Assessing the Reproducibility of an Analytical Method

Jason J.Z. Liao*, Robert C. Capen, and **Timothy L. Schofield**
Merck Research Laboratories, P.O. Box 4, WP37C-305, West Point, PA 19486

Abstract

The reproducibility of a validated analytical method may require reassessment because of various reasons, such as the transfer between laboratories or companies, changes in the instruments or software platforms (or both), or changes in critical reagents, among others. This paper is a demonstration of an assay bridging study in evaluating reproducibility. The approach is simple but very informative and offers many advantages over existing approaches.

Introduction

It is well known that selective and sensitive analytical methods for the quantitative evaluation of drugs and their metabolites (analytes) are critical for the successful conduct of preclinical and clinical pharmacology studies. A validated analytical method is often modified to suit the requirement of the laboratory performing the assay. This can occur in all phases of drug development. During the course of development, an analytical method may require changes to support specific studies, and, as a result, different strategies may be needed to demonstrate the satisfactory performance of the assay. These strategies are classified into three categories: full-validation, partial-validation, and cross-validation (1). A full-validation may be needed when developing and implementing an analytical method for the first time or for a new drug entity, or when adding new metabolites into an existing assay for quantitation. A cross-validation may be acceptable when two or more analytical methods are used to generate data within the same study or across different studies, or when the comparison is for the revised method to the original validated method. A cross-validation could also be acceptable when data are generated using different analytical techniques in different studies or generated at more than one site or laboratory. A partial-validation may be adequate if the validation is for the modified version of an already validated method. Typical method changes that fall into this category include (1): transfers between laboratories/analysts; changes in detection system, matrix within specimen (plasma to urine), sampling processing procedure, species within matrix (rat plasma to mouse plasma), and instrument or software platform (or both); or demonstrations of an analyte in the presence of concomitant medications/specific metabolites. Per the FDA guidelines (1), one of the fundamental parameters for these revalidations is the reproducibility of the method. In other words, agreement in the results obtained by the method before and after the change need to be confirmed.

Consider an assay bridging study, in which a new assay was developed to replace the current assay. Agreement was to be assessed by having each assay test a common sample set ranging in relative potency from 0.4 to 3.2. For this purpose, 32 paired samples across the entire selected potency range were tested. It was therefore important to know how the concordance of these two assays should be determined. Accepting the new assay as concordant with the current assay, when in fact they do not agree, would mostly likely lead to an incorrect decision regarding the disposition of batches tested by the new assay. This increases manufacturing costs and possibly puts the public health at risk. Wrongly concluding that the two assays disagree, when in fact they are concordant, simply means that the lab needs to re-optimize the new assay. Although this could be a time-consuming activity, it is more cost effective than needlessly investigating out-of-specification occurrences or recalling product from the market.

The agreement problem has a long history and can be traced back over 100 years to Pearson, who proposed the correlation coefficient to measure agreement. The existing approaches can be classified into three categories.

The first category is the hypothesis testing type approach such as the regression analysis by testing the departure from the perfect agreement (i.e., intercept = 0 and slope = 1). This type of approach depends on the residual variance, which can reject a reasonably good agreement when the residual errors are small but accept a poor agreement when the residual errors are large.

The second category is an index approach, such as the intraclass correlation coefficient, the concordance correlation coefficient, and an improved concordance correlation coefficient (4). In assessing agreement, both Lin (5) and Liao (4) assumed observations from a bivariate normal distribution with a fixed mean and constant covariance. However, the mean values at different potency levels in the assay bridging study were different, which is usually the case in real examples. Therefore, current indices on

* Author to whom correspondence should be addressed: email jason_liao@merck.com.

agreement are not appropriate here. Furthermore, any index is very sensitive to the data range. Usually, only one single index is unacceptable to measure the agreement (3). When an index indicates a poor agreement, there is no clue what went wrong. In addition, there are no guidelines for defining an acceptable range of values for an agreement index. For poor agreement results, people in practice would like to know what went wrong and what the biases [fixed or proportional (or both) biases] were. Any agreement index cannot answer these questions.

The third category is a graphical approach. Bland and Altman (2) proposed a mean-difference graphic plot that plots the difference against the mean of the two measurements along with the 95% confidence limits of the difference. This approach is a step in the right direction. It evaluates the agreement in each individual level; it is simple. However, this method is not appropriate when a mixture of fixed, proportional bias, and proportional error occurs (6).

When there is a fixed or proportional bias (or both) between these two measurements, the mean from these two measurements is not a good metric for the true value. In addition, the mean of the two measurements used in their approach is always a random variable even if one of the two measurements is a "gold" standard. The variance of the mean from the two measurements can be larger than the variance of the difference from these two measurements. The simple 95% confidence interval of the difference will not give much information about the concordance of two methods because the confidence limits will cover 95% of all the differences. When a poor agreement conclusion is reached, any bias, which is a very important practical issue, cannot be assessed directly by this approach.

Based on the previously mentioned arguments, any new method in evaluating the concordance should be very informative. When given a poor concordance, it should easily indicate what went wrong and what the biases [fixed or proportional (or both) biases] were so that a calibration can be implemented if needed. In the next section, an approach is described in detail through the data analysis from an assay bridging study to evaluate the reproducibility.

Assay Bridging Study

In practice, a simple measure of agreement for each individual pair is preferred. An obvious starting point is the difference between measurements for each pair. That is to say, we can judge the agreement of two measurement methods by deriving an agreement interval and then showing that the difference of paired measurements falls within the specified interval. In other words, an agreement interval (Δ) is defined, and a pair of measurements is claimed to be "in agreement" at a specified level if their difference is within the interval. This is similar to interpreting "agreement" as an "in-control" process, in which being "in-control" occurs if no observation falls outside the limits of a Shewhart control chart (7). That means that two measurement methods agree only if all the paired differences fall within the agreement interval (Δ). The graphical description of this approach is in Figure 1.

Consider the assay bridging example in the Introduction. A new assay was developed to replace the existing relative potency assay. The goal of this study was to assess the concordance of these two assays. For this purpose, 32 pairs of measurements from the two assays were chosen. The range for the current relative potency

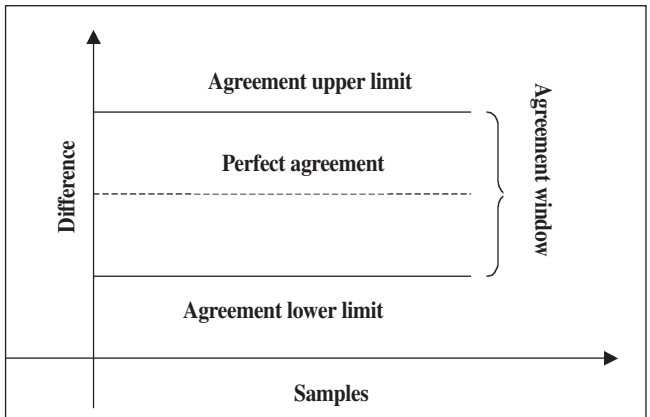

Figure 1. Graphical description of the quality control type approach.

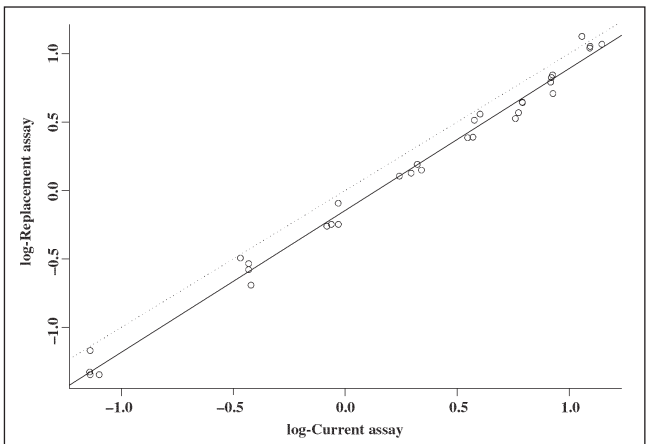

Figure 2. Agreement scatterplot of log(current) and log(replacement). Dotted line, perfect agreement line log(current) = log(replacement); solid line, regression line.

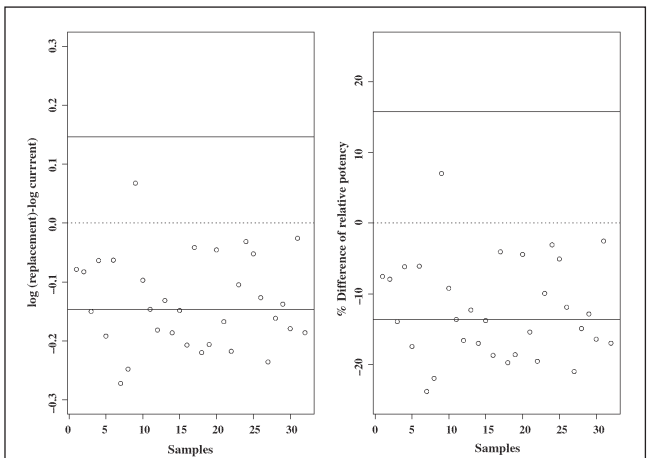

Figure 3. Concordance assessment. Dotted line, perfect agreement line; solid line, agreement limits. Left graph, difference in log-scale; right graph, %difference in raw scale.

assay was 0.4 to 3.2. Therefore, eight samples (0.4, 0.8, 1.2, 1.6, 2.0, 2.4, 2.8, and 3.2) were created that spanned this range. These eight samples were divided into four aliquots each and labeled as A, B, C, or D for a total of 32 samples. A single reportable result was generated for each of the 32 samples of both the current assay and the replacement assay. The results were paired together for analysis. Scatter plot of the data in a log-scale is shown in Figure 2, where the dotted line is the perfect agreement line.

To assess the agreement of two assays, the agreement window needs to be constructed. Using equations 2 and 3 (see Appendix A), $\hat{a}_0 = -0.145$, $\hat{b}_0 = 1.037$, and $\hat{\sigma} = 0.051 = 5.1\%$. Based on this determination of σ and 31 degrees of freedom, the values for lower and upper agreement limits, as defined in equation 1 of Appendix A, were calculated to be –0.1465 and 0.1465, respectively. That is, $\Delta = (-0.1465, 0.1465)$ in the log-scale and the agreement interval was (–13.627, 15.777) in terms of the percent potency difference in raw scale. To visually interpret the results, the arithmetic differences of log(*replacement*) – log(*current*) were plotted against the observational numbers in the left graph of Figure 3, and the percent potency differences were plotted against the observation numbers in the right graph of Figure 3. The figures indicate that there are a total of 17 of 32 values outside the agreement interval which correspond to samples 3A, 5A, 7A, 8A, 3B, 4B, 6B, 7B, 8B, 2C, 3C, 5C, 6C, 3D, 4D, 6D, and 8D. Therefore, the two assays did not show agreement.

Even though there are values outside the calculated agreement intervals, Figures 2 and 3 clearly indicate that there is a constant relative bias between the two assays. These figures all indicate that the replacement assay produces a lower reportable value than the current assay, with the exception of one sample, 1B. Please note that sample 1B is a kind of "outlier" relative to all remaining samples. However, it does not induce any bias between the two assays relative to that of the remaining samples. To truly assess the bias, all samples except sample 1B, were used again in equation 3 (Appendix A), which gives $\hat{a}_0 = -0.148$ and $\hat{b}_0 = 1.027$. The log-bias can be best represented by the equation $-0.148 + 0.027 \times X$, where X represents the log-value for the current assay. Because the log-value of interest for the current assay ranges approximately –1.0 to 1.0, as indicated in Figure 2, the proportional log-bias $0.027 \times X$ is relatively small compared with the fixed log-bias of –0.148. For example, when log-value for the current assay = –1 (the raw reportable value 0.37), the predicted log-value for the replacement assay is $-0.148 + 1.027 \times (-1) = -1.175$ (raw value is 0.31, thus, %bias = –16.2%). When the log-value for the current assay is 1 (the raw scale reportable value is 2.72), the predicted log-value for the replacement assay is 0.879 (the raw value is 2.41, thus, %bias = –11.4%). If the proportional bias is ignored, the percent bias is estimated to be $100 \times [\exp(-0.148) - 1]\% = -13.8\%$ across the entire range of raw reportable values. As a result, the relative bias between the two assays can be considered fixed and corrected with a constant percent offset value. In terms of the relative percent difference, the replacement assay gives a value lower than that of the current assay by 13.8%. In other words, to correct for this offset, multiply the replacement assay by exp(0.148) = 1.16. Using equation 9 (Appendix A), the 95% confidence interval of the fixed bias is (–0.156, –0.140) in terms of the arithmetic difference in the log-scale (13.07, 14.40) in terms of the relative percent potency difference in the raw scale, and (1.15, 1.17) in terms of the multiply factor.

Conclusion

In this paper, a simple but informative method was detailed in the data analysis for an assay bridging study to assess concordance. This approach is similar in spirit to the concept of determining if a process is in control through the use of a Shewhart control chart. The proposed method for assessing concordance overcomes the drawbacks and offers many improvements over Bland and Altman's (2) graphical approach. It uses an interval to evaluate each individual difference and can easily catch any existing bias (fixed, proportional, or both). Given a poor concordance conclusion, the proposed approach clearly indicates what type of bias exists and provides a way to estimate it for calibration purposes, which is the most important issue for practitioners.

The approach in this paper has other advantages as well. For example, in an animal potency assay transfer (or bridging) study, the ED_{50} obtained from a Probit analysis is often used to assess the agreement between two laboratories (assays). However, this approach generally requires a large number of animals. Using the approach in this paper, the original observations [e.g., optical densities (OD)] can be used instead to assess the agreement of the two laboratories (assays). Therefore, the number of animals required to evaluate concordance is greatly reduced because many more OD measurements are obtained per assay run.

Appendix A. Statistical Formula

Let $(X_i, Y_i); i = 1 \ldots n$, be the n pairs of observations, which might represent reportable values or transformed values such as the log-transformed values. Both X and Y contain measurement error. The agreement interval (Δ) is defined as follows:

$$\Delta = (-t_{1-\alpha/2,n-1} \times \sqrt{2}\hat{\sigma}, + t_{1-\alpha/2,n-1} \times \sqrt{2}\hat{\sigma}) \qquad \text{Eq. 1}$$

where $t_{1-\alpha/2,n-1}$ is the $100 \times (1 - \alpha/2)$th quantile of a t-distribution with degrees of freedom $n - 1$ and:

$$\hat{\sigma}^2 = \frac{1}{n}\left(S_{xx} - \frac{S_{xy}}{\hat{b}_0}\right) \qquad \text{Eq. 2}$$

$$\hat{b}_0 = \frac{-(S_{xx} - S_{xy}) + \sqrt{(S_{xx} - S_{xy})^2 + 4S_{xy}^2}}{2S_{xy}}, \hat{a}_0 = \overline{Y} - \hat{b}_0 \times \overline{X} \qquad \text{Eq. 3}$$

and

$$S_{xx} = \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 \qquad \text{Eq. 4}$$

$$S_{xy} = \sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right) \qquad \text{Eq. 5}$$

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X \qquad \text{Eq. 6}$$

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \qquad \text{Eq. 7}$$

If the two measurement methods are discordant, then bias information may be needed to calibrate the two methods. the bias can be assessed by:

$$\hat{a}_0 + (\hat{b}_0 - 1) \times X \qquad \text{Eq. 8}$$

where a_0 is the fixed bias and $(b_0 - 1) \times X$ is the proportional bias. The $100 \times (1 - \gamma)\%$ confidence interval for the fixed bias is:

$$\hat{a}_0 - \frac{z_{\gamma/2}\hat{\sigma}_a}{\sqrt{n}} \le a \le \hat{a}_0 + \frac{z_{\gamma/2}\hat{\sigma}_a}{\sqrt{n}} \qquad \text{Eq. 9}$$

where $z_{\gamma/2}$ is the $100 \times (\gamma/2)$th quantile of a standardized normal distribution

$$\hat{\sigma}_a^2 = \frac{\hat{\sigma}^2}{n} + \overline{X}^2\hat{\sigma}_b^2 \qquad \text{Eq. 10}$$

$$\hat{\sigma}_b^2 = \frac{\left(1 + \hat{b}_0^2\right)\left(S_{xx}S_{yy} - S_{xy}^2\right)}{\left(S_{xx} - S_{yy}\right)^2 + 4S_{xy}^2} \qquad \text{Eq. 11}$$

$$S_{yy} = \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 \qquad \text{Eq. 12}$$

Similiarly, the $100 \times (1 - \gamma)\%$ confidence interval for the proportional bias is:

$$[(b_0^L - 1) \times X, (b_0^U - 1) \times X] \qquad \text{Eq. 13}$$

where

$$b_0^L = \hat{b}_0 - \frac{z_{\gamma/2}\hat{\sigma}_b}{\sqrt{n}} \qquad \text{Eq. 14}$$

and

$$b_0^U = \hat{b}_0 + \frac{z_{\gamma/2}\hat{\sigma}_b}{\sqrt{n}} \qquad \text{Eq. 15}$$

Acknowledgments

The authors thank the editor and three referees for their comments, which improved the presentation.

References

1. U.S. Department of Health and Human Services, Food and Drug Administration. *Guidance for Industry: Bioanalytical Method Validation.* FDA, Washington, D.C., 2001.
2. J.M. Bland and D.G. Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* **2:** 307–10 (1986).
3. R.A. Deyo, P. Diehr, and D.L. Patrick. Reproducibility and responsiveness of health status measures: statistics and strategies for evaluation. *Controlled Clin. Trials* **12:** 142S–58S (1991).
4. J.J.Z. Liao. An improved concordance correlation coeffcient. *Pharm. Statistics* **2(4):** 253–61 (2003).
5. L.I.-K. Lin. A concordance correlation coeffcient to evaluate reproducibility. *Biometrics* **45:** 255–68 (1989).
6. J. Ludbrook. Comparing methods of measurement. *Clin. Exp. Pharm. Physiol.* **24:** 193–203 (1997).
7. D.C. Montgomery. *Introduction to Statistical Quality Control,* 3rd ed. John Wiley & Sons, Inc., New York, NY, 1996.

Manuscript received June 14, 2005;
revision received January 9, 2006.