Add the following:

■⟨1034⟩ ANALYSIS OF BIOLOGICAL ASSAYS

1. INTRODUCTION

Although advances in chemical characterization have reduced the reliance on bioassays for many products, bioassays are still essential for the determination of potency and the assurance of activity of many proteins, vaccines, complex mixtures, and products for cell and gene therapy, as well as for their role in monitoring the stability of biological products. The intended scope of general chapter *Analysis of Biological Assays* ⟨1034⟩ includes guidance for the analysis of results both of bioassays described in the *United States Pharmacopeia* (*USP*), and of non-USP bioassays that seek to conform to the qualities of bioassay analysis recommended by USP. Note the emphasis on analysis—design and validation are addressed in complementary chapters (*Development and Design of Bioassays* ⟨1032⟩ and *Biological Assay Validation* ⟨1033⟩, respectively).

Topics addressed in ⟨1034⟩ include statistical concepts and methods of analysis for the calculation of potency and confidence intervals for a variety of relative potency bioassays, including those referenced in *USP*. Chapter ⟨1034⟩ is intended for use primarily by those who do not have extensive training or experience in statistics and by statisticians who are not experienced in the analysis of bioassays. Sections that are primarily conceptual require only minimal statistics background. Most of the chapter and all the methods sections require that the nonstatistician be comfortable with statistics at least at the level of *USP* general chapter *Analytical Data—Interpretation and Treatment* ⟨1010⟩ and with linear regression. Most of sections 3.4 *Nonlinear Models for Quantitative Response* and 3.6 *Dichotomous (Quantal) Assays* require more extensive statistics background and thus are intended primarily for statisticians. In addition, ⟨1034⟩ introduces selected complex methods, the implementation of which requires the guidance of an experienced statistician.

Approaches in ⟨1034⟩ are recommended, recognizing the possibility that alternative procedures may be employed. Additionally, the information in ⟨1034⟩ is presented assuming that computers and suitable software will be used for data analysis. This view does not relieve the analyst of responsibility for the consequences of choices pertaining to bioassay design and analysis.

2. OVERVIEW OF ANALYSIS OF BIOASSAY DATA

Following is a set of steps that will help guide the analysis of a bioassay. This section presumes that decisions were made following a similar set of steps during development, checked during validation, and then not required routinely. Those steps and decisions are covered in general information chapter *Design and Development of Biological Assays* ⟨1032⟩. Section 3 *Analysis Models* provides details for the various models considered.

1. As a part of the chosen analysis, select the subset of data to be used in the determination of the relative potency using the prespecified scheme. Exclude only data known to result from technical problems such as contaminated wells, non-monotonic concentration–response curves, etc.
2. Fit the statistical model for detection of potential outliers, as chosen during development, including any weighting and transformation. This is done first without assuming similarity of the Test and Standard curves but should include important elements of the design structure, ideally using a model that makes fewer assumptions about the functional form of the response than the model used to assess similarity.
3. Determine which potential outliers are to be removed and fit the model to be used for suitability assessment. Usually, an investigation of outlier cause takes place before outlier removal. Some assay systems can make use of a statistical (noninvestigative) outlier removal rule, but removal on this basis should be rare. One approach to “rare” is to choose the outlier rule so that the expected number of false positive outlier identifications is no more than one; e.g., use a 1% test if the sample size is about 100. If a large number of outliers are found above that expected from the rule used, that calls into question the assay.
4. Assess system suitability. System suitability assesses whether the assay Standard preparation and any controls behaved in a manner consistent with past performance of the assay. If an assay (or a run) fails system suitability, the entire assay (or run) is discarded and no results are reported other than that the assay (or run) failed. Assessment of system suitability usually includes adequacy of the fit of the model used to assess similarity. For linear models, adequacy of the model may include assessment of the linearity of the Standard curve. If the suitability criterion for linearity of the Standard is not met, the exclusion of one or more extreme concentrations may result in the criterion being met. Examples of other possible system suitability criteria include background, positive controls, max/min, max/background, slope, IC_{50} (or EC_{50}), and variation around the fitted model.
5. Assess sample suitability for each Test sample. This is done to confirm that the data for each Test sample satisfy necessary assumptions. If a Test sample fails sample suitability, results for that sample are reported as “Fails Sample Suitability.” Relative potencies for other Test samples in the assay may still be reported. Most prominent of sample suitability criteria is similarity, whether parallelism for parallel models or equivalence of intercepts for slope-ratio models. For nonlinear models, similarity assessment involves all curve parameters other than EC_{50} (or IC_{50}).
6. For those Test samples in the assay that meet the criterion for similarity to the Standard (i.e., sufficiently similar concentration–response curves or similar straight-line subsets of concentrations), calculate relative potency estimates assuming similarity between Test and Standard, i.e., by analyzing the Test and Standard data together using a model constrained to have exactly parallel lines or curves, or equal intercepts.
7. A single assay is often not sufficient to achieve a reportable value, and potency results from multiple assays can be combined into a single potency estimate. Repeat steps 1–6 multiple times, as specified in the assay protocol or monograph, before determining a final estimate of potency and a confidence interval.
8. Construct a variance estimate and a measure of uncertainty of the potency estimate (e.g., confidence interval). See section 4 *Confidence Intervals*.

A step not shown concerns replacement of missing data. Most modern statistical methodology and software do not require equal numbers at each combination of concentration and sample. Thus, unless otherwise directed by a specific monograph, analysts generally do not need to replace missing values.

3. ANALYSIS MODELS

A number of mathematical functions can be successfully used to describe a concentration–response relationship. The first consideration in choosing a model is the form of the assay response. Is it a number, a count, or a category such as Dead/Alive? The form will identify the possible models that can be considered.

Other considerations in choosing a model include the need to incorporate design elements in the model and the possible benefits of means models compared to regression models. For purposes of presenting the essentials of the model choices, section 3 *Analysis Models* assumes a completely randomized design so that there are no design elements to consider and presents the models in their regression form.

3.1 Quantitative and Qualitative Assay Responses

The terms *quantitative* and *qualitative* refer to the nature of the response of the assay used in constructing the concentration–response model. Assays with either quantitative or qualitative responses can be used to quantify product potency. Note that the *responses* of the assay at the concentrations measured are not the relative potency of the bioassay. Analysts should understand the differences among responses, concentration–response functions, and relative potency.

A quantitative response results in a number on a continuous scale. Common examples include spectrophotometric and luminescence responses, body weights and measurements, and data calculated relative to a standard curve (e.g., cytokine concentration). Models for quantitative responses can be linear or nonlinear (see sections 3.2–3.5).

A qualitative measurement results in a categorical response. For bioassay, qualitative responses are most often quantal, meaning they entail two possible categories such as Positive/Negative, 0/1, or Dead/Alive. Quantal responses may be reported as proportions (e.g., the proportion of animals in a group displaying a property). Quantal models are presented in section 3.6. Qualitative responses can have more than two possible categories, such as end-point titer assays. Models for more than two categories are not considered in this general chapter.

Assay responses can also be counts, such as number of plaques or colonies. Count responses are sometimes treated as quantitative, sometimes as qualitative, and sometimes models specific to integers are used. The choice is often based on the range of counts. If the count is mostly 0 and rarely greater than 1, the assay may be analyzed as quantal and the response is Any/None. If the counts are large and cover a wide range, such as 500 to 2500, then the assay may be analyzed as quantitative, possibly after transformation of the counts. A square root transformation of the count is often helpful in such analyses to better satisfy homogeneity of variances. If the range of counts includes or is near 0 but 0 is not the preponderant value, it may be preferable to use a model specific for integer responses. Poisson regression and negative binomial regression models are often good options. Models specific to integers will not be discussed further in this general chapter.

Assays with quantitative responses may be converted to quantal responses. For example, what may matter is whether some defined threshold is exceeded. The model could then be quantal—threshold exceeded or not. In general, assay systems have more precise estimates of potency if the model uses all the information in the response. Using above or below a threshold, rather than the measured quantitative responses, is likely to degrade the performance of an assay.

3.2 Overview of Models for Quantitative Responses

In quantitative assays, the measurement is a number on a continuous scale. Optical density values from plate-based assays are such measurements. Models for quantitative assays can be linear or nonlinear. Although the two display an apparent difference in levels of complexity, parallel-line (linear) and parallel-curve (nonlinear) models share many commonalities. Because of the different form of the equations, slope-ratio assays are considered separately (section 3.5 *Slope-Ratio Concentration–Response Models*).

Assumptions—The basic parallel-line, parallel-curve, and slope-ratio models share some assumptions. All include a residual term, e, that represents error (variability) which is assumed to be independent from measurement to measurement and to have constant variance from concentration to concentration and sample to sample. Often the residual term is assumed to have a normal distribution as well. The assumptions of independence and equal variances are commonly violated, so the goal in analysis is to incorporate the lack of independence and the unequal variances into the statistical model or the method of estimation.

Lack of independence often arises because of the design or conduct of the assay. For example, if the assay consists of responses from multiple plates, observations from the same plate are likely to share some common influence that is not shared with observations from other plates. This is an example of intraplate correlation. A simple approach for dealing with this lack of independence is to include a block term in the statistical model for plate. With three or more plates this should be a random effects term so that we obtain an estimate of plate-to-plate variability.

In general, the model needs to closely reflect the design. The basic model equations given in sections 3.3–3.5 apply only to completely randomized designs. Any other design will mean additional terms in the statistical model. For example, if plates or portions of plates are used as blocks, one will need terms for blocks.

Calculation of Potency—A primary assumption underlying methods used for the calculation of relative potency is that of similarity. Two preparations are similar if they contain the same effective constituent or same effective constituents in the same proportions. If this condition holds, the Test preparation behaves as a dilution (or concentration) of the Standard preparation. Similarity can be represented mathematically as follows. Let F_T be the concentration–response function for the Test, and let F_S be the concentration–response function for the Standard. The underlying mathematical model for similarity is:

$$F_T(z) = F_S(\rho\ z), \qquad [3.1]$$

where z represents the concentration and ρ represents the relative potency of the Test sample relative to the Standard sample.

Methods for estimating ρ in some common concentration–response models are discussed below. For linear models, the distinction between parallel-line models (section 3.3 *Parallel-Line Models for Quantitative Response*) and slope-ratio models (section 3.5 *Slope-Ratio Concentration–Response Models*) is based on whether a straight-line fit to log concentration or concentration yields better agreement between the model and the data over the range of concentrations of interest.

3.3 Parallel-Line Models for Quantitative Responses

In this section, a linear model refers to a concentration–response relationship, which is a straight-line (linear) function between the logarithm of concentration, x, and the response, y. y may be the response in the scale as measured or a transformation of the response. The functional form of this relationship is y = a + bx. Straight-line fits may be used for portions of nonlinear concentration–response curves, although doing so requires a method for selecting the concentrations to use for each of the Standard and Test samples (see ⟨1032⟩).

Means Model versus Regression—A linear concentration–response model is most often analyzed with least squares regression. Such an analysis results in estimates of the unknown coefficients (intercepts and slope) and their standard errors, as well as measures of the goodness of fit [e.g., R^2 and root-mean-square error (RMSE)].

Linear regression works best where all concentrations can be used and there is negligible curvature in the concentration–response data. Another statistical method for analyzing linear concentration–response curves is the *means model*. This is an analysis of variance (ANOVA) method that offers some advantages, particularly when one or more concentrations from one or more samples are not used to estimate potency. Because a means model includes a separate mean for each unique combination of sample and dose (as well as block or other effects associated with the design structure) it is equivalent to a saturated polynomial regression model. Hence, a means model provides an estimate of error that is independent of regression lack of fit. In contrast, a regression residual based estimate of error is a mixture of the assay error, as estimated by the means model, combined with lack of fit of the regression model. At least in this sense, the means model error is a better estimate of the residual error variation in an assay system.

Parallel-Line Concentration–Response Models—If the general concentration–response model (3.1 *Quantitative and Qualitative Assay Responses*) can be made linear in x = log(z), the resulting equation is then:

$$y = \alpha + \beta\log(z) + e = \alpha + \beta x + e,$$

where e is the residual or error term, and the intercept, α, and slope, β, will differ between Test and Standard. With the parallelism (equal slopes) assumption, the model becomes

$$y_S = \alpha + \beta\log(z) + e = \alpha_S + \beta x + e$$

$$y_T = \alpha + \beta\log(\rho z) + e = [\alpha + \beta\log(\rho)] + \beta x + e = \alpha_T + \beta x + e, \qquad [3.2]$$

where S denotes Standard, T denotes Test, $\alpha_S = \alpha$ is the y-intercept for the Standard, and $\alpha_T = \alpha + \beta\log(\rho)$ is the y-intercept for the Test (see *Figure 3.1*).

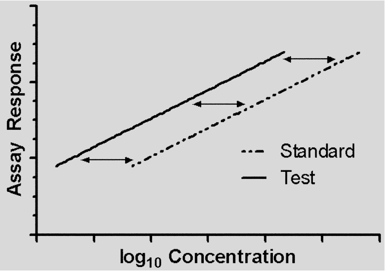

Figure 3.1. Example of parallel-line model.

Where concentration–response lines are parallel, as shown in *Figure 3.1*, a separation or horizontal shift indicates a difference in the level of biological activity being assayed. This horizontal difference is numerically log(ρ), the logarithm of the relative potency, and is found as the vertical distance between the lines α_T and α_S divided by the slope, β. The relative potency is then

$$\rho = \text{antilog}\left(\frac{\alpha_T - \alpha_S}{\beta}\right)$$

Estimation of Parallel-line Models—Parallel-line models are fit by the method of least squares. If the equal variance assumption holds, the parameters of equation [3.2] are chosen to minimize

$$\sum\left(y - \hat{\alpha}_S - \hat{\delta}T - \hat{\beta}x\right)^2 \qquad [3.3]$$

where the carets denote estimates. This is a linear regression with two independent variables, T and x, where T is a variable that equals 1 for observations from the Test and 0 for observations from the Standard. The summation in equation [3.3] is over all observations of the Test and Standard. If the equal variance assumption does not hold but the variance is known to be inversely proportional to a value, w, that does not depend on the current responses, the y's, and can be determined for each observation, then the method is weighted least squares

$$\sum w\left(y - \hat{\alpha}_S - \hat{\delta}T - \hat{\beta}x\right)^2 \qquad [3.4]$$

Equation 3.4 is appropriate only if the weights are determined without using the response, the y's, from the current data (see ⟨1032⟩ for guidance in determining weights). In equations [3.3] and [3.4] β is the same as the β in equation [3.2] and $\delta = \alpha_T - \alpha_S = \beta\log\rho$. So, the estimate of the relative potency, ρ, is

$$\hat{\rho} = \text{antilog}\left(\hat{\delta}\big/\hat{\beta}\right)$$

Commonly available statistical software and spreadsheets provide routines for least squares. Not all software can provide weighted analyses.

See section 4 for methods to obtain a confidence interval for the estimated relative potency. For a confidence interval based on combining relative potency estimates from multiple assays, use the methods of section 4.2. For a confidence interval from a single assay, use Fieller's Theorem (section 4.3) applied to $\hat{\delta}/\hat{\beta}$.

Measurement of Nonparallelism—Parallelism for linear models is assessed by considering the difference or ratio of the two slopes. For the difference, this can be done by fitting the regression model,

$$y = \alpha_S + \delta T + \beta_S x + \gamma x T + e$$

where $\delta = \alpha_T - \alpha_S$, $\gamma = \beta_T - \beta_S$, and T = 1 for Test data and T = 0 for Standard data. Then use the standard t-distribution confidence interval for γ. For the ratio of slopes, fit

$$y = \alpha_S + \delta T + \beta_S x(1 - T) + \beta_T x T + e$$

and use Fieller's Theorem, equation [4.3], to obtain a confidence interval for β_T/β_S.

3.4 Nonlinear Models for Quantitative Responses

Nonlinear concentration–response models are typically S-shaped functions. They occur when the range of concentrations is wide enough so that responses are constrained by upper and lower asymptotes. The most common of these models is the four-parameter logistic function as given below.

Let y denote the observed response and z the concentration. One form of the four-parameter logistic model is

$$y = D + \frac{A - D}{1 + \left(\frac{z}{C}\right)^B} + e \qquad [3.5]$$

One alternative, but equivalent, form is

$$y = a_0 + \frac{d}{1 + \text{antilog}\left[M(\log z - b)\right]} + e$$

The two forms correspond as follows:

Lower asymptote: D = a_0
Upper asymptote: A = a_0 + d
Steepness: B = M (related to the slope of the curve at the EC_{50})
Effective concentration 50% (EC_{50}): C = antilog(b) (may also be termed ED_{50}).

Any convenient base for logarithms is suitable; it is often convenient to work in log base 2, particularly when concentrations are twofold apart.

The four-parameter logistic curve is symmetric around the EC_{50} when plotted against log concentration because the rates of approach to the upper and lower asymptotes are the same (see *Figure 3.2*). For assays where this symmetry does not hold, asymmetrical model functions may be applied. These models are not considered further in this general chapter.

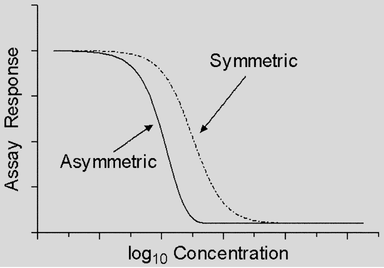

Figure 3.2. Examples of symmetric (four-parameter logistic) and asymmetric sigmoids.

In many assays the analyst has a number of strategic choices to make during assay development (see *Development and Design of Biological Assays* 〈1032〉). For example, the responses could be modeled using a transformed response to a four-parameter logistic curve, or the responses could be weighted and fit to an asymmetric sigmoid curve. Also, it is often important to include terms in the model (often random effects) to address variation in the responses (or parameters of the response) associated with blocks or experimental units in the design of the assay. For simple assays where observations are independent, these strategic choices are fairly straightforward. For assays performed with grouped dilutions (as with multichannel pipets), assays with serial dilutions, or assay designs that include blocks (as with multiple plates per assay), it is usually a serious violation of the statistical assumptions to ignore the design structure. For such assays, a good approach involves a transformation that approximates a solution to non-constant variance, nonnormality, and asymmetry combined with a model that captures the important parts of the design structure.

Parallel-Curve Concentration–Response Models—The concept of parallelism is not restricted to linear models. For nonlinear curves, parallel or similar means the concentration–response curves can be superimposed following a horizontal displacement of one of the curves, as shown in *Figure 3.3* for four-parameter logistic curves. In terms of the parameters of equation [3.5], this means the values of A, D, and B for the Test are the same as for the Standard.

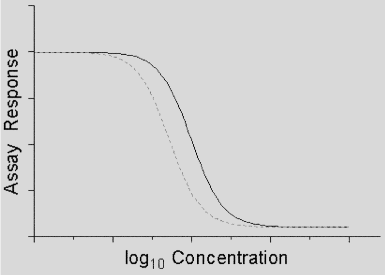

Figure 3.3. Example of parallel curves from a nonlinear model.

The equations corresponding to the figure (with error term, e, added) are

$$y_S = D + \frac{A - D}{1 + \left(\frac{z}{C}\right)^B} + e$$

$$y_T = D + \frac{A - D}{1 + \left(\frac{\rho z}{C}\right)^B} + e$$

or

$$y_S = D + \frac{A - D}{1 + \text{antilog}\left[M(\log z - b)\right]} + e$$

$$y_T = D + \frac{A - D}{1 + \text{antilog}\left[M(\log z - b + \log \rho)\right]} + e$$

Log ρ is the log of the relative potency and the horizontal distance between the two curves, just as for the parallel-line model. Because the EC_{50} of the standard is antilog(b) and that of the Test is antilog(b – log ρ) = antilog(b)/ρ, the relative potency is the ratio of EC_{50}'s (standard over Test) when the parallel-curve model holds.

Estimation of Parallel-Curve Models—Estimation of nonlinear, parallel-curve models is similar to that for parallel-line models, possibly after transformation of the response and possibly with weighting. For the four-parameter logistic model, the parameter estimates are found by minimizing:

$$\sum\left(y - \hat{D} - \frac{\hat{A} - \hat{D}}{1 + \text{antilog}\left[\hat{M}\left(\log z - \hat{b} + \hat{r}T\right)\right]}\right)^2$$

without weighting, or

$$\sum w\left(y - \hat{D} - \frac{\hat{A} - \hat{D}}{1 + \text{antilog}\left[\hat{M}\left(\log z - \hat{b} + \hat{r}T\right)\right]}\right)^2 \qquad [3.6]$$

with weighting. (As for equation [3.4], equation [3.6] is appropriate only if the weights are determined without using the responses, y's, from the current data.) In either case, the estimate of r is the estimate of the log of the relative potency. For some software, it may be easier to work with d = A – D.

The parameters of the four-parameter logistic function and those of the asymmetric sigmoid models cannot be found with ordinary (linear) least squares regression routines. Computer programs with nonlinear estimation techniques must be used.

Analysts should not use the nonlinear regression fit to assess parallelism or estimate potency if any of the following are present: a) inadequate asymptote information is available; or b) a comparison of pooled error(s) from nonlinear regression to pooled error(s) from a means model shows that the nonlinear model does not fit well; or c) other appropriate measures of goodness of fit show that the nonlinear model is not appropriate (e.g., residual plots show evidence of a "hook").

See section 4 for methods to obtain a confidence interval for the estimated relative potency. For a confidence interval based on combining relative potency estimates from multiple assays, use the methods of section 4.2. For a confidence interval from a single assay, advanced techniques, such as likelihood profiles or bootstrapping are needed to obtain a confidence interval for the log relative potency, r.

Measurement of Nonparallelism—Assessment of parallelism for a four-parameter logistic model means assessing the slope parameter and the two asymptotes. During development (see 〈1032〉), a decision should be made regarding which parameters are important and how to measure nonparallelism. As discussed in 〈1032〉, the measure of nonsimilarity may be a composite measure that considers all parameters together in a single measure, such as the parallelism sum of squares (see 〈1032〉), or may consider each parameter separately. In the latter case, the measure may be functions of the parameters, such as an asymptote divided by the difference of asymptotes or the ratio of the asymptotes. For each parameter (or function of parameters), confidence intervals can be computed by bootstrap or likelihood profile methods. These methods are not presented in this general chapter.

3.5 Slope-Ratio Concentration–Response Models

If a straight-line regression fits the nontransformed concentration–response data well, a slope-ratio model may be used. The equations for the slope-ratio model assuming similarity are then:

$$y_S = \alpha + \beta z + e = \alpha + \beta_S z + e$$
$$y_T = \alpha + \beta(\rho z) + e = \alpha + \beta_S \rho z + e = \alpha + \beta_T z + e \qquad [3.7]$$

An identifying characteristic of a slope-ratio concentration–response model that can be seen in the results of a ranging study is that the lines for different potencies from a ranging study have the same intercept and different slopes. Thus, a graph of the ranging study resembles a fan. *Figure 3.4* shows an example of a slope-ratio concentration–response model. Note that the common intercept need not be at the origin.

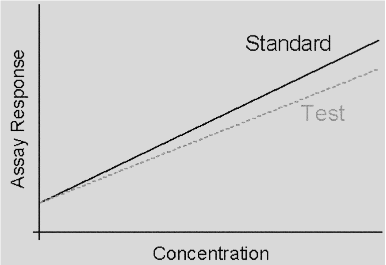

Figure 3.4. Example of slope-ratio model.

An assay with a slope-ratio concentration–response model for measuring relative potency consists, at a minimum, of one Standard sample and one Test sample, each measured at one or more concentrations and, usually, a measured response with no sample (zero concentration). Because the concentrations are not log transformed, they are typically equally spaced on the original, rather than log, scale. The model consists of one common intercept, a slope for the Test sample results, and a slope for the Standard sample results as in equation [3.7]. The relative potency is then found from the ratio of the slopes:

$$\text{Relative Potency} = \text{Test sample slope/Standard sample slope} = \beta\rho/\beta = \rho$$

Assumptions for and Estimation of Slope-Ratio Models—The assumptions for the slope-ratio model are the same as for parallel-line models: The residual terms are independent, have constant variance, and may need to have a normal distribution. The method of estimation is also least squares. This may be implemented either with or without weighting, as demonstrated in equations [3.8] and [3.9], respectively.

$$\sum\left(y - \hat{\alpha} - \hat{\beta}_S z(1-T) - \hat{\beta}_T zT\right)^2 \qquad [3.8]$$

$$\sum w\left(y - \hat{\alpha} - \hat{\beta}_S z(1-T) - \hat{\beta}_T zT\right)^2 \qquad [3.9]$$

Equation [3.9] is appropriate only if the weights are determined without using the response, the y's, from the current data. This is a linear regression with two independent variables, z(1 − T) and zT, where T = 1 for Test data and T = 0 for Standard data. $\hat{\beta}_T$ is the estimated slope for the Test, $\hat{\beta}_S$ the estimated slope for the Standard, and then the estimate of relative potency is

$$R = \hat{\beta}_T \Big/ \hat{\beta}_S$$

Because the slope-ratio model is a linear regression model, most statistical packages and spreadsheets can be used to obtain the relative potency estimate. In some assay systems, it is sometimes appropriate to omit the zero concentration (e.g., if the no-dose controls are handled differently in the assay) and at times one or more of the high concentrations (e.g., if there is a hook effect where the highest concentrations do not have the highest responses). The discussion about using a means model and selecting subsets of concentrations for straight parallel-line bioassays applies to slope-ratio assays as well.

See section 4 for methods to obtain a confidence interval for the estimated relative potency. For a confidence interval based on combining relative potency estimates from multiple assays, use the methods of section 4.2. For a confidence interval from a single assay, use Fieller's Theorem (section 4.3) applied to

$$\hat{\beta}_T / \hat{\beta}_S$$

Measurement of Nonsimilarity—For slope-ratio models, statistical similarity corresponds to equal intercepts for the Standard and Test. To assess the similarity assumption it is necessary to have at least two nonzero concentrations for each sample. If the intercepts are not equal, equation [3.7] becomes

$$y_S = \alpha_S + \beta_S z + e$$
$$y_T = \alpha_T + \beta_T z + e$$

Departure from similarity is typically measured by the difference of intercepts, $\alpha_T - \alpha_S$. An easy way to obtain a confidence interval is to fit the model,

$$y = \alpha_S + \delta T + \beta_S z(1 - T) + \beta_T zT + e,$$

where $\delta = \alpha T - \alpha_S$ and use the standard t-distribution-based confidence interval for δ.

3.6 Dichotomous (Quantal) Assays

For quantal assays the assay measurement has a dichotomous or binary outcome, e.g., in animal assays the animal is dead or alive or a certain physiologic response is or is not observed. For cellular assays, the quantal response may be whether there is or is not a response beyond some threshold in the cell. In cell-based viral titer or colony-forming assays, the quantal response may be a limit of integer response such as an integer number of particles or colonies. When one can readily determine if any particles are present—but not their actual number—then the assay can be analyzed as quantal. Note that if the reaction can be quantitated on a continuous scale, as with an optical density, then the assay is not quantal.

Models for Quantal Analyses—The key to models for quantal responses is to work with the probability of a response (e.g., probability of death), in contrast to quantitative responses for which the model is for the response itself. For each concentration, z, a treated animal, as an example, has a probability of responding to that concentration, P(z). Often the curve P(z) can be approximated by a sigmoid when plotted against the logarithm of concentration, as shown in *Figure 3.5*. This curve shows that the probability of responding increases with concentration. The concentration that corresponds to a probability of 0.5 is the EC_{50}.

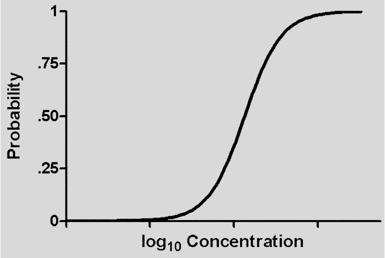

Figure 3.5. Example of sigmoid for P(z).

The sigmoid curve is usually modeled based on the normal or logistic distribution. If the normal distribution is used, the resulting analysis is termed probit analysis, and if the logistic is used the analysis is termed logit or logistic analysis. The probit and logit models are practically indistinguishable, and either is an acceptable choice. The choice may be based on the availability of software that meets the laboratory's analysis and reporting needs. Because software is more commonly available for logistic models (often under the term logistic regression) this discussion will focus on the use and interpretation of logit analysis. The considerations discussed in this section for logit analysis (using a logit transformation) apply as well to probit analysis (using a probit transformation).

Logit Model—The logit model for the probability of response, P(z), can be expressed in two equivalent forms. For the sigmoid,

$$\begin{aligned} P(z) &= \frac{1}{1+\text{antilog}\,[-\beta_0-\beta_1\log(z)]} \\ &= \frac{1}{1+(z/ED_{50})^{-\beta_1}} \end{aligned}$$

where $\log(ED_{50}) = -\beta_0/\beta_1$. An alternative form shows the relationship to linear models:

$$\text{logit transform of P} = \log\left(\frac{P}{1-P}\right) = \beta_0+\beta_1\log(z) \qquad [3.10]$$

The linear form is usually shown using natural logs and is a useful reminder that many of the considerations, in particular linearity and parallelism, discussed for parallel-line models in section 3.3 *Parallel-Line Models for Quantitative responses* apply to quantal models as well.

For a logit analysis with Standard and Test preparations, let T be a variable that takes the value 1 for animals receiving the Test preparation and 0 for animals receiving the Standard. Assuming parallelism of the Test and Standard curves, the logit model for estimating relative potency is then:

$$\log\left(\frac{P}{1-P}\right) = \beta_0+\beta_1\log(z)+\beta_2 T$$

The log of the relative potency of the Test compared to the Standard preparation is then β_2/β_1. The two curves in *Figure 3.6* show parallel Standard and Test sigmoids. (If the corresponding linear forms equation [3.10] were shown, they would be two parallel straight lines.) The log of the relative potency is the horizontal distance between the two curves, in the same way as for the linear and four-parameter logistic models given for quantitative responses (sections 3.3 *Parallel-Line Models for Quantitative Responses* and 3.4 *Nonlinear Models for Quantitative Responses*).

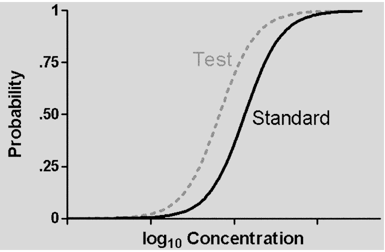

Figure 3.6. Example of Parallel Sigmoid Curves.

Estimating the Model Parameters and Relative Potency—Two methods are available for estimating the parameters of logit and probit models: maximum likelihood and weighted least squares. The difference is not practically important, and the laboratory can accept the choice made by its software. The following assumes a general logistic regression software program. Specialized software should be similar.

Considering the form of equation [3.10], one observes a resemblance to linear regression. There are two independent variables, x = log(z) and T. For each animal, there is a yes/no dependent variable, often coded as 1 for yes or *response* and 0 for *no* or *no response*. Although bioassays are often designed with equal numbers of animals per concentration, that is not a requirement of analysis. Utilizing the parameters estimated by software, which include β_0, β_1, and β_2 and their standard errors, one obtains the estimate of the natural log of the relative potency:

$$\text{Estimate of log of relative potency} = \frac{\hat{\beta}_2}{\hat{\beta}_1}$$

See section 4 for methods to obtain a confidence interval for the estimated relative potency. For a confidence interval based on combining relative potency estimates from multiple assays, use the methods of section 4.2. For a confidence interval from a single assay, use Fieller's Theorem (section 4.3) applied to $\hat{\beta}_2/\hat{\beta}_1$. The confidence interval for the relative potency is then [antilog(L), antilog(U)], where [L, U] is the confidence interval for the log relative potency.

Assumptions—Assumptions for quantal models have two parts. The first concerns underlying assumptions related to the probability of response of each animal or unit in the bioassay. These are difficult to verify assumptions that depend on the design of the assay. The second part concerns assumptions for the statistical model for P(z). Most important of these are parallelism and linearity. These assumptions can be checked much as for parallel-line analyses for quantitative responses.

In most cases, quantal analyses assume a standard binomial probability model, a common choice of distribution for dichotomous data. The key assumptions of the binomial are that at a given concentration each animal treated at that concentration has the same probability of responding and the results for any animal are independent from those of all other animals. This basic set of assumptions can be violated in many ways. Foremost among them is the presence of litter effects, where animals from the same litter tend to respond more alike than do animals from different litters. Cage effects, in which the environmental conditions or care rendered to any specific cage makes the animals from that cage more or less likely to respond to experimental treatment, violates the equal-probability and independence assumptions. These assumption violations and others like them (that could be a deliberate design choice) do not preclude the use of logit or probit models. Still, they are indications that a more complex approach to analysis than that presented here may be required (see ⟨1032⟩).

Checking Assumptions—The statistical model for P(z) assumes linearity and parallelism. To assess parallelism, equation [3.10] may be modified as follows:

$$\log\left(\frac{P}{1-P}\right)=\beta_0+\beta_1\log(z)+\beta_2 T+\beta_3 T*\log(z)$$

Here, β_3 is the difference of slopes between Test and Standard and should be sufficiently small. [The T*log(z) term is known as an *interaction term* in statistical terminology.] The measure of nonparallelism may also be expressed in terms of the ratio of slopes, $(\beta_1 + \beta_3)/\beta_1$. For model-based confidence intervals for these measures of nonparallelism, bootstrap or profile likelihood methods are recommended. These methods are not covered in this general chapter.

To assess linearity, it is good practice to start with a graphical examination. In accordance with equation [3.10], this would be a plot of log[(y + 0.5)/(n − y + 0.5)] against log(concentration), where y is the total number of responses at the concentration and n is the number of animals at that concentration. (The 0.5 corrections improve the properties of this calculation as an estimate of log[P/(1 − P)].) The lines for Test and Standard should be parallel straight lines as for the linear model in quantitative assays. If the relationship is monotonic but does not appear to be linear, then the model in [3.10] can be extended with other terms. For example, a quadratic term in log(concentration) could be added: $[\log(\text{concentration})]^2$. If concentration needs to be transformed to something other than log concentration, then the quantal model analogue of slope-ratio assays is an option. The latter is possible but sufficiently unusual that it will not be discussed further in this general chapter.

Outliers—Assessment of outliers is more difficult for quantal assays than for quantitative assays. Because the assay response can be only *yes* or *no*, no individual response can be unusual. What may appear to fall into the outlier category is a single response at a low concentration or a single no-response at a high concentration. Assuming that there has been no cause found (e.g., failure to properly administer the drug to the animal), there is no statistical basis for distinguishing an outlier from a rare event.

Alternative Methods—Alternatives to the simple quantal analyses outlined here may be acceptable, depending on the nature of the analytical challenge. One such challenge is a lack of independence among experimental units, as may be seen in litter effects in animal assays. Some of the possible approaches that may be employed are Generalized Estimating Equations (GEE), generalized linear models, and generalized linear mixed-effects models. A GEE analysis will yield standard errors and confidence intervals whose validity does not depend on the satisfaction of the independence assumption.

There are also methods that make no particular choice of the model equation for the sigmoid. A commonly seen example is the Spearman–Kärber method.

4. CONFIDENCE INTERVALS

A report of an assay result should include a measure of the uncertainty of that result. This is often a standard error or a confidence interval. An interval (c, d), where c is the lower confidence limit and d is the upper confidence limit, is a 95% confidence interval for a parameter (e.g., relative potency) if 95% of such intervals upon repetition of the experiment would include the actual value of the parameter. A confidence interval may be interpreted as indicating values of the parameter that are consistent with the data. This interpretation of a confidence interval requires that various assumptions be satisfied. Assumptions also need to be satisfied when the width or half width [(d-c)/2] are used in a monograph as a measure of whether there is adequate precision to report a potency. The interval width is sometimes used as a suitability criterion without the confidence interpretation. In such cases the assumptions need not be satisfied.

Confidence intervals can either be *model-based* or *sample-based*. A model-based interval is based on the standard errors for each of the one or more estimates of log relative potency that come from the analysis of a particular statistical model. Model-based intervals should be avoided if sample-based intervals are possible. Model-based intervals require that the statistical model correctly incorporate all the effects and correlations that influence the model's estimate of precision. These include but are not be limited to serial dilution and plate effects. Section 4.3 *Model-Based Methods* describes Fieller's Theorem, a commonly used model-based interval.

Sample-based methods combine independent estimates of log relative potency. Multiple assays may arise because this was determined to be required during development and validation or because the assay procedure fixes a maximum acceptable width of the confidence interval and two or more independent assays may be needed to meet the specified width requirement. Some sample-based methods do not require that the statistical model correctly incorporate all effects and correlations. However, this should not be interpreted as dismissing the value of addressing correlations and other factors that influence within-assay precision. The within-assay precision is used in similarity assessment and is a portion of the variability that is the basis for the sample-based intervals. Thus minimizing within-assay variability to the extent practical is important. Sample-based intervals are covered in section 4.2 *Combining Independent Assays (Sample-Based Confidence Interval Methods).*

4.1 Combining Results from Multiple Assays

In order to mitigate the effects of variability, it is appropriate to replicate independent bioassays and combine their results to obtain a single reportable value. That single reportable value (and not the individual assay results) is then compared to any applicable acceptance criteria. During assay development and validation, analysts should evaluate whether it is useful to combine the results of such assays and, if so, in what way to proceed.

There are two primary questions to address when considering how to combine results from multiple assays:

Are the assays mutually independent?

A set of assays may be regarded as mutually independent when the responses of one do not in any way depend on the distribution of responses of any of the others. This implies that the random errors in all essential factors influencing the result (for example, dilutions of the standard and of the preparation to be examined or the sensitivity of the biological indicator) in one assay must be independent of the corresponding random errors in the other assays. Assays on successive days using the original and retained dilutions of the Standard, therefore, are not independent assays. Similarly, if the responses, particularly the potency, depend on other reagents that are shared by assays (e.g., cell preparations), the assays may not be independent.

Assays need not be independent in order for analysts to combine results. However, methods for independent assays are much simpler. Also, combining dependent assay results may require assumptions about the form of the correlation between assay results that may be, at best, difficult to verify. Statistical methods are available for dependent assays, but they are not presented in this general chapter.

Are the results of the assays homogeneous?

Homogeneous results differ only because of random within-assay errors. Any contribution from factors associated with intermediate precision precludes homogeneity of results. Intermediate precision factors are those that vary between assays within a laboratory and can include analyst, equipment, and environmental conditions. There are statistical tests for heterogeneity, but lack of statistically significant heterogeneity is not properly taken as assurance of homogeneity and so no test is recommended. If analysts use a method that assumes homogeneity, homogeneity should be assessed during development, documented during validation, and monitored during ongoing use of the assay.

Additionally, before results from assays can be combined, analysts should consider the scale on which that combination is to be made. In general, the combination should be done on the scale for which the parameter estimates are approximately normally distributed. Thus, for relative potencies based on a parallel-line, parallel-curve, or quantal method, the relative potencies are combined in the logarithm scale.

4.2 Combining Independent Assays (Sample-Based Confidence Interval Methods)

Analysts can use several methods for combining the results of independent assays. A simple method described below (Method 1) assumes a common distribution of relative potencies across the assays and is recommended. A second procedure is provided and may be useful if homogeneity of relative potency across assays can be documented. A third alternative is useful if the assumptions for Methods 1 and 2 are not satisfied. Another alternative, analyzing all assays together using a linear or nonlinear mixed-effects model, is not discussed in this general chapter.

Method 1—Independent Assay Results From a Common Assay Distribution—The following is a simple method that assumes independence of assays. It is assumed that the individual assay results (logarithms of relative potencies) are from a common normal distribution with some nonzero variance. This common distribution assumption requires that all assays to be combined used the same design and laboratory procedures. Implicit is that the relative potencies may differ between the assays. This method thus captures inter-assay variability in relative potency. Note that the individual relative potencies should not be rounded before combining results.

Let R_i denote the logarithm of the relative potency of the ith assay of N assay results to be combined. To combine the N results, the mean, standard deviation, and standard error of the R_i are calculated in the usual way:

$$\text{Mean } \bar{R} = \sum_{i=1}^{N} R_i / N$$

$$\text{Standard Deviation } S = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(R_i - \bar{R})^2}$$

$$\text{Standard Error } SE = S / \sqrt{N}$$

A 100(1 − α)% confidence interval is then found as

$$\bar{R} \pm t_{N-1,\alpha/2}SE,$$

where $t_{N-1,\alpha/2}$ is the upper α/2 percentage point of a t-distribution with N − 1 degrees of freedom. The quantity $t_{N-1,\alpha/2}SE$ is the expanded uncertainty of $\bar{R}$. The number, N, of assays to be combined is usually small, and hence the value of t is usually large.

Because the results are combined in the logarithm scale, the combined result can be reported in the untransformed scale as a confidence interval for the geometric mean potency, estimated by antilog($\bar{R}$),

$$\text{antilog}\left(\bar{R} - t_{N-1,\alpha/2}SE\right),\ \text{antilog}\left(\bar{R} \pm t_{N-1,\alpha/2}SE\right)$$

Method 2—Independent Assay Results, Homogeneity Assumed—This method can be used provided the following conditions are fulfilled:

(1) The individual potency estimates form a homogeneous set with regard to the potency being estimated. Note that this means documenting (usually during development and validation) that there are no contributions to between-assay variability from intermediate precision factors. The individual results should appear to be consistent with homogeneity. In particular, differences between them should be consistent with their standard errors.

(2) The potency estimates are derived from independent assays.

(3) The number of degrees of freedom of the individual residual errors is not small. This is required so that the weights are well determined.

When these conditions are not fulfilled, this method cannot be applied and Method 1, Method 3, or some other method should be used. Further note that Method 2 (because it assumes no inter-assay variability) often results in narrower confidence intervals than Method 1, but this is not sufficient justification for using Method 2 absent satisfaction of the conditions listed above.

Calculation of Weighting Coefficients—It is assumed that the results of each of the N assays have been analyzed to give N estimates of log potency with associated confidence limits. For each assay, i, the logarithmic confidence interval for the log potency or log relative potency and a value L_i are obtained by subtracting the lower confidence limit from the upper. (This formula, using the L_i, accommodates asymmetric confidence intervals such as from Fieller's Theorem, section 4.3 *Model-Based Methods*). A weight W_i for each value of the log relative potency, R_i, is calculated as follows, where t_i has the same value as that used in the calculation of confidence limits in the ith assay:

$$W_i = \frac{4t_i^2}{L_i^2} \qquad [4.1]$$

Calculation of the Weighted Mean and Confidence Limits—The products W_iR_i are formed for each assay, and their sum is divided by the total weight for all assays to give the weighted mean log relative potency and its standard error as follows:

$$\text{Mean } \bar{R} = \sum_{i=1}^{N} W_i R_i \Big/ \sum_{i=1}^{N} W_i$$

$$\text{Standard Error SE} = 1 \Big/ \sqrt{\sum_{i=1}^{N} W_i}$$

A 100(1 – α)% confidence interval in the log scale is then found as

$$\bar{R} \pm t_{k,\alpha/2}\text{SE} \qquad [4.2]$$

where $t_{k,\alpha/2}$ is the upper α/2 percentage point of a t-distribution with degrees of freedom, k, equal to the sum of the number of degrees of freedom for the error mean squares in the individual assays. This confidence interval can then be transformed back to the original scale as for Method 1.

Method 3—Independent Assay Results, Common Assay Distribution Not Assumed—Method 3 is an approximate method that may be considered if the conditions for Method 1 (common assay distribution) or Method 2 (homogeneity) are not met.

The observed variation then has two components:

- the intra-assay variation for assay i:

$$s_i^2 = 1/W_i$$

- the inter-assay variation:

$$s_B^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(R_i - \bar{R}\right)^2 - \frac{1}{N}\sum_{i=1}^{N} s_i^2$$

For each assay, a weighting coefficient is then calculated as

$$W_i' = \frac{1}{s_i^2 + s_B^2}$$

which replaces W_i in equation [4.1] and where t in equation [4.2] is often approximated by the value 2.

4.3 Model-Based Methods

Many confidence intervals are of the form:

Confidence interval = value ± k times the standard error of that value.

For such cases, as long as the multiplier k can be easily determined (e.g., from a table of the t-distribution), reporting the standard error and the confidence interval are largely equivalent because the confidence interval is then easily determined from the standard error. However, the logarithms of relative potencies for parallel-line models and some parameterizations of nonlinear models and the relative potencies from slope-ratio models are ratios. In such cases, the confidence intervals are not symmetric around the estimated log relative potency or potency, and Fieller's Theorem is needed. For these asymmetric cases the confidence interval should be reported because the standard error by itself does not capture the asymmetry.

Fieller's Theorem is the formula for the confidence interval for a ratio. Let R = a/b be the ratio for which we need a confidence interval. For the estimates of a and b, we have their respective standard errors, SE_a and SE_b, and a covariance between them, denoted Cov. (The covariance is a measure of the degree to which the estimates of a and b are related and is proportional to the correlation between the estimates of a and b.) The covariance may be 0, as for some parameterizations of standard parallel-line analyses, but it need not be. The confidence interval for R then is as follows:

$$(R_L, R_U) = \frac{\left\{\hat{R} - \frac{g\text{Cov}}{\text{SE}_b^2} \pm \frac{t}{\hat{b}}\sqrt{(1-g)\text{SE}_a^2 + \hat{R}^2\text{SE}_b^2 - 2\hat{R}\text{Cov} + \frac{g\text{Cov}^2}{\text{SE}_b^2}}\right\}}{1-g}$$

where

$$g = \frac{t^2\text{SE}_b^2}{\hat{b}^2}$$

and t is the appropriate t deviate value that will depend on the sample size and confidence level chosen (usually 95%). If g > 1, it means that the denominator, $\hat{b}$, is not statistically significantly different from 0 and the use of the ratio is not sensible for those data.

For those cases where the estimates of a and b are statistically uncorrelated (Cov = 0), the confidence interval formula simplifies to

$$(R_L, R_U) = \frac{\left\{\hat{R} \pm \frac{t}{\hat{b}}\sqrt{(1-g)\text{SE}_a^2 + \hat{R}^2\text{SE}_b^2}\right\}}{1-g} \qquad [4.3]$$

5. ADDITIONAL SOURCES OF INFORMATION

A variety of statistical methods can be used to analyze bioassay data. This chapter presents several methods, but many other similar methods could also be employed. Additional information and alternative procedures can be found in the references listed below and other sources.

1. Bliss CI. *The Statistics of Bioassay.* New York: Academic Press; 1952.
2. Bliss CI. Analysis of the biological assays in U.S.P. XV. *Drug Stand.* 1956;24:33–67.
3. Böhrer A. One-sided and two-sided critical values for Dixon's outlier test for sample sizes up to n = 30. *Econ Quality Control.* 2008;23:5–13.
4. Brown F, Mire-Sluis A, eds. *The Design and Analysis of Potency Assays for Biotechnology Products.* New York: Karger; 2002.
5. Callahan JD, Sajjadi NC. Testing the null hypothesis for a specified difference—the right way to test for parallelism. *Bioprocessing J.* 2003:2;71–78.
6. DeLean A, Munson PJ, Rodbard D. Simultaneous analysis of families of sigmoidal curves: application to bioassay, radioligand assay, and physiological dose–response curves. *Am J Physiol.* 1978;235:E97–E102.
7. European Directorate for the Quality of Medicines. *European Pharmacopoeia,* Chapter 5.3, Statistical Analysis. Strasburg, France: EDQM; 2004:473–507.
8. Finney DJ. *Probit Analysis.* 3rd ed. Cambridge: Cambridge University Press; 1971.
9. Finney DJ. *Statistical Method in Biological Assay.* 3rd ed. London: Griffin; 1978.
10. Govindarajulu Z. *Statistical Techniques in Bioassay.* 2nd ed. New York: Karger; 2001.
11. Hauck WW, Capen RC, Callahan JD, et al. Assessing parallelism prior to determining relative potency. *PDA J Pharm Sci Technol.* 2005;59:127–137.
12. Hewitt W. *Microbiological Assay for Pharmaceutical Analysis: A Rational Approach.* New York: Interpharm/CRC; 2004.
13. Higgins KM, Davidian M, Chew G, Burge H. The effect of serial dilution error on calibration inference in immunoassay. *Biometrics.* 1998;54:19–32.
14. Hurlbert, SH. Pseudo replication and the design of ecological field experiments. *Ecological Monogr.* 1984;54:187–211.

15. Iglewicz B, Hoaglin DC. *How to Detect and Handle Outliers.* Milwaukee, WI: Quality Press; 1993.

16. Nelder JA, Wedderburn RWM. Generalized linear models. *J Royal Statistical Soc, Series A.* 1972;135:370–384.

17. Rorabacher DB. Statistical treatment for rejection of deviant values: critical values of Dixon's "Q" parameter and related subrange ratios at the 95% confidence level. *Anal Chem.* 1991;63:39–48.

APPENDIX–GLOSSARY

[NOTE—This glossary is applicable to 〈111〉, 〈1032〉, 〈1033〉, and 〈1034〉.]

GLOSSARY

The following is a glossary pertinent to biological assays. For some of this document's terms, the derivation may be clear. Rather than claiming originality, the authors seek to associate with this work a compendial perspective that will broadly provide clarity going forward; consistency with previous authoritative usage; and a useful focus on the bioassay context. In many cases the terms cited here have common usages or are defined in USP general chapter *Validation of Compendial Procedures* 〈1225〉, and the International Conference on Harmonization (ICH) Guideline Q2(R1), *Text on Validation of Analytical Procedures* (*1*). In such cases, the authors seek to be consistent, and they have made notes where a difference arose due to the bioassay context. Definitions from 〈1225〉 and ICH Q2 are identified as "1225" if taken without modification or "adopted from 1225" if taken with minor modification for application to bioassay. (Q2 and 〈1225〉 agree on definitions.) Most definitions are accompanied by notes that elaborate on the bioassay context.

I. General Terms Related to Bioassays

Analytical procedure (adopted from Q2A)—Detailed description of the steps necessary to perform the assay.

Notes: 1. The description may include but is not limited to the sample, the reference standard and the reagents, use of the apparatus, generation of the standard curve, use of the formulas for the calculation, etc. 2. An FDA Guidance provides a list of information that typically should be included in the description of an analytical procedure (*2*).

Assay—Analysis (as of a drug) to determine the quantity of one or more components or the presence or absence of one or more components.

Notes: 1. *Assay* often is used as a verb synonymous with *to determine*, as in, "I will assay the material for impurities." In this glossary, *assay* is a noun and is synonymous with the *analytic procedure (protocol)*. 2. The phrase "to run the assay" means to perform the analytical procedure(s) as specified.

Assay data set—The set of data used to determine a single potency or relative potency for all samples included in the bioassay.

Notes: 1. The definition of an assay data set can be subject to interpretation as necessarily a *minimal* set. It is important to understand that it may be possible to determine a potency or relative potency from a set of data but not to do this well. It is *not* the intent of this definition to mean that an assay data set is the *minimal* set of data that can be used to determine a relative potency. In practice, an assay data set should include, at least, sufficient data to assess similarity (q.v.). It also may include sufficient data to assess other assumptions. 2. It is also not an implication of this definition that assay data sets used together in determining a reportable value (q.v.) are necessarily independent from one another, although it may be desirable that they be so. When a run (q.v.) consists of multiple assay data sets, independence of assay sets within the run must be evaluated.

Bioassay, biological assay (these terms are interchangeable)—Analysis (as of a drug) to quantify the biological activity/activities of one or more components by determining its capacity for producing an expected biological activity, expressed in terms of units.

Notes: 1. Typically a bioassay involves controlled administration of the drug substance to living matter, in vivo or in vitro, followed by observation and assessment of the extent to which the expected biological activity has been manifested. 2. The description of a bioassay includes the analytic procedure, which should include the statistical design for collecting data and the method of statistical analysis that eventually yields the estimated potency or relative potency. 3. Bioassays can be either direct or indirect.

Direct bioassays—Measure the concentration of a substance that is required to elicit a specific response. For example, the potency of digitalis can be directly estimated from the concentration required to stop a cat's heart. In a direct assay, the response must be distinct and unambiguous. The substance must be administered in such a manner that the exact amount (threshold concentration) needed to elicit a response can be readily measured and recorded.

Indirect bioassays—Compare the magnitude of responses for nominally equal concentrations of reference and test preparations rather than test and reference concentrations that are required to achieve a specified response. Most biological assays in *USP* are indirect assays that are based on either quantitative or quantal (yes/no) responses.

Potency—[21 CFR 600.3(s)] The specific ability or capacity of the product, as indicated by appropriate laboratory tests or by adequately controlled clinical data obtained through the administration of the product in the manner intended, to effect a given result.

Notes: 1. A wholly impotent sample has no capacity to produce the expected specific response, as a potent sample would. Equipotent samples produce equal responses at equal dosages. Potency is typically measured relative to a reference standard or preparation that has been assigned a single unique value (e.g., 100%) for the assay; see *relative potency*. At times, additional qualifiers are used to indicate the physical standard employed (e.g., "international units"). 2. Some biological products have multiple uses and multiple assays. For such products there may be different reference lots that do not have consistently ordered responses across a collection of different relevant assays. 3. [21 CFR 600.10] Tests for potency shall consist of either in vitro or in vivo tests, or both, which have been specifically designed for each product so as to indicate its potency in a manner adequate to satisfy the interpretation of potency given by the definition in 21 CFR 600.3(s).

Relative potency—A measure obtained from the comparison of a Test to a Standard drug substance on the basis of capacity to produce the expected biological activity.

Notes: 1. A frequently invoked perspective is that relative potency is the degree to which the test preparation is diluted or concentrated relative to the standard. 2. Relative potency is unitless and is given definition, for any test material, solely in relation to the reference material and the assay.

Reportable value—The potency or relative potency estimate of record that is intended to achieve such measurement accuracy and precision as are required for use.

Notes: 1. The reportable value is the value that will be compared to a product specification. The specification may be in the *USP* monograph, or it may be set by the company, e.g., for product release. 2. The term *reportable value* is inextricably linked to the "intended use" of an analytical procedure. Tests are performed on samples in order to yield results that can be used to evaluate some parameter of the sample in some manner. One type of test may be configured in two different ways because the resulting data will be used for two different purposes (e.g., lot release versus stability). The reportable value would likely be different even if the mechanics of the test itself were identical. Validation is required to support the properties of each type of reportable value. In practice there may be one physical document that is the analytical procedure used for more than one application, but each application must be detailed separately within that document. Alternatively, there may be two separate documents for the two applications. 3. When the inherent variability of a biological response, or that of the log potency, precludes a single assay data set's attaining a value sufficiently accurate and precise to meet an assay specification, the assay may consist of multiple blocks or complete replicates, as necessary. The number of blocks or complete replicates needed depends on the assay's inherent accuracy and precision and on the intended use of the reported value. It is practical to improve the precision of a reported value by reporting the geometric mean potency from multiple assays. The number of assays used is determined by the relationship between the precision required for the intended use and the inherent precision of the assay system.

Run—That performance of the analytical procedure that can be expected to have consistent precision and trueness; usually, the assay work that can be accomplished by a laboratory team in a set time with a given unique set of assay factors (e.g., standard preparations).

Notes: 1. There is no necessary relationship of *run* to *assay data set* (q.v.). The term *run* is laboratory specific and relates to the physical capability of a team and its physical environment. An example of a run is given by one analyst's simultaneous assay of several samples in one day's bench work. During the course of a single run, it may be possible to determine multiple reportable values. Conversely, a single assay or reportable value may include data from multiple runs. 2. From a statistical viewpoint, a run is one realization of the factors associated with intermediate precision (q.v.). It is good practice to associate runs with factors that are significant sources of variation in the assay. For example, if cell passage number is an important source of variation in the assay response obtained, then each change in cell passage number initiates a new run. If the variance associated with all factors that could be assigned to runs is negligible, then the influence of runs can be ignored in the analysis and the analysis can focus on combining independent analysis data sets. 3. When a run contains multiple assays, caution is required regarding the independence of the assay results. Factors that are typically associated with runs and that cause lack of independence include cell preparations, groups of animals, analyst, day, a common preparation of reference material, and analysis with other data from the same run. Even though a strict sense of independence may be violated because some elements are shared among the assay sets within a run, the degree to which independence is compromised *may* have negligible influence on the reportable values obtained. This should be verified and monitored.

Similar preparations (similarity)—The property that the Test and Standard contain the same effective constituent, or the same effective constituents in fixed proportions, and all other constituents are without effect.

Notes: 1. Similarity is often summarized as the property that the Test behaves as a dilution (or concentration) of the Standard. 2. Similarity is fundamental to methods for determination of relative potency. Bioassay similarity requires that the reference and test samples should be sufficiently similar for legitimate calculation of relative potency. Given demonstration of similarity, a relative potency can be calculated, reported, and interpreted. Relative potency is valuable in assessing consistency and also intra- and intermanufacturer comparability in the presence of change. In the absence of similarity, a meaningful relative potency cannot be reported or interpreted. 3. The practical *consequence* of similarity is a comparable form of dose and/or concentration–response behavior. 4. Failure to statistically demonstrate dissimilarity between a reference and a test sample does not amount to demonstration of similarity. To assess similarity it is not sufficient to fail to find evidence that a reference and a test sample are not similar.

II. Terms Related to Performing a Bioassay

Configuration, assay (also known as assay format)—The arrangement of experimental units (q.v.) by number, position, location, temporal treatment, etc. and the corresponding test, control, or reference sample dilution that will be applied to each.

Notes: 1. The assay configuration must be specified in the formalized assay protocol. 2. Assay configuration can include nested dimensions like plate design, multiple plates per day, single plates on multiple days, etc. The configuration will depend on what the variance analysis (performed during assay development) reveals regarding sources of variability on assay response.

Out of specification—The property of a measurement in which it falls outside its acceptable range.

Sample suitability—A sample is suitable (may be described as having a potency) if its response curve satisfies certain properties defined in the protocol.

Note: Most significant of these properties is that of similarity to the standard response curve. If this property of similarity is satisfied, then the sample is suitable for the assay and can be described via a relative potency estimate.

System suitability—The provision of assurance that the laboratory control procedure is capable of providing legitimate measurements as defined in the validation report.

Notes: 1. System suitability may be thought of as an assessment of current validity achieved at the time of assay performance. An example is provided by positive and negative controls giving values within their normal ranges, ensuring that the assay system is working properly. 2. As described in *USP* general chapter *Validation of Compendial Procedures* ⟨1225⟩ and ICH Q2, system suitability testing is an integral part of many analytical procedures. The tests are based on the concept that the equipment, electronics, analytical operations, and samples to be analyzed constitute an integral system that can be evaluated as such. System suitability test parameters to be established for a particular procedure depend on the type of procedure being validated. *USP–NF* is a source of many system suitability tests.

III. Terms Related to Precision and Accuracy

Accuracy (1225)—An expression of the closeness of agreement between the value that is accepted either as a conventional true value or an accepted reference value and the value found.

Notes: 1. ICH and ISO give the same definition of accuracy. However, ISO specifically regards accuracy as having two components, bias and precision (*3*). That is, to be accurate as used by ISO, a measurement must be both "on target" (have low bias) and precise. In contrast, ICH Q2 says that accuracy is sometimes termed "trueness" but does not define trueness. ISO defines trueness as the "closeness of agreement between the average value obtained from a large series of test results and an accepted reference value" and indicates that "trueness is usually expressed in terms of bias." The 2001 FDA guidance on Bioanalytical Method Validation defines accuracy in terms of "closeness of *mean* test results" (emphasis added) and is thus consistent with the ICH usage. This glossary adopts the USP/ICH approach. That is, it uses the phrase "accurate and precise" to indicate low bias (accurate) and low variability (precise). 2. Considerable caution is needed when using or reading the term *accuracy.* In addition to the inconsistency between USP/ICH and ISO, common usage is not consistent.

Error, types of—Two sources of uncertainty that affect the results of a biological assay are systematic and random error.

A **systematic error** is one that happens with similar magnitude and consistent direction repeatedly. This introduces a *bias* in the determination. Effective experimental design, including randomization and/or blocking, can reduce systematic error.

A **random error** is one whose magnitude and direction vary without pattern. Random error is an inherent variability or uncertainty of the determination. Transformation of systematic into random error will increase the robustness of a biological assay and allow a comparatively simple analysis of assay data.

Format (configuration) variability—Predicted variability for a particular assay format.

Geometric standard deviation (%GSD)—The variability of the log-transformed values of a log normal response expressed as a percent in the untransformed scale.

Note: For example, if the standard deviation of log potency is σ using log base 2, the %GSD of potency is 100 * 2^{σ}.

Intermediate precision (adopted from 1225)—Expresses within-laboratory precision associated with changes in operating conditions.

Notes: 1. Factors contributing to intermediate precision involve anything that can change within a given laboratory and that may affect the assay, including different days, different analysts, different equipment, etc. Intermediate precision is thus "intermediate" in scope between the extremes of repeatability and reproducibility. 2. Any statement of intermediate precision should include clarification about which factors varied. For example, "The intermediate precision associated with changing equipment and operators is ..." 3. There can also be value in separately identifying the precision associated with each source, e.g., interanalyst precision. This may be part of assay development and validation when there is value in identifying which are the important contributors to intermediate precision. 4. When reporting intermediate precision, particularly for individual sources, analysts should take care to distinguish between intermediate precision variance and components of that variance. The variance includes repeatability and thus must be necessarily at least as large as the repeatability variance. A variance component, e.g., for analyst, is also a part of the intermediate precision variance for analyst, but it could be negligible and need not be larger in magnitude than the repeatability variance.

Precision (1225)—The closeness of agreement (degree of scatter) between a series of measurements obtained from multiple sampling of the same homogeneous sample under the prescribed conditions.

Notes: 1. Precision may be considered at three levels: repeatability (q.v.), intermediate precision (q.v.), and reproducibility (q.v.). 2. Precision should be investigated using homogeneous, authentic samples. However, if it is not possible to obtain a homogeneous sample, precision may be investigated using artificially prepared samples or a sample solution. 3. Precision is usually expressed as the variance, standard deviation, coefficient of variation, or geometric standard deviation.

Relative bias—Degree of difference from the true value expressed as a percent.

Repeatability (1225)—The expression of the precision under the same operating conditions over a short interval of time.

Notes: 1. ICH Q2A says that repeatability is also termed "intra-assay" precision. In the bioassay context, the better term is "intra-run," and a "short interval of time" is meant to connote "within-run." 2. The idea of a "short interval of time" can be problematic with bioassay. If a run takes multiple weeks and consists of a single assay set, then intra-run precision cannot be determined. Alternatively, if a run consists of two assay data sets and a run can be done in a single day, repeatability of the relative potency determination can be assessed. 3. Operating conditions include, but not limited to, equipment and analyst.

Reproducibility (1225)—Expresses the precision between laboratories.

Notes: 1. Reproducibility includes contributions from repeatability and all factors contributing to intermediate precision, as well as any additional contributions from interlaboratory differences. 2. Reproducibility applies to collaborative studies such as those for standardization or portability of methodology. Depending on the design of the collaborative study, it may be possible to separately describe variance components associated with intra- and interlaboratory sources of variability.

Specificity (1225)—The ability to assess unequivocally the analyte in the presence of components that may be expected to be present.

Note: Typically these components may include impurities, degradants, matrix, etc.

IV. Terms Related to Validation

Detection limit (adopted from 1225)—The lowest amount of analyte in a sample that can be detected but not necessarily quantified or quantified to any given level of precision and accuracy.

Linearity, dilutional (adopted from 1225)—The ability (within a given range) of a bioassay to obtain log relative potencies that are directly proportional to the log relative potency of the sample.

Notes: 1. Dilutional linearity, sometimes called bioassay linearity, is demonstrated across a range of known relative potency values by considering a plot of true log potency versus observed log potency. If that plot yields an essentially straight line with a y-intercept of 0 and a slope of 1, the assay has direct proportionality. If that plot yields an essentially straight line but either the y-intercept is not 0 or the slope is not 1 (or both), the assay has a proportional linear response. 2. To assess whether the slope is (near) 1.0 requires an a priori equivalence or indifference interval. It is not proper statistical practice to test the null hypothesis that the slope is 1.0 against the alternative that it is not 1.0 and conclude a slope of 1.0 if this is not rejected. Assay linearity is separate from consideration of the shape of the concentration–response curve. Linearity of concentration–response is *not* a requirement of assay linearity.

Quantitation limits (adopted from 1225)—The limits of true relative potencies between which the assay has suitable precision and accuracy.

Note: This applies to assay results (log potency) rather than the reportable value.

Range (adopted from 1225)—The interval between the upper and lower true relative potencies for which the bioassay is demonstrated to have a suitable level of precision, accuracy, and assay linearity.

Note: This applies to assay results (log potency) rather than the reportable value.

Robustness (1225)—A measure of an analytical procedure's capacity to remain unaffected by small but deliberate variations in method parameters.

Notes: 1. Robustness is an indication of a bioassay's reliability during normal usage. For example, a cell culture assay system that is robust to the passage number of the cells would provide potency values with equivalent accuracy and precision across a consistent range of passage numbers. 2. ICH Q2 states:

> the evaluation of robustness should be considered during the development phase and depends on the type of procedure under study. It should show the reliability of an analysis with respect to deliberate variations in method parameters. If measurements are susceptible to variations in analytical conditions, the analytical conditions should

be suitably controlled, or a precautionary statement should be included in the procedure. One consequence of the evaluation of robustness should be that a series of system suitability [q.v.] parameters is established to ensure that the validity of the analytical procedure is maintained whenever used.

Validation, assay—A formal, archived demonstration of the analytical capacity of an assay that provides justification for use of the assay for an intended purpose and a range of acceptable potency values.

Note: Formal validations are conducted prospectively according to a written, approved plan.

V. Terms Related to Statistical Design and Analysis

Analysis of variance (ANOVA)—A statistical tool used to assess contributions of variability from experimental factors.

Blocking—The grouping of related experimental units in experimental designs.

Notes: 1. Blocking is often used to reduce the variability of a measure of interest. 2. Blocks may consist of groups of animals (a cage, a litter, or a shipment), individual 96-well plates, sections of 96-well plates, or whole 96-well plates grouped by analyst, day, or batch of cells. 3. The goal is to isolate a systemic effect, such as cage, so that it does not obscure the effects of interest.

A **complete block design** occurs when all levels of a treatment factor (in a bioassay, the primary treatment factors are sample and concentration) can be applied to experimental units for that factor within a single block. Note that the two treatment factors, sample and concentration, may have different experimental units. For example, if the animals within a cage are all assigned the same concentration but are assigned unique samples, then the experimental unit for concentration is cage and the experimental unit for sample is animal; cage is a blocking factor for sample.

An **incomplete block design** occurs when the number of levels of a treatment factor exceeds the number of experimental units for that factor within the block.

Confidence interval—A statistical interval expressing the likely value of a parameter.

Confounded design—Two factors are confounded if their levels vary together (they are not crossed).

Notes: 1. For example, in a bioassay validation experiment in which one analyst performs assays on a set of samples for three days using cells from one passage number, then another analyst performs assays on the same set of samples for another three days using cells from a different passage number, the passage number of the cells and the analysts are confounded. [Also note that days are nested (q.v.) within analyst and cell passage number.] When factors are confounded one cannot tell which of the factors has caused an observed experimental difference. 2. Fractional factorial designs (q.v.), in which factors are only partially crossed, also are partially confounded. A full factorial design also can be confounded if the number of treatment combinations (sample and concentration) is greater than the block size.

Crossed (and partially crossed)—Two factors are crossed (or fully crossed) if each level of each factor appears with each level of the other factor. Two factors are partially crossed when they are not fully crossed but multiple levels of one factor appear with a common level of the other factor.

Notes: 1. For example, in a bioassay in which all samples appear at all dilutions, samples and dilutions are (fully) crossed. In a bioassay validation experiment in which two of four analysts each perform assays on the same set of samples on each of six days and a different pair of analysts is used on each day the analysts are partially crossed with days. 2. Each factor may be applied to different experimental units, and the factors may be both fully crossed and nested (q.v.), creating a split-unit or split-plot design (q.v.). 3. Experiments with factors that are partially crossed require particular care for proper analysis. 4. A randomized complete block design (RCBD) (q.v.) is a design in which the block factor (which often is treated as a random effect) is crossed with the treatment factor (which is usually treated as a fixed effect).

Design of experiments (DOE)—A systematic approach for studying multiple factors.

Note: DOE is used in bioassay development and validation; see ⟨1032⟩ and ⟨1033⟩.

Equivalence test—A test of conformance to interval-based target acceptance criteria.

Notes: 1. An equivalence test differs from most common statistical tests in the nature of the statistical hypotheses. For most common tests, the statistical null hypothesis is no difference and the alternative is that there is *some* difference, without regard to the magnitude or importance of the difference. In equivalence testing the alternative hypothesis is that the difference is sufficiently small so there is no important difference. 2. A common statistical procedure used for equivalence tests is the two one-sided test (TOST) procedure.

Expected mean square—A mathematical expression of variances estimated by an ANOVA mean square.

Experimental design—The structure of assigning treatments to experimental units.

Notes: 1. Blocking (q.v.), randomization (q.v.), replication (q.v.), and specific choice of design (cf. general chapter *Design and Development of Biological Assays* ⟨1032⟩) are some aspects of experimental design. 2. Important components of experimental design include the number of samples, the number of concentrations, and how samples and concentrations are assigned to experimental units and are grouped into blocks. 3. The experimental design influences which statistical methodology should be used to achieve the analytical objective.

Experimental unit—The smallest unit to which a distinct level of a treatment is randomly allocated.

Notes: 1. Randomization of treatment factors to experimental units is essential in bioassays. 2. Different treatment factors can be applied to different experimental units. For example, samples may be assigned to rows on a 96 well plate while dilutions are assigned to columns on the plate. In this case, rows are the experimental units for samples, columns are the experimental units for concentrations, and wells are the experimental units for the interaction of sample and concentration. 3. An experimental unit needs to be distinguished from a sampling unit, the smallest unit on which a distinct measurement is recorded (e.g., a well). Because the sampling unit is often smaller than the experimental unit, it is an easy mistake to treat sampling units as if they are experimental units. This mistake is called pseudoreplication (q.v.).

Factor—An assay design element that may affect assay response and that varies in an experiment.

Note: In a bioassay there will be at least two treatment factors—sample and concentration.

Fixed factor (fixed effect) is a factor that is deliberately set at specific levels in an experiment. Inference is made only to the levels used in the experiment. In a bioassay, sample and concentration are both fixed factors.

Random factor (random effect) is one for which its levels represent a sample of ways in which that factor might vary. In a bioassay, the test organisms, plate, and day often are considered random factors.

Factorial design—One in which there are multiple factors and the factors are partially or fully crossed.

In a **full factorial design**, each level of a factor appears with each combination of levels of all other factors. For example, if factors are sample (test and reference), concentration, and analyst, for a full factorial design each analyst must analyze all combinations of sample and concentration.

A **fractional factorial design** is one in which some factors are deliberately partially confounded with interactions associated with other combinations of factors.

General linear model—A statistical linear model that relates study factors, which can be continuous or discrete, to experimental responses.

Independence—For two measurements or observations A and B (raw data, assay sets, or relative potencies) to be independent, values for A must be unaffected by B's responses and vice versa.

Note: A consequence of nonrecognition of lack of independence is poor characterization of variance. In practice this means that if two potency or relative potency measurements share a common factor that might influence assay outcome such as analyst, cell preparation, incubator, group of animals, or aliquot of Standard samples, then the correct initial assumption is that these relative potency measurements are not independent. As assay experience is gained, an empirical basis may be established so that it is reasonable to treat potency measures as independent even if they share a common level of a factor. The same concern for lack of independence holds if the two potency or relative potency measurements are estimated together from the same model or are in any way associated without including in the model some term that captures that there are two or more potency measurements.

Interaction—Two factors are said to interact if the effect of one factor depends on the level of the other factor.

Level—A location on the scale of measurement of a factor.

Notes: 1. Factors have two or more distinct levels. For example, if a bioassay contains two samples, test and reference, then there are two levels for the factor sample. 2. Levels of a factor in a bioassay may be quantitative, such as concentration, or categorical, such as sample (i.e., test and reference).

Log normal distribution—A skewed distribution characterized by increased variability with increased level of response.

Note: A normal distribution is generated by taking the log of the response.

Mean square—A calculation in ANOVA representing the variability associated with an experimental factor.

Mixed-effects model—A statistical model including both fixed and random effects.

Modeling, statistical—The mathematical specification of the concentration–response relationship and important sources of variation in the bioassay.

Notes: 1. Modeling includes methods to capture the dependence of the response on the samples, concentration, and groups or blocking factors in the assay configuration. 2. Modeling of bioassay data includes making many choices, some of which are driven by data. For continuous data there is a choice between linear and nonlinear models. For discrete data there is a choice among logit/log models within a larger family of generalized linear models. In limiting dilution assays there is published literature advocating Poisson models and Markov chain binomial models. One can use either fixed-effects models or mixed-effects models for bioassay data. The fixed-effects models are more widely available in software and are somewhat less demanding for statisticians to set up. On the other hand, mixed models have advantages over fixed ones. The former are more accommodating of missing data and, more importantly, can allow each block to have different slopes, asymptotes, median effective concentrations required to induce a 50% effect (EC_{50}), or relative potencies. Particularly when the analyst is using straight-line models fit to nonlinear responses or in assay systems in which the concentration–response curve varies from block to block, the mixed model captures the behavior of the assay system in a much more realistic and interpretable way. 3. It is essential that any modeling approach for bioassay data use all available data simultaneously to estimate the variation (or, in a mixed model, each of several sources of variation). It may be necessary to transform the observations before this modeling; to include a variance model; or to fit a "means" model (in which there is a predicted effect for each combination of sample and concentration) to get pooled estimate(s) of variation.

Multiplicity—The property of compound risk with multiple independent events, all with fixed risk.

Nested—A factor A is nested within another factor B if the levels of A are different for every level of *B*.

Notes: 1. For example, in a bioassay validation experiment two analysts may perform assays on the same set of samples on each of six days when no analyst performs the assay on more than one day (this requires 12 analysts who are qualified to perform the assay); these analysts are nested within days. 2. Nested factors have a hierarchical relationship. 3. For two factors to be nested they must satisfy the following: a) be applied to different-sized experiment units; b) the larger experimental unit contains more than one of the smaller experimental units; and c) the factor applied to the smaller experimental unit is not fully crossed with the factor applied to the larger experimental unit. When conditions (a) and (b) are satisfied and the factors are partially crossed, then the experiment is partially crossed and partially nested. Experiments with this structure require particular care for proper analysis.

Parallelism (of concentration–response curves)—The concentration–response curves of the test and standard are identical in shape and differ only in a constant horizontal difference.

Notes: 1. When test and reference preparations are similar (q.v.) and assay responses are plotted against log concentrations, the resulting curve for the test preparation will be the same as that for the standard but shifted horizontally by an amount that is the logarithm of the relative potency. Because of this relationship, similarity (q.v.) is generally referred to as *parallelism.* Note that similarity is the primary concept and that parallelism is not necessary for similarity. See slope-ratio models in general chapter *Analysis of Biological Assays* 〈1034〉 in which samples with similar concentration–response relationships have a common (or nearly common) y-intercept but may differ in their slopes. 2. In practice, it is not possible to demonstrate that the shapes of two curves are exactly the same. Instead, the two curves are shown to be sufficiently similar (equivalent) in shape. Note that *similar* should be interpreted as "we have evidence that the two values are close enough" rather than "we don't have evidence that the two values are different." 3. The assessment of parallelism depends on the type of function used to fit the response curve. Parallelism for a nonlinear assay using a four-parameter logistic fit means that: a) the slopes of the rapidly changing parts of the sample and reference standard curves (that is, slope at tangent to the curve, where the first derivative is at a maximum) should be similar; and b) the upper and lower asymptotes of the response curves (plateaus) should be similar. For straight-line analysis, the slopes of the lines should be similar.

Point estimate—A single-value estimate obtained from statistical calculations.

Note: Examples are the average, standard deviation, and relative potency.

P-value (significance probability)—A statistical calculation representing the probability associated with observing an experimental outcome that is different from expectation.

Notes: 1. The P-value is the probability of observing what was seen or something more extreme under the assumption that the statistical null hypothesis is true. "More extreme" means further from the null hypothesis. 2. Commonly, $P < 0.05$ is taken as indicating statistical significance, though any value may be used.

Randomization—A process of assignment of treatment to experimental units based on chance so that all equal-sized groups of units have an equal chance of receiving a given treatment.

Notes: 1. The chance mechanism may be an unbiased physical process (rolling unbiased dice, flipping coins, drawing from a well-mixed urn), random-number tables, or computer-generated randomized numbers. Care must be taken in the choice and use of method. Good practice is to use a validated computerized random-number generator. 2. The use of randomization results in systematic error becoming random error not associated with particular samples or a dilution pattern but distributed throughout the assay. In 96-well bioassays, plate effects can be substantial and cause bias or trending, particularly in assays involving long-term cell culturing or multiple addition and wash steps. In animal studies, a variety of factors associated with individual animals can influence responses. If extraneous factors that influence either plate assays or animal assays are not routinely demonstrated to have been eliminated or minimized to be negligible, randomization is essential to obtain unbiased data required for the calculation of true potency. Randomization is central to the experimental design and analysis of data obtained from most biological assays.

Replication—A process in which multiple independent experimental units receive the same level of a treatment factor.

Notes: 1. The purpose of replication is to minimize the effects of uncontrollable sources of random variability. 2. Replication can occur either completely at random or across blocks. Generally, replication within blocks is pseudoreplication (see below).

True replicates—Samples based on independent experimental units.

Pseudoreplication—Is the identification of samples from experimental units as independent and thus true replicates when they are actually not independent.

Notes: 1. Pseudoreplication results in wrong inferences and the appearance of more replicates than are actually present. 2. Pseudoreplication is dangerous because it is an easy mistake to make, it is easy to overlook, and the consequences can be serious. For example, pseudoreplicates commonly arise when analysts are making a dilution series for each sample in tubes (the dilution series can be made with serial dilutions, by single-point dilutions, or with any convenient dilution scheme). The analyst then transfers each dilution of each sample to several wells on one or more assay plates. The wells are then pseudoreplicates because they are simply aliquots of a single dilution process. 3. In general, pseudoreplication should be avoided because, unless it is properly addressed in the analysis, it leads to underestimation of replicate variance. 4. The simple way to analyze data from pseudoreplicates is to average over the pseudoreplicates (if a transformation of the observed data is used, the transformation should be applied before averaging over pseudoreplicates) before fitting any sort of concentration–response model. In many assay systems averaging over pseudoreplicates will leave the assay without any replication. A more complex way to use data containing pseudoreplicates is to use a mixed model that treats the pseudoreplicates as a separate random effect. The only case in which pseudoreplication is useful is when the pseudoreplicate (i.e., well-to-well) variation is very large compared to the variation associated with replicates and the cost of pseudoreplicates is much lower than the cost of replicates.

Standard error of estimate—The variability associated with an estimate of a reportable value or other parameter.

Note: The standard error is also known as the standard uncertainty.

Statistical process control (SPC)—A set of statistical tools used to monitor for shifts and trends in a process.

Type I error (α)—The error made in judging data analysis, wherein the alternative hypothesis is accepted when it is false.

Type II error (β)—The error made in judging data analysis, wherein the alternative hypothesis is rejected when it is true.

Variance component analysis—A statistical analysis that divides total variability into its component parts.

GLOSSARY REFERENCES

1. ICH. Q2(R1): Text on Validation of Analytical Procedures: Text and Methodology. 2005. Available at: http://www.ich.org/fileadmin/Public Web Site/ICH Products/Guidelines/Quality/Q2 R1/Step4/Q2 RI Guideline.pdf. Accessed 22 July 2011.
2. FDA. Guidance for Industry. Analytical Procedures and Methods Validation: Chemistry, Manufacturing, and Controls Documentation. 2000. Available at: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM070489.pdf. Accessed 22 July 2011.
3. ISO. International Standard 5725-1. Accuracy (Trueness and Precision) of Measurement Methods and Results—Part 1: General Principles and Definitions. Geneva, Switzerland; 1994.■1S *(USP35)*

Add the following:

■〈1105〉 IMMUNOLOGICAL TEST METHODS—SURFACE PLASMON RESONANCE

Introduction

Surface plasmon resonance (SPR) optical detection is a useful method for the label-free assays (procedures) that study biomolecular interactions. Commercially available SPR biosensors that incorporate these assays can collect real-time, information-rich data from binding events. These data can be used widely from basic research to drug discovery and development to manufacturing and quality control (QC). SPR can characterize binding events with samples ranging from proteins, nucleic acids, and small molecules to complex mixtures, lipid vesicles, viruses, bacteria, and eukaryotic cells. Typical quality and safety attributes addressed with SPR analysis include:

- Interaction specificity
- Interaction affinity
- Kinetic binding parameters
- Thermodynamic parameters
- Biologically active concentration of an analyte

This chapter provides an overview of the physics underlying SPR and common instrument configurations, as well as the range of molecules that can be studied and general considerations for experimental design as determined by the assay objective.

Overview

History

The physical principles of SPR were first explained in the early 1900s, starting with a description of the uneven distribution of light in a diffraction grating spectrum caused by the excitation of surface plasmon waves. A landmark series of experiments showed the optical excitation of surface plasmons under conditions of total internal reflection and fostered detailed studies of the application of SPR for chemical and biological sensing. Since then, SPR's potential for characterizing thin films and monitoring interactions at metal interfaces has been recognized, and significant research and development have yielded instruments that can quantitatively evaluate the binding interactions of small and large molecules.

Physics

SPR is an optical phenomenon that occurs when a thin conducting film is placed between two media that have different refractive indices. In many commercially available instruments, the two media are glass and the sample solution, and the conducting film is preferentially a gold layer applied to the glass, although other conducting metals such as silver have been used. The glass–metal component comprises a solid support that is often referred to as a *sensor*.

Light applied to the glass under conditions of total internal reflection produces an electromagnetic component that is called an *evanescent wave*. The evanescent wave penetrates the medium of lower refractive index (typically the sample solution) without losing net energy. The amplitude of the evanescent wave decays exponentially with distance from the surface, roughly one-half of the wavelength of the incident light (e.g., for a light source of 760 nm the evanescent wave penetrates approximately 300 nm).

For a specific combination of wavelength and angle of incident light, electron charge density waves called *plasmons* are excited in the gold film. As energy is absorbed via the evanescent wave, a decrease in the intensity of the reflected light at a specific angle (the SPR angle) is observed. Analysts can conduct an SPR experiment by fixing the wavelength and varying the angle of incident light.

An increase in mass at the sensor surface caused by a binding interaction between two or more molecules causes a change in the local refractive index (RI) that gives rise to an SPR response, which is observed as a shift in the SPR angle. By monitoring the shift in the SPR angle as a function of time, an analyst can generate a *sensorgram* (*Figure 1*). The change in RI is very similar for different proteins, so the SPR measurement depends primarily on the mass change at the sensor surface and is relatively independent of the nature of the molecules being measured.

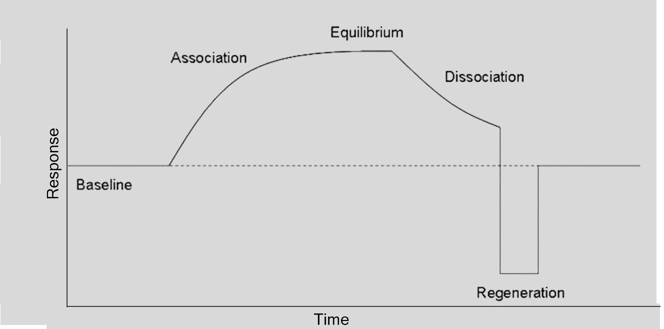

Figure 1. Representative sensorgram.

Instruments

The main components of commercially available SPR instruments are (1) a light source, typically a high-efficiency light-emitting diode, (2) an optical detector such as a diode-array or charge-coupled device camera, (3) a solid support containing the conducting film and some means for attaching molecules, (4) a sample delivery system, frequently a microfluidic device capable of delivering samples using single serial or parallel injections via single or multiple needles, and (5) a computer with appropriate software for instrument control, data collection, and analysis.

Prism-based and diffraction-grating instrument systems are commercially available. Most prism-based systems follow the Kretschmann configuration (*Figure 2*). The light is focused onto the sensor surface (away from the samples) via a prism with a refractive index matching that of the surface. In this configuration the incident light does not penetrate the sample solution, which permits SPR measurements for heterogeneous, turbid, or opaque samples. In systems that utilize a diffraction grating (*Figure 3*) the analyte solution is placed over a plastic surface on which a metal has been deposited. The plastic acts as an attenuated total internal reflection prism in which light reflected from the grating is reflected many times back to the grating surface. In this configuration light passes through the analyte sample solution, and thus turbid or opaque samples are not suitable for measurement. The diffraction grating does permit sampling of a larger surface area and is applicable for SPR measurements of arrays.

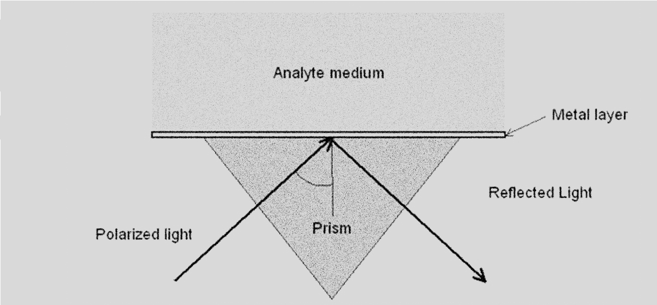

Figure 2. Kretschmann SPR configuration.