Review

Recommendations for the validation of immunoassays used for detection of host antibodies against biotechnology products

Gopi Shankar[a], Viswanath Devanarayan[b], Lakshmi Amaravadi[c], Yu Chen Barrett[d], Ronald Bowsher[e], Deborah Finco-Kent[f], Michele Fiscella[g], Boris Gorovits[h], Susan Kirschner[i,1], Michael Moxness[j], Thomas Parish[k], Valerie Quarmby[l], Holly Smith[m], Wendell Smith[n], Linda A. Zuckerman[o], Eugen Koren[p,*]

[a] *Clinical Pharmacology Sciences, Centocor Research & Development Inc., Radnor, PA 19087, USA*
[b] *Global Exploratory Statistics, Abbott Laboratories, Parsippany, NJ 07054, USA*
[c] *Preclinical & Clinical Development Sciences, Biogen Idec, Cambridge, MA 02142, USA*
[d] *Clinical Discovery, Bristol-Myers Squibb Company, Princeton, NJ 08543, USA*
[e] *Millipore, BioPharma Services Division, 15 Research Park Dr., St. Charles, MO 63304, USA and B2S Consulting, Beech Grove, IN 46107, USA*
[f] *Drug Safety Research and Development, Pfizer, Groton, CT 06340, USA*
[g] *Human Genome Sciences, 14200 Shady Grove Road, Rockville, MD 20850, USA*
[h] *Drug Safety and Metabolism, Wyeth, Pearl River, NY 10965, USA*
[i] *Office of Biotechnology Products, CDER, Food and Drug Administration, Bethesda, MD 20892, USA*
[j] *Clinical Immunology, Medical Sciences, Amgen Inc., Thousand Oaks, CA 91320, USA*
[k] *Analytical Science Department, Procter & Gamble Pharmaceuticals, Norwich, NY 13815, USA*
[l] *BioAnalytical Research & Development, Genentech, South San Francisco, CA 94080, USA*
[m] *Investigative Toxicology, Eli Lilly and Company, Greenfield, IN 46140, USA*
[n] *B2S Consulting, 1305 Chapman Drive, Greenfield, IN 46140, USA*
[o] *PreClinical Development Department, ZymoGenetics Inc., Seattle, WA 98102, USA*
[p] *Bioanalytical R&D, Abbott Vascular Inc., Santa Clara, CA 95054, USA*

ARTICLE INFO

ABSTRACT

Most biological drug products elicit some level of anti-drug antibody (ADA) response. This antibody response can, in some cases, lead to potentially serious side effects and/or loss of efficacy. In humans, ADA often causes no detectable clinical effects, but in the instances of some therapeutic proteins these antibodies have been shown to cause a variety of clinical consequences ranging from relatively mild to serious adverse events. In nonclinical (preclinical) studies, ADA can affect drug exposure, complicating the interpretation of the toxicity, pharmacokinetic (PK) and pharmacodynamic (PD) data. Therefore, the immunogenicity of therapeutic proteins is a concern for clinicians, manufacturers and regulatory agencies.

In order to assess the immunogenic potential of biological drug molecules, and be able to correlate laboratory results with clinical events, it is important to develop reliable laboratory test methods that provide valid assessments of antibody responses in both nonclinical and clinical studies. For this, method validation is considered important, and is a necessary bioanalytical component of drug marketing authorization applications. Existing regulatory guidance documents dealing with the validation of methods address immunoassays in a limited manner, and in particular lack information on the validation of immunogenicity methods. Hence this article provides scientific recommendations for the validation of ADA immunoassays. Unique validation performance characteristics are addressed in addition to those provided in existing regulatory documents pertaining to bioanalyses. The authors recommend experimental and statistical approaches for the validation of immunoassay performance characteristics; these recommendations should be considered as examples of best practice and are intended to foster a more unified approach to antibody testing across the biopharmaceutical industry.

© 2008 Elsevier B.V. All rights reserved.

---

* Corresponding author. Tel.: +1 408 845 1181; fax: +1 408 845 3689.
E-mail address: eugen.koren@av.abbott.com (E. Koren).
[1] The views expressed here represent those of the author and not necessarily those of the Food and Drug Administration.

Contents

1. Introduction

Biopharmaceutical products differ from conventional small molecule drugs in that they are larger in size (i.e., typically >1–3 kDa), are biopolymers of amino acids, carbohydrates or nucleic acids, and are often manufactured by human or non-human cells or microorganisms. Because of these differences, biopharmaceuticals have a greater potential for inducing immune responses [1,2]. The immunogenic potential of a biopharmaceutical is governed by product-intrinsic factors (e.g., species-specific epitopes, degree of foreignness, glycosylation status, extent of aggregation or denaturation, impurities and formulation), product-extrinsic factors (e.g., route of administration, acute or chronic dosing, pharmacokinetics, and existence of endogenous equivalents), and patient-specific factors (e.g., autoimmune disease, immunosuppression, and replacement therapy) [3].

While often benign, the induction of anti-drug antibodies (ADA) can result in adverse clinical sequelae including hypersensitivity or autoimmunity, and altered pharmacokinetics (for example, drug neutralization, abnormal biodistribution, or enhanced drug clearance rates, potentially resulting in altered efficacy of the treatment). Immune response caused by drug treatment is, therefore, a major safety and efficacy concern for regulatory agencies, drug manufacturers, clinicians, and patients [4]. Consequently, the United States Food & Drug Administration (FDA) as well as regulatory authorities in the European Union, Canada, Japan and Australia require that ADA be evaluated and correlated with any pharmacological and/or toxicological observations [5–7].

Correlations between immunogenicity and clinical sequelae depend on an objective detection and characterization of antibodies against biological therapeutics in nonclinical and clinical studies. Hence bioanalytical methods used for immunogenicity testing should be properly developed and validated before testing is initiated with study samples. Recommendations on method development and optimization [8,9] and strategies for the detection and characterization of ADA are provided in prior publications [10,11]. *Validation is defined as a process of demonstrating, through the use of specific laboratory investigations, that the performance characteristics of an analytical method are suitable for its intended analytical use* [12,13]. In the case of ADA detection methods, validation constitutes proof that the assay will reliably (i.e., consistently and reproducibly) detect low amounts of drug-specific antibodies in a complex biological matrix, such as serum or plasma. Validation should be performed in the 'pre-study' phase (i.e., before clinical or nonclinical study samples are analyzed), but it is equally important to demonstrate that the assay remains valid or 'in control', during the 'in-study' phase (i.e., when clinical or nonclinical study samples are analyzed) as well; only then can the results of test samples be deemed acceptable.

The intent of this paper is to present the performance characteristics important for the validation of ADA immunoassays and to recommend appropriate and objective methodological

approaches of validation. *The recommendations in this paper should be considered as examples of best practice; alternate methodological approaches may also be acceptable, as long as scientific rationale and objectivity are maintained and uncompromised irrespective of assay-specific practical considerations.* It is advised that the acceptability of alternative approaches be discussed with regulatory authorities. Cellular function-based neutralizing ADA (NAb) bioassays and assays for cell-mediated immune responses are outside the scope of this paper.

2. Methodology

2.1. ADA detection

Clinical and nonclinical immunogenicity is generally evaluated via detection and characterization of treatment-induced ADA. A number of analytical formats and detection methods are available for the detection of ADA, including enzyme-linked immunosorbent assay (ELISA), radioimmunoassay (RIA) or radioimmunoprecipitation assay (RIPA), surface plasmon resonance (SPR), and electrochemiluminescence (ECL). Each of these formats has its relative merits and limitations, and these have been discussed in recent publications [8,14]. Irrespective of the analytical format, once a prototypic assay is developed and optimized, it needs to undergo formal validation to ensure that the method will be suitable for its intended purpose. It is therefore anticipated that the recommendations for validation herein are generally applicable to the majority of anti-drug antibody immunoassays. Researchers should consider unique aspects of their assay systems to determine the appropriate validation scheme for their methods.

Four types of methods are generally performed for the detection and characterization of ADA for clinical studies. ADA detection assays include screening and specificity confirmation (confirmatory) assays, whereas characterization assays typically include titration and neutralizing antibody detection [10]. For application with study samples, ADA assays require establishment of two critical decision parameters a 'screening cut point' for the screening assay and a 'specificity cut point' for the specificity confirmation assay. The screening cut point enables the classification of ADA results as either antibody negative or reactive samples (reactive samples are sometimes referred to as "potential positive" samples). The reactive samples undergo further characterization to categorize them as positive versus negative ("non-specific reactive") samples by means of a specificity confirmation assay (e.g., signal inhibition by competition with drug). The magnitude of signal inhibition required for a sample to be deemed as containing drug-specific ADA is termed as specificity cut point and should be experimentally established.

Immunoassays used for ADA detection are generally non-quantitative assays (sometimes referred to as quasi-quantitative) because standardized, species-specific (especially human) polyclonal ADA reference materials to use as calibrators are unavailable [15]. It is very unlikely that positive controls, generally developed in-house (monoclonal antibodies or hyperimmune sera derived polyclonal antibodies), will be identical to all ADA detected in subjects. If reporting of ADA levels in mass units is intended, acceptable parallelism should be demonstrated between a standard calibrator (positive control) and test samples in order to determine ADA concentration with acceptable accuracy [8,10]. In the absence of demonstrated parallelism between the sample and calibrator, accuracy is questionable. Another approach to assess antibody level is titration, which is also susceptible to potential lack of dilutional parallelism between different study samples. However, it is notable that the estimation of antibody

levels by titration has been found useful, over decades, in diagnosing and treating infectious and autoimmune diseases as well as vaccination [16–22]. Titration method generally requires less validation work and makes it easier to compare ADA responses because it does not require detailed characterization and demonstration of comparability and/or parallelism between calibrator antibodies developed for different products. Furthermore, the mass units approach requires re-qualification whenever the calibrator is changed, which is not the case with the titration method. As stated above, neither method provides accurate quantitative data. Therefore, sponsors have a choice between determining ADA levels as "relative concentrations" expressed as mass units, or in terms of more traditional and, clinically well established, titers. The latter approach may also be preferred by some regulatory agencies.

In the titer approach, the reciprocal of the lowest dilution whose signal falls below the cut point is considered the titer. Instead, the highest dilution that remains above the cut point can be considered the titer as long as it is ensured that at least one of the serial dilutions produces assay signal below the cut point (i.e., the signals from the serially diluted samples should span the cut point). Alternatively, some laboratories interpolate the titer at a pre-established value/cut point (generally the screening cut point), which requires fitting the dilution profile with an appropriate regression curve. In general, the screening cut point is used for titer determination; however, in assays where matrix has a blocking effect such that its signal is lower than that of the assay diluent, the titer of serially diluted samples is determined using a diluent-based "titer cut point".

During method development and optimization, methodological decisions shall have been made and documented, such as the selection of requisite reagents, their optimal concentrations, and the minimum required dilution of the intended samples. It is left up to the investigator to decide whether those decisions will need to be repeated or simply confirmed during pre-study validation. In this paper the authors assume that the following assay attributes have been established prior to validation: (1) immunoassay format and design, (2) type of assay matrix, (3) minimum required dilution (MRD), (4) optimal concentrations of reagents, (5) assessment of plate uniformity (e.g., location effects, drift, etc.). Supportive data on such determinations (e.g., MRD) may be requested by regulatory agencies. It is also useful if a preliminary understanding of method variability is obtained from the performance of control samples during development and optimization.

2.2. Application of statistics

A major focus of this paper is to convey that subjectivity during assay validation must be reduced, if not eliminated. Hence, one must rely upon statistical means to ensure objectivity. Because most researchers do not have access to the services of a trained statistician, simple yet sufficiently rigorous and valid statistical methods are provided in this paper. These statistical computations can be applied with the help of user-friendly commercial software, without the need for a formal training in statistics. However, the assistance of a statistician for planning validation experiments and analyses of data can lead to the application of more rigorous and elegant statistics than suggested herein.

3. Pre-study assay validation

The validation of an assay before commencing sample bioanalysis for nonclinical or clinical studies is called 'pre-study validation'; it describes in mathematical and quantifiable terms the performance characteristics of an assay [8]. This should not be confused

with the colloquial term "prevalidation" used to describe any preparatory work performed before initiating pre-study validation. On the other hand, in-study validation refers to the monitoring of assay performance throughout its use to assure that the assay remains in a valid state and the resulting bioanalytical data are reliable.

Thorough method development and optimization activities should result in an assay that the investigator believes is "ready for validation", meaning that there's a significant amount of method optimization data indicating potential reliability of the assay and suitability for its intended purpose. Reliable performance of the assay is dependent upon properly functioning analytical equipment and computer systems, as well as training and dexterity of the analysts. In essence, the assay is a "system" comprising several elements other than assay reagents alone. Therefore, method validation establishes "system-suitability", which should also be maintained during the in-study phase.

A validation plan or a validation standard operating procedure (SOP) is recommended before initiating pre-study validation experiments. The validation plan should state the intended purpose of the method, a detailed description of the immunoassay, a summary of the performance characteristics to be validated, and *a priori* target acceptance criteria for precision, robustness, stability and when appropriate, ruggedness. It is advised that some experimental detail and data handling procedures be presented in the validation plan because it provides a clear guidance to the validation analysts, ensuring better control over the resulting data.

For ADA methods, acceptance criteria should be established that help ensure that assays performed during the in-study phase remain in a valid state. To do this, acceptance ranges (system suitability criteria) for quality controls should be established by statistical evaluation of experimental data acquired during the validation. In addition, it is useful to apply acceptance criteria for controls and study samples based on inter-replicate precision during the in-study phase. While data from assays that fail acceptance criteria during the in-study phase should be rejected, criteria for passing or failing assays in pre-study validation experiments should be avoided because these can potentially lead to the exclusion of some validation data, resulting in an inaccurate estimate of analytical error. *All assays during pre-study validation should be included; the only exceptions should be those rejected for an assignable cause (e.g., technical error) and those that deviated intentionally (e.g., certain robustness experiments) or unintentionally from the method protocol.*

Current FDA, International Conference on Harmonization (ICH), and United States Pharmacopeia guidance documents describe general performance characteristics that should be investigated for analytical validation of quantitative assays [23–25]. Some of these characteristics may not be appropriate to investigate for ADA assays because of their quasi-quantitative nature. For ADA immunoassay validation, the following nine analytical performance characteristics are relevant:

1. Screening cut point
2. Specificity cut point
3. Sensitivity[*]
4. System suitability control (QCs) acceptance criteria
5. Selectivity/interference[*]
6. Precision
7. Robustness
8. Stability[*]
9. Ruggedness, when applicable

When adequately demonstrated during the method development/optimization and documented appropriately, the performance characteristics marked above with an asterisk (*) may not require repeat verification during pre-study validation, although relevant data should be provided in the validation report. *Care must be taken to ensure that the assay procedure was not altered after these characteristics were evaluated during method development/optimization; if not, they will likely have to be determined during pre-study assay validation.*

The use (intended purpose) of ADA detection assays is likely to change during development and post-licensure. For example, early in development only a single analyst may be performing an assay, whereas later in development and post-licensure multiple analysts are likely to perform the assay. Consistent with assay use, validation requirements may change throughout the product life cycle. Nevertheless, validation of the above listed assay performance characteristics is invariably required during the application for drug licensure.

The traditional analytical approach [23] to dilutional linearity testing in general does not apply to ADA assays. When ADA level is estimated as a titer, it is important to demonstrate using the positive control(s), or preferably with accrued ADA positive samples, that there is no unusual non-linearity within a reasonable dilution range. Typically this is a component of MRD selection during method development and in that case, retesting during validation may be unnecessary. If prozone (hook) effects are observed, it is advised that the screening assay use an MRD that is free from prozone, or employ more than one dilution of the sample. Dilutional linearity information is also useful in the design of appropriate "system suitability" quality controls (Appendix E).

3.1. Screening cut point

The screening cut point is defined as the level of response of the screening assay at and above which a sample is defined to be a "reactive" (often called "potential positive") for the presence of ADA, and below which it is probably negative. A valid assay cut point is established during pre-study validation by a systematic and statistical evaluation of assay responses for a subset of samples that are judged to be representative of drug-naïve target patient/subject population.

Using a risk-based approach to immunogenicity evaluations, it is more appropriate to have false positives than false-negatives during screening [8,10]. A screening assay that does not identify any reactive samples whatsoever can cast doubt on the ability of the assay to detect low positive samples. A screening assay that picks up some (e.g., ≥5%) positives that can subsequently show to be non-specific in a confirmatory assay provides assurance that true low positives can be detected [8]. In practice, identification of any false positives is better than none at all (i.e., false-positive rate in studies may not equal 5%).

3.1.1. Types of screening cut point

There are three types of screening cut point that can be calculated for application during the in-study phase—fixed, floating and dynamic. These are defined below:

(a) *Fixed cut point*: a cut point that is determined in pre-study validation and the same value is used for the in-study phase. The fixed cut point is used for analyses of test samples until there is a need to revalidate or change the cut point (e.g., critical change in the assay, assay transfer to another laboratory, upgraded instruments, etc.). The cut point value can be fixed within a given study, for a target population, or across studies and multiple target populations. In order to use this approach, one should demonstrate similar means and variances across assay runs during pre-study validation.

(b) *Floating cut point*: a cut point calculated by multiplying a specific normalization factor, determined from the pre-study validation data, to the biological background obtained during the in-study phase. Alternatively, biological background value can be added to the normalization factor, if log transformation was not necessary in the analysis. The biological background may be represented by the negative control (pool of matrix from subjects that are negative for anti-drug antibody), the assay diluent, or the pre-dose subject sample (subject-specific cut point). A floating cut point may be plate specific, unique for each run (common for a number of plates within a run), or unique for each subject (using the subject's pre-treatment/"baseline" sample result). If subject-specific floating cut point is used, then the pre-treatment and post-treatment samples should be tested in the same plate (or at least in the same run). *Because the method for determining floating cut point uses the variation estimate from the pre-study validation, one should demonstrate homogeneity of sample variance across runs*. When negative control is used for normalization, one should also ensure appropriately that the negative control results represent the drug-naïve matrix sample results of the target population, i.e., by verifying that the negative control mean and the mean of biological matrix sample results correlate across assay runs. This can be addressed by examining the scatter plot of mean values from different runs. If the means are correlated, then use of the negative control mean to calculate a floating cut point would be considered appropriate. If not, the use of assay diluent for normalization, or pre-treatment subject ("baseline") sample results may be more appropriate.

(c) *Dynamic cut point*: a cut point that changes between the plates in a run, between runs in a study, or between studies, and *does not use the variation estimates from pre-study validation*. The latter characteristic differentiates it from a floating cut point. This approach is necessary only when the variability between sample results from each run is significantly different between assay runs. A practically limiting factor is that a significant number of samples are needed in each plate/run to compute this cut point. In practice, therefore, the use of a fixed or floating cut point is recommended whenever appropriate. If differences between runs (means and/or variances) are encountered, an investigation of the source of such differences is recommended before instituting a dynamic cut point. For example, if these differences were primarily due to analysts or instruments, then one should consider the appropriateness of analyst or instrument-specific fixed or floating cut points instead of a dynamic cut point. Because this is an uncommon and unconventional cut point approach, it is advised that the use of dynamic cut points be cautiously instituted, preferably with regulatory consultation.

3.1.2. Samples for cut point evaluation

It is recommended that samples from an appropriate population be utilized for the development of an assay cut point. In some cases it may not be practical, or feasible, to obtain matrix samples from a population having a target disease prior to initiating pre-study validation experiments. Consequently, the common practice is to evaluate samples from healthy drug-naïve subjects to establish an initial cut point. This approach is the preferred one for a conventional Phase I study that is conducted in normal volunteers. When the clinical program progresses beyond Phase I and samples from the target disease population become available, it is appropriate to re-evaluate cut point data for the respective target population. If the distribution of assay responses with respect to both the mean and variability are comparable between the target population and the normal volunteers, then the same cut point

can be used. If not, target disease-specific cut points should be utilized.

For estimation of a cut point, an adequate number of samples should be analyzed during validation to provide a statistically valid assessment of biological and assay variability at the selected MRD. Typically in clinical studies, matrix samples from $\geq$50 individual human subjects are analyzed. Due to practical considerations in nonclinical studies, at least 15 samples might be sufficient. *The use of pooled matrix samples for cut point determination is inappropriate because testing replicate samples from a pooled mix measures analytical variation but not biological variation*. A 'balanced' experimental design that allows reliable assessment of potential sources of variability is recommended (see Appendix A). If multiple analysts will conduct analyses of test samples during the study, the cut point investigation during pre-study validation should include at least two analysts. This is also required if study samples are tested by a single analyst, but one different from the analyst who validated the assay. In addition, replicate tests of each sample may be needed depending on how a result is reported during in-study bioanalysis (e.g., using duplicate or triplicate wells in a microtiter plate).

3.1.3. Exclusion of outliers

Occasional reactive samples from drug-naïve subjects should be confirmed for drug specificity via methods such as drug inhibition. At this point, a specificity cut point will not be available to truly confirm positives. Scientific judgment should be used instead. When unsure, it is better to exclude such a sample. ADA positive samples and statistical outliers (samples with unusually low or high signals) should be excluded from the statistical evaluation of the assay cut point. Outliers in the lower extreme should also be excluded because these tend to inflate the variability and hence usually the cut point as well. A lower cut point resulting from the elimination of such high and low outlier samples will usually err on the side of caution by increasing the rate of false positives (the false positives may subsequently be proven negative by the specificity confirmation test). *If robust alternatives are used (e.g., median-based methods, Tukey's biweight function) in the cut point calculation, then the exclusion of outliers is not required*.

3.1.4. Determining a screening cut point

In order to determine the optimal type of screening cut point, a series of assessments are required, as shown in Fig. 1. First, if a parametric approach is used, the distribution of drug-naïve matrix sample results should be assessed (test for normality), appropriate data transformation should be selected if the test for normality fails, and the outliers should be evaluated (Appendix B.1). If a nonparametric approach is used (95th percentile), transformation of the data is not necessary, but the outliers should be evaluated. Next, the assay run means and variances should be compared using ANOVA-based statistical methods for assessing the suitability of a fixed or floating cut point, or the potential need for a dynamic cut point (Appendix B.2). If a negative control is used for normalization in a floating cut point calculation, it should be compared with the mean of the target population's matrix samples to determine its suitability for the calculation. Finally, based on findings from these assessments, appropriate statistical methods should be used to calculate the cut point (Appendix B.3).

3.2. Specificity cut point

Specificity is the property of an analytical method to unequivocally detect the target analyte in the presence of other matrix components [23]. From the screening assays, samples are selected
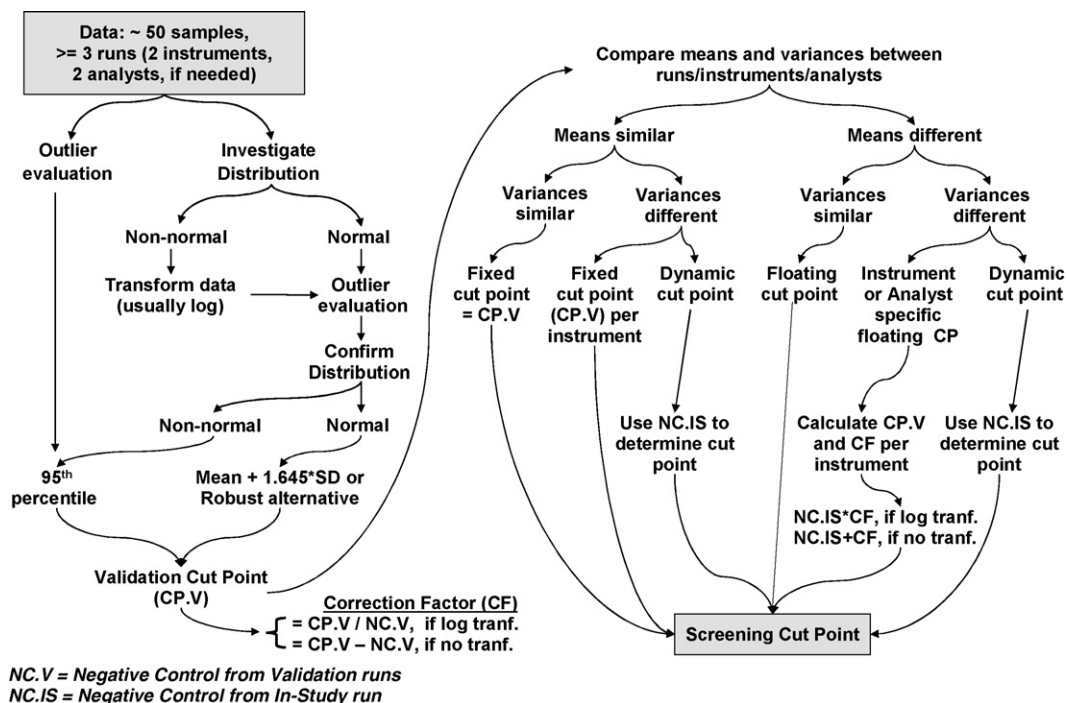
Fig. 1. Scheme for evaluating cut point samples and calculating screening cut point.

that are "reactive" (often called "potential positive") for the presence of ADA; however, some of these may be non-specific because of the 5% false-positive rate built into the screening cut point. Therefore, the confirmation of positives among the reactive samples requires the demonstration of specific reactivity to the drug. Testing the specificity of ADA in human samples is a common practice; this is also recommended for animal samples.

The specificity confirmation assay is usually a competitive inhibition test in which the data are evaluated for a change in assay signal of a sample with or without preincubation (in liquid phase) with the study drug. Alternatively, reactivity in wells coated with an immunochemically unrelated protein of similar size and charge in comparison to drug-coated wells may be assessed. Instead of these approaches, some researchers compare the signal between drug-coated and uncoated wells in plate-based assays; however, we recommend that the acceptability of such an approach be proactively discussed with regulatory agencies. In any of these approaches, the amount of change in assay signal that determines a positive confirmation is referred to as the "specificity cut point". *The validity of the specificity cut point, which distinguishes specific binding versus non-specific binding of antibodies to the drug, is very critical and must be determined objectively*. The use of subjective criteria, such as ≥50% inhibition of signal, is discouraged. This is particularly important when one has to evaluate ADA positive samples with low signals that are slightly above the assay cut point. High ADA positives, given that excess drug is used for inhibiting the signal, are generally not prone to this problem. In the instance of a low signal ADA positives, even a minor decrease in the signal of a drug-spiked sample may result in a 50% or greater decrease relative to the unspiked counterpart leading to a false-positive assessment. Furthermore, it may only be possible to reduce the signal of an ADA positive sample down to the assay background, but not any lower, which might be less than 50% of the signal of the uninhibited sample; in such cases, false-negatives would occur. Therefore, the specificity cut point should be determined by an objective approach, in the context of assay variability near the low positive

range of the assay. There are several experimental approaches that are being used in determining the specificity cut point.

It is advised that the specificity cut point be determined during validation of the screening cut point. Drug-naïve samples from the population used for determination of the screening cut point (excluding the statistical higher/upper outlier samples) should be spiked with an excess of drug and analyzed in identical fashion. The mean percent change from the unspiked sample (inhibition) and SD are calculated. The mean inhibition plus 3.09 SD (if a 0.1% false-positive rate is desired) represents the specificity cut point. As in determination of the screening cut point, outliers are eliminated in order to make the specificity cut point more conservative. Further details and the steps for calculating the specificity cut point are provided in Appendix C.

In some labs, a number of samples (≥25 recommended) below the screening cut point are spiked with a positive control resulting in a signal at or just above the cut point. After incubation with excess drug, mean percent inhibition is calculated and diminished by '×' SD to establish the specificity cut point (where '×' is chosen to correspond to an appropriate false-positive rate). In this approach, it is critical to ensure that the positive control preparation does not produce a signal far above the cut point because that can raise the likelihood of false-negatives. Also, one caveat to this approach is that the specificity cut point can vary highly based upon the positive control reagent(s) used. It should be noted that low affinity ADA that can occur within study samples (especially multivalent IgM), may not be inhibited as well as affinity matured IgG. Therefore, a study sample may not be reliably classified as having specific or non-specific reactivity based on a specificity cut point derived using a single positive control.

A *t*-test based approach proposed by Neyer et al. [26], which compares signals of the native and drug-spiked sample, represents an alternative determination of ADA specificity. However, results of the competitive inhibition in this assay are typically based on a very limited number of observations and it does not take into account the biological (inter-subject) variability.

Nevertheless, any of the above outlined approaches are certainly more objective and reliable than the arbitrary ≥50% inhibition. Some researchers perform antibody immunodepletions (e.g., with protein A), but that does not necessarily address specificity to the drug; instead it only proves that signal is immunoglobulin based. While this approach may serve as a supporting test, it is not recommended as a determinant of drug specificity.

3.3. Sensitivity

Unlike fully quantitative methods that utilize an appropriate reference standard curve to estimate the analyte levels in the study samples, sensitivity of ADA assays is highly dependent upon the positive control reagent(s) used to characterize it. For example, the use of a high affinity positive control will produce a better sensitivity value than the use of a lower affinity positive control in the same assay. Hence when more than one positive control reagent is available, it is a common finding that each reagent produces a different sensitivity value. Furthermore, no positive control can be expected to represent the spectrum of immune response observed in individuals treated with study compounds. Despite these caveats, assay sensitivity is a regulatory expectation because it provides a general sense of relevance of the assay, and thus it is determined and reported. It is particularly useful during assay development for choosing an optimal ADA detection method (comparisons between methods during initial development) and for the determination of a low positive control for validation (see Section 3.4).

Sensitivity of ADA assays is defined by the lowest concentration at which a positive control antibody preparation *consistently* provides a positive signal in the assay. For example, this can be defined as the concentration of the control antibody at which the assay result is expected to be above the screening cut point at least 95% of the time, or 50% of the time, depending on the level of consistency preferred. See Appendix D for an experimental approach to determining assay sensitivity.

3.4. System suitability controls

System suitability controls "ensure that the validity of the analytical procedure is maintained whenever used" [24]. Each assay typically contains a set of quality controls (high and low positive controls and negative controls) that assure performance consistency, thereby supporting the validity of results obtained over time. In other words, the results of these controls aid in assuring that the assay remains valid during the in-study phase. Thus, these are quality controls that help determine whether an assay run passes or fails the validated acceptance criteria. See Appendix E for details on developing these controls and the required acceptance criteria for in-study runs.

3.5. Selectivity/interference

Whether components in the sample prevent the assay from detecting ADA is an important concern. Selectivity is the ability of an assay to measure the analyte of interest in the presence of other constituents in the sample [27]. It is characterized by the recovery of analyte (represented by a mock positive control sample) from matrix samples containing the potential interferent(s). When the positive control and the matrix originate from different species, as is often the case, anti-framework and anti-xenogeneic antibodies may often hinder recovery. It should be noted that selectivity of ADA assays, when assessed using the positive control, is not likely to reflect the selectivity of the assay when applying actual nonclinical or clinical samples. Nevertheless, it is important to gain an understanding of ADA detection under conditions when the matrix could possess interfering factors such as study drug or its endogenous counterpart, concomitant medications, rheumatoid factor, etc.

3.5.1. Interference by matrix components

Assessment of recovery in ADA assays involves testing whether components in the matrix inhibit ADA from binding to drug under assay conditions, thereby affecting the assay response. When possible, selectivity investigations should include a comparison of the positive control ADA recovery within normal and disease-state sera in view of the possibility that interfering substances may be prevalent in some populations or disease states. If a significant proportion of test samples are hemolyzed or contain high lipid content, these conditions may also require recovery evaluation. To determine recovery from target matrix samples, both low and high concentrations of specific ADA control preparations (polyclonal or monoclonal) should be added into assay buffer (which is free of matrix components), biological matrix (serum or plasma) samples from healthy individuals, and treatment population subjects. The responses in assay buffer are then compared to responses observed in matrix, and typically up to 20% difference between the two is considered acceptable, *but this will vary based on the positive control used.* It is useful to select donors that express high and low non-specific backgrounds in the assay. This will not only test the assay recovery but also check the matrix interference and confirm the selection of the MRD.

3.5.2. Study drug interference (drug tolerance)

The main interferent in an ADA assay is usually the drug itself. It is expected that samples containing drug will exhibit interference due to competition for product-specific antibodies between the drug and that used as a capture reagent in the assay system, thereby producing false-negative results. Thus it is a common practice to determine the concentration of drug that inhibits detection of a positive control antibody, and to apply this "tolerance limit" in decision-making regarding the ADA status of study samples. However, this practice is flawed because drug tolerance is highly dependent upon the positive control(s) used to characterize it, such that the use of a high affinity positive control can result in low drug tolerance whereas the use of a low affinity positive control can result in a high drug tolerance in the same assay. Because individuals will vary in their ADA repertoire, the true drug tolerance in the context of antibodies in the study samples may be different from that obtained with the positive controls. In fact, when more than one positive control reagent is available, it is a common finding that each reagent produces a different drug tolerance value. Thus, the "true" drug tolerance of an assay is intangible. Thus, like assay sensitivity, drug tolerance is generally limited to an understanding of whether the drug interferes at all in the assay, and is useful mainly for comparisons of different assay formats during assay development and optimization. Nevertheless, determining drug tolerance is a regulatory expectation.

To determine the tolerance of an assay to drug interference, the low ADA QC is preincubated with serial dilutions of drug, and then tested in the assay. The lowest concentration of drug that prevents the detection of the low QC signal characterizes the "drug tolerance limit" of the assay, *but this will vary based on the positive control used.*

For study phase bioanalytical reporting, when drug is detected in a sample that produces a negative ADA result in the assay, it is recommended that the negative ADA status of that sample should be accompanied with a statement of possible drug interference [10].

3.6. Precision

Precision is a quantitative measure of the random variation between a series of measurements from a method. It is recom-

mended that the acceptance criteria applied against precision be within the range commonly expected for immunoassays; these criteria should also be appropriate for the platform used (for example, the criteria may be tighter for electrochemiluminescence assays as opposed to ELISAs). The use of a balanced experimental design is recommended for the performance of precision validation, as described in Appendix A. Approaches for determining precision of the screening, specificity confirmation, and titration methods are presented below. When feasible, it is recommended that precision be determined in experiments that are scaled-up approximately to the level of intended use during the in-study phase.

3.6.1. Screening assay precision

Precision of the screening assay can be determined using data from at least six independent runs of the assay controls (negative control, low positive and high positive control). The data for this evaluation can be obtained from the experiment described in Appendix D for determining sensitivity. From that experiment calculate the imprecision around a high positive control concentration and a low one that is just above the cut point. The imprecision from these assay signal data can be reported in terms of %CV for characterizing relevant sources of variability. The estimates of intra-assay precision (also called intra-run or intra-batch precision) and inter-assay precision (also called inter-run precision, inter-batch precision, intermediate precision, or overall precision) of the assay signal measurements should be reported as %CV, for example using the formulae from a recent publication [27]. If analyst-specific cut points are used for testing study samples, then the inter-analyst CV should also be determined.

3.6.2. Specificity confirmation assay precision

The objective of precision assessment for the specificity confirmation assay is to test the reproducibility of signal inhibition. In order to calculate %inhibition, an appropriate level of drug should be added to the low and high control samples from at least six independent runs, as determined during the specificity cut point experiment. Determination of inter-assay precision is sufficient for this assay, and is calculated in a manner identical to the method described in Section 3.6.1.

3.7. Robustness

Robustness is an indication of the reliability of an assay, assessed by the capacity of the assay to remain unaffected by small, but deliberate, variations in method parameters [23]. The focus of robustness is to elucidate assay consistency under relevant, real life changes in standard laboratory situations. The robustness parameters to test during validation should be based on the knowledge of the assay and its associated risks [28]. One should assess which conditions are likely to vary in an assay in a particular laboratory, and design appropriate tests to examine the parameters that are deemed critical. These may include changes in microtiter plate lots, incubation times, temperature, number of plates per run, reagent lot and concentrations, or instrumentation. Study samples, or the positive control samples, can be used to test assay robustness. The use of acceptance criteria for system suitability controls during robustness validation (computed from the assay development/optimization data, or validated system suitability control acceptance criteria) is recommended.

3.8. Stability

Stability studies evaluate assay performance under the intended sample storage conditions. Ideally, stability testing conditions should mimic the expected sample and reagent handling conditions, storage temperature(s), and varying lengths of storage time.

From an analyte stability standpoint, the nature of ADA is worth pondering. Whether the ADA analyte is anti-drug X antibody or anti-drug Y antibody, it is a polyclonal antibody just the same. Therefore, it is reasonable to assume that the stability of ADA is the same whether it is specific to drug X or drug Y. With this logic, the stability of ADA can be approximated by the stability of serum or plasma immunoglobulin *specific to any antigen*. It is recommended that available clinical and nonclinical matrix samples, from each target species, be characterized separately for stability and the ensuing results be extended to ADA assays of all drug programs in a research laboratory. It may also be necessary to do this with matrix from certain target indications such as rheumatoid arthritis. *Thus, it is not stipulated that sample stability for each drug-specific assay be separately validated*. While a caveat to this approach is the fact that the administrations of different drugs can elicit different proportions of IgM, IgG_1, IgG_2, IgG_3, or IgG_4 in different subjects, it is impossible to know and mimic the various proportions of isotypes and subclasses of immunoglobulin in any population of subjects. Hence, this approach is reasonable for the evaluation of analyte stability.

Alternatively, drug-naïve matrix spiked with the positive control(s) can be considered as mock positive sample(s) for stability testing, but whether such samples reasonably mimic the intended test samples is a scientific judgment. Multiple positive samples at different concentrations are preferable, if available; but in the least, the samples should include the high-signal positive control and the low-signal positive control. See Appendix F for further details.

Stability characterization may also include stability of assay-critical reagents such as the quality controls, the coated assay plate or chip (if applicable), and other critical reagents (such as conjugates). However, this is a business decision rather than a stipulated validation characteristic because the ADA assays are stability indicating (i.e., loss of stability of critical reagents can be detected by poor assay performance, or assay failure, monitored via the system suitability control).

3.9. Ruggedness

Ruggedness refers to the reliability of an assay when performed by more than one laboratory. Thus it refers to the inter-laboratory studies needed to demonstrate that an assay is valid. The term 'ruggedness' is not present in FDA and ICH method validation guidances [13,23,24] while the United States Pharmacopeia (USP) [25] describes it as including inter-laboratory precision and inter-analyst precision. Inter-analyst precision is described within 'intermediate precision' in the ICH documents [23,24], which classify inter-laboratory variation as another element of precision termed 'reproducibility'. Ruggedness is often erroneously interpreted as 'routine changes' such inter-equipment imprecision, which is in fact a robustness element.

Because some companies may have multiple bioanalytical testing laboratories, or due to growing trends such as outsourcing, it is recommended that ruggedness be validated in those situations. Ruggedness becomes very useful for assessing the "transferability" of an assay, i.e., the validity of testing samples in two or more laboratories, and comparability of data produced by them. For example, assay sensitivity using a single positive control, cut point values and control ranges should be evaluated; blinded samples could be generated in the sponsor lab and tested by the receiving lab. When only a single laboratory performs the ADA assay, however, ruggedness becomes a non-issue and need not be validated until the method is to be transferred to another laboratory.

4. In-study validation (monitoring) and assay revalidation

In-study validation (monitoring for maintenance of system suitability) and revalidation are critical components of any bioanalytical method. Hence, the validation of a method actually does not end until the method is ultimately retired from analytical use.

For in-study performance of quantitative bioanalytical methods, acceptance criteria for precision and accuracy are generally required [13]. Since accuracy is not applicable for ADA methods, monitoring the performance of quality control samples, as described in Section 3.4, reassures that the assay system is "suitable for intended use", i.e., the assay remains valid and is performing as well as it had during pre-study validation. The use of a low positive control assures the assay remains sensitive. Details on sample and plate acceptance criteria are explained in Appendix E2.

It is also advised that assays be re-validated on an as-needed basis to reduce assay "drift". For example, when there are changes in critical method components, equipment, or samples (disease indications), an assay revalidation may be required. The revalidation may cover some or all validation characteristics (i.e., a partial or whole assay revalidation). Use of lots or batches of assay critical reagents that are different from those used in pre-study validation do not require assay revalidation, but must be supported by appropriate experimental qualification to ensure maintenance of system suitability.

5. Conclusion

The first of three papers in this series on immunogenicity evaluation described common approaches for developing and optimizing immunoassays for antibodies to biotechnology products [8], following which the second paper described strategies for the evaluation of ADAs [10]. This paper described assay performance characteristics that are important to ADA immunoassay validation, and provided recommendations on objective approaches for determining them. These are intended to facilitate a standardized approach for assessing the immunogenicity of therapeutic proteins across the biopharmaceutical industry. The recommendations in all the three papers should be considered as examples of best practice; alternate methodological approaches may also be acceptable, as long as scientific rationale and objectivity are maintained despite assay-specific practical considerations.

It is expected that the recommended approaches presented herein will result in the production of high quality ADA data that enables a better understanding of the clinical impact of immunogenicity. It is also hoped that this publication will lead to the production of specific guidance documents by regulatory agencies worldwide.

Appendix A. Use of a balanced experimental design

One important consideration in the design of validation experiments is to eliminate or reduce confounding factors. That is, the effect of each variable such as analyst, batch/run, etc., assessed in the experiment should be entirely due to those factors alone without bias from the effect of other factors. *It is recommended that the influence of confounding factors be reduced via the use of a 'balanced' experimental design for the determination of screening cut point, specificity cut point, and precision.*

For example, when evaluating precision, the difference between analysts, assay runs, and plates should be considered. Suppose that three plates per run are tested over three runs by each of two analysts. The experiment (Table 1) can be designed as follows in order to estimate the difference between analysts, assay runs and plates separately:

1. Divide the samples into three equal sized subgroups.
2. For each analyst, test each subgroup of samples exactly once in each assay run and on each plate.

An imbalanced design can have a negative impact on the conclusions from a validation experiment. Suppose that each subgroup of samples is tested in the same plate across assay runs, the differences between plates are confounded with the difference between sample subgroups. That is, any difference in results between plates cannot be attributed fully to the plate testing order; it may be due to a difference between the sample subgroups. Similarly, if each subgroup of samples is not tested in every run, then the differences between runs are confounded with the difference between sample subgroups. Finally, if every analyst does not test each subgroup of samples, then the analyst difference will be confounded with the sample differences.

Thus it is critical to try to ensure during the planning of the validation experiments that the experimental design or layout for

Table 1

Balanced experimental design. This table illustrates an experimental design that is balanced for the key assay variables considered such as the analyst, assay run, plate testing order, and sample groups tested. In this scenario, each of two analysts conducts three assay runs, testing 60 drug-naïve matrix samples in each run such that three groups of 20 samples are tested in each of three plates. Note that all samples are tested in each run and each sample group is tested in every plate testing order by each analyst. Such a balanced design ensures that the difference between the levels of a factor (say, between assay runs) is not confounded by the effect of another factor (say, sample groups), thus providing a more reliable assessment of key assay variables. The plate testing order here refers to the order in which the plates are tested within a run (assumed to be P_1, P_2, and P_3).

Analyst	Assay run	Assay plate	Validation serum samples		
			S_1–S_{20}	S_{21}–S_{40}	S_{41}–S_{60}
A_1	R_1	P_1	X		
		P_2		X	
		P_3			X
	R_2	P_1		X	
		P_2			X
		P_3	X		
	R_3	P_1			X
		P_2	X		
		P_3		X	
A_2	R_4	P_1	X		
		P_2		X	
		P_3			X
	R_5	P_1		X	
		P_2			X
		P_3	X		
	R_6	P_1			X
		P_2	X		
		P_3		X	

running the control samples and validation matrix samples is balanced with respect to all the key factors under consideration.

Appendix B

B.1. Cut point evaluation: investigate distribution of results and exclude outliers

The assessment of the distribution of results from the population of drug-naïve ADA negative matrix samples is an important first step prior to establishing a screening cut point.

Evaluation of the distribution includes:

- Assessment of normality.
- Selection of an appropriate data transformation, such as logarithmic (if needed).
- Identification of statistical outliers.

Data transformation is often needed in order to satisfy the distributional assumptions of the statistical analysis (e.g., ANOVA). Low and high outliers arising from analytical or biological abnormalities should preferably be excluded [8], or appropriately down-weighted (e.g., by use of Median and Median Absolute Deviation or Tukey's biweight function) in the determination of a screening cut point. *This is important because not doing so can inflate the variability and therefore the screening cut point, making the assay more prone to false-negatives.*

A couple of popular methods for detecting outliers are as follows:

1. Box-plots of the reported results.
2. Studentized residual plots from ANOVA.

The studentized residual plot approach might require the assistance of a statistician whereas the box-plot approach is simpler and also acceptable.

In the statistical software JMP® (SAS Institute Inc., Cary, NC, U.S.A), the box-plot approach can be carried out by the use of the outlier box-plot. The plot identifies all the points ("high outliers") that are above the 75th percentile (Q3) plus 1.5 times the inter-quartile range (Q3–Q1) and all the points ("low outliers") that are below the 25th percentile (Q1) minus 1.5 times inter-quartile range. These extreme results are then plotted individually in the graph.

The distribution of results in Fig. 2 is plotted in both the original scale and logarithmic scale, and the test for normality is carried out using the Shapiro–Wilk test. Since the distribution of the original data is highly skewed to the right, the logarithmic transformation is considered, and the outliers are assessed on the log-transformed scale.

Despite transforming the data to logarithmic scale, the Shapiro-Wilk test for normality indicates a non-normal distribution. This is due to the presence of several outliers in the logarithmic scale. After excluding the outliers identified from the box-plot approach, the Shapiro-Wilk test now provides evidence of a normal distribution, as shown in Fig. 3. At this point, the assay run means and variances can be compared as described in Section 3.1.4.

B.2. Cut point evaluation: compare assay run means and variances to determine the suitability of a floating or fixed cut point, or the need for a dynamic cut point

- Use a fixed cut point when assay means and variances are not statistically different between runs, however, floating or dynamic cut point can also be used if preferred.
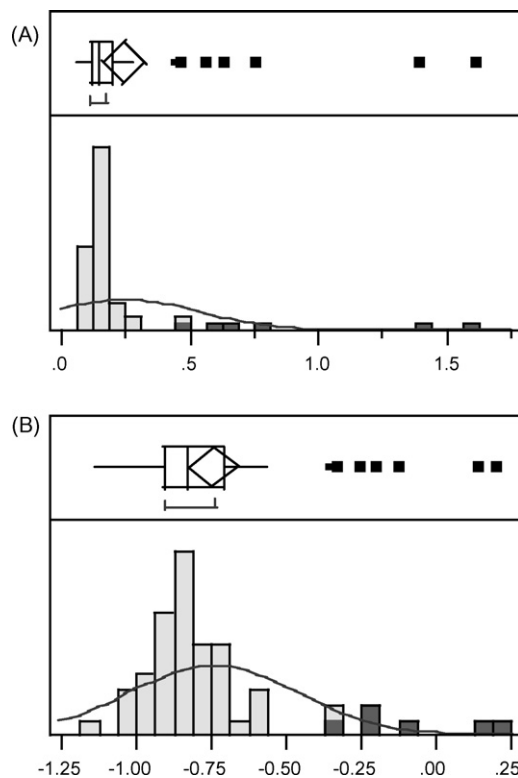

Fig. 2. Distribution of the drug-naïve matrix sample results. Using JMP outlier boxplot, the distribution of the drug-naïve matrix sample results is plotted in both the original scale of the reported optical density results (Panel A) and logarithmic scale (Panel B). The ends of the box correspond to the 25th and 75th percentiles, also called the quartiles (Q1 and Q3). The difference between these quartiles is the inter-quartile range (IQR). The line across the middle of the box indicates the median. The center of the diamond indicates the sample mean, and its length corresponds to the 95% confidence interval. The lines extends from both ends of the box to the outer-most data point that falls within the distances computed as [Q3 + (1.5 × IQR)] and [Q1 − (1.5 × IQR)]. The bracket along the edge of the box identifies the shortest half, which is the densest 50% of the observations. The normal density curve estimated using the data is displayed. Since the distribution of the original data is highly skewed to the right, the logarithmic transformation is considered, and so the outliers are assessed in the log-transformed scale. The outliers can be identified from either the box-plots of the reported results provided in this figure, or using the studentized residual plots from the ANOVA (see Appendix B.1). The outlier boxplot individually lists the outlier points, as evident from these figures. Among these points, the smallest one was not identified as an outlier by the studentized residual plots.

- Use a floating cut point when assay means are statistically different but the variances are not statistically different among runs, however, dynamic cut point can also be used if preferred.
- Use a dynamic cut point when assay variances are statistically different between runs, regardless of the assay run means.
- Use analyst specific or instrument specific cut points when the drug-naïve matrix sample results are significantly different.

The difference between assay run means may be formally assessed within the framework of ANOVA by treating the assay runs as a fixed effect in the statistical model. First assess the p-value for the equality of assay run means. Next, assess the p-value from the Levene test for homogeneity of variances across assay runs. If variances across assay runs are homogenous and assay run means are not significantly different, then the use of a fixed cut point is justified. A more rigorous "mixed effects ANOVA" with assay runs and samples as "random effects" is required to evaluate mean differences and/or variance heterogeneity specific to other design factors such as population (normal versus disease samples) or analyst.
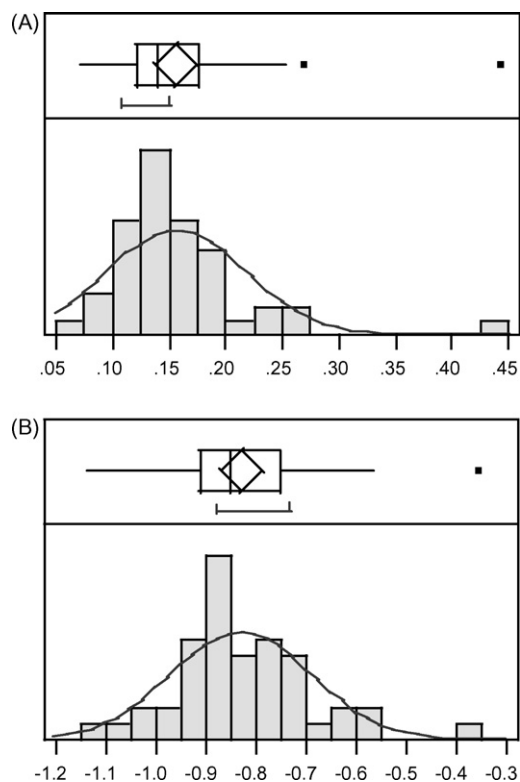
Fig. 3. Distribution of the drug-naïve matrix sample results without the outliers. These data are similar to Fig. 2A and B, without the outliers. After excluding the outliers identified from the box-plot approach, data in the logarithmic scale are closer to a normal distribution and is confirmed by the Shapiro–Wilk test for normality.

The comparison of assay runs is illustrated in Fig. 4. The p-value corresponding to the run number is 0.3068 suggesting that the run means are not statistically different at alpha = 0.05 significance level. The p-value from the Levene test is 0.6565 suggesting that the variances across assay runs are also not significantly different at alpha = 0.05 significance level. So this supports the use of
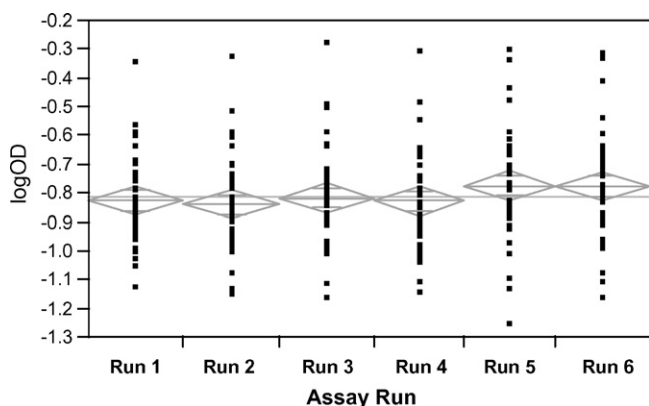

Fig. 4. Comparison of assay run means and variances. This illustrates the comparison of optical density readings in log scale (represented by log (OD) in the vertical axis) between six assay runs. Each point represents the result from a drug-naïve matrix sample. The vertical span of the diamonds represent the 95% confidence interval for the corresponding run mean. The horizontal lines near the top and bottom of the diamonds are useful for pair wise comparisons of the assay runs. As is evident from this figure, and as confirmed by the ANOVA F-test, the assay run means are not statistically significantly different. Also, the Levene's test for the equality of variances among assay runs indicated that the variances are not significantly different. This justifies the use of a fixed cut point for the in-study phase, or use of a floating cut point if preferred (see Appendix B.2).

a fixed cut point, however, a floating cut point may also be used if preferred.

B.3. Cut point evaluation: calculate a screening cut point

Alternative statistical methods for calculating the screening cut point using data from pre-study validation are listed below.

1. *Parametric method*: mean + 1.645SD. The mean and standard deviation (SD) are estimated using ANOVA. The parametric method assumes normality and exclusion of outliers. If data do not follow a normal distribution, this calculation should be performed on transformed scale. The calculation of SD should take into consideration of both the biological and analytical variability. A simple, but approximate way of accomplishing this is by calculating the variance of each run, and then pooling them using a weighted average, where the weights are the degrees of freedom from each run. If the number of samples is the same for each run, then the weights are same, and therefore the pooled variance is simply the arithmetic mean. The square root of the pooled variance results in the pooled SD which then can be used in this cut point calculation. A more elegant way of incorporating both the biological and analytical variability is by performing a variance component analysis using restricted maximum likelihood method within the framework of random-effects ANOVA. This typically requires the assistance of a statistician.

2. *Robust parametric method*: median + 1.645 × (1.483 × MAD), where MAD is median absolute deviation. This calculation resembles the formula given in parametric approach in that the median is used in lieu of mean and 1.483 × MAD is used in lieu of SD. Other alternatives such as the Tukey's biweight method may be considered. The robust parametric approach is most useful when certain data points are suspected to be outliers, but not identified by statistical tests. This approach assumes normality but exclusion of outliers is not required. If data do not follow a normal distribution, this calculation should be performed on transformed scale.

3. *Non-parametric method: Empirical 95th percentile*. This is robust against non-normality, but sensitive to outliers. Therefore, no transformation is needed when using this method, however, outliers should be removed as described in Appendix B.1. For example, the 95th percentile of a group of 60 samples will be the 57th sample after sorting the results in ascending order.

From our experience, both analytical and biological outliers are almost always associated with extremely high values. These values are often more than 6 standard deviations from the relevant mean value even after data transformation. Such values are not usually representative of samples from the drug-naïve normal or patient population based on which we determine the screening cut point. Failure to delete or down-weight these values would result in an unacceptably high cut point. Consequently, the deletion of outliers may be more practical and appropriate for the risk-based approach here than for other applications.

If a fixed cut point is chosen, the run-specific means and standard deviations are used to determine the pooled mean and pooled standard deviation. If a mixed effects ANOVA is used, then these estimates can be obtained automatically from the model. The overall fixed cut point is then calculated as described above using the pooled mean and standard deviation. No further calculations are necessary related to the screening cut point during the in-study phase.

If a floating cut point is chosen, and if the negative control is used as the biological background for normalization, then a normaliza-

Table 2a

Illustration of screening cut point calculation based on validation data. If a fixed cut point is used during the in-study phase, then it will be 0.192 (bottom-right of this table). If a floating cut point is preferred or necessary, then these cut point calculations should be used to determine the normalization/correction factor illustrated in Table 2b. See Appendix G for details.

Assay number	Analyst	Log 10-transformed estimates			Sample absorbance cut point
		Biweight mean	Biweight SD	Sample cut point	
1	A	−0.892	0.222	−0.527	0.297
2	A	−1.098	0.210	−0.753	0.177
3	A	−1.082	0.202	−0.750	0.178
4	B	−1.154	0.200	−0.824	0.150
5	B	−1.151	0.173	−0.866	0.136
6	B	−1.064	0.204	−0.730	0.186
	A	−1.024	0.212	−0.676	0.211
Pooled intra-assay	B	−1.124	0.193	−0.807	0.156
	Combined	−1.074	0.202	−0.741	0.182
Inter-assay	Combined	−1.076	0.219	−0.716	0.192

tion factor should be determined based on the pre-study validation data. Note that this is done only during the validation phase and used subsequently for calculating floating cut point during the in-study phase. This factor is simply the fixed cut point minus the average of the negative control sample values obtained during the pre-study validation phase. The floating cut point for an in-study assay run is then the run mean for the negative control sample plus the normalization factor. If the transformation is required, the calculation should be done on the transformed scale and the floating cut point established on the transformed scale needs to be back transformed prior to reporting.

Most often, logarithmic transformation is used in the computation of cut point. It can be shown using simple algebra that the normalization factor becomes multiplicative in the original scale and is defined in terms of the ratio as the cut point (determined from pre-study validation data) divided by the average of the negative control samples run during the validation phase. The floating cut point is then the average of the negative control samples run during the in-study phase multiplied by this normalization factor.

If the negative control is found to be inappropriate for calculating the floating cut point, then the pre-dose patient sample may be used to determine the patient-specific floating cut point. The floating cut point is then the average of the pre-dose sample replicates for each patient plus 1.645 times the SD, where SD is determined from the pre-study validation data described above using the parametric or robust parametric method (steps 1 and 2). However, it is important to note that the pre-dose and the post-dose samples should be run in the same plate or at least in the same run.

If a dynamic cut point is necessary, the parametric method from step 1 is used, where the mean and SD are determined from only the in-study samples in each assay run. As noted in Section 3.1.1, this typically suffers from limited sample size in practice and further investigation of the assay design factors (e.g., analyst, instrument) may be worthwhile. The use of dynamic cut points is laborious, impractical for large scale applications, and uncommon; thus its application should be instituted cautiously, preferably by soliciting feedback from regulatory agencies.

An example of the screening cut point calculation is provided in Appendix G. Per the procedure outlined in Appendix B.1, log transformation of the data was used for evaluating the variability and the calculation of cut point values. Similarly, using the analysis described in Appendix B.2, the floating cut point method using the negative control pool for normalization was adopted for the determination of cut point for the in-study phase. As suggested in Section 3.1.1, the relevance of the negative control was established by investigating the correlation between the negative control mean and the biological matrix sample means from each run. The cut point calculations were based on log-transformed data during the validation phase where each of two analysts tests the drug-naïve matrix samples in three runs each (hence a total of six assay runs).

The cut point calculation procedure and the formula used in Table 2a correspond to step 2 in this Appendix B.3. Robust alternatives to the mean and standard deviation using the Tukey's biweight method were used in the calculation. While it was not necessary to exclude the outliers, the assessment of normality and logarithmic transformation of the data were critical because the cut point calculation formula relies on the assumption of a normal distribution.

The calculation of normalization factor for use in floating cut point evaluations during the in-study phase is illustrated in Table 2b. Note that because these data were log transformed, as explained above, the normalization factor is additive in the log

Table 2b

Illustration of normalization/correction factor for the calculation of floating cut point. See Appendix G for details.

Assay number	Analyst	Log 10-transformed estimates			Absorbance multiplicative normalization factor
		Negative control mean	Sample cut point	Additive normalization factor	
1	A	−1.1531	−0.5266	0.6255	4.231
2	A	−1.2924	−0.7528	0.5396	3.464
3	A	−1.2697	−0.7503	0.5194	3.306
4	B	−1.2234	−0.8242	0.3992	2.507
5	B	−1.1807	−0.8661	0.3146	2.064
6	B	−1.1612	−0.7296	0.4316	2.701
	A	−1.2384	−0.6763	0.5621	3.648
Pooled intra-assay	B	−1.1884	−0.8069	0.3815	2.407
	Combined	−1.2134	−0.7408	0.4726	2.969

scale, and multiplicative in the original (untransformed) scale. During the in-study phase, one would multiply the mean of the negative control mean from each run by the multiplicative factor from Table 2b to derive the floating cut point for that run. Given the difference between the multiplicative factors corresponding to each analyst, it may be preferable to use the analyst-specific multiplicative factor, and hence the analyst-specific floating cut point in this scenario.

Appendix C. Calculation of a specificity cut point for a competitive inhibition test format

The following steps for calculating a fixed specificity cut point are suggested:

1. *Determine a single concentration of drug that can inhibit high levels of ADA in samples*: Using the available positive control(s), determine a single concentration of drug (in molar excess of ADA) that can inhibit the highest assay signal (i.e., the upper end of the dynamic range of the reading instrument). For example in an ELISA format, prepare mock high positive samples in matrix or assay buffer that result in an assay signal near 3.0 OD units in a plate reader that has a dynamic range of 0.0–4.0. Then add drug at serially increasing concentrations to identify the one concentration that inhibits the mock high positive samples to levels below the screening cut point. A higher concentration of drug should be used for inhibition during validation and in-study (e.g., 10-fold of the determined value, to reduce the likelihood of false-negatives due to affinity differences during the in-study phase). To assure that this drug concentration does not produce any unusual effects at lower concentrations of mock ADA positive samples, it should be confirmed in a second experiment where it is added to various lower concentrations of mock positive samples. This concentration of drug is thereby validated for use in inhibiting samples (during pre-study validation as well as subsequent in-study testing) that might have a screening OD result of up to 3.0 units. This means that study samples with screening OD results above 3.0 will have to be diluted to below 3.0 OD units before spiking the drug for specificity confirmation.
2. *Compute the mean percent inhibition and SD (standard deviation) that approximates that of "potentially positive" non-specific samples*: Pre-incubate the samples chosen for the screening cut point experiment at the predetermined excess study drug level (from step 1 above), ideally in the same experiment (together) as the screening cut point experiment. If this is not done in the screening cut point experiment, we recommend that this be done using a similar design recommended for the screening cut point experiment. Calculate percent signal inhibition for each sample with the following formula:

$$\%\text{Signal inhibition} = 100 \left[1 - \left(\frac{\text{study drug inhibited sample}}{\text{uninhibited sample}} \right) \right]$$

Compute the mean percent inhibition from these data. Then compute the SD of the % inhibition data. Typically a log transformation is necessary when analyzing the %inhibition data. Negative inhibition values may present a challenge when taking logarithm. One way to get around this is by analyzing the logarithm of the ratio of drug spiked versus unspiked samples.
3. *Calculate a fixed specificity cut point*: Applying the mean and SD from step 2, the specificity cut point is now defined as mean + 3.09 × SD when log transformation is not applied. If the data were log transformed, compute the mean and SD of the logarithm of the ratio of drug spiked versus unspiked sam-

ples as suggested in step 2; compute mean − 3.09 × SD of these log ratios, and then take the anti-log to obtain the specificity cut point via the formula: 100 × (1 − antilog value). (Note that the value 3.09 corresponds to the 99.9th percentile of the normal distribution. Since this is a confirmation assay, only the samples that are truly positive and specific to the study drug should be reported, limiting the false-positive error rate to around 0.1%. Instead, if 1% false-positive rate is preferred, the threshold of 2.33 should be used in place of 3.09 in the above formula.)

Appendix D. Determination of assay sensitivity

Experimentally, we suggest preparing five or more serial dilutions of the positive control spanning the screening cut point, and tested totally in at least six assay runs (by more than one analyst if multiple analysts will be involved during the in-study phase) in order to measure analytical variation. Positive controls should be prepared in undiluted pooled matrix, and then diluted according to the assay design. *Pooled matrix is recommended instead of individual matrix samples because the biological variation among the latter can confound the result, whereas the use of a pool ensures the measurement of analytical variation alone.* In some instances this preparation in the matrix may be proven unfeasible, as in the case of protein-A/G purified human mAb drug-specific positive controls prepared in animals; in such preparations, the presence of human Ig framework specific antibodies can significantly lower recovery from human matrix. In such instances, it may be acceptable to determine sensitivity in assay diluent rather than matrix.

Negative control (pooled drug-naïve matrix) and diluent (if applicable) should also be included in each run to provide data for the evaluation of precision (Section 3.6.1) and for the application of floating cut point if needed. Each of these dilution curves should be fitted by an appropriate regression model to interpolate the concentration corresponding to the screening assay cut point (fixed or floating, as determined to be appropriate for the method). Note that when a floating cut point is used, interpolation should be done using run-specific cut points.

Sensitivity can be defined as the mean of the interpolated positive control concentrations from these assay runs. In this case it should be understood that a sample at this sensitivity level would be detected as positive just around half of the time (i.e., 50% consistency).

If 95% consistency for reporting sensitivity is preferred, the mean and SD of the interpolated positive control concentrations corresponding to these assay runs are used to establish the sensitivity of the assay, defined as the mean plus $t_{0.05,\text{df}} \times \text{SD}$. Here the mean and SD should be determined in log scale, $t_{0.05,\text{df}}$ is the critical value determined from the t-distribution corresponding to a 5% false-positive rate and "df" is the degrees of freedom that depends on the number of samples and runs used in the calculation. The mean + $t_{0.05,\text{df}} \times \text{SD}$ is then transformed back to the original scale. It should be noted that we suggest the use of the t-distribution threshold of $t_{0.05,\text{df}}$ instead of the more familiar normal distribution threshold of 1.645 that we used for the calculation of cut points *because the sensitivity determination is based on very limited number of samples.*

Assay sensitivity should be reported for undiluted matrix and expressed as mass of antibody detectable per unit volume (milliliter). That is, sensitivity of the assay must be reported after factoring-in the screening assay MRD. When more than one positive control antibody is used for determining assay sensitivity it is acceptable to report a range of observed sensitivity values.

Appendix E. Developing system suitability controls (QCs) and acceptance criteria for the in-study phase

E.1. Step 1: Defining the QCs

A high positive control is generally chosen as one that produces a signal that is on the upper end of the linear response of the assay. This control is, arguably, of minimal use in ADA immunoassays and is therefore optional, but it can be useful for methods prone to hook effects, and in tracking assay performance, reagent qualifications, and troubleshooting. The concentration of the high positive control should be chosen from the linear range of the dilution curve, usually just below the upper plateau of the curve or at the higher end of the study sample range if there is *a priori* information on the ADA range in the study samples.

Unlike the optional high positive control, a low signal control ensures reliable performance of the assay and is therefore mandatory. The low signal positive control should produce a response that can be reproducibly seen above the cut point, but it may sometimes result in a signal that is below the cut point (thereby failing/invalidating the assay). On the other hand, choosing an unreasonably high concentration for a low positive control may produce an assay signal that is substantially above the cut point, which is inappropriate. To provide objectivity to the selection of a low positive control concentration, it is useful to think in terms of assay rejection rates, i.e., the percentage of assays (plates) that fail because the low positive control produces a result below the cut point [8]. For example, a 1% rejection rate may be a reasonable target for a low positive control. This is calculated as mean + $t_{0.01,df} \times$ SD, where mean and SD are determined using the data from the sensitivity experiment (Section 3.3 and Appendix D) or related assay development data, and $t_{0.01,df}$ is the critical value determined from the t-distribution corresponding to a 1% false-positive rate and "df" is the degree of freedom that depends on the number of samples and runs used in the calculation. This theoretically implies that about 99% of the data from the low positive controls will be at or above the cut point.

A matrix (usually a pool) negative control is useful in monitoring the non-specific background of the assay.

An assay diluent control may also be necessary in some assays as an additional negative control because matrix proteins sometimes have a blocking effect that can lower the signal in contrast to the true assay background with the diluent alone.

E.2. Step 2: Establishing in-study assay acceptance criteria for QCs

For QCs, assay acceptance criteria derived from the pre-study validation data must be applied in the in-study phase. To develop these criteria, data from each control (positive, matrix negative, and diluent negative, if applicable) from at least 3 runs, 3 plates per run, by two or more analysts (if multiple analysts will be involved in the in-study phase), and the same number of replicates as intended for study samples, should be used. If data from the performance of controls in other validation experiments is available, it should be utilized in these calculations.

Acceptance criteria for the negative controls should have an upper limit while a lower limit is optional. The positive controls generally have upper and lower limits, although the upper limits are arguably less critical than the lower limits (when the signal is within a linear range of the instrument), because the consequence of 'higher than normal' assay signal can result in a higher incidence of reactive samples, which can subsequently prove to be non-specific (by the specificity confirmation assay) and therefore does not affect the false negative rate of detection.

For a 1% failure rate of the matrix negative control and diluent negative control, the upper limit is the mean response + $t_{0.01,df} \times$ SD, where the mean and SD are derived from the experiment described above, and $t_{0.01,df}$ is defined as above in step 1. For the positive control(s), if both upper and lower limits are defined, then the acceptance range would be the mean response $\pm t_{0.005,df} \times$ SD, where $t_{0.005,df}$ assumes a 1% failure rate with respect to the lower or high extremes. If a different failure rate is preferred, then the thresholds can be changed accordingly. Some scientists also prefer a set of titration controls in titration assays to confirm dilutional linearity of a positive control, but these are generally unnecessary for in-study performance of the assay. In addition, it is acceptable to have overlapping acceptance limits for the negative control and low positive control, so long as these limits were derived objectively.

If a floating cut point approach is used for the screening cut point during the in-study phase, the system suitability criteria can be defined in terms of the ratio of low positive control to negative control, and also the ratio of high positive control to negative control if a high QC is used, using similar statistical formulae as above (log transformation might be necessary). The negative control used here (diluent or matrix pool) should be the same as the negative control used for calculating screening cut point during the in-study phase. If the negative, low and high positive control samples trend in a similar direction such that the ratios of the controls are within the system suitability limits, the use of a floating cut point approach will help ensure the validity of the assay. The use of an upper limit for the negative control in addition to these limits on the ratios may also be considered.

When there is high variability in the raw assay signal (optical density or relative luminescence units) some laboratories prefer the use of endpoint titer criteria for positive controls rather than acceptance criteria based upon the raw assay signal. In this case the positive control is expected to result in a predetermined minimum endpoint titer in a successful assay run. The acceptance criteria in this case should be validated appropriately.

Generally during in-study sample analysis, the inter-replicate precision of results (CV) of positive controls as well as test samples (with assay signal at or higher than the screening cut point) is also controlled using suitable acceptance criteria (per plate) to ensure that meaningful data are consistently obtained. Results below the cut point, however, may not be required to meet CV limit criteria. It is recommended that the acceptance criteria applied for precision be within the range commonly expected for immunoassays; these criteria should also be appropriate for the platform used (for example, the criteria may be tighter for electrochemiluminescence assays as opposed to ELISAs), guided by assay development data and experience with the technology platform and assay method.

Appendix F. Testing analyte stability in matrix

Comparing freshly prepared controls or samples with those that have been stored under conditions to be used for nonclinical or clinical studies can determine the stability of samples and controls. Based upon intended sample handling and storage conditions, common stability testing conditions include: freeze–thaw cycles, −20 °C storage, refrigerated (4 °C) storage, and the stability of diluted samples at specified storage temperatures. The evaluation of stability at or below −60 °C may not be required because literature references are available to support the stability of antibodies at temperatures ≤20 °C for 2 years or longer [29,30].

In general, the results of stored versus fresh positive controls should be compared. The recommended acceptance criteria for the stored controls are that they should generate titer values that fall

within a serial dilution of the titer of the fresh control. For example, when 2-fold serial dilutions are used for titration, if the titer of a fresh positive control is 40, then a titer of a stored positive control within 20–80 will be considered equivalent. Alternatively, this comparison between stored and fresh positive controls can be done using raw assay signal or concentration values, in which case mean recovery of 80–120% is typically considered acceptable. In addition, statistical analyses may be used to evaluate the significance of these comparisons. Furthermore, it may be necessary to repeat a stability sample with a result beyond its acceptance criteria to assure that the result was not anomalous and indeed due to analyte instability.

Due to heterogeneity of the ADA response, it should be noted that stability of the positive control may or may not reflect stability of an actual study sample. When possible, the stability of study samples should be determined by testing stored (incurred) samples and comparing the results of end point titer, raw assay signal, or calculated concentration (as appropriate for the assay) with the original results from the same samples recorded prior to the storage (i.e., when first tested).

Appendix G. Illustration of the screening cut point calculation

Tables 2a and 2b provide an illustration of the screening cut point calculation based on log transformed data values during the validation phase where each of two analysts test drug-naïve matrix samples in three runs each (total of six assay runs). Fifty (50) samples were assayed for the cut point determinations in the example depicted in Table 2. Two samples were identified as biological outliers, and the results of these two samples from all runs were excluded from the cut point determination. In addition, results for two other samples were deleted as analytical outliers from assay run 6. Therefore, the number of samples was 48, 48, 48, 48, 48, and 46, respectively, for the six runs. This differing sample size across runs was taken into consideration when calculating the pooled cut points.

Log transformation ensured approximate normality of the data (Appendix B.1). The assay run means were not significantly different, and the variances were fairly similar (Appendix B.2). As per Fig. 1, use of a fixed cut point is justified for the in-study phase of this assay; however, a floating cut point can also be used. The calculation procedure outlined in Appendix B.3 and the formula provided in step 2 of that section is illustrated in Table 2a. The calculation of normalization factor (correction factor) to be used for floating cut point during the in-study phase is illustrated in Table 2b. *Because the cut point calculation is performed on the log-transformed data, the normalization factor is additive in the log scale, and multiplicative in the untransformed scale.* During the in-study performance of the assay, one would multiply the mean of the negative control mean from each run by the multiplicative factor from Table 2b to derive the floating cut point for that run. Given the differences between the multiplicative factors from each analyst, it may be preferable to use the analyst-specific multiplicative factor, and hence the analyst-specific floating cut point in this scenario.

References

[1] H. Schellekens, Neurology 61 (2003) S11–S12.
[2] E. Koren, L.A. Zuckerman, A.R. Mire-Sluis, Curr. Pharm. Biotech. 3 (2002) 349–360.
[3] H. Schellekens, Nephrol. Dial. Transplant. 20 (2005) vi3–vi9.
[4] H. Schellekens, Clin. Ther. 24 (2002) 1720–1740.
[5] G. Shankar, E. Shores, C. Wagner, A.R. Mire-Sluis, Trends Biotech. 24 (2006) 274–280.
[6] Proceedings of the International Conference on Harmonization (ICH) of Technical Requirements for Registration of Pharmaceuticals for Human Use: Tripartite guideline, S6: Preclinical Safety Evaluation of Biotechnology-Derived Pharmaceuticals, July 1997. http://www.ich.org/LOB/media/MEDIA503.pdf.
[7] Committee for Medicinal Products (CHMP) For Human Use: Concept paper on immunogenicity assessment of therapeutic proteins, February 2006, EMEA/CHMP/BMWP/246511/2005. http://www.emea.europa.eu/pdfs/human/biosimilar/24651105en.pdf.
[8] A.R. Mire-Sluis, Y.C. Barrett, V. Devanarayan, E. Koren, H. Liu, M. Maia, T. Parish, G. Scott, G. Shankar, E. Shores, S.J. Swanson, G. Taniguchi, D. Wierda, L.A. Zuckerman, J. Immunol. Meth. 289 (2004) 1–16.
[9] S. Gupta, S.R. Indelicato, V. Jethwa, T. Kawabata, M. Kelley, A.R. Mire-Sluis, S.M. Richards, B. Rupp, E. Shores, S.J. Swanson, E. Wakshull, J. Immunol. Meth. 321 (2007) 1–18.
[10] E. Koren, H. Smith, W. Shores, G. Shankar, D. Finco-Kent, B. Rupp, Y.C. Barett, V. Devanarayan, B. Gorovits, S. Gupta, T. Parish, V. Quarmby, M. Moxness, S. Swanson, G. Taniguchi, L.A. Zuckerman, C. Stebbins, A.R. Mire-Sluis, J. Immunol. Meth. 333 (2008) 1–9.
[11] G. Shankar, C. Pendley, K.E. Stein, Nat. Biotechnol. 25 (2007) 555–561.
[12] V.P. Shah, K.K. Midha, S.V. Dighe, I.J. McGiveray, J.P. Skelley, A. Yacobi, T. Layloff, C.T. Viswanathan, C.E. Cook, R.D. McDowall, K.A. Pittman, S. Spector, Pharm. Res. 9 (1992) 588–592.
[13] U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research, Center for Veterinary Medicine. Guidance for Industry: Bioanalytical Method Validation (May 2001). http://www.fda.gov/CDER/GUIDANCE/4252fnl.pdf.
[14] M. Wadhwa, C. Bird, P. Dilger, R. Gaines-Das, R. Thorpe, J. Immunol. Meth. 278 (2003) 1–17.
[15] J.W. Findlay, W.C. Smith, J.W. Lee, G.D. Nordblom, I. Das, B.S. DeSilva, M.N. Khan, R.R. Bowsher, J. Pharm. Biomed. Anal. 21 (2000) 1249–1273.
[16] I. Goldschneider, E.E.C. Gotschlich, M.S. Artenstein, J. Exp. Med. 129 (1969) 1307–1326.
[17] N. Andrews, R. Borrow, E. Miller, Clin. Diagn. Lab. Immunol. 10 (2003) 780–786.
[18] R. Borrow, N. Andrews, D. Goldblatt, E. Miller, Infect. Immun. 69 (2001) 1568–1573.
[19] J. Zhao, R. Yang, Z. Cui, M. Chen, M.H. Zhao, H.Y. Wang, Nephron Clin. Pract. 107 (2007) c56–c62.
[20] N.L. Lamb, B.M. Patten, Muscle Nerve 14 (1991) 1021–1027.
[21] D.M. Jensen, I.G. McFarlane, B.S. Portmann, A.L. Eddleston, R. Williams, N. Engl. J. Med. 299 (1978) 1–7.
[22] M.R. Halili, I. Spigland, N. Foster, N.A. Ghossein, J. Surg. Oncol. 10 (1978) 457–463.
[23] International Conference on Harmonization (ICH) of Technical Requirements for Registration of Pharmaceuticals for Human Use ICH Harmonized Tripartite Guideline, Q2A – Text on validation of analytical procedures, March 1995. http://www.fda.gov/cder/Guidance/ichq2a.pdf.
[24] International Conference on Harminization (ICH) of Technical Requirements for Registration of Pharmaceuticals for Human Use ICH Harmonized Tripartite Guideline –Q2B Validation of analytical procedures: methodology, November 1996. http://www.fda.gov/Cder/Guidance/1320fnl.pdf.
[25] The United States Pharmacopeia and National Formulary 2002 (USP 25/NF 20). <1225> Validation of compendial methods. United States Pharmacopeial Convention, Inc., Rockville, MD, U.S.A. http://www.pharmacopeia.cn/v29240/usp29nf24s0_c1225.html.
[26] L. Neyer, J. Hiller, K. Gish, S. Keller, I. Caras, J. Immunol. Meth. 315 (2006) 80–87.
[27] B. DeSilva, W. Smith, R. Weiner, M. Kelley, J. Smolec, B. Lee, M. Khan, R. Tacey, H. Hill, A. Celniker, Pharm. Res. 20 (2003) 1885–1900.
[28] D. Geng, G. Shankar, A. Schantz, M. Rajadhyaksha, H. Davis, C. Wagner, J. Pharm. Biomed. Anal. 39 (2005) 364–375.
[29] N.S. Fineberg, S.E. Fineberg, J.H. Anderson, M.A. Birkett, R.G. Gibson, S. Hufferd, Diabetes 45 (1996) 1750–1754.
[30] E. Harlow, D. Lane, Antibodies: A Laboratory Manual, Cold Spring Harbor Laboratory, New York, 1988, p. 119.