Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://amstat.tandfonline.com/loi/uasa20

Statistical Methods in Assessing Agreement

Lawrence Lin, A. S.Hedayat, Bikas Sinha and Min Yang
Lawrence Lin is Senior Research Scientist (Statistics), Baxter Healthcare International Corp. and Adjunct Professor of Statistics, Department of Mathematics, Statistics, and Computer Science, University of Illinois, Chicago, IL 60607 . A. S. Hedayat is Professor of Statistics, Department of Mathematics, Statistics, and Computer Science, University of Illinois, Chicago, IL 60607 . Bikas Sinha is Professor of Statistics, Indian Statistical Institute, Calcutta, India (Email: . Min Yang is Ph.D. student, Department of Mathematics, Statistics, and Computer Science, University of Illinois, Chicago, IL 60607 . This research was primarily sponsored by National Science Foundation grants DMS-9803596 and DMS-0103727, National Cancer Institute grant P01-CA48112-08, and National Institutes of Health (NIH) National Center for Complementary and Alterative Medicine grant P50-AT00155. The contents are solely the responsibility of the authors and do not necessarily represent the official views of NIH. The authors acknowledge numerous constructive comments from four anonymous referees and an associate editor that have been extremely helpful in clarifying the presentation of the ideas related to agreement measurements.

Available online: 31 Dec 2011

To cite this article: Lawrence Lin, A. S.Hedayat, Bikas Sinha and Min Yang (2002): Statistical Methods in Assessing Agreement, Journal of the American Statistical Association, 97:457, 257-270

To link to this article: http://dx.doi.org/10.1198/016214502753479392

Statistical Methods in Assessing Agreement: Models, Issues, and Tools

Lawrence LIN, A. S. HEDAYAT, Bikas SINHA, and Min YANG

Measurements of agreement are needed to assess the acceptability of a new or generic process, methodology, and formulation in areas of laboratory performance, instrument or assay validation, method comparisons, statistical process control, goodness of fit, and individual bioequivalence. In all of these areas, one needs measurements that capture a large proportion of data that are within a meaningful boundary from target values. Target values can be considered random (measured with error) or fixed (known), depending on the situation. Various meaningful measures to cope with such diverse and complex situations have become available only in the last decade. These measures often assume that the target values are random. This article reviews the literature and presents methodologies in terms of "coverage probability." In addition, analytical expressions are introduced for all of the aforementioned measurements when the target values are fixed and when the error structure is homogenous or heterogeneous (proportional to target values). This article compares the asymptotic power of accepting the agreement across all competing methods and discusses the pros and cons of each. Data when the target values are random or fixed are used for illustration. A SAS macro program to compute all of the proposed methods is available for download at http://www.uic.edu/~hedayat/.

KEY WORDS: Accuracy; Concordance correlation coefficient; Coverage probability; Mean squared deviation; Precision; Target values; Total deviation index.

1. INTRODUCTION

Measurements of *agreement* are needed to assess the acceptability of a new or generic process, methodology, and formulation in areas of laboratory performance, instrument or assay validation, method comparisons, statistical process control, goodness of fit, and individual bioequivalence. Examples include the agreement of laboratory measurements collected in various laboratories, the agreement of a newly developed method with a gold standard method, the agreement of manufacturing process measurements with specifications, the agreement of observed values with predicted values, and the agreement in bioavailability of a new or generic formulation with a commonly used formulation. In all of these areas, one needs measurements that capture a large proportion of data that are within a meaningful boundary from target values. Examples of target values include mean, gold standard, quality control specification, predicted, and common formulation values. Target values can be considered random (measured with error) or fixed (known), depending on the situation. There are also situations in which one is interested in comparing two methods without a designated gold standard or target values.

When the agreement measurements show evidence of lack of agreement, we need to address the sources of the deficiencies. When there is a disagreement between the two marginal distributions, the source is defined as constant and/or scale "shift," or lack of "accuracy." When there is a disagreement due to large within-sample variation, the source is defined as lack of "precision."

The question of assessment of agreement has received considerable attention in the literature. Cohen (1960, 1968) discussed this problem in the context of categorical data. Bland and Altman (1986) proposed a simple and meaningful graphical approach for assessing the agreement between two clinical measurements. In a series of articles, Lin (1989, 1992, 1997, 2000) and Lin and Torbeck (1998) examined this problem critically in the framework of method reproducibility and suggested a few measures and studied their properties.

In the context of bioequivalence, similar studies have been reported by Anderson and Hauck (1990), Sheiner (1992), Holder and Hsuan (1993), Schall and Luus (1993), Schall (1995), Schall and Williams (1996), and Lin (2000). In the context of goodness of fit, Vonesh, Chinchilli, and Pu (1996) and Vonesh and Chinchilli (1997) have modified Lin's approach for choosing models that have better agreement between the observed and the predicted values.

The article is organized as follows. In Section 2 we briefly discuss existing methods and add analytical solutions for these methods when target values are fixed. In Section 3 we propose methods in terms of coverage probability (CP). In Section 4 we compare the power of accepting the agreement among all competing methods, and in Section 5 we perform a simulation study to examine the finite-sample performance of the methods. In Section 6 we present two examples based on real data, one when the target values are random and one when the target values are fixed. We discuss general extension problems and future studies in Section 7, followed by a conclusion in Section 8.

Lawrence Lin is Senior Research Scientist (Statistics), Baxter Healthcare International Corp. and Adjunct Professor of Statistics, Department of Mathematics, Statistics, and Computer Science, University of Illinois, Chicago, IL 60607 (E-mail: *lawrence_lin@baxter.com*). A. S. Hedayat is Professor of Statistics, Department of Mathematics, Statistics, and Computer Science, University of Illinois, Chicago, IL 60607 (E-mail: *hedayat@uic.edu*). Bikas Sinha is Professor of Statistics, Indian Statistical Institute, Calcutta, India (E-mail: *bksinha@isical.ac.in*). Min Yang is Ph.D. student, Department of Mathematics, Statistics, and Computer Science, University of Illinois, Chicago, IL 60607 (E-mail: *yangmin@math.uic.edu*). This research was primarily sponsored by National Science Foundation grants DMS-9803596 and DMS-0103727, National Cancer Institute grant P01-CA48112-08, and National Institutes of Health (NIH) National Center for Complementary and Alterative Medicine grant P50-AT00155. The contents are solely the responsibility of the authors and do not necessarily represent the official views of NIH. The authors acknowledge numerous constructive comments from four anonymous referees and an associate editor that have been extremely helpful in clarifying the presentation of the ideas related to agreement measurements.

2. METHODS

2.1 Mean Squared Deviation

2.1.1 When the Target Values are Random. A meaningful statistic to measure the agreement of observations (Y) with their target values (X) has been the mean squared deviation (MSD). Let $D = Y - X$; then

$$\text{MSD} = \epsilon^2 = E(D^2) = E(Y - X)^2. \tag{1}$$

We assume that the joint distribution of Y and X has finite second moments with means μ_y and μ_x, variances σ_y^2 and σ_x^2, and covariance σ_{yx}. Next, note that (1) can be expressed as

$$\epsilon^2 = (\mu_y - \mu_x)^2 + \sigma_y^2 + \sigma_x^2 - 2\sigma_{yx}, \tag{2}$$

and the sample counterpart of MSD can be computed as

$$e^2 = (\bar{y} - \bar{x})^2 + s_y^2 + s_x^2 - 2s_{yx},$$

where $\bar{y}$, $\bar{x}$, s_y^2, s_x^2, and s_{yx} represent the usual sample based on n paired observations on (y, x) and each with divisor n.

The bootstrap method has been proposed for inference based on the MSD estimate for individual bioequivalence (Schall and Luus 1993). Lin (2000) showed that $W = \ln(e^2)$ has an asymptotic normal distribution with mean $\omega = \ln(\epsilon^2)$ and variance

$$\sigma_W^2 = \frac{2[1 - (\mu_x - \mu_y)^4/e^4]}{n-2}. \tag{3}$$

For lesser bias of estimating ω and σ_W^2, we use $\tilde{e}^2$ to estimate ϵ^2, where

$$\tilde{e}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - x_i)^2. \tag{4}$$

Throughout the article, we use $n-1$, $n-2$, or $n-3$ instead of n in the denominator for lesser bias with smaller sample size, based on results of the simulation study in Section 5.

2.1.2 When the Target Values are Fixed. When X is fixed, we consider $\{(y_i|x_i) \mid i = 1, \ldots, n\}$ as observations in a random sample from the model

$$Y = \beta_0 + \beta_1 X + e_Y. \tag{5}$$

Here e_Y is the residual error with mean 0 and variance σ_e^2. The familiar quantities b_0, b_1, s_e^2, and r should be used as the estimators for β_0, β_1, σ_e^2, and ρ.

We compute $E(Y - X)^2$ for each x_i by model (5) and then take the average across x_i. Therefore, the MSD becomes

$$\epsilon_{|X}{}^2 = (\mu_{y|\bar{x}} - \bar{x})^2 + s_x^2(1 - \beta_1)^2 + \sigma_e^2, \tag{6}$$

where $\mu_{Y|\bar{x}} = \beta_0 + \beta_1\bar{x}$. The MSD estimate remains the same as in the case where X is random, except that the variance of W becomes smaller. It can be shown that (3) is replaced by

$$\sigma_{W|X}^2 = \frac{2}{n-2}\left(1 - \frac{[(\mu_{y|\bar{x}} - \bar{x})^2 + s_x^2(1-\beta_1)^2]^2}{\epsilon_{|X}{}^4}\right). \tag{7}$$

For variance estimation, we proceed as usual.

2.2 Concordance Correlation Coefficient

2.2.1 When the Target Values are Random. Lin (1989, 1992) presented another measurement, the concordance correlation coefficient (CCC), for measuring agreement (reproducibility) between two methods. The MSD in (2) can be standardized as a correlation coefficient to yield the CCC denoted by ρ_c,

$$\rho_c = 1 - \frac{\epsilon^2}{\epsilon^2|\rho = 0} = \frac{2\sigma_{yx}}{\sigma_y^2 + \sigma_x^2 + (\mu_y - \mu_x)^2}.$$

Here $\frac{\epsilon^2}{\epsilon^2|\rho=0}$ is the ratio of the mean square of within-sample total deviation and the total deviation. The mean square of the within-sample total deviation contains the within-sample variance and the bias square. The mean square of the total deviation contains the largest possible variance among nonnegative correlated samples and the bias square. The CCC can be written as the product of the accuracy and the precision coefficients $\rho_c = \chi_a\rho$. The accuracy coefficient is $\chi_a = \frac{2}{\varpi + 1/\varpi + \upsilon^2}$, where $\upsilon^2 = \frac{(\mu_y - \mu_x)^2}{\sigma_y\sigma_x}$ and $\varpi = \frac{\sigma_y}{\sigma_x}$. Here the marginal distributions of Y and X are equal (i.e., both means and variances are equal) if and only if the accuracy coefficient is 1. The precision coefficient is the Pearson correlation coefficient.

The CCC translates the MSD into a correlation coefficient that measures the agreement along the identity line, in which a value of 1 represents a perfect agreement ($Y = X$), a value of -1 represents a perfect disagreement ($Y = -X$), and a value of 0 represents no agreement. The sample counterpart of CCC is given as

$$r_c = \frac{2rs_ys_x}{s_y^2 + s_x^2 + (\bar{y} - \bar{x})^2}.$$

Lin (1989) showed that $Z = .5\ln\left(\frac{1+r_c}{1-r_c}\right)$ has an asymptotic normal distribution with mean

$$\zeta = \tanh^{-1}(\rho_c) = .5\ln\left(\frac{1+\rho_c}{1-\rho_c}\right),$$

where $\tanh^{-1}(\cdot)$ is the inverse hyperbolic tangent function. Its variance is

$$\sigma_Z^2 = \frac{1}{n-2}\left[\frac{(1-\rho^2)\rho_c^2}{(1-\rho_c^2)\rho^2} + \frac{2\upsilon^2(1-\rho_c)\rho_c^3}{(1-\rho_c^2)^2\rho} - \frac{\upsilon^4\rho_c^4}{2(1-\rho_c^2)^2\rho^2}\right]. \tag{8}$$

2.2.2 When the Target Values are Fixed. When X is fixed, under (5), the CCC is computed in the same way as in the case when X is random. However, the variance of the Z transformation of the CCC estimate becomes smaller,

$$\begin{aligned}\sigma_{Z|X}^2 &= \frac{\rho_c^2(1-\rho^2)}{(n-2)\rho^2(1-\rho_c^2)^2}\left[\rho_c^2\upsilon^2\varpi + (\rho_c\rho\varpi - 1)^2 + \frac{1}{2}\rho_c^2\varpi^2(1-\rho^2)\right] \\ &= \frac{\rho_c^2(1-\rho^2)}{(n-2)\rho^2(1-\rho_c^2)^2}\left\{\frac{\rho_c^2[\beta_0 + (\beta_1 - 1)\bar{x}]^2}{s_x^2} + (\rho_c\beta_1 - 1)^2 + \frac{\rho_c^2\beta_1^2(1-\rho^2)}{2\rho^2}\right\}. \end{aligned} \tag{9}$$

2.3 Precision and Accuracy

When agreement measurements show evidence of lack of agreement, we need to address the sources of the deficiencies. It is important to know whether the deficiencies are coming from the large within-sample variation (imprecision) or from the shift in marginal distributions (inaccuracy). The former would become a variance reduction exercise, which typically is more cumbersome than the latter. The latter most likely is a calibration problem. The CCC has meaningful components of precision (ρ) and accuracy (χ_a). In a way, the precision squared is in the same scale as accuracy, where 0 represents no agreement and 1 represents perfect agreement. The inference on ρ when X is random has been routinely used for decades. However, the inference on ρ when X is fixed is not known to be addressed in the literature. To develop a large-sample inference for ρ, we can simply let $\upsilon = 0$, $\varpi = 1$, and $\rho_c = \rho$ in (8) and (9) when X is random and fixed. Thus asymptotic variance of the Z transformation of r is $\frac{1}{n-3}$ when X is random, and $\frac{1}{n-3}(1-\frac{\rho^2}{2})$ when X is fixed.

We now present the inference on χ_a. Let the accuracy estimate be

$$c_a = \frac{2}{\mathrm{u}^2 + \mathrm{v} + 1/\mathrm{v}},$$

where $\mathrm{u}^2 = \frac{(\bar{y}-\bar{x})^2}{s_y s_x}$ and $\mathrm{v} = \frac{s_y}{s_x}$. Then the logit function of c_a is $L = \ln(\frac{c_a}{1-c_a})$. The random variable L has an asymptotic normal distribution with mean

$$\Lambda = \ln\left(\frac{\chi_a}{1-\chi_a}\right)$$

and variance (Robieson 1999)

$$\sigma_L^2 = \frac{\chi_a^2 \upsilon^2(\varpi + 1/\varpi - 2\rho) + \chi_a^2(\varpi^2 + 1/\varpi^2 + 2\rho^2)/2 + (1+\rho^2)(\chi_a \upsilon^2 - 1)}{(n-2)(1-\chi_a)^2}.$$

When X is fixed (known), the foregoing variance estimate becomes smaller,

$$\sigma_{L|X}^2 = \frac{\upsilon^2 \varpi \chi_a^2(1-\rho^2) + (1-\varpi\chi_a)^2(1-\rho^4)/2}{(n-2)(1-\chi_a)^2} = \frac{[\beta_0 + (\beta_1 - 1)\bar{x}]^2 \chi_a^2(1-\rho^2)/s_x^2 + (1-\beta_1\chi_a/\rho)^2(1-\rho^4)/2}{(n-2)(1-\chi_a)^2}.$$

2.4 Total Deviation Index

An intuitively clear measurement of agreement is a measure that captures a large proportion of data within a predetermined boundary from target values. In other words, we want the probability of the absolute value of D less than the boundary, κ, to be large. For example, consider the agreement assessment between the digital instrument used at home and the manual instrument used in a hospital for measuring diastolic blood pressure. In this case, a widely acceptable criterion is that at least 90% of the digital observations must be within 10 mmHg measured from the manual instrument. There are two approaches to measure agreement. We can fix the predetermined κ value (10 mmHg in this example), compute the coverage probability π (CP), and compare this CP with the predetermined probability level (90% in this example). We can also fix the predetermined CP value, compute the κ value, and compare this κ value with the predetermined boundary (10 mmHg in this example). In Section 3 we present the method for estimating π for a given κ. In this section we present the method for estimating κ for a given π.

Assuming that the distribution of D is normal with mean $\mu_d = \mu_y - \mu_x$ and variance $\sigma_d^2 = \sigma_y^2 + \sigma_x^2 - 2\sigma_{yx}$, the proportion of the population with $|D|$ (absolute difference) less than κ, $\kappa > 0$, becomes

$$\pi = \Pr(D^2 < \kappa^2) = \chi^2\left[\kappa^2, 1, \frac{\mu_d^2}{\sigma_d^2}\right],$$

where $\chi^2(\cdot)$ is the cumulative noncentral chi-squared distribution with 1 degree of freedom and noncentrality parameter $\frac{\mu_d^2}{\sigma_d^2}$. This noncentrality parameter is the relative bias squared. The total deviation index (TDI) for measuring the boundary κ is defined as

$$\mathrm{TDI} = \sqrt{\chi^{2^{(-1)}}\left[\pi, 1, \frac{\mu_d^2}{\sigma_d^2}\right]}, \tag{10}$$

where $\chi^{2^{(-1)}}(\cdot)$ is the inverse function of $\chi^2(\cdot)$. Inference based on estimate of this TDI is intractable. According to Chebyshev's inequality, this probability has a lower bound of

$$\Pr(D^2 < \kappa^2) > 1 - \frac{\epsilon^2}{\kappa^2}.$$

Therefore, the lower bound of the κ^2 value is proportional to ϵ^2, the MSD, in (2). Lin (2000) suggested using the TDI to approximate the value of κ that yields $P(D^2 < \kappa^2) = \pi$, which is

$$\mathrm{TDI}_\pi = \kappa_\pi = \Phi^{-1}\left(1 - \frac{1-\pi}{2}\right)|\epsilon|, \tag{11}$$

where $\Phi^{-1}(\cdot)$ is the inverse cumulative normal distribution and $|\cdot|$ is the absolute value. This can be estimated by replacing ϵ in (11) with $\tilde{e}$ in (4). Lin (2000) showed that the approximation is good under the following conditions: $\pi = .75$ and $\frac{\mu_d^2}{\sigma_d^2} \le \frac{1}{2}$, $\pi = .80$ and $\frac{\mu_d^2}{\sigma_d^2} \le 8$, $\pi = .85$ and $\frac{\mu_d^2}{\sigma_d^2} \le 2$, $\pi = .90$ and $\frac{\mu_d^2}{\sigma_d^2} \le 1$, and $\pi = .99$ and $\frac{\mu_d^2}{\sigma_d^2} \le \frac{1}{2}$. We may then use the sample counterpart and perform statistical inference in the same way as shown in Section 2.1 on MSD through the asymptotic normality of $W = \ln(\tilde{e}^2)$ when X is either random or fixed. Note that when X is fixed, we would compute the κ_π^2 for each x_i by model (5) and take the average across x_i. The square root of this average is equivalent to (11).

The TDI_π is proportional to the square root of the MSD and is intuitively much clearer than the MSD. It is the 100π percentile of the absolute difference of paired observations. The TDI_π is similar in concept to the prediction limit, and its confidence limit is similar in concept to the tolerance limit for capturing individual observations. The difference is that the boundary is set to deviate from target values, instead of from the mean.

2.5 Intraclass Correlation Coefficient

The intraclass correlation coefficient (ICC) (Fisher 1925) in its original form is the ratio of between-sample variance and the total variance (between- and within-sample) to measure precision under the model of equal marginal distributions. However, when the marginal distributions are not equal (inaccuracy), the ICC captures the deviations and considers those as imprecision. In contrast, the Pearson correlation coefficient ignores the inaccuracy component, and the CCC could segregate inaccuracy from imprecision. The ICC is closely related to the CCC. The subtle difference is that the ICC value remains the same when some pairs of y_i and x_i are interchanged, whereas the CCC does not. Unlike the CCC, the ICC does not have meaningful components of accuracy and precision. For these reasons, we prefer to use CCC. However, we expect the performance of the ICC and the special forms that have evolved to be very similar to that of CCC.

3. COVERAGE PROBABILITY

In this section we present a method to compute π for a given κ. When the target values are considered random, Anderson and Hauck (1990) proposed using this method to assess individual bioequivalence. They used nonparametric counting and inference for the assessment. Schall (1995) later advocated such a proposal through the normal distribution, and proposed using bootstrap estimation for inference. From (11), one can also approximate π for a given κ. Here we have $\pi_\kappa \cong 1 - 2[1 - \Phi(\kappa/|\epsilon|)] = \chi^2(\kappa^2/\epsilon^2, 1)$. Like the TDI, the approximation is subject to the restriction of reasonable relative bias squared values. This section provides the exact parametric estimation through the normal distribution and provides the asymptotic variance analytically for inference when the target values are random or fixed.

3.1 When the Target Values are Random

Consider the problem of assessment of agreement when D is from the normal distribution with mean μ_d and variance σ_d^2. Then the coverage probability for a given κ, as discussed in Section 2.4, is

$$\begin{aligned} \mathrm{CP}_k = \pi_k &= \Pr(|Y - X| < \kappa) \\ &= \chi^2\left(\kappa^2, 1, \frac{\mu_d^2}{\sigma_d^2}\right) \\ &= \Phi\left[\frac{\kappa - \mu_d}{\sigma_d}\right] - \Phi\left[\frac{-\kappa - \mu_d}{\sigma_d}\right], \end{aligned} \tag{12}$$

where $\Phi(\cdot)$ is the cumulative normal distribution function. The estimates of μ_d and σ_d^2 are $\hat{\mu}_d = \bar{y} - \bar{x}$ and $s_d^2 = \frac{n}{n-3}(s_y^2 + s_x^2 - 2s_{yx})$. Furthermore, $\hat{\mu}_d$ and s_d^2 are independent. Consequently, an estimator of CP_κ can be taken to be

$$p_\kappa = \Phi\left[\frac{\kappa - \hat{\mu}_d}{s_d}\right] - \Phi\left[\frac{-\kappa - \hat{\mu}_d}{s_d}\right].$$

Next, we can use the first-order approximation to compute $E(p_\kappa)$, and $V(p_\kappa)$,

$$\begin{aligned} p_\kappa = {} & \Phi\left(\frac{\kappa - \mu_d}{\sigma_d}\right) - \frac{1}{\sigma_d}\phi\left(\frac{\kappa - \mu_d}{\sigma_d}\right)(\hat{\mu}_d - \mu_d) \\ & - \frac{\kappa - \mu_d}{\sigma_d^2}\phi\left(\frac{\kappa - \mu_d}{\sigma_d}\right)(s_d - \sigma_d) - \Phi\left(\frac{-\kappa - \mu_d}{\sigma_d}\right) \\ & + \frac{1}{\sigma_d}\phi\left(\frac{-\kappa - \mu_d}{\sigma_d}\right)(\hat{\mu}_d - \mu_d) \\ & - \frac{\kappa + \mu_d}{\sigma_d^2}\phi\left(\frac{-\kappa - \mu_d}{\sigma_d}\right)(s_d - \sigma_d) \\ & + O[(\hat{\mu}_d - \mu_d)^2] + O[(s_d - \sigma_d)^2] \\ & + O[(\hat{\mu}_d - \mu_d)(s_d - \sigma_d)], \end{aligned}$$

where $\phi(x)$ is the density function of standard normal distribution and $\lim_{x \to 0} \frac{O(x)}{x} < \infty$.

Therefore, it is clear that

$$E(p_\kappa) = \pi_\kappa + O\left(\frac{1}{n}\right),$$

and $V(p_\kappa)$ becomes

$$\begin{aligned} \sigma_p^2 = {} & \frac{1}{n-3}\left\{\left[\phi\left(\frac{-\kappa - \mu_d}{\sigma_d}\right) - \phi\left(\frac{\kappa - \mu_d}{\sigma_d}\right)\right]^2\right. \\ & \left. + \frac{1}{2}\left[\frac{\kappa - \mu_d}{\sigma_d}\phi\left(\frac{\kappa - \mu_d}{\sigma_d}\right) + \frac{\kappa + \mu_d}{\sigma_d}\phi\left(\frac{-\kappa - \mu_d}{\sigma_d}\right)\right]^2\right\} \\ & + O\left(\frac{1}{n^2}\right). \end{aligned} \tag{13}$$

Because CP_κ is bounded by 0 and 1, it is better to use the logit transformation for inference. Let $T = \ln\left(\frac{p_\kappa}{1 - p_\kappa}\right)$. Its asymptotic mean is $\tau = \ln\left(\frac{\pi_\kappa}{1 - \pi_\kappa}\right)$, and its asymptotic variance is $\sigma_T^2 = \frac{\sigma_p^2}{\pi_\kappa^2(1 - \pi_\kappa)^2}$.

3.2 When the Target Values are Fixed

Consider the problem of assessment of agreement when target values X are fixed. If in model (5) we further assume that e_Y has the normal distribution with mean 0 and variance σ_e^2, then the coverage probability of the ith observation is

$$\begin{aligned} \pi_{\kappa i} &= \Pr(|y_i - x_i| < \kappa) \\ &= \Phi\left[\frac{\kappa - \beta_0 - (\beta_1 - 1)x_i}{\sigma_e}\right] \\ &\quad - \Phi\left[\frac{-\kappa - \beta_0 - (\beta_1 - 1)x_i}{\sigma_e}\right]. \end{aligned} \tag{14}$$

We define the overall coverage probability as

$$\pi_{\kappa|X} = \frac{1}{n}\sum_{i=1}^{n} \pi_{\kappa i}. \tag{15}$$

Suppose that we have a random sample $\{(y_i, x_i) \mid i = 1, \ldots, n\}$ and that β_0, β_1, and σ_e^2 are estimated by b_0, b_1,

and s_e^2; then b_0 and b_1 are independent of s_e. Here we use $s_e^2 = \frac{n}{n-3}(1-r^2)s_y^2$. An estimate of $\pi_{\kappa i}$ is

$$p_{\kappa i} = \Phi\left[\frac{\kappa - b_0 - (b_1-1)x_i}{s_e}\right] - \Phi\left[\frac{-\kappa - b_0 - (b_1-1)x_i}{s_e}\right],$$

and an estimate of $\pi_{\kappa|X}$ is

$$p_{\kappa|X} = \frac{1}{n}\sum_{i=1}^{n} p_{\kappa i}.$$

By the same method as in Section 3.1, it can be shown that

$$E(p_{\kappa|X}) = \pi_{\kappa|X} + O\left(\frac{1}{n}\right),$$

and that the asymptotic variance of $p_{\kappa|X}$ is

$$\sigma^2_{p|X} = \frac{1}{n-3}\left\{\frac{C_0^2}{n^2} + \frac{(C_0\bar{x} - C_1)^2}{n^2 s_x^2} + \frac{C_2^2}{2n^2}\right\}. \tag{16}$$

Here

$$C_0 = \sum_{i=1}^{n}\left[\phi\left(\frac{-\kappa - \beta_0 - (\beta_1-1)x_i}{\sigma_e}\right) - \phi\left(\frac{\kappa - \beta_0 - (\beta_1-1)x_i}{\sigma_e}\right)\right],$$

$$C_1 = \sum_{i=1}^{n}\left[\phi\left(\frac{-\kappa - \beta_0 - (\beta_1-1)x_i}{\sigma_e}\right) - \phi\left(\frac{\kappa - \beta_0 - (\beta_1-1)x_i}{\sigma_e}\right)\right]x_i,$$

and

$$C_2 = \sum_{i=1}^{n}\left[\frac{-\kappa - \beta_0 - (\beta_1-1)x_i}{\sigma_e}\phi\left(\frac{-\kappa - \beta_0 - (\beta_1-1)x_i}{\sigma_e}\right) - \frac{\kappa - \beta_0 - (\beta_1-1)x_i}{\sigma_e}\phi\left(\frac{\kappa - \beta_0 - (\beta_1-1)x_i}{\sigma_e}\right)\right].$$

The logit transformation is again used for inference, and the corresponding expression is the same as in Section 3.1.

3.3 Sectional Coverage Probability

In certain situations, agreement requirements might be different for certain sections of an analytical range. For example, in measuring a clinical marker for cancer cells, there might be a threshold value for which certain therapies would be prescribed when the marker value exceeds the threshold value. Therefore, we might require stricter agreement criteria near the threshold window. Suppose that the analytical range in the interval of a to b can be classified into m sections based on the gold standard method. Then the section d_i becomes

$$d_i = a_i - a_{i-1},\ i = 1, 2, \ldots, m,$$

where $a_0 = a$, $a_i > a_{i-1}$, $a_m = b$, and a_i values are prespecified. For d_i, an acceptable form of agreement would mean that Y should *largely* lie inside an interval of length $2I_i$ centered around X. Under this assumption, we need to consider the sectional coverage probability. For this purpose, the integrated sectional coverage probability (ISCP) between Y and X over those intervals is defined as

$$\text{ISCP} = \pi_I = \frac{\sum_{i=1}^{m} d_i \Pr[|Y - X| < I_i]}{b-a}. \tag{17}$$

This coverage probability involves only the distribution of $D = Y - X$ for both random and fixed X, which for the purpose of this article is assumed to have a normal distribution with parameters μ_d and σ_d^2. Thus (17) becomes

$$\pi_I = \frac{\sum_{i=1}^{m}(a_i - a_{i-1})\left[\Phi\left(\frac{I_i-\mu_d}{\sigma_d}\right) - \Phi\left(\frac{-I_i-\mu_d}{\sigma_d}\right)\right]}{b-a}. \tag{18}$$

The estimator of π_I, denoted by p_I, is obtained by substituting $\hat{\mu}_d$ and s_d for μ_d and σ_d in (18). By the same method as in Section 3.1, it can be shown that p_I is a consistent estimator of π_I with asymptotic variance

$$\sigma^2_{p_I} = \frac{1}{n(b-a)^2}\left\{\left[\sum_{i=1}^{m}(a_i - a_{i-1})\left(\phi\left(\frac{I_i-\mu_d}{\sigma_d}\right) - \phi\left(\frac{I_i+\mu_d}{\sigma_d}\right)\right)\right]^2 + \frac{1}{2}\left[\sum_{i=1}^{m}(a_i - a_{i-1})\left(\frac{I_i-\mu_d}{\sigma_d}\phi\left(\frac{I_i-\mu_d}{\sigma_d}\right) + \frac{I_i+\mu_d}{\sigma_d}\phi\left(\frac{I_i+\mu_d}{\sigma_d}\right)\right)\right]^2\right\}. \tag{19}$$

Again, the logit transformation should be used for inference.

3.4 Proportional Error Case

In practice, both Y and X are usually positive-valued variables. However, $\frac{Y}{X}$ has a bounded variance. Here we evaluate the proportion change $(\frac{Y}{X})$ rather than absolute difference $(Y - X)$, because the error is proportional to the measurement. Let $100\theta\%$ represent the percent change between Y and X. A simplified approach in this case is to assume that $\ln Y$ and $\ln X$ have a bivariate normal distribution. Thus the probability that Y lies in the interval $\frac{X}{1+\theta}$ to $X(1+\theta)$ is given by

$$\begin{aligned}\text{CP}_\theta = \pi_\theta &= \Pr\left[\frac{X}{1+\theta} < Y < X(1+\theta)\right] \\ &= \Pr\left[\frac{1}{1+\theta} < \frac{Y}{X} < 1+\theta\right] \\ &= \Pr[|\ln Y - \ln X| < \ln(1+\theta)].\end{aligned}$$

Let $D = \ln Y - \ln X$ and $\kappa = \ln(1+\theta)$. Then all algorithms of the previous section can be applied to the logs to obtain CP_κ and TDI_π. Here the TDI_π is a percent,

$$\text{TDI}_\pi\% = 100\theta_\pi\% = 100(e^{\kappa_\pi} - 1)\%.$$

In this proportional error case, we should also compute the CCC and MSD from the log transformation of the data.

4. ASYMPTOTIC POWER

In this section we investigate the asymptotic power of accepting agreement among estimates of MSD, CCC, and CP. Inference based on TDI can be assessed through MSD. Therefore, the asymptotic power of the estimate of TDI is the same

as that of MSD. For comparison, we must establish a common agreement criteria. We say that Y and X are in agreement if $v^2 < v_0^2$, $\frac{1}{\varpi_0} < \varpi < \varpi_0$, and $\rho > \rho_0$ where v_0, ϖ_0, and ρ_0 are prespecified values. We refer to these as null values. We compute the chance of declaring agreement under the alternative values at $v^2 = v_1^2$, $\varpi = \varpi_1$, and $\rho = \rho_1$. We refer to these as power. In addition, let the null and alternative values of $\sigma_y\sigma_x$ be $\sigma_{y0}\sigma_{x0}$ and $\sigma_{y1}\sigma_{x1}$, and let $h = \frac{\sigma_{y0}\sigma_{x0}}{\sigma_{y1}\sigma_{x1}}$. Let the null and alternative values of $\mu_y - \mu_x$ be $\mu_{y0} - \mu_{x0}$ and $\mu_{y1} - \mu_{x1}$. Furthermore, let $sign_0$ be positive when $\mu_{y0} - \mu_{x0} \geq 0$ and negative otherwise, and let $sign_1$ be positive when $\mu_{y1} - \mu_{x1} \geq 0$ and negative otherwise.

4.1 When the Target Values are Random

Let the corresponding log MSD, Z value of CCC, and logit CP values evaluated at $v^2 = v_m^2$, $\varpi = \varpi_m$, and $\rho = \rho_m$, $m = 0, 1$, be

$$\omega_m = \ln(\epsilon_m^2),$$
$$\zeta_m = \ln\left(\frac{1+\rho_{cm}}{1-\rho_{cm}}\right),$$

and

$$\tau_m = \ln\left(\frac{\pi_{\kappa m}}{1-\pi_{\kappa m}}\right), \tag{20}$$

where

$$\epsilon_m^2 = \sigma_{ym}\sigma_{xm}(v_m^2 + \varpi_m + \frac{1}{\varpi_m} - 2\rho_m),$$
$$\rho_{cm} = \frac{2\rho_m}{v_m^2 + \varpi_m + 1/\varpi_m},$$
$$\pi_{km} = \Phi\left(\frac{k}{\sigma_{dm}} - \delta_m\right) - \Phi\left(-\frac{k}{\sigma_{dm}} - \delta_m\right),$$
$$\sigma_{dm}^2 = \sigma_{ym}\sigma_{xm}(\varpi_m + 1/\varpi_m - 2\rho_m),$$

and

$$\delta_m = sign_m \frac{v_m}{(\varpi_m + 1/\varpi_m - 2\rho_m)^{\frac{1}{2}}}.$$

Let

$$\gamma_m = 2\left[1 - \frac{v_m^4}{(v_m^2 + \varpi_m + 1/\varpi_m - 2\rho_m)^2}\right], \tag{21}$$

$$\eta_m = \frac{(1-\rho_m^2)\rho_{cm}^2}{(1-\rho_{cm}^2)\rho_m^2} + \frac{2v_m^2(1-\rho_{cm})\rho_{cm}^3}{(1-\rho_{cm}^2)^2\rho_m} - \frac{v_m^4\rho_{cm}^4}{2(1-\rho_{cm}^2)^2\rho_m^2}, \tag{22}$$

and

$$\psi_m = \frac{1}{\pi_{km}^2(1-\pi_{km})^2} \times \left\{\left[\phi\left(\frac{k}{\sigma_{dm}} - \delta_m\right) - \phi\left(-\frac{k}{\sigma_{dm}} - \delta_m\right)\right]^2 + \frac{1}{2}\left[\left(\frac{k}{\sigma_{dm}} - \delta_m\right)\phi\left(\frac{k}{\sigma_{dm}} - \delta_m\right) + \left(\frac{k}{\sigma_{dm}} + \delta_m\right)\phi\left(\frac{k}{\sigma_{dm}} + \delta_m\right)\right]^2\right\}. \tag{23}$$

The asymptotic variances of W, Z, and T statistics evaluated at v_m^2, ϖ_m, and ρ_m become

$$\sigma_{Wm}^2 = \frac{\gamma_m}{n-2},$$
$$\sigma_{Zm}^2 = \frac{\eta_m}{n-2},$$

and

$$\sigma_{Tm}^2 = \frac{\psi_m}{n-3}.$$

The regions for declaring agreement by using W, Z, and T at the α significance level are

$$W \leq \omega_0 - \Phi^{-1}(1-\alpha)\sigma_{W0},$$
$$Z \geq \zeta_0 + \Phi^{-1}(1-\alpha)\sigma_{Z0},$$

and

$$T \geq \tau_0 + \Phi^{-1}(1-\alpha)\sigma_{T0}.$$

The asymptotic powers of accepting agreement at $v^2 = v_1^2$, $\varpi = \varpi_1$, and $\rho = \rho_1$ by using W, Z, and T are

$$P_W = \Phi\left[\frac{\omega_0 - \omega_1 - \Phi^{-1}(1-\alpha)\sigma_{W0}}{\sigma_{W1}}\right],$$
$$P_Z = 1 - \Phi\left[\frac{\zeta_0 - \zeta_1 + \Phi^{-1}(1-\alpha)\sigma_{Z0}}{\sigma_{Z1}}\right],$$

and

$$P_T = 1 - \Phi\left[\frac{\tau_0 - \tau_1 + \Phi^{-1}(1-\alpha)\sigma_{T0}}{\sigma_{T1}}\right].$$

The required sample sizes yielding the same power $1-\beta$ by using W, Z, and T are

$$n_W = \left[\frac{\Phi^{-1}(1-\beta)\sqrt{\gamma_1} + \Phi^{-1}(1-\alpha)\sqrt{\gamma_0}}{\omega_0 - \omega_1}\right]^2 + 2,$$
$$n_Z = \left[\frac{\Phi^{-1}(1-\beta)\sqrt{\eta_1} - \Phi^{-1}(1-\alpha)\sqrt{\eta_0}}{\zeta_0 - \zeta_1}\right]^2 + 2,$$

and

$$n_T = \left[\frac{\Phi^{-1}(1-\beta)\sqrt{\psi_1} - \Phi^{-1}(1-\alpha)\sqrt{\psi_0}}{\tau_0 - \tau_1}\right]^2 + 3.$$

We do not show the asymptotic power and sample size for precision and accuracy here; one can easily follow the same principles for precision and accuracy.

4.2 When the Target Values are Fixed

When X is fixed, we simply use $\bar{x}$ and s_{xm}^2 in place of μ_{xm} and σ_{xm}^2. Equations (14), (7), (9), and (16) are evaluated at both the null and alternative values to obtain the fixed X equivalents of equations (20), (21), (22), and (23).

4.3 Comparisons of Asymptotic Powers

The above powers and sample sizes for TDI (MSD) and CP when X is random and for TDI when X is fixed depend on the value of $h = \frac{\sigma_{y0}\sigma_{x0}}{\sigma_{y1}\sigma_{x1}}$. To compare these powers, we assume that $h = 1$.

The power of CP depends on the choice of κ. We let $\kappa = 1.5\sigma_{d0}, 2\sigma_{d0}, 2.5\sigma_{d0}$ when X is random and $\kappa = \sigma_{e0}$, $1.5\sigma_{e0}, 2\sigma_{e0}$ when X is fixed. We let $v_0 = .15$, $\varpi_0 = 1.15$,

and $\rho_0 = .8, .9, .95, .99$ and evaluate the power at $\upsilon_1 = .05$ and .10, $\varpi_1 = 1.05$ and 1.10, and $\rho_1 = \tanh[\tanh^{-1}(\rho_0) + .1]$ and $\tanh[\tanh^{-1}(\rho_0) + .2)]$. The null hypotheses correspond to a 15% location shift per $(\sigma_y\sigma_x)^{1/2}$ and a 15% scale shift for various precision values. These null hypotheses are compared to alternatives with two levels of improved precision, location, and scale shifts. The ranges were chosen so that most power values are in the broad range of .2 to .99. When X is fixed, from model (5) we let $X = -1, 0$, and 1. We let Y have values equally distributed among the three levels of X values. We let $\beta_1 = \rho\varpi$, $\beta_0 = (\upsilon^2 s_x^2 \varpi)^{1/2}$, and $\sigma_e^2 = s_x^2\varpi^2(1-\rho^2)$ for the corresponding υ, ϖ, and ρ values in the null and alternative cases.

Table 1 presents the asymptotic powers among TDI, CCC, and CP when X is random and fixed for $n = 30$ and $\alpha = .05$. The results indicate that the asymptotic powers of TDI, $CP_{\kappa 1}$, $CP_{\kappa 2}$, and $CP_{\kappa 3}$ are in general quite similar whenever X is random or fixed. Therefore, the choice of κ has little impact on the power.

The asymptotic power of CCC is inferior to that of TDI (MSD) in all test cases when X is random or fixed. Presumably, the estimation for the denominator of the CCC increases the noise. Therefore, for statistical inferences, CP and TDI are always preferable to CCC, especially with normally distributed data. However, in the case of a higher correlation coefficient (.99) when X is fixed, the CCC has a similar power. Regardless of the fact that the power performance is not very appealing, the CCC, precision, and accuracy remain as useful descriptive tools.

5. SIMULATION

For statistical inference based on any of the foregoing agreement measurement estimates, we would replace the parameters with their sample counterparts in the respective variance expressions. To assess the asymptotic normality and power of the methods, we performed two Monte Carlo simulations for X when random and fixed. In each simulation, we examined two cases: one case representing the null hypothesis and the other representing the alternative hypothesis. For comparison, we selected the same cases as shown in Table 1. The simulations also attempted to cover those CP, CCC, precision, and accuracy values near their boundaries.

For the simulation when X is random, paired samples were generated from each of the following bivariate normal distribution cases:

1. The null hypothesis case with mean (.15, 0), variance (1.15, 1/1.15), and correlation ρ_0. Here $\upsilon_0 = .15$ and $\varpi_0 = 1.15$.
2. The alternative hypothesis case with mean (.1, 0), variance (1.1, 1/1.1), and correlation $\tanh[\tanh^{-1}(\rho_0 + .2)]$. Here $\upsilon_1 = .1$, $\varpi_1 = 1.1$, and $h = 1$.

For the simulation when X is fixed, we generated univariate normal samples, with equal sample size among the $X = -1, 0, 1$ values. We let $\beta_1 = \rho\varpi$, $\beta_0 = (\upsilon^2 s_x^2 \varpi)^{1/2}$, and $\sigma_e^2 = s_x^2\varpi^2(1-\rho^2)$ under (5) for the corresponding υ, ϖ, and ρ values in the following null and alternative cases:

1. The null hypothesis case with $\upsilon_0 = .15$, $\varpi_0 = 1.15$, and correlation ρ_0.
2. The alternative hypothesis case with $\upsilon_1 = .1$, $\varpi_1 = 1.1$, and correlation $\rho_1 = \tanh[\tanh^{-1}(\rho_0 + .2)]$.

In each of the these random and fixed cases, we let $\rho_0 = .95$ and .99. The samples generated correspond to $n = 15$, $n = 30$, and $n = 60$. We have a total of 24 situations: 2 cases by 2 levels of precision by 3 levels of sample size for X when random and when fixed. For each situation, 5,000 runs were performed. Tables 2–5 present the simulation results for cases where X is random and fixed. In all of these tables, the fourth column presents the theoretical value of precision, accuracy, CCC, TDI, and CP for the null and alternative hypotheses.

To assess the robustness of each agreement statistics estimate, in each run, we calculated estimates of the respective transformation of precision (Z), accuracy (logit), CCC (Z), TDI_9 (ln MSD), $CP_{\kappa 1}$, and $CP_{\kappa 3}$ (logit). The $\kappa 1$ and $\kappa 3$ values correspond with those in Table 1. The mean and standard deviation of each estimate based on 5,000 runs were computed. The respective antitransformation of each mean estimate is reported in the fifth column. Comparisons between the fourth and fifth columns were used to assess the robustness of the estimates. The standard deviation of each estimate is reported in the sixth column ("Std of estimate"). The mean of the standard deviation estimate of each estimate based on 5,000 runs was also computed. This is reported in the seventh column ("Mean of std"). Comparisons between the sixth and seventh columns were used to assess the robustness of the variance estimates.

To assess the asymptotic normality for the significance level (case 1) and power (case 2) of each estimate at $\alpha = .05$, for each run we computed the proportion of each estimate among 5,000 runs that falls into the respective rejection region (accepting agreement) in Section 4. These proportions are reported in the eighth column ("Proportions in reject region"). The proportions in case 1 represent the significance level while the proportions in case 2 represent the power of accepting case 2 against the null hypothesis of case 1 at $\alpha = .05$. For comparison, the corresponding theoretical probabilities are shown in the last column.

For point estimates, the results showed that all but the precision estimates are robust (i.e., have little bias) for all 24 situations in this study, even when $n = 15$. The precision estimate performs well when X is random, but tends to overestimate when X is fixed and the sample size is smaller.

For standard deviation estimates, there was practically no discrepancy between the sixth and seventh columns in all situations. For significance level and power, the results showed that all but the precision estimates are accurate for all 24 situations in this study, even when $n = 15$. The precision estimate performs well when X is random, but tends to reject more often due to overestimates when X is fixed and the sample size is smaller. The power of the CCC estimate tends to be larger in the simulation study than its theoretical value.

6. EXAMPLES

This section presents two examples based on real data. One example compares the agreement of two instruments in measuring blood counts in human samples, and the other compares the agreement of an assay to measure factor V III against

Table 1. Asymptotic Power of Accepting the Agreement, Where $\kappa 1 = 1.5\sigma_{d0}, \kappa 2 = 2\sigma_{d0}$, and $\kappa 3 = 2.5\sigma_{d0}$ (Random); $\kappa 1 = \sigma_{e0}, \kappa 2 = 1.5\sigma_{e0}$, and $\kappa 3 = 2\sigma_{e0}$ (Fixed); and $\upsilon_0 = .15, \varpi_0 = 1.15, n = 30$, and $\alpha = .05$

				Random					Fixed				
ρ_0	υ_1	ϖ_1	ρ_1	*TDI*	*CCC*	$CP_{\kappa 1}$	$CP_{\kappa 2}$	$CP_{\kappa 3}$	*TDI*	*CCC*	$CP_{\kappa 1}$	$CP_{\kappa 2}$	$CP_{\kappa 3}$
.80	.05	1.05	.8332	.2601	.1936	.3128	.3258	.3330	.3905	.2623	.4246	.4468	.4625
			.8614	.5149	.3666	.5781	.5916	.5986	.6609	.5120	.6793	.6947	.7048
		1.10	.8332	.2368	.1783	.2854	.2975	.3043	.2981	.2357	.3294	.3488	.3627
			.8614	.4798	.3432	.5451	.5592	.5667	.5589	.4815	.5875	.6070	.6197
	.10	1.05	.8332	.2341	.1787	.2822	.2946	.3017	.3586	.2399	.3920	.4141	.4300
			.8614	.4757	.3421	.5413	.5562	.5645	.6238	.4786	.6459	.6637	.6756
		1.10	.8332	.2126	.1643	.2564	.2677	.2743	.2702	.2144	.2993	.3178	.3313
			.8614	.4417	.3196	.5082	.5235	.5320	.5201	.4477	.5510	.5721	.5863
.90	.05	1.05	.9174	.3752	.2644	.4379	.4513	.4574	.5209	.4008	.5512	.5706	.5828
			.9318	.6478	.4551	.6931	.7017	.7054	.7771	.6763	.7794	.7869	.7914
		1.10	.9174	.3217	.2286	.3805	.3933	.3991	.3928	.3438	.4264	.4468	.4601
			.9318	.5815	.4060	.6362	.6467	.6514	.6585	.6203	.6766	.6910	.6996
	.10	1.05	.9174	.3155	.2290	.3738	.3880	.3958	.4553	.3437	.4873	.5090	.5239
			.9318	.5738	.4026	.6298	.6430	.6506	.7145	.6086	.7240	.7374	.7464
		1.10	.9174	.2679	.1968	.3200	.3328	.3398	.3333	.2900	.3647	.3849	.3992
			.9318	.5082	.3562	.5700	.5846	.5927	.5872	.5495	.6110	.6303	.6434
.95	.05	1.05	.9589	.5814	.4117	.6323	.6391	.6391	.7259	.6291	.7357	.7431	.7454
			.9662	.8146	.6112	.8239	.8243	.8228	.9036	.8522	.8882	.8854	.8829
		1.10	.9589	.4712	.3321	.5302	.5386	.5390	.5560	.5234	.5836	.5979	.6042
			.9662	.7134	.5202	.7441	.7475	.7467	.7851	.7734	.7860	.7909	.7924
	.10	1.05	.9589	.4585	.3293	.5182	.5326	.5397	.6131	.5102	.6313	.6486	.6595
			.9662	.7020	.5089	.7356	.7468	.7532	.8249	.7532	.8153	.8234	.8290
		1.10	.9589	.3588	.2598	.4164	.4296	.4355	.4390	.4058	.4684	.4881	.5006
			.9662	.5894	.4211	.6399	.6535	.6608	.6730	.6555	.6836	.7001	.7110
.95	.05	1.05	.9918	.9936	.9486	.9787	.9726	.9675	.9993	.9985	.9963	.9923	.9870
			.9933	.9989	.9813	.9902	.9866	.9839	.9999	.9998	.9986	.9965	.9937
		1.10	.9918	.9270	.8388	.9118	.8982	.8830	.9805	.9796	.9705	.9630	.9514
			.9933	.9697	.9186	.9502	.9401	.9301	.9951	.9951	.9867	.9824	.9763
	.10	1.05	.9918	.9241	.7875	.9099	.9195	.9220	.9856	.9711	.9676	.9692	.9640
			.9933	.9712	.8702	.9544	.9598	.9612	.9966	.9923	.9848	.9858	.9828
		1.10	.9918	.7201	.5880	.7376	.7493	.7496	.8803	.8658	.8702	.8689	.8598
			.9933	.8240	.6957	.8249	.8376	.8402	.9473	.9413	.9243	.9246	.9230

NOTE: $h = 1$ for TDI and CP when X is random, and for TDI when X is fixed.

known target values in test tubes (in vitro). The former is the constant error case when X is random, and the latter is the proportional error case when X is fixed.

6.1 Constant Error When the Target Values are Random

Diaspirin crosslinked hemoglobin (DCLHb) is a solution containing oxygen-carrying hemoglobin. The solution was created as a blood substitute to treat acute trauma patients and to replace blood loss during surgery. Measurements of DCLHb in patient's serum after infusion are routinely performed using a Sigma instrument. A method of measuring hemoglobin called the HemoCue photometer was modified to reproduce the Sigma instrument DCLHb results. To validate this modified method, serum samples from 299 patients over the analytical range of 50–2000 mg/dL were collected. DCLHb values of each sample were measured simultaneously with the HemoCue and Sigma methods. Agreement was defined as having at least 90% of pair observations over the analytical range of 50–2000 mg/dL within 150 mg/dL of each other and a within-sample total deviation not more than 15% of the total deviation. This translates into a least acceptable CCC of .9775 $(1 - .15^2)$.

Figure 1 presents the plot of HemoCue versus Sigma measurements of DCLHb. The plot indicates that the within-sample error is relatively constant across the clinical range. The plot also indicates that the HemoCue accuracy is excellent and that the precision is adequate.

Table 6 presents the agreement statistics and the appropriate 95% upper or lower confidence limits. The CCC estimate is .9866, which means that the within-sample total deviation is about 11.6% of the total deviation. The CCC one-sided lower confidence limit is .9838, which is greater than .9775. The precision estimate is .9867 with a one-sided lower confidence limit of .9839. The accuracy estimate is .9999 with a one-sided lower confidence limit of .9989. The MSD estimate is 6,007 with a one-sided upper confidence limit of 6,875. The TDI_9 estimate is 127.5 mg/dL, which means that 90% of HemoCue observations were within 127.5 mg/dL of their target values. The one-sided upper confidence limit for TDI_9 is 136.4 mg/dL, which is less than 150 mg/dL. Finally, the CP_{150} estimate is .9463, which means that 94.6% of HemoCue observations are within 150 mg/dL of their target values. The

Table 2. Results of the Simulation Study When X is Random for Moderate Precision

Case	Sample size	Statistics	Theoretical value	Mean of estimate	Std of estimate	Mean of std	Proportion in reject region	Theoretical probability
H_0	15	Precision	.9500	.9531	.2848	.2887	.0590	.05
		Accuracy	.9794	.9767	.9218	.9669	.0498	.05
		TDI	.6200	.6213	.3696	.3804	.0486	.05
		CCC	.9304	.9248	.2531	.2437	.0364	.05
		$CP_{\kappa 1}$	.8303	.8143	.5971	.6098	.0474	.05
		$CP_{\kappa 3}$	.9789	.9765	1.2977	1.2469	.0596	.05
	30	Precision	.9500	.9516	.1913	.1925	.0548	.05
		Accuracy	.9794	.9783	.6448	.6391	.0558	.05
		TDI	.6200	.6209	.2607	.2616	.0532	.05
		CCC	.9304	.9280	.1727	.1690	.0354	.05
		$CP_{\kappa 1}$	.8303	.8220	.4017	.4096	.0492	.05
		$CP_{\kappa 3}$	.9789	.9776	.8381	.8353	.0592	.05
	60	Precision	.9500	.9506	.1314	.1325	.0528	.05
		Accuracy	.9794	.9788	.4324	.4375	.0518	.05
		TDI	.6200	.6211	.1799	.1825	.0488	.05
		CCC	.9304	.9290	.1188	.1186	.0368	.05
		CP_{κ_1}	.8303	.8267	.2789	.2837	.0482	.05
		CP_{κ_3}	.9789	.9785	.5775	.5786	.0550	.05
H_1	15	Precision	.9662	.9685	.2824	.2887	.1926	.1777
		Accuracy	.9905	.9892	1.0575	1.1993	.1736	.2554
		TDI	.4843	.4848	.3756	.3841	.3250	.3568
		CCC	.9571	.9539	.2552	.2505	.2084	.2594
		$CP_{\kappa 1}$	.9220	.9132	.8592	.8286	.3088	.4347
		$CP_{\kappa 3}$	.9969	.9965	2.0338	1.9178	.3486	.4682
	30	Precision	.9662	.9673	.1901	.1925	.2952	.2786
		Accuracy	.9905	.9901	.7743	.7884	.2890	.3744
		TDI	.4843	.4856	.2610	.2634	.5712	.5894
		CCC	.9571	.9555	.1734	.1739	.3764	.4211
		$CP_{\kappa 1}$	.9220	.9180	.5679	.5577	.5534	.6399
		$CP_{\kappa 3}$	.9969	.9967	1.3489	1.2914	.5946	.6608
	60	Precision	.9662	.9669	.1327	.1325	.4744	.4513
		Accuracy	.9905	.9904	.5293	.5310	.5088	.5575
		TDI	.4843	.4831	.1825	.1838	.8626	.8516
		CCC	.9571	.9566	.1230	.1221	.6442	.6613
		$CP_{\kappa 1}$	.9220	.9192	.3817	.3840	.8434	.8573
		$CP_{\kappa 3}$	.9969	.9967	.8922	.8868	.8604	.8613

NOTE: $v_0 = .15, \varpi_0 = 1.15, \rho_0 = .95; v_1 = .1, \varpi_1 = 1.1, \rho_1 = .9662$.

one-sided lower confidence limit for CP_{150} is .9276, which is greater than .9. The agreement between HemoCue and Sigma is acceptable with excellent accuracy and adequate precision. The relative bias squared is estimated to be .003, and so the approximation of TDI should be excellent.

6.2 Proportional Error When the Target Values are Fixed

A study was designed to evaluate Dade International's reagent test system for clottable factor VIII (FVIII) assay. The FVIII assay involved measuring modified activated partial thromboplastin time (APTT) with varying dilutions of plasma and specific factor-deficient substrate. Using a reference plasma of known FVIII activity, a standard curve was prepared. The dilution scheme of the standard curve started at either 1:5 or 1:10, and serial dilutions were prepared until the target values were reached. The percent of FVIII activity present in plasma was determined by the degree of correction of the APTT. The reagent test system consisted of Dade actin-activated cephaloplastin reagent and Dade factor assay reference plasma. The target values were 3%, 8%, 38%, 91%, and 108%. Each level was assayed for six FVIII observations starting at 1:5 and 1:10. One FVIII value started at 1:5 at the 91% target value was missing.

Figure 2 presents the results started at 1:5, and Figure 3 presents the same at 1:10 for the plots of observed FVIII assay results versus targeted values in $\log_2$ scale. Note that in Figure 2, four 3% and two 2% were observed at the target value of 3%. Circles at the target value of 8% represents duplicate readings of 8%, 9%, and 10%. Duplicate readings of 45% were observed at target values of 38%. Also note that in Figure 3, four 5% and two 4% were observed at the target value of 3%, three 11% and two 12% were observed at the target value of 8%, duplicate readings of 49% were observed at target values of 38%, and duplicate readings of 124% were observed at target values of 91%. The plots indicate that the within-sample error was relatively constant across the target values in log scale. The precision was good for both, but the accuracy was not as good for the assay started at 1:10.

The client defined an acceptable agreement as having 80% of FVIII assay values over the analytical range of 3%–108% within 50% from the target percentage values. The client also wanted, in log scale due to proportional error by dilution, the within-sample total deviation to be not more than 15% of the total deviation. This translated into a least acceptable CCC of .9775. Table 7 presents the agreement statistics and their 95% upper or lower confidence limits for the assays started at 1:5 and 1:10.

Table 3. Results of the Simulation Study When X is Random for High Precision

Case	Sample size	Statistics	Theoretical value	Mean of estimate	Std of estimate	Mean of std	Proportion in reject region	Theoretical probability
H_0	15	Precision	.9900	.9907	.2820	.2887	.0618	.05
		Accuracy	.9794	.9774	.5252	.5086	.0390	.05
		TDI	.4098	.4104	.3567	.3592	.0558	.05
		CCC	.9696	.9671	.2176	.2108	.0320	.05
		$CP_{\kappa 1}$	.7600	.7448	.5103	.5373	.0462	.05
		$CP_{\kappa 3}$	.9590	.9563	1.0542	1.0348	.0624	.05
	30	Precision	.9900	.9903	.1893	.1925	.0568	.05
		Accuracy	.9794	.9784	.3556	.3486	.0440	.05
		TDI	.4098	.4116	.2490	.2465	.0520	.05
		CCC	.9696	.9683	.1472	.1455	.0350	.05
		$CP_{\kappa 1}$	.7600	.7545	.3472	.3606	.0462	.05
		$CP_{\kappa 3}$	.9590	.9586	.6958	.6934	.0672	.05
	60	Precision	.9900	.9902	.1301	.1325	.0554	.05
		Accuracy	.9794	.9790	.2434	.2432	.0436	.05
		TDI	.4098	.4103	.1724	.1721	.0544	.05
		CCC	.9696	.9691	.1005	.1016	.0394	.05
		$CP_{\kappa 1}$	.7600	.7571	.2483	.2483	.0500	.05
		$CP_{\kappa 3}$	.9590	.9587	.4875	.4750	.0632	.05
H_1	15	Precision	.9933	.9938	.2831	.2887	.1986	.1788
		Accuracy	.9905	.9897	.5879	.5753	.3690	.4594
		TDI	.2965	.2978	.3636	.3664	.5166	.5481
		CCC	.9839	.9826	.2235	.2205	.3704	.4400
		$CP_{\kappa 1}$	.9031	.8957	.7845	.7728	.4778	.5949
		$CP_{\kappa 3}$	.9959	.9957	2.0080	1.8336	.5410	.6382
	30	Precision	.9933	.9936	.1898	.1925	.3096	.2807
		Accuracy	.9905	.9902	.3924	.3908	.6678	.7043
		TDI	.2965	.2969	.2494	.2518	.8270	.8240
		CCC	.9839	.9834	.1531	.1521	.6696	.6957
		$CP_{\kappa 1}$	.9031	.8998	.5075	.5166	.8182	.8249
		$CP_{\kappa 3}$	.9959	.9958	1.2364	1.2190	.8500	.8402
	60	Precision	.9933	.9934	.1301	.1325	.4782	.4548
		Accuracy	.9905	.9904	.2772	.2730	.9138	.9218
		TDI	.2965	.2968	.1743	.1758	.9850	.9797
		CCC	.9839	.9837	.1074	.1065	.9124	.9231
		$CP_{\kappa 1}$	.9031	.9018	.3539	.3564	.9834	.9697
		$CP_{\kappa 3}$	.9959	.9959	.8481	.8397	.9898	.9691

NOTE: $\upsilon_0 = .15, \varpi_0 = 1.15, \rho_0 = .99$; $\upsilon_1 = .1, \varpi_1 = 1.1, \rho_1 = .9933$.

For the results started at 1:5, the CCC was estimated to be .9917, which means that the within-sample total deviation is about 9.1% of the total deviation. The one-sided lower confidence limit was .9875, which is greater than .9775. The precision was estimated to be .9942 with a one-sided lower confidence limit of .9908, the accuracy was estimated to be .9975 with a one-sided lower confidence limit of .9935, and the MSD was estimated to be .0356 with a one-sided upper confidence limit of .0549 (both at log scale). The $TDI_{.8}\%$ was estimated to be 27.3%, which means that 80% of observations were within a 27.3% change from the target values (percentage of percentage values). The one-sided upper confidence limit was 35%, which is less than 50%. Finally, the $CP_{50\%}$ was estimated to be .9653, which means that 96.5% of observations were within a 50% change from target values. The one-sided lower confidence limit was .8921, which is greater than .8. The agreement between the FVIII assay and the actual concentration was acceptable, with good precision and accuracy. The relative bias squared was estimated to be .009, so that the approximation of TDI should be excellent.

For the results started at 1:10, the CCC was estimated to be .9669, which means that the within-sample total deviation is about 18.2% of the total deviation. The one-sided lower confidence limit was .9584, which is less than .9775. The precision was estimated to be .9947 with a one-sided lower confidence limit of .9917, the accuracy was estimated to be .9721 with a one-sided lower confidence limit of .9638, and the MSD was estimated to be .1308 with a one-sided upper confidence limit of .1677 (both at log scale). The $TDI_{.8}\%$ was estimated to be 58.9%, which means that 80% of observations were within a 58.9% change from the target percentage values. The one-sided upper confidence limit was 69.0%, which is greater than 50%. Finally, the $CP_{50\%}$ was estimated to be .7016, which means that 70.2% of the observations were within a 50% change from the target values. The one-sided lower confidence limit was .5898, which is less than .8. The agreement between the FVIII assay and actual concentration was not acceptable with good precision and mediocre accuracy. The relative bias squared was estimated to be 3.747, so the approximation of TDI should be acceptable.

7. DISCUSSION AND FUTURE STUDY

7.1 Agreement Measurements Summary

Under the normal or log-normal distribution, each of the agreement measurements (MSD, CCC, TDI, and CP) basically measures the same information but from different perspectives. Note that the asymptotic variances of MSD, CCC, precision, and accuracy were derived from the covariance matrix of the sample moments, in which the normality assumption was

Table 4. Results of the Simulation Study When X is Fixed for Moderate Precision

Case	Sample size	Statistics	Theoretical value	Mean of estimate	Std of estimate	Mean of std	Proportion in reject region	Theoretical probability
H_0	15	Precision	.9500	.9597	.2106	.2126	.1282	.05
		Accuracy	.9794	.9800	.8585	.8914	.0670	.05
		TDI	.6339	.6320	.3751	.3706	.0546	.05
		CCC	.9304	.9351	.1740	.1710	.0760	.05
		$CP_{\kappa 1}$	.8302	.8191	.6109	.6066	.0504	.05
		$CP_{\kappa 3}$	.9339	.9305	.9218	.8825	.0594	.05
	30	Precision	.9500	.9543	.1417	.1422	.0992	.05
		Accuracy	.9794	.9797	.5895	.5878	.0662	.05
		TDI	.6339	.6359	.2564	.2569	.0516	.05
		CCC	.9304	.9322	.1189	.1189	.0670	.05
		$CP_{\kappa 1}$	.8302	.8222	.3996	.4044	.0486	.05
		$CP_{\kappa 3}$	.9399	.9305	.5856	.5840	.0552	.05
	60	Precision	.9500	.9525	.0959	.0980	.0792	.05
		Accuracy	.9794	.9796	.4019	.4011	.0650	.05
		TDI	.6339	.6327	.1779	.1800	.0496	.05
		CCC	.9304	.9317	.0823	.0832	.0612	.05
		$CP_{\kappa 1}$	.8302	.8276	.2761	.2811	.0484	.05
		$CP_{\kappa 3}$	.9339	.9331	.4014	.4060	.0546	.05
H_1	15	Precision	.9662	.9730	.2115	.2099	.4034	.2357
		Accuracy	.9905	.9904	.9707	1.1114	.2094	.2824
		TDI	.4843	.4822	.3732	.3754	.3896	.4264
		CCC	.9571	.9602	.1779	.1774	.4654	.4076
		$CP_{\kappa 1}$	.9285	.9235	.8696	.8562	.3696	.4537
		$CP_{\kappa 3}$	.9841	.9837	1.4270	1.3586	.4078	.5052
	30	Precision	.9662	.9692	.1407	.1403	.5204	.4028
		Accuracy	.9905	.9906	.7395	.7750	.3494	.4141
		TDI	.4843	.4862	.2596	.2600	.6340	.6730
		CCC	.9571	.9581	.1229	.1232	.6828	.6555
		$CP_{\kappa 1}$	.9285	.9238	.5703	.5677	.6226	.6836
		$CP_{\kappa 3}$	.9841	.9830	.9098	.8926	.6498	.7110
	60	Precision	.9662	.9678	.0980	.0966	.7380	.6548
		Accuracy	.9905	.9906	.4951	.4975	.5750	.6093
		TDI	.4843	.4844	.1843	.1821	.9060	.9088
		CCC	.9571	.9577	.0876	.0862	.9154	.8987
		$CP_{\kappa 1}$	.9285	.9267	.4012	.3951	.9026	.9006
		$CP_{\kappa 3}$	.9841	.9837	.6344	.6205	.9104	.9023

NOTE: $\upsilon_0 = .15, \varpi_0 = 1.15, \rho_0 = .95; \upsilon_1 = .1, \varpi_1 = 1.1, \rho_1 = .9662$.

used, even though the point estimates do not depend on the normality assumption. None of the methods proposed in this article is expected to be robust against outliers and/or large deviation from normality.

The interpretation of MSD is difficult to understand. The TDI is desirable as an alternative because of its straightforward interpretation. The CP is the most intuitively clear approach; it mirrors the information provided by the TDI. Both TDI and CP depend on the normality assumption and offer better power for inference than the CCC. The CP would have difficulty discriminating among instruments or assays that have excellent agreement, all because the CP values would be very close to 1. In this case, the TDI can be used to discriminate among these.

When a meaningful clinical range is known and the study is conducted over that range, the CCC offers a meaningful geometric interpretation and is unit free. Furthermore, the accuracy and precision components of the CCC offer more insight. Therefore, the CCC, accuracy, and precision remain very useful tools. Note that when Y and X are not linearly related, the CCC will capture the total deviation. However, it will treat the nonlinear deviation as imprecision rather than inaccuracy.

The CCC, ICC, and Pearson correlation coefficient depend largely on the analytical range and the intrasample variation. This property makes sense when we want to assess agreement over the range of all potential outcomes. Good agreement over a small range of measurements (e.g., at low concentration only) can not be extrapolated to infer good agreement over a larger range of measurements (e.g., at higher concen-

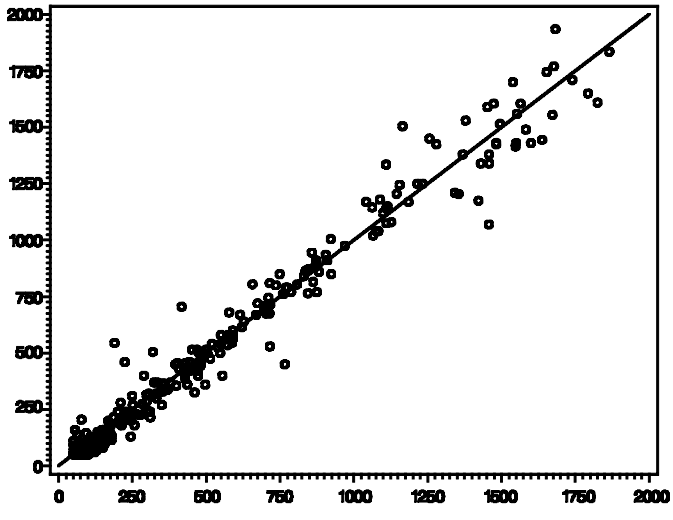

Figure 1. HemoCue (Vertical) and Sigma Readings (Horizontal) on Measuring DCLHb.

Table 5. Results of the Simulation Study When X is Fixed for High Precision

Case	Sample size	Statistics	Theoretical value	Mean of estimate	Std of estimate	Mean of std	Proportion in reject region	Theoretical probability
H_0	15	Precision	.9900	.9919	.2010	.2059	.1200	.05
		Accuracy	.9794	.9795	.3690	.3599	.0586	.05
		TDI	.4190	.4256	.2923	.3013	.0456	.05
		CCC	.9696	.9707	.1342	.1311	.0702	.05
		$CP_{\kappa 1}$	.7534	.7473	.4232	.4527	.0520	.05
		$CP_{\kappa 3}$	.8879	.8880	.6513	.6695	.0632	.05
	30	Precision	.9900	.9910	.1364	.1373	.1048	.05
		Accuracy	.9794	.9795	.2556	.2529	.0566	.05
		TDI	.4190	.4214	.2030	.2059	.0492	.05
		CCC	.9696	.9703	.0932	.0922	.0660	.05
		$CP_{\kappa 1}$	.7534	.7514	.2899	.3011	.0538	.05
		$CP_{\kappa 3}$	.8879	.8888	.4337	.4421	.0634	.05
	60	Precision	.9900	.9905	.0943	.0946	.0872	.05
		Accuracy	.9794	.9795	.1773	.1784	.0558	.05
		TDI	.4190	.4199	.1410	.1434	.0500	.05
		CCC	.9696	.9700	.0648	.0651	.0630	.05
		$CP_{\kappa 1}$	.7534	.7527	.2010	.2065	.0536	.05
		$CP_{\kappa 3}$	.8879	.8885	.2963	.3020	.0610	.05
H_1	15	Precision	.9933	.9947	.2021	.2053	.4276	.2509
		Accuracy	.9905	.9906	.4564	.4379	.6090	.6351
		TDI	.2965	.2993	.3165	.3198	.6960	.7452
		CCC	.9839	.9847	.1494	.1450	.7802	.7236
		$CP_{\kappa 1}$	.9101	.9109	.7608	.7546	.7250	.7165
		$CP_{\kappa 3}$	.9809	.9832	1.3544	1.2804	.7614	.7437
	30	Precision	.9933	.9940	.1381	.1369	.5596	.4276
		Accuracy	.9905	.9906	.3104	.3079	.8886	.8730
		TDI	.2965	.2984	.2172	.2205	.9482	.9473
		CCC	.9839	.9842	.1020	.1022	.9664	.9413
		$CP_{\kappa 1}$	.9101	.9096	.4879	.4944	.9556	.9243
		$CP_{\kappa 3}$	.9809	.9817	.8371	.8259	.9644	.9230
	60	Precision	.9933	.9937	.0945	.0943	.7766	.6853
		Accuracy	.9905	.9906	.2169	.2166	.9942	.9874
		TDI	.2965	.2968	.1554	.1542	.9992	.9985
		CCC	.9839	.9841	.0732	.0721	.9994	.9982
		$CP_{\kappa 1}$	.9101	.9106	.3422	.3397	.9992	.9954
		$CP_{\kappa 3}$	.9809	.9816	.5759	.5643	.9992	.9933

NOTE: $v_0 = .15, \varpi_0 = 1.15, \rho_0 = .99$; $v_1 = .1\,\varpi_1 = 1.1, \rho_1 = .9933$.

tration). However, caution must be taken when using these correlation coefficients. Comparisons among these coefficients are meaningful only when the clinical study ranges are similar. Ranges by different units can be compared as long as they have similar clinical interpretations.

7.2 Categorical Data

The CP has long been used in categorical data by summing the diagonal elements of the joint probability matrix based on assigning subjects' scores to two raters. Cohen (1960) proposed using the kappa coefficient to correct for the probability of agreement by chance. Cohen (1968) later improved the kappa coefficient by assigning different weights according to the degree of disagreements. Interestingly, CCC becomes the weighted kappa proposed by Cohen (1968). Therefore, we can use the CCC for categorical data in most situations (Robieson 1999; King and Chinchilli 2001).

Table 6. Agreement Statistics and Their Confidence Limits for Example 1

Statistics	Estimate	95% Confidence limit	Allowance
CCC	.9866	.9838	.9775
Accuracy	.9999	.9989	
Precision	.9867	.9839	
MSD	6,007	6,875	
$TDI_{.9}$	127.5	136.4	150
CP_{150}	.9463	.9276	.9

NOTE: Relative bias squared was estimated to be .003 ($\leq$1; see sec. 2.4).

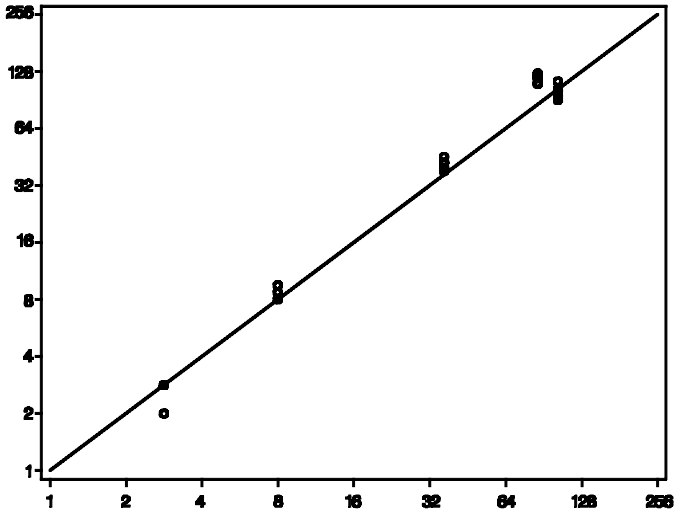

Figure 2. Observed FVIII Assay Results (Vertical, %) Versus Targeted Values (Horizontal, %) Started at 1:5.

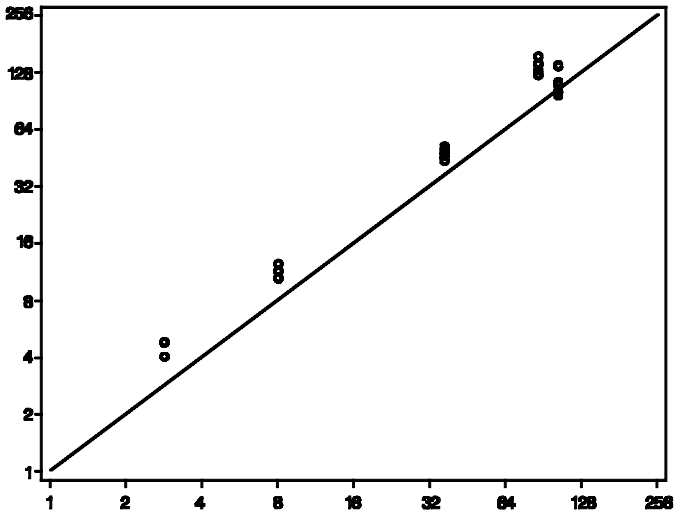

Figure 3. Observed FVIII Assay Results (Vertical, %) Versus Targeted Values (Horizontal, %) Started at 1:10.

7.3 More General Cases

This article provides a systematic treatment of modeling agreement measurement methodology based on the most basic bivariate model when X is random and on the regression model when X is fixed. When X is random, the model does not allow us to assess agreement among repeated estimates of the target method. Therefore, it is not known whether any mediocre agreement between two methods may have resulted from imprecision in the target method. In an individual bioequivalence (IBE) study, measures of the bioavailability are recorded when a patient is given a test formulation twice (T_1 and T_2) and a reference formulation twice (R_1 and R_2) by a four-period crossover design (Schall and Williams 1996). The current FDA guideline is based on the absolute and relative measures of $E(T-R)^2 - E(R_1-R_2)^2$. The denominator of the relative measure is $E(R_1-R_2)^2$.

The many gold standard assays include immunoassays or enzyme assays, which are imprecise. The methodologies and designs used in an IBE study should be adopted for imprecise assays. More importantly, to prove that a new method is better than a gold standard method, one should adopt the four-period design and principles used in the IBE study. Here we would be in a position to demonstrate an acceptable $E(T-R)^2$ and further that $E(T_1-T_2)^2 < E(R_1-R_2)^2$. We would then translate the measurement into CCC, precision, accuracy, TDI, and CP for better interpretation. Our future articles will address those more general cases.

When assessing the agreement coefficients, there might be situations where some other controllable factors could be present. For example, in a four period crossover bioequivalence study, treatment-order effects might be present. In these situations, we can eliminate these effects by fitting a mixed-effects model with these fixed effects and a random subject effect in the model. Then the methods presented here can be applied asymptotically by letting Y and X represent the respective residuals computed from the model.

Chinchilli Martel, Kumanyka, and Lloyd (1996) extended Lin's CCC to repeated-measures designs by using a weighted CCC. The CCC on multiple raters with robust estimates has been studied by King and Chinchilli (2001). For more general approaches, the use of generalized estimating equations (GEEs) possibly could be used to model the functions of the agreement coefficients. Barnhart and Williamson (2001) used GEEs to model CCC with good results. They used three estimating equations, one to model location sums, one to model the sums of squares, and one to model the cross-products with Z transformation of CCC values computed from functions of the foregoing. Potentially, GEEs also can be used to model MSD, TDI, and CP. This approach is flexible enough to allow comparisons of multiple agreement coefficients in the presence of some explanatory covariates. Such an approach would allow us to compare, for example, the agreements of two competing methods (A and B) with a gold standard method (C) or, in other words, to compare the agreement of A and C to the agreement of B and C. Thus GEE will fit nicely into the framework of an IBE study for comparing the agreement of T and R relative to the agreement of R_1 and R_2, while controlling for period and order effects.

Table 7. Agreement Statistics and Their Confidence Limits for Example 2

Statistics	Estimate	95% Confidence limits	Allowance
	Started at 1:5		
CCC	.9917	.9875	.9775
Accuracy	.9975	.9935	
Precision	.9942	.9908	
MSD	.0356	.0549	
$TDI_{.8}$%	27.347	35.010	50
$CP_{50\%}$	.9653	.8921	.8
	Started at 1:10		
CCC	.9669	.9584	.9775
Accuracy	.9721	.9638	
Precision	.9947	.9917	
MSD	.1308	.1677	
$TDI_{.8}$%	58.949	69.007	50
$CP_{50\%}$	.7016	.5898	.8

NOTE: Relative bias squared was estimated to be .009 for the assay started at 1:5 and 3.747 for the assay started at 1:10 ($\leq$8; see sec. 2.4).

8. CONCLUSION

We have summarized various methods for assessing the agreement among individual paired samples when X is random or fixed and when error is constant or proportional. We suggest using CCC, TDI, and CP to summarize the agreement results. These offer the same information from different perspectives. In addition, the coefficients of accuracy and precision should also accompany the results to identify the sources of any disagreement. When we are confident that the data have normal or lognormal distribution, inference should be based on TDI and CP for better power of accepting the agreement. We plan to provide more general approaches regarding all of the above in the future.

For convenience, a validated SAS macro is provided at http://www.uic.edu/ $\sim$ hedayat/ that computes the estimates and confidence limits for CCC, precision, accuracy, TDI, and CP. We can specify when X is random or fixed and the error is constant or proportional, along with the confidence level, CCC, CP, and TDI allowances. The program also generates

the agreement plot of Y versus X with the identity line under a customized scale.

[Received December 2000. Revised July 2001.]

REFERENCES

Anderson, S., and Hauck, W. W. (1990), "Considersation of Individual Bioequivalence," *Journal of Pharmacokinetics and Biopharmaceutics*, 18, 259–273.

Barnhart, H. X., and Williamson, J. M. (2001), "Modeling Concordance Correlation via GEE to Evaluate Reproducibility," *Biometrics*, 57, 931–940.

Bland. J. M., and Altman, D. (1986), "Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement," *Lancet*, 8, 307–310.

Chinchilli, V. M., Martel, J. K., Kumanyka, S., and Lloyd, T. (1996), "A Weighted Concordance Correlation Coefficient for Repeated Measurement Designs," *Biometrics*, 52, 341–353.

Cohen, J. (1960), "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, 20, 37–46.

——— (1968), "Weighted Kappa: Nominal Scale Agreement With Provision for Scaled Disagreement or Partial Credit," *Psychological Bulletin*, 70, 213–220.

Fisher, R. A. (1925), *Statistical Methods for Research Workers*, Edinburgh: Oliver & Boyd.

Holder, D. J., and Hsuan, F. (1993), "Moment-Based Criteria for Determining Bioequivalence," *Biometrika*, 80, 835–846.

King, T. S., and Chinchilli, V. M. (2001), "A Generalized Concordance Correlation Coefficient for Continuous and Categorical Data," *Statistics in Medicine*, 20, 2131–2147.

Lin. L. I. (1989), "A Concordance Correlation Coefficient to Evaluate Reproducibility," *Biometrics*, 45, 255–268.

——— (1992), "Assay Validation Using the Concordance Correlation Coefficient," *Biometrics*, 48, 599–604.

——— (1997), Rejoinder to the Letter to the Editor by Atkinson and Nevill, *Biometrics*, 53, 777–778.

——— (2000), "Total Deviation Index for Measuring Individual Agreement: With Application in Lab Performance and Bioequivalence," *Statistics in Medicine*, 19, 255–270.

Lin. L. I., and Torbeck, L. D. (1998), "Coefficient of Accuracy and Concordance Correlation Coefficient: New Satistics for Method Comparison," *PDA Journal of Pharmaceutical Science and Technology*, 52, 55–59.

Robieson, W. Z. (1999), "On the Weighted Kappa and Concordance Correlation Coefficient," Ph.D. dissertation, University of Illinois at Chicago.

Schall. R. (1995), "Assessment of Individual and Population Bioequivalence Using the Probability That Bioavailities are Similar," *Biometrics*, 51, 615–626.

Schall, R., and Luus, H. G. (1993), "On Population and Individual Bioequivalence," *Statistics in Medicine*, 12, 1109–1124.

Schall, R., and Williams, R. L. (1996), "Towards a Practical Strategy for Assessing Individual Bioequivalence," *Journal of Pharmacokinetics and Biopharmaceutics*, 24, 133–149.

Sheiner, L. B. (1992), "Bioequivalence Revisted," *Statistics in Medicine*, 11, 1777–1788.

Vonesh. E. F., and Chinchilli, V. M. (1997), *Linear and Nonlinear Models for the Analysis of Repeated Measurement*, New York: Marcel Dekker.

Vonesh, E. F., Chinchilli, V. M., and Pu, K. (1996), "Goodness of Fit in Generalized Nonlinear Mixed-Effect Models," *Biometrics*, 52, 572–587.