ELSEVIER

Journal of Pharmaceutical and Biomedical Analysis
21 (2000) 1249–1273

JOURNAL OF
PHARMACEUTICAL
AND BIOMEDICAL
ANALYSIS

www.elsevier.com/locate/jpba

Validation of immunoassays for bioanalysis: a pharmaceutical industry perspective

J.W.A. Findlay [a], W.C. Smith [b], J.W. Lee [c], G.D. Nordblom [d], I. Das [a,*], B.S. DeSilva [e], M.N. Khan [f], R.R. Bowsher [g]

[a] *Metabolism and Safety Evaluation, Searle, 4901 Searle Parkway, Skokie, IL 60077, USA*
[b] *Statistical and Mathematical Sciences, Lilly Research Laboratories, Eli Lilly and Company, Greenfield, IN 46140, USA*
[c] *MDS Harris, 621 Rose Street, Lincoln, NE 68502, USA*
[d] *Department of Pharmacokinetics, Dynamics and Metabolism, Parke-Davis Pharmaceutical Research, Division of Warner-Lambert Co., 2800 Plymouth Road, Ann Arbor, MI 48105, USA*
[e] *Bioanalytical, Procter and Gamble Pharmaceuticals, Route 320 Woods Corners, Norwich, NY 13815, USA*
[f] *Clinical Testing Laboratory, Quality Control Department, MedImmune, 636 Research Drive, Frederick, MD 21703, USA*
[g] *Drug Disposition, Lilly Research Laboratories, Eli Lilly and Company, Lilly Corporate Center, Indianapolis, IN 46285, USA*

Received 10 September 1999; received in revised form 12 October 1999; accepted 13 October 1999

Abstract

Immunoassays are bioanalytical methods in which quantitation of the analyte depends on the reaction of an antigen (analyte) and an antibody. Although applicable to the analysis of both low molecular weight xenobiotic and macromolecular drugs, these procedures currently find most consistent application in the pharmaceutical industry to the quantitation of protein molecules. Immunoassays are also frequently applied in such important areas as the quantitation of biomarker molecules which indicate disease progression or regression, and antibodies elicited in response to treatment with macromolecular therapeutic drug candidates. Currently available guidance documents dealing with the validation of bioanalytical methods address immunoassays in only a limited way. This review highlights some of the differences between immunoassays and chromatographic assays, and presents some recommendations for specific aspects of immunoassay validation. Immunoassay calibration curves are inherently nonlinear, and require nonlinear curve fitting algorithms for best description of experimental data. Demonstration of specificity of the immunoassay for the analyte of interest is critical because most immunoassays are not preceded by extraction of the analyte from the matrix of interest. Since the core of the assay is an antigen–antibody reaction, immunoassays may be less precise than chromatographic assays; thus, criteria for accuracy (mean bias) and precision, both in pre-study validation experiments and in the analysis of in-study quality control samples, should be more lenient than for chromatographic assays. Application of the SFSTP (Societe Francaise Sciences et Techniques Pharmaceutiques) confidence interval approach for evaluating the total error (including both accuracy and precision) of results from validation samples is recommended in considering the acceptance/rejection of an immunoassay procedure resulting from validation experiments. These recommendations for immunoassay validation are presented in the hope that their

* Corresponding author. Fax: +1-847-982-4857.
E-mail address: ira.das@monsanto.com (I. Das)

PII: S0731-7085(99)00244-7

consideration may result in the production of consistently higher quality data from the application of these methods.

Keywords: Pharmaceutical; Drug development; Immunoassays; Biomarkers; Antibodies; Assay validation; Bioanalytical; Calibration; Immunogenicity

1. Introduction

It is now nearly a decade since the December, 1990, Crystal City conference on the validation of bioanalytical methods, that was co-sponsored by the regulatory agencies in the United States (Food and Drug Administration) and Canada (Health Protection Branch), as well as by several scientific organizations. The proceedings and recommendations of that meeting were subsequently published [1], and have essentially become de facto guidelines for validation of bioanalytical methods in the pharmaceutical industry. The conference addressed bioanalytical methods in general, but also acknowledged differences between chromatographic assays and biologically based assays, and included some specific recommendations for validation of immunoassays and microbiological assays.

Since 1990, the general issue of analytical method validation has also been discussed at the level of the International Conference on Harmonization [2,3]. In addition, there has been an analysis of the original Crystal City conference report [4], and additional discussions of the validation of chromatographic methods [5,6]. Validation of immunoassays has also been reviewed in general [7,8] and specifically for assays for macromolecules [9]. Other papers have focused on the statistical aspects of validation [10,11] and on the use of quality control (QC) samples for acceptance of analytical runs, as proposed in the Crystal City conference report [12].

The collective opinion of the current authors is that the published conference proceedings [1] do not adequately address the special issues pertaining to the validation of immunoassays. Following a roundtable discussion of these issues at the 1998 annual meeting of the American Association of Pharmaceutical Scientists (AAPS), we agreed that a more rigorous discussion was needed. This opinion was reinforced by the recent publication by the FDA of a draft Guidance for Industry on Bioanalytical Methods Validation for Human Studies which, after a brief mention of immunological techniques, provides no guidance on appropriate validation approaches for these methods [13]. Thus, this paper presents a discussion of important issues that are specific to the validation of immunoassays, provides guidance on how to deal with these issues and identifies additional issues which require further consideration.

As for any bioanalytical method, the extent to which an immunoassay should be validated depends on the intended application of the method. Thus, an immunoassay developed to support early R&D discovery, when rapid turnaround of results is needed, does not need to be fully validated. On the other hand, immunoassays to support GLP-compliant preclinical safety studies or clinical pharmacokinetic and bioequivalence studies should be fully validated prior to the analysis of study samples.

This paper focuses on topics that the authors consider most important to the validation of immunoassays. Section 2 provides some background information concerning immunoassays and their application to the bioanalysis of pharmaceutical products. Section 3 describes some fundamental differences between chromatographic assays and immunoassays, highlights unique considerations for the validation for immunoassays and addresses specific limitations of current guidelines[1] as they relate to immunoassays. Section 4 presents information concerning the optimization of the calibration model for immunoassays, while Section 5 describes recommended procedures and acceptance criteria for pre-study validation and

[1] For this publication, validation guidelines refer to information in guidance and draft guidance documents included as Refs [1] and [13], respectively.

Section 6 includes specific recommendations for run acceptance during the analyses of study samples (in-study validation). Sections 7 and 8 provide discussions of, and recommendations for, the immunoanalysis of biomarkers and antibodies (immunogenicity), respectively. Section 9 summarizes the discussions and provides concluding remarks. We believe that broad application of the proposals described herein will help standardize immunoassay validation procedures and ensure high quality bioanalytical data for support of preclinical and clinical studies. Whenever possible, the ICH standard terminology is used in this document [2], and these terms are included in a glossary provided in Appendix A. Information about validation statistics is included in Appendix B.

2. Background

Immunoassays are analytical methods based on signal responses generated as a consequence of an antibody–antigen reaction. The response signal is generated from a label (e.g. enzymatic, fluorescent, luminescent or radioisotopic) attached to either the analyte (antigen) or antibody, or from a secondary, high affinity binding reaction, usually involving another labeled antibody or the well-characterized biotin–avidin system [14]. The concentration of unlabeled analyte in a test sample can be calculated by interpolation from a calibration (standard) curve. Immunoassays are capable of quantifying a variety of compounds, ranging from traditional small molecule xenobiotic drugs to large macromolecules. Antibody assays also provide the basis for assessing the pharmacokinetics of antibody-based therapeutics and the immunogenicity of candidate macromolecular drugs [15].

During the past decade, the use of immunoassays for the bioanalysis of low molecular weight drugs has declined in the pharmaceutical industry [16]. The single greatest contributor to this change, namely modern mass-spectrometry, may ultimately spur a resurgence in the use of immunoassays. Mass-spectrometry can serve as a reference method for verifying the specificity of an immunoassay [16] and, thereby, address the major limitation of small molecule immunoassays, namely their perceived lack of specificity. Once specificity has been confirmed, immunoassays offer a cost-effective alternative to LC-MS-MS for the analysis of samples from Phase III/IV clinical studies while requiring less sample volume and freeing expensive LC-MS-MS instrumentation for other projects [16]. In some instances where total concentration of the pharmacologically active parent drug and metabolite is needed, an immunoassay with cross-reactivity to both parent compound and metabolically active metabolite may be preferred to a chromatographic method [7]. Furthermore, the recent surge of interest in therapeutic proteins and other biomacromolecules has assured a role for immunoassays in pharmaceutical development, since immunoassay is the methodology of choice for measuring therapeutic proteins in biological matrices [7]. Immunoassays are also used to quantify biomarkers for the assessment of drug pharmacodynamics and disease progression or regression.

3. Analytical issues specific to immunoassays

Some differences between immunoassays and chromatographic assays illustrate the need for additional guidance on the validation of immunologically based assays (Table 1). Chromatographic assays depend on chemical or physicochemical properties of the molecule for detection, whereas the critical component of an immunoassay is the binding reaction between the analyte (antigen) and antibody, coupled with a suitable endpoint detection system. The antibody reagent is derived from an animal source, with the attendant variability typical of such reagents. Also, the time to develop a new immunoassay may be months, because of the need to elicit the desired antibody by immunization. In contrast, a prototype chromatographic assay can often be established in days. However, once suitable reagents are available, an immunoassay can be established in a timeframe that is competitive with chromatographic assays. Calibration curves for chromatographic assays are typically linear, whereas most

Table 1
Differences between chromatographic assays and immunoassays

	Chromatographic assays	Immunoassays
Basis of measurement	Physicochemical properties of analyte	Antigen–antibody reaction
Analytical reagents	Well-characterized and widely available	Unique and usually not widely available
Analytes	Small molecules	Small molecules and macromolecules
Detection method	Direct	Indirect
Sample pretreatment	Yes	Usually no
Calibration model	Linear	Nonlinear
Assay environment	Contains organic solvents	Aqueous
Time required for development	Weeks	Months (due to time needed for Ab generation)
Intermediate (inter-assay) imprecision	Low (<10%)	Moderate (<20%)
Source of imprecision	Intra-assay	Inter-assay
Assay working range	Broad	Limited
Cost of equipment	Expensive	Inexpensive
Analysis mode	Series, batch	Batch
Assay throughput	Good	Excellent

immunoassay curves are inherently nonlinear. As a result of these fundamental differences, there are limitations in the current validation guidelines [1,13] when applied to immunoassays; these limitations are particularly acute for assays for macromolecules, such as those for proteins and some biomarkers.

Proteins differ from traditional small molecule xenobiotic drugs in that their disposition is determined largely by factors that govern their in vivo physiology [17–19]. For this reason, the development and validation of immunoassays for proteins present unique challenges. One common issue for therapeutic proteins is the presence of endogenous equivalents in biological matrices (see Section 3.4.2) [7,20]. This not only poses a bioanalytical challenge, but also complicates the design of pharmacokinetic studies and the analysis and interpretation of pharmacokinetic data [19–21]. Immunoassay issues in the bioanalysis of proteins have been discussed in several recent review articles [7,9,20–23].

The Crystal City conference laid excellent groundwork for the validation of bioanalytical methods in general. Specific bioanalytical and statistical issues, which differentiate immunoassays from chromatographic assays, will be discussed in the following subsections and demonstrate the need for special considerations in validation of immunoassays. A suggested scheme for the development (pre-validation), validation (pre-study validation) and implementation (in-study validation) of immunoassays is presented in Fig. 1. This figure emphasizes that method development and validation should be viewed as a continuum.

3.1. Unique reagents

Immunoassay reagents are generally not available commercially, but must be acquired or produced in adequate amounts and characterized sufficiently for assay development, validation and subsequent analysis of test samples. Some reagents are also subject to lot-to-lot variation (e.g. antisera, conjugates and radiolabeled components). It is, therefore, important to document the quality of these reagents.

One characteristic that should be investigated during the generation of immunological reagents is stability. Assay performance and sensitivity generally deteriorate with degradation of immunological reagents. Therefore, it is important to investigate storage conditions to ensure the integrity of key assay reagents for the anticipated duration of time that they will be needed. In the absence of stability information, it is recommended that reagents be stored refrigerated or frozen, since biologics are generally more stable at low temperatures [23].

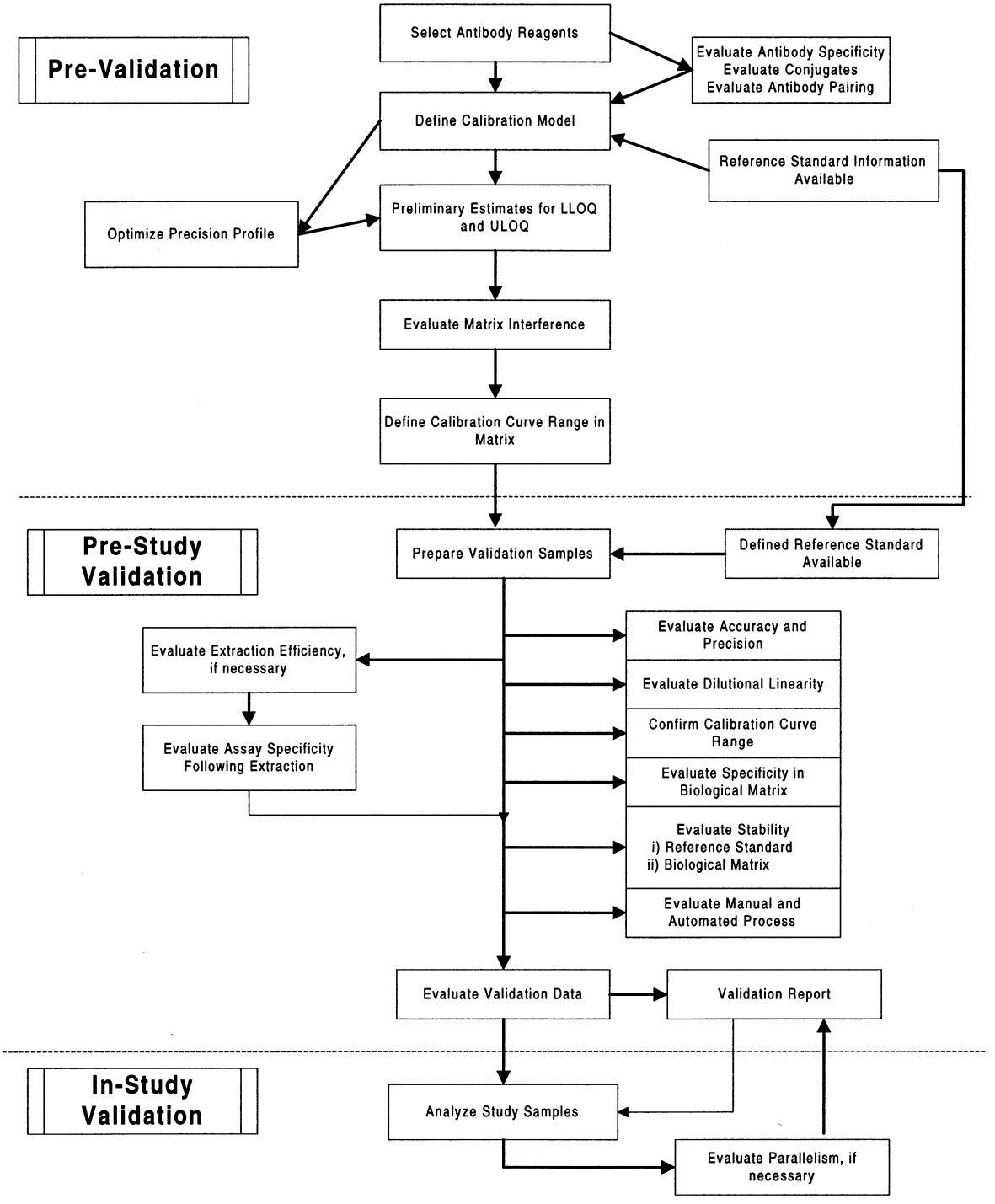

Fig. 1. A schematic of pre-validation, pre-study- and in-study validation processes for immunoassays.

3.1.1. Reference standards

A reference standard for most small molecule xenobiotic drugs is usually well characterized. For some proteins, reference standards are not as rigorously characterized, and proteins from different sources can vary in their potency and immunoreactivity. In some instances, the reference material may not be truly representative of the administered protein due to differences in post-translational modifications, such as the extents of deamidation and glycosylation. When a reference standard does not have well-defined physical characteristics, its biological activity may be better defined, and the biological activity (e.g. binding or enzyme activities) per unit weight or volume should then be used and documented. For example, an activity that has been standardized by an organization, such as the World Health Organization (WHO) or United States Pharmacopeia (USP), can be used in place of mass per volume. Unfortunately, not all proteins are characterized in terms of standardized biological activity units [24–26].

3.2. Nonlinear calibration curve models

Immunoassay techniques have intrinsic characteristics that make selection of a calibration curve model more complicated than for chromatographic assays. For most immunoassays (RIA, IRMA, ELISA, EIA), the mean response is a nonlinear function of analyte concentration, and the variance in replicate response measurements is a nonconstant function of the mean response. A statistical model that accurately describes the functional relationships will maximize the reliability of results derived from the calibration curve. Although numerous articles have been published on models for fitting immunoassay data [27,28], issues associated with fitting and interpreting data from nonlinear heterogeneous concentration–response relationships are often not well understood and ignored in practice.

Current guidance recommendations regarding calibration curves may be inappropriate for immunoassays. For example, the draft guidance [13] states that the lower limit of quantitation (LLOQ) should serve as the lowest concentration on the standard curve. In the case of immunoassays, inclusion of low- and high-calibrators (outside the dynamic range of quantitation) with responses near the two 'flat' asymptotic regions of the sigmoidal curve are often beneficial in curve-fitting and can improve the accuracy (mean bias) and precision of interpolated results near the limits of quantitation.

3.3. Precision

Immunoassays are inherently less precise than chromatographic assays [29]. Consequently, specific aspects of current guidance recommendations for the design of validation experiments to assess precision are not completely appropriate for immunoassays. Due to their greater inter-assay imprecision, immunoassays may require more validation runs than for chromatographic methods to achieve the same level of confidence in the estimates of assay performance [10]. Given the higher inherent imprecision, a higher percentage of in-study immunoassay runs would be expected to fail the 4-6-20 quality control acceptance criteria [12]. Thus, recent publications have recommended that the 20% acceptance limit be made less restrictive for immunoassays (e.g. 25%) [9,29].

3.4. Specificity

Specificity is the ability to measure the analyte unequivocally in the presence of other components, either exogenous or endogenous [2,30]. In general, chromatographic methods are selective since they separate, detect and quantify multiple analytes in a complex biological milieu. In contrast, immunoassays need to be 'specific', because they are used to measure analytes in biological matrices, preferentially without prior sample extraction. A number of approaches have been recommended for the validation of immunoassay specificity [1].

Nonspecificity can be classified as either 'specific nonspecificity' or as 'nonspecific nonspecificity' [31]. Specific nonspecificity is analytical interference that is caused by substances that have physicochemical similarity to the analyte, including metabolites, degraded forms of the analyte,

isoforms, and variants that differ in their post-translational modifications. Host anti-idiotypic antibodies are also capable of causing specific assay interference.

Nonspecific nonspecificity is analytical interference caused by factors other than those listed above, which also may affect the antigen–antibody binding reaction [31]. Nonspecific nonspecificity is commonly referred to as 'matrix effects'. Immunoassays are more prone to matrix interference than are chromatographic assays, because immunoassays are performed routinely without sample extraction. Some factors that interfere nonspecifically with the antigen–antibody binding reaction include, hyperlipidemia, hemolysis, ionic strength, pH, cations, sample viscosity, serum proteins (e.g. complement and rheumatoid factor), anticoagulants, proteases, binding proteins, autoantibodies and heterophilic anti-IgG antibodies.

We recommend that, when possible, at least six different lots of biological matrix be evaluated for nonspecific interference. Matrix interference may be assessed by comparing the concentration–response relationship of spiked and unspiked biological matrix to those in a buffer matrix. Dilution with buffer containing chaotropic or chelating agents (e.g. Tween-20, Triton X-100, and/or EDTA) may be effective in minimizing nonspecific interference. If dilution fails to decrease the interference, further sample clean up, such as protein precipitation, liquid–liquid or solid-phase extraction, or HPLC separation may be useful. If sample pretreatment is employed, the same procedure should be applied to the calibrators (standards), QC samples, and test samples to ensure analytical consistency and eliminate the need for a correction factor when calculating the analyte concentrations in unknown test samples. As a last resort, pre-dose samples from each subject may be used to construct the calibrators and QC samples for quantitation of specific samples from that subject.

3.4.1. Low molecular weight xenobiotic molecules

Even with the advent of quantitative LC-MS procedures, the relatively low cost and high throughput of immunoassays still make these assays potentially attractive as alternatives for bioanalytical support of late stage clinical studies of low molecular weight chemical entities. Immunoassays for small molecules are subject to all of the validation requirements described in detail elsewhere in this manuscript.

For small molecule xenobiotic drugs, specificity is of paramount importance for immunoassay validity. By the time a conventional xenobiotic drug reaches late stage clinical development, metabolic pathways have normally been elucidated and the circulating concentrations of parent and metabolites are known. This information permits the investigation of potential interferences from circulating metabolites separately or in combination with the parent molecule. Analytical interference from co-administered medications and, in some cases, from endogenous molecules that are structurally related to the drug of interest should also be investigated.

A recommended approach to investigate immunoassay specificity for a conventional xenobiotic drug involves evaluating the cross-reactivity of some or all metabolites, endogenous compounds and concomitant medications, followed by comparative analysis of actual study samples by another validated bioanalytical method (e.g. LC-MS-MS), if available. Initial experiments should assess the degree of cross-reactivity from each potential interferent and express the cross-reactivity as the ratio of EC_{50} values. Further tests to investigate cross-reactivity should then be conducted with the compounds in a mixture(s) to reflect the in vivo situation. Results from these cross-reactivity studies, considered along with knowledge of the concentrations of the analyte and potential interferents, should lead to a preliminary conclusion regarding the validity of the immunoassay for its intended application.

The comparative analysis approach mentioned above requires that criteria for comparing the relative accuracy of the immunoassay to the reference method be defined a priori. If the immunoassay meets these predefined criteria (i.e. immunoassay-derived values are within a predefined range of reference method values), it is established as being equivalent to the reference method. Samples chosen for this comparison

should be actual study samples or pooled study samples to reflect relevant metabolite concentrations. The samples should be collected two or more times following drug administration (e.g. approximate time of maximum plasma concentration and a subsequent time corresponding to several elimination half lives) to allow immunoassay performance to be assessed in the presence of varying amounts of metabolites.

3.4.2. Therapeutic proteins

The ability to define the specificity of immunoassays for therapeutic proteins is more limited than for small molecule drugs, since the catabolic and metabolic pathways are often poorly defined and sufficiently sensitive comparator assays are lacking [9,22]. Specificity is an important consideration when immunoassays are used to assess the pharmacokinetics, bioequivalence and toxicokinetics of therapeutic proteins. Biotransformation (e.g. proteolysis) may or may not alter the antigenicity of a therapeutic protein. In some cases a small change in protein structure (e.g. proteolytic clip) can result in a significant decrease in antigenicity. Conversely, a change in primary or secondary structure may not result in markedly reduced immunoreactivity. Hence, the question often remains whether or not the immunoassay is able to differentiate between biologically active and inactive forms of a therapeutic protein [9,20]. Inferences about assay specificity can be gained from epitope-mapping experiments or by interfacing high resolution separation techniques (e.g. HPLC, FPLC®, electrophoresis and mass-spectrometry) with immunoassay to characterize the nature of the immunoreactivity present in the biological samples. Discordance in measured serum concentrations often results when immunoassays differ in their ability to detect intact and proteolytically cleaved forms (metabolites) of a therapeutic protein.

3.4.3. Endogenous analytes

Since many therapeutic proteins are analogs (recombinant versions) of human proteins [32], it is common for the endogenous protein to be present in the test samples. These endogenous analytes complicate immunoassay development and validation, because calibration curves are prepared preferentially in matrix from the same species as the test samples [1,7,9].

Various strategies may be used to limit or eliminate interference from endogenous analytes. First, if the pharmacological concentrations of the therapeutic protein are substantially higher than the basal concentrations of the endogenous protein, e.g. area under the curve (AUC or total exposure) of endogenous levels $< 5\%$ of total AUC, then the concentration of the endogenous protein will introduce only a small degree of assay bias. Another approach for high concentrations of therapeutic proteins is to dilute the test samples in a physiological buffer or diluted biological matrix and prepare the calibration curve in the same matrix [7]. A third strategy is to prepare 'analyte-free' matrix for preparation of calibrators and QC samples by degradation of the endogenous analyte [33], removal by nonspecific adsorption (e.g. charcoal) [34], or specific removal by immunoaffinity chromatography [7,35]. Another useful strategy is to prepare the calibration curve in a heterologous biological matrix, such as equine or pig serum, that is devoid of, or contains low endogenous levels of, the analyte [7]. Regardless of the strategy employed, it is necessary for the calibration curve to be representative (negligible bias) of the analyte in the test sample. In some cases, it may not be possible to remove the endogenous analyte or use a surrogate matrix. In such cases, the calibration curve may be 'adjusted' by either subtracting the endogenous level from the calibrators or assigning a 'corrected' concentration to each of the calibrators.

Endogenous analytes also present an issue in the preparation of control samples, since treatment of a biological matrix to remove an endogenous analyte may alter the matrix so that it is no longer representative of the test sample matrix. While spiking reference standard into a pool of treated or heterologous matrix can be used during pre-study validation to establish the validated range of the assay, QC samples to be used during assay implementation should be prepared in the untreated spiked matrix to assess bias and analyte stability. When interfering endogenous analytes are present, analytical bias can also be evaluated by the method of standard addition [6].

3.5. Protein binding

Pharmacological effects of drugs are usually attributed to the unbound or 'free' fraction of drug in the biological system [36]. Most small molecule drugs are bound to a limited number of plasma proteins such as albumin and α_1-acid glycoprotein in a low affinity and high capacity manner [37]. In general, immunoassays for these small molecules measure 'total' concentrations (i.e. protein-bound and free). Some small molecule xenobiotic drugs can bind with high affinity binding to specialized proteins in the circulation, e.g. progestagens to sex hormone binding globulin [38] and tacrolimus to FK506 binding protein [39]. In these cases, sample pretreatment to release the small ligand analytes is usually performed prior to immunoanalysis.

Protein binding of circulating therapeutic proteins is more complicated and less well-defined than for conventional small molecules. Therapeutic proteins can bind to both low (e.g. albumin, α_2-macroglobulin [27,40,41] and protease inhibitors [27]) and high affinity binding proteins (e.g. soluble receptors [42,43] and carrier proteins [27,44,45]). Binding proteins are known to cause interference in immunoassays for some proteins, including cytokines [46,47], growth hormone [48], tissue plasminogen activator [26] and insulin-like growth factors-I and -II [49,50]. Disagreement in measured serum concentrations may occur when immunoassays differ in their ability to detect the free and bound forms of a therapeutic protein. Extraction procedures used to release small molecules from serum proteins are usually not applicable to therapeutic proteins. Autoantibodies resulting from the repeated administration of a therapeutic protein are also capable of causing analytical interference and altering the pharmacokinetics and pharmacodynamics of the therapeutic protein [5,6,51].

4. Recommendations for calibration model optimization

In this section, procedures are proposed for dealing with immunoassay specific calibration issues that should be addressed prior to the initiation of definitive pre-study validation experiments.

4.1. Calibration model selection

A calibration model that accurately fits the concentration-response relationship for standard (calibration) samples should be determined prior to generating results for independently prepared validation samples. Selection of a model requires the specification of (1) an algebraic equation that represents the mean concentration–response relationship, and (2) a second equation (possibly constant) that characterizes the response–error relationship (i.e. the relationship between the mean response and the variance of replicate measurements about the mean). Insight into the two functional relationships can be gained by plotting the observed concentration–response data and the descriptive statistics (e.g. sample standard deviation (S.D.) versus sample mean) from responses at each concentration level [52].

4.1.1. Calibration models

A commonly acknowledged 'reference' model for fitting immunoassay data uses the four-parameter logistic (4PL) equation to fit the mean concentration–response relationship and the power-of-the mean (POM) equation to fit the response–error relationship [53]. The 4PL function provides an accurate representation of the sigmoidal relationship between the measured response and the logarithm of concentration observed for many immunoassays (RIA, IRMA, ELISA, EIA, etc.). Similarly, the POM function accurately represents the response–error relationship in which the variance in replicate response measurements is an increasing function of the mean response.

For noncompetitive assays in which no response values are observed above the EC_{50}, it may be necessary to constrain the 4PL parameter that defines the upper asymptote to be a constant (e.g. for an IRMA, set the asymptote equal to the total CPM). In other applications, it may be necessary to add a fifth parameter to the 4PL equation to accommodate 'asymmetry' in the mean concentra-

tion–response curve [53]. Equations other than the logistic function (e.g. cubic spline, logit–log, etc.) that result in a continuous monotonically increasing (or decreasing) calibration curve may be adequate for some applications. However, unlike chromatographic methods the simple linear model is seldom acceptable for immunoassays.

4.1.2. Curve-fitting algorithm

An appropriate curve-fitting algorithm must be selected for calibration curve estimation. A weighted, nonlinear, least-squares method is generally recommended for fitting dose–response data from immunoassays [27,53]. Weighting is needed to account for the heterogeneity of response variances evident in the response–error relationship. It is recommended that weights be computed using smoothed variance estimates (e.g. based on the POM function) rather than individual weights from replicate measurements. Individual weights lead to undesirable statistical properties in the estimates of calibration curve parameters, particularly in the case of duplicate measurements [54]. The arbitrary use of weights such as the inverse of the response $(1/Y)$, or the inverse of concentration $(1/X)$ is not appropriate without an evaluation of the response–error relationship. Failure to weight responses properly will result in greater bias and imprecision in analytical results, particularly at analyte concentrations near the limits of quantitation.

It is generally preferable to use results from multiple runs to estimate the response–error relationship, because of the limited replication present in a single analytical run [53,55]. The estimation procedure will depend upon the function chosen to fit the variance relationship. For example, the POM model assumes that the expected standard deviation $S(Y)$ is related to the expected response $E(Y)$ by the equation $S(Y) = \sigma E(Y)^{\theta}$, where σ is a proportionality constant and θ is a shaping parameter that is considered stable across runs. Consequently, it is recommended that a value for θ be estimated by pooling information from multiple runs. A common estimation procedure is to use simple linear regression to obtain an estimate of θ for each run [27,56], and then to set θ equal to the mean value of all runs. A more efficient (less variable) estimate of θ for each run can be obtained by using generalized least squares [57]. Periodic re-evaluation of the value for θ should be completed during routine assay use.

4.1.3. Model acceptance

Acceptability of a proposed model should be verified for a particular application by evaluating the relative bias (% R.E.) between back-calculated and nominal concentrations of the calibration samples [58]. For a model to be acceptable, we suggest that the mean % R.E. of calibrators within the anticipated validated range be generally no more than 10%. Two or more candidate models can be compared informatively using the % R.E. of back-calculated standards [58]. In some cases, if differences in bias and precision are practically unimportant, it may be preferable to choose a simpler model over a more accurate but complex model. Standard curves from a minimum of three pre-validation analytical runs should be evaluated in selecting a model.

Classical statistical tests for goodness-of-fit are often too restrictive in model assessment because they are designed to test for zero lack-of-fit, which is seldom a requirement for analytical procedures. A desired model can produce acceptable results even in the presence of a statistically significant lack-of-fit. Use of the correlation coefficient is also not recommended for model validation [6]. Even for a linear model, unacceptable calibration bias can exist despite a correlation > 0.9999 [11]. Correlation measures are even less informative when assessing models for nonlinear dose–response relationships.

4.2. Preliminary estimate of validated range (limits of quantitation)

The validated range is the range of concentrations, from the lower limit of quantitation (LLOQ) to the upper limit of quantitation (ULOQ), for which interpolated results have an acceptable level of total error[2] (accuracy and precision). One approach for obtaining initial esti-

[2] In this publication, the term total error refers to the closeness between a measured test result and its nominal value; it is a combination of systematic (mean bias) and random (precision) error components.

mates of the limits of quantitation is to include additional sets (e.g. ≥ 2) of standard samples in the runs used to assess model acceptance. These additional samples are treated as test samples with analyte concentrations calculated from the fitted calibration curve. The mean bias (% R.E.) and intermediate (inter-assay) precision (% C.V.) of these results are computed and plotted against the nominal concentrations. The plots provide insight into positional biases and the precision profile, respectively. Preliminary estimates of the LLOQ and ULOQ are given by the lowest and highest concentrations, respectively, for which the two-sided 90% SFSTP confidence limits for percent relative error are within 25% of the nominal value (Fig. 2).

4.3. Calibrators outside the validated range

The guidance states [13] that the lower limit of quantitation (LLOQ) should serve as the lowest standard on the calibration curve, meaning that the calibration curve range and the assay range should be the same. This statement regarding the calibration curve may be inappropriate for immunoassays since 'anchoring' calibration points are beneficial in curve-fitting and lead to improved accuracy and precision (total error) at the lower and upper limits of quantitation [7]. 'Anchoring' points that fall outside the limits of quantitation should be retained in the calibration set unless it can be demonstrated that their removal does not adversely affect the bias or precision within the validated range. Blank samples should not be included as zero calibrators.

5. Recommendations for pre-study validation of immunoassays

For any analytical procedure, the goal of pre-study validation is to document that the procedure will produce reliable analytical results that meet requirements of the laboratory and the intended user(s). The validation process requires a specification of (1) 'theoretical' acceptance limits for the unknown value of assay parameters (e.g. mean bias and precision), (2) a design for completing validation experiments, (3) appropriate statistical analysis procedures, and (4) data-based statistical acceptance criteria that define rules for

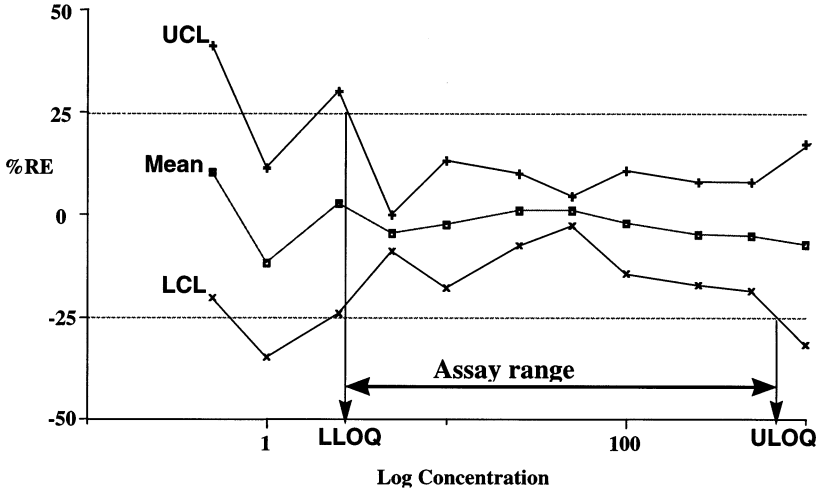

Fig. 2. A plot of the percent relative error (% R.E.) for a typical immunoassay. Estimates of the mean bias (mean) and the lower (LCL) and upper (UCL) SFSTP confidence limits are plotted versus the logarithm of the nominal sample concentration. Estimates of the lower and upper limit of quantitation are represented by LLOQ and ULOQ, respectively.

Table 2
Statistical acceptance criteria for immunoassays[a]

Assay characteristic	Statistical acceptance criteria[b]
Pre-study validation	
Total error	$\lvert (100/\mu_T) \cdot [(\bar{z}_{..} - \mu_T) \pm t_{v,0.95} \cdot s_{IP}] \rvert \leq 25$
Dilutional linearity	% R.E. < 20 (each dilution level)
Parallelism	% R.S.D. < 20 (among dilution levels)
Analyte stability (biological matrices)	$\lvert 100 \cdot (\bar{z}_{stored} - \mu_T)/\mu_T \rvert \leq 20$
	$\lvert 100 \cdot (\bar{z}_{stored} - \bar{z}_{fresh})/\bar{z}_{fresh} \rvert \leq 20$ (when using fresh control)
In-study validation	
Total error	4-6-25 rule

[a] All statistics are defined in Appendix B.
[b] Limits may be less restrictive for some applications.

accepting or rejecting a method. In this section, recommendations for each of these specifications are provided for immunoassays.

5.1. Target acceptance limits

The design, analysis, and interpretation of validation experiments should be guided by the specification of acceptable limits for assay parameters. The specified limits define the minimal performance required of a validated method, and they provide a reference for the evaluation of statistical acceptance criteria (Table 2). Guidance documents, scientific publications, and previous laboratory experiences should be used to determine a priori limit values for a given assay. For immunoassays, we recommend minimal acceptance limits of 20% (25% at the limits of quantitation) for accuracy (mean bias) and precision. Previous limits of 15% (20% at the LLOQ) recommended in the 1990 Crystal City conference report [1] are often too restrictive for immunoassay applications [9,23], including biomarkers and antibody assays. Limits greater than 20% (25% at the LLOQ) are permissible if a scientific rationale exists.

5.2. Design considerations

Information from pre-validation (i.e. assay development) experiments (e.g. analytical runs to investigate the mean concentration–response relationship) should be used to optimize the design of pre-study validation experiments.

5.2.1. Calibrator concentrations

The optimal number of calibrator concentrations and replicates depends on the nature of the concentration–response relationship, and other factors such as constraints imposed by the assay format (e.g. 96-well microtiter plate for ELISAs). It is recommended that a minimum of six nonzero concentrations be used when fitting a calibration curve to the nonlinear (sigmoidal) concentration–response. Each concentration level of calibrator should be analyzed at least in duplicate. The concentrations should span the anticipated assay range with levels approximately equally spaced on a logarithmic scale and chosen to optimize the precision of reported analytical results.

Optimal concentration levels for calibration based on the 4PL function have been derived for an ELISA application [59]. In general, a calibrator should be approximately at the EC_{50} and the remaining calibrators should bracket the EC_{50} based on a logarithmic progression. Assuming a constant response variance (i.e. $\theta = 0.0$ for the power of the mean function), the optimal midpoint of the concentration series is equal to the EC_{50} [59]. For the more typical case in which the response variance is an increasing function of the mean response ($\theta > 0.0$), the optimal midpoint is greater than the EC_{50} for noncompetitive assays (e.g. ELISA) and less than the EC_{50} for competitive assays (e.g. RIA).

5.2.2. Validation sample concentrations, dilutional linearity and analyte stability

Validation samples should include a minimum of three analyte concentrations, including the anticipated LLOQ, ULOQ, and the approximate midrange (on logarithmic scale) of the calibration curve. If some study sample analyte concentrations are expected to exceed the ULOQ, then one

or more additional validation samples in this concentration range should be included to assess dilutional linearity. We recommend that these dilution samples be prepared in the range of expected maximum concentrations in the study samples to assess analyte recovery after dilution. Assessment of dilutional linearity is particularly important when the calibrator diluent differs in composition from the test sample matrix.

Although a direct evaluation of test sample storage stability for short- or intermediate-term is possible within a single analytical run, the continuing nature of storage stability experiments means that this approach is impractical for longer storage time periods, such as 6 months and 1 year. Consequently, a comparison of stability test results to an original assay value must be made relative to the intermediate (inter-assay) precision of the assay. Stability assessments can be made more precise by performing multiple runs, or by examining the trend of stability sample values over time. Alternatively, stability experiments can be designed to eliminate inter-run variability by analyzing the stored sample(s) and a freshly prepared control within the same run.

5.2.3. Number of validation runs/replicates

Since immunoassays have increased inter-assay imprecision, we recommend at least six validation runs. Assays should be performed over several days, with no more than two runs per day per analyst by the method being validated. At least three sets of validation samples should be included in each run, with each sample analyzed in, at least, duplicate. A set contains validation samples at all respective concentration levels (e.g. LLOQ, midrange, ULOQ).

A validation run should be rejected only if there is an analytical problem with an assignable cause (e.g. an error in reagent preparation, instrument failure, pipetting error), or if the data from a run are so erratic that the values could only have resulted from an unexplained analytical error(s). Otherwise, analytical results from validation experiments may not be representative of values generated during routine application of a procedure.

5.3. Statistical analysis of validation data

For validation designs in which replicate measurements are made over multiple analytical runs, a one-way random effects model is commonly used to estimate the accuracy (mean bias) and precision of a method. The model has been described in statistical texts [60] and other publications [10,61,62]. A brief review of the model, with statistical formulae representing one approach to estimating model parameters for a balanced study design, is presented in Appendix B. The formulae are provided primarily as a convenient reference for interpreting the statistical acceptance criteria in Table 2. More general formulae for unequal numbers of replicates, including confidence intervals for precision estimates, are available in statistical textbooks [60].

5.4. Statistical acceptance criteria

The specification of statistical acceptance criteria for method validation is often a source of confusion. Since the desired outcome of a validation study is to accept a method, standard statistical procedures that are designed to reject a point hypothesis (e.g. bias equal to 0) are inappropriate. An alternate approach that is used currently for many applications is to define acceptance criteria based only on point estimates of assay parameters without an assessment of uncertainty. With this approach, the risks of accepting an unsuitable assay and rejecting a suitable assay are unknown and uncontrolled [4,10]. A better approach that controls these risks is to use confidence intervals and equivalence testing procedures [10,63]. Another approach (referred to as the SFSTP approach in this publication) is to use confidence intervals for total measurement error (including both accuracy and precision) so criteria for method acceptance are consistent with those for run acceptance [10,64].

For immunoassays, our recommended statistical criteria for acceptance of a procedure are summarized in Table 2. The criterion for total error is equal to the two-sided 90% confidence interval proposed by the SFSTP Commission [64], except for the use of Satterthwaite's degrees of

freedom for the intermediate precision estimate in lieu of the degrees of freedom for the intra-run variance component. This modification is important because of the greater inter-assay imprecision of immunoassays. The criteria for dilutional linearity and sample stability are defined in terms of the agreement (% R.E.) between observed results and a nominal target value. Parallelism criteria are specified in terms of the agreement among analytical results (% R.S.D.) obtained at different dilution levels of a sample. It is also recommended that point estimates and confidence limits for accuracy (mean bias), repeatability (intra-assay precision), and intermediate (inter-assay) precision be computed and reviewed in the overall assessment of method performance.

The SFSTP criteria for total error for method acceptance have many advantages over criteria based on point estimates of accuracy (mean bias) and precision [64]. The confidence interval approach provides greater control of the risks associated with accepting an unsuitable procedure and rejecting an acceptable procedure. In addition, the SFSTP confidence interval is compatible with in-study acceptance criteria (Section 6). Despite the many advantages of the SFSTP approach, certain issues must be addressed before universal adoption and implementation is possible. These include educating bioanalytical scientists, who need to understand and interpret these criteria, obtaining a consensus among the pharmaceutical industry and regulatory agencies on applicability of the approach, and the widespread availability of validated software. Our view is that the benefits of the SFSTP approach make the resolution of these issues a worthwhile endeavor.

6. Recommendations for in-study validation of immunoassays

Based on arguments presented earlier, we recommend that the 20% limit in the 4-6-20 QC rule for run acceptance be changed to 4-6-25 for immunoassays. For some applications, a limit greater than 25% may be appropriate.

7. Biomarkers

7.1. Rationale

Biomarkers are useful indicators of the pathogenic process of a disease and of the potential effect of drug intervention. During drug development, when clinical end points (outcomes) are far in the future, biomarkers can provide an early measure of efficacy or toxicity, enabling earlier go/no-go decision-making [65,66]. Biomarkers may be classified into two distinct categories, one being discrete analytes that can be measured in concentration units, and the second being binding macromolecules or enzymes that are quantitated in units of activity, with the total activity being the sum of all activities of possible different forms. Selectivity requirements of assays for enzyme activity may not need to be as stringent as those for discrete analytes, but potency (or activity) should be defined for each batch of analyte(s).

7.2. Considerations for assay development and validation

Endogenous levels of biomarkers add complexity to the establishment of assay ranges. If an analyte-free matrix cannot be obtained, establishment of the LLOQ is difficult, as with therapeutic proteins (Section 3.4.2). Standard curves and LLOQ validation pools may be prepared by choosing and pooling matrix from individuals with low baseline concentrations, diluting baseline samples with a protein-based buffer, or using an alternate species matrix with negligible concentrations of the analyte. The ULOQ can be established by fortifying the baseline sample with the analyte.

Standard curves to measure the specific biomarker should bracket the normal and disease-state levels, if possible. If the expected concentration exceeds the dynamic range of the assay, dilutional linearity of validation samples with concentrations approximating the highest expected concentrations in patient samples should be established. In addition, the endoge-

nous levels of the biomarker for healthy subjects should be investigated for both intra-subject (such as circadian or seasonal fluctuations) and inter-subject variability. The recommended number of individual samples to be tested is 25 or greater. It is important to keep in mind that there may be wide inter-subject variability and the number needed to define the disease-state level of biomarkers should be carefully considered. In addition, some biomarkers do not have a graded concentration relationship to the disease status and/or to drug intervention, i.e. they may have a quantal, or all-or-none phenomenon. In this situations, a rough 'cut-off' criterion may be established for decision-making.

Since biomarkers are endogenous, standards may be prepared from an analyte-free protein-based buffer. Whenever possible, the appropriate biological matrix should be used for QC sample preparation. This could be a systemic matrix, such as whole blood or plasma, or target tissue-specific matrix such as sputum, cerebrospinal fluid, aqueous humor, platelets, T-cells or tissues. Alternatively, if no matrix effect can be demonstrated, quality control pools may be prepared in an 'analyte-free' protein-based buffer. The concentration levels in the QC samples should be representative of the concentration levels in physiological and pathological disease states.

7.3. Recommendation on acceptance criteria

In order to obtain clinically meaningful data, the same total error (accuracy and precision) criteria described elsewhere in this paper should be required for most biomarkers. However, QC samples should be permitted higher bias because of the endogenous nature of the biomarkers. The true (target) values of the QC samples will be determined by the baseline values during validation of the matrix pool. In general, a target value of 25% for R.E. and C.V. is recommended. More lenient acceptance criteria than the 4-6-25 rule recommended in this paper can be justified based on statistical rationale developed from experimental data.

8. Detection of antibodies to macromolecules

Immunogenicity is an important property distinguishing most biologic products from most small drug molecules. In recent years, recombinant human (rh) protein therapeutics have been an intense area of investigation and a few of these compounds have already reached the market, while many others are in various stages of development. The spectrum of biotechnological products encompasses hormones (growth hormone, insulin), enzymes (DNnase, asparaginase), cytokines (interleukins -1, -2, -11, interferons), growth factors (G-CSF, GM-CSF), clotting factors (factor V111), vaccines (hepatitis B), thrombolytics (tPA), monoclonal antibodies (OKT3) and novel fusion proteins such as PIXY321 and Enbrel [67,68]. These products are either recombinant versions of human proteins, analogs of human proteins containing minor changes in their primary sequence and/or altered post-translational modifications, or are re-engineered novel proteins. The administration of these recombinant proteins to animals and humans may result in their recognition by the host's immune system as 'non-self', resulting in an antibody response.

In vivo production of antibodies in response to treatment can potentially interfere with antibody and activity-based assays for the proteins in biological matrices, thereby, preventing the accurate quantitation of circulating drug. In preclinical and clinical studies, antibodies produced in response to treatment with the candidate drug may alter the pharmacokinetics of the drug, neutralize its pharmacodynamic effects, and confound the interpretation of the safety data [15]. In addition, the presence of pre-existing antibodies (autoantibodies) to endogenous proteins can further complicate the assessment of safety of the drug. An immune response against the therapeutic protein may also result in abrogation of the biological activity of the endogenous equivalent protein, if cross-reactivity occurs. Antibodies may also interfere in imaging and diagnostic procedures utilizing antibodies; for example, human anti-mouse antibodies (HAMA) in the serum of patients treated with a murine antibody-based therapeutic may interfere in diagnostic assays using murine monoclonal antibodies.

8.1. Considerations in assay development

Several factors may contribute to immunogenicity and should be considered during method development. These include product characteristics, such as the impurity profile, post-translation modifications and fragments and aggregates of the administered protein. Additionally, timing of the study sample collection should be considered, since the presence of high levels of circulating macromolecules or macromolecule complexes, may interfere with the assay. In some instances, cross-reactivity of the antibody detected may need to be evaluated against sub-classes of the macromolecule (e.g. interferon), since the binding affinities of each sub-type may vary greatly and may influence the interpretation of the results. Other factors that should be considered are the genetic make-up of the patient population (allotypic determinants as in RF factor) and their disease states (example autoimmune disease). Pre-existing antibodies can also limit the utility of another antibody of the same species in assay development (e.g. as in detection antibody). Concomitant medications can also be potential interferents in the assay.

8.2. Assay format

Detection of antibodies is based on the antibody–antigen reaction and the endpoint of the reaction can be detected by several techniques such as precipitin reactions, agglutination, competitive immunoassays (e.g. RIA), noncompetitive immunoassays (e.g. ELISA) and Western blot. The selection of an assay format is dependent on the purpose of the assay and availability of appropriate reagents. To detect high affinity antibodies selectively, thereby conferring high specificity, a competitive assay is an appropriate format; however, if a high sensitivity assay is needed to detect antibodies of all affinities, a non-competitive ELISA format is preferred. Method selection should also evaluate the potential loss of epitopes when the protein is directly adsorbed on the surface of the plate, the possibility of circulating antigen–antibody complexes, circulating antigen aggregates, and the source species of antibodies for capture and detection.

8.3. Assay qualification

Antibody assays have generally not been applied truly quantitatively, due to the difficulties in obtaining well-characterized, species-specific, polyclonal anti-drug antibody reference materials to be used as calibrators. Even though these assays are not quantitative, many of the validation parameters for quantitative assays should be considered. Assay development should include optimization of incubation times and temperatures and antigen coating concentration, as well as studies of the binding of the protein antigen to the microtiter plate surface to verify randomness of epitope presentation and immunoreactivity of epitopes. Qualification experiments should also include extensive evaluation of intermediate precision, to determine assay variance due to variation in the day, plate, differing analysts, lots of plates and reagents, and position on the plate. Assay specificity evaluation should include assessment of any non-specific binding of the antibody to the microtiter plate, effects of concomitantly administered drugs, and of the administered protein, endogenous protein analogs, and antigen–antibody complex or cross-reacting antibodies that may be present in the sample under evaluation. Additional experiments include freeze–thaw and storage stability of both samples and reagents. The qualification experiments for the detection of antibody response should be conducted at several dilutions of the antibody-positive and antibody-negative control samples. The recommended acceptance criterion for the evaluation of intermediate precision is $< 25\%$ C.V. and, for specificity and stability, accuracy of $100 \pm 25\%$ for absorbance (in an ELISA) of the positive control value in the pseudo-linear portion of the dilution curve.

Ideally, QC samples should consist of the analyte of interest (species and antigen specific antibodies) in the matrix to be analyzed. In the early stages of macromolecule drug development, no species-specific QC samples for the detection of antibodies are available, since no animal species has been exposed to the compound. Reagent controls, consisting of affinity-purified protein-specific antibodies, can be used to demonstrate

consistent assay performance over time, several lots of plate preparations, reagent preparation and different lots of enzyme substrate. In clinical trials for new biological entities, no antibody-positive control materials are available and therefore, serum from primates exposed to drug can serve as potential positive controls for assays, provided that the assay system can be demonstrated clearly to detect the primate antibody by use of an anti-human immunoglobulin, as would be used in human studies. A pool of human plasma from healthy volunteers can serve as a negative control.

8.4. Development of antibody negative cut-off

In preparation for clinical study support, it is important, to describe the volunteer and patient population serum sample background absorbance readings to establish a cut-off value to distinguish between antibody-negative and -positive classifications. Specificity studies should include evaluation of at least 25 individual serum samples from healthy volunteers and 25 serum samples from appropriate patients to determine the frequency of false positive responses.

However, the most important application for these assays is the identification of true positive antibody responses. In the preclinical setting, data for antibody-negative controls are derived by analysis of serum samples from untreated animals (frequently primates). The response data from such samples tends to cluster closely together, and a negative control level in an ELISA can be defined as mean absorbance ± 3 S.D. at a given dilution factor(s). However, when specificity studies are conducted in the human volunteer and patient populations, much greater variability in background values is often noted, making definition of a negative cut-off value and range much more difficult. In a clinical setting where one is attempting to detect any antibody response, the cut-off should initially be set such that there are no, or minimal, false negatives and all immune responses could be detected. The cost of this approach will, of course, be the detection of an increased number of false positives. The assay validation scheme should include a process to distinguish true responses from false positives. This is particularly important since autoantibodies against various proteins may be present in otherwise healthy individuals. Several approaches can be taken to elucidate whether an apparent antibody response is truly positive, including an alternative method for detecting the antibody (e.g. Western blotting). In some cases, where antibody response is evaluated against several different antigens, true positives may be distinguished from false positives by cross-reactivity patterns. Finally, examination of the response prior to drug administration, and the change in this response over time, with continued exposure to the agent, will normally distinguish between true and false positives. In addition, in vitro neutralizing activity assays and clinical effects (in vivo neutralization) provide further support for the presence of drug specific antibodies.

9. Conclusions

Despite the widespread availability of mass spectrometry-based bioanalytical procedures for low molecular weight drug candidates, immunoassays remain of critical importance for certain bioanalytical applications in support of drug development. This is particularly true for the quantitation of protein therapeutic drug candidates, for biomarkers/surrogate markers and for the assessment of antibody responses to treatment with macromolecules. Current proposed guidelines from regulatory agencies for the validation of bioanalytical methods focus on chromatographic procedures, and do not adequately address those aspects of immunoassays which differentiate them from chromatographic assays.

Many of these differences emanate from the facts that immunoassays depend on the reaction of a key biological reagent, namely an antibody, and that the analytes assayed are frequently proteins. These factors result in a number of issues that require special attention for validation of immunoassays, as enumerated below:

1. Key reagents, such as the antibody directed against the analyte, frequently are not commercially available, resulting in longer assay development times, and also need more strin-

gent storage conditions than many chemical reagents used in chromatographic assays.
2. Calibration curves for immunoassays, which are preferentially established in the same matrix as that of the study samples, are inherently nonlinear and need special attention to select the best model fit to describe the data. Most frequently this is adequately and correctly provided by a four-parameter logistic model.
3. 'Anchoring' points above and below the validated range of the immunoassay may be used to optimize the fit of the calibration curve data to the selected model.
4. Due to the core reaction between analyte and a biological reagent (antibody), immunoassays may have more inherent imprecision than chromatographic assays. Thus, we recommend that more lenient target acceptance criteria of 20% (25% at the limits of quantitation) for accuracy (mean bias) and precision be adopted for immunoassay applications, including biomarker and antibody assays.
5. Specificity of immunoassays is critically dependent on the specificity of the antibody directed against the analyte. However, care must also be taken to ensure the absence of nonspecific interferences related to the matrix (i.e. matrix effects).
6. Special challenges arise in the immunoassay of endogenous analytes, such as naturally occurring proteins, steroids, prostaglandins, etc. Analyte-free matrices for establishment of calibration curves are sometimes difficult to obtain, so that alternative matrices may need to be used, and difference/addition methods applied to analysis of validation samples and QC samples.
7. Particular consideration needs to be given to assays for biomarkers as pharmacodynamic indicators of disease progression or amelioration upon treatment with candidate drugs. In these cases, attention needs to be paid to appropriate selection of calibration standards and QC samples, so that the latter reflect the range of biomarker molecule concentration or activity present in the pathologic state under study.
8. Assays for evaluating antibody responses to treatment with candidate protein therapeutic molecules are normally semi-quantitative in nature, so that complete validation is usually not possible. Particularly important for these assays, however, is the establishment of a titer cut-off value to permit the classification of clinical responses as antibody-negative or -positive.
9. The use of two-sided 90% confidence intervals for the total error of validation samples (SFSTP approach) is recommended in considering acceptance/rejection of an immunoassay procedure resulting from validation experiments. Such an approach reflects more directly the performance of individual assays when using the '4-6-25 rule' for acceptance of in-study runs and will result in fewer rejected in-study runs than the current procedure that compares point estimates of observed bias and precision with the 15% (20%) target acceptance criteria.

Acknowledgements

The authors gratefully acknowledge R. Dillard, Department of Preclinical Statistics, at Searle for his critical review, advice and constructive comments on this manuscript.

Appendix A. Glossary of terms

For the purpose of standardizing terminology, we recommend use of the following definitions for immunoassays.

Accuracy	A concept that expresses the closeness of agreement between a measured test result and its theoretical true value. In this paper the

	term accuracy refers to systematic error (mean bias)
Batch	A set of standard curve (calibrators), validation samples, and/or quality control samples, and/or study samples that is analyzed together as a single group using the method to be validated. Batch is synonymous with run
Bias	Systematic difference between measured test results and the theoretical true value. Bias is expressed either as a relative difference (% relative error, % R.E.) or as a ratio (% recovery).
Calibration curve	The relationship between analyte concentration in the standard samples (calibrators) and the binding response. Calibration curve is synonymous with standard curve
Calibrator	An aliquot of matrix spiked with the analyte of interest at a predetermined concentration using a well-characterized reference material. Calibrator is synonymous with standard
Coefficient of variation (C.V.)	A quantitative measure of precision expressed relative to the mean result (also referred to as the Relative Standard Deviation). See R.S.D.
Dilutional linearity	A special case of parallelism in which the sample being analyzed at multiple dilutions has a stated nominal concentration (e.g. validation sample). Dilutional linearity is assessed during pre-study validation by examining the % R.E. of the sample at multiple dilutions
EC_{50}	An abbreviation for the 'concentration necessary to produce a response of 50%'. For a competitive assay, the EC_{50} is the concentration of the analyte that is necessary to produce a 50% displacement of the Tracer. For a noncompetitive assay, the EC_{50} is the concentration of analyte necessary to produce a response of 50% or one-half of the observed maximum binding
Four-parameter logistic (4PL) function	A versatile function that is recognized as the 'reference standard' for fitting the mean concentration–response for immunoassays. The function is defined by the equation $$E(Y) = D + \frac{(A-D)}{[1+(X/C)^B]}$$ where $E(Y)$ is the expected response; X, concentration; A, response at zero concentration; D, response at infinite concentration; C, concentration resulting in a response halfway between A and D (ED_{50}); B, slope parameter that is typically near 1.0
Intermediate precision	Precision of repeated measurements within a laboratory taking into account all relevant sources of variation affecting the results (e.g. runs, days, analysts, equipment, reagents, etc.). Intermediate precision is also termed inter-assay or inter-batch precision
Limit of detection (LOD)	Lowest concentration of analyte for which the response can be reliably distinguished from background noise
Linearity	Ability of the analytical method within a specified concentration range to obtain test results that are proportional to the concentration of the analyte in the test sample

Lower limit of quantitation (LLOQ)	The lowest concentration of analyte that has been demonstrated to be measurable with acceptable levels of accuracy (mean bias) and precision
Nominal concentration	A stated or theoretical concentration that may or may not differ from the true concentration
Parallelism	Parallelism of the calibration (standard) curve and the concentration–response for analyte in a test sample (obtained after in vivo administration of the compound of interest) is a necessary condition for validity of an analytical result when the mean response is defined as a function of log dose (e.g. the sigmoidal concentration–response relationships observed in immunoassays). Parallelism can be assessed in combination with other validity requirements by analyzing the test sample at multiple dilutions and then examining for a constant recovery of the analyte (e.g. as measured by the % R.S.D. among results at different dilutions)
Power-of-the-mean (POM) function	A flexible function for modeling the response–error relationship observed in most immunoassays. The function is defined by the equation $S(Y) = \sigma E(Y)^{\theta}$ with $S(Y)$ the S.D. in replicate response values; $E(Y)$, expected mean response; σ, proportionality constant that must be estimated for each assay run, and θ, shaping parameter that is assumed to be stable across runs
Precision	A quantitative measure (usually the S.D. or the C.V.) of the random variation between repeated measurements from multiple sampling of the same homogenous sample under specified conditions
Precision profile	A graphical representation of the relationship between the coefficient of variation (C.V.) in analytical measurements of a sample under specified conditions and the nominal concentration of analyte in the sample
Pre-study validation	Procedures used prior to the analysis of study samples to establish that an analytical method is suitable for its intended application
Pre-validation	Assay development experiments conducted prior to formal validation in order to enhance immunoassay performance
Quality control samples	Biological matrix samples spiked with the analyte of interest at predetermined concentrations. Alternatively, the analyte may be present endogenously in the biological matrix. A set of quality control samples is assayed with each analytical run during the analysis of in-study samples. Quality control samples are used to evaluate the acceptability of analytical runs
Recovery	A quantitative measure of the closeness of an observed result to its theoretical true value, expressed as a percent of the nominal (theoretical) concentration $\%\text{Recovery} = \left(\frac{\text{Observed}}{\text{Nominal}}\right) \times 100$

Relative error (R.E.)	A quantitative measure of the closeness of an observed result to its theoretical true value, expressed as a percent relative difference from the nominal (theoretical) concentration $$\%\text{R.E.} = \left[\left(\frac{\text{Observed}}{\text{Nominal}}\right) - 1\right] \times 100$$
Relative standard deviation (R.S.D.)	A quantitative measure of precision (also referred to as the coefficient of variation) expressed relative to the observed or theoretical (nominal) mean value $$\%\ \text{R.S.D.} = \left[\frac{\text{S.D.}}{\text{Mean}}\right] \times 100$$ For repeatability, the S.D. is computed from replicate analyses within a single validation run. For intermediate precision, the S.D. is computed from replicate analyses over multiple validation runs within the same laboratory
Repeatability	Precision of repeated measurements within the same analytical run (or batch). Repeatability is also termed intra-assay or intra-batch precision
Response error relationship (RER)	The relationship between the S.D. in replicate response values (e.g. counts per minute) and the mean response
Run	A run represents a set consisting of standard curve (calibrators) and validation samples and/or quality control samples and/or study samples that are analyzed together as a single group using the method to be validated. Run is synonymous with batch
Total error	A concept that expresses the closeness of agreement between a measured test result and its theoretical true value. The term total error describes a combination of systematic (mean bias) and random (precision) error components. In other publications, the term total error is also referred to as accuracy e.g. ISO definition
Upper limit of quantitation (ULOQ)	The highest concentration of analyte that has been demonstrated to be measurable with stated levels of accuracy (mean bias) and precision
Validated range	The interval of analyte concentrations over which the assay method has been validated. This interval includes concentrations from LLOQ to the ULOQ
Validation samples	Biological matrix samples spiked with the analyte of interest at predetermined concentrations. Alternatively, the analyte may be present endogenously in the biological matrix. A set of validation samples is used during pre-study validation to assess accuracy (mean bias) and precision

Appendix B. Validation sample statistics

For each validation sample concentration, the random effects model assumes that the measured result

(z_{ij}) from the jth replicate of the ith batch run is described by the equation

$$z_{ij} = \mu_M + b_i + \varepsilon_{ij} \qquad (i = 1, 2, \ldots t; j = 1, 2, \ldots n)$$

where t is the total number of runs, n is the number of replicates per run, μ_M is the unknown true analytical mean for the method, and b_i and ε_{ij} denote random effects that explain the variability in the ith run and jth replicate. Random effects are assumed to be normally and independently distributed with mean zero and variance components equal to σ_b^2 for between runs and σ_w^2 for replicates within a run. Statistical analysis of the measured z_{ij} values leads to estimates of the unknown parameters μ_M, σ_w^2, and σ_b^2, and other assay parameters derived from them. The following summary provides formulae for parameter estimation, based on a one-way analysis of variance (ANOVA), when the number of replicates are equal for all runs. Alternative methods of estimation are also acceptable [61,62]. Formulae for the unbalanced case (i.e. unequal number of replicates) can be found in statistical texts [60].

Analytical parameters		
'Known'	Nominal (target) concentration	μ_T
Unknown	'True' mean of analytical method	μ_M
	'True' intra-run variance component	σ_w^2
	'True' inter-run variance component	σ_b^2

Statistics		
Descriptive statistics	Sample mean for ith run	$\bar{z}_{i.} = \frac{1}{n}\sum_{j=1}^{n} z_{ij}$
	Sample S.D. for ith run	$s_i = \sqrt{\frac{1}{(n-1)}\sum_{j=1}^{n} (z_{ij} - \bar{z}_{i.})^2}$
	Overall sample mean	$\bar{z}_{..} = \frac{1}{t}\sum_{i=1}^{t} \bar{z}_{i.}$
ANOVA mean square errors	Within-run	$\mathrm{MS_w} = \frac{1}{t}\sum_{i=1}^{t} s_i^2$
	Between-run	$\mathrm{MS_b} = \frac{n}{(t-1)}\sum_{i=1}^{t} (\bar{z}_{i.} - \bar{z}_{..})^2$
	Total	$\mathrm{MS_t} = \frac{1}{(tn-1)}\sum_{i=1}^{t}\sum_{j=1}^{n} (\bar{z}_{ij} - \bar{z}_{..})^2$

Variance components	Repeatability (within-run)	$s_w^2 = MS_w$
	Between-run	$s_b^2 = \frac{(MS_b - MS_w)}{n}$
	Intermediate precision (total random error)	$s_{IP}^2 = s_w^2 + s_b^2$
Degrees of freedom	Between-run mean square error (MS_b)	$\eta = t - 1$
	Intermediate precision (s_{IP}^2) (Satterthwaite approximation)	$v = \frac{[(n-1)\cdot MS_w + MS_b]^2}{\left[\left(\frac{n-1}{t}\right)\cdot MS_w^2 + \left(\frac{1}{t-1}\right)\cdot MS_b^2\right]}$
Assay characteristics	Mean bias (% R.E.)	$\%R.E. = 100 \times \left(\frac{\bar{\bar{z}}_{..} - \mu_T}{\mu_T}\right)$
	Repeatability (% C.V.)	$\%C.V. = 100 \times \left(\frac{s_w}{\mu_T}\right)$
	Intermediate precision (% C.V.)	$\%C.V. = 100 \times \left(\frac{s_{IP}}{\mu_T}\right)$
Other statistics	αth percentile of Student's t-distribution with k degrees of freedom	$t_{k,\alpha}$
	αth percentile of χ^2 distribution with k degrees of freedom	$\chi^2_{k,\alpha}$
	Estimated variance of $\bar{\bar{z}}_{..}$	$\mathrm{var}(\bar{\bar{z}}_{..}) = \frac{MS_b}{tn}$
	Two-sided 95% confidence limits for mean bias (% R.E.)	$(100/\mu_T)\cdot[(\bar{\bar{z}}_{..} - \mu_T) \pm t_{\eta,\,0.975}\cdot\sqrt{\mathrm{var}(\bar{\bar{z}}_{..})}]$
	Lower two-sided 95% confidence limit for intermediate precision (% C.V.)	$\frac{(100/\mu_T)\cdot v\cdot s_{IP}^2}{\chi^2_{v,0.975}}$

Upper two-sided 95% confidence limit for intermediate precision (% C.V.)	$\frac{(100/\mu_T) \cdot v \cdot s_{IP}^2}{\chi^2_{v,0.025}}$

References

[1] V.P. Shah, K.K. Midha, S. Dighe, et al., Pharm. Res. 9 (1992) 588–592.
[2] Food and Drug Administration, Fed. Reg. 60 (1995) 11260–11262.
[3] Food and Drug Administration, Fed. Reg. 62 (1997) 27464–27467.
[4] C. Hartmann, D.L. Massart, R.D. McDowall, J. Pharm. Biomed. Anal. 12 (1994) 1137–1343.
[5] E. Chapuzet, N. Mercier, S. Bervos-Martin, et al., S.T.P. Pharma Prat. 7 (1997) 169–194.
[6] C. Hartmann, J. Smeyers-Verbeke, D.L. Massart, et al., J. Pharm. Biomed. Anal. 17 (1998) 193–218.
[7] J.W.A. Findlay, I. Das, J. Clin. Ligand Assay 20 (1997) 49–55.
[8] J.W.A. Findlay, in: H.H. Blume, K.K. Midha (Eds.), Bio-International, vol. 2, Bioavailability, Bioequivalence and Pharmacokinetic Studies, Medpharm Scientific Publishers, Stuttgart, 1995, p. 361.
[9] J.W.A. Findlay, I. Das, J. Clin. Ligand Assay 21 (1998) 249–253.
[10] R.O. Kringle, R.C. Khan-Malek, Proceedings of the Biopharmaceutical Section of the American Statistical Association, 1994, pp. 510–514.
[11] W.C. Smith, G.S. Sittampalam, J. Biopharm. Stat. 8 (1998) 509–532.
[12] R.O. Kringle, Pharm. Res. 11 (1994) 556–560.
[13] Food and Drug Administration, Fed. Reg. 64 (1999) 517.
[14] E.P. Diamandis, T.K. Christopoulos, Clin. Chem. 37 (1991) 625–638.
[15] P.K. Working, Pharmaceutical Biotechnology, in: B.L. Ferraiolo, et al. (Eds.), Protein Pharmacokinetics and Metabolism, vol. 1, Plenum, New York, 1992, pp. 73–92.
[16] J.D. Gilbert, J. Clin. Ligand Assay 20 (1997) 23–31.
[17] Y. Sugiyama, M. Hanano, Pharm. Res. 6 (1989) 192–202.
[18] C.A. Gloff, L.Z. Benet, Adv. Drug Deliv. Rev. 4 (1990) 359–386.
[19] R.J. Wills, B.L. Ferraiolo, Clin. Pharmacokinet. 23 (1992) 406–414.
[20] S. Toon, Eur. J. Drug Metab. Pharmacokinet. 21 (1996) 93–103.
[21] A.B. Chen, D.L. Baker, B.L. Ferraiolo, Pharmacokinetics and pharmacodynamics, in: P.D. Garzone, W.A. Colburn, M. Mokotoff (Eds.), Peptides, Peptoids, and Proteins, vol. 3, Harvey Whitney, Cincinnati, OH, 1991, pp. 54–71.
[22] B.L. Ferraiolo, M.A. Mohler, Pharmaceutical biotechnology, in: B.L. Ferraiolo, et al. (Eds.), Protein Pharmacokinetics and Metabolism, vol. 1, Plenum, New York, 1992, pp. 1–33.
[23] T. Chard, An introduction to radioimmunoassay and related techniques, in: R.H. Burdon, M. van Knippenberg (Eds.), Laboratory Techniques in Biochemistry and Molecular Biology, vol. 6, Part 2, third ed., Elsevier, New York, 1987, pp. 38–39.
[24] M. Wadhwa, R. Thorpe, Cytokine 9 (1997) 791–793.
[25] S. De Kossodo, V. Houba, G.E. Grau, J. Immunol. Methods 182 (1995) 107–114.
[26] A. Ledur, C. Fitting, B. David, et al., J. Immunol. Methods 186 (1995) 171–179.
[27] D. Rodbard, G.R. Frazier, Methods Enzymol. 37 (1975) 3–22.
[28] M.C. Haven, P.J. Orsulak, L.L. Arnold, G. Crowley, Clin. Chem. 33 (1987) 1207–1210.
[29] S. Braggio, R.J. Barnaby, P. Grossi, et al., J. Pharm. Biomed. Anal. 14 (1996) 375–388.
[30] J.W.A. Findlay, R.F. Butz, E.C. Jones, Clin. Chem. 27 (1981) 1524–1527.
[31] T. Chard, An introduction to radioimmunoassay and related techniques, in: R.H. Burdon, van Knippenberg (Eds.), Laboratory Techniques in Biochemistry and Molecular Biology, vol. 6, Part 2, third ed., Elsevier, New York, 1987, pp. 175–193.
[32] M.E. Gosse, M.A. Manocchia, Drug Inform. J. 30 (1996) 991–1001.
[33] A.R. Buick, M.S. Doig, S.C. Jeal, et al., J. Pharm. Biomed. Anal. 8 (1990) 629–637.
[34] P. Carter, Clin. Chem. 24 (1978) 362–364.
[35] J. Das Sarma, C. Duttagupta, E. Ali, et al., J. Immunol. Methods 184 (1995) 7–14.
[36] J.D. Wright, F.D. Boudinot, M.R. Ujhelyi, Clin. Pharmacokinet. 30 (1996) 445–462.
[37] C.K. Svensson, M.N. Woodruff, J.G. Baxter, et al., Clin. Pharmacokinet. 11 (1986) 450–469.
[38] W. Kuhnz, M. Pfeffer, G. Al-Yacoub, J. Steroid Biochem. 35 (1990) 313–318.

[39] K. Tamura, T. Fujimura, S. Sakuma, et al., Biochem. Biophys. Res. Commun. 202 (1994) 437–443.
[40] W. Borth, A. Urbanski, R. Prohaska, et al., Blood 75 (1990) 2388–2395.
[41] Y. Du, B. Ni, M. Glinn, et al., J. Neurochem. 69 (1997) 299–305.
[42] D.W. Leung, S.A. Spencer, G. Cachianes, et al., Nature 330 (1987) 537–543.
[43] Y. Ma, H.E. Hurst, R. Fernandex-Botran, J. Pharmacol. Exp. Ther. 279 (1996) 340–350.
[44] J. Zapf, Eur. J. Endocrinol. 132 (1995) 645–654.
[45] F.B. Diamond Jr., D.C. Eichler, G. Duckett, et al., Biochem. Biophys. Res. Commun. 233 (1997) 818–822.
[46] A.R. Mire-Sluis, R. Gaines-Das, R. Thrope, J. Immunol. Methods 186 (1995) 157–160.
[47] S.C. Piscitelli, W.G. Reiss, W.D. Figg, et al., Clin. Pharmacokinet. 32 (1997) 368–381.
[48] S. Fisker, L. Ebdrup, H. Orskov, Scand. J. Clin. Lab. Invest. 58 (1998) 373–382.
[49] W.H. Daughaday, M. Kapadia, I. Mariz, J. Lab. Clin. Med. 109 (1987) 355–363.
[50] P. Bank, R.C. Baxter, W.F. Blum, et al., J. Endocrinol. 143 (1994) C1–C2.
[51] B.N. Rehlaender, M.J. Cho, Pharm. Res. 15 (1989) 1652–2022.
[52] W. Penninckx, C. Hartmann, D.L. Massart, et al., J. Anal. At. Spectrom. 11 (1996) 237–246.
[53] R.A. Dudley, P. Edwards, R.P. Ekins, et al., Clin. Chem. 31 (1985) 1264–1271.
[54] R.J. Carroll, D.B.H. Cline, Biometrika 75 (1988) 35–43.
[55] D.J. Finney, P. Phillips, Appl. Stat. 26 (1978) 312–320.
[56] D.J. Finney, Statistical Method in Biological Assay, third ed., Charles Griffin, London, 1978.
[57] M. Dividian, R.J. Carroll, J. Am. Stat. A. 82 (1987) 1079–1091.
[58] H.T. Karnes, C. March, J. Pharm. Biomed. Anal. 9 (1991) 911–918.
[59] D.M. Rocke, G. Jones, Technometrics 39 (1997) 162–170.
[60] S.R. Searle, G. Casella, C.E. McCulloch, Variance Components, Wiley, New York, 1992.
[61] D. Rodbard, Clin. Chem. 20 (1974) 1255–1270.
[62] S.C. Chow, S.K. Tse, Stat. Med. 10 (1991) 1543–1553.
[63] C. Hartmann, J. Smeyers-Verbeke, W. Penninckx, et al., Anal. Chem. 67 (1995) 4494–4499.
[64] P. Hubert, P. Chiap, J. Crommen, et al., Anal. Chim. Acta 391 (1999) 135–148.
[65] W.A. Colburn, J. Clin. Pharmacol. 37 (1988) 355–362.
[66] J.W. Lee, J.D. Hulse, W.A. Colburn, J. Clin. Pharmacol. 35 (1995) 464–470.
[67] The Pharmaceutical Research and Manufacturers of America (PhRMA), 1998 Survey, New Medicines in Development Biotechnology, 1998, pp. 1–36.
[68] A.S. Lubiniecki, S.A. Vargo, in: A.S. Lubiniecki, S.A. Vargo (Eds.), Regulatory Practice for Biopharmaceutical Production, Wiley, New York, 1994, pp. 1–12.