BioProcessing JOURNAL
TRENDS & DEVELOPMENTS IN BIOPROCESS TECHNOLOGY

Defining Therapeutic Window for Viral Vectors: A Statistical Framework to Improve Consistency in Assigning Product Dose Values

By Nancy Sajjadi and Janice D. Callahan

Abstract

Pre-clinical and clinical trials conducted to establish the minimum effective dose and the maximum tolerated dose of a viral vector assume that the assigned dose values are comparable across studies. Toxicity has been associated with high dose administration of both adenovirus and adeno-associated virus-based vectors, and increased attention must be paid to assays used to measure dose. High assay variability can be mitigated by replication and the reporting of a mean value for product lot release. The establishment of a dose specification and a testing strategy must take into account the risk of errant quality control decisions. This can be accomplished by linking assay qualification information to measurement uncertainty through a statistical framework. By adopting an equivalence approach, the risk of releasing lots with unacceptably high or low dose values is minimized by reducing measurement uncertainty. This article provides a worked-through example to introduce applicable statistical concepts and the equations necessary to facilitate their implementation in the field.

Introduction

The overarching goals of a viral vector product development program are: (1) to define therapeutic window, the range of doses that exist between the minimum effective dose and the maximum tolerated dose; and (2) to identify the product's critical quality attributes. The recent deaths of young patients enrolled in a gene therapy trial[1] are tragic reminders of the potential risks associated with administration of high dose viral vectors[2], and they highlight the importance in assigning dose values to all types of viral vector products for use in pre-clinical and clinical studies. A call to action in response to these unfortunate events includes a plea to "*calmly sort out what happened and put in place safeguards to prevent this from happening again*" and acknowledges that the field of adeno-associated virus (AAV)-based gene therapy "*does not have a universal standardized assay to titer AAV, thus these (and all AAV doses) should be considered approximate.*"[3] Uncertainty is associated with every form of measurement. Therefore, any test result is an approximation of its true value. Because dose determining assays for viral vectors are more variable than mass measurements used to assign dose for other types of products, they are expected to be appropriately qualified for use as quality control (QC) lot release tests.

According to recently issued FDA guidance[4], investigational new drug applications (INDs) for human gene therapies should include specifications for measuring an appropriate dose level at Phase 1. The guidance states that: "*To ensure consistent dosing in your clinical investigations, assays used to determine dose (e.g., vector genome titer by quantitative polymerase chain reaction [qPCR], transducing units, plaque-forming units, flow cytometry for transduced cells) should be qualified as suitable for use prior to initiating clinical studies.*" Expectations regarding qualification are further clarified in the same document[4]: "*In your original IND submission, you should provide a detailed description of the qualification protocol (e.g., samples; standards; positive/negative controls; reference lots; and controls evaluated, such as operators, reagents, equipment, dates) and data supporting the accuracy, reproducibility, sensitivity, and specificity of the method.*" Although a recombinant AAV reference standard material for serotype 2 (rAAV2 RSM) has been thoroughly characterized and is available for use as a calibrator for in-house reference materials[5,6], there is a paucity of information on how to *specifically* demonstrate the suitability of a dose determining method for use in assigning dose to product that will be released for dose escalation studies. Regardless, appropriate statistical analyses must be performed and used to convey the uncertainty associated with dose value assignments if consistency is to be achieved. Demonstrating assay suitability begins with an assessment of the risks associated with measurement uncertainty.

Measurement Uncertainty

Measurement uncertainty is generally defined in terms of the dispersion of values within which the true value is purported to lie. However, some standard-setting documents[7] ask for "*a range that includes an allowance for both random and systematic errors*." This paper focuses first on establishing statistical hypothesis testing and statistical error as means for improving consistency in label dose assignments, in dealing with random assay error, and then illustrates how to incorporate allowances for offsets introduced by assay bias and/or manufacturing variation. The primary purpose of this paper is to educate non-statisticians and to elevate their discussions with statisticians who will fully understand the approaches described herein for controlling the risks associated with dose measurements. One logical starting point for these important discussions is "dose response studies," which require precise measures of vector dose.

From a statistical standpoint, experiments designed to measure dose responses *in vitro* and *in vivo* assume that dose, the predictor variable (independent variable) is accurate (*i.e.*, fixed with no distribution). Random variation in the response (dependent variable) is expected. However, in reality, dose values are neither error-free nor fixed—they are derived from measurements that are subject to systematic and random error introduced by the assay system that is used to determine product concentration. While it is impossible to meet the assumption of error-free "true values," it is critical to generate dose value *approximations* with quantifiable, high statistical confidence. Choosing to accept test results from a dose-defining assay and applying a label concentration to a product lot are important QC decisions that should be based on formal statistical hypothesis testing. As will be shown, equivalence testing is the correct approach to use in establishing lot release limits and qualifying assays to assign dose values to product lots. Because most bench scientists have a basic understanding of statistical difference testing, we begin with a review of that approach as an introduction to the concepts of "statistical error."

Statistical Hypothesis Testing and Statistical Error

It is common in vector product research and development work to design experiments to look for important differences. For example, to demonstrate the efficacy of a candidate viral vector to induce a desired response in an animal model, studies are designed to compare a vector treated group to a control group. Results from animals receiving a vector treatment are *expected* to give different results than the animals in the control group (placebo) because there is reason to believe (from previous research, literature, *etc.*) that the product will be efficacious. In fact, the primary goal of the animal study is to *prove* (*i.e.*, demonstrate with a high level of confidence) that the anticipated conclusion is true. The corresponding "null hypothesis" in this experiment is the opposite outcome (*i.e.*, that there is no vector treatment effect). However, there will be variability associated with the mean response values obtained for both of the animal groups. To address the question as to whether the mean responses obtained for each group are truly different, this variation in the mean responses must be taken into account.

Specifically, the variation associated with the mean responses (due to a combination of animal-to-animal variation and measurement uncertainty) is used to calculate the probability of obtaining the resultant difference between the groups under the *assumption* that the vector treatment had no effect at all. Stated another way, if the true value of the difference between the treated and untreated groups is zero, the probability of getting the observed difference by random sampling chance can be calculated. Consequently, the bigger the measured difference between the means of the two groups, the lower the probability of it occurring by sampling chance alone. That calculated probability is called a *p-value* and it is *never* zero because no matter how large the measured group difference value is, there is *always* a theoretical possibility that the *true* difference between the groups is zero. Given that this is the case, it follows that the null hypothesis—that the measured differences between groups are due to sampling (or experimental) error and *not* a true treatment effect—can never be rejected and nothing can ever be proven to be 100% true. *So how does statistical hypothesis testing actually work?* To answer this question, the term "statistically significant" must be introduced.

A *statistically significant* difference is derived from an arbitrarily chosen probability cut-off, an α value (commonly 0.05 or 5%) that sets a *critical p-value* ($p = 0.05$ for 5% α), which the scientists choose in advance. If the measured difference between the groups returns a *p-value* that is smaller, it is considered too unlikely to have occurred by chance, and the more reasonable explanation is that the null hypothesis is not true (*i.e.*, that the differences are very likely to be real). Thus, by *rejecting the null hypothesis*, the conclusion is to *accept the alternative hypothesis* and assert, with the selected "confidence level" that the group differences are due to the effect of the vector treatment. This is justified by the *big difference* in the measurements of the two groups; a measured difference that is greater than the *smallest difference to be considered significant* (defined in advance by choosing an α).

Data returning a *p-value* smaller than the *critical p-value* are considered very strong evidence (often called "proof") of the claim that the vector treatment is efficacious. It is important to remember that statistical hypothesis testing is grounded in probability theory and that the null hypothesis is the theoretical reference point scientists use to quantify the confidence they have in their beliefs about what is true. Note that for an α set at 5%, there is a 5% chance that the conclusion is wrong (*i.e.*, the null is rejected when it is actually true), which is a statistical error. Choosing an α quantifies the risks of making an errant conclusion. Because decision-makers in quality and regulatory roles also draw conclusions from data, it is imperative that they understand the risks associated with measurement uncertainty and its implications in trusting product test results.

Assay precision data, the replication strategy used to generate a "reportable value" (RV), and the choice of acceptance limits in assigning dose are all integral to ensuring consistency in dose values. The working assumption for the purposes of this paper is that a QC decision for dose assignment will be based on a RV using "*n*" replicates to calculate a mean value and, if that value is deemed acceptable, the target dose value will be used to assign the label concentration for the lot. As such, qualifying a dose-determining assay and establishing lot release limits are inextricably linked through the concepts of statistical hypothesis testing and statistical error.

Table 1 introduces "Type I" and "Type II" errors. In the animal study described previously, the Type I error is associated with deciding there is a difference when there isn't one (α). The Type II error (β) is just the opposite, concluding that there is not enough evidence to reject the null hypothesis when it is false, and there truly is a difference. Making errant decisions creates risk, which means there is the potential for significant losses or harm.

TABLE 1. Statistical hypothesis testing and decision errors.

		TRUTH	
		Null Hypothesis (H_0)	**Alternative Hypothesis (H_A)**
DECISION	**Do Not Reject Null Hypothesis (H_0)**	Correct	Type II Error
	Reject Null Hypothesis (H_0)	Type I Error	Correct

It is critical to understand that there are *always* statistical hypothesis errors. It is only the definition of the hypotheses that change. There are several different forms of hypothesis testing, and choosing the appropriate statistical test is dependent on the question being answered by the analysis. Difference testing is used when the goal is to demonstrate a difference when making comparisons, whereas statistical equivalence testing is applicable when small but inconsequential differences are expected and the things being compared can be shown to be similar enough to be accepted as the same.[8, 9] Statistical equivalence is the framework that QC release testing for dose assignment operates within. Lot release acceptance limits for product dose must allow for some variation in both the manufacturing process and the testing process. Beyond those limits, the differences in reportable values are defined as "consequential" and a reason to reject a lot of vector. Within those limits, a vector lot is considered "close enough" to be labeled as the target dose.

The hypotheses for both difference tests and equivalence tests for the two applications discussed can be formally written in statistical language as follows:

Difference Test

H_A: Mean for vector treatment group − mean for control group ≠ 0

H_0: Mean for vector treatment group − mean for control group = 0

Equivalence Test

H_A: Target dose − $\delta \le$ mean $\le$ target dose + δ

H_0: Mean < target dose − δ Or mean > target dose + δ

Where H_A is the alternative hypothesis, H_0 is the null hypothesis, and δ is defined as the "equivalence margin" (to be described later).

For equivalence testing, α (Type I error), the probability of rejecting the null hypothesis when it is true, translates into deciding that the vector lot is within the equivalence interval (= target dose ± δ) when the product lot is *not* similar (either too low or too high). Stated another way, α (Type I error) is the probability of assigning the target dose to a lot when it is *not* equivalent. For β (Type II error), this translates to the probability of not rejecting H_0 when it is false, which equates to the probability of failing a lot when it actually *is* equivalent. Each of these errors is the opposite of a correct decision. "Statistical power" (a very important term) is expressed mathematically as $1-\beta$.

In the context of QC decision, power is the probability of releasing an acceptable lot and correctly assigning the target dose. It is obvious then that as Type II error is reduced, power is increased. It is also important to note that there is no statistical term to describe the other correct decision, the probability of rejecting a lot when it is truly not close enough to the target dose. For dose assignment, this probability matters much more than power because if the doses are incorrectly assigned, conclusions from dose response studies will likely not be valid. Although throwing away acceptable lots is a loss to be avoided, it does not affect interpretation of dose escalation studies. To illustrate how these abstract concepts are interrelated and how they can be applied to assay qualification, imaginary data from a fictitious dose-determining assay and viral vector product are used to create a worked-through example.

Linking Assay Qualification Information to QC Lot Release Limits and Statistical Error

The simulated dose assay data in the example are assumed to follow a lognormal distribution. Log_{10}-transformed values are used in describing dose targets and acceptance ranges. Although no back transformation of results is provided in this paper, it is acceptable to appropriately do so for the purposes of returning data to the original units of scale for more intuitive interpretation. Assay precision values used throughout the working example are in $\log_{10}$ units and are considered intermediate precision estimates, for ease of interpretation. Correspondingly, "sample size" (n) refers to the number of independent assay runs. Precision values are considered to be reasonable estimates of the true assay precision, having been derived from well-designed qualification studies and/or historical assay performance metrics. It is beyond the scope of this article to describe how such assay qualification studies should be done and to review the statistical treatment of lognormal data. For comprehensive information on these topics, USP information chapters[10–12] and articles describing their reduction to practice[13,14] serve as excellent primers.

To begin, let's imagine that three dose levels for a given vector are being prepared to achieve dose levels of 1, 3.3, and 10 dose units/mL. After $\log_{10}$ transformation, these three dose target values become 0, 0.5, and 1 $\log_{10}$ dose units/mL, respectively. Note that the vector "dose" is used as shorthand for viral vector concentration in "dose units"/mL, and the two terms are used interchangeably for convenience throughout this paper. **Figure 1** provides a schematic illustrating the *sampling probability distributions* for log-transformed target mean (μ) values at three evenly spaced levels (0.5 $\log_{10}$) showing putative acceptance limits for a reportable value, the widest allowable for target dose 0.5 set at the midpoint between the three targets. In this approach, an average ($\bar{x}$) of n replicates (a reportable value) is then compared to the acceptance range. If $\bar{x}$ is within the limits, the lot is labeled as 0.5, and if not, the lot is rejected.

This seems like reasonable decision-making, except that without specifying a statistical hypothesis framework, statistical error is not taken into account. **Figure 2** includes the same curves as **Figure 1**, but with two additional distributions centered at the two midpoints between the original doses. Imagine that a newly produced lot of vector has a

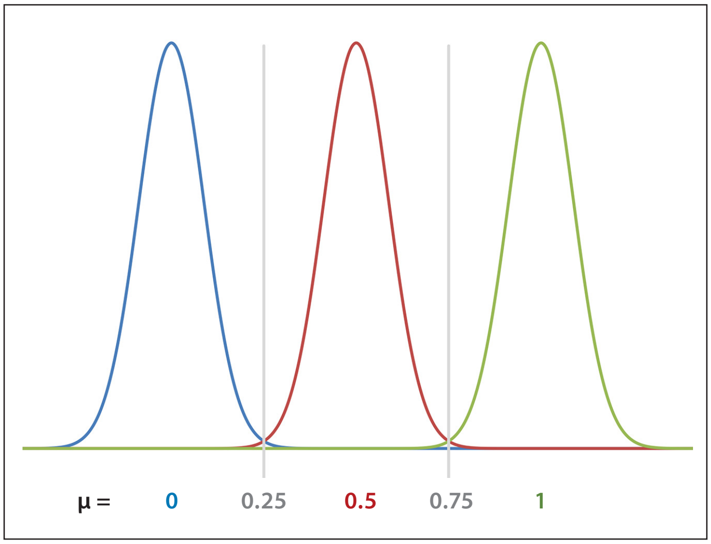

FIGURE 1. Probability distributions for mean values at three target vector doses. The target values are indicated as the mean (μ) of a normal sampling distribution. The mean is at the center of the normal distribution. The vertical bars represent example lower and upper acceptance limits around the target dose of 0.5 for $\bar{x}$. These bounds also represent the upper and lower acceptance limits around the target doses of 0 and 1, respectively. The other bounds for the 0 and 1 doses have not been included to allow for a focus on the example 0.5 curve.

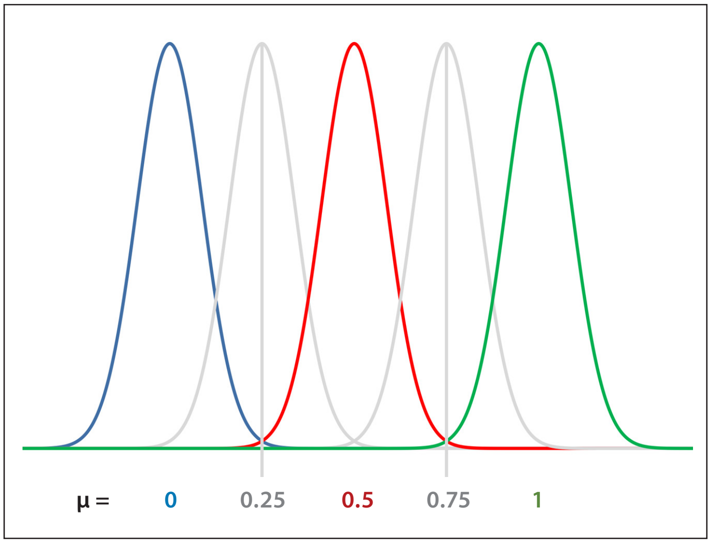

FIGURE 2. Probability distributions of target doses and the midpoints between them. All five distributions overlap to varying degrees. By centering the distributions at the midpoint between target doses (0.25 lower, 0.75 upper), it becomes obvious that by sampling error alone, a vector product lot, with a true value midway between the target doses, has an equal chance of returning a mean value that falls within the lower and higher flanking doses, if the midpoints (gray vertical lines) are chosen as the acceptance limits.

dose value that is actually at one of those midpoints. It should be obvious that half of the time, the QC test results for either of these lots would fall into the acceptance limits set for the 0.5 dose target. In using this approach, QC has inadvertently allowed for 50% error! *How can that problem be fixed?* The answer is tied to linking assay precision with the concept of "standard error."

Standard error is a key component of communicating measurement uncertainty associated with a reported value, and because it is a function of the assay precision, it becomes one of the critical assay performance parameters in assay qualification. In both **Figures 1** and **2**, the width of the dose curves is described by the standard deviation of the sampling distribution of a mean. This statistic is also called the "standard error of the mean" (SEM), which is calculated as follows:

Equation <1>

$$\mathbf{SEM} = \frac{\boldsymbol{\sigma}}{\sqrt{\mathbf{n}}}$$

Where: σ = standard deviation; and n = sample size.

In our example, σ is the data dispersion, as measured by assay intermediate precision, and n is the number of independent assay results that will be averaged to determine the reportable value, which is a mean ($\bar{x}$). Commonly, routine QC tests are considered independent if the assays are run on different days or by different analysts using separate vials of product for each test. To reiterate, using this approach, a random sample of size (n) is being taken and $\bar{x}$ is being calculated. In the initial strategy, a vector lot with target dose 0.5 undergoing QC lot release testing will pass if $0.25 < \bar{x} < 0.75$ and the *decision* that the lot is equivalent to 0.5 means that it will be labeled as such. Note, by referring back to **Figure 2**, the probability of this occurring increases as the curves get narrower, which is to say that the SEM gets smaller. The other thing that happens is that the statistical errors become less likely to occur. This is a win-win situation for QC and a goal of assay optimization and qualification! We will return to this again, but for now let's turn our attention to a graphical display of statistical errors.

Consider again the testing of a vector lot produced to achieve the middle dose. **Figure 3** isolates the middle three curves from **Figure 2** and includes two vertical gray bars that are provided as markers to visualize the effects of changing α and β values. The vertical bars have been placed intentionally at the two halfway points between the 0.5 target dose and the midpoints (0.25 and 0.75) such that the probabilities of Type I and Type II errors are equal. The Type I error, α, has been shaded gray under the middle red curve and the Type II errors, $\beta/2$, have been shaded pink under both of the gray curves. If the vertical gray bars were to be moved to the center of the middle dose, then α is zero and each $\beta/2$ is 50%. In contrast, moving the vertical bars away from the middle target dose increases α and decreases $\beta/2$. As they reach the centers of the two surrounding distributions at 0.25 and 0.75, α approaches 50% and $\beta/2$ approaches 0%. It is now also easy to see that if the vertical bars remain in place but the curves get narrower, SEM decreases and both types of errors become smaller.

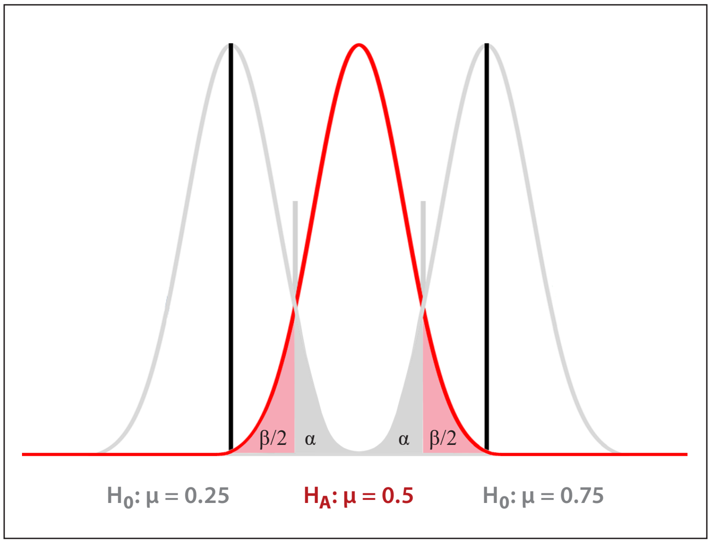

FIGURE 3. Target dose at 0.5 with two surrounding null hypothesis tests. The center curve represents the distribution of the target dose for 0.5. The two flanking curves correspond to the lower and upper midpoints using 0.5 log intervals for target doses. The shaded areas labeled α and $\beta/2$ represent the errors. The vertical gray bars are provided as markers to visualize the effects of changing α and $\beta/2$. The width of the curves reflects the same constant SEM as shown in **Figures 1** and **2**.

Figure 3 is also a visual display of equivalence testing which uses "two one-sided tests" (TOST) to reject a null and accept an alternative, at the α and β levels shown. The means of the outlying gray curves can now be labeled as null hypotheses and the red curve mean labeled as the alternative hypothesis. The vertical bars are drawn at the critical values for rejecting the two null hypotheses. The distance between the target dose and each null hypothesis is called the equivalence margin (δ). The shaded areas under the curves visually represent the probabilities for Type I errors (α, gray shading) and Type II errors ($\beta/2$, red shading), respectively. Note that in **Figure 3**, there is only one α in the name but two shaded areas under H_A distribution and there are two $\beta/2$ values. Because only one of the two null hypotheses can be falsely rejected within one test, then **there is only one $\boldsymbol{\alpha}$ that can occur**. But if H_A is incorrectly accepted, both $\beta/2$ areas have occurred.

Proving that a vector lot is equivalent to a target dose of 0.5 in this example can only be concluded by rejecting both null hypotheses. Thus, in our example using a 0.5 target dose with δ set to 0.25, the first null hypothesis is written as H_0: $\mu \leq 0.25$ with the alternative hypothesis being H_A: $\mu > 0.25$. The second null hypothesis is H_0: $\mu \geq 0.75$ with the alternative hypothesis H_A: $\mu < 0.75$. If either one of these null hypotheses is not rejected, then the dose is unacceptable. Simultaneously rejecting both null hypotheses proves that $0.25 < \mu < 0.75$, thus proving that the dose is equivalent to 0.5.

To control the error in dose assignment, the equivalence interval ($2 * \delta$) must be established and the α and β values both assigned before data are collected. In our example, the equivalence bounds (= null hypotheses) are set as the largest possible (*i.e.*, the midpoints between the lower and upper dose values flanking 0.5). Setting α and $\beta/2$ as equal is convenient and practical for the calculations that follow in our worked-through example (**Figure 3**), and is the recommended starting point for those without access to an experienced statistician. For QC lot release assays used to measure dose, the main driver in the decision-making for dose assignment should be in establishing an appropriately low α value. In other words, this should be the first step taken before deciding on acceptance limits and the suitability of measurement uncertainty associated with a reported value. This is because it is very important to prevent a lot that is *not at target dose* from being *incorrectly assigned as the target dose.*

To recap what was discussed earlier, if that mistake (Type I error) occurs, then any dose escalation studies using such a vector lot would be compromised. In contrast, by making a mistake in the other direction (Type II error), the consequence is throwing away (or repurposing) good lots. This is costly to the manufacturer ("the manufacturer's risk") but it does not introduce any error into the dose response analysis.

To conduct a hypothesis test for a reportable value, a *z*-value must be calculated. The *z*-value is called a "test statistic." It is a way of expressing differences between the measured and target values in multiples of the SEM, as shown in **Equation <2>**. Note that a *z*-value increases when the reportable $\bar{x}$ value moves away from the target and/or the SEM decreases:

Equation <2>

$$z = \frac{\bar{x} - (\mu \pm \delta)}{\text{SEM}}$$

Where: $\bar{x}$ = the reportable value; μ = the target dose; and δ = the established equivalence margin. SEM is derived from assay precision (σ) and the number of measurements averaged (n) per **Equation <1>**. δ is the established equivalence margin.

The acceptance limits for lot release based on a reportable value are derived from this test statistic and developed stepwise using the 0.5 dose example. They are expressed mathematically from **Equation <2>** as follows. For every target dose at 0.5 log intervals, there are two midpoints, one below and one above. Consider the lower midpoint and perform a hypothesis test of H_0: $\mu \leq 0.25$ versus H_A: $\mu > 0.5$. Reject this H_0 if:

$$\text{Lower: } z = \frac{\bar{x} - (0.5 - \delta)}{\text{SEM}} > z_\alpha \;\text{ Or }\; \bar{x} > (0.5 - \delta) + \text{SEM} * z_\alpha$$

Where z_α is a tabled *z*-value based on the chosen α.

A little arithmetic then shows that the other H_0 can be rejected if:

$$\text{Upper: } z = \frac{\bar{x} - (0.5 + \delta)}{\text{SEM}} < z_\alpha \;\text{ Or }\; \bar{x} > (0.5 + \delta) - \text{SEM} * z_\alpha$$

Taken together, the hypothesis testing can be expressed as TOST (a statistical eqivalence test) as shown:

Equation <3>

$$(0.5 - \delta) + \text{SEM} * z_\alpha < \bar{x} < (0.5 + \delta) - \text{SEM} * z_\alpha$$

By rearranging **Equation <3>** algebraically and continuing with the working example ($\delta = 0.25$), a target dose is proven to be equivalent to 0.5 if:

$$0.25 < \bar{x} \pm z_\alpha * \text{SEM} < 0.75$$

The 0.25 and 0.75 values are still the original acceptance limits from **Figure 1**, but by framing lot acceptance as an equivalence test, we have taken "error" into account. Instead of comparing the reported $\bar{x}$ value directly to the acceptance limits, we now require that $\bar{x} \pm z_\alpha * \text{SEM}$ (called a "confidence interval" [CI] on a mean) be narrow enough to fit within the equivalence limits. However, it is very important to understand that the standard deviation (σ) is the assay intermediate precision estimate, *not* the sample standard deviation for the reportable value. A narrow CI can be achieved in two ways with: (1) a small z_α value; and/or (2) a small SEM.

A tabled z_α is dictated by the chosen α, and the SEM is dependent on assay precision (σ) and sample size (n) (**Equation <1>**). As such, increasing sample size decreases both values simultaneously. Again, this increases power and is the statistical reward for doing more work. Also, note that the $1 - 2 * \alpha$ value is called the confidence level. The higher this value, the higher the confidence (*i.e.*, the lower the chance of making a Type I error) and the larger the corresponding z_α value. This means that the higher the confidence desired, the greater the demand on the assay performance to yield a small SEM for a CI that falls within established limits.

Remember the assumption that dose is error-free? That translates to 100% confidence, which is impossible, so the issue becomes deciding what level of confidence is good enough. It is also important to note that the widest acceptable CI occurs when the true value of the lot is at the target dose (center of the equivalence bounds). As the true value shifts to the left or the right of the target center (or the dose-determining assay introduces bias into the results), the associated CI must become smaller to fit within the bounds. Introducing an offset to account for this will be applied later, but for now, what should be very clear is that the SEM is central to hypothesis testing and statistical error.

The probabilities of all four decision outcomes (listed in **Table 1**) are dependent on SEM, which in turn is dependent on assay precision and reportable value measurement replication. The suitability of a dose-defining assay is therefore tied to the reportable value (assumed to be a mean) and the SEM associated with that value. By establishing risk tolerances for error (setting the equivalence bounds and choosing acceptable α and β values), a specific SEM can become the minimum target in dose-defining assay qualification. Therefore, the critical decisions involved in assay development become the ones made to appropriately allocate resources, in accordance with timelines and other practical constraints, to achieve a minimally acceptable SEM. To reduce the SEM, assays should be optimized to improve assay precision (reduce σ). However, the effects of high assay variability can be mitigated by increasing replication (running more assays and combining more values to generate the RV).

Reportable Value Sample Size Calculations

Using an equivalence-based QC decision making framework, the n required to achieve the assigned power for a given assay precision can be determined using a sample-size calculation:

Equation <4>

$$n = \left(\frac{(z_{\alpha} + z_{\beta/2}) \times \sigma}{\delta} \right)^2$$

Where: z_{α} and $z_{\beta/2}$ are tabled normal z values; σ = assay precision; and δ = equivalence margin.

To solve for n using **Equation <4>**, the value for σ must be known and an effect size selected. The value for σ is the intermediate precision, as measured during the assay qualification with an $n \sim 30$.

Using a simulated assay precision value of 0.15 and the equivalence margins set at the midpoints (0.25 – 0.75; target 0.5 with $\delta = 0.25$), the minimum number of replicates required to generate a mean reportable value to meet an 80% power target with $\alpha = 10\%$ would be $n = 3$.

For those cases in which the intermediate precision has been estimated with an $n < 30$, both **Equations <3>** and **<4>** must be changed with the z_{α} becoming $t_{\alpha,n\text{-}1}$. **Equation <3>**, for the equivalence CI, then becomes:

Equation <5>

$$(0.5 - \delta) + \mathrm{SEM} * t_{\alpha,n-1} < \bar{x} < (0.5 + \delta) - \mathrm{SEM} * t_{\alpha,n-1}$$

Correspondingly, **Equation <4>** becomes **Equation <6>**, which requires iteration:

Equation <6>

$$n = \left(\frac{(t_{\alpha,n-1} + t_{\beta/2,n-1}) \times s}{\delta} \right)^2$$

Where: $\alpha = \beta/2$; $t_{\alpha,n\text{-}1}$ = tabled t-value with n-1 degrees of freedom; s = assay precision; and δ = equivalence margin.

In **Equations <5>** and **<6>**, the z-values are replaced by t-values and the σ is replaced with s. Using the n from **Equation <3>** as a starting point, the $t_{\alpha,n\text{-}1}$ is used to calculate an updated n. This process must be repeated until the updated n equals the previous n. For the example above that returned an $n = 3$ using **Equation <3>**, the final result is $n = 6$ and $t_{\alpha/2,\,n\text{-}1} = 2.015$.

Note that because of the symmetry in our example, as evidenced in **Figure 3** by the placement of the vertical gray bars at the halfway points, for the sample size calculation, $\alpha = \beta/2$. This means that by selecting an α, β becomes $2 * \alpha$ and power = (1-2 $* \beta/2$). Applied to lot release, power translates to the probability of assigning the target dose to a lot when it is equivalent. The other correct decision, the probability of failing a lot when it actually is not equivalent, becomes 1-α. Note that the probability of failing a lot when it is not equivalent is a probability at the hypothesis error level and is thus 1-α, whereas the confidence level for constructing the confidence interval is 1-2*α around the reported value.[15]

For an assay to be considered qualified, the minimum sample size required for generating a reportable value must be practical. **Table 2** provides some example sample sizes for

TABLE 2. The minimum number of measurements (n; sample size) required to control statistical errors as a function of assay intermediate precision.

Precision (σ)	0.200	0.150	0.100	0.075	0.050	α
	0.400	0.300	0.200	0.150	0.100	β
0.05	1	1	1	1	1	Sample Size Values
0.1	1	1	2	2	2	
0.15	2	2	3	3	4	
0.2	2	3	5	6	7	
0.25	3	5	7	9	11	
0.3	5	7	10	12	16	

NOTE: All non-integer values are rounded up.

select α, β, and intermediate precision (σ) with δ set at 0.25. If the sample size required for the initial parameters selected is too large for routine testing, we recommend focusing efforts to improve assay precision (lowering σ) to reduce the sample size to the practical limit. This decision is costly, but it is the price to be paid for maintaining appropriately low α and β values.

Incorporating an Offset to Account for Allowable Assay Bias and Process Variability

Notice that **Equations <4>** through **<6>** assume that $\bar{x}$, the reportable value, will be exactly the target value. This is highly unlikely. The sample size equation can be modified to incorporate an offset in the $\bar{x}$ measurement which will result in a CI that is not centered at the target dose but is still within the equivalence bounds. The offset can be used to account for expected differences in $\bar{x}$ introduced by small shifts in true dose values and/or assay bias. Calculating sample size with an offset included must be performed using:

Equation <7>

$$n = \left(\frac{(z_{\alpha} + z_{\beta/2}) \times \sigma}{\delta - \text{Offset}}\right)^2$$

Where: $\alpha = \beta/2$; z_α and $z_{\beta/2}$ = tabled z-values; σ = assay precision; δ = equivalence margin; offset = amount $\bar{x}$ can acceptably deviate from the target dose and still have a CI within the equivalence margins.

Establishing a Lot Release Specification for a Target Dose

The choice of equivalence bounds and α and β values drive the requirements for assay performance. For a given assay precision, sample size is used to achieve the desired SEM. An offset must be introduced to allow for deviations of the reportable value from the target dose. Sample size is determined after the allowance for an offset is selected and the acceptance limits on the reportable value become the target ± the offset value. Recall that we began the discussion of lot release acceptance by imagining a scenario that did not use a statistical framework.

In that first scenario, the reportable value was considered acceptable if it fell within the midpoint values. Initially, there was no consideration given to assay precision, replication strategy, or error. Next it was shown that by using a statistical equivalence approach, a reportable value of 0.5 could be considered acceptable if the associated CI fell within the original bounds. However, in calculating a minimal sample size necessary to achieve this, it was necessary to assume that the reported value was actually at the target dose. Lastly, an offset was introduced to allow for reportable values other than 0.5. By applying the offset to the target dose (0.5 ± offset value), an acceptance range for the reportable value is established. To achieve this, the sample size was suitably increased to ensure that mean values very close to the reportable value lot release limits will also have CIs that fit within the established equivalence interval (0.25–0.75 in the worked-through example).

The relationship between σ and n and the SEM, as provided in **Equation <1>**, means that higher assay variation also necessitates higher replication to achieve the same statistical power. It must be reiterated that σ (or s) is the assay precision estimate from assay qualification studies, not the sample standard deviation associated with an individual reportable value. Furthermore, the sample size equation works for all target doses only if the standard deviation (assay precision) is the same for all doses. Assays with a constant "coefficient of variation" (CV) have increasing standard deviation with increasing mean. Typically, log transformation of dose determining assay values makes the standard deviations the same for all means. As indicated at the outset, statistical analyses assume constant variance and normality. Unless data can be shown to follow a normal distribution, a lognormal distribution should be assumed for dose determining assays, and statistical analyses should *always* be performed on log-transformed data.

Lastly, in some lot release applications (*e.g.*, primary and working reference standard dose assignment), it may be important to reduce δ. Because the equivalence margin is in the denominator in **Equation <4>**, note that a reduction in δ from 0.25 to 0.125 would lead to a quadrupling of the unrounded sample size values in **Table 2**.

Conclusion

The assignment of label dose is a critical part of establishing the safety and efficacy of viral vectors. The suitability of a dose-determining assay is dependent on assay precision, the number of measurements that will be made in calculating a reportable value, and the lot release limits established. Using log-transformed data, a lot release acceptance range surrounding a target dose is appropriately symmetrical. By setting criteria using statistical power to exclude a putative product lot with a specified true value (*e.g.*, midpoints flanking the 0.5 target dose), the confidence and consistency in assigning a target dose is controlled by the SEM associated with a reported value. The use of an equivalence approach has been invoked. The lot release acceptance limits for a reported value can be derived from the risk tolerance for error and allowable offsets, which are then directly tied to assay qualification.

To recap, risk is controlled by selecting appropriate α (Type I error), β (Type II error), and δ values. To qualify a dose-determining assay, the intermediate precision estimate (σ or s) can be evaluated by using it to calculate the corresponding minimal n needed to achieve the resultant statistical power such that a reported value and its CI will fit within the equivalence limits. **Table 2** provides *minimum sample sizes* calculated for a range of assay precision values and different combinations of α and β values with δ set at 0.25. The calculations in **Table 2** were performed using **Equation <4>**. If an offset were to be included, **Equation <6>** would need to be applied. In this case, the values in **Table 2** would increase.

An assay is then considered qualified if the SEM necessary can be achieved with the precision and replication strategy that meets the limits imposed by the risk tolerance. The QC lot release acceptance limits for $\bar{x}$, the reportable value, are derived by adding and subtracting an allowable offset value to the target dose. Vector lots with reportable values in this range can be considered equivalent to the target dose and can receive the target dose label assignment. For example, using an offset of 0.125, the specification for a dose label assignment of 0.5, developed with equivalence bounds at 0.25 and 0.75, would be that the reported value must fall within 0.375 and 0.625. To achieve 90% power (α is set to 0.05), the minimum sample size for an assay with a precision of 0.1 would be 16. This is quadruple the n required if $\bar{x}$ were to be centered at the target of 0.5 ($n = 4$).

The equivalence bounds and the allowable offsets selected are somewhat arbitrary, but they tie directly to the issue this paper addresses. There is an implicit assumption made in any dose response evaluation that the dose is a fixed value. In practice, that means that dose values are assumed to be consistent across studies. Equivalence limits give meaning to what are understood to be dose "approximations." The narrower the equivalence limits (*i.e.*, smaller δ), the more consistent the dose values will be, and the greater the demands placed on the assay performance to avoid statistical errors. We believe that widespread adoption of the framework proposed in this paper could lead to greater consistency in dose estimates for defining the therapeutic window of viral vector products.

References

[1] Joshua Frase Foundation. RECENSUS / INCEPTUS / ASPIRO Studies. https://www.joshuafrase.org/get-involved/recensus-study.php

[2] Hollon T. Researchers and regulators reflect on first gene therapy death. *Nature Medicine*, 2000: 6, 6. https://doi.org/10.1038/71545

[3] Paulk N. Gene therapy: It's time to talk about high-dose AAV. *Genet Eng Biotechnol News*, 2020; 40(9). https://www.genengnews.com/commentary/gene-therapy-its-time-to-talk-about-high-dose-aav/

[4] FDA (CBER) — Guidance for industry: *Chemistry, manufacturing, and control (CMC) information for human gene therapy investigational new drug applications (INDs)*. Jan 2020. https://www.fda.gov/media/113760/download

[5] Lock M *et al*. Characterization of a recombinant adeno-associated virus type 2 reference standard material. *Human Gene Therapy*, 2010; 21(10): 1273–85. https://doi.org/10.1089/hum.2009.223

[6] Werling NJ, Satkunanathan S, Thorpe R, Zhao Y. Systematic comparison and validation of quantitative real-time PCR methods for the quantitation of adeno-associated viral products. *Human Gene Therapy Methods*, 2015; 26(3): 82–92. http://doi.org/10.1089/hgtb.2015.013

[7] Produced jointly with EUROLAB, Nordtest and the UK RSC Analytical Methods Committee. *Measurement uncertainty arising from sampling: A guide to methods and approaches*. First edition 2007. https://www.scribd.com/document/328410311/EURACHEM1-tcm18-102815-pdf

[8] Lakens D. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Soc Psychol Personal Sci*, 2017; 8(4): 355–62. https://doi.org/10.1177/1948550617697177

[9] Ialongo C. The logic of equivalence testing and its use in laboratory medicine. *Biochemia Medica,* 2017; 27(1): 5–13. https://doi.org/10.11613/BM.2017.001

[10] United States Pharmacopeia (USP) and the National Formulary (NF). USP43-NF38: <1033> *Biological assay validation*. Feb 2020. https://www.uspnf.com

[11] United States Pharmacopeia (USP) and the National Formulary (NF). USP43-NF38: <1010> *Analytical data – interpretation and treatment*. Feb 2020. https://www.uspnf.com

[12] United States Pharmacopeia (USP) and the National Formulary (NF). USP43-NF38 <1210> *Statistical tools for procedure validation*. Feb 2020. https://www.uspnf.com

[13] Bower KM. Statistical assessments of bioassay validation acceptance criteria. *BioProcess Int*, 2018; 16(5): 42–4.

[14] Tan CY. RSD and other variability measures of the lognormal distribution. *Pharmacopeial Forum*, 2005; 31(2): 653–5.

[15] Harris AH, Fernandes-Taylor S, Giori N. "Not statistically different" does not necessarily mean "the same": the important but underappreciated distinction between difference and equivalence studies. *J Bone Jt Surg*, 2012 Mar 7;94(5):e29. https://doi.org/10.2106/JBJS.K.00568

About the Authors

Nancy Sajjadi, MSc*, Independent Quality Consultant, Biopharmaceuticals, Fairfax, Virginia USA

Janice D. Callahan, PhD, Consulting Statistician, Callahan Associates Inc., La Jolla, California USA

Corresponding Author:

Email: sajjadiconsulting@gmail.com | LinkedIn: https://www.linkedin.com/in/nancy-sajjadi-702a4b9/