BioProcessing JOURNAL
TRENDS & DEVELOPMENTS IN BIOPROCESS TECHNOLOGY

Practical Considerations in Using an Equivalence Approach to Establish Lot Release Limits for Vector Dose

By Nancy Sajjadi and Janice Callahan

Abstract

To demonstrate that a dose-determining assay is fit for purpose, the measurement uncertainty associated with a reported release test result must be suitably small. The establishment of a corresponding product specification is inextricably linked to the tolerance for error in assigning a dose value for a vector lot. By adopting an equivalence-based lot release model which includes a total error approach to assay qualification, specific testing strategies can be evaluated quantitatively for dose error and lot release decision risks throughout the drug development process. This article aims to reinforce how the concepts tied to an equivalence-based lot release model are interrelated and applied in practice. It provides in-depth explanations of fundamental concepts and clarifies common misunderstandings for quality control, quality assurance, and regulatory affairs personnel held accountable for decisions made in vector dose assignment and product lot release.

Introduction

The approval of several gene therapy products and gene-modified cell therapies over the last six years[1] has led to increasing numbers of investigational new drug applications (INDs) using viral vectors. However, these successes have been tempered by the risks of dose-related toxicities including 11 patient deaths attributed to adeno-associated viral vectors being evaluated in clinical trials for multiple disease indications.[2-4] The therapeutic window for a product is derived from pre-clinical and clinical dose-response models, which assume statistically that measurements of dose are exact. Whether vector is administered directly or used as a critical raw material to prepare a gene-modified cellular product, the assignment of a label concentration to a vector batch is critical for establishing consistency of product used in preclinical and clinical development.

Measurement uncertainty (MU) and volume delivered contribute to potential dose error. To avoid confounding dose with volume, vector dose escalation for *in vivo* applications is often done by preparation of vector at target dose levels corresponding to the escalation increments (*e.g.*, half-log) and administering the same volume of product. We have previously published a paper advocating for an equivalence approach to lot release for viral vector dose in that context.[5] Implementation of this statistical technique has raised awareness of the importance of minimizing dose error and has been a welcome tool to address the practical challenges of dealing with MU. However, it has also highlighted several important issues where further clarification is needed.

This article is organized to provide information that allows the reader to respond to the questions listed below. The discussions in this paper will illustrate that to achieve a high lot acceptance rate and make minimal dose volume adjustments while maintaining a low tolerance for dose error, both low process variability and low MU are required.

1. How can the total error (TE) approach to assay validation be applied when using an equivalence-based lot-release model?
2. What is the "order of operations" when it comes to establishing targets, equivalence bounds, offsets, Type I and Type II risk tolerance and assay precision qualification?
3. What does a high coefficient of determination (R^2-value) for a dose-response model curve fit mean, in terms of dose error?
4. When using a forecasted confidence interval (CI) to establish lot release acceptance, how should the sample standard deviation be evaluated in each assay run?
5. How is the propagation of error caused by testing bulk drug substance and final container accounted for in setting the offset for final container?
6. What is the impact on dose volume adjustments when the CIs on the dose measurements for any two batches don't overlap?

Beginning with the End in Mind

Our original article[5] details the statistical concepts, terminology, and equations required to implement an equivalence model for lot release and vector dose label assignment, and should serve as a companion guide to this paper. For convenience, the abbreviations and some equations are included again. We used the same subheadings in the original order to organize information to address the questions raised and added sidebars where more in-depth explanations were warranted. Our expectation is that this integrated summary will provide professionals in the field sufficient comprehension of the subjects covered to independently answer the six questions that catalyzed and framed this article. Furthermore, in conjunction with the prior paper, those responsible for preparation and review of regulatory submissions will be equipped to justify adopting and accepting an equivalence model.

Measurement Uncertainty

Assignment of vector concentration is accomplished using the results from a dose-determining test method. Release of a vector product lot for use in pre-clinical studies or clinical trials requires that this test method is fit for purpose.[6] Definition and demonstration of assay suitability is inextricably linked to a quality control (QC) lot release strategy.[7] A test result that is compared to a specification and recorded on the certificate of analysis for a vector lot is called the reportable value (RV) and ideally is a mean ($\bar{x}$) of several independent replicate measurements (*i.e.*, combined results from multiple assay runs). RVs are considered approximations of the true value of the material being tested. From a statistical standpoint, the true value of a vector lot dose is a theoretical concept—the mean value that would be derived from infinite measures of a vector product lot. The infinite measures then constitute a population from which a lot-release test comprising a finite number of measurements (n) would be considered a sample. The concept of assay precision comes from the variability that would be associated with the distribution of infinite measures. As detailed in our original article, these measurement distributions are assumed to be Gaussian (normal). Assay validation, when done well, is designed to get a good estimate of σ, the run-to-run precision. Having a reasonable estimate is important because confidence in a reported result is dependent on the true variability of the method (σ; standard deviation; assay precision) and the number of values used to generate the average (n; sample size = number of replicate measurements). A reasonable approximation of the true assay precision can be achieved from monitoring assay performance over time (*e.g., n*>15 assay runs capturing a stable homogeneous control sample over a period of at least one year) and generating data from prospective experiments that vary conditions such as reagent lots, analyst, equipment, *etc.* that will contribute to assay variability over the long term.

The standard error of the mean (SEM; $\sigma/\sqrt{n}$) can be used to quantify MU by calculating a CI on the sample mean value. The CI describes the dispersion of values within which the true mean value is likely to lie for a chosen level of confidence. The importance of evaluating MU (*i.e.*, σ for the sampling distribution of the mean RVs) rather than assay precision itself is described in detail elsewhere.[8] Using an equivalence approach to lot release for dose, it is the RV and its associated forecasted CI that are used to determine whether a dose test result demonstrates the equivalence of a vector lot to its intended target. The sample standard deviation (s; derived from n replicates) is used as a sample suitability criterion, as described in detail later.

The CI half-width (CI HW) is calculated using α, or Type I error probability, to derive the appropriate t-value and the CI on a mean is derived from the CI HW calculation shown in **Equation <1>** and adding and subtracting that value from the sample mean, as shown in **Equation <2>**.

Equation <1>: $\text{CI HW} = t_{\alpha,df} * \text{SEM}$

where $t_{\alpha,df}$ is the appropriate t-value derived from the chosen level of confidence (α) and the sample size (n)

Equation <2>: $\text{CI} = \bar{x} \pm \text{CI HW}$

However, the meaning of MU may go beyond the random assay variation, σ, captured by the SEM term in the CI HW equation and can include an estimate of systematic errors that lead to a bias in the RV. The most simplistic approach to expanding MU is to add an estimate of absolute assay bias to the CI half-width (HW) thus widening it and further reducing the reliability of the RV.[9] Expanded MU performed this way creates a single calculation to incorporate both sources of measurement error and can be expressed as **Equation <3>:**

Equation <3>: $\text{Expanded CI} = \bar{x} \pm (\text{CI HW} + |\text{assay bias}|)$

The USP also describes the utility of evaluating the combined impact of precision and accuracy in evaluating test method performance[10] and incorporates the concept referred to as “total error,” “total analytical error (TAE),” and “total allowable analytical error” into the analytical procedure lifecycle process.[7] The existing assay validation guidance has emphasized measuring and reporting both accuracy and precision[11, 12], but the limits established do not consider their combined effect on MU. Note that the term “TE” is also defined in current bioanalytical method validation (BMV) guidance[13, 14] but it is not calculated in the same way as **Equation <3>**; TE in the BMV is derived from adding the assay % coefficient of variation (CV) to the % relative error (RE) which returns a different result than **Equation <3>**. More recently, TAE has also been introduced into the proposed revision to the ICH guidance on analytical method validation[15] and is defined as follows:

> *Total analytical error (TAE) represents the overall error in a test result that is attributed to imprecision and inaccuracy. TAE is the combination of both systematic error of the procedure and random measurement error.*

This definition of TAE is consistent with the current USP definition.[7] Although an equation is not included in the definition of either source, the intention is clearly to have a single limit that considers the combined effect of measurement inaccuracy and imprecision. It is important to note that although we did not directly address or use the terms "TE" or "TAE" in our original article, we intentionally achieved a TAE approach in an alternative manner.

Rather than add the assay bias estimate to the CI associated with the RV, as could be done by following **Equation** <3>, we have elected to add assay bias to the allowable difference of an RV from the center target dose value. Conceptually, we believe it is more intuitive to think about how the distance of the product true value from the target value (contributed by drift in the manufacturing process) and the distance of the measured value from the true value (contributed by the skewing of the assay results from the true product lot value) add together to create the RV acceptance limits. Then by setting limits on the RV and establishing equivalence bounds, an acceptance limit on the MU of the RV can be derived by calculating the corresponding maximum CI that would be allowable to demonstrate equivalence of an RV at the limits of lot acceptance. Recall that the CI is a function of both σ and n, which means a forecasted CI can always be made smaller by increasing independent replication (larger sample size, n-runs). **Figure 1** provides a schematic showing these relationships between equivalence bounds, RV release limits, and CI length, and drawing a conclusion of equivalence to target dose.

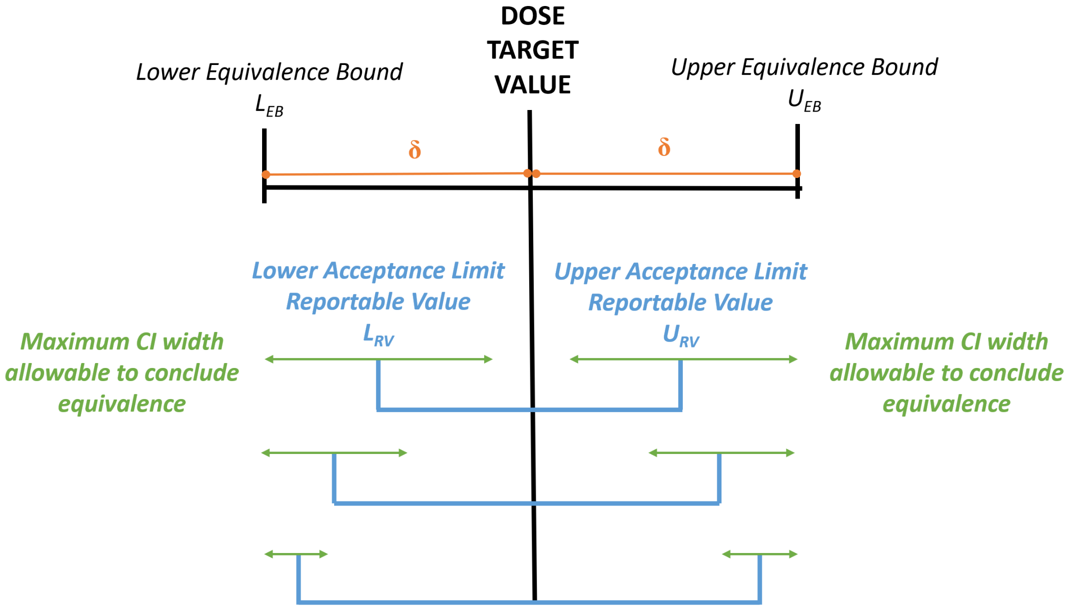

DOSE TARGET VALUE
Lower Equivalence Bound
L_{EB}
Upper Equivalence Bound
U_{EB}
δ
δ
Lower Acceptance Limit Reportable Value
L_{RV}
Upper Acceptance Limit Reportable Value
U_{RV}
Maximum CI width allowable to conclude equivalence
Maximum CI width allowable to conclude equivalence

FIGURE 1: Relating equivalence bounds, RV release limits, and CI length to concluding equivalence to target dose.

To assign the target dose to a vector lot, the reported value and the forecasted CI must fall within the established equivalence bounds. As shown schematically, widening of the acceptance limits on a RV to allow for drift in the manufacturing process means that there must be a corresponding narrowing of the associated CI to conclude equivalence. The heavy blue lines represent the acceptable range for an RV which is calculated as target dose value ± [|assay bias|+|maximum allowable drift from target|]. We refer to |assay bias|+|maximum allowable drift from target| as an offset value. To calculate the maximum CI HW that corresponds to the RV at its limits, this offset is subtracted from δ (the equivalence HW value shown in orange). From the CI HW, the minimum number of replicates is determined per **Equation** <7> in the original paper. The CIs are the green horizontal lines with arrows at the ends. These are displayed only at the lower and upper limits for the RVs because that's where the requirement of narrowing them becomes more obvious. The RV is considered equivalent if its CI falls within the equivalence bounds. Recall that reducing error in dose is achieved by minimizing the equivalence bounds.

By creating what we originally termed as an "offset," the further the true dose value for the lot drifts from center, and greater the bias introduced into the measurement by the assay, the closer the RV gets to the equivalence bounds and thus, the smaller the associated CI must be to prove equivalence of an RV at its acceptance limits. This is discussed in more detail later in sections **Incorporating Offset to Account for Allowable Assay Bias** and **Process Variability**.

Statistical Hypothesis Testing and Statistical Error

To briefly recap, our approach uses a formal statistical hypothesis test to establish whether the reported result from the QC release assay for vector concentration demonstrates equivalence of the product lot to the specification target for dose. Using statistical equivalence as a model for vector dose assignment during pre-clinical and clinical development can provide more meaningful relationships between dose administered and the *in vivo* outcomes for safety and efficacy. More accurate and consistent dose values translate to better data used to define the therapeutic window; these are the data ultimately used to justify the final specifications when the product reaches licensure. Our expectation is that lot release for commercial products would also be conducted using an equivalence strategy.

This is because the expectations of regulators and patients is that the measured vector product concentration will be equivalent to the label concentration as it is the basis of the dose volume calculation. It is universally understood that not all manufacturing runs will produce vector that is exactly at the target value; in fact, very few, if any, runs are likely to achieve this. The whole purpose of the lot release specification and the associated statistical model used is to establish limits for the fluctuation in dose values that can be reasonably tolerated without making dose volume adjustments. The end-goal is to ensure consistency in the manufacture and testing of product and provide regulators and sponsors the necessary assurances that vector batches released for distribution will perform within the agreed upon margins for safety and efficacy.[16, 17]

Protecting patients from receiving unacceptable products is paramount throughout the drug development process and requires quantifying statistical errors that lead to incorrect lot release decisions. It is important to understand that statistical hypothesis tests are used to provide proof of an expected outcome. The expected outcome always defines the alternative hypothesis and the corresponding hypothesis test. Concluding that the alternative is true only occurs when its opposite, the null hypothesis, is rejected. If the test results fail to reject the null hypothesis, the only conclusion to be drawn is "insufficient evidence to reject the null." The underlying truth is never known so it is possible to reject, or not reject, the null when the truth would dictate the opposite. For example, to conclude a difference, the null (no difference) must be rejected. If the null is rejected when there truly is no difference, then a Type I error occurs (concluding a difference that is not there). If the null is not rejected when there is truly a difference, a Type II error has occurred (concluding insufficient evidence to support a difference that has occurred). Most importantly, statistical power is always the probability of correctly concluding what is to be proven when it is true.

Tables 1A and 1B (on page 5) are expanded versions of the **Table 1** provided in the original article showing the difference in their application to both an equivalence and a difference model for lot release. Statistical power corresponds to the bottom right corner of both tables—making a correct decision to accept the alternative hypothesis by rejecting the null hypothesis when the alternative hypothesis is indeed true. Stated differently, statistical power is used to quantify the probability of concluding what one intends to prove. It is expressed algebraically as $1-\beta$. By always subtracting the Type II error (the probability of not rejecting the null hypothesis when it should be rejected) from 1 (100% probability), power becomes the probability of rejecting the null when the alternative hypothesis is indeed true. As shown in **Table 1A**, for product lot release using equivalence means concluding equivalence of the reported value to a specified target value.

It is critical to understand that equivalence is exactly what a manufacturer needs to prove because it would be expected that the process and the testing will be well-controlled by design and optimization during development. Specifically, statistical power in the context of a lot-release model for vector concentration using an equivalence approach is the probability of releasing an acceptable lot and correctly assigning the target dose, due to assay variability alone, if the RV is within its acceptance limits. The term "correct" in this context means concluding equivalence when the measured product dose is truly equivalent to the target dose.

If a Type II error occurs, then a vector product lot that is truly equivalent to target will be rejected. This error poses a risk for the manufacturer because it means failing to release an acceptable batch of vector product for pre-clinical or clinical use. However, in early-phase development, research batches and engineering runs are often used for IND enabling studies. If a lot "fails" lot release criteria, the batch can often be appropriately repurposed, for example, being used as a control in an assay or for accelerated stability studies. In other words, failure to hit a dose target may only disqualify its use in a dose escalation study.

It is also critical to note, however, that it is the CI width that dictates whether a reported value is considered equivalent. As shown in **Equation <1>**, the CI HW is calculated using α, or Type I error probability, to derive the appropriate t-value, and the CI is derived by adding and subtracting the CI HW, as shown in **Equation <2>**. The smaller the α-value, the larger the t-value and the higher the confidence in the

TABLE 1: Statistical hypothesis testing and decision errors.

A		TRUTH	
		Not equivalent Null Hypothesis (H_0)	Equivalent Alternative Hypothesis (H_A)
DECISION	Do Not Reject Null Hypothesis (H_0)	Insufficient evidence to conclude equivalent $1-\alpha$ Correct QC REJECTS NOT EQUIVALENT	Insufficient evidence to conclude equivalent Type II Error $=\beta$ QC REJECTS EQUIVALENT Manufacturer RISK
	Reject Null Hypothesis (H_0)	Equivalent Type I Error $=\alpha$ QC RELEASES NOT EQUIVALENT Patient RISK	Equivalent $1-\beta=$ Power Correct QC RELEASES EQUIVALENT LOT

Equivalence Lot Release Model. Statistical power is a measure of being able to prove what is expected. For lot release, the expectation is that product will be manufactured in a consistent manner. In practice, this means that vector batches will have a concentration that can be considered equivalent to the target value. Thus, the equivalence hypothesis test is the appropriate test. However, it is possible that errant decisions will be made. The primary goal is to protect the patient from receiving an unacceptable lot (reduce dosing errors; patient risk) and secondarily to minimize rejection of acceptable batches (reduce losses; manufacturer risk). These probabilities are controlled by specifying limits on Type I and Type II error based on risk tolerance. The lower the Type I error, the larger the *t*-values and the wider the CI value becomes as all other factors are held constant. By factoring in and minimizing Type II error in the sample size calculation for a given Type I error, more replicates will be required to conclude equivalence for a given reported value. Note that when the null hypothesis is not rejected, the only conclusion to be drawn is insufficient evidence to reject the null. However, correctly rejecting the null yields the desired (expected) conclusion and aligns with statistical power, the reward for achieving better process and assay control.

B		TRUTH	
		Not different Null Hypothesis (H_0)	Different Alternative Hypothesis (H_A)
DECISION	Do Not Reject Null Hypothesis (H_0)	Insufficient evidence to conclude different $1-\alpha$ Correct QC RELEASES NOT DIFFERENT	Insufficient evidence to conclude different Type II Error $=\beta$ QC RELEASES DIFFERENT Patient RISK
	Reject Null Hypothesis (H_0)	Different Type I Error $=\alpha$ QC REJECTS NOT DIFFERENT Manufacturer RISK	Different $1-\beta=$ Power Correct QC REJECTS DIFFERENT

Difference Lot Release Model. Statistical power is a measure of being able to prove what is expected. For lot release, the expectation is that product will be manufactured in a consistent manner. Although a difference model is often assumed, applying a difference model is not correct. Note that the manufacturer and patient risks align with the Type I and Type II errors differently than for an equivalence test. While it is still true that when the null hypothesis is not rejected, the only conclusion to be drawn is insufficient evidence to reject the null hypothesis, using a difference model means that good batches are only accepted based on "insufficient evidence to reject the null hypothesis." Correspondingly, statistical power is gained in proving that lots are different from target. Note that this is not what is expected and thus, difference testing is the wrong statistical hypothesis test for lot release.

reported value. For a fixed SEM, this means that the interval will be wider as higher confidence levels are selected, making it less likely to conclude equivalence as the RV moves away from the target. The combination of high confidence with narrow intervals and narrow equivalence bounds leads to less dose error; achieving this is a primary objective.

For an equivalence model, dose error is considered a patient risk. Because Type I error is the probability of releasing a lot when it is not actually equivalent to the target dose, the probability of rejecting a lot when it is truly not equivalent becomes 1–α. Although equivalence testing uses two one-sided tests (referred to as TOST), only one side can falsely fail. In our original paper, we arbitrarily set manufacturer risk to be equal to patient risk so that selection of α automatically leads to a fixed β. To limit dose error, and ultimately the patient's risk, α is first selected, and then the width of the CI is controlled by assay precision and replication levels as well as factoring in manufacturer's risk by including a corresponding β if needed (**Table 1A**). Note that reducing manufacturer's risk requires an increase in power (1–β). The **Linking Assay Qualification Information to QC Lot Release Limits and Statistical Error** is detailed later.

Returning to the subtle but critical concept that selection of a statistical hypothesis test begins with knowing what is to be proven, it important to show that this has been overlooked even when experienced statisticians have published proposed lot release decision rules. For example, Burgess *et al.*[18] invoke the use of a statistical test for difference rather than equivalence. The authors do not directly state that their lot release decisions are based on a statistical difference hypothesis, but this is the underlying model they are using. The Type I error for a difference test does not correspond to patient risk but instead to manufacturer's risk. Specifically, they describe α as the probability of concluding a product lot is out of specification (OOS). In practice, this means that an acceptable lot will fail. Returning to **Table 1B**, this means rejecting the null when one should not and thus, by doing so, one is committing a Type I error. This can only be the case for a difference test—errantly rejecting the null of no-difference when the null is true. In other words, concluding a difference (*i.e.*, concluding the batch is OOS) when there is no difference from target.

But recall that the goal of product lot release (*i.e.*, the expectation) is not to prove a difference. The expectation is that the manufacturing process and the assay are operating in a state of appropriate control. By logical extension, the goal is to demonstrate using statistical proof, that the product is within an established acceptable range. Unfortunately, the proposed use of a difference-based release model by Burgess *et al.*[18] and others[19] promote the common critical mistake[20] made by non-statisticians which is to conclude "not different" when they fail to reject the null hypothesis of "no difference" when applying a difference test for decision-making. Thus, the entire set of rules described in the Burgess paper is fundamentally flawed. Notice that increasing the power in statistical hypothesis testing methods means that the likelihood of rejecting the null hypothesis increases. Since rejecting the null hypothesis is not the desirable outcome for difference testing for release, this clearly points to logical error. Increasing the power for a statistical test, by definition, increases the probability of the 'desirable result.'

To reiterate and emphasize—statistical "proof" only comes from a decision to reject the null hypothesis and thus, conclude the alternative hypothesis is true. This is the case, regardless of the statistical hypothesis used. That said, we are often asked what "failure to reject the null hypothesis" *does* mean. To be clear, the *only* conclusion that can be made for a statistical hypothesis test (equivalence, difference, superiority, or any other form) when one fails to reject the null is this—*there is insufficient evidence to reject the null.* Also, because "not different" (difference = 0; the null hypothesis for a difference test) can never be proven in a difference test, the only possibility for accepting a lot that is different but considered close enough to target is to choose a statistical equivalence strategy and to define *what level of difference* is acceptable. Again, it is to be expected that the process and the assay will introduce variation into the production lots.

What statistics or a statistician cannot do is to *define* what it means for a vector batch dose value to be considered equivalent or close enough to the target. The key is to appropriately limit product and assay variability to protect patients, which requires at the outset, some knowledge of the biological behavior and the associated risks anticipated. By using equivalence, the Type I error is assigned to patient risk rather than manufacturer's risk and can become the sole focus during pre-clinical and clinical studies.

It is also very important to notice that equivalence testing results naturally in both patient risk and manufacturing risk, meaning that both Type I and II error rates must eventually be stipulated to establish the maximum CI HW. Equivalence testing is the only release methodology for which both error rates are built into the specifications. The assignment of Type I and Type II error values for a licensed product will be inextricably linked to the setting of specifications and the demonstration that the dose assay is suitable for intended use. This package of information should be part of the product license application and subject to regulatory review and input.

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

By formally incorporating Type I error (the patient risk when using an equivalence model) into the calculation of a CI associated with an RV, product specifications can provide assurance that both conforming and non-conforming lots will be appropriately judged and dispositioned as such. Our recommendation is to ignore Type II error risk in establishing release limits to be used in pre-clinical studies and clinical trials. Referring to **Equations <6> and <7>** in the original paper, it can be readily seen that the addition of the t_{β}-value in the numerator will lead to much higher replicates. To reduce the burden on QC testing, a lot that is falsely rejected (*i.e.*, a good batch is considered OOS) could always be repurposed if considered suitable for some other application in research and development. Note also that the probability of a lot falling within specification is dependent upon both assay and process variability and bias, but the probability of accepting a lot with a true value at the maximum allowable drift from target is dependent on assay variability alone.

In this manner, the width of the CI will be dependent on SEM once the Type I error limit is assigned (*i.e.*, α is chosen). Referring again to **Equation <1>** it is obvious that the width of the CI is derived from the $t_{\alpha,df}$-value, the assay precision and the replication. Because $t_{\alpha,df}$ decreases with increasing n, once the estimate for assay precision is made, the CI HW can be calculated for varying levels of n. The choice of n will impact how wide the offset for the RV can be for a lot to still be considered acceptable (**Figure 1**). Once the assay precision is estimated, the minimum number of replicates needed to achieve a passing result for a reported value at its limits can be determined. By setting a maximum forecasted CI in this way, a target SEM-value that is practical (*i.e.*, n is not excessive) becomes the goal of assay development and qualification process. Constraints driven by resources and timelines will inform the choice of where effort needs to be placed to preserve the goal of high confidence in dose value assignment. The strategy for achieving this is covered in the following sections on **Reportable Value Sample Size Calculations** and **Incorporating Offset to Account for Allowable Assay Bias and Process Variability**.

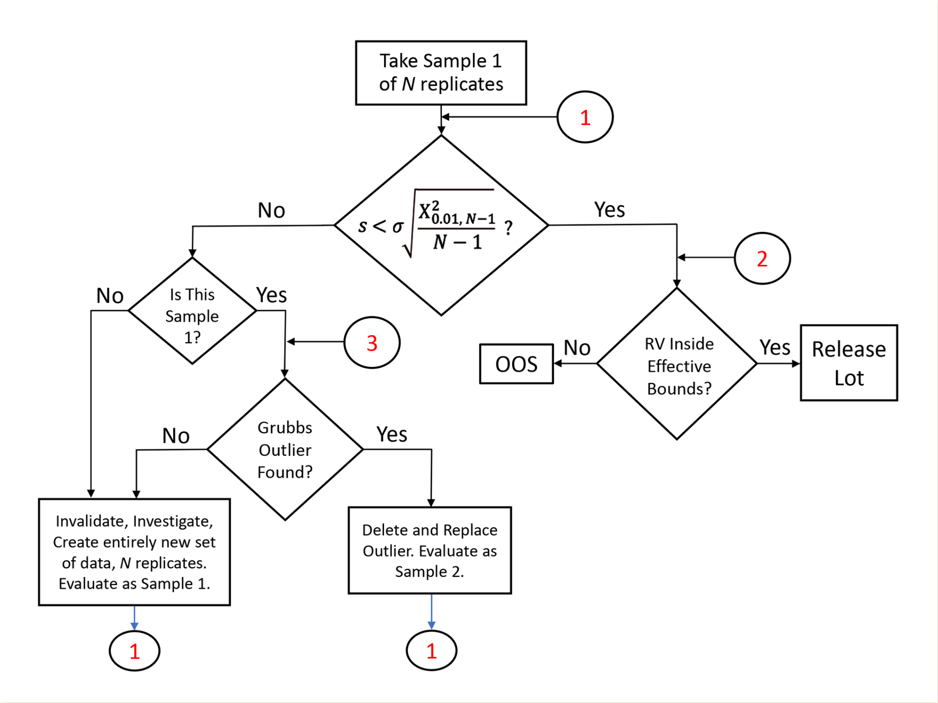
In our original article, we emphasized that the maximum CI for a reported value is one that is forecasted from an assay precision measurement (estimate of population standard deviation, σ). The MU for future RVs is predicted using **Equation <2>**. It is important to note that the CI HW calculation (**Equation <1>**) presumes that the estimate of assay precision, σ, has been derived from both prospectively designed studies and evaluation of any historical assay performance data such that the forecasted CI around the RV is a reasonable estimate. However, after running a QC test on a product lot sample with n replicate measurements, a sample standard deviation, s, can be calculated for the RV. It is necessary to set an upper limit on this s value because of the assumptions being made about the uncertainty of the measurement, which is built into the effective bounds on the reported value. Therefore, the s value should be evaluated as part of a sample suitability assessment, as described in detail in the sidebar and briefly justified as follows.

SIDEBAR 1: Evaluation of Sample Suitability

The true precision of a method must account for all sources of variability over time and is something that can only be estimated through prospective studies and confirmed through ongoing trending in real time. Validation and early assay performance monitoring represent a snapshot used to determine the suitability of the method for its intended use. However, it may be possible, over short periods, to observe smaller variability than the estimated true precision. This is because all the factors contributing to variability may not captured in the replicates used to calculate the RV. By using the forecasted value for σ instead of s in this case, we are ensuring a more conservative estimate of the CI and by extension limiting the drift in the target value to that established in specification development. In other words, we are not allowing for a smaller s value to allow for additional offsets. However, it is also possible, due to random chance, that the measured variability for a sample will also be greater than the forecasted precision value. This scenario is a legitimate cause for concern and is the basis for establishing sample suitability requirements for the RV value replicate variability.

In our proposed evaluation and management of sample results (**Figure 2**), an allowance is made up front...[CONTINUED ON PG. 8]

FIGURE 2: Decision tree for evaluating suitability of a mean value from *N* replicates. Because the CI HW used is one forecasted from the assay precision estimate and the sample size, there needs to be a sample suitability test before a value can be reported and be compared to the product specification. The first step is to test the measured standard deviation, s, using a Chi square test where σ is the estimated assay variability used in the forecast. This should be performed at the 1% level. If s is suitably small (*i.e.*, a "Yes" at decision point 1), then the CI associated with the RV is considered valid and the RV itself can be compared to the acceptance limits in the product specification. At decision point 2, the product lot is either released or deemed to be OOS. However, if the test sample s fails the Chi square test (a "No" at decision point 1), then a Grubbs outlier test (decision point 3) can be performed at the 1% level to determine if an outlier is present. If no outlier is detected, then the test sample fails sample suitability and the RV is considered invalid. If an outlier is identified, then it is deleted and replaced with a value from a single repeat assay run and a new s-value is calculated. The decision point 1 procedure is then repeated with the new sample. If an outlier is identified a second time (sample 2 yields a "No" prior to decision point 3), then the test result with the single replace value is considered invalid. Otherwise, decision point 2 is reached and the RV is considered a valid basis for an equivalence-based lot release decision. We propose that the two allowable Chi square tests are performed at the 1% level in the evaluation to reduce the possibility of rejecting due to multiple testing.


Take Sample 1 of N replicates
1
No
$s < \sigma \sqrt{\frac{X^2_{0.01,N-1}}{N-1}}$?
Yes
2
No
Is This Sample 1?
Yes
3
OOS
No
RV Inside Effective Bounds?
Yes
Release Lot
No
Grubbs Outlier Found?
Yes
Invalidate, Investigate, Create entirely new set of data, N replicates. Evaluate as Sample 1.
Delete and Replace Outlier. Evaluate as Sample 2.
1
1

The s-values are expected to follow a Chi square distribution.[21] As such, limits can be placed on an upper bound to ensure that the forecasted CI is valid for a given sample. That is, an s-value greater than σ is expected at a frequency that can be calculated *a priori* and thus, the forecasted CI can be considered acceptable if the s-value for a given RV falls within predefined limits based on probabilities. In no case are we proposing that the forecast CI is replaced with a CI derived from sample data. Thus, all sample s-values that fall under the upper limit are considered valid for accepting the forecasted CI. The corollary is also true, any s-values below the limit cannot be used to justify expanding RV acceptance limits during the lot release decision process.

SIDEBAR 1 (cont.):

Evaluation of Sample Suitability

. . . for some degree of "variability in the variability estimate" using the Chi square distribution. Thus, if the sample variability exceeds the Chi Square limit, a statistical outlier test can be used to discard a single aberrant point and serve as justification for the generation of a single replacement replicate. The replacement is necessary to maintain the "n" used for direct comparison to the forecasted CI.

A sequential testing approach, as outlined in **Figure 2,** should not be considered one of "testing into compliance," rather it should be viewed as a QC lot release testing framework in which there is a "penalty to be paid" for unexpectedly high variability for a given sampling. The goal of the equivalence approach for lot release is to demonstrate acceptable confidence in the measured value. Being able to average more data provides a better estimate of the mean and greater confidence in the measurement. This is best viewed as the "statistical reward" for the practical price incurred for needing additional replicates due to the occasional "bad luck" with low probability events leading to high sample variance. However, to be very clear, the omission of an outlier data point and the need to do additional replication should indeed be a relatively rare event if the assay precision estimated from development and validation studies is a good approximation of the true precision of the method. If the precision estimate is underestimated due to poorly designed studies or unforeseen variability issues, the s-value could routinely exceed the Chi square limit, and this would serve as a signal to revisit the validity of the predicted precision value.

We recommend using Grubb's test to identify outliers[22] and limiting outliers to a single-value removal. Note that the Grubbs test requires Gaussian or normal data, so it must be emphasized that outlier tests on log normal data must be performed on log-transformed results to be statistically valid. If an outlier is detected, rather than simply discarding the value and averaging the remaining replicates, we advocate for allowing up-front for a single repeat test to replace the discarded outlier value.

Reportable Value Sample Size Calculations

Once equivalence bounds are set, the Type I error risk tolerance is established (*i.e.*, α chosen), a value for the offset determined, and assay precision estimates appropriately derived through well-defined qualification studies, then the necessary replication to minimally achieve a corresponding specified statistical power can also be calculated. This requires defining the Type II error risk tolerance and thus, the probability that an acceptable lot will return an OOS result based on assay variability. The stage of product development and the available resources will force evaluation of the trade-offs, costs, and timelines when choosing whether to replicate more during lot release testing or reduce variability. In some instances, organizations may choose to accept greater risks in lot release decisions. Ignoring Type II error in early development will lead to a cost savings, in terms of fewer replicates needed to demonstrate equivalence, while maintaining a low probability (α) that an unacceptable lot is released for use (**Table 1A**), and by extension, retaining a high probability of rejecting a lot that is not truly equivalent.

Remember, the end-goal in early development is to arrive at reported dose value approximations that can be accepted as reliable enough to satisfy, in practice, the statistical assumption that dose values used in a dose-response study are exact and error-free. By reducing to practice something that is theoretically impossible, end-users must find ways to figure out what is pragmatic yet rigorous enough to reasonably work within the statistical framework. Conventions can provide convenient defaults but ultimately, these decisions come down to accountability for risk. The choices are context and consequence-driven and rely on end-users having sufficient understanding of the framework, and trade-offs being quantified and communicated with data.

The primary challenge here is in establishing the equivalence bounds around dose input levels. Once those are established, the replication strategy can be calculated as described using **Equation** <7> in our original paper (or modifying **Equation <6>** by incorporating an offset and allowing for smaller n). Because high levels of replicates are undesirable yet may be required, some non-statisticians have turned to using an acceptable R^2-value for a dose-response model as evidence that lower levels of replication in determining X-values could be acceptable. This approach underscores the dangers of failing to understand the meaning of a statistical tool or technique and misapplying it to arrive at invalid conclusions. The R^2 for a dose-response regression has a very specific meaning and utility.

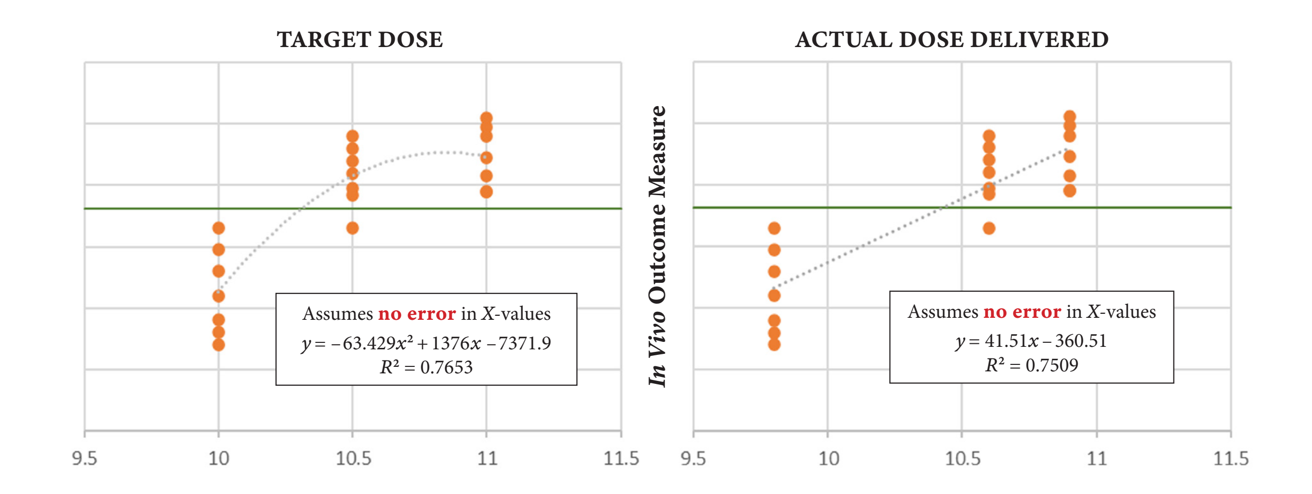
The therapeutic window for a product is defined by the difference between the dose that is efficacious and a dose that causes unacceptable adverse effects. It is typically derived from *in vivo* preclinical dose escalation studies and human clinical trials. Variability in responses to dose levels administered will be a function of individual recipient differences and variability in efficacy and safety outcome measurements made. To determine how well dose level (x input) predicts response outcomes (y readout), a dose-response relationship is evaluated. By fitting a mathematical model to study data, a corresponding goodness-of-fit can be quantified using R^2-value or root mean square error (RMSE). For both of these metrics of predictive value, the difference between the observed y results and those y-values predicted using the regression fit to the curve model selected, are used in the calculation. As detailed in the second side bar, model fit metrics cannot be used as a measure of the exactness in x input. Although, for many gene therapy products, single-dose clinical studies, rather than dose escalation studies, are used to support regulatory filings, consistency in dose assignment throughout the drug development process is still important in defining the product dose.

SIDEBAR 2: Dose Errors Cannot be Identified by Model Fit Metrics

If the values obtained in the study are plotted as a function of dose but the dose values are not correct, then the mathematical function that fits the observed data and/or the estimates of the model parameters will also be incorrect (**Figure 3**). The exactness of the dose input cannot be deduced from an R^2-value or the RMSE reported from a regression for the chosen model. This is because these measures simply describe how far the observed y-values fall from the y-values that are predicted using the model parameters derived from the regression process. This statistical fitting of dose-response data to a model assumes *a priori* that the input x-values are exact and accurate. Furthermore, an R^2-value is only valid for linear models, and it expresses how much of the variation in the y-values can be explained by fitting the data to the selected model, in comparison to using the mean of the y-values (the simplest model which would assumes that x does not predict y at all).

R^2 is a measure of how well an x-input can predict a y-outcome—and obviously x cannot predict y at all if the actual x-input is not the assigned, assumed x-value. An R^2-value close to 0 indicates a model with very little explanatory (predictive) power and an R^2-value close to 1 indicates a model with more explanatory (predictive) power; both assume that the x-values are exact. With a clear understanding of this, it is easy to see by extension that dose values must also be consistently measured from proof-of-concept research studies through pre-clinical and clinical studies, if they are to be compared. This is true whether mathematical models are formally fit to data, or the data are only visually evaluated. Without consistency in dose assignment, predictive power is lost because neither the model, nor the goodness-of-fit metric, make any statistical sense.

FIGURE 3: Dose errors cannot be identified by model fit metrics. The two graphs use the same example outcome data from a pre-clinical study to demonstrate that the R^2-value obtained fitting a dose-response model, based on the assumption that the target dose is correct, is virtually the same as what is obtained using a different model fit for the actual dose delivered. Note also that most non-statisticians think of the $y = mx + b$ straight line function as a linear model. While it is a linear model, it is not the only linear model. Here, linear refers to the parameters of the model. For a straight line, the model parameters are slope (m) and y-intercept (b). However, a polynomial such as $y = ax^2 + bx + c$ is also a linear model with the coefficients, a, b, and c being linear parameters (remember, x and x^2 are known). Non-linear models also include the four-parameter and five-parameter logistic equations commonly used in potency and bioanalytic assays. Because the regression process is different for fitting linear and non-linear models, the assumptions required for R^2 cannot be made for non-linear regressions and hence are meaningless. RMSE can be used as a goodness of fit tool for both types of models and is what we recommend be used. Again, if actual dose inputs are grossly incorrect, then the model chosen will also not reflect the true relationship between dose and the response measured. However, as the graphs depict, that incorrect model could have a good fit to the data.



Incorporating Offset to Account for Allowable Assay Bias and Process Variability

As already explained, reduced drift from center means that the offset value applied to the target can be smaller, so the acceptance limits on the RV form a tighter specification (**Figure 1**). From an assay development standpoint, reduced assay variability means that QC can run fewer replicates to achieve lot release objectives without compromising product quality. As discussed in the **Measurement Uncertainty** section, TAE must be considered. To account for assay bias (the difference between the true dose value of the vector lot and the measured value) and its impact, we have combined the drift in process (the difference between the true value of the lot and the target) with the assay bias to yield a maximum allowable offset. This approach not only highlights the trade-offs that exist in the burdens shared by the process development, assay development, and QC teams, it provides a framework to easily quantify them. By adjusting the values that can be chosen, and assuming that further efforts will not be made to improve assay precision, the minimum number of replicates needed to conclude equivalence of an RV to the established bounds can be calculated. Eliminating manufacturer's risk as a concern early on allows for a reduced burden on QC and/or the establishment of narrower equivalence bounds. Again, we believe the latter should be the driving factor in the early stages of product development.

The equivalence model for lot release establishes the framework to evaluate risks associated with QC decision-making and in the earliest stages, allows for ignoring or accepting higher manufacturing risk to compensate for the focus on reducing the risks of Type I error to minimize dose assignment error in establishing therapeutic window. The efforts to control the manufacturing process and improve assay accuracy and precision all require the allocation of significant resources. Thus, it should be obvious now that statistical power is something purchased. It generally does not come cheap, but by being focused on evaluating the trade-offs in quantitative terms, more efficient progress toward greater statistical power can be achieved. It should also be noted that lots that do not meet the release requirements for dose during development may be suitable for research purposes, as assay controls, or for accelerated stability studies. This means that lots that "fail" for dose can be repurposed rather than discarded during development.

Also, it is worth noting that making dilutions and/or formulating drug product based on a concentration measurement made on the bulk drug substance assumes that the concentration value is exact (*i.e.*, is the basis of a dilution calculation). In practice, it makes more sense to increase the stringency on the dose measurement at the bulk drug stage as it is often the source of drift from target at the final container stage. In other words, dilution error-based on reported values with high MU is likely one of the main causes of missing the target dose at fill. By applying an equivalence model at both bulk and final container stages, then the HW of the CI on the RV for drug substance is mathematically added to the offset value for final container to account for the propagation of error introduced by measuring twice.

Establishing a Lot Release Specification for Dose

Pre-clinical development studies and human clinical trials that can demonstrate that a drug product is safe and effective are necessary but not sufficient for commercial licensure. Among other regulatory requirements, it is critical that the production process and the lot release procedures can be shown to operate within a state of control. The lot release specifications for drug substance and final product are informed by the risks to patients posed and by the manufacturing process and lot release assay variability. Risks to patients are evaluated based on *in vivo* studies that link safety and efficacy outcomes to dose level(s) such that a target dose can be defined, and limits on product drift from it can be established. Therefore, one of the most critical assays for viral vectors is the one used to determine the product concentration—the volume of product required to deliver a specific dose is calculated from this value.

The predictive capacity of pre-clinical and clinical data dose-response models is dependent on the statistical assumption that dose input levels are exact. As acknowledged already, it is impossible for dose levels to be error-free because dose is based on a measurement. However, restrictions on MU can be used to set limits on the amount of error that can be introduced. By using an equivalence approach to setting lot release acceptance limits for product to be used in pre-clinical and clinical development, the dose targets serve as center-points for the specification. The spacing of doses corresponds to the intended dose escalation, such as half-log (3.16-fold increase). In our original paper, we provided an example set of calculations based on using equivalence bounds derived from the addition and subtraction of 0.25 log dose units from the target values. This HW value represents the maximum value that can be used to set equivalence bounds, but it was not meant to be a recommendation.

The actual equivalence bounds chosen by sponsors should be based on the level of dose escalation error that can be calculated from assuming that vector could be at a concentration corresponding to each of the upper and lower equivalence limits. Assuming, for example, that a low dose of 10 log units is considered a starting point, then a half-log dose escalation would mean the next dose up would be 10.5 log units. A range of potential dose escalations allowing for a specified error tolerance can then be calculated.

TABLE 2: Dose escalation error limits.

Accepted Dose Error at Equivalence Limits	Dose Targets for Half-Log Escalation from Starting Point Using 10 as an Example		Dose Escalation Range Possible Using Lowest and Highest Values from Equivalence Limits in Worst Case Combinations		Equivalence Bounds
	10	10.5	Log	Fold	
Lowest	9.75	10.25	0	1	Target ±
Highest	10.25	10.75	1	10	0.25
	10	10.5	Log	Fold	
Lowest	9.875	10.375	0.25	1.8	Target ±
Highest	10.125	10.625	0.75	5.6	0.125
	10	10.5	Log	Fold	
Lowest	9.9375	10.4375	0.375	2.4	Target ±
Highest	10.0625	10.5625	0.625	4.2	0.0625
	10	10.5	Log	Fold	
Lowest	9.9688	10.4688	0.4375	2.7	Target ±
Highest	10.0313	10.5313	0.5625	3.7	0.0313
	10	10.5	Log	Fold	
Lowest	9.9844	10.4844	0.4688	2.9	Target ±
Highest	10.0156	10.5156	0.5313	3.4	0.0156

Because the width of the equivalence interval defines the range of values that are considered equivalent to the target, taking the highest theoretical value of the lower dose (*e.g.*, 10 log units target plus equivalence half-width used) and the lowest value of the next higher dose (*e.g.* 10.5 log units target minus the equivalence HW used), the lowest dose escalation possible can be determined as shown in both log units and fold differences (calculated by back transformation $10^{\text{log difference}}$). The values shaded in blue represent the minimal dose escalation possible; they come from subtracting the upper right blue from the lower left blue for each set of equivalence bounds. Similarly, for the pink-shaded values, the upper left value is subtracted from the lower right value to determine the maximal dose escalation possible. As the equivalence bounds narrow, the width of the range decreases and approaches the intended value (for 0.5 log; 3.16-fold) thus reducing the dose error that can occur. Using the ± 0.25 limits proposed in our original paper, the range of dose errors spans from 0 (1-fold) to 1 (10-fold). By reducing the equivalence bounds by half, that range narrows to between 1.8 and 5.6-fold.

Table 2 provides example dose escalation ranges for five different equivalence bound limits showing the corresponding range of dose escalation errors that can occur. As the bounds narrow the range of the potential escalation errors narrows from 1–10 for ± 0.25 and approaches the target value of 0.5 log or 3.16-fold with 2.9–3.4 being achievable using + 0.0156 limits. The actual equivalence bounds chosen by sponsors should be based on the level of dose escalation error that can be tolerated, taking into consideration the specific risks associated with the disease and patient population that the product is designed to treat.

It is important to note that although the equivalence bounds define the limits on how different the RV for product dose could be from target and still be considered acceptable (*i.e.*, of no meaningful consequence), these are not the limits placed on the RV. This is because, to conclude equivalence requires that both the reported value and its CI fall within the bounds (**Figure 1**). Thus, it is the width of the CI that drives how much processing drift and how much assay bias can be tolerated in a reported value and *vice versa*. **Table 2** takes the RV values to the extremes (*i.e.*, the equivalence bounds) by assuming an exact RV to generate the most extreme scenarios. Referring again to **Figure 1**, it can be readily seen that the greater the amount of drift from the target value caused by process and measurement bias together, the smaller the CI must be to conclude equivalence. Again, remember that the choice of equivalence limits defines *a priori* what is considered acceptable. The more well-behaved the measuring system, the greater the allowable variability in the production process. And the more consistent the process, the more MU can be tolerated. However, always keep in mind, the end-goal is to ensure that both the process and the assay are well-controlled, and that dose error is kept to a minimum.

Because process variability measurement includes assay variability, the capability of the assay should be determined early on. By repeatedly measuring samples of a stable homogeneous material, process variability can be removed from the variability equation, and assay precision alone can be estimated. Statistically designed experiments can yield variance components analyses and provide an indication of which factors contribute most to assay variability. Once identified, the most critical robustness factors should be optimized for minimal impact on the TE associated with the RV. When the assay procedure is ready to be implemented for lot release, a pragmatic testing strategy should be developed to maximize the amount of drift in process that can be tolerated for the established equivalence bounds. By increasing the number of values averaged and strategically replicating across the highest sources of variability, greater confidence in a mean reported value can be achieved. At the same time, efforts to control the process should be made to increase the probability of hitting the desired target dose.

As already discussed, an equivalence-based lot release model based on a statistical difference test is flawed, whereas an equivalence model is appropriate. Our approach is further supported by the official USP chapter <1210>[10] in which a TOST is given as an example for an equivalence-based lot release strategy for product strength. As we have also indicated regarding a Type I error for equivalence (**Table 1A**), the USP <1210> currently notes that selecting an α represents the maximum risk of stating that the acceptance criterion is satisfied when a lot is not truly equivalent. Interestingly, difference hypothesis testing is sometimes applied to data after equivalence has been demonstrated. Such practices have led to the conclusion that the equivalence approach is flawed. This is because in some instances, if the equivalence bounds are at the maximum (± 0.25 log dose with back-to-back specification spaced at half-log intervals), then it is possible that with enough replication and low MU, RVs within a given specification may have non-overlapping CIs, and an RV CI may not include the target value. There are indeed four different outcomes that can occur when difference and equivalence tests are applied to the same data set using the same α-value[23]

(1) Equivalent and not different
(2) Equivalent and different
(3) Not equivalent and not different
(4) Not equivalent and different

Statistically significant differences can arise, but it critical to recognize that in an equivalence approach, there is a decision *a priori*, as to the level of error that will be considered "inconsequential" (*i.e.*, a meaningless difference). Thus, it is scenario (2) that is the cause for confusion. From the possibility of concluding both equivalence and difference, it follows for some that the equivalence approach is wrong, or it should be modified to allow for dose volume adjustments when RVs are different. However, what is required by adopting an equivalence-based strategy for lot release is to establish the equivalence bounds such that they reflect an acceptable level of dose error. The equivalence interval becomes the limits within which dose volume adjustments will not be allowable. This means accepting statistically significant differences between batches for doses within the range (and/or different from the target) in some cases. At a minimum, the goal is to ensure that no reverse dose escalation is occurring (**Table 2**; lowest escalation is 0 or no escalation). This is accomplished by creating gaps between acceptable ranges on the RVs across the dose levels such that lots that do not fall within the targeted dose ranges will be rejected and repurposed rather than used in dose escalation.

If the specifications are broken down into "smaller buckets" with smaller equivalence bounds and smaller spacing in-between target values, then testing can be set up such that a passing lot can be adjusted for dose volume to achieve dose escalation. In this strategy, QC must work to reduce the width of the CI associated with the reported value. There are three ways they can achieve reduced uncertainty in the RV:

(1) Improve the assay precision.
(2) Increase the number of replicates combined to yield a RV.
(3) Reduce the confidence level (*i.e.*, increase acceptable Type I error = allow for less exactness in *x*).

Alternatively, the equivalence bounds are narrowed on the target values while maintaining the target dosing at dose escalation intervals such that greater consistency in manufacturing will be required to maintain a single volume across dosing levels. In this scenario, QC is still required to work to keep CIs narrow, but pressure is also placed on manufacturing to hit the target values. Ultimately, this is the goal for both process and assay validation—to demonstrate that the manufacturer is operating within an acceptable state of control when a product reaches a Biologics License Application (BLA) submission stage.

Conclusions and Recommendations

To calculate a dose volume, the nominal assigned target value for vector concentration (dose) can be used when an equivalence approach is adopted. By demonstrating that a reported value is equivalent to the intended target specification, stakeholders can be assured that released products will perform within the established safety and efficacy profile. Note that when an equivalence model is adopted, the Type I and Type II errors are the opposite of what they would be for a difference model such that Type I error becomes patient's risk and Type II, the manufacturer's risk. In all statistical hypothesis tests, statistical power—the ability to prove what is to be proven when that outcome reflects the underlying truth—is increased when variability and bias are decreased and sample size is increased. As we have published previously regarding the use of equivalence for parallelism testing for bioassays[24] when greater statistical power returns a counterintuitive outcome, that can be taken as evidence that the wrong statistical hypothesis framework is being used.

This is precisely what happens using a statistical difference test for lot release. Higher statistical power means that it becomes easier to prove that lots that fall outside of the range would be considered different, which is not the conclusion that is intended to be routinely reached. By using an equivalence approach, high confidence in rejecting an unacceptable lot leads to greater confidence in interpreting dose-response data and thus, protects patients while greater statistical power results in an increased ability to prove that an acceptable lot is within the established specification and equivalent to the target. With the equivalence approach to lot release, any measured difference within bounds is considered meaningless. The balancing of patient and manufacturer risk is then used appropriately in establishing the width of the CIs associated with a reported value. To provide further clarity, specific recommendations and a succinct summary, the order of operations thus, becomes as follows.

All measurements are assumed to be normal for statistical analyses therefore and thus, must be in log units for lognormal data. Because this is most often the case, the examples are in log units. The first step is to assign target values that will be the center of the specification(s). The target values will be spaced according to the dose volume adjustment strategy. If dose volume adjustments will not be allowed, the target values are spaced at the same interval as the dose escalation (*e.g.*, half-log). If allowances for volume adjustments will be made, then smaller increments may be used, as described above. Either way, the next step is to establish the equivalence bounds. We recommend that these are done symmetrically by adding and subtracting a value that represents the range of values within which dose volume adjustment is not considered necessary.

These equivalence bounds set limits on allowable dose error. In the original units of measure, this error can be expressed as the maximal fold difference in values that is considered acceptable. With half-log dose targets and ±0.25 bounds, this means that RVs up to 1.78-fold from target are considered acceptable. Dropping those bounds to ±0.125, this fold difference becomes 1.33. Recall that with a half- log escalation, the intended increments are 3.16-fold. Using ±0.25 log unit bounds allow for the lower dose and next level up dose to be off by as little as 0 and as much as $4*0.25$ (from lower −0.25 to next level up +0.25 for a total of 1 log unit in dose difference), which translates into a range of possible dose escalations from 1–10, as shown in **Table 2**. This arises if the low dose is on the high side and the higher dose on the low side or *vice versa*. However, no reverse dose escalation should occur. For ±0.125-derived bounds, that range is further restricted from 1.8 to 5.6-fold. As the equivalence bounds approach 0, then dose escalation approximates the intended dose spacing.

Because the reported value and the CI must fall within the bounds for the vector batch to be considered equivalent to target, the width of the CI will determine how much skew in the process (drift of true dose from target) can be allowed, as well as how much bias from assay can be tolerated. The CI width is derived from the risk tolerance for Type I, the precision of the dose determining assay (σ, run-to-run standard deviation), and the number of replicate runs combined to yield a reported value. The higher the confidence desired, the greater the number of replicates needed to achieve a narrow interval. A higher probability of passing lots that are truly equivalent means more usable lots. This goal can be met by reducing Type II error. Thus, the third step in the order of operations is deciding α and β-values.

In early-phase development, we suggest calculating the width of a 90% CI for different numbers of replicates using reasonable estimates of precision. The 90% CI corresponds to a conventional 5% error of assigning a batch as equivalent to the target value when in fact, it is not (*i.e.*, a Type I error; risk of errant dosing). Note that 5% corresponds to a 90% CI, rather than 95% CI, because equivalence testing uses two simultaneous one-sided difference tests.[25] Because adding in Type II error drives up the number of replicates needed to assure that good batches will fall within specification and because Type II error does not introduce error into dose (rather the manufacturer does not use an acceptable lot), we recommend ignoring the risk of failing (or having to repurpose) an acceptable lot during development. This means delaying the assignment of both final α and β-values and finalizing the corresponding replication strategy until clinical studies and assay performance assessments are complete, and when commercial product specifications are being established.

In other words, we do not recommend factoring in Type II error at the beginning, although **Equations <4>, <6>, and <7>** in the original paper includes $t_{\beta,df}$ in the calculation for minimum number of replicates. Using this strategy, assay qualification means demonstrating that the assay precision and corresponding replication strategy yields an offset and corresponding RV range that are deemed satisfactory. The wider the range required, the smaller the corresponding CI. The balance comes from evaluating the relative capabilities of the manufacturing process to hit the target and the assay used to return a tight estimate of dose. Lack of consistency in manufacture will reduce the probability of an RV falling within its limits. Process bias leads to a larger offset or higher risk of OOS results. It also burdens QC to compensate for wider RV acceptance limits within the same established equivalence bounds. Similarly, a lack of consistency in measurement burdens manufacturing with hitting the target or QC to compensate with higher replication, as described already. The widening of equivalence bounds relieves the pressure for both but also increases the risk of dose error. By conducting the exercise of implementing the proposed lot release strategy, the need to shift focus on improving manufacturing and/or assay capabilities may become evident.

Correspondingly, it is important to understand that in our approach to TE, we combine the assay bias and the process drift into a single allowable offset for a reported value. It is this value around which the CI is applied. By subtracting the allowable offset from the established equivalence bounds, we create “effective” bounds for the RV which require increasingly smaller CIs as the effective bounds shrink away from the original equivalence bounds. Thus, the further the true value deviates from center, and the greater the assay bias, the more confidence in the result required to conclude equivalence. The confidence in the result is derived from the confidence level (chosen Type I error tolerance) and the assay precision and replication. Choosing a higher confidence will widen the interval. To compensate and maintain a narrow interval, precision must be better and/or greater replicates must be included.

The TE, as embodied in our model, directly connects equivalence bounds to RV drift from target (*i.e*, offset) that is contributed by both process variability and assay bias. Assay qualification/validation is achieved by setting limits on the CI width, which links reported values to equivalence limits set for studies to establish therapeutic window (or at licensure, derived from clinical data). From a theoretical standpoint, equivalence bounds are statements about how far the true value of the lot can drift from the center and still be considered clinically equivalent to target. Our original article addressed the challenges of assigning lot acceptance and dose assignment for vector batches destined for pre-clinical studies and clinical trials. However, it is our expectation that the same equivalence approach will be applied for commercial products.

At that point, the final product specification for product dose will reflect a consensus between regulators and manufacturers that risks to patients are minimized while expectations of benefits to them are maintained, as measured by a consistent supply of high-quality product. The clinical data are used to establish the equivalence bounds on the target dose to ensure that product with unacceptable safety and efficacy limits are not reached. Once established, the capability of the process and the assay to ensure supply chain demands are met must be demonstrated. Using an equivalence model, product may drift from the target to a certain degree to allow for process variability, but as the true lot value moves away from the target value, the CI must narrow to conclude equivalence of the reported value within established bounds. Remember that the specification sets limits on the reported value. These acceptance limits reflect the range of reported values and their associated CI fitting within the wider, pre-established equivalence bounds. The probability of errant decisions is a function of both process and assay variability, but the probability of an errant decision based on true RV limits is restricted to assay variability. In other words, the probability of an RV landing within limits is a function of process variability, whereas the probability of concluding the RV is within bounds is a function of RV MU.

To ensure that our proposed approach to dose assignment and lot release leads to more accurate scientific conclusions and greater protection of patients, it is critical that those accountable for its use have a solid understanding of statistical hypothesis testing, dose-response models, and MU. Although we encourage ongoing review of other sources for introductory statistics to improve comprehension, our experience indicates that the challenge of applying statistical techniques arises when the relationship between multiple tools required is not obvious. Clear communication between those who generate and review dose data and statisticians responsible for ensuring appropriate analyses is key to achieving valid outcomes.

Demonstrating that an assay is fit for purpose means defining both what fitness means and precisely what the purpose is, in terms of a statistical hypothesis framework. We have undertaken to further clarify why an equivalence model for lot release and dose assignment is correct, and to make the connections between the calculations required to adopt the approach clearer. In so doing, we have provided answers to the six questions posed in the introduction and equipped both non-statisticians and statisticians with the knowledge needed to be accountable for protecting patients and justifying an equivalence approach. We end with acknowledging that using an equivalence approach for lot release raises questions related to other applications including, but not limited to stability monitoring and method-bridging studies. These and other considerations were beyond the scope of this paper and will be addressed in future articles.

References

[1] FDA CBER—https://www.fda.gov/vaccines-blood-biologics/cellular-gene-therapy-products/approved-cellular-and-gene-therapy-products Last accessed as content—current as of: 06/30/2023

[2] Zhao Z, Anselmo AC, Mitragotri S. Viral vector-based gene therapies in the clinic. *Bioeng Transl Med*, 2021 Oct 20;7(1):e10258. https://doi.org/10.1002/btm2.10258. PMID: 35079633; PMCID: PMC8780015.

[3] Shen W, Liu S, Ou L. rAAV immunogenicity, toxicity, and durability in 255 clinical trials: a meta-analysis. *Front Immunol*, 2022; Oct 27; 13:1001263. https://doi.org/10.3389/fimmu.2022.1001263. Erratum in: *Front Immunol*, 2023 Jan 18; 13:1104646. PMID: 36389770; PMCID: PMC9647052.

[4] Stone D, Aubert M, Jerome KR. Adeno-associated virus vectors and neurotoxicity—lessons from preclinical and human studies. *Gene Ther*, 2023 May 10. Online ahead of print. https://doi.org/10.1038/s41434-023-00405-1

[5] Sajjadi N, Callahan JD. Defining therapeutic window for viral vectors: a statistical framework to improve consistency in assigning product dose values. *BioProcess J*, 2020; 19. https://doi.org/10.12665/J19OA.Sajjadi

[6] FDA Guidance for Industry—*Chemistry, Manufacturing, and Control (CMC) Information for Human Gene Therapy Investigational New Drug Applications (INDs)*. January 2020. https://www.fda.gov/media/113760/download

[7] Pharmacopeial Forum (PF) PF46(5) <1220> *Analytical Procedure Life Cycle*. May 2022. https://doi.org/10.31003/USPNF_M10975_02_01

[8] Schofield T *et al*. Distinguishing the analytical method from the analytical procedure to support the USP analytical procedure life cycle paradigm. *2019 Pharmacopeia Forum*. Vol. 45(6).

[9] Theodorsson E, Magnusson B, Leito I. Bias in clinical chemistry. *Bioanalysis*, 2014; 6(21):2855–75. https://doi.org/10.4155/bio.14.249. PMID: 25486232.

[10] United States Pharmacopeia (USP) and National Formulary USP43-NF38 <1210> *Statistical Tools for Procedure Validation*. May 2018. https://doi.org/10.31003/USPNF_M8646_07_01

[11] United States Pharmacopeia (USP) and National Formulary USP43-NF38 <1225> *Validation of Compendial Procedures*. Aug 2017. https://doi.org/10.31003/USPNF_M99945_04_01

[12] FDA Guidance for Industry—Q2(R1) *Validation of Analytical Procedures: Text and Methodology*. Sep 2021. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/q2r1-validation-analytical-procedures-text-and-methodology-guidance-industry/

[13] FDA Guidance for Industry—*Bioanalytical Method Validation*. May 2018. https://www.fda.gov/media/70858/download

[14] European Medical Agency—*ICH guideline M10 on bioanalytical method validation and study sample analysis*. Jul 2022. https://www.ema.europa.eu/en/documents/scientific-guideline/ich-guideline-m10-bioanalytical-method-validation-step-5_en.pdf

[15] European Medical Agency—*ICH guideline Q2(R^2) on validation of analytical procedures*.
https://www.ema.europa.eu/en/documents/scientific-guideline/ich-guideline-q2r2-validation-analytical-procedures-step-2b_en.pdf

[16] FDA CBER— *Q6B Specifications: Test Procedures and Acceptance Criteria for Biotechnological/Biological Products*. Aug 1999. https://www.fda.gov/media/71510/download

[17] European Medical Agency—*ICH guideline Q10 on pharmaceutical quality system*. Jun 2008. *ICH guideline Q10 on pharmaceutical quality system—Step 5*. Sept 2015. https://www.ema.europa.eu/en/documents/scientific-guideline/international-conference-harmonisation-technical-requirements-registration-pharmaceuticals-human_en.pdf

[18] Burgess C, Curry P, LeBlond DJ, Gratzl GS, Kovacs E, Martin GP, McGregor PL, Netthercote P, Pappa H, Weitzel J. Fitness for use: decision rules and target measurement uncertainty. 2016. *Pharmacopeial Forum*. Vol. 42(2).

[19] United States Pharmacopeia (USP) and National Formulary USP43-NF38 <1033> *Biological Assay Validation*. https://doi.org/10.31003/USPNF_M912_01_01

[20] Sackett DL. Superiority trials, non-inferiority trials, and prisoners of the 2-sided null hypothesis. *BMJ Evidence-Based Medicine*, 2004;9:38–39. http://dx.doi.org/10.1136/ebm.9.2.38

[21] National Institute of Standards and Technology (NIST)—*Chi-Square Test for the Variance*. https://www.itl.nist.gov/div898/handbook/eda/section3/eda358.htm

[22] Pharmacopeial Forum (PF) PF46(4) <1010> *Analytical Data—Interpretation and Treatment*. Dec 2021. https://doi.org/10.31003/USPNF_M99740_05_01

[23] Lakens D. Equivalence tests: a practical primer for *t* tests, correlations, and meta-analyses. *Soc Psychol Personal Sci*, 2017 May;8(4):355-362. https://doi.org/10.1177/1948550617697177

[24] Callahan JD, Sajjadi NC. Testing the null hypothesis for a specified difference—the right way to test for parallelism. *BioProcess J*, 2003; 2(2): 71–77. https://doi.org/10.12665/J22.Callahan

[25] 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*, Principal Quality Consultant,
Sajjadi Consulting, Fairfax, Virginia USA

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

***Corresponding Author:**

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