to be valid in the context of bioanalytical methods. Not all definitions will be consistent with terminology from other disciplines.

1. Accuracy: Per the FDA Guidance on Bioanalytical Method Validation (May, 2001), Accuracy of an analytical method describes the closeness of mean test results obtained by the method to the true value (concentration) of the analyte. This is sometimes referred to as Trueness or Bias.

2. Advanced Validation: A method validation that requires more rigor and thorough investigation, both in validation tasks and documentation, to support pivotal studies or critical decisions; e.g., differentiating subtle graded drug effects, monitoring drug safety, or for submission to regulatory agencies for drug approval.

3. Biomarker: A characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacologic response to a therapeutic intervention.

4. Clinical Endpoint: A characteristic or variable that reflects how a patient feels, functions, or survives.

5. Clinical Qualification: The evidentiary and statistical process linking biologic, pathologic and clinical endpoints to the drug effect, or linking a biomarker to biologic and/or clinical endpoints.

6. Definitive Quantitative Assay: An assay with well-characterized reference standards, which represents the endogenous biomarker, and uses a response–concentration standardization function to calculate the absolute quantitative values for unknown samples.

7. Dilution(al) Linearity: A test to demonstrate that the analyte of interest, when present in concentrations above the range of quantification, can be diluted to bring the analyte concentrations into the validated range for analysis by the method. Samples used for this test are, in general, the ones containing high concentrations of spiked analyte, not endogenous analyte.

8. Dynamic Range: The range of the assay that is demonstrated from the prestudy validation experiments to be reliable for quantifying the analyte levels with acceptable levels of bias, precision, and total error.

9. Exploratory Validation: Method validation that is less rigorous but adequate to meet study needs; e.g., looking for big effects in drug candidate screen, mechanism exploration, or internal decision with relatively minor impact to the final product, and not used for submission to regulatory agencies.

10. Interference: (1) Analytical interference: presence of entities in samples that causes a difference in the measured concentration from the true value. (2) Physicochemical interference (matrix interference): A change in measured physical chemical property of the specimen (e.g., excess bilirubin or hemoglobin, ionic strength, and pH) that causes a difference between the population mean and an accepted reference value.

11. Intermediate Precision: Closeness of agreement of results measured under changed operating conditions within a laboratory; e.g., different runs, analysts, equipments, or plates, etc. This is one of the three types of Precision.

12. Limit of Detection: A concentration resulting in a signal that is significantly different (higher or lower) from that of background. Limit of detection is commonly calculated from mean signal at background ± 2 or 3 standard deviations. This is often described as the analytical "sensitivity" of the assay in a diagnostic kit.

13. Limit of Quantification: Highest and lowest concentrations of analyte that have been demonstrated to be measurable with acceptable levels of bias, precision, and total error. The highest concentration is termed the Upper Limit of Quantification, and the lowest concentration is termed the Lower Limit of Quantification.

14. Minimum Required Dilution: The minimum dilution required to dilute out matrix interference in the sample for acceptable analyte recovery.

15. Parallelism: Relative accuracy from recovery tests on the biological matrix, incurred study samples, or diluted matrix against the calibrator calibrators in a substitute matrix. It is commonly assessed with multiple dilutions of actual study samples or samples that represent the same matrix and analyte combination of the study samples.

16. Pharmacodynamic: The relationship between drug concentrations and biochemical and physiological effects of drugs and mechanisms of drug action.

17. Precision: Precision is a quantitative measure (usually expressed as standard deviation and coefficient of variation) of the random variation between a series of measurements from multiple sampling of the same homogenous sample under the prescribed conditions. If it is not possible to obtain

a homogenous sample, it may be investigated using artificially prepared samples or a sample solution. Precision may be considered at three levels: 1. Repeatability, 2. Intermediate Precision, and 3. Reproducibility.

18. Precision Profile: A plot of the coefficient of variation of the calibrated concentration *vs.* the concentration in log scale. It provides preliminary estimates of the quantification limits and feasibility assessments on the intended range.

19. Quality Controls: A set of stable pools of analyte, prepared in the intended biological matrix with concentrations that span the range claimed for the test method, used in each sample assay run to monitor assay performance for batch acceptance.

20. Qualitative Assay: The assay readout does not have a continuous proportionality relationship to the amount of analyte in a sample; the data is categorical in nature. Data may be nominal (positive or negative) such as presence or absence of a gene or gene product. Alternatively, data might be ordinal, with discrete scoring scales (1 to 5, −+, +++, etc.), such as immunohistochemistry assays.

21. Quasiquantitative Assay: (Quasi: "possesses certain attributes") A method that has no calibrator, has a continuous response, and the analytical result is expressed in terms of a characteristic of the test sample. An example would be an antidrug antibody assay that is express as titer or % bound.

22. Relative Quantitative Assay: A method which uses calibrators with a response–concentration calibration function to calculate the values for unknown samples. The quantification is considered relative because the reference standard is either not well characterized, not available in a pure form, or is not fully representative of the endogenous biomarker.

23. Relative Accuracy: For relative quantitative methods, absolute accuracy is not possible to evaluate due to the unknown nature of the endogenous biomarker. Relative accuracy is the recovery (see below) of the reference standard spiked into the study matrix.

24. Recovery: The quantified closeness of an observed result to its theoretical true value, expressed as a percent of the nominal (theoretical) concentration. Recovery is often used as a measure of accuracy.

25. Repeatability: Closeness of agreement between results of successive measurements of the same samples carried out in the same laboratory under the same operating condition within short intervals of time. It is also termed intraassay or intrabatch precision. This is one of the three types of Precision.

26. Reproducibility: Closeness of agreement of results measured under significantly changed conditions; e.g., inter laboratory, alternate vendor of a critical reagent. This is also referred to as cross validation.

27. Robustness of the assay: A measure of the capacity of a method to remain unaffected by small, but deliberate changes in method parameters and provides an indication of its reliability during normal run conditions.

28. Selectivity: The ability of a method to determine the analyte unequivocally in the presence of components that may be expected to be present in the sample.

29. Sensitivity: The lowest concentration of analyte that an analytical method can reliably differentiate from background (limit of detection).

30. Specificity: The ability of assay reagents (e.g., antibody) to distinguish between the analyte, to which the reagents are intended to detect, and other components.

31. Surrogate Endpoint: A biomarker that is intended to substitute for a clinical endpoint. A surrogate endpoint is expected to predict clinical benefit or harm (or lack of benefit or harm) based on epidemiologic, therapeutic, pathophysiologic, or other scientific evidence.

32. Target Range: Range of analyte concentrations where the study samples are expected to fall.

33. Total Error: The sum of all systematic bias and variance components that affect a result; i.e., the sum of the absolute value of the Bias and Intermediate Precision. This reflects the closeness of the test results obtained by the analytical method to the true value (concentration) of the analyte.

34. Validation: It is the confirmation via extensive laboratory investigations that the performance characteristics of an assay are suitable and reliable for its intended analytical use. It describes in mathematical and quantifiable terms the performance characteristics of an assay.

35. Validation samples: Test samples in biological matrix mimicking study samples, endogenous and/or spiked, used in prestudy validation to provide characterizations of assay performances; e.g., intra- and inter-run accuracy and precision, and analyte stability.

APPENDIX A

Precision profiles of calibration curves

The Precision Profile is a plot of the coefficient of variation of the calibrated concentration *vs.* the true concentration in log scale. The standard error of the calibrated concentration should include the variability of both the assay response and the calibration curve. Complex computation programs can be implemented with the help of a statistician (29) or simple calculations can be performed in an Excel spreadsheet.

A Precision Profile of an ELISA standard curve based on a four-parameter logistic (4PL) model with weighting is shown in Fig. 2a. In this example, the quantitative limits at 10 and 2,000 pg/mL do not approximate the bounds of "linear" portion of the calibration curve that is typically symmetric around EC_{50}. This is attributable to the relatively lower variability (standard deviation) of the low response values and the relatively higher variability of the higher response values. Thus the quantitative range of the curve/assay may not be at the apparently "linear" region of the curve/assay (see Fig. 2b). This is usually true for immunoassays and most other assay formats where the response error variability is not constant across the entire range of the response.

APPENDIX B

Calibration curve model selection

Using data from a validation experiment, we describe a computation procedure for selecting the appropriate calibration curve model. Some of the calculations below, such as bias, precision and total error, can be easily implemented in

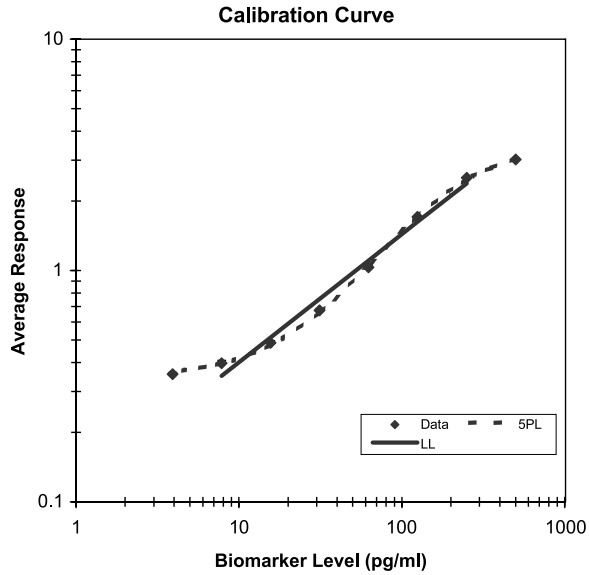

Fig. 3. Calibration model selection based on % bias and intermediate precision. The solid line represents the linear model in the log scale of both the concentration and response (LL) that is fit to the linear part of the calibrator range (i.e., 6 out of 8 points were used). The dashed line is the weighted five-parameter logistic model (5PL) that is fit to all the data.

Microsoft Excel using available formulae (20), in consultation with a statistician.

(1) Fit the calibration curve data from each run using various statistical models and use these models to calibrate the concentrations of the VS (Fig. 3).

(2) For each VS level, compute the Total Error = Absolute value of Bias + Intermediate Precision from the calibrated results, corresponding to each calibration curve model, as described in DeSilva *et al.* (20).

(3) Plot the "Total Error profile" of total error *vs.* the validation sample concentrations, for each calibration curve model (Fig. 4).

(4) Select the calibration curve model that has the lowest total error overall across the concentration range of interest.

Total Error profiles from the qualification runs should be evaluated as illustrated in the ELISA example in Figs. 3 and 4. The two models considered here for the calibration curve are the linear model in the log scale (LL) and the weighted five-parameter logistic model (5PL). These Total Error profiles are derived from four validation runs. Each run contained eight calibrator standards in triplicate from 500 to 4 pg/mL along with eight VS levels in six replicates. The linear model was fit to only the linear part of the calibrator range (6 out of 8 points were used), resulting in an R^2 of 98.5%. As the clinical test samples from this assay were expected to fall within the range of 14–450 pg/mL, the Total Error profiles were compared for the two calibration curve models within this range. It is evident that the 5PL model is more appropriate, mostly attributable to improved Bias in the lower and higher ends of the curve. The widely used 4PL model, although not plotted here, showed similar performance to the log-linear model.

Some important points regarding this model selection process:

(1) If a quick decision on a "working" calibration curve model has to be made based on just one run, independent VS instead of calibrators should be used for the evaluation. The use of calibrator samples alone to select the optimal model will bias a more complex model due to overfitting. In addition, VS levels should not duplicate calibrator concentrations.

(2) Although widely reported, R^2 is not useful for evaluating the quality of a calibration curve model because it does not penalize model complexity and consequently encourages overfitting. Alternative model selection criteria such as the Akaike's Information Criterion and Schwarz's Bayesian Information Criterion are abstruse, and neither is explicitly designed to choose models with optimal calibration properties. Our proposed method uses an independent set of validation samples to compare the models objectively with respect to the assay performance characteristics such as bias, precision, and total error.

APPENDIX C

Importance of weighting for calibration curves

Figure 5 shows the bias (% relative error) and precision (error bars) of VS from a prestudy validation of an ELISA, corresponding to the weighted and unweighted 4PL models. The range of the VS covers the range of the calibration curve in Fig. 2b, where the study samples are expected to fall. It is clear from this example that weighting significantly improves

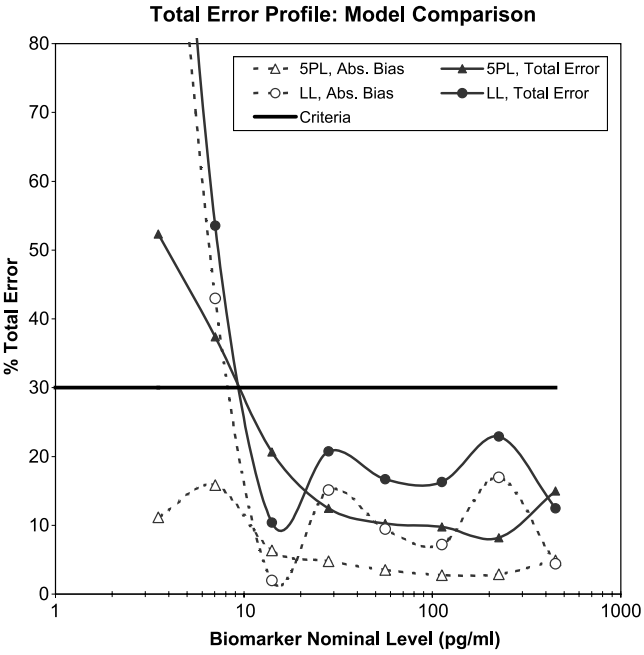

Fig. 4. Calibration model selection based on Total Error (TE). The solid and dashed lines with circles represent TE and Absolute Bias, respectively for a linear model in log-scale (LL), the solid and dotted lines with triangles for a five-parameter logistic model (5PL). TE and Absolute Bias are determined using VS from four independent assay runs.

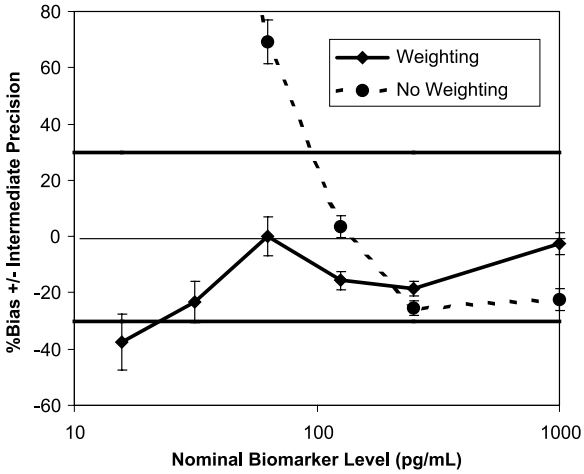

Fig. 5. Weighting for calibration curves. The solid line and the dashed line are the % bias corresponding to the weighted and unweighted four-parameter logistic model, respectively. The error bars represent the intermediate precision of the assay. The % relative error and the intermediate precision are determined using VS from four independent assay runs.

the recovery (total error) of the VS across the range of interest, and thus can have a tremendous impact on the performance of an analytical method.

In general, the weights should be estimated based on multiple runs of replicate data using one of the methods described by Carroll and Ruppert (29) in consultation with a statistician. We recommend this practice during method development and prestudy validation. The estimated weighting factor/parameter derived from the prestudy validation should then be used to fix the weights for the calibration curves generated during the sample analysis.

APPENDIX D

Parallelism experiments

The following are points to consider in constructing parallelism experiments. Some of the following experiments will not be part of exploratory validation, but may be carried out in subsequent stages of validation. The experiments should be performed on a case-by-case and matrix-by-matrix basis with a fit-for-purpose approach.

(1) The parallelism sample(s) should be serially diluted to result in a set of samples having analyte concentrations that fall within the quantitative range of the assay. Ideally, the diluent should be the same as the intended sample matrix if an analyte-free matrix is available. In practice, the diluent used should be the same as that used for the calibrators.

(2) Samples from at least three individual donors should be used. They should not be pooled unless the sample volumes are too small. Pooled samples could result in potential assay interferences (e.g., aggregates) and false nonparallelism. Lot-to-lot consistency of parallelism should be noted.

(3) When only limited samples are available at a high analyte concentration, the approach of mixing a high- with a low-concentration sample at various ratios and correlating the concentration to the proportion may be considered.

(4) When no samples of high analyte concentration are available, the endogenous level may be increased by stimulation of transfected cells, or by use of a nontarget population source (e.g., differing in gender or species), if applicable.

(5) Varying cell counts may be used instead of dilution if the analyte is intracellular.

The measured concentrations of the dilution samples can be plotted against 1/dilution factor using log scales and a linear regression performed. Some individuals use observed results *vs.* expected where there is confidence in the actual concentration of the analyte in the samples used. Parallelism is proven when the results show a slope of nearly 1. The broken line in Fig. 6a approached parallel with a slope of 0.995, compared to the less parallel, solid line with a slope of 0.877.

Alternatively, the coefficient of variation (CV) among the recovered concentrations at different dilutions of the test sample can be used to verify parallelism (30). The value of this CV can be adapted on a case-by-case basis based on considerations of other assay performance parameters as well. Figure 6b shows a plot with the "recovered" concentrations replaced by "dilution adjusted" concentrations

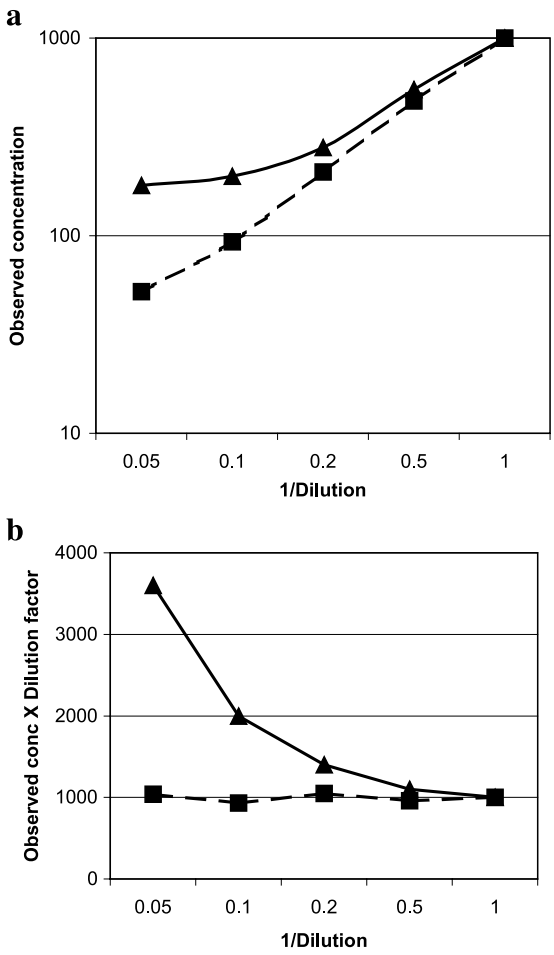

Fig. 6. Parallelism. Data from two individual samples to show parallelism (■) or the lack of it (▲). The same dataset was plotted as (a) observed concentration *vs.* 1/dilution factor, or (b) dilution adjusted concentrations *vs.* 1/dilution factor.

(observed concentration × dilution factor) because the absolute concentration is usually unknown. CV was 5.1% for the matrix that showed parallelism, and 58.7% for the one that did not.

REFERENCES

1. J. W. Lee, R. S. Weiner, and J. M. Sailstad et al. Method validation and measurement of biomarkers in nonclinical and clinical samples in drug development. A conference report. *Pharm. Res.* **22**:499–511 (2005).
2. J. A. DiMasi, R. W. Hansen, and H. G. Grabowski. The price of innovation: new estimates of drug development costs. *J. Health Econ.* **22**:151–185 (2003).
3. J. M. Reichert. Trends in development and approval times for new therapeutics in the United States. *Nat. Rev. Drug. Discov.* **2**:695–702 (2003).
4. FDA Mar 2004 report. Innovation or Stagnation? Challenge and opportunity on the critical path to new medical products. Available at http://www.fda.gov/oc/initiatives/criticalpath/.
5. E. Zerhouni. Medicine. The NIH roadmap. *Science* **302**:63–72 (2003).
6. G. J. Downing (ed). Biomarkers and surrogate endpoints: clinical research and applications. Proceedings of the NIH–FDA Conference held on April 15–16, 1999. Elsevier, New York, 2000.
7. G. Levy. Mechanism-based pharmacodynamics modeling. *Clin. Pharmacol. Ther.* **56**:356–358 (1994).
8. C. C. Peck, W. H. Barr, and L. Z. Benet *et al.* Opportunities for integration of pharmacokinetics, pharmacodynamics, and toxicokinetics in rational drug development. *Pharm. Sci.* **81**:600–610 (1992).
9. W. A. Colburn. Selecting and validating biologic markers for drug development. *J. Clin. Pharmacol.* **37**:355–362 (1997).
10. E. S. Vesell. Advances in pharmacogenetics and pharmacogenomics. *J. Clin. Pharmacol.* **40**:930–938 (2000).
11. Guidance for industry on bioanalytical method validation: availability. *Fed. Regist.* **66**:28526–28527 (2001).
12. Code of Federal Regulations, Title 21, Vol. 1. Good Laboratory Practice for Nonclinical Laboratory Studies. Revised April 1, 2001.
13. Code of Federal Regulations, Title 42, Vol. 3. Clinical Laboratory Improvement Amendment. Revised October 1, 2001.
14. National Committee for Clinical Laboratory Standards (CLSI), Document EP5-A: Evaluation of Precision Performance of Clinical Chemistry Devices: Approved Guideline. 1999. Document EP6-P: Evaluation of the Linearity of Quantitative Analytical Method: Proposed Guideline. 1986. Document EP7-P: Interference Testing in Clinical Chemistry: Proposed Guideline. 1986. Document EP9-A: Method Comparison and Bias Estimation Using Patient Samples: Approved Guideline. 1995.
15. Biomarkers Definitions Working Group, Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin. Pharmacol. Ther.* **69**:89–95 (2001).
16. J. A. Wagner. Overview of biomarkers and surrogate endpoints in drug development. *Dis. Markers* **18**:41–46 (2002).
17. J. A. Wagner. Bridging preclinical and clinical development: biomarker validation and qualification, in R. Krishna and D. Howard (eds.), *Dose Optimization in Drug Development*, Marcel Dekker, New York, in press.
18. J. W. Lee, W. C. Smith, G. D. Nordblom, R. R. Bowsher. Validation of assays for the bioanalysis of novel biomarkers. In C. Bloom, R. A. Dean (eds.), *Biomarkers in Clinical Drug Development*, Marcel Dekker, New York, 2003, pp. 119–149.
19. A. R. Mire-Sluis, Y. C. Barrett, and V. Devanarayan, *et al.* Recommendations for the design and of immunoassays used in the detection of host antibodies against biotechnology products. *J. Immunol. Methods* **289**:1–16 (2004).
20. B. DeSilva, W. Smith, and R. Weiner, *et al.* Recommendations for the bioanalytical method validation of ligand-binding assays to support pharmacokinetic assessments of macromolecules. *Pharm. Res.* **20**:1885–1900 (2003).
21. V. P. Shah, K. K. Midha, and S. Dighe, *et al.* Analytical methods validation: bioavailability, bioequivalence, and pharmacokinetic studies. *Pharm. Res.* **9**:588–592 (1992).
22. ICH Guidelines. Text on validation of analytical procedures, Q2A; International Conference on Harmonization, Geneva, Switzerland, 1994.
23. D. Borderie, C. Roux, and B. Toussaint, *et al.* Variability in urinary excretion of bone resorption markers: limitations of a single determination in clinical practice. *Clin. Biochem.* **34**: 571–577 (2001).
24. C. A. Ray, R. R. Bowsher, W. C. Smith, *et al.* Development, validation and implementation of a multiplex immunoassay for the simultaneous determination of five cytokines in human serum. *J. Pharm. Biomed. Anal.* **36**:1037–1044 (2001).
25. J. O. Westgard, P. L. Barry, and M. R. Hunt, *et al.* A multi-rule Shewhart chart for quality control in clinical chemistry. *Clin. Chem.* **27**:493–501 (1981).
26. M. Feinber, B. Boulanger, and W. Dewe, *et al.* New advances in method validation and measurement uncertainty aimed at improving the quality of clinical data. *Anal. Bioanal. Chem.* **380**:502–514 (2004).
27. J. A. Rogers. Statistical limits for assessment of accuracy and precision in immunoassay validation: a review. *J. Clin. Ligand Assay* **27**:256–261 (2005).
28. C. A. Ray, C. Dumaual, and M Willey, *et al.* Optimization of analytical and pre-analytical variables associated with an *ex vivo* cytokine secretion assay. *J. Pharm. Biomed. Anal.,* In press.
29. R. J. Carroll, D. Ruppert. *Transformation and Weighting in Regression*, Chapman & Hall, New York, 1988.
30. J. W. A. Findlay, W. C. Smith, and J. W. Lee, *et al.* Validation of immunoassays for bioanalysis: a pharmaceutical industry perspective. *J. Pharm. Biomed. Anal.* **21**:1249–1273 (2000).