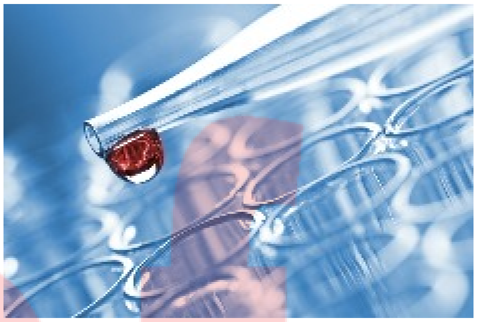
Biopharmaceutical Best Practices Association

Assay Acceptance Criteria for Multiwell-Plate–Based Biological Potency Assays

C. Jane Robinson, Michael Sadick, Stanley N. Deming, Sian Estdale, Svetlana Bergelson, and Laureen Little



For most biopharmaceuticals, potency is assessed in a bioassay by comparing the dose–response curve of the test material with that of a reference standard. As with all analytical techniques, such assays require criteria by which their execution can be judged objectively to be valid, regardless of whether the desired or expected result is obtained for the test sample. The purpose of this paper is to provide guidance on setting assay acceptance criteria (AAC) for potency assays using multiwell plates.

Multiple components of the overall assay system — from instruments to incubation media — need to be within defined limits to permit execution of a valid assay, so they are tested for suitability either before or during the assay. Because of the complexity of bioassay systems, many relevant factors may not be controlled. Thus, it is necessary to rely strongly on analysis of data produced by each individual assay to determine whether that assay was executed correctly. This paper discusses criteria that can be applied to the assay results and the importance of assay design in selecting useful acceptance criteria. A draft version of the paper was published for consultation (1) and the contents were presented at international scientific meetings. This revised version takes into account the comments received during the consultation.

KEYWORDS:
Acceptance Criteria, Potency Assays, Bioassays, Immunoassays, Reference Standards, Control Samples, System Suitability, Product Release

SCOPE

WHY MULTIWELL-PLATE–BASED ASSAYS SPECIFICALLY? Analytic dilution assays using multiwell plates constitute the most common platform for measurement of biological activity by in vitro bioassay or measurement of immunoreactivity by immunoassay. The multiwell format provides a convenient means of handling the necessary number of doses and replicates, and it is supported by the availability of a wide range of plate types with a standardized footprint and supporting equipment and measurement systems. Multiwell-plate formats introduce specific artifacts to the measured responses. Thus, assay design — especially the positioning of individual samples within and between plates — is crucial to obtaining valid results and needs to be considered when setting acceptance criteria for an assay.

Much of the following discussion is based on cell-based assays in 96-well plates, which reflects their wide use in potency testing. With some adaptation, however, many of the ar-

ABBREVIATIONS AND TERMS

4PL: four-parameter logistic nonlinear regression model

AAC: assay acceptance criteria

CV: coefficient of variation

ED$_{50}$: the dose of a drug that induces a response that is 50% of the maximum response

PLA: commercial software package designed for the analysis of bioassay and immunoassay dose-response curves (Stegmann Systems GmbH)

R^2: coefficient of determination

RSS: the residual sum of squares

SAC: sample acceptance criteria

Shewhart SPC: control chart named after Walter A. Shewhart

SoftMax Pro: Data Acquisition and Analysis Software (Molecular Devices LLC)

SPC: statistical process control

StatLIA: commercial software package designed for the analysis of bioassay and immunoassay dose-response curves (Brendan Technologies, Inc.)

Summary

This paper is restricted to discussion of tests applied to the responses of a bioassay system to reference standards, controls and test samples obtained during performance of a potency assay.

Assay acceptance criteria are based primarily on comparison of dose–response curves of control samples with a reference standard, all of which should be well characterized in the assay system.

At least one control sample should be known to behave similarly to the reference standard in the assay system. Both this control and the reference standard should be known to behave similarly to the expected behavior of test samples.

We propose the name assay control sample for the control material that behaves similarly to the reference standard and the expected behavior of the test samples.

The origin of the assay control sample should be as independent of the reference standard and test samples as is possible within the constraint that all behave similarly in the assay system.

For plate-based biological potency assays, we propose the following two separate sets of acceptance criteria: assay acceptance criteria (AAC) and sample acceptance criteria (SAC).

We propose a two-level, sequential assessment of acceptance criteria. First, AAC are assessed for the assay, or subsection of the assay. Failure means that the assay, or subsection, is invalid. There is no processing of the corresponding test sample data. Passing AAC allows processing of test sample data. Second, each test sample potency determination is then subjected to its own SAC. If it passes, then that test sample potency measurement is valid. If it fails, then that particular test sample potency quantification fails. Other test sample determinations may still be valid.

Similarity of dose–response curves of reference standards and assay control samples is an essential AAC. Similarity of dose-response curves of the reference standard and test sample is an essential SAC.

AAC and SAC applied to an assay should be demonstrated to be useful in judging the validity of the assay. Both the criteria and the limits set on their values should be reviewed — and modified, if appropriate — as more assays are performed and further data are accumulated.

guments can be applied to other assay systems and plate formats for plate-based potency assays.

What Types of Assay? Considerations for setting AAC apply to a number of different assay types, including

- functional assays (in which a biological response is measured)
- biochemical assays (such as clotting-factor activity assays), considered as functional assays or as a separate class
- binding assays (that measure binding of a ligand, receptor, cofactor, and so on), which may be potency assays, in which binding is the mode of action (MOA), or surrogate potency assays
- Immunoassays such as enzyme-linked immunosorbent assays (ELISAs) that measure binding of an antibody preparation (monoclonal or polyclonal) to one or more epitopes — distinct from functional assays in which binding of an antibody induces a functional biological response
- hybrid assays (e.g., immunoassay-cofactor binding assays).

Many assays consist of multiple steps. For example, a functional assay may involve stimulation of cytokine secretion followed by immunoassay measurement of the secreted protein. AAC are generally set for the overall output, but may, in addition, be set for some individual steps. Potency assays for some types of products — e.g., advanced therapy medicinal products (ATMPs) — may require atypical assay designs and, consequently, atypical acceptance criteria.

AAC or System Suitability? Guidelines for testing a system to demonstrate its suitability for an analytical procedure have been developed primarily for physicochemical techniques. ICH Q2(R1) states that “a series of system suitability parameters (e.g., resolution test) is established to ensure that the validity of the analytical procedure is maintained whenever used” and then provides specific examples for liquid and gas chromatography (2). USP chapter <621> on chromatography states, “To ascertain the effectiveness of the final operating system, it should be subjected to a suitability test prior to use. The essence of such a test is the concept that the electronics, the equipment, the specimens and the analytical operations constitute a single analytical system, which is amenable to an overall test of system function” (3). A useful definition, also from the field of liquid chromatography, is: “System suitability is the checking of a system to ensure system performance before or during the analysis of unknowns” (4). So a system suitability test determines whether an analytical system is fit for use. In physicochemical analyses, making that decision may be possible before test samples have been committed to analysis.

As with physicochemical analyses, multiple components of the overall assay system need to be within defined limits to permit execution of a valid bioassay. For example, instruments, media, and environmental conditions are tested for suitability either before or during an assay. However, guidelines and examples designed for physicochemical techniques are not necessarily appropriate or sufficient for bioassays. Bioassays and (generally to a lesser extent) immunoassays tend to be susceptible to a greater number of factors than are most physicochemical analytical techniques. Some of those factors may be poorly controlled or not identified. For example, biological media may contain unidentified or undetected components that vary from batch to batch and affect the response of an assay system. Because such uncontrollable or unidentified sources of variability can cause assay-to-assay variability, it is necessary to rely strongly on analyzing data produced by each individual assay to determine whether that assay was executed correctly. Which tests of these data will be useful in determining the validity of the assay should be investigated during assay development and characterization.

The general term “acceptance criteria” has been defined to be “conditions which must be fulfilled before an operation, process or item, such as a piece of equipment, is considered to be satisfactory or to have been completed in a satisfactory way” (5).

This paper is not intended to be a rigid set of rules, but rather to identify the major issues to be addressed in setting appropriate AAC and SAC. Unusual and novel assay systems and designs may require additional factors to be considered.

The importance of justifying the selection acceptance criteria and the limits set on their values using data from assay development, validation and on-going monitoring and trending is emphasized as is the importance of providing clear definition of terms used.

What is the distinction between AAC and system suitability tests? The former are often considered to be a subclass of the latter. Both are set based on data acquired during assay development, characterization, and validation. System suitability may be checked either before or during the performance in which a test sample is assayed. The term “assay acceptance criteria” is generally used to describe conditions that must be met by data derived from an actual assay in which a test sample is assayed.

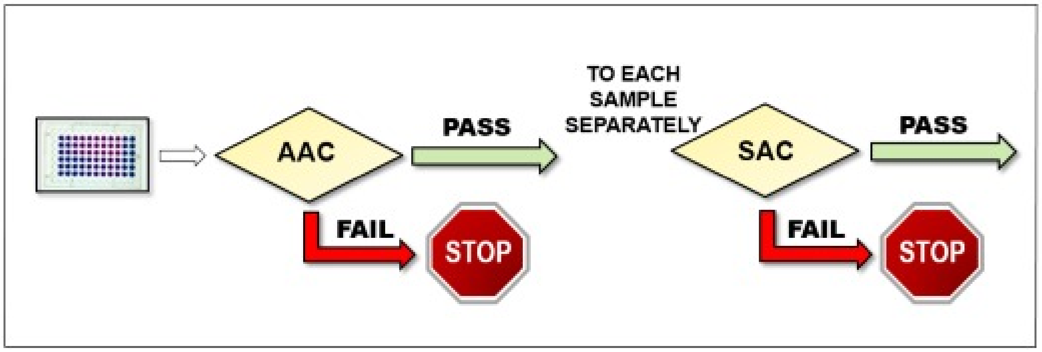
Because biological assay systems are more variable than physicochemical systems, it is generally necessary for more system testing to be run simultaneously with (rather than prior to) sample testing. For potency assays using multiwell plates, some tests are generally considered to measure system suitability and others are generally considered to assess assay acceptance criteria. The important consideration for both system suitability tests and AAC is that they should be appropriate to the specific assay system, the precise purpose of the particular assay being performed, and the assay design. Classification of a particular test as a system suitability test or AAC may not be necessary or particularly helpful.

This paper is restricted to discussion of tests applied to data derived from the responses of test samples, controls, and reference standards obtained during performance of an assay intended to yield a potency value for a test sample.

Pharmacopeial Requirements: With pharmacopeial assays, all specified AAC need to be met for those assays performed under relevant quality systems. The specified criteria can provide useful guidance in setting criteria for similar assays and similar products. However, pharmacopeial assays generally specify fewer AAC than would be considered necessary for assays used for product release, so manufacturers would probably need to set additional AAC for release assays and develop appropriate criteria for novel assays and products.

ASSAY ACCEPTANCE AND SAMPLE ACCEPTANCE

For plate-based biological potency assays, we propose two separate sets of acceptance criteria: AAC and sample acceptance criteria (SAC). The former are based on responses of control samples and reference standards, and the latter are based on the responses of each separate test sample. Making this distinction permits the validity of an assay (or assay subsection) to be judged separately from that of each separate test sample. If an assay fails, then there is no processing of its test sample data. If analysis of data from one test sample fails the SAC, then that particular test sample potency quantification fails. Other test sample determinations should be assessed separately and may still be valid.



Two separate sets of acceptance criteria are applied to the bioassay data: Assay Acceptance Criteria (AAC), based on responses of control samples and reference standard, and Sample Acceptance Criteria (SAC), based on responses of each separate sample. If the plate/block/assay fails AAC, then there is no processing of test sample data from that plate/block/assay. If one test sample fails its SAC, then that particular test sample potency quantification fails. Other test samples are assessed separately.

SAC for the bioassay should not be confused with product specifications. For example, a test sample might meet SAC for the bioassay and a valid potency measurement be obtained, but it would not necessarily follow that the product met specifications: the measured potency of the sample might indicate that the potency of the batch of product lay outside the limits of the product specifications. There have been suggestions that replacement of the word “acceptance” by “validity”, giving the term “sample validity criteria”, could avoid any such confusion. However, the word “validity” has specific connotations and could be problematic for some groups. We propose that the term “sample acceptance criteria” be used as the default and that if an alternative term is used, it should be clearly defined in the assay protocol and other relevant documentation.

A particular feature of multiwell-plate–based assays is well-to-well variability within a plate and plate-to-plate variability. This imposes constraints on assay design and consequently on the application of AAC and SAC. AAC may include criteria that are applied to subsections of an assay, for example, an individual plate, a block of plates, or a block of wells within a plate. These subsets of AAC may be referred to as “plate acceptance criteria”, “block acceptance criteria”, and so on. If any such terms are used, they need to be defined in the assay protocol or other appropriate documentation. This is discussed further in the section “Assay Design”.

ASSAY ACCEPTANCE: AAC are based on control samples and a reference standard. These are materials that are well characterized in the assay system and are independent of the test samples. At least one control sample should be known to behave similarly to the reference standard in this assay system, and both that control and the reference standard should be known to behave similarly to the expected behavior of the test samples. (See the section "Similarity of Dose–Response Curves" below.)

The control material that behaves similarly to the reference standard and test samples is often called the assay control, control material, or control sample. We propose adopting the term "assay control sample". Other types of controls include those designed to exclude responses of a system that are not caused by specific actions of the test samples and those designed to demonstrate a positive response by a different mode of action. The former are often called negative controls and include materials such as excipients of the test sample or antibodies of a different specificity. Whereas the assay control sample and reference standard must be tested at multiple dilutions, other controls may be tested at only a few concentrations or a single concentration.

AAC are based primarily on comparing dose–response curves of the assay control sample(s) with a reference standard. The origin of the assay control sample should be as independent of the reference standard and test samples as is possible within the constraint that all should behave similarly in the assay. Different lots of material from the same production method are commonly used. Two independent dilutions of the reference standard material (one serving as the reference and one as the assay control sample) are not sufficient; using two dilution series of the same material is testing only the dilution procedure and subsequent handling. In early assay development, however, an independent assay control sample may not be available, so using two samples of the reference standard could be the only option. In this case, the two samples should be processed as independently as possible, ideally starting from separate vials — and, if possible, using alternative intermediate dilution schemes.

Ideally, the assay control sample and reference standard should be as similar as possible to the test samples in excipient formulation, concentration, aliquot volume, container, and so on. The assay control sample, reference standard, and test samples should be prepared and tested in exactly the same way. However, an initial step such as reconstitution of a lyophilized reference standard or dilution of a more concentrated solution may be necessary. Commonly, the reference standard may be presented in a formulation buffer different from that of the test samples to permit its storage as a frozen solution. It then requires dilution (concentration permitting) in the formulation buffer of the test samples to render it as similar as possible to the test samples. Similarly, the formulation of the assay control sample and other control samples should be brought as close as possible to that of the test samples. Minimizing any differences in the treatment of the various samples and reference standard reduces the possibilities for the assay procedure to be responsible for any differences in the measured potencies.

The assay control sample must be compared with the reference standard and evaluated in terms of curve shape and potency. Assay control sample potency is evaluated against a statistically generated assay control chart. The concept of a control chart for the assay control sample is essentially simple: Multiple independent executions of an assay are performed in which the assay control sample is included as a test sample, with no preexisting expectations other than curve similarity to the associated reference standard. After a certain number of assay repetitions (e.g., n = 25 or n = 50), the arithmetic average and associated standard deviation (SD) of that value are determined for the assay control sample potency. A control chart acceptance criterion is then established using the mean assay control sample potency, with plus or minus some multiple of the SD defining the upper and lower limits of
acceptability.

Although it is common practice to set statistical process control (SPC) limits at plus or minus three times the estimated SD ("Shewhart limits"), in many bioassay applications the estimated SD is based on only a few measurements and is thus a very uncertain estimate. An approach to setting SPC limits that takes this uncertainty into account is to use tolerance intervals (6). These intervals are wider when based on only a few measurements and (usually) become narrower as more and more data are acquired until, with an infinite amount of data, they become the same as the ±3 SD limits associated with typical Shewhart SPC charts.

In the article "Specification Setting: Setting Acceptance Criteria from Statistics of the Data" (7), upper and lower control limits (tolerance limits) are provided for as few as five assays (with a corresponding multiplier of ±10.75) to as many as 199 assays (with a corresponding multiplier of ±3.03). Using the table provided in this reference, one can establish a reasonable initial control chart with 25 assays (with a corresponding multiplier of ±4.05), and then revise those upper and lower control limits as the number of collected values for the assay control sample relative potency increases.

Sample Acceptance: SAC are tests applied separately to each test sample. Passing or failing of each test sample is independent of other test samples. Similarity of sample curve shape to that of the reference curve is a standard test. Other SAC (e.g., variability of replicates) are usually similar to those for reference and assay control samples, but some may be different. For example, with some assay systems (particularly those that may not use the whole dose–response curve), the SAC may allow for fewer doses to be used in the curve fit than for the reference standard or assay control sample. Limits may be set for the ED_{50} of test samples to ensure that the dose–response curve falls within the validated dose range.

The reportable value is defined in the testing protocol. It is typically the mean of a specified minimum number of results from valid independent assays, with "independent" being defined in the protocol. Most protocols specify that failure of the assay to meet AAC or of a sample to meet SAC more than a specified maximum number or percentage of times will trigger specified actions. It must be recognized that on statistical grounds a certain percentage of assays or samples would be expected to fail, with the actual percentage expected to fall outside the limits of the acceptance criteria depending on the specified range.

ASSAY DESIGN

Use of multiwell plates introduces particular sources of potential artifacts in the measured responses of an assay. Wells in the corners, on the edges, and in the center of each plate have different environments, and the individual plates may be subject to slightly different conditions, any of which differences may affect responses. These effects are observed so commonly that the terms “well effects”, “edge effects”, and “plate effects” are widely used in the bioassay field. In many assays — particularly cell-based assays with long incubation times — the edge effects are so significant that often only the 60 center wells of each 96-well plate are used for measurements and the edge wells are filled with media or other solutions. Evaporation of medium from external wells, even in a humidified atmosphere, is a major contributor to edge well effects. Recently developed 96-well plates with reservoir troughs around the edge can reduce the edge effect and, depending on the assay,

Most Commonly Used AAC and SAC

Similarity of dose–response curves is an absolute requirement for determining relative potency. In current practice, some form of assessment is almost universally applied. It is recommended that at least one control sample known to behave similarly to both the reference standard and the expected behavior of test samples in the particular assay system (the assay control sample) be run as a full dose–response curve. Similarity of dose–response curves for the reference standard and assay control sample is an essential assay acceptance criterion, and similarity of dose–response curves of the reference standard and test sample is an essential sample acceptance criterion. Depending on the assay, whole or partial response curves (when doses do not permit an asymptote to be reached) and linear responses may be measured, and various combinations of criteria can be applied (see below). 4PL is commonly used for curve analysis. In-house programs and commercial software (e.g., PLA, StatLIA and SoftMax Pro) are used. In some cases, the F-test is used initially to judge similarity and then, as historical data are acquired, it is replaced by equivalence testing. Against the recommendations of this paper, some assays do not include an assay control sample or other controls. In such cases, AAC are based on the reference standard dose–response curve or comparison of test sample and reference standard.

Curve Slope: Slope is an acceptance criterion applied in most assays, but in a few cases it is used for trending only. Most often, the ratio of slope is used, forming part of curve similarity testing. Limits on the ratio can be defined as confidence intervals or ranges (e.g., 0.80–1.25). In some cases, absolute values are defined for the slopes and used either in addition to the ratios or alone. Absolute values are based on validation or control chart data (e.g., mean ±3 SD from historical data). When setting limits of ±3 SD, care should be taken to use a sufficiently large data set. Alternatively, tolerance intervals may be used. AAC limits can differ from those for SAC.

Lower asymptote is a useful AAC and SAC for some assays. Limits can be set as ratios between the values for the reference standard, assay control sample and test samples, forming part of similarity testing, and as absolute values based on historical data. Lower asymptote values may be used only for trending.

Upper Asymptote: The same comments apply as for the lower asymptote, with the additional point that an upper limit may be placed on the value to account for instrument limitations, particularly when optical density (OD) read-outs are used. In the case of OD, when the read-out is the difference between OD at two wavelengths, care should be taken that neither exceeds the limit of the instrument.

Goodness of Fit: Goodness of fit is an essential AAC and SAC and is almost universally applied. Individual replicates or the means can be fitted. R^2 is widely used, with limits commonly set from $R^2 = 0.95$ to $R^2 = 0.98$ for replicates, and possibly tighter if fitting the means. If replicates are fitted, R^2 will reflect the spread of the data as well as the model fit. Testing R^2 is relatively simple, but R^2 may not provide a very sensitive indicator of goodness of fit. During assay development, for selecting the most appropriate model, it can be useful to plot residuals against dose to reveal whether there is a random or a systematic deviation from the model. Residual sum of squares (RSS) and chi-squared are also used, the latter primarily for quantal assays. The most appropriate model and test depend on the individual assay and should, as with all acceptance criteria, be checked during routine use.

Potency of Control: With the exception of a few unusual assays, an assay control sample (run as full dose–response curve) should be included in every assay — usually on every plate. At present, some assays do not include any control samples and base the AAC on the reference standard alone or in comparison with the test sample. Other assays include a positive control at only one dose level or a few dose levels. Limited space on assay plates is one reason for the latter case. Limits ranging from 70–130% (or 70-143%) to 90–110% (or 90-111%) are typical, with 80–125% being very common (for some assays, eg viral potency, with dilution steps possibly on a log10 scale, potency limits may be much broader). Limits are often set based on historical data (e.g., mean ± 3 SD). When setting limits of ±3 SD, care should be taken to use a sufficiently large data set. Care should be taken to define assay blanks and negative controls. Blanks usually refer to the absence of the active pharmaceutical ingredient (API), corresponding to the addition of a sample of the formulation buffer. Negative controls usually consist of relevant substances without a specific action in the particular assay (e.g., in the assay of a therapeutic antibody, a negative control might be antibody without specific binding to the ligand used in the assay).

(Continued)

may permit use of all 96 wells of each plate.

During assay development, use of plate uniformity tests (the same dose applied to all wells, including the edge wells) can facilitate assay optimization by permitting assessment and possibly reduction or elimination of plate-positional effects, including edge anomalies. Nonrandom distribution of samples and doses — for example, placing all dilution curves for a particular sample in edge rows of the plates or reference standard in one plate and test sample in another — can introduce bias to measured relative potencies. Good assay design should reduce such bias. Completely random distribution of samples and doses is generally not feasible, so block or other structured designs can be used (8).

Within plates, replicate dilution series of the reference standard, assay control sample, and test samples should be placed in nonequivalent and nonsimilar positions (avoiding locating replicates of the dilution series all in edge wells, or all in center wells, in adjacent rows, all in the top of the plate, all in the bottom of the plate, and so on). Most assays are constrained by the number of wells and plates that can

Most Commonly Used AAC and SAC (Continued)

Variability of Replicates: Assessing the variability of replicates is essential, with limits normally set as AAC and SAC. Replicates commonly consist of the following:

- Replicate wells containing the same solution (aliquots of the same dilution point in a single dilution series)
- Replicate dilution series from a common starting solution
- Replicate aliquots, vials, ampoules, and so on.

The data may be analyzed from one multiwell plate or across several plates.

When variability of replicates is quoted, what is replicated must be defined exactly. In the first case above, the variability measures the errors in pipetting solution into the wells and errors introduced in subsequent steps. It does not assess errors in preparation of the starting solution or the dilution series. For measuring relative potency, ideally the handling of each sample should be the same, and each replicate should capture independently as many steps as possible from the handling process.

Variability depends on the assay. It is commonly expressed as %CV. For enzyme-linked immunosorbent assays (ELISAs) and cell-based potency assays for products such as cytokines or MAbs, AAC and SAC are generally set at %CV 10–30%. For vaccine assays, in which dilution series steps of $\log_{10}$ may be used, considerably broader limits may be acceptable depending on the protocol and requirements of a given assay. An alternative option is to set a SAC as the relative 95% confidence interval around the mean value for test sample relative potency (e.g., at 80–125% or 75–133%).

The requirements for the limits on assay variability are often subject to requirements for the precision of the assay result. If greater precision is required, one approach can be to modify the assay protocol to include more replicates.

Although variability between replicate wells of one dose (measured by CV) is the most widely used assessment of replicate variability, other assessments may be used (e.g., variability between replicate potency determinations from single dose curves) depending on the assay design. Variability of replicates is one of the most important tools for trending.

Minimum Number of Doses Used in Curve Fitting: Usually set as an AAC and SAC, the value depends on the particular assay and whether it is a linear or full curve fit. Usually all doses are required, but in some cases exclusion of one or more points is permitted (see “Maximum Number of Statistical Outliers Excluded” section below). Depending on assay design, this generally results in a minimum of 6–10 doses, with 8 being very common. For some assays (e.g., those in which wide ranges of test sample potencies may be encountered) a smaller number of doses, specified as being consecutive, may be set as a SAC.

Minimum Number of Doses in “Linear” Part of Dose–Response Curve: Usually set as an AAC and SAC, both for linear and full curve fits; values generally range from 3 to 6, with 3 and 4 being most common.

Minimum Number of Doses in Upper and/or Lower Asymptote: This criterion is used in some assays with the most common value being 2. This criterion is sometimes used during development and subsequently discarded.

Minimum Dose Range Used in Curve Fit: Sometimes set explicitly as an assay acceptance and sample acceptance criterion, this commonly specifies that a minimum dose range (e.g., 50–200%) should lie within the linear range. In many assays, the minimum dose range is set implicitly as an acceptance criterion because the protocol specifies doses tested and the minimum number of doses used in curve fitting.

Maximum Number of Statistical Outliers Excluded: For many assays, a maximum number of statistical outliers that can be excluded is set as an assay acceptance and sample acceptance criterion. This can be defined as the number of doses (all replicates) or individual points that can be excluded per curve, or the number of replicates per dose, or a combination of these. The statistical test used to identify an outlier should be defined and have been shown to be suitable for the assay. Limits on the number of permitted exclusions depend strongly on the assay. Typical examples are one dose or one point per 8-point curve or per sample. Many assays allow no exclusion of outliers; others allow exclusion only if an anomaly can be attributed to experimental error. To prevent bias, any point suspected of being subject to experimental error should normally be excluded before its value is determined. However, it is recognized that experimental error may sometimes be identified only after a point is observed to be anomalous.

Less Commonly Used AAC and SAC

ED_{50}: The ED_{50} is an absolute value. ED_{50} is widely used for trending but not commonly used as an AAC or SAC. The ratio of ED_{50} of reference to ED_{50} of assay control sample or test sample is sometimes used to assess potency.

Inflection Point: The inflection point is not commonly used as an independent AAC or SAC, but proves useful in a few cases. In 4PL, this corresponds to the ED_{50}.

Ratio of Upper Asymptote to Lower Asymptote: an absolute value with limits can be assigned to this ratio, or limits can be set on the ratio between a reference and assay control sample or test sample of values for this ratio (ratio of ratios). The ratio of upper to lower asymptote is sometimes used for trending (e.g., indicating changes with cell passage number). The ratio of upper to lower asymptote may, or may not, be meaningful in reflecting the underlying biology of the response. Use of the ratio as an acceptance criterion should be considered with caution.

Ratio of Values for Different Control Samples: Use of more than one positive control is not common, but the ratios of high to medium or high to low positive controls, for example, can be useful AAC. The value of the ratio of positive to negative control responses will vary greatly with changes in response of the negative control if the negative control is very small and the ratio may not reflect the underlying biology of the response.

be used, so assay design is a compromise between the ideal and the feasible. The adequacy of a proposed design should be investigated during assay development and characterization.

The assay design determines the selection of the AAC and SAC. For example, if replicate curves are made on a single plate, then acceptance criteria might include limits on the coefficient of variation (CV) for the slopes of the replicate curves. If separate potency determinations are made from each plate in a multiplate assay, then acceptance criteria limits might be set on the CV of these potency determinations. Commonly used acceptance criteria and some typical assigned values are listed in the table “Most Commonly Used AAC and SAC”.

Commonly, plates are treated independently, with a reference standard and an assay control sample dilution series included on each plate, and a set of AAC (possibly referred to as “plate acceptance criteria”) applied separately to each plate. Failure of the plate to meet these AAC means that there is no processing of the test sample data from that plate. Alternatively, replicate curves of reference standard and assay control sample may be analysed across several plates in a block or the whole assay. The AAC can specify whether individual plates can fail or the block (or assay) fails as a whole. Similarly, SAC may be applied to replicate curves within a plate or across several plates, depending on the assay design.

In most cases, plate-to-plate variability requires that a reference standard and an assay control sample dose dilution series be included on every plate, whether or not AAC are applied to each plate independently. There may be cases where it is possible to demonstrate sufficiently low plate-to-plate variability that several plates can be treated as a block within which only some plates contain a reference standard and an assay control sample dilution series and the block passes or fails the AAC (possibly referred to as “block acceptance criteria”) as a single unit. This reduces the proportion of wells required for reference standard and assay control sample, increasing the number of wells available for test samples, but it does not commonly prove a suitable design.

AAC may have different limits when applied to subsections of an assay (e.g. plate acceptance criteria) compared with those applied to the assay as a whole.

Definition of terms: A problem common to many analytical techniques is lack of a common interpretation of terms such as “test”, “assay” and “assay run”. Clear definition of such terms, how they apply to a given assay design and how they relate to the reportable value, should be included in the assay protocol and other appropriate documentation.

Necessity for an assay control sample: In the past, some assay designs have not included an assay control sample. In such cases, assay acceptance would be based solely on the reference standard dilution curve. As stated in the section “Assay Acceptance”, in cases where an independent assay control sample is not available, such as early assay or product development, it may be necessary to use two samples of the reference standard. However, these should be processed as independently as possible and an assay control sample should be developed as soon as possible.

It is recognized that inclusion of an assay control sample can reduce the number of test samples or the number of replicates that can be included in an assay. This has been presented as an argument for using data from the test samples to judge assay validity: if the results from the test samples are as expected, the assay is judged to be valid. This approach is completely contrary to the concept of determining assay validity objectively, independently of whether the expected or desired results are obtained for the test samples.

Another argument that has been presented is that a problem with the assay control sample could cause an assay to fail the AAC when the test sample results would have been valid. This risk and cost has to be weighed against the risk and cost of a false result being accepted for a test sample due to an undetected problem with the assay. In general, the consequences of accepting a false result for a test sample are more serious than those of falsely failing an assay.

During routine analysis, when the test sample result is as expected, the usefulness of an assay control sample may not be apparent. Its value is more apparent when an assay performs outside, or at the extreme of, the normal parameters. In this case, understanding the performance of a sample which is similar to the test sample, but with well-characterized response characteristics in the assay, greatly enhances the ability to perform a root cause analysis and decide with confidence whether the test sample result is true.

The use of an independent assay control sample, which has historical data on its performance in the validated assay system, with a defined range of acceptable values, provides a rigorous and objective means of assessing the assay validity and is essential for statistical quality control.

MONITORING AND TRENDING

Values observed for assay (and sample) acceptance criteria can be used for assay (and sample) monitoring and trending. For the simplest case of assay monitoring, visual inspection of run charts (trend charts) displaying observed data in a time sequence can reveal aberrant results and indicate whether an assay is stable or drifting. Comparison with the results of system suitability tests (e.g., cell density before harvest for an assay or degree of cell confluence before dosing) and records of critical reagents and equipment can help reveal the cause of drift or aberrant results.

In addition to assay and sample acceptance criteria, some parameters are reported "for information only." These parameters do not determine the acceptance or failure of an assay or sample, but they can be useful in monitoring and trending. Monitoring and trending of as wide a range as feasible of assay response characteristics is useful, particularly in the early stages of assay development and use. As data accumulate over a large number of assays, incorporating variability in operating conditions (such as reagent batches and operators), it can become evident which characteristics may be useful for continued monitoring and trending and which should be adopted or retained as acceptance criteria.

Inclusion of an assay response characteristic in monitoring or trending should not carry an automatic expectation that it should be adopted as an acceptance criterion and have limits set. Such an expectation can discourage the investigation of response characteristics and eventual selection of optimum sets of acceptance criteria.

When statistical analysis is applied to observed values, a statistical process control (SPC) chart can be obtained, permitting an objective analysis of the variation in assay performance and allowing limits to be set to indicate when action must be taken to prevent an assay drifting into failure. "The use of statistical control charts to map the ongoing performance and stability of reference material during routine assays can be a useful quality control tool allowing for early detection of adverse trends" (9) and "SPC charts are a powerful tool for showing auditors the continual validation of an assay" (10).

EVOLUTION OF ACCEPTANCE CRITERIA DURING DEVELOPMENT

Assay and sample acceptance criteria (and their assigned values) will change during assay and product development processes as a result of improvement in assay performance, accumulation of more data on that performance, and stricter requirements at later stages of product development. Even when an established assay platform is applied to an additional product, it should be expected that there will be a development process for establishing assay and sample acceptance criteria.

During assay development and on accumulation of performance data, it would normally be expected that limits on some criteria would be tightened. However, it may become apparent that initial data collected over a short period of time did not reflect the full variation in assay conditions (e.g., variation in reagent batches) that will be encountered over a longer period. The criteria may therefore be initially set too tightly and subsequently need to be widened.

Additional criteria may be included later on. Conversely, with accumulation of data, it may become evident that some criteria initially set do not reflect the assay's validity. These criteria should be removed as they do not contribute to assessing assay validity, and their inclusion could result in rejection of assays that are fit for purpose. Their values can be recorded for the purposes of monitoring and trending.

ACCEPTANCE CRITERIA

Absolute and Relative Values: A potency assay is comparative, measuring the potency of a test sample or control sample relative to that of a reference standard. A comparative assay is based on the assumption that a factor that affects the response of a system to a test sample should affect the response of that system to a standard sample equally. Therefore, absolute values for characteristics of a response curve should not be critical. In practice, however, most assay systems function adequately over only a limited range of conditions reflected by a limited range of acceptable values for some of the dose–response curve characteristics. An unusually high or low value for the ED_{50}, for example, can indicate that an assay system is not behaving as usual. Limits might be set as acceptance criteria and/or as action limits in the process control.

Similarity of Dose–Response Curves: The response of a bioassay system to a sample can be affected by a variety of external factors, some of which may not be controlled, so potency measurement cannot be an absolute value. Bioassays are comparative, with the biological activity of a test material measured relative to that of a reference preparation (11). If two preparations are sufficiently similar, then their responses should be affected equally by any variation in the system, and relative potency should remain constant. It is a fundamental requirement for obtaining a valid relative potency that the reference standard and test sample must behave similarly in the assay system and hence their dose–response curves must have the same mathematical form. Any displacement between these curves along the log-concentration axis must be constant at all responses. This constant displacement is used to estimate relative potency. Nonsimilarity of two preparations may lead to dose–response curves of different mathematical forms with variation in the magnitude of displacement between the curves. In this case, any attempted measurement of relative potency would vary depending on the response level at which it was measured.

Similarity of dose response is thus an essential assay acceptance criterion. To determine whether a reference standard and test sample demonstrate the same dose–response relationship in a given bioassay, it is necessary to measure the response of each at several doses spread over an appropriate range. Assessing the similarity of dose–response curves depends on statistical analysis of the data. For some pharmacopeial assays, the method of analysis may be specified precisely. In many cases, however, only general guidance can be obtained from pharmacopeial sources. It is beyond the scope of this paper to discuss the relative merits of different statistical methods, such as the F-test and equivalence testing, for assessing similarity in various circumstances. Further information can be found in the literature (12-16).

For many well-characterized bioassays, a mathematical transformation of the response is selected to give a linear relation (over a range of doses) with log dose. This is the parallel-line assay and, in this case, dose–response curve similarity is referred to as parallelism. Even when parallel-line analysis is used, it may be appropriate to use asymptotic values as additional assay acceptance criteria.

With the increasing availability of software packages, four-parameter logistic (4PL) curve fitting is widely used. The four parameters are minimum asymptote (A), Hill slope (curve steepness, B), inflection point (C) and maximum asymptote (D). Each of these parameters may provide an assay acceptance criterion for determining similarity of the dose–response curves of test samples and reference standards and/or as an absolute value. 5PL fits are used less commonly but can provide a better fit for dose–response curves that are not symmetrical around the inflection point.

A common question is the use of constrained versus non-constrained curve fitting, i.e. the use of a single value, shared by all samples, for a particular parameter such as an asymptotic value. The choice requires an understanding of the assay design and of the underlying biological meaning of the assay readout.

Consultation with a statistician who is experienced specifically in analysis of bioassay data is strongly recommended during assay development and validation to assist in establishing methods of data analysis appropriate to a particular assay and setting appropriate AAC.

The table “Most Commonly Used AAC and SAC” lists the most commonly used AAC and SAC, including comments on their utility and some typical values. AAC are applied to reference standard and control samples, whereas SAC are applied to each individual test sample.

Common Practices That Cause Problems: Certain common practices in setting acceptance criteria are the cause of frequent problems. One such practice is acceptance being made too dependent on an individual data point. A protocol illustrating this point involves standard, assay control sample and test samples tested in duplicate dose-response curves on a plate. Upper asymptotes are determined as the mean of duplicate responses at maximum dose with a maximum acceptable percentage difference between upper asymptotes on one plate. The reportable value is the mean of the potencies determined on two replicate plates, with a maximum acceptable percentage difference between the two potencies. One aberrant well at the maximum dose of the assay control sample on one plate results in the upper asymptote of the assay control sample being aberrant compared with the reference standard so the plate is rejected. No reportable value is obtained because of one aberrant well out of 120.

Another common cause of problems is failure to ensure that each of the acceptance criteria is valuable in judging the validity of the assay as established during the assay characterization and continued monitoring. Inclusion of unnecessary criteria may result in the rejection of assays that are in fact fit for purpose. For example, in some systems, ED_{50} may vary widely between assays that are fit for purpose, so setting tight limits on an absolute value for ED_{50} could result in pointless rejection of an assay.

The consultation process: Information provided by delegates at the 2013 Biopharmaceutical Emerging Best Practices Association (BEBPA) bioassay conference (17) was used to compile the draft paper for consultation published January 2014 in the online and printed copy of free access journal BioProcess International (1), and posted on the BEBPA website (www.bebpa.org). Presentations, updated to take account of comments received, were made at the conferences CASSS Bioassays, March 2014, (18), IBC's 24th International Biological Assay conference, May 2014, (19), BEBPA HCP Workshop May 2014 (20), BEBPA Biological Assays, September 2014 (21) and to Health Canada, July 2014 (22).

Points raised in the discussions following presentations, and comments received by the authors, were given due consideration in compiling this revised version of the paper. The paper is not intended to be a rigid set of rules, but rather to identify the major issues to be addressed in setting appropriate AAC and SAC. Unusual and novel assay systems and designs may require additional factors to be considered. The importance of justifying the selection acceptance criteria and the limits set on their values with data from assay development, validation and on-going monitoring and trending must be emphasized as must be the importance of providing clear definition of terms used.

The authors would like to thank everyone who has contributed to the discussions and acknowledge especially Shea Watrin, Amgen Inc., and Zeban Kolen, Synthon Biopharmaceuticals, for their comments.

REFERENCES

1. Robinson CJ, Sadick M, Deming SN, Estdale S, Bergelson S, Little L. Assay Acceptance Criteria for Multiwell-Plate–Based Biological Potency Assays - Draft for Consultation BioProcess International 12(1), 2014: 30-41

2. ICH Q2(R1): Validation of Analytical Test Procedures: Text and Method-

ology. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use: Geneva, Switzerland, October 1994/November 1996; www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Quality/Q2_R1/Step4/Q2_R1__Guideline.pdf.

3.Chapter <621> Chromatography. USP29–NF24. US Pharmacopeial Convention: Rockville, MD.

4.Shabir GA. Validation of High-Performance Liquid Chromatography Methods for Pharmaceutical Analysis: Understanding the Differences and Similarities Between Validation Requirements of the US Food and Drug Administration, the US Pharmacopeia, and the International Conference on Harmonization. J. Chromatography A 987, 2003: 57–66.

5.United Nations Office on Drugs and Crime. Guidance for the Validation of Analytical Methodology and Calibration of Equipment Used for Testing of Illicit Drugs in Seized Materials and Biological Specimens. United Nations: New York, NY 2009; www.unodc.org/documents/scientific/validation_E.pdf.

6.Hahn GJ, Meeker, WQ. Statistical Intervals: A Guide for Practitioners. Wiley: New York, NY, 1981.

7.Orchard T. Specification Setting: Setting Acceptance Criteria from Statistics of the Data. BioPharm Int. 19(11) 2006: 22–29; www.biopharminternational.com/biopharm/Article/Specification-Setting-Setting-Acceptance-Criteria-/ArticleStandard/Article/detail/390955.

8.Lansky D. Strip-Plot Designs, Mixed Models, and Comparisons Between Linear and Nonlinear Models for Microtitre Plate Bioassays in the Design and Analysis of Potency Assays. Dev. Biol. 107, 2002: 11–23.

9.CBER. Guidance for Industry: Potency Tests for Cellular and Gene Therapy Products. US Food and Drug Administration: Rockville, MD, 2011; http://www.fda.gov/downloads/BiologicsBloodVaccines/GuidanceComplianceRegulatoryInformation/Guidances/CellularandGeneTherapy/UCM243392.pdf

10.Deming S. Trending, Statistical Process Control and Continual Validation. NEPDA Newsletter 3(1) 2008: 1–3; www.pda.org/Chapters/North-America-cont/New-England/Chapter-News/January-2008-Newsletter.aspx.

11.Lightbown JW. Biological Standardization and the Analyst: A Review. J. Soc. Anal. Chem. 86, 1961: 216–230.

12.Chapter 5.3: Statistical Analysis of Results of Biological Assays and Tests. European Pharmacopoeia 8. European Directorate for the Quality of Medicines and Health Care: Strasbourg, France.

13.Chapter <1034> Analysis of Biological Assays. USP36–NF31. US Pharmacopeial Convention: Rockville, MD.

14.Hauck WW, et al. Assessing Parallelism Prior to Determining Relative Potency. PDA J. Pharmaceut. Sci. Technol. 59, 2005: 127–137.

15.Callahan JD, Sajjadi NC. Testing the Null Hypothesis for a Specified Difference: The Right Way to Test for Parallelism. BioProcessing J. 2(2), 2003: 71-77

16.Plikaytis BD, et al. Determination of Parallelism and Nonparallelism in Bioassay Dilution Curves. J. Clin. Microbiol. 32, 1994: 2441–2447.

17.BEBPA Bioassay Conference, 25–27 September 2013, Basel, Switzerland. BioPharmaceutical Emerging Best Practices Association: Citrus Heights, CA; www.bebpa.org.

18.CASSS Bioassays 2014, 24 – 25 March 2014, Silver Spring, MD, USA. The California Separation Science Society: Emeryville, CA; www.casss.org

19.IBC's 24th International Biological Assay conference , 05-07 May 2014, Berkeley, CA, USA. IBC Life Sciences: Westborough, MA; www.ibclifesciences.com

20.BEBPA HCP Workshop, 15-16 May 2014, Dubrovnik, Croatia. BioPharmaceutical Emerging Best Practices Association: Citrus Heights, CA; www.bebpa.org

21.BEBPA Biological Assays, 24-26 September 2014, Barcelona, Spain. BioPharmaceutical Emerging Best Practices Association: Citrus Heights, CA; www.bebpa.org

22.Webinar presentation 11 July 2014 to Health Canada: Ottawa, Ontario; www.hc-sc.gc.ca

Corresponding author C. Jane Robinson, formerly Principal Scientist at the National Institute for Biological Standards and Control in South Mimms, UK, is currently Scientific Liaison, Biopharmaceutical Emerging Best Practices Association, Citrus Heights, CA (jane.robinson@bebpa.org). Michael Sadick is Senior Manager in Large Molecule Analytical Chemistry at Catalent Pharma Solutions in Kansas City, MO (mike.sadick@catalent.com). Stanley N. Deming is President of Statistical Designs in Houston, TX (standeming@statisticaldesigns.com). Sian Estdale is Principal Scientist in Large-Molecule Biopharm CMC at Covance Laboratories Limited in Harrogate, UK (Sian.Estdale@covance.com). Svetlana Bergelson is Director of Analytical Development at Biogen Idec in Cambridge, MA (svetlana.bergelson@biogenidec.com), and Laureen Little is Principal Consultant at Quality Services in Seattle, WA (biotech@ix.netcom.com).