Commentary

Statistical Considerations for Assessment of Bioanalytical Incurred Sample Reproducibility

David Hoffman^{1,2}

Received 27 May 2009; accepted 29 July 2009; published online 8 August 2009

Abstract. Bioanalytical method validation is generally conducted using standards and quality control (QC) samples which are prepared to be as similar as possible to the study samples (incurred samples) which are to be analyzed. However, there are a variety of circumstances in which the performance of a bioanalytical method when using standards and QCs may not adequately approximate that when using incurred samples. The objective of incurred sample reproducibility (ISR) testing is to demonstrate that a bioanalytical method will produce consistent results from study samples when re-analyzed on a separate occasion. The Third American Association of Pharmaceutical Scientists (AAPS)/Food and Drug Administration (FDA) Bioanalytical Workshop and subsequent workshops have led to widespread industry adoption of the so-called “4–6–20” rule for assessing incurred sample reproducibility (i.e. at least 66.7% of the re-analyzed incurred samples must agree within $\pm 20\%$ of the original result), though the performance of this rule in the context of ISR testing has not yet been evaluated. This paper evaluates the performance of the 4–6–20 rule, provides general recommendations and guidance on appropriate experimental designs and sample sizes for ISR testing, discusses the impact of repeated ISR testing across multiple clinical studies, and proposes alternative acceptance criteria for ISR testing based on formal statistical methodology.

KEY WORDS: bioanalysis; containment proportion; incurred samples; reproducibility; tolerance interval.

INTRODUCTION

Bioanalytical methods for the quantitative determination of drugs and their metabolites in biological matrices provide critical support for the evaluation and interpretation of bioavailability, bioequivalence, pharmacokinetic, and toxicokinetic studies. The quality and integrity of these studies is inherently reliant on the quality and integrity of the accompanying bioanalytical data. Well-characterized and fully validated bioanalytical methods are thus essential to ensure the safety and efficacy of pharmaceuticals.

Scientific and regulatory guidelines for bioanalytical method validation have continued to evolve since the first bioanalytical method validation workshop held in 1990 (1). More recently, the Third American Association of Pharmaceutical Scientists (AAPS)/Food and Drug Administration (FDA) Bioanalytical Workshop resulted in a conference report which sought to propose best practices in bioanalytical method validation for both small molecules and macromolecules (2). While noting that the current FDA bioanalytical guidance (3) remains valid, the conference report recommends that certain additional validation studies be performed. In particular, laboratories are now expected to

demonstrate the reproducibility of bioanalytical methods using study samples from dosed subjects (incurred samples).

The goal of incurred sample reproducibility (ISR) testing is to demonstrate that the bioanalytical method will produce consistent results from study samples when re-analyzed on a separate occasion. The term reproducibility is often used to refer to the precision of a bioanalytical method between two laboratories (3). In the context of ISR testing, reproducibility will be defined here as the agreement of results obtained from the analysis of incurred samples on two (or more) separate occasions within the same laboratory.

Generally, bioanalytical method validation studies are conducted using standards and quality control (QC) samples which are prepared to be as similar as possible to the study samples which are to be analyzed. However, there are a variety of circumstances in which the performance of the bioanalytical method when using standards and QCs may not adequately approximate that when using incurred samples. These may include, among others, conversion of unstable metabolites to parent, protein-binding differences in patient samples, recovery issues, sample inhomogeneity, and matrix effects (2). Such considerations, along with instances of poor reproducibility observed by FDA inspectors, gave rise to the recommendation that laboratories now routinely perform ISR testing.

For preclinical toxicology studies conducted to good laboratory practice (GLP) standards, it is proposed that ISR testing be performed once for each animal species. Recognizing that the likelihood of incurred sample irreproducibility is greater in humans than in animals, it is further proposed that ISR testing be performed in multiple clinical studies. Specific

¹ Preclinical and Research Biostatistics, sanofi-aventis, 200 Crossing Blvd. P.O. Box 6890 Mailcode BX2-416A Bridgewater, New Jersey 08807, USA.

² To whom correspondence should be addressed. (e-mail: david.hoffman@sanofi-aventis.com)

studies in which to conduct ISR testing may depend on the known characteristics of the drug, its metabolism, and its clearance, but will generally include all bioequivalence studies, and may also include first-in-man, proof-of-concept in patient populations, special population (e.g., renal or hepatic impairment), and drug–drug interaction studies.

While the Third AAPS/FDA Bioanalytical Workshop conference report did not specifically address acceptance criteria for ISR testing, subsequent workshops and discussion have led to a widespread industry adoption of the common “4–6–X” (or “two-thirds”) rule, with $\pm 20\%$ acceptance limits for small molecules or $\pm 30\%$ acceptance limits for large molecules (4). That is, at least 66.7% of the re-analyzed incurred samples must agree within $\pm 20\%$ of the original result (or $\pm 30\%$ for large molecules). It is noted that a recent workshop report (4) recommends that the re-analyzed incurred samples should agree within $\pm 20\%$ of the mean of the original and repeat results, rather than within $\pm 20\%$ of the original result. The rationale for assessing agreement relative to the mean result, rather than to the original result, is unclear. Such a practice will tend to attenuate the lack of agreement when the true relative bias between the repeat and original results is positive, and exaggerate the lack of agreement when the true relative bias is negative (though there would be little impact if the true relative bias is negligible). Thus, for the remainder of this paper, the “4–6–X” rule will be assumed to apply to the percent differences calculated by (repeat – original)/original $\times 100\%$.

The acceptance limits of $\pm 20\%$ were likely chosen with reference to the acceptance criteria for in-study monitoring contained in the FDA bioanalytical guidance (3) that at least 66.7% of QC samples must be within $\pm 15\%$ of their respective nominal concentration. The expansion from $\pm 15\%$ acceptance limits to $\pm 20\%$ acceptance limits is an apparent attempt to account for the variability in the original result. For convenience, this approach (i.e., “4–6–X” rule with $\pm 20\%$ acceptance limits) will be referred to as the “4–6–20” rule throughout the remainder of this paper.

However, the deficiencies of ad-hoc approaches such as the 4–6–20 rule have been well-documented (5,6) and there has been no evaluation of the performance of such an approach in the context of ISR testing. Further, there has been little consideration of appropriate experimental designs for ISR testing, the impact of repeated ISR testing over multiple clinical studies, or the use of rigorous statistical methodology for evaluating incurred sample reproducibility. The purpose of this paper is to address each of these issues in order to provide general guidance and recommendations on the design and analysis of ISR experiments.

EXPERIMENTAL DESIGN

ISR experiments will typically be conducted by selecting individual study samples which are representative of the drug's pharmacokinetic profile, and should generally include one or more samples near the peak of the profile and one or more samples near the end of the elimination phase. It has been further recommended that samples be selected from amongst several dosed subjects, rather than simply selecting the entire pharmacokinetic profile of relatively few subjects, due to the potential for inter-subject variability in matrix composition

(4,7). Another reasonable approach would be to select individual study samples by random selection. Such a selection strategy could be easily performed by readily available software packages and would ensure a representative sampling of the individual subjects and pharmacokinetic profile. Regardless of the selection strategy employed, samples should be selected such that the dynamic range of the assay is covered and inter-subject variability is represented. Additionally, any sample selected for ISR testing must have sufficient volume to allow for the repeat analysis to be performed.

While the proper selection of individual samples to be included in ISR testing is an important consideration, it is also vital to consider the impact of the number of analytical runs over which the samples will be analyzed, as well as the total number of samples to include in the ISR experiment.

Number of Analytical Runs

All calculated concentrations obtained in the course of ISR testing will be subject to both within-run (intra-batch) and between-run (inter-batch) random variability intrinsic to the bioanalytical method. As a practical matter, calculated repeat and original concentrations for an incurred sample will be obtained in separate analytical runs. Differences in calculated repeat and original concentrations are thus partially (or wholly) comprised of differences in the random errors (both within-run and between-run) associated with each analytical run. However, the impact of the number of analytical runs and the relative magnitudes of between-run and within-run analytical variability in ISR testing has not been explored.

One simple scenario for performing an ISR test would be to obtain all original concentrations in a single analytical run and all repeat concentrations in a separate single analytical run. In this simple scenario, all original concentrations are correlated by a single between-run random error. Likewise, all repeat concentrations are correlated by a single between-run random error. While perhaps seemingly innocuous, such a scenario may have a profound impact on the assessment of incurred sample reproducibility, as differences in these between-run random errors will be indistinguishable from a true lack of reproducibility. That is, the assessment of incurred sample reproducibility in this scenario may simply reflect the between-run variability of the bioanalytical method rather than true non-reproducibility between the original and repeat results.

To illustrate the impact of the number of analytical runs in the context of ISR testing, a simulation study was performed. For simplicity, all original concentrations and all repeat concentrations were assumed to be obtained over an identical number of analytical runs (e.g., all original concentrations obtained over two analytical runs and all repeat concentrations obtained over two separate analytical runs). Calculated original and repeat concentrations were assumed to follow the statistical model (Model 1):

$$\begin{aligned} Y_{ij}^O &= \mu_i + b_j^O + \varepsilon_{ij}^O \\ Y_{ik}^R &= \mu_i + b_k^R + \varepsilon_{ik}^R \end{aligned}$$

where Y_{ij}^O is the original concentration for the i th ($i=1,2,\dots,N$) incurred sample and assayed in the j th ($j=1,2,\dots,J$) analytical run; Y_{ik}^R is the repeat concentration for the i th incurred sample and assayed in the k th ($k=1,2,\dots,K$) analytical run; μ_i is the

true (unknown) concentration for the i th incurred sample; b_j^O and b_k^R are the random errors for the j th and k th analytical runs, respectively; and ε_{ij}^O and ε_{ik}^R are the random errors for Y_{ij}^O and Y_{ik}^R , respectively. Without loss of generality, the true concentrations μ_i were assumed to be uniformly distributed on the range (0, 100).

The between-run random errors b_j^O and b_k^R were assumed to be independently and normally distributed with mean zero and variance σ_B^2 . The within-run random errors ε_{ij}^O and ε_{ik}^R were assumed to be independently and normally distributed with mean zero and variance σ_E^2 . These variances, σ_B^2 and σ_E^2 , correspond to the between-run and within-run variability of the bioanalytical method, respectively. The total analytical variability is then given by $\sigma_{TOT}^2 = \sigma_B^2 + \sigma_E^2$. The proportion of total variability due to between-run variability is given by $\rho = \sigma_B^2 / (\sigma_B^2 + \sigma_E^2)$.

Simulated original and repeat concentrations were assumed to follow the models given above. The number of incurred samples was fixed at $N=40$ samples. For simplicity, the true intermediate precision or total coefficient of variation (CV) for the bioanalytical method was assumed to be 12% and constant across the range of true concentrations μ_i . Thus, the between-run and within-run random errors for each simulated concentration were expressed relative to the true concentration. Various combinations of the number of original and repeat analytical runs ($J=1, 2, 4, 5, 8, 10$, or 20 runs with $K=J$) and proportion of total variability due to between-run variability ($\rho=0.00, 0.25$, or 0.50) were considered. For each combination of number of analytical runs and ρ , 10,000 datasets were simulated and the probability of failing an ISR test based on the 4–6–20 rule (i.e., less than 66.7% of repeat concentrations within $\pm 20\%$ of original concentration) was estimated. All simulations were performed using SAS (version 9.1) software.

Figure 1 gives the probability of failing an ISR test based on the 4–6–20 rule *versus* the number of analytical runs, for different values of ρ .

The results in Fig. 1 unsurprisingly show that when the true between-run variance is zero (i.e. $\rho=0$), the number of analytical runs has no impact on the probability of failing the 4–6–20 rule. The probability of failure is constant, regardless of the number

of analytical runs performed. However, most bioanalytical methods will typically exhibit substantial between-run variability and a value of $\rho=0$ is likely an unrealistic expectation. For realistic values of $\rho=0.25$ and 0.50, Fig. 1 shows a marked increase in the probability of failing the 4–6–20 rule when the number of analytical runs is small. As the number of analytical runs increase, the probability of failure decreases. Note that if the number of analytical runs were equal to the number of incurred samples (i.e., one incurred sample per analytical run), the probability of failing the 4–6–20 rule would be identical regardless of the value of ρ , though such an experimental design would rarely be feasible in practice.

Figure 1 clearly shows the impact of both the number of analytical runs and the relative magnitude of the between-run variability (ρ) on the probability of failing an ISR test based on the 4–6–20 rule. However, these are not the only factors which determine the probability of ISR test failure. Other relevant factors include the sample size (assumed to be 40 incurred samples in Fig. 1), the true total coefficient of variation (assumed to be 12% in Fig. 1), and the true relative bias between the original and repeat results (assumed to be 0% in Fig. 1). The combination of each of these factors will determine the probability of an ISR failure. Thus, specific requirements or recommendations regarding the appropriate number of analytical runs to perform will be influenced by the anticipated values of the other relevant factors. This could be accomplished via simulation techniques as illustrated in Fig. 1. Nevertheless, a general recommendation is to avoid analyzing all incurred samples in a single or relatively few analytical runs, as this is likely to entail an increase in the risk of ISR test failure. Incurred samples should be analyzed over as many analytical runs as practicable within a laboratory.

Number of Incurred Samples

The impact of the number of analytical runs and relative magnitude of the between-run variability (ρ) has been illustrated above. For simplicity, assume hereafter that the number of analytical runs has been chosen appropriately in order to render the impact of the between-run variability negligible (i.e., the number of analytical runs is sufficiently large). The primary experimental design issue remaining is to determine the number of incurred samples to include in the ISR experiment.

A common procedure for sample size selection in an ISR experiment is to use a fixed percentage, say 5–10%, of the total number of study samples (4). However, the number of incurred samples to include in an ISR experiment should ideally be chosen to yield small risks of incorrect decision-making (i.e. incorrectly rejecting a truly reproducible method or incorrectly accepting a truly non-reproducible method). As noted previously, non-statistical rules such as the 4–6–20 rule do not strictly control the risks of such incorrect decision-making. In order to evaluate the risks of incorrect decision-making with the 4–6–20 rule, it is first necessary to define the performance characteristics (i.e., true precision and relative bias) of truly reproducible bioanalytical methods.

Defining Truly Reproducible Bioanalytical Methods

One obvious definition for performance characteristics which constitute a truly reproducible bioanalytical method

Fig. 1. Probability of failing ISR test *versus* number of analytical runs, for various $\rho = \sigma_B^2 / (\sigma_B^2 + \sigma_E^2)$. Sample size is 40 incurred samples. True total CV is 12% and true relative bias is 0%. Acceptance criteria based on the 4–6–20 rule

can be derived from the 4–6–20 rule itself. Under this definition, a bioanalytical method is truly reproducible if the true proportion of incurred sample repeat concentrations which will be within $\pm 20\%$ of the original concentration is at least 66.7%. However, such a definition is inconsistent with the current rule used for in-study monitoring of QC samples (i.e., 4–6–15 rule). That is, the performance characteristics required to satisfy the 4–6–20 rule for incurred samples are inconsistent with those required to satisfy the 4–6–15 rule for QC samples.

Figure 2 illustrates the acceptance regions (i.e., combinations of true precision and relative bias) defined by the 4–6–15 rule for QC samples and the 4–6–20 rule for incurred samples.

Based on the 4–6–15 rule for QC samples, truly reproducible bioanalytical methods have true total coefficients of variation and relative biases that lie within the dashed curve in Fig. 2. These are methods such that the true proportion of QC sample concentrations which will be within $\pm 15\%$ of the nominal value is at least 66.7%. Likewise, truly non-reproducible methods have true total coefficients of variation and relative biases which lie on or outside of the dashed curve in Fig. 2.

However, based on the 4–6–20 rule for incurred samples, truly reproducible bioanalytical methods have true total coefficients of variation and relative biases that lie within the solid curve in Fig. 2. These are methods such that the true proportion of incurred sample repeat concentrations which will be within $\pm 20\%$ of the original concentration is at least 66.7%. Truly non-reproducible methods have true total coefficients of variation and relative biases which lie on or outside of the solid curve in Fig. 2.

Note that the acceptance regions defined by the 4–6–15 rule for QC samples and 4–6–20 rule for incurred samples do not coincide. For example, consider a bioanalytical method with true relative bias of 0%. Under the 4–6–15 rule for QC samples, such a method is truly reproducible (i.e., true

proportion of QC sample concentrations within $\pm 15\%$ of nominal value is at least 66.7%) so long as the true total coefficient of variation is less than (approximately) 15.5%. However, under the 4–6–20 rule for incurred samples, the method is truly reproducible (i.e. true proportion of incurred sample repeat concentrations within $\pm 20\%$ of original concentration is at least 66.7%) only if the true total coefficient of variation is less than (approximately) 14.6%. Though this difference in acceptance regions is fairly small, the effect is to require bioanalytical methods to quantify incurred samples with greater precision than that required for QC samples. To avoid such inconsistencies, truly reproducible bioanalytical methods will be defined here as those methods with true total coefficient of variation and relative bias that lie within the dashed curve in Fig. 2. Truly non-reproducible methods will be defined as those methods with true total coefficient of variation and relative bias that lie on or outside the dashed curve in Fig. 2.

Performance of 4–6–20 Rule

Having defined above the performance characteristics which constitute truly reproducible and non-reproducible bioanalytical methods, a simulation study was conducted to evaluate the performance of the 4–6–20 rule in the context of ISR testing and provide some general guidance on the number of incurred samples to include in an ISR experiment.

Calculated original and repeat concentrations were assumed to follow the statistical model given previously, but with $\sigma_B^2 = 0$ to reflect the assumption that the number of analytical runs is sufficiently large to render the impact of the between-run variability negligible (Model 2):

$$\begin{aligned} Y_i^O &= \mu_i + \varepsilon_i^O \\ Y_i^R &= \mu_i + \varepsilon_i^R \end{aligned}$$

Fig. 2. Acceptance regions defined by 4–6–15 rule for QC samples and 4–6–20 rule for incurred samples. *Dashed curve* gives combinations of true relative bias and total CV such that the true proportion of QC sample concentrations within $\pm 15\%$ of nominal value is 66.7%. *Solid curve* gives combinations of true relative bias and total CV such that the true proportion of incurred sample repeat concentrations within $\pm 20\%$ of original concentration is 66.7%

As before, the true total coefficient of variation for the bioanalytical method was assumed to be constant across the range of true concentrations μ_i and within-run random errors for each simulated concentration were thus expressed relative to the true concentration. For simplicity, the true relative bias is fixed at 0% for all simulations (note that a non-zero relative bias would further increase the probabilities of failing an ISR test based on the 4–6–20 rule in the following simulation results). Various combinations of the number of incurred samples ($N=20$ to 160 incurred samples) and true total coefficient of variation (CV=10.0%, 11.0%, 12.0%, 15.5%, 17.5%, and 20.0%) were considered. Note that true coefficients of variation equal to 10.0%, 11.0%, and 12.0% correspond to truly reproducible methods, while true coefficients of variation equal to 15.5%, 17.5%, and 20.0% correspond to truly non-reproducible methods. For each combination of number of incurred samples and true total CV, 10,000 datasets were simulated and the probability of failing an ISR test based on the 4–6–20 rule was estimated. All simulations were performed using SAS (version 9.1) software.

Figure 3 gives the probability of failing an ISR test based on the 4–6–20 rules *versus* the number of incurred samples, for true total CV=15.5%, 17.5%, and 20.0%.

The true total coefficients of variation considered in Fig. 3 correspond to bioanalytical methods which are truly non-reproducible. Thus, it is desirable to have a high probability of rejecting such methods (i.e., the probability of ISR test failure should be high). The results in Fig. 3 indicate that when the true total CV is 20.0%, there is a high probability (> 90%) of ISR test failure with a sample size as small as 20 incurred samples. For a true total CV of 17.5%, the probability of ISR failure is approximately 80% with a sample size of 20 incurred samples and approximately 90% with 60 incurred samples. For a true total CV of 15.5%, the probability of ISR failure is less than 80% even with a sample size as large as 160 incurred samples.

Figure 4 gives the probability of failing an ISR test based on the 4–6–20 rules *versus* the number of incurred samples, for true total CV=10.0%, 11.0%, and 12.0%.

The true total coefficients of variation considered in Fig. 4 correspond to bioanalytical methods which are truly reproducible. Thus, it is desirable to have a low probability of rejecting such methods (i.e. the probability of ISR failure should be low). The results in Fig. 4 indicate that when the true total CV is 10.0%, the probability of ISR test failure is less than 1% with a sample size as small as 40 incurred samples. For a true total CV of 11.0%, the probability of ISR test failure is 2% with a sample size of 40 incurred samples and less than 1% with 60 incurred samples. For a true total CV of 12.0%, the probability of ISR test failure is approximately 4% with a sample size of 60 incurred samples and approximately 1% with 100 incurred samples.

While the 4–6–20 rule does not provide strict control over the risks of incorrect decision-making, the results in Figs. 3 and 4 may be used to provide some general guidance with regard to sample size selection in ISR experiments. For example, a sample size of approximately 40 incurred samples would be sufficient to provide a high probability (>90%) of correctly rejecting a non-reproducible bioanalytical method with true total CV of 20% (and true relative bias of 0%) and a low probability (<1%) of incorrectly rejecting a reproducible bioanalytical method with true total CV of 10% (and true relative bias of 0%). This may form a reasonable basis for sample size selection in routine ISR testing with the 4–6–20 rule.

Fig. 3. Probability of failing ISR test *versus* number of incurred samples, for true total CV=15.5%, 17.5%, and 20.0%. True relative bias is 0%. Acceptance criteria based on the 4–6–20 rule

Fig. 4. Probability of failing ISR test *versus* number of incurred samples, for true total CV=10.0%, 11.0%, and 12.0%. True relative bias is 0%. Acceptance criteria based on the 4–6–20 rule

MULTIPLE ISR TESTING

While ISR testing may be performed only once for each animal species during preclinical drug development, it is proposed that ISR testing be performed in multiple clinical studies. This multiple-testing requirement will impact the overall probabilities of incorrectly rejecting truly reproducible methods or incorrectly accepting truly non-reproducible methods with the 4–6–20 rule.

Assume that for a given bioanalytical method, all ISR experiments are independent of one another and that the true performance characteristics (i.e., precision and relative bias) of the method remain unchanged across all experiments. The assumption of independence will generally be satisfied in practice, though the true performance characteristics of a method may potentially be study-dependent (e.g., first-in-man *versus* drug–drug interaction studies). Nonetheless, these assumptions are made here merely to facilitate the following discussion of multiple ISR testing.

In the previous section, the probability of failing an ISR test as a function of the number of incurred samples and true total CV was explored. Now, consider the probability of ISR test failure across multiple ISR tests. For simplicity, further assume that each ISR test is performed with an identical number of incurred samples (so that the probability of an ISR test failure remains the same from one test to the next).

Let the probability of failing an ISR test be denoted by p . Then the probability of failing at least one ISR test among t independent ISR tests is simply given by:

$$\text{Prob}(\text{at least one ISR failure}) = 1 - (1 - p)^t$$

Clearly, the probability of observing at least one ISR test failure increases as the number of ISR tests (t) increases. To further illustrate the impact of multiple ISR testing, consider the simulated data described above in the previous section. The probabilities of failing an ISR test estimated from this simulated data can be utilized to determine the probability of at least one ISR test failure as a function of the total number of ISR tests. Let the sample size of each ISR test be fixed at

40 incurred samples, and consider up to $t=15$ total number of ISR tests.

Figure 5 gives the probability of failing at least one ISR test based on the 4–6–20 rule *versus* the total number of ISR tests, for true total CV=15.5%, 17.5%, and 20.0% (and true relative bias of 0%). Figure 6 gives the probability of failing at least one ISR test based on the 4–6–20 rule *versus* the total number of ISR tests, for true total CV=10.0%, 11.0%, and 12.0% (and true relative bias of 0%). As noted previously, the true total coefficients of variation considered in Fig. 5 correspond to bioanalytical methods which are truly non-reproducible, while those in Fig. 6 correspond to methods which are truly reproducible.

The results in Fig. 5 indicate that for bioanalytical methods with true total CV $\geq 17.5\%$, the probability of at least one ISR test failure is nearly 100% after as few as three ISR tests (based on a sample size of 40 incurred samples per test). For methods with true total CV=15.5%, the probability of at least one ISR test failure is greater than 99% after as few as five ISR tests. Thus, one consequence of the multiple ISR testing requirement is an increased probability of correctly rejecting a truly non-reproducible bioanalytical method (i.e., at least one ISR test will result in failure).

Figure 6 indicates that for bioanalytical methods with true total CV=12.0%, the probability of at least one ISR failure is greater than 20% after only three ISR tests and greater than 50% after nine ISR tests (based on a sample size of 40 incurred samples per ISR test). This illustrates another consequence of the multiple ISR testing requirement: an increased probability of incorrectly rejecting a truly reproducible bioanalytical method. For methods with true total CV $\leq 10\%$, this may be of somewhat lesser impact. Figure 6 indicates that for a method with true total CV=10%, the probability of at least one ISR failure is approximately 3% after 15 ISR tests.

The requirement of performing ISR tests on multiple clinical studies has a clear impact on the risks of incorrect decision-making. The magnitude of this impact increases as the number of ISR tests performed increases. Noting that the

Fig. 5. Probability of failing at least one ISR test *versus* number of ISR tests performed, for true total CV=15.5%, 17.5%, and 20.0%. True relative bias is 0%. Sample size is 40 incurred samples per ISR test. Acceptance criteria based on the 4–6–20 rule

Fig. 6. Probability of failing at least one ISR test *versus* number of ISR tests performed, for true total CV=10.0%, 11.0%, and 12.0%. True relative bias is 0%. Sample size is 40 incurred samples per ISR test. Acceptance criteria based on 4–6–20 rule

number of required ISR tests will generally be unknown at the beginning of a clinical development program and that the eventual number of ISR tests may be quite large, the potential impact of multiple ISR testing should not be disregarded or ignored.

One possible approach for dealing with the potential impact of multiple ISR testing may be to perform an additional “confirmatory” ISR test subsequent to any ISR test failure. The objective of such a “confirmatory” test would be to provide assurance that the observed ISR test failure is indicative of true bioanalytical method non-reproducibility, and data generated from the confirmatory test would be assessed independently of the initial (failing) ISR test. The confirmatory test would require acceptance criteria based on rigorous statistical methodology which controls the risk of incorrect decision-making. Two appropriate statistical approaches are described in the following section.

STATISTICAL APPROACHES

Two statistical approaches which are readily applicable to the problem of ISR assessment are (1) tolerance intervals and (2) containment proportions. Both approaches have previously been advocated or applied in the context of bioanalytical method validation (8,9). Unlike the 4–6–20 rule (or similar approaches), these statistical approaches provide strict control over the risk of incorrectly accepting truly non-reproducible bioanalytical methods.

Both the tolerance-interval and containment-proportion approaches described below are based on the assumption that the underlying data are independent and normally distributed, though minor departures from this assumption are generally of little practical consequence. Noting that the incurred samples selected should span a wide range of concentrations and that bioanalytical precision is generally proportional to the true concentration, it is suggested that the calculated repeat and original concentrations be log-transformed prior to application of the tolerance-interval and containment-proportion approaches described below. In practice, the differences

(repeat – original) in log-transformed concentrations will likely approximate a normal distribution, though gross departures from this assumption can be assessed via graphical techniques or statistical hypothesis tests. Further, the assumption of independence may be reasonably satisfied by appropriate choice for the number of analytical runs as described earlier. Gross departures from the assumption of independence may inflate the risk of ISR test failure.

For the remainder of this section, the following notation will be used:

$$\begin{aligned}\Delta_i &= \log(Y_i^R) - \log(Y_i^O) \\ \bar{\Delta} &= \frac{1}{N} \sum_{i=1}^N \Delta_i \\ \hat{\sigma}_{\Delta}^2 &= \frac{1}{N-1} \sum_{i=1}^N (\Delta_i - \bar{\Delta})^2\end{aligned}$$

where Δ_i is the (repeat – original) difference in log-transformed concentration for the i th ($i=1,2,\dots,N$) incurred sample, $\bar{\Delta}$ is the mean of the differences in log-transformed concentration, and $\hat{\sigma}_{\Delta}^2$ is the variance of the differences in log-transformed concentration.

Tolerance Interval

A two-sided β -content, γ -confidence tolerance interval is a statistical interval (L, U) such that at least a proportion β of a population will lie within the interval with $\gamma\%$ confidence. Two-sided β -content, γ -confidence tolerance intervals provide lower (L) and upper (U) limits such that a specified proportion β of measurements will lie within the interval (L, U), with specified confidence coefficient γ .

In the context of incurred sample reproducibility, a two-sided β -content, γ -confidence tolerance interval may be used to determine an interval (L, U) such that a proportion β of the (repeat – original) measurement differences lie within the interval, with a specified confidence coefficient γ . This interval (L, U) can then be compared to appropriately chosen acceptance limits (A, B). Such an approach provides a statistical framework for controlling the risk of incorrectly accepting bioanalytical methods for which less than a proportion β of the (repeat – original) measurement differences lie within acceptance limits (A, B).

A proposed tolerance-interval approach is as follows:

1. Construct a two-sided β -content tolerance interval (L, U) with desired confidence level γ (say, 90%).
2. Compare the interval (L, U) to the acceptance limits (A, B).
3. If (L, U) falls completely within (A, B), the ISR test is passed; otherwise, the ISR test is failed.

A two-sided β -content tolerance interval with confidence coefficient γ is given by (10,11):

$$\bar{\Delta} \pm Z_{(1+\beta)/2} \sqrt{1 + N^{-1}} \sqrt{(N-1) \hat{\sigma}_{\Delta}^2 / \chi_{N-1, 1-\gamma}^2}$$

where $Z_{(1+\beta)/2}$ is the upper $(1+\beta)/2$ quantile of the standard normal distribution and $\chi_{N-1, 1-\gamma}^2$ is the lower γ quantile of the chi-square distribution with $N-1$ degrees of freedom.

Note that this application of tolerance intervals has the structure of a statistical hypothesis test. The null hypothesis (H_0) is that less than a proportion β of the (repeat – original) measurement differences will fall within the acceptance limits (A, B). The alternative hypothesis (H_A) is that at least a proportion β of the (repeat – original) measurement differences will fall within (A, B). The tolerance-interval approach is to reject the null hypothesis (and therefore accept the bioanalytical method) if the two-sided β -content, γ confidence tolerance interval falls completely within the acceptance limits (A, B).

Implementation of a tolerance-interval approach requires appropriate choices of content level (β), confidence level (γ), and acceptance limits (A, B). For assessment of incurred sample reproducibility, 66.7% content and 90% confidence are logical choices. A rational choice for the acceptance limits may be derived from the current acceptance criteria for in-study monitoring of QC samples specified in the FDA bioanalytical guidance (i.e., 4–6–15 rule). Noting that the coefficient of variation for a difference (repeat – original) of measurements is larger than that for an individual measurement by a factor of $\sqrt{2}$, limits of $\pm(15\sqrt{2})\% = 21.2\%$ are suggested. For log-transformed data, this corresponds to acceptance limits of $\pm\log(1.212)$. These choices of content level and acceptance limits directly correspond to the definition of truly reproducible bioanalytical methods described previously and used throughout this paper. Thus, the proposed tolerance-interval approach consists of constructing a two-sided $\beta=66.7\%$ content, $\gamma=90\%$ confidence tolerance interval on the differences of log-transformed measurements. If the resulting tolerance limits are completely within the $\pm\log(1.212)$ acceptance limits, the ISR test is passed; otherwise, the ISR test is failed.

Unlike the 4–6–20 rule, the tolerance-interval approach proposed above strictly controls the risk of incorrectly accepting a truly non-reproducible method. Regardless of the sample size chosen, this risk is no greater than 5% (i.e., $\frac{(100-\gamma)\%}{2}$) for the tolerance-interval approach; in fact, the risk will generally be even less than 5%. This can be seen by noting that the confidence level of a two-sided β -content tolerance interval is a probability statement regarding the content level of the resulting interval (i.e., $\gamma\%$ of the intervals constructed in this manner will have content of at least β); it is not a probability statement regarding the content contained within the acceptance limits. If the two-sided tolerance interval falls completely within the acceptance limits, then at least a proportion β of (repeat – original) differences are contained within the acceptance limits with confidence γ . However, if the true proportion of (repeat – original) differences contained within the acceptance limits is β , it does not necessarily follow that a two-sided tolerance interval will fall within the acceptance limits with confidence γ . Thus, the tolerance-interval approach can be somewhat conservative (i.e., the risk of incorrectly accepting a truly non-reproducible method is less than $\frac{(100-\gamma)\%}{2}$).

While the risk of incorrectly accepting a truly non-reproducible method is strictly controlled, the risk of incorrectly rejecting a truly reproducible method with the tolerance-interval approach must be controlled by appropriate choice of sample size. A simulation study was conducted to provide general guidance on the number of incurred samples required for ISR

Fig. 7. Probability of failing ISR test *versus* number of incurred samples, for true total CV=10.0%, 11.0%, and 12.0%. True relative bias is 0%. Acceptance criteria based on tolerance-interval approach

experiments when the acceptance criteria are based on the proposed tolerance-interval approach.

Calculated original and repeat concentrations were simulated according to Model 2 given previously. Various combinations of the number of incurred samples ($N=60$ to 200 incurred samples) and true total coefficient of variation (CV=10.0%, 11.0%, and 12.0%) were considered. For each combination of number of incurred samples and true total CV, 10,000 datasets were simulated and the probability of failing an ISR test based on the tolerance-interval approach was estimated. All simulations were performed using SAS (version 9.1) software.

Figure 7 gives the probability of failing an ISR test based on the tolerance-interval approach *versus* the number of incurred samples, for true total CV=10.0%, 11.0%, and 12.0%.

The results in Fig. 7 clearly indicate that the tolerance-interval approach requires larger sample sizes than that needed for the 4–6–20 approach (as shown previously in Fig. 3). For a true total CV of 10%, the probability of ISR test failure is approximately 3% with a sample size of 100 incurred samples. For a true total CV of 11%, the probability of ISR test failure is approximately 5% with a sample size of 200 incurred samples. For a true total CV of 12%, the probability of ISR test failure is nearly 40% even with a sample size of 200 incurred samples. Thus, for truly reproducible bioanalytical methods with true total CV greater than 11%, the sample sizes necessary to ensure a low probability of ISR test failure may be prohibitively large.

Containment Proportion

The tolerance-interval approach above consists of constructing an interval which contains at least a proportion β of the (repeat – original) measurement differences, with specified confidence level, and comparing the interval to pre-defined acceptance limits. An alternative approach is to directly estimate the proportion of the (repeat – original) measurement differences which lie within the pre-defined acceptance limits, with specified confidence level, and com-

pare this proportion to β . That is, the proportion π of (repeat – original) measurement differences which are “contained” within the pre-defined acceptance limits is estimated, with specified confidence level, and compared to the required proportion β .

A proposed containment-proportion approach is as follows:

1. Calculate a one-sided lower confidence bound, π_L , for the proportion of (repeat – original) differences which are contained within the acceptance limits (A, B), with desired confidence level $1-\alpha$ (say, 95%)
2. Compare π_L to the required proportion β (say, 66.7%)
3. If $\pi_L \geq \beta$, the ISR test is passed; otherwise, it is failed.

A point estimate ($\hat{\pi}$) for the proportion contained within the interval (A, B) is given by:

$$\hat{\pi} = \Phi\left(\omega \frac{B - \bar{\Delta}}{\hat{\sigma}_{\Delta}}\right) - \Phi\left(\omega \frac{A - \bar{\Delta}}{\hat{\sigma}_{\Delta}}\right)$$

where $\omega = \sqrt{N/(N-1)}$ and $\Phi(\bullet)$ denotes the standard normal distribution function.

The variance ($\hat{\sigma}_{\hat{\pi}}^2$) of the point estimate $\hat{\pi}$ can be approximated by the following:

$$\hat{\sigma}_{\hat{\pi}}^2 = \left(\frac{1}{N-1}\right)(\varphi_A - \varphi_B)^2 + \left(\frac{\omega^2}{2(N-1)\hat{\sigma}_{\Delta}^2}\right)\left((B - \bar{\Delta})\varphi_B - (A - \bar{\Delta})\varphi_A\right)^2$$

where $\varphi_A = \varphi\left(\omega \frac{A - \bar{\Delta}}{\hat{\sigma}_{\Delta}}\right)$, $\varphi_B = \varphi\left(\omega \frac{B - \bar{\Delta}}{\hat{\sigma}_{\Delta}}\right)$, and $\varphi(\bullet)$ denotes the standard normal density function.

A $(1-\alpha)$ one-sided lower confidence bound (π_L) for the proportion contained within the interval (A, B) is then given by (12):

$$\pi_L = \left(\frac{1}{1+\eta}\right)\left\{\left(\hat{\pi} + 0.5\eta\right) - \sqrt{\eta\hat{\pi}(1-\hat{\pi}) + 0.25\eta^2}\right\}$$

Fig. 8. Probability of failing ISR test *versus* number of incurred samples, for true total CV=10.0%, 11.0%, and 12.0%. True relative bias is 0%. Acceptance criteria based on containment-proportion approach

where $\eta = \frac{Z_{1-\alpha}^2 \hat{\sigma}_\theta^2}{\hat{\pi}(1-\hat{\pi})}$ and $Z_{1-\alpha}$ is the upper $(1-\alpha)$ quantile of the standard normal distribution.

Note that the containment-proportion approach corresponds to the same statistical hypothesis test as the tolerance-interval approach. The null hypothesis is $H_0: \pi < \beta$, while the alternative hypothesis is $H_A: \pi \geq \beta$. The containment-proportion approach is to reject the null hypothesis (and therefore accept the bioanalytical method) if the $(1-\alpha)$ one-sided lower confidence bound π_L is greater than or equal to β .

Similar to the tolerance-interval approach, implementation of a containment-proportion approach requires appropriate choices of the required proportion (β), one-sided confidence level $(1-\alpha)$, and acceptance limits (A , B). For assessment of incurred sample reproducibility, 66.7% required proportion and 95% confidence are logical choices. As previously described above, acceptance limits of $\pm \log(1.212)$ for log-transformed data are proposed. Thus, the proposed containment-proportion approach consists of constructing a 95% one-sided lower confidence bound for the proportion of log-scale differences which are contained within the $\pm \log(1.212)$ acceptance limits. If the resulting lower confidence bound is greater than or equal to the required proportion $\beta=66.7\%$, the ISR test is passed; otherwise, the ISR test is failed.

Like the tolerance-interval approach, the containment-proportion approach proposed above strictly controls the risk of incorrectly accepting a truly non-reproducible method. Regardless of the sample size, this risk is no greater than 5% (i.e., $\alpha\%$). The risk of incorrectly rejecting a truly reproducible method with the containment-proportion approach must be controlled by appropriate choice of sample size. The simulation study described above for the tolerance-interval approach was also used to provide general guidance on the number of incurred samples required for ISR experiments

when the acceptance criteria are based on the proposed containment-proportion approach.

Figure 8 gives the probability of failing an ISR test based on the containment-proportion approach *versus* the number of incurred samples, for true total CV=10.0%, 11.0%, and 12.0%.

The results in Fig. 8 indicate that while the containment-proportion approach requires larger sample sizes than that needed for the 4–6–20 rule, the containment-proportion approach requires somewhat less than that needed for the tolerance-interval approach. For a true total CV of 10%, the probability of ISR test failure with the containment-proportion approach is approximately 3% with a sample size of 60 incurred samples. For a true total CV of 11%, the probability of ISR test failure is approximately 3% with a sample size of 120 incurred samples. For a true total CV of 12%, the probability of ISR test failure is nearly 15% with a sample size of 200 incurred samples. Thus, the containment-proportion approach yields a lower risk of incorrectly rejecting a truly reproducible method than does the tolerance-interval approach (i.e., the containment approach has greater power to correctly accept truly reproducible methods). This reflects the conservatism of the tolerance-interval approach noted previously.

EXAMPLES

The 4–6–20 rule, tolerance-interval, and containment-proportion approaches are each illustrated by application to data from two actual ISR experiments. The data for each example are calculated repeat and original concentrations (nanograms per milliliter) of incurred plasma samples analyzed by an LC-MS/MS method. The first example consists of incurred plasma samples taken from a clinical study conducted in healthy volunteers while the second example

Table I. Repeat and Original Concentrations (nanograms per milliliter) with Percentage Difference for Example No. 1

Sample	Original	Repeat	% Difference	Sample	Original	Repeat	% Difference
1	5.10	5.53	8.4	25	27.0	31.7	17.4
2	40.1	39.7	-1.0	26	99.5	107	7.5
3	49.6	46.7	-5.8	27	129	141	9.3
4	48.7	54.2	11.3	28	96.9	102	5.3
5	64.4	65.8	2.2	29	75.2	80.8	7.4
6	35.7	38.0	6.4	30	79.4	92.4	16.4
7	162	190	17.3	31	86.6	96.1	11.0
8	76.5	79.4	3.8	32	30.3	36.7	21.1
9	36.7	38.1	3.8	33	42.9	51.9	21.0
10	120	131	9.2	34	26.1	32.8	25.7
11	96.9	124	28.0	35	31.2	39.2	25.6
12	140	148	5.7	36	19.0	21.6	13.7
13	99.5	179	79.9	37	6.33	8.96	41.5
14	85.3	64.0	-25.0	38	2.68	3.68	37.3
15	170	189	11.2	39	1.51	1.82	20.5
16	59.0	77.8	31.9	40	61.2	69.0	12.7
17	79.4	82.3	3.7	41	92.6	111	19.9
18	49.6	58.3	17.5	42	66.5	68.4	2.9
19	45.0	45.9	2.0	43	98.0	108	10.2
20	9.47	11.4	20.4	44	63.7	55.6	-12.7
21	1.63	1.87	14.7	45	49.0	62.4	27.3
22	12.0	13.0	8.3	46	126	132	4.8
23	65.8	69.1	5.0	47	79.5	83.6	5.2
24	63.7	60.9	-4.4	48	22.2	25.3	14.0

Table II. Repeat and Original Concentrations (ng/mL) with Percentage Difference for Example No. 2

Sample	Original	Repeat	% Difference
1	4,030	4,070	1.0
2	14,100	12,600	-10.6
3	2,120	1,530	-27.8
4	21.6	19.9	-7.9
5	859	761	-11.4
6	192	215	12.0
7	2,710	2,790	3.0
8	13,000	10,000	-23.1
9	886	787	-11.2
10	9.46	9.01	-4.8
11	401	381	-5.0
12	123	122	-0.8
13	3,520	3,370	-4.3
14	3,430	3,410	-0.6
15	2,580	2,630	1.9
16	13.7	13.2	-3.6
17	1,410	1,320	-6.4
18	437	433	-0.9
19	3,240	3,250	0.3
20	11,000	12,200	10.9
21	879	799	-9.1
22	6.68	6.24	-6.6
23	384	313	-18.5
24	57.0	64.8	13.7
25	2,930	3,150	7.5
26	11,400	11,600	1.8
27	2,310	2,270	-1.7
28	6.11	6.22	1.8
29	881	900	2.2
30	193	169	-12.4
31	3,870	4,340	12.1
32	16,600	16,200	-2.4
33	1,250	1,180	-5.6
34	10.3	9.62	-6.6
35	581	673	15.8
36	196	188	-4.1

consists of incurred plasma samples taken from a preclinical toxicology study conducted in dogs.

Note that under the “confirmatory” testing strategy described previously, the statistical approaches would be applied only to data generated from a confirmatory ISR test (in the event of an initial ISR failure based on the 4–6–20 rule). However, for illustration purposes, each approach (4–6–20 rule, tolerance-interval, and containment proportion) is applied to the examples below.

Example 1

Table I gives the repeat and original concentrations, as well as the simple percentage difference calculated by (repeat – original)/original × 100%.

Note that 35 of the 48 repeat concentrations (72.9%) are within ±20% of the original concentration. Thus, the ISR test passes the 4–6–20 rule acceptance criteria.

To apply the tolerance-interval approach, the repeat and original concentrations are log-transformed. After log-transformation, the following statistics can be calculated: $\bar{\Delta} = 0.11274$ and $\hat{\sigma}_{\Delta}^2 = 0.01722$. The appropriate standard normal and chi-square quantiles can be easily obtained from tabulated values or a statistical software package, and are as follows: $Z_{0.8335} = 0.96809$ and $\chi_{47,0.10}^2 = 35.0814$. A two-sided $\beta = 66.7\%$ content, $\gamma = 90\%$ confidence tolerance interval is then given by:

$$0.11274 \pm 0.96809 \sqrt{1 + 48^{-1}} \sqrt{(47)0.01722/35.0814}$$

The resulting two-sided tolerance interval is given by (–0.0358, 0.2613). The interval is not entirely contained within the acceptance limits $\pm \log(1.212) = \pm 0.1923$. Thus, the ISR test fails the tolerance-interval acceptance criteria.

To apply the containment-proportion approach, the following additional statistics are calculated: $\hat{\pi} = 0.7205$ and $\hat{\sigma}_{\hat{\pi}}^2 = 0.002715$. The appropriate standard normal quantile is given by $Z_{0.90} = 1.6448$, which gives $\eta = 0.03647$. A 95% one-sided lower confidence bound for the proportion of log-scale differences contained within the $\pm \log(1.212)$ acceptance limits is then given by:

$$\left(\frac{1}{1.03647}\right) \left\{ (0.7205 + 0.5(0.03647) - \sqrt{0.03647(0.7205)(1 - 0.7205) + 0.25(0.03647^2)}) \right\}$$

The resulting one-sided lower confidence bound is given by 0.6282 (or 62.82%). The lower bound is not greater than the required proportion of 66.7%. Thus, the ISR test fails the containment proportion acceptance criteria.

Example 2

Table II gives the repeat and original concentrations, as well as the simple percentage difference calculated by (repeat – original)/original × 100%.

Note that 34 of the 36 repeat concentrations (94.4%) are within ±20% of the original concentration. Thus, the ISR test passes the 4–6–20 rule acceptance criteria.

To apply the tolerance-interval approach, the repeat and original concentrations are log-transformed. After log-transformation, the following statistics can be calculated: $\bar{\Delta} = -0.03350$ and $\hat{\sigma}_{\Delta}^2 = 0.01034$. The appropriate standard normal and chi-square quantiles are as follows: $Z_{0.8335} = 0.96809$ and $\chi_{35,0.10}^2 = 24.7966$. A two-sided $\beta = 66.7\%$ content, $\gamma = 90\%$ confidence tolerance interval is then given by:

$$-0.03350 \pm 0.96809 \sqrt{1 + 36^{-1}} \sqrt{(35)0.01034/24.7966}$$

The resulting two-sided tolerance interval is given by (–0.1521, 0.0851). The interval is entirely contained within the acceptance limits $\pm \log(1.212) = \pm 0.1923$. Thus, the ISR test passes the tolerance-interval acceptance criteria.

To apply the containment-proportion approach, the following additional statistics are calculated: $\hat{\pi} = 0.9312$ and $\hat{\sigma}_{\hat{\pi}}^2 = 0.00110$. The appropriate standard normal quantile is given by $Z_{0.90} = 1.6448$, which gives $\eta = 0.04631$. A 95% one-sided lower confidence bound for the proportion of log-scale

differences contained within the $\pm \log(1.212)$ acceptance limits is then given by:

$$\left(\frac{1}{\sqrt{0.04631}}\right) \left\{ (0.9312 + 0.5(0.04631)) - \sqrt{0.04631(0.9312)(1 - 0.9312) + 0.25(0.04631^2)} \right\}$$

The resulting one-sided lower confidence bound is given by 0.8556 (or 85.56%). The lower bound is greater than the required proportion of 66.7%. Thus, the ISR test passes the containment proportion acceptance criteria.

CONCLUSIONS

Laboratories are now expected to demonstrate the reproducibility of bioanalytical methods using incurred samples. The Third AAPS/FDA Bioanalytical Workshop and subsequent workshops have lead to widespread industry adoption of an ISR-acceptance criteria based on the 4–6–20 rule. However, there has been little consideration of appropriate experimental designs for ISR testing, the performance of the 4–6–20 rule, the impact of repeated ISR testing, or the use of formal statistical methods for assessing reproducibility.

The number of analytical runs over which the incurred samples are analyzed can substantially impact the probability of failing an ISR test. This impact increases as the relative magnitude of the between-run analytical variability increases, and can lead to marked increases in the probability of ISR failure when the number of analytical runs is small. To avoid or mitigate this impact, incurred samples should be analyzed over as many analytical runs as practicable within a laboratory.

While the 4–6–20 rule does not provide strict control over the risks of incorrectly rejecting a truly reproducible method or incorrectly accepting a truly non-reproducible method, these risks of incorrect decision-making can be estimated as a function of the true bioanalytical performance characteristics and the number of incurred samples included in the ISR testing. A moderate sample size of 40 incurred samples is sufficient to yield a high probability (>90%) of correctly rejecting a truly non-reproducible method with true total CV of 20% (and true relative bias of 0%) and low probability (<1%) of incorrectly rejecting a truly reproducible method with true total CV of 10% (and true relative bias of 0%). The risks associated with other choices of sample size can be determined from the figures provided in this paper or from additional simulation studies.

The requirement of performing ISR tests in multiple clinical studies decreases the risk of incorrectly accepting a truly non-reproducible method but also increases the risk of incorrectly rejecting a truly reproducible method. As the number of ISR tests performed increases, the probability of observing at least one ISR failure increases accordingly. This can lead to a high probability of failing at least one ISR test even for truly reproducible methods. One simple approach for attempting to account for the impact of multiple ISR

testing is to perform a “confirmatory” ISR test subsequent to any ISR test failure. However, such a confirmatory test must apply formal statistical methods for assessing reproducibility in order to ensure that the risk of incorrectly accepting a truly non-reproducible method is strictly controlled.

Both tolerance-interval and containment-proportion approaches provide formal statistical frameworks for assessing incurred sample reproducibility. Each approach strictly controls the risk of incorrectly accepting non-reproducible methods, though the risk of incorrectly rejecting truly reproducible methods must be controlled by the choice of sample size. Either approach requires larger sample sizes than that required for the simple 4–6–20 rule, though the containment-proportion approach generally requires fewer samples than the tolerance-interval approach.

REFERENCES

1. Shah VP, Midha KK, Dighe SV, McGilveray IJ, Skelly JP, Yacobi A, *et al.* Analytical methods validation: bioavailability, bioequivalence, and pharmacokinetic studies. *Pharm Res.* 1992;9:588–92.
2. Viswanathan CT, Bansal S, Booth B, DeStefano AJ, Rose MJ, Sailstad J, *et al.* Workshop/conference report - quantitative bioanalytical methods validation and implementation: best practices for chromatographic and ligand binding assays. *AAPS J.* 2007;9(1):E30–42.
3. Food and Drug Administration. Draft guidance for industry: bioanalytical method validation. Rockville, MD: US Food and Drug Administration; 1999.
4. Fast D, Kelley M, Viswanathan CT, O’Shaughnessy J, King S, Chaudhary A, *et al.* Workshop report and follow-up—AAPS workshop on current topics in GLP bioanalysis: assay reproducibility for incurred samples—implications of Crystal City recommendations. *AAPS J.* 2009;. doi:10.1208/s12248-009-9100-9.
5. Kringle R. An assessment of the 4–6–20 rule for acceptance of analytical runs in bioavailability, bioequivalence, and pharmacokinetic studies. *Pharm Res.* 1994;11:556–60.
6. Kringle R, Hoffman D, Newton J, Burton R. Statistical methods for assessing stability of compounds in whole blood for clinical bioanalysis. *Drug Inf J.* 2001;35:1261–70.
7. Rocci M, Devanarayan V, Haughey D, Jardieu P. Confirmatory reanalysis of incurred bioanalytical samples. *AAPS J.* 2007;9(3):E336–43.
8. Hoffman D, Kringle R. A total error approach for the validation of quantitative analytical methods. *Pharm Res.* 2007;24:1157–64.
9. Boulanger B, Dewe W, Gilbert A, Govaerts B, Maumy-Bertrand M. Risk management for analytical methods based on the total error concept: conciliating the objectives of the pre-study and in-study validation phases. *Chemometr Intell Lab Syst.* 2007;86:198–207.
10. Wald A, Wolfowitz J. Tolerance limits for a normal distribution. *Ann Math Stat.* 1946;17:208–15.
11. Howe WG. Two-sided tolerance limits for normal distributions—some improvements. *J Am Stat Assoc.* 1969;64:610–20.
12. Mee R. Estimation of the percentage of a normal distribution lying outside a specified interval. *Commun Stat., Theory Methods.* 1988;17:1465–79.