White Paper

BioTechniques

For reprint orders, please contact: reprints@futuremedicine.com

Best practices in bioassay development to support registration of biopharmaceuticals

John R White*[1], Marla Abodeely[2], Sammina Ahmed[3], Gaël Debauve[4], Evan Johnson[5], Debra M Meyer[6], Ned M Mozier[6], Matthias Naumer[7], Alessandra Pepe[8], Isam Qahwash[9], Edward Rocnik[5], Jeffrey G Smith[10], Elaine SE Stokes[11], Jeffrey J Talbot[12] & Pin Yee Wong[13]

ABSTRACT

Biological activity is a critical quality attribute for biopharmaceuticals, which is accurately measured using an appropriate relative potency bioassay. Developing a bioassay is a complex, rigorous undertaking that needs to address several challenges including modelling all of the mechanisms of action associated with the biotherapeutic. Bioassay development is also an exciting and fast evolving field, not only from a scientific, medical and technological point of view, but also in terms of statistical approaches and regulatory expectations. This has led to an industry-wide discussion on the most appropriate ways to develop, validate and control the bioassays throughout the drug lifecycle.

KEYWORDS

best practice • bioassay • biologicals • control • development • outlier • plate design • potency • replication strategy • statistics • similarity

*[1]Biopharmaceutical Development GlaxoSmithKline R&D, 1250 S Collegeville Rd, Collegeville, PA 19426, USA; [2]Analytical Development Department, Shire, 200 Shire Way, Lexington, MA 02421, USA; [3]Analytical Services, Lonza Biologics plc, 228 Bath Rd, Slough, SL1 4DX, UK; [4]Bioassay Development, Analytical Development Sciences for Biologicals, UCB Pharma, Chemin du Foriest, B-1420 Braine-l'Alleud, Belgium; [5]Analytical Development – Biologics, Takeda Pharmaceuticals International Co., 300 Massachusetts Ave, Cambridge, MA 02139, USA; [6]Biotherapeutics Pharmaceutical Sciences, Pfizer Inc, 700 Chesterfield Parkway West, Chesterfield, MO 63017, USA; [7]NBE Analytical R&D, AbbVie Deutschland GmbH & Co. KG, Knollstrasse, Ludwigshafen 67061, Germany; [8]Analytical Development Biotech department, Merck Serono SpA, 11 Via Luigi Einaudi, Guidonia Montecelio (RM) 00012, Italy; [9]Biologics Development, Molecular & Analytical Development, Bristol-Myers Squibb, 311 Pennington-Rocky Hill Road, Pennington, NJ 08534, USA; [10]Analytical R&D, Pfizer Inc., 1 Burtt Road, Andover, MA 01810, USA; [11]Development Group, BioPhorum Operations Group, The Gridiron building, 1 Pancras Square, London, N1C 4AG, UK; [12]Analytical Sciences, Regeneron Pharmaceuticals Inc., 777 Old Saw Mill River Road, Tarrytown, NY 10591, USA; [13]Analytical Development and Quality Control, Genentech, Inc., 1 DNA Way, South San Francisco, CA 94080, USA; *Author for correspondence: john.r.white@gsk.com

BioTechniques 67: 00-00 (September 2019) 10.2144/btn-2019-0031

The BioPhorum Development Group (https://www.biophorum.com) is an industry-wide consortium, enabling the sharing of best practices for the development of biopharmaceuticals. The intent of this paper is to present a collective opinion of the BioPhorum Development Group and to discuss the inherent challenges and industry-proven approaches leading to the successful development, registration and implementation of a bioassay to support commercialization.

For biotherapeutics, a selective, physiologically relevant bioassay is essential to report on the product's potency and stability, by providing an assessment of the molecule's biological activity. Bioassays, in principle, can range from recognition of a particular antigen in a simple binding method, through systems as complex as blocking an inhibitory ligand that restores a co-stimulatory effect. Selection of an appropriate method has its challenges rooted not only in the need to mimic the mechanism of action (MOA), but also because bioassays can be costly to develop, perform, transfer and maintain. Despite efforts to implement measures to ensure method control, cell-based bioassays can be inherently variable and often lack the precision and robustness of biophysical methods simply because they use living organisms, tissues or cells.

It is generally agreed by regulators that a phased approach to the development of bioassays be implemented. It is often advantageous to start with a binding method for the early phases of product development, such as enzyme-linked immunosorbent assays (ELISAs) or surface-plasmon resonance (SPR) techniques. This approach allows time for the development of more complex bioassays (typically cell-based) by later phases. However, the sooner a relevant MOA-based bioassay is developed the better, not only to gain greater process and product understanding but also to gain a better understanding of method performance prior to pivotal clinical trials. Cell-based bioassays should be qualified and monitored over the span of clinical development to have an accurate understanding of the critical steps and components of the assay. In recent years, there has been more in-depth discussion between industry and regulators on whether a cell-based bioassay is always required before registration. The growing consensus seems to be that the decision should be driven by a product's therapeutic MOA. In cases where the MOA is simply binding to a target, a surrogate method, such as a protein binding or competitive binding assay, may be sufficient for the determination of potency. In some cases, regulatory agencies have been amenable to implement surrogate, non-cell-based bioassays if an existing cell-based bioassay is demonstrated to be too variable or not amenable for a quality control environment [1]. In these cases, it may be advisable to move the cell-based bioassay to the characterization panel.

MECHANISM OF ACTION & ITS INFLUENCE ON POTENCY & CONTROL STRATEGIES

Design strategies for bioassays are driven by the drug's intended physiological MOA. Unlike other analytical techniques, bioassays are almost always unique for each therapeutic. A well-designed bioassay will accurately capture the biological activity of a drug candidate. Common MOAs of therapeutics include direct binding to soluble targets (e.g., ligands, cytokines and enzymes), or to cell-surface receptors in either an inhibitory or agonistic manner. Recombinant antibodies represent the majority of current clinical biotherapeutics [2–5]. As a result, the majority of the following guidance will reflect recommended approaches for antibody therapeutics, with additional modalities specifically highlighted where appropriate. Figure 1 depicts four common MOAs typical for monoclonal antibodies including target sequestering, antibody-dependent cellular phagocytosis, antagonism of cell-surface receptors and Fc-mediated immune cytotoxicity (e.g., antibody-dependent cellular cytotoxicity or complement-dependent cytotoxicity).

Each MOA will require a different approach when considering the bioassay design. In the case of monoclonal antibodies and related recombinant proteins, secondary, tertiary, or synergistic MOAs may be discovered during development. This biological complexity further contributes to the challenge of developing MOA-reflective assays to wholly capture the candidate molecule's putative therapeutic biological activity [6–10]. In some cases where multiple MOAs exist in a single molecule, a combination assay that measures all MOAs in a single assay may be suitable for release and stability testing with secondary characterization assays developed to measure the individual activities of each MOA.

Across the spectrum of analytical assays, quantitation of a drug's MOA and implied higher order structure, potency and efficacy is unique to the bioassay, and as such is an expected part of any analytical package for a biotherapeutic [11].

As the biotherapeutic progresses through the clinical lifecycle, the analytical package is refined as part of the overall process and analytical control strategy (PACS) for the molecule, a step in the quality-by-design approach to process development. The PACS provides the product-specific portion of the overall manufacturing control strategy, providing a detailed justification and description of the control elements to be applied to the control of critical quality attributes (CQAs). In the early clinical stages, assessment of criticality is understood to be relatively uncertain due to limited understanding of the molecule's characteristics, susceptibility to process variations and changes, and the subsequent clinical impact due to limited data. However, in preparation for BLA registration, a cross-functional effort across the organization is typically initiated for a more comprehensive characterization of the molecule's structure along with greater clinical data, with the goal to gain more insight into CQAs and develop more robust PACSs.

While many therapeutic antibody candidates fall under the guidance of the Centre for Drug Evaluation and Research (CDER), the sister organization Centre for Biologic Evaluation and Research (CBER) provides clear guidance regarding bioassay life-cycle management with the expectation that as candidate molecules progress through the clinic, a cell-based functional bioassay that more closely models the molecule's therapeutic activity will be evaluated [12,13]. To evaluate the fitness of the cell-based bioassay to replace or supplement a binding assay, comparability experiments must be conducted to successfully bridge from binding to a cell-based potency bioassay. At a minimum, a bridging study should evaluate precision, accuracy, linearity and stability indicating properties using degraded and other samples [13]. The best bioassays to support commercial release (binding or cell-based) recognize degraded product as well as high-potency samples, and have adequate precision and accuracy to support a release specification range. It is common practice to inform, discuss and seek pre-approval of novel or non-routine potency strategies with the relevant regulatory agencies prior to filings to ensure seamless and successful BLAs.

INNOVATIVE ASSAY FORMATS

In recent years, innovative assay formats (e.g., reporter gene and second messenger assays) have emerged as drug discovery tools [12]. Some of these assays have subsequently found their way into cell-based assays for drug development [14]. Characteristics of these assays include their ability to measure the regulation of cellular signalling events upon drug treatment instead of analyzing the respective downstream output processes such as proliferation, toxicity or cytokine release. New non-cell-based technologies have also emerged (e.g., homogeneous proximity-based readouts); however, the current practice by companies reported in the survey indicates that these formats have not been widely adopted.

Commonly used cell-based reporter assays are based on stably-transfected vectors encoding luciferase or other reporter genes. Gene expression is then controlled by an inducible promoter where the production of reporter protein is directly related to the binding of transcription factors involved in the drug's MOA to their corresponding response elements. The outputs of these assays are easily quantitated via readily available plate readers.

The simplicity of these assays, combined with shorter assay times (hours vs days), improved reliability and high sensitivity, as well as high dynamic ranges, are some of the key advantages of these assay formats. According to the BioPhorum development group bioassay point share (BPDG-BPS) members surveyed, reporter and second messenger assays are formats that have been accepted by regulatory agencies as long as they are reflective of the drug candidate's MOA.

TACTICAL DESIGN

Once the therapeutic's MOA(s) is defined and the *in vitro* system selected, method development can begin. As previously discussed, cell-based bioassays are inherently complex and variable. To control for this variability, a methodical, stepwise approach is used to design and select the assay type that best reflects the MOA(s) of the therapeutic. In this section, we discuss each of the components of a robust bioassay, beginning with selection of cell type and moving through assay design, plate selection, plate layout, plate bias, data analysis, and mitigation strategies. Bioassays evolve over time and necessitate the development of a method life cycle management strategy to ensure uninterrupted commercial supply for patients. Figure 2 outlines the recommended bioassay method development process.

Selection of cell type in bioassays

The choice of cell type, cell handling and bioassay design will influence bioassay performance (e.g., robustness, accuracy, reproducibility, linearity, and so on). In addition, the transfer of a bioassay to a secondary testing site with less expertise than the developing site should also be taken into consideration when deciding on the cell type to use and bioassay complexity.

The choice of cell type is usually between a primary cell typically isolated from human blood or tissue, and an immortalized established cell line derived from human or nonhuman origin. Cell lines can also be modified to overexpress the receptor or ligand that is targeted by the therapeutic. Cell line choice should ideally be relevant to the MOAs of the therapeutic. For example, when constructing a reporter cell line in which a target receptor is transfected, the receptor should be of human origin to reflect the intermolecular interactions between the drug and its target.

Primary cells may be required when the MOA of the therapeutic agent is complex or multiple receptors are involved in the MOA and a transfected cell line cannot fully represent all interactions. Unless primary cells are absolutely required, they should be avoided due to considerable donor-to-donor variability [16]. The primary cell variability may be partially circumvented by freezing banks of cells that can be used in the bioassay following thawing [17].

When cells are cultured, it is important to monitor for changes in morphology, magnitude of response, half maximal effective concentration (ED_{50}), doubling time and cell viability, among other parameters. Cell culture and assay performance should be monitored over time and a limit established for the allowable number of cell passages, before routinely starting a new culture.

Plate layout & the identification of bias

Microtiter plates are an essential component of most biochemical or cell-based bioassays. The setup of both binding assays and cell-based bioassays are typically amenable to a 96-well plate format. This format allows analysts to set up assays using multichannel pipettes or, alternatively, automated liquid handling systems to prepare or add dilutions, add reagents, and so on. It also allows for the use of plate washers and plate readers that facilitate the performance of the assays. Plates with higher well count (384 and 1536) require automation to assist with set up and thus require an investment in automation for routine use. When automation is anticipated, it should be undertaken early in bioassay development. There may be differences between the manual and automated methods (e.g., dilution order, order of reagent addition). However, both the manual and the automated method should use the same reagents in order to allow flexibility as to which bioassay is implemented in a quality environment.

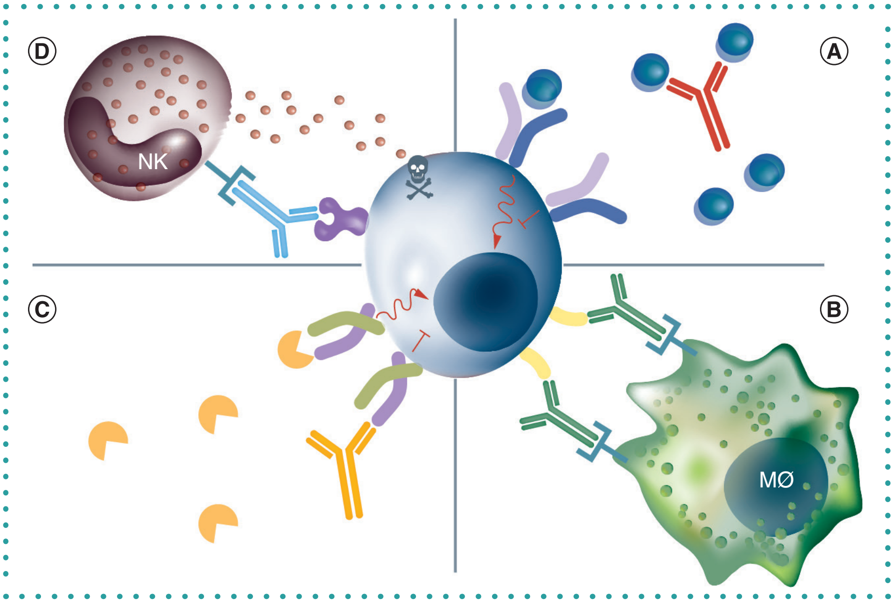


Figure 1. Four common biopharmaceutical mechanisms of action exhibited by monoclonal antibodies. (A) Inhibition of soluble ligand binding to a cell surface receptor. (B) Antibody-dependent cell phagocytosis typically mediated by macrophages binding to the Fc region of the antibody and reverse signaling through FcγRIIA. (C) Inhibition of receptor–ligand interaction via a blocking antibody binding to the receptor and inhibiting signaling. (D) Antibody-dependent cell cytotoxicity mediated by NK cells interacting with the FcγRIIIA receptor on NK cells.
NK: Natural killer.

Microtiter plate assays have multiple potential sources of variability that can affect bioassay performance and thus impact the accuracy of results. These include variations in cell plating and cell growth rate, inconsistent cell response, biased results due to the location of the sample in assay plates, order of addition of standard, control sample(s), test samples and critical reagent, analyst-to-analyst, plate-to-plate and run-to-run variability. Among these variables, the microtiter plate is a dominant contributing source of location-based error. For example, the most common plate-related phenomenon is the so-called 'edge effect', where the response from peripheral wells differs from the response observed from the inner wells of a microtiter plate (Figure 3). To minimize or protect against potential plate location effects, different approaches have been reported in the literature [18–23] and discussed by BPDG-BPS members. The most common practices include: 1) the use of techniques that help minimize the edge effect (e.g., plate hotels, heat-transfer plate and others, see section on plate effect); 2) the inclusion of replicates (wells or plates); 3) careful consideration on the placement of standard, control, and test samples in the plate (plate layout); 4) the use of randomization or pseudo-randomization (row or column); 5) exclusion of outer rows and/or columns; 6) use of automation.

Plate effect

An integral part of bioassay development is the assessment of plate positional bias to help derive the final assay plate layout. To evaluate positional and edge effects, the two approaches described below may be used. In addition, it is important to perform these studies when possible using the same equipment and loading order (e.g., using a 12-channel from top to bottom or 8-channel from left to right) used by QC during routine testing to capture positional effects that may result from loading. For one approach, a single drug dilution targeting the ED_{50} is loaded into all wells of multiple microtiter plates. Responses in the exterior wells are compared to those of the inside wells and graphed to determine the extent of the bias (Figure 3). Alternatively, a full dose–

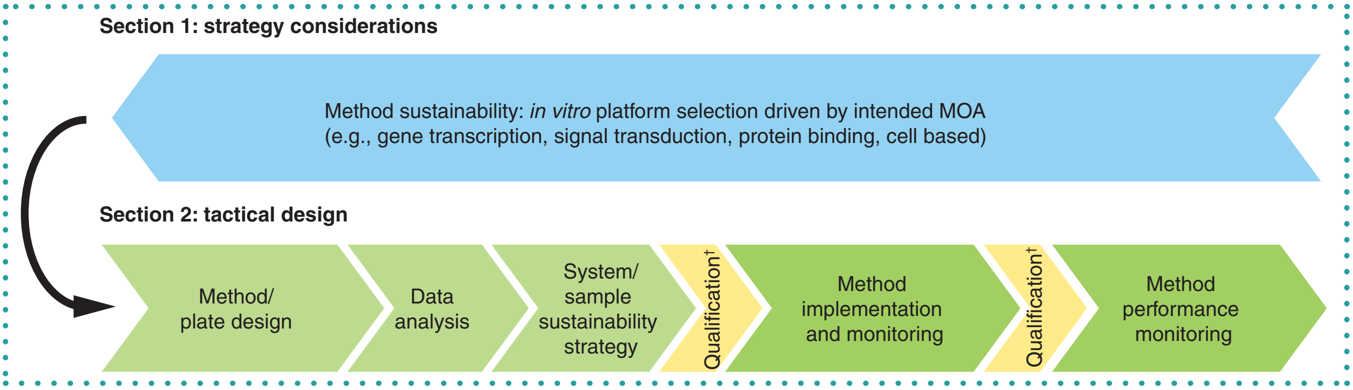

Section 1: strategy considerations
Method sustainability: in vitro platform selection driven by intended MOA
(e.g., gene transcription, signal transduction, protein binding, cell based)
Section 2: tactical design
Method/ plate design
Data analysis
System/ sample sustainability strategy
Qualification†
Method implementation and monitoring
Qualification†
Method performance monitoring

Figure 2. Bioassay development considerations.
†Qualification describes activities that demonstrate suitability of a method for use in early stages of product development while full method validation using a preapproved protocol with acceptance criteria is performed prior to using a method for later stages of development and commercialization [13,15].
MOA: Mechanism of action.

response range can be prepared, using reference standard, and loaded into columns 1–12 or rows A–H of multiple plates, depending on the bioassay design. If full-dose response curves are used, then these can be processed as per the method to establish ED_{50} values, averaged for all plates used and then compared across multiple plates and experiments for bias. If no 'edge effect' is observed, all wells of the microtiter plate may be used. If an edge effect is present, this bias may be addressed by utilizing only the inner-80 or inner-60 wells of the plate (Figure 3).

Although this is the most common approach applied, using only a portion of the assay plate significantly decreases sample throughput. To avoid this, other approaches (i.e., preincubation of cell-seeded plates at room temperature [18], use of circumferential built-in-moat microtiter plates [22], and others) have been reported to minimize plate bias or edge effect while enabling the use of the full microplate.

Randomization, replication & plate design

Randomization, pseudo-randomization, replication and plate design are also standard approaches that can be used to protect against or to minimize plate location and plate-to-plate bias, even in cases where no obvious plate positional bias exist.

Pseudo-randomization can be applied both within a single plate and between plates. Within the plate, dose replicates can be loaded in alternating rows or columns and plate location effects can be mitigated by ensuring that test and reference samples have equal exposure to the edges. At the same time, each plate can be designed with a different randomized or pseudo-randomized sample loading pattern to further protect against location effects and reduce overall variability [20].

The choice of the number of replicates may vary based on the application and variability of the bioassay (release, stability, comparability or characterization), and the expected throughput and precision needed. Usually, within a plate, at least two replicates per dose level are employed and, likewise, sometimes two or more plates are used to generate the sample reportable value. The exact number of plates used should be established through statistical analysis to achieve the desired method performance to support the specification. The final bioassay format (number of wells per dose level within a plate and number of plates to generate a sample reportable value) is often derived and supported by the analysis of variance decomposition (ANOVA of variance components) of method validation data and/or control trending data, and by consulting with a biostatistician [20,23,24].

However, even with the most well-designed plate layout and replication schemes, outliers can still occur. Setting a precision acceptance criterion on the replicates or the use of statistical tests could be applied to detect outliers and support removal of the outlier data point or plate when appropriate.

Analysis of data

As a summary of the USP <1034> [21] approach, calculation of the relative potency is based on the prerequisite that test and standard drug preparations behave similarly in the bioassay system. Consequently, the test and standard dose–response curves should share common functional parameters (i.e., Hill slope, upper and lower asymptotes) and ideally would only differ by a horizontal displacement, representative of a gain or loss of activity. If sufficient similarity (or parallelism as these two terms can be used interchangeably) cannot be demonstrated, then the relative potency calculation obtained from the two curves cannot be confidently interpreted as an indication of biological relative potency and should not be reported as such (Figure 4). Depending on the bioassay, whole (full-curve model) or partial response curves (e.g., linear part only when asymptote cannot be reached) are used in the assessment. However, even if a linear model allows sufficient assessment for lot release, it is recommended to assess parallelism using the full-curve model to support stability, comparability, or to qualify reference material or critical reagents [3].

Implementation of the similarity assessment: a 3-step process

Similarity assessment can be implemented in three steps:

Step 1: Select a mathematical model that fits the data (goodness of fit [GOF])
Step 2: Select relevant measures of similarity
Step 3: Define acceptance criteria based on the similarity measures

Step 1: Selection of a curve fit model (GOF)

Bioassays usually present a nonlinear relationship between the response and the analyte such as log-concentration with a sigmoidal shape. The model commonly used for curve-fitting analysis of symmetric sigmoidal curves is the 4-Parameter Logistic (4-PL) regression model described by the following equation:

$$Y = \frac{(A-D)}{\left(1+\left(\frac{x}{C}\right)^{B}\right)} + D$$

Where Y is the response, A is the response at zero analyte concentration, D is the response at infinite analyte concentration, C is the inflection point (also known as ED_{50} that corresponds to the point where "Y = (A + D)/2"), B is the Hill slope that defines the steepness of the curve and x is the analyte concentration.

A 4-PL function requires a sufficient number of concentrations or dilutions to fit the model (the BPDG-BPS recommends a minimum of eight concentration points to define a 4-PL curve). At least one, and preferably two, concentrations are commonly used to define each asymptote (parameters A and D) and at least three, preferably four, concentration points in the linear part of the curve. The 'linear' region of a 4-PL function is often defined as concentrations near the center of the response region where test sample input produces a direct linear assay signal output. Mathematically, the ideal Hill slope (B parameter) within the linear range should be around 1, meaning that the concentration points are spatially well distributed when dilutions are even and symmetric. However, this might be difficult to achieve due to intrinsic characteristics of the drug being evaluated (i.e., affinity of the drug to its target, mode of action). Therefore, a development objective would be to optimize experimental conditions and dilution scheme so that concentration points create a well-defined linear range. In addition, each dose–response curve should be made of individual values at each concentration point and not from averaged replicate values.

The validity of a given mathematical model as a descriptor of the concentration–response relationship should be assessed. To achieve this, analysis of the residuals is recommended. Residuals are the differences between the observed response and the response predicted by the fitted model at a given concentration. If the fitted model is appropriate, residuals are randomly and independently distributed around zero. Residuals normality can be assessed through, for example, a residual plot (residuals vs concentration) or normality test (e.g., Shapiro-Wilk) [25].

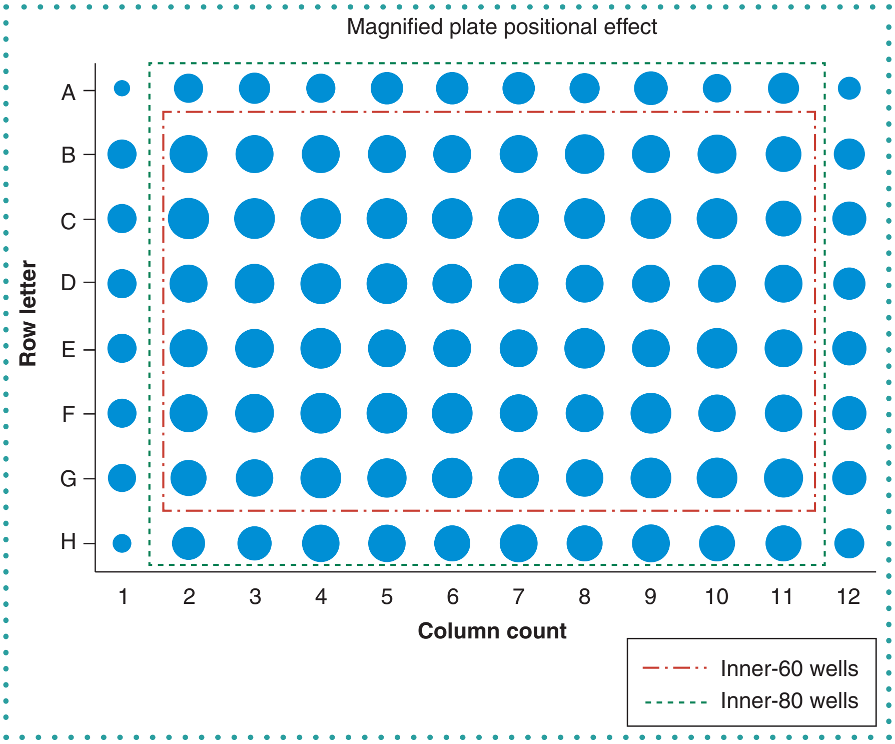
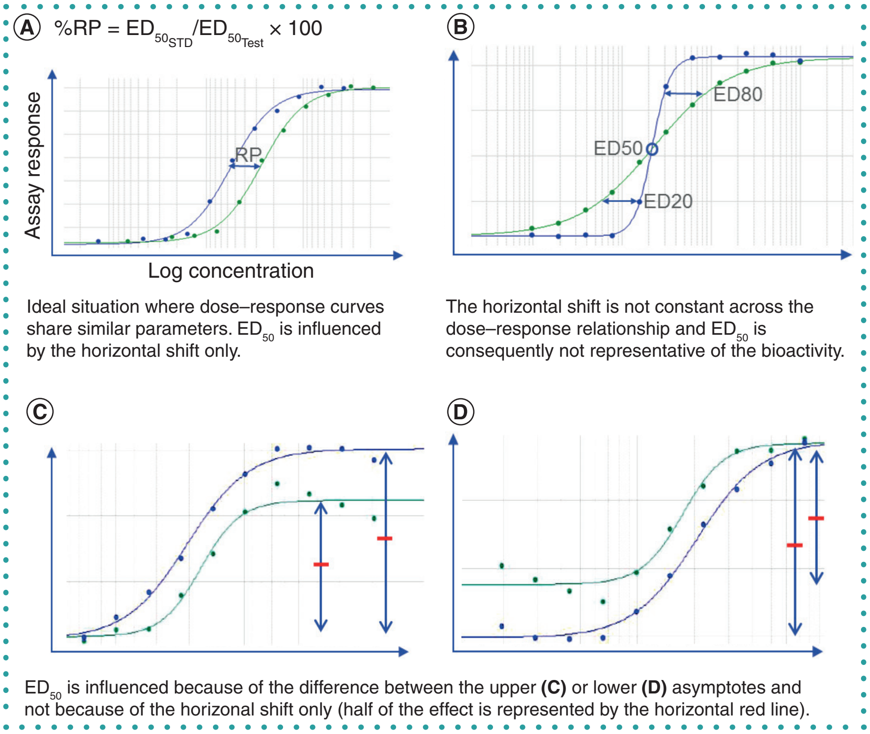


Figure 3. Potency response map of individual wells in a cell-based bioassay. Each well was set to target 50% (ED_{50}) of the response in the bioassay. Samples were applied to the plate in the same order using the same pipetting technique as would be used in the qualified assay. The size of the well/circle represents the results of individual wells and was magnified to differentiate the smallest and largest responses. For this assay plate the boundaries of the analysis were set so the smallest circle represents 31.9% of the ED_{50} while the largest circle represents 60% of the ED_{50} response. This plate is a typical result from a multi-plate experiment.

As described earlier, the 4-PL model is typically employed first as it captures the salient features of most biological dose–response curves. When the biological MOA produces a curve that is not symmetric around the midpoint, evaluation of an alternative model may be necessary, and typically a 5-PL model will produce a better fit for asymmetric curves. As such, evaluating model GOF is a key piece of early assay development. Different approaches to evaluate the GOF are proposed in the literature [19,26]. Sum of squared errors (SSE) is widely used as a representation of GOF. This approach consists of evaluating the difference between the observed response at a considered concentration and the fitted model. R^2 is another way of evaluating GOF based on the ratio between sum of squares regression and total sum of squares. Plotting residuals against dose from the model would indicate whether there is a random deviation (suggesting that the model is appropriate) or a systematic deviation (suggesting that the model is not appropriate). Weighting can be considered, particularly if the variance of the response increases with magnitude. This approach will serve to increase the precision of regression parameter estimates by mitigating the effect of unequal variance [27].

Step 2: Measurement of similarity

For nonlinear models, USP <1032> proposes two methods for evaluating similarity: 1) curve parameters (i.e., slope, upper and lower asymptotes, and asymmetry factor for a 5-PL fitting model); or 2) based on a single composite measure (i.e., residual-SSE [RSSE]) [28]. The majority of BPDG-BPS member organizations evaluate similarity through the ratio of curve parameters (10 of 16 respondents), while 3 of 16 respondents use a single composite measure; the remaining respondents use a combination of approaches.



Ideal situation where dose–response curves share similar parameters. ED_{50} is influenced by the horizontal shift only.

The horizontal shift is not constant across the dose–response relationship and ED_{50} is consequently not representative of the bioactivity.

ED_{50} is influenced because of the difference between the upper **(C)** or lower **(D)** asymptotes and not because of the horizonal shift only (half of the effect is represented by the horizontal red line).

Figure 4. Relative potency should only be considered if similarity between the considered dose–response curves is demonstrated. (A) Ideal situation where dose–response curves are well defined and share similar parameters (upper/lower asymptotes and slope). Relative potency is then influenced by the horizontal shift between the two curves and calculated as the ratio between the two inflection points (ED_{50}). (B–D) Unsuitable situations where parameters are dissimilar and parallelism is not met, meaning that the distance between Hill slopes (B), upper asymptotes (C), and lower asymptotes (D) is outside a predefined acceptance range. Consequently, relative potency is not solely a reflection of the horizontal displacement between the two curves. Therefore, calculating relative potency as the ratio between the two inflection points is not appropriate.

Evaluation of similarity based on curve parameters

The most popular way of assessing similarity is through the evaluation of curve parameter ratios (slope, upper and lower asymptotes). The way the curve ratio approach is applied differs slightly from company to company. The standard approach in the evaluation of similarity based on curve parameters is to consider the following three ratios:

1.

$$\textit{Hill slope ratio} = \frac{B_S}{B_T}$$

2.

$$\textit{Lower asymptote ratio} = \frac{A_S}{A_T}$$

3.

$$\textit{Upper asymptote ratio} = \frac{D_S}{D_T}$$

Where T and S are the test sample and standard dose response curves, respectively, while A, B and D are the curve parameters corresponding to lower asymptote, Hill slope and upper asymptote, respectively.

In some cases, a higher variability with the lower asymptote ratio can be observed due to the large impact of small variations in the ratios of standard and test sample. To overcome this potential drawback, Yang *et al.* suggest not considering the lower asymptote alone, but together with the upper asymptote as the ratio of upper to lower asymptote:

$$\textit{Upper to lower asymptote ratio} = \frac{(D-A)_S}{(D-A)_T}$$

[25,29]. They also recommend considering the ratio of standard slope at inflection point:

$$\textit{Standard slope ratio} = \frac{-((D-A)*B/4)_S}{-((D-A)*B/4)_T}$$

rather than the Hill slope ratio. However, even if this last proposal is more accurate from a mathematical point of view, the impact of both approaches on the assessment of slope similarity is probably limited in most cases.

The main advantage of evaluating similarity based on curve parameters is the ease of implementation, as it does not require complex statistical tools or formulas and it is straightforward to interpret. However, it requires defining limits on each considered ratio and it does not consider the interdependence between the curve parameters [25,28–30].

Evaluation of similarity based on a single composite measure

Non-parallelism RSSE ($RSSE_{nonPar}$) is a composite measure that considers all parameters (slope, upper and lower asymptotes) together in a single measure [27,31]. It is a direct measure of the level of (non) parallelism between two dose–response curves and it ranges from 0 (perfect parallelism) to ∞ (nonparallel). It measures the difference between the residual variability when the parameters of the two curves (slope and asymptotes) are constrained to be equal (constrained model) and the residual variability when the parameters (slope and asymptotes) are different for each curve (unconstrained model; Figure 5): $(RSSE_{nonPar}) = (RSSE_{Constrained}) - (RSSE_{Unconstrained})$

Therefore, '$RSSE_{nonPar}$' corresponds to the energy required to force two dose–response curves to be similar in a constrained model. The more energy it takes, the more dissimilar the curves were in the unconstrained model.

The $RSSE_{nonPar}$ criterion is not without drawbacks. As a composite measure, it can be hard to interpret. Statistically, it is possible for $RSSE_{nonPar}$ to identify nonparallelism when there are no practically important differences in any one parameter.

In conclusion, both approaches ($RSSE_{nonPar}$ and curve parameter ratios) have different advantages, and both are viable options for parallelism assessment.

Step 3: Define acceptance criteria on the similarity measures

Once similarity measures are defined, an acceptance criterion needs to be established to discriminate between parallel and nonparallel response function curves. Equivalence testing implies conformance to an interval acceptance criterion and is one statistical test that does not penalize results that are too precise [28,32,33]. Fleetwood *et al.* highlighted the importance of the size of the historical dataset used to define reliable equivalence limits [34]. While a dataset of at least 100 assays is proposed by Fleetwood *et al.*, a more general recommendation would be to have a phase-appropriate dataset size. In addition, parameters for similarity testing can be established and re-established as the assay is developed and used. The BPDG-BPS ideally recommends periodically reassessing the equivalence limits.

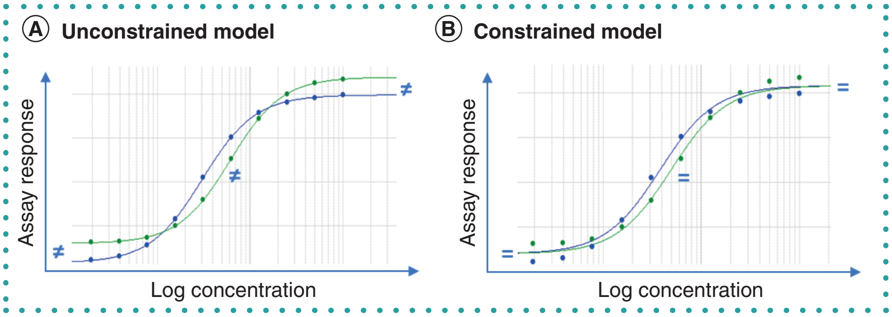


Figure 5. Model used for the calculation of non-parallelism residual sum of squared errors ($RSSE_{nonPar}$). (A) Unconstrained model (full, unrestricted: fit independent 4-PL curves to the dose–response data). (B) Constrained model (reduced, restricted: fit 4-PL curves having common A, D and B parameters) [31].

Relative potency calculation

Once curve similarity is demonstrated, the relative potency of the test sample can be calculated. The choice between the constrained or unconstrained model could be based on qualification data (qualification meaning 'early-phase validation'). The BPDG-BPS recommends analyzing repeatability and linearity data derived from the qualification set through constrained and unconstrained models to determine which of the two models best fit the data. In cases where there is little difference in the performance of the two methods, companies are justified in adopting either approach. Under standard statistical assumptions, constrained fitting produces less variable estimates. However, deviation from these assumptions could subject constrained fitting to unpredictable biases which unconstrained fits avoid. The ideal approach would exhibit no systematic bias and low variability of the relative potency estimates.

For calculation of the relative potency using only the linear range, a linear regression with a common slope and x-intercepts is fitted on the linear range selected from the 4-PL curve of each standard and test sample. The relative potency (RP; in %) is computed as:

$$RP = 100 \times antilog\left(\frac{A_T - A_S}{B}\right)$$

Where A_T is the test x-intercept, A_s is the standard x-intercept and B is the common slope.

In a full-curve model, for each pair of standard and test samples the RP (%) is computed as:

$$RP = 100 \times \left(\frac{BD50_S}{BD50_T}\right)$$

Where ED_{50S} and ED_{50T} are the ED_{50} parameters of the standard sample and of the test sample, respectively.

The reportable value is then calculated as the arithmetic mean or geometric mean (taking into consideration the geometric distribution of the relative potency value) from the corresponding sample replicates.

System & sample suitability considerations

Design & application of system & sample suitability criteria

Prior to calculating the potency of a test article, an assessment of the assay performance should be conducted. System and sample suitability criteria are a panel of assay performance criteria, established during assay development and confirmed during assay validation, to ensure the assay was in a controlled state during the testing. System and sample suitability are applied individually to each plate, and each test article on the plate. Initially described in ICH Q2(R1), system suitability is intended to control for variability by implementing criteria specific to each analytical procedure [19]. Guidance was updated in USP <1032> to capture bioassay-specific information on system and sample suitability criteria, reflecting the advances in bioassay development and testing across the industry [28]. It is expected that system and sample suitability criteria will evolve alongside the maturing assay.

Multiple factors influence the choice of system and sample suitability criteria (assay design, molecule MOA, technology platform, life-cycle stage, availability of reagents, presence of a control sample, and so on). System suitability criteria typically characterize the quality of reference standard and control sample curves (e.g., GOF), proper functioning of the bioassay system (e.g., max to min ratio) and/or measurement of variance in the bioassay run (Table 1). Conversely, sample suitability criteria compare the performance of the sample to the performance of the reference standard (e.g., curve parameter ratios). If these criteria are not met, no further calculations are performed, and the sample and/or assay is considered invalid and must be repeated [35].

During early development, small data sets make generating statistically derived values for initial system and sample suitability criteria challenging. As such, using a predictive statistical method that addresses the uncertainty associated with small data sets such as tolerance intervals is appropriate [36]. This approach can prevent unduly stringent initial criteria. Over time as data sets expand, they may support potentially more stringent criteria, at which point they can appropriately be evaluated using confidence intervals. In addition to the tightening of criteria, it may become beneficial to remove criteria that are no longer indicative of critical assay parameters, or add criteria whose need becomes apparent due to greater variations seen in the expanded data sets [19].

Signal-to-noise ratio is also usually evaluated as an indicator of assay performance. A general recommendation is to have a development objective set for the minimum signal-to-noise ratio that would ensure better control on assay accuracy and precision. ▶

Table 1. Examples of system, control, and sample suitability criteria.

Criteria	System suitability	Control suitability	Sample suitability
Statistical evaluation of fit[†]	X	X	X
%CV of ED_{50} standard values[‡]	X	X	
ED_{50} range[§]	X	X	
Max/min ratio[¶]	X		
Measure of similarity between standard and test sample (curve parameter ratios analyzed independently or in combination[#] and/or single composite measure; e.g., $RSSE_{nonPar}$)		X	X
Measure of replicate variability[††]		X	X

[†] Goodness of fit test by comparing curves, which typically relies on a database of historical assay data to establish the acceptance parameters (e.g., R^2/residual sum of squares/root mean square).

[‡] Relevant for Standard and control when run on same plate since standard and control are well-characterized samples. Below a certain %CV ED_{50} ensures proper functioning of bioassay system.

[§] ED_{50} applies to assay control and standard, and the need for the result to fall within a statistically-derived bioassay control range. This parameter is only used when the assay has demonstrated a highly consistent ED_{50} and is useful when only one control or standard curve is run on a plate.

[¶] Ensures assay system is functioning and responding properly. It has the ability to indicate contamination, incorrect dilutions or cell numbers.

[#] Ensure standard and test sample dose–response curves share similar parameters using an equivalence approach. Curve parameters can be evaluated independently (e.g., ratio of slope, ratio of upper asymptote, ratio of lower asymptote or ratio of upper to lower asymptote).

[††] Bioassay reportable values are typically made of multiple replicates. Due to space constraints on a single plate more than one plate may be required. Placing limits on acceptable variability (e.g., %CV, CI) between replicate series allows confidence in the data to be controlled and ability to generate a potency value reflecting the true state of the sample.

$RSSE_{nonPar}$: Non-parallelism residual sum of squared errors.

Management of outliers

Outlier evaluation methodology is detailed in USP <1032> as part of bioassay development and is used to varying degrees by the BPDG-BPS member organizations. Visual outlier inspection as an initial identification tool must be supplemented with an objective statistically derived approach to avoid bias. Commonly used statistical tools may include Grubbs or Dixon's Q tests, as well as replication- or model-based approaches [28,37].

A replication-based approach can be used when multiple replicates are performed at concentrations of test and standard preparations. To identify outliers, an additional variability criterion can be defined (i.e., %CV or ED_{50} range) based on the observed variability among replicates in a historical dataset.

A model-based approach uses the residuals from the fit of an appropriate model to identify outliers. Example decision criteria for outlier detection and removal could include internally studentized residuals and z-scores [38].

Within BPDG-BPS member organizations, the majority, nine of 16 respondents, have clearly defined outlier removal strategies for sample and/or dose–response curves. Although there was no consensus as to the best outlier methodology, Grubbs test was the most popular (six of 14 respondents), and regardless of approach, members overwhelmingly agreed that consistent, non-subjective, application was key for a robust outlier strategy (14 of 16 respondents).

Removal of outlier points does have an impact on influencing the confidence of a reportable value, especially given a limited number of replicate data points. A general recommendation would be to define an acceptable minimum number of replicates used to compute the reportable value. This minimum number of replicates could be supported by a variance decomposition analysis performed on validation data, computing the expected %CV of reportable values obtained from different replicate numbers.

Monitoring bioassay performance

Incorporation of controls

General assay controls (e.g., blanks, positive/negative controls) are used to ensure the method was properly executed and performing as intended. Within the BPDG-BPS members, 13 of 16 respondents incorporate an internal control in their bioassays at some point during development. The internal control is either an independent batch of material, or reference material having a known activity, which is run as a test sample and monitors performance of the assay. The internal control has specific acceptance criteria assigned to further evaluate whether the bioassay is technically valid and if the test sample results should be reported. Exactly how this process is executed varied among the BPDG-BPS members. While not universal, most members do employ an internal control despite the challenges of maintaining this critical reagent in terms of cost, documentation, plate occupancy and assay throughput.

In some cases, members who do not use an internal control (three of 16) use alternative control strategies to ensure adequate performance of their bioassays, for example, reference material ED_{50} tracking and failure mode and effects analysis. In addition to the selection of an internal control, there are differences in the timing of implementation (phase of product development) and specific system suitability criteria applied.

Those using internal controls are typically required to do so for each assay plate. If using an independent lot, the material selected to serve as an internal control sample should be suitable and able to demonstrate an acceptable level of accuracy and precision versus the reference standard. The internal control sample should be tested with a full dose–response curve consistent with the reference standard and test samples, as well as using a similar sample dilution format, unless there are key drivers for implementing an alternative approach (e.g., test samples are not analyzed with a full dose–response curve). Within the BPDG-BPS, 10 of the 17 respondents that implement an internal control do so during earlier phases of bioassay testing, regulatory toxicology through to Phase I clinical batch testing. Five of the 17 respondents implement use of an internal control sample at a later stage of development or bioassay testing (Phase II/III clinical batch testing), and two of the 17 do not implement until the commercial testing stage. The majority of the BPDG-BPS members see an advantage to implementing a suitable internal control in their bioassay testing as early as feasible and by late-stage clinical development (Phase III) or commercial. It is common to implement the internal control program prior to method validation; however, some companies use validation data to set internal control acceptance ranges and in this case implement the internal control post validation.

According to the BPDG-BPS members surveyed, selection of material for a bioassay control sample varies. For those BPDG-BPS companies where the internal control sample is derived from a batch independent of the reference material, it typically comes from a batch produced and formulated similar to the reference material and may be at a similar concentration, to avoid bias from variable dilution practices. Whether derived from the reference or an independent lot, the internal control receives identical treatment to the test sample(s) being analyzed. Both approaches are common and require monitoring (e.g., curve parameters and other attributes) to determine the stability of the material. While there is no single acceptable path for how to select a material for the internal control sample, it is important to select an appropriate material for the development stage and incorporate in bioassay testing with acceptance criteria for run suitability evaluation.

Regardless of the material used, specific, numerical assay acceptance criteria should be defined for the control sample, which when applied can help guide determination of the conformity of an assay and if a result can be considered or not. The acceptance criteria range for an internal control attribute monitored during early stages of development, when experience with the bioassay is limited, will typically be set fairly wide. This range will narrow over time with experience gained and additional method optimization performed.

Trending of key assay parameters

Once the initial stages of bioassay development are complete, monitoring key assay performance parameters with a chronologically ordered trend chart can provide valuable insight into assay functioning. Bioassay monitoring may point out drift, indicate unexpected results due to instability of critical reagents, or when used in conjunction with an assay control sample, identify influences from formulation excipients and matrix effects. Suitable assay parameters to track can include those derived from the 4-PL curves such as upper and lower asymptote, ED_{50} and reportable results, as well as the Max/Min signal ratio and R^2. Tracking these indicators can bring an additional level of understanding to the bioassay development process.

DISCUSSION

Bioassays are an essential part of developing biopharmaceutical drugs and are used to determine potency, stability and comparability between different processes. The development and implementation of a bioassay can be complicated and, therefore, the concept of a white paper focused on the common practices in bioassay development evolved from the 22 BPDG-BPS biopharmaceutical member companies. These companies represent a wide perspective of current bioassay development in the biopharmaceutical industry. Therefore, the idea of capturing a condensed and harmonized view on the common challenges and best practices in bioassay development covering both strategic and tactical aspects came to fruition.

The bioassay design is directly influenced by the drug's structure and the stage of development. In addition, it should balance the inherent variability linked to the use of biological material and the reliability requirements of a quality control environment. If the drug is determined to have multiple MOAs, it is acknowledged that each of the MOAs should be monitored with one or more methods. To support the characterization of MOAs, innovative bioassay formats (e.g., reporter and second messenger assays) have emerged as robust bioassay design options.

FUTURE PERSPECTIVE

The aim of this paper was to align regulatory expectations with best industrial practices in bioassay design. From the BPDG-BPS discussions while writing this manuscript, specific points emerged that need additional consideration. The first point was related to specification range setting. It is proposed that bioassay product specifications should be defined according to the clinical stage of development, starting with a reasonably broad range, but with an expectation of tightening the specifications after a sufficient number of representative drug batches are available. Therefore, bioassay development should focus on characterizing and minimizing variability. The second point highlighted was related to the analytical data required to justify use of a binding assay versus a cell-based bioassay as the primary potency release test. While there was no clear consensus on the minimum requirement to support the binding assay over the cell-based assay, several aspects were considered, including how the binding assay reflects on the drug's MOA(s), how the bioassay reflects on stability indicating conditions, and how the bioassay differentiates between different product degradants.

We must recognize the effort being made by regulators to stimulate discussion and idea exchange through workshops and conferences. There will certainly be evolution in the potency bioassay field over the next 5–10 years as companies bring more complex biotherapeutic modalities into clinical development such as bispecific antibodies and gene and cell therapies. New technologies and bioassay platforms are available to enable the development of MOA-reflective, robust bioassays to ▶

support potency assessment throughout clinical development and commercialization. We are at the beginning of the journey and each question stimulates a new question, which promises an exciting future and new directions to come.

AUTHOR CONTRIBUTIONS

All authors contributed to the concept, writing and review of the manuscript.

ACKNOWLEDGMENTS

The authors wish to acknowledge Adi Atlas (Ferring), Petra Bennington (MSD), Wolfgang Ebel (Eisai Inc), Mingfang Shen (ImmunoGen), Sara Varricchio (Sanofi), Wei Zhang (Biogen), Joseph Boyer (GSK) and other members of the BPDG-BPS for the useful discussions that have contributed to this article. Since its inception in 2004, the BioPhorum has become a trusted environment where senior leaders of the biopharma industry come together to openly share and discuss the emerging trends and challenges facing their industry. BioPhorum currently comprises more than 2000 active participants in seven Phorums – Drug Substance, The Development Group, Cell and Gene Therapy, Fill Finish, The Technology Roadmap, IT Group, and Supply Partners. The Bioassay point share is part of the Development Group. The manuscript is a composite view of opinions shared by the whole of the BPDG-BPS and should not be attributed to the individual positions of the participating companies.

FINANCIAL & COMPETING INTERESTS DISCLOSURE

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

No writing assistance was utilized in the production of this manuscript.

OPEN ACCESS

REFERENCES

Papers of special note have been highlighted as: • of interest

1. Karlsson R, Fridh V, Frostell A. Surrogate potency assays: comparison of binding profiles complements dose response curves for unambiguous assessment of relative potencies. *J. Pharmaceut. Anal.* 8(2), 138–146 (2018).
2. Ecker DM, Jones SD, Levine HL. The therapeutic monoclonal antibody market. *mAbs* 7(1), 9–14 (2015).
3. Graul AI, Pina P, Cruces E, Stringer M. The year's new drugs and biologics 2018: Part I. *Drugs Today (Barc.)* 55(1), 35–87 (2019).
4. Grilo AL, Mantalaris A. The increasingly human and profitable monoclonal antibody market. *Trends Biotechnol.* 37(1), 9–16 (2019).
5. Singh S, Kumar NK, Dwiwedi P *et al.* Monoclonal antibodies: a review. *Curr. Clin. Pharmacol.* 13(2), 85–99 (2018).
6. Cymer F, Beck H, Rohde A, Reusch D. Therapeutic monoclonal antibody N-glycosylation – structure, function and therapeutic potential. *Biologicals* 52, 1–11 (2018).
7. Liu H, Saxena A, Sidhu SS, Wu D. Fc Engineering for developing therapeutic bispecific antibodies and novel scaffolds. *Front. Immunol.* 8, 38 (2017).
8. Ricklin D, Mastellos DC, Reis ES, Lambris JD. The renaissance of complement therapeutics. *Nat. Rev. Nephrol.* 14(1), 26–47 (2018).
9. Shepard HM, Phillips GL, Thanos CD, Feldmann M. Developments in therapy with monoclonal antibodies and related proteins. *Clin. Med. (Lond.)* 17(3), 220–232 (2017).
10. Schnueriger A, Grau R, Sondermann P, Schreitmueller T, Marti S, Zocher M. Development of a quantitative, cell-line based assay to measure ADCC activity mediated by therapeutic antibodies. *Mol. Immunol.* 48(12–13), 1512–1517 (2011).
11. ICH. International Congress on Harmonization Q6B: specifications: test procedures and acceptance criteria for biotechnological/biological products. Geneva, Switzerland (1999).
12. Michelini E, Cevenini L, Mezzanotte L, Coppa A, Roda A. Cell-based assays: fuelling drug discovery. *Anal. Bioanal. Chem.* 398(1), 227–238 (2010).
   - **Highlights the importance and variety of cell-based assays used in drug discovery and development.**
13. Ritter N, Russell R, Schofield T *et al.* Bridging analytical methods for release and stability testing: technical, quality and regulatory considerations. *Bioprocess Int.* 14(2), 13–23 (2016).
   - **As methods develop and are used in later stages of drug development, the importance of a good bridging strategy becomes important.**
14. Wang L, Xu GL, Gao K *et al.* Development of a robust reporter-based assay for the bioactivity determination of anti-VEGF therapeutic antibodies. *J. Pharm. Biomed. Anal.* 125, 212–218 (2016).
15. Ritter N, Advant SJ, Hennessey J, Simmerman H, McEntire J, Mire-Sluis A, Joneckis C. What is Test Method Qualification. *Bioprocess Tech.* 2, 32–47 (2004).
16. Bruhns P, Iannascoli B, England P *et al.* Specificity and affinity of human Fcgamma receptors and their polymorphic variants for human IgG subclasses. *Blood* 113(16), 3716–3725 (2009).
17. Gazzano-Santoro H, Chan L, Ballard M, Young J. Ready-to-use cryopreserved primary cells. *BioProcess Int.* 12(2), 28–39 (2014).
   - **Good example of the cryopreserved cell strategy with difficult-to-culture primary cells.**
18. Lundholt BK, Scudder KM, Pagliaro L. A simple technique for reducing edge effect in cell-based assays. *J. Biomol. Screen.* 8(5), 566–570 (2003).
19. Robinson J, Sadick M, Deming S, Estdale S, Bergelson S, Little L. Assay acceptance criteria for multiwell-plate–based biological potency assays. *BioProcess Tech.* 12(1), 30–41 (2014).
20. Roselle C, Verch T, Shank-Retzlaff M. Mitigation of microtiter plate positioning effects using a block randomization scheme. *Anal. Bioanal. Chem.* 408(15), 3969–3979 (2016).
21. Schlain B, Jethwa H, Subramanyam M, Moulder K, Bhatt B, Molloy M. Designs for bioassays with plate location effects. *BioPharm. Int.* 14(11), 40–44 (2001).
22. Wagener J, Plennevaux C. Eppendorf 96-Well cell culture plate – a simple method of minimizing the edge effect in cell-based assays. *Eppendorf Application Note* 326 (2014).
23. Williams J, Birch J, Walfish S. A statistical method to account for plate-to-plate variability in multiple-plate bioassays. *BioPharm. Int.* 16, 44–54 (2003).
24. Schlain B, Jethwa H, Subramanyam M *et al.* Designs for bioassays with plate location effects. *BioPharm. Int.* 14(11), 40–44 (2001).
25. Sondag P, Joie R, Yang H. Comment and completion: implementation of parallelism testing for four-parameter logistic model in bioassays. *PDA J. Pharmaceut. Sci. Technol.* 69(4), 467–470 (2015).
26. Gottschalk PG, Dunn JR. The five-parameter logistic: a characterization and comparison with the four-parameter logistic. *Anal. Biochem.* 343(1), 54–65 (2005).
27. Gottschalk PG, Dunn JR. Measuring parallelism, linearity, and relative potency in bioassay and immunoassay data. *J. Biopharmaceut. Stat.* 15(3), 437–463 (2005).
28. USP. USP1032: design and development of biological assays. *United States Pharmacopeial Convention.*
29. Yang H, Kim HJ, Zhang L, Strouse RJ, Schenerman M, Jiang XR. Implementation of parallelism testing for four-parameter logistic model in bioassays. *PDA J. Pharmaceut. Sci. Technol.* 66(3), 262–269 (2012).
   - **In-depth discussion on the importance of establishing parallelism of test sample curves before deriving the potency of the test article.**
30. Hauck WW, Capen RC, Callahan JD *et al.* Assessing parallelism prior to determining relative potency. *PDA J. Pharmaceut. Sci. Technol.* 59(2), 127–137 (2005).
   - **One of the early papers presenting equivalence tests prior to the new USP bioassay chapters**
31. Bortolotto E, Rousseau R, Teodorescu B, Wielant A, Debauve G. Assessing similarity with parallel-line and parallel-curve models: implementing the USP development/validation approach to a relative potency assay. *Bioprocess Int.* 13(6), 26–37 (2015).
   - **A recent paper discussing the implementation of USP bioassay chapter guidelines and case study.**
32. Callahan J, Sajjadi N. Testing the null hypothesis for a specified difference – the right way to test for parallelism. *BioProcess J.* 2(2), 71–77 (2003).
33. Jonkman JN, Sidik K. Equivalence testing for parallelism in the four-parameter logistic model. *J. Biopharmaceut. Stat.* 19(5), 818–837 (2009).
34. Fleetwood K, Bursa F, Yellowlees A. Parallelism in practice: approaches to parallelism in bioassays. *PDA J. Pharmaceut. Sci. Technol.* 69(2), 248–263 (2015).
   - **A comparison of several statistical approaches to determine parallelism**
35. Li R, Cai W, Zocher M. A Novel lack-of-fit assessment as a system suitability test for potency assays. *PDA J. Pharmaceut. Sci. Technol.* 71(5), 368–378 (2017).
36. Hahn GJ, Meeker WQ. Statistical intervals: a guide for practitioners. John Wiley & Sons, 92 (2011).
37. Verma SP, Diaz-Gonzalez L, Rosales-Rivera M, Quiroz-Ruiz A. Comparative performance of four single extreme outlier discordancy tests from Monte Carlo simulations. *Sci. World J.* 2014, 746451 (2014).
38. Little T. Essentials in bioassay design and relative potency determination. *BioPharm. Int.* 29(4), 49–52 (2016).