Journal of Biopharmaceutical Statistics, 19: 345–359, 2009

ISSN: 1054-3406 print/1520-5711 online
DOI: 10.1080/10543400802622576

Taylor & Francis
Taylor & Francis Group

ASSESSMENT OF INDIVIDUAL AGREEMENTS WITH REPEATED MEASUREMENTS BASED ON GENERALIZED CONFIDENCE INTERVALS

Jorge Quiroz[1] and Richard K. Burdick[2]
[1]*Statistics Department, Schering-Plough Research Institute, Kenilworth, New Jersey, USA*
[2]*Quality Engineering and Improvement Amgen, Inc., Longmont, Colorado, USA*

Individual agreement between two measurement systems is determined using the total deviation index (TDI) or the coverage probability (CP) criteria as proposed by Lin (2000) and Lin et al. (2002). We used a variance component model as proposed by Choudhary (2007). Using the bootstrap approach, Choudhary (2007), and generalized confidence intervals, we construct bounds on TDI and CP. A simulation study was conducted to assess whether the bounds maintain the stated type I error probability of the test. We also present a computational example to demonstrate the statistical methods described in the paper.

Key Words: Bootstrap-*t* confidence intervals; Coverage probability; Equivalence studies; Individual agreement; Generalized confidence intervals; Total deviation index; Variance components.

1. INTRODUCTION

An important problem that arises in many method comparison studies is the assessment of individual agreement of two competing methods, processes, formulations, instruments, assays, or two measurement systems. Very frequently, the new method measures an attribute where direct measurement is difficult or impossible. Since the true value remains unknown, the new method is evaluated by assessing its agreement to an established (reference) method instead of the true quantity. If the new and the reference measurement systems agree sufficiently well, the reference may be replaced or both measurement systems used interchangeably. Equivalence testing is an approach commonly used to determine the acceptability of a new method against a reference method.

If the methods are used to make decisions based on individuals rather than on population characteristics, a measure of individual agreement is preferable. Lin (2000) and Lin et al. (2002) developed the total deviation index (TDI) and the coverage probability (CP) for assessing individual agreements. These criteria are intuitive measures of agreement that indicate a proportion of the data are within

Received August 23, 2007; Accepted July 11, 2008
Address correspondence to Jorge Quiroz, Statistics Department, Schering-Plough Research Institute, Kenilworth, New Jersey, USA; E-mail: jorge.quiroz@spcorp.com

a boundary. Both the TDI and CP are attractive criteria because they are easy to interpret. Lin et al. (2007) and Barhhart et al. (2007) extended the TDI and CP to data with multiple methods and multiple observations on the same subject. These authors also discuss other measures of agreements.

Paired measurements on the same subject occur naturally in method comparison studies where individual agreement is of interest. Experiments with paired measurements are repeated multiple times on the same subject and can be described by repeated measurement designs. These designs allow one to estimate the variability of each method. This variability can be an important component in the estimation of a repeatability parameter. The TDI and CP criteria can be used in this repeated measurement design. The model proposed originally by Lin (2000) and Lin et al. (2002) was not easily adaptable to this setting. Choudhary (2007) used an analysis of variance (ANOVA) approach to assess individual agreement using the TDI and CP criteria in a model with repeated measurements. The ANOVA approach is easy to use and widely available to practitioners. Choudhary and Nagaraja (2007) developed a bootstrap approach to construct bounds on TDI and CP. The bootstrap approach was later applied to mixed models with repeated measurements by Choudhary (2007). However, this bootstrap approach is based on large-sample properties and may not perform well for small sample data that are frequently an issue in method comparison studies. In this paper, we propose constructing bounds for the TDI and CP criteria using a variance components approach and generalized confidence intervals. Simulation study suggests that these bounds are good alternatives to the bootstrap approach.

In the following sections we define the ANOVA model using repeated measurements and describe the TDI and CP to measure individual agreement of two competing methods. We present tests on TDI and CP based on a bootstrap method and generalized confidence intervals. We also present a simulation study to compare the performance of the bounds and a computational example to demonstrate the tests.

2. THE EXPERIMENTAL DESIGN

Consider a study to assess the individual agreement between two methods. The standard experiment employs a repeated measurements design. On each occasion, simultaneous, continuous measurements on the same subject are obtained using a competing and a reference method. The experiment is repeated n times. The resulting responses are paired continuous measurement replicates. We also assume that interaction between the methods and the subject is present. For simplicity, we denote the competing system as test (T) and the standard system as reference (R). A measurement is described by the ANOVA model

$$Y_{ijk} = \mu_i + S_j + (MS)_{ij} + E_{ijk} \quad i = T, R; \quad j = 1, \ldots, s; \quad k = 1, \ldots, n \qquad (1)$$

where Y_{ijk} is the kth measurement from subject j obtained with method i, μ_T and μ_R are the means of the test and reference systems, respectively, S_j are independent normal random variables with mean 0 and variance σ_S^2, $(MS)_{ij}$ are independent normal random variables with mean 0 and variance σ_{iS}^2, and E_{ijk} are independent normal random variables with mean 0 and variance σ_i^2. The random interaction and error terms are independent from each other.

Table 1 A repeated measurements design with three subjects and two paired measurements per subject

Subject	Replicate	Reference	Test	Paired difference
1	1	Y_{R11}	Y_{T11}	D_{11}
	2	Y_{R12}	Y_{T12}	D_{12}
2	1	Y_{R21}	Y_{T21}	D_{21}
	2	Y_{R22}	Y_{T22}	D_{22}
3	1	Y_{R31}	Y_{T31}	D_{31}
	2	Y_{R32}	Y_{T32}	D_{32}

Table 1 displays a schematic illustration of the ANOVA design with three subjects and two paired measurements per subject.

Model (1) in matrix form is

$$\underline{Y} = X\underline{\mu} + Z_1\underline{U}_1 + Z_2\underline{U}_2 + \underline{\epsilon} \tag{2}$$

where $\underline{Y}^\top = [Y_{T11}, \ldots, Y_{Tsn}, \ldots, Y_{R11}, \ldots, Y_{Rsn}]$, $X = I_2 \otimes \underline{1}_s \otimes \underline{1}_n$, $Z_1 = \underline{1}_2 \otimes I_s \otimes \underline{1}_n$, $Z_2 = I_2 \otimes I_s \otimes \underline{1}_n$, $\underline{\mu}^\top = [\mu_T \quad \mu_R]$, $\underline{U}_1^\top = [S_1, \ldots, S_n]$, $\underline{U}_2^\top = [(MS)_{T1}, \ldots, (MS)_{Ts}, (MS)_{R1}, \ldots, (MS)_{Rs}]$, and $\underline{\epsilon}$ is multivariate normally distributed with mean $\underline{0}_{1\times 2sn}$, and variance

$$\mathrm{Var}(\underline{\epsilon}) = \Sigma_E \otimes I_s \otimes I_n, \tag{3}$$

where

$$\Sigma_E = \begin{bmatrix} \sigma_T^2 & 0 \\ 0 & \sigma_R^2 \end{bmatrix}. \tag{4}$$

In the matrix notation used above, the vector $\underline{1}_p$ is a $p \times 1$ vector of ones, J_p is a $p \times p$ matrix of ones, I_p is an identity matrix of order p, $P_p = (1/p)J_p$, $Q_p = I_p - P_p$, and $\otimes$ is the right direct (Kronecker) product operator.

Under the assumption of normality, the vector $\underline{Y}$ is multivariate normal with mean $X\underline{\mu}$ and variance–covariance matrix

$$\mathrm{Var}(\underline{Y}) = \sigma_S^2(J_2 \otimes I_s \otimes J_n) + \Sigma_{MS} \otimes I_s \otimes J_n + \Sigma_E \otimes I_s \otimes I_n, \tag{5}$$

where

$$\Sigma_{MS} = \begin{bmatrix} \sigma_{TS}^2 & 0 \\ 0 & \sigma_{RS}^2 \end{bmatrix}. \tag{6}$$

The measure of individual agreement is based on the difference $D_{jk} = Y_{Tjk} - Y_{Rjk}$, where Y_{Tjk} and Y_{Rjk} are measurements obtained on the same subject at the same time using the test and reference methods, respectively. Following an idea from Choudhary (2007), we used the difference of paired measurements to construct

bounds on TDI and CP. Based on the specification of model (1), the difference of paired measurements is described by the ANOVA model

$$D_{jk} = \mu_D + I_j + N_{jk} \quad j = 1, \ldots, s; \quad k = 1, \ldots, n, \tag{7}$$

where D_{jk} is the difference of paired measurements on the same subject, I_j are independent normal random variables with mean 0 and variance $\gamma_I = \sigma^2_{TS} + \sigma^2_{RS}$, and N_{jk} are independent normal random variables with mean 0 and variance $\gamma_E = \sigma^2_T + \sigma^2_R$. From the assumption of model (1), the random error terms are independent from each other.

Model (7) in matrix form is

$$\underline{D} = X_D \mu_D + Z_D \underline{U}_D + \underline{\eta}, \tag{8}$$

where $\underline{D}^\top = [D_{11}, D_{12}, \ldots, D_{1n}, \ldots, D_{s1}, \ldots, D_{sn}]$, $X_D = \underline{1}_s \otimes \underline{1}_n$, $Z_D = I_s \otimes \underline{1}_n$, the scalar $\mu_D = \mu_T - \mu_R$, $\underline{U}_D^\top = [I_1, \ldots, I_s]$, and $\underline{\eta}$ is multivariate normally distributed with mean $\underline{0}_{1\times sn}$ and variance $\gamma_E I_{sn}$.

The ANOVA for model (7) is shown in Table 2 where the sum of squares in quadratic forms $SS_I = \underline{D}^\top (Q_s \otimes P_n)\underline{D}$, and $SS_E = \underline{D}^\top (I_s \otimes Q_n)\underline{D}$.

Under the assumption of model (7), the mean squares S_I^2 and S_E^2 are independent. Additionally, $n_I S_I^2/\theta_I$ and $n_E S_E^2/\theta_E$ are chi-squared random variables with n_I and n_E degrees of freedom, respectively.

From the assumption above, the paired difference D_{jk} is normally distributed with mean μ_D and variance $\sigma_D^2 = \gamma_I + \gamma_E$. The random variable $\mathfrak{D}^2 = D_{jk}^2/\sigma_D^2$ has a noncentral chi-squared distribution with one degree of freedom and noncentrality parameter

$$\lambda_D = \frac{\mu_D^2}{\sigma_D^2}. \tag{9}$$

Other random variables of interest are $\overline{D}_{**}$, and $\overline{D}_{**}^2$, where $\overline{D}_{**} = \sum_{j=1}^{s}\sum_{k=1}^{n} D_{jk}/(sn)$. The random variable $\overline{D}_{**}$ is normally distributed with mean $\mu_D = \mu_T - \mu_R$ and variance $\theta_I/(sn)$. From the delta method, the random variable $\overline{D}_{**}^2$ is approximately normally distributed with mean μ_D^2 and variance $4\mu_D^2\theta_I/(sn)$.

3. MEASURE OF INDIVIDUAL AGREEMENTS

Consider a method comparison study to determine the agreement of two methods. Without loss of generality, the two methods can be described by two measuring systems, two formulations, two biological assays, or two processes.

Table 2 ANOVA for Model (7)

Source of variation	DF	MS	EMS
Subjects (I)	$n_I = s - 1$	$S_I^2 = SS_I/n_I$	$\theta_I = n\gamma_I + \gamma_E$
Error (N)	$n_E = s(n-1)$	$S_E^2 = SS_E/n_E$	$\theta_E = \gamma_E$

Often, the true value measured with the methods is unknown, and the new method is compared to an established method. Thus, both methods contain some measurement error. The competing method is represented by T and the reference method by R.

Lin (2000) and Lin et al. (2002) proposed the TDI and CP criteria as two user-friendly measures of individual agreement for paired measurements. Measures of individual agreements extended to more than two methods is discussed in Lin et al. (2007) and Barhhart et al. (2007). In this paper, we restrict our discussion to comparison of two methods. The TDI describes a boundary such that a majority of the differences of paired observations are within the boundary. In particular, the TDI for a particular percentage, π_0, is the boundary κ_{π_0} such that

$$\pi_0 = Pr[|Y_{Tjk} - Y_{Rjk}| \leq \kappa_{\pi_0}]. \tag{10}$$

Note that

$$\pi_0 = Pr[|D_{jk}| \leq \kappa_{\pi_0}] = Pr\left[\mathfrak{D}^2 \leq \frac{\kappa_{\pi_0}^2}{\sigma_D^2}\right]. \tag{11}$$

Under the assumption above, the TDI becomes

$$\kappa_{\pi_0} = \sqrt{\sigma_D^2 \chi^{2(-1)}(\pi_0, 1, \lambda_D)}, \tag{12}$$

where $\chi^{2(-1)}(\pi_0, 1, \lambda_D)$ is the π_0th percentile of a noncentral chi-squared random variable with one degree of freedom and noncentrality parameter $\lambda_D = \mu_D^2/\sigma_D^2$.

The TDI criterion can easily translate to an equivalence specification. To illustrate, consider the specification that two methods are equivalent if a large proportion of absolute paired differences are within a specified value, κ_0. More specifically, two methods are equivalent if at least 90% of the absolute paired differences are less than 10. We can use the following hypothesis test to demonstrate equivalence:

$$H_0 : \kappa_{0.90} \geq 10 \quad \text{vs.} \quad H_a : \kappa_{0.90} < 10. \tag{13}$$

We can test this set of hypotheses by constructing a 95% upper bound on $\kappa_{0.90}$. If the upper bound is less than 10, the two methods are declared equivalent.

Conversely, the coverage probability (CP) is the probability that the absolute paired differences are within a specified boundary. From Equation (10), we can define the CP for a specified boundary κ_0 as

$$\pi_{\kappa_0} = \Phi\left(\frac{\kappa_0 - \mu_D}{\sqrt{\sigma_D^2}}\right) - \Phi\left(\frac{-\kappa_0 - \mu_D}{\sqrt{\sigma_D^2}}\right), \tag{14}$$

where Φ is a cumulative normal distribution with mean 0 and variance 1.

The CP criterion also translates into an equivalence specification. Using the equivalence specification defined above, we can use the following hypothesis test to demonstrate equivalence:

$$H_0 : \pi_{10} \leq 0.90 \quad \text{vs.} \quad H_a : \pi_{10} > 0.90. \tag{15}$$

To test this hypothesis, one can construct a 95% lower bound on π_{10}. Equivalence is declared if the lower bound is greater than 0.90.

For specified κ_0 and π_0, the hypotheses $\kappa_{\pi_0} \geq \kappa_0$ and $\pi_{\kappa_0} \leq \pi_0$ are equivalent. Thus, individual agreement can be assessed using either one of the hypotheses in Equations (13) and (15). The choice whether to use the TDI or CP criterion should be based on whether the emphasis of the comparison method study is on the boundary or the percentage.

The hypothesis in Equation (13) allows one to control the consumer's risk. The consumer's risk is the probability of rejecting the inequivalence hypothesis (the null hypothesis) when it is true (Type I error) at a predetermined level (usually 5%). That is, the consumer's risk is the probability of accepting a truly bad competitor method. The hypothesis in Equation (15) also controls consumer's risk.

4. STATISTICAL INFERENCE

We discuss two methods for constructing confidence intervals on TDI and CP. The first method is based on a parametric bootstrap as proposed by Choudhary (2007) and the second method on the concept of generalized confidence intervals introduced by Tsui and Weerahandi (1989). Both methods are based on Monte Carlo simulation.

4.1. Bootstrap Confidence Intervals

The bootstrap method as proposed by Choudhary (2007) is based on the large sample property and the delta method. This method was also used in a study without repeated measurements by Choudhary and Nagaraja (2007). The method is performed by replacing the parameters μ_D, σ_D^2, and λ_D with their maximum likelihood estimators (MLE) and using some large sample properties. The MLEs are shown in Table 3. See the Appendix for more details.

To demonstrate, the algorithm used to compute bootstrap intervals on κ_{π_0} is as follows:

1. Compute the MLE for μ_D, γ_I, γ_E, σ_D^2, and λ_D for the collected data and denote them as $\tilde{\mu}_D$,$\tilde{\gamma}_I$, $\tilde{\gamma}_E$, $\tilde{\sigma}_D^2$, and $\tilde{\lambda}_D$ as indicated in Table 3.
2. Compute $\tilde{\kappa}_{\pi_0}$ using Equation (12) by replacing the parameters μ_D, σ_D^2, and λ_D with their respective MLE for the collected data. Also, compute $\tilde{\sigma}^2_{\ln\kappa}$ using Equation (28) in the Appendix.

Table 3 Maximum likelihood estimators

Result	Parameter	Maximum likelihood solutions
1	γ_I	$\tilde{\gamma}_I = \frac{(1-1/s)S_I^2 - S_E^2}{n}$
2	γ_E	$\tilde{\gamma}_E = S_E^2$
3	μ_D	$\tilde{\mu}_D = \overline{D}_{**}$
4	$\sigma_D^2 = \gamma_I + \gamma_E$	$\tilde{\sigma}_D^2 = \tilde{\gamma}_I + \tilde{\gamma}_E$
5	$\lambda_D = \frac{\mu_D^2}{\sigma_D^2}$	$\tilde{\lambda}_D = \frac{\tilde{\mu}_D^2}{\tilde{\sigma}_D^2}$

3. Generate a set of $\underline{D}^*$ by replacing the values μ_D, γ_I, γ_E with the MLE values computed in Step 1 in model (7). With the generated values $\underline{D}^*$, compute $\tilde{\kappa}^*_{\pi_0}$ using Equation (12) by replacing their MLE computed for the simulated values.
4. Using the sample estimate from Step 2 and the bootstrap estimate from Step 3, compute

$$Z^*_B = \frac{(\ln \tilde{\kappa}^*_{\pi_0} - \ln \tilde{\kappa}_{\pi_0})}{\sqrt{\tilde{\sigma}^2_{\ln \kappa}}}. \tag{16}$$

5. Repeat Steps 3 and 4, B times. In our simulation and example, we set $B = 1999$, as used by Choudhary (2007).
6. Order the B values obtained in Step 4 from least to greatest.
7. Select the value in the $B \times \alpha$ position of the ordered set in Step 7 and denote it by $z_B(\alpha)$.
8. Define the upper bound on κ_{π_0} for a $100(1-\alpha)\%$ interval as

$$U_\kappa = \exp\left(\ln \tilde{\kappa}_{\pi_0} - z_B(\alpha)\sqrt{\tilde{\sigma}^2_{\ln \kappa}}\right). \tag{17}$$

A $100(1-\alpha)\%$ lower bound on π_{κ_0} can be constructed graphically as follows. Generate values U_{κ_i} for different π_i, define the lower bound, L_π, as the smallest π_j with $U_{\kappa_j} \geq \kappa_0$.

4.2. Generalized Confidence Intervals

Generalized confidence intervals require the construction of generalized pivotal quantities. The construction of generalized confidence intervals on the TDI and CP criteria is performed by replacing the parameters μ_D, σ^2_D, and λ_D with generalized pivotal quantities (GPQs). The GPQ is an extension of the standard notion of a pivotal quantity. For more on the development and properties of the GCI method, the reader is referred to Weerahandi (1993, 1995), and Hannig et al. (2006).

We determine the GPQs following a method described in Burdick et al. (2005, Appendix B). One must first express the variance as a linear combination of the variance components as

$$\sigma^2_D = \frac{\theta_I + (n-1)\theta_E}{n}. \tag{18}$$

The GPQs are shown in Table 4, where $\overline{d}_{**}$, ss_I, and ss_E are realized values of $\overline{D}_{**}$, SS_I, and SS_E, respectively, Z_1 and Z_2 are independent normal random variables with mean zero and variance one, W_I and W_{I1} are independent chi-squared random variables with degrees of freedom $s-1$, and W_E is a chi-squared random variable with degrees of freedom $s(n-1)$.

To demonstrate, the algorithm used to compute a generalized confidence interval for κ_{π_0} or π_{κ_0} is as follows:

1. Compute $\overline{D}_{**}$, SS_I, and SS_E for the collected data and denote the realized values as $\overline{d}_{**}$, ss_I, and ss_E, respectively.

Table 4 Generalized pivotal quantities

Result	Parameter	GPQ
1	σ_D^2	$\text{GPQ}_{\sigma_D^2} = \left(\frac{ss_I}{W_I} + \frac{(n-1)ss_E}{W_E}\right)/n$
2	$\sigma_{\overline{D}}^2 = \theta_I/sn$	$\text{GPQ}_{\sigma_{\overline{D}}^2} = ss_I/snW_{I1}$
2	$\mu_D = \mu_T - \mu_R$	$\text{GPQ}_{\mu_D} = \overline{d}_{**} - Z_1\sqrt{\text{GPQ}_{\sigma_{\overline{D}}^2}}$
2	μ_D^2	$\text{GPQ}_{\mu_D^2} = \max\left[0, \overline{d}_{**}^2 - 2Z_2\lvert\text{GPQ}_{\mu_D}\rvert\sqrt{\text{GPQ}_{\sigma_{\overline{D}}^2}}\right]$
3	$\lambda_D = \frac{\mu_D^2}{\sigma_D^2}$	$\text{GPQ}_{\lambda_D} = \text{GPQ}_{\mu_D^2}/\sqrt{\text{GPQ}_{\sigma_{\overline{D}}^2}}$

2. Simulate N values of each GPQ shown in Table 4 by simulating N independent values each of Z_1, Z_2, W_I, W_{I1}, and W_E.
3. Compute N values of κ_{π_0} or π_{κ_0} (depending on the approach used) using Equations (12) or (14) by replacing unknown parameters with the GPQ formed in Step 2.
4. Order the N values obtained in Step 3 from least to greatest.
5. Define the upper bound on κ_{π_0}(lower bound on ϕ_{κ_0}) for a $100(1-\alpha)\%$ interval as the value in the position $N \times (1-\alpha)$ [$(N \times \alpha)$ for lower bound] of the ordered set in Step 4.

Similar to the bootstrap approach, an alternative way to construct a lower bound on CP is to generate upper bounds on TDI, $\widehat{U}_{\pi_i}$ for different values π_i, and define the lower bound on CP as the smallest value π_j with $\widehat{U}_{\pi_j} \geq \kappa_0$.

5. SIMULATION STUDY

Individual agreement as discussed in this paper is measured with either the TDI or the CP criterion. The procedure to measure individual agreement with either of the criteria is to calculate a 95% bound on the parameter of interest. In particular, when the agreement is determined using TDI with the hypothesis in Equation (13), the new method is accepted if the computed 95% upper bound on TDI is less than a specified value κ_0 (10 in Equation (13)). When agreement is determined using CP with the hypothesis in Equation (15), the new method is accepted if a 95% lower bound on CP is greater than a specified value π_0 (0.90 in Equation (15)). Since the exact distribution of these bounds is unknown, a simulation study was conducted to assess whether these bounds maintain the stated type I error probability of the tests. Since for specified values κ_0 and π_0, the hypotheses $\kappa_{\pi_0} \geq \kappa_0$ and $\pi_{\kappa_0} \leq \pi_0$ are equivalent, we conducted our simulation study based on the TDI.

Simulation designs were established by selecting values for $\mu_T - \mu_R$, σ_S^2, σ_{Ts}^2, σ_{Rs}^2,σ_T^2, and σ_R^2. We set $\sigma_S^2 = 100$ for all combinations because it is not required for constructing the bounds. The examined parameter combinations are shown in Table 5. TDI and CP were computed using Equations (12) and (14) and replacing unknown parameters with their respective values in the design. To estimate the type I error probability of the tests, we arbitrarily defined one acceptability criterion. The designs were chosen so that at least 93.5% of the paired differences in absolute value were less than 10. The value 93.5% was selected to facilitate selection of

Table 5 Simulation designs

$\mu_T - \mu_R$	σ^2_{Ts}	σ^2_{Rs}	σ^2_T	σ^2_R	TDI	CP	$\lambda_D = \frac{\mu^2_D}{\sigma^2_D}$
1.200	8.000	8.000	6.000	6.000	10.011	0.935	0.051
2.500	1.000	0.900	9.000	12.500	10.015	0.935	0.267
2.980	5.000	5.000	5.500	5.500	10.003	0.935	0.423
4.150	3.750	3.750	3.750	3.750	10.018	0.934	1.148
5.000	0.900	1.500	4.500	4.000	9.999	0.935	2.294
5.000	1.500	0.900	4.000	4.500	9.999	0.935	2.294
5.800	5.000	0.600	2.000	0.100	10.001	0.935	4.369
5.800	0.600	5.000	0.100	2.000	10.001	0.935	4.369
7.000	0.950	0.950	0.950	0.950	9.952	0.938	12.895
8.500	0.250	0.250	0.250	0.250	10.014	0.933	72.250
9.000	0.250	0.100	0.050	0.050	10.016	0.932	180.000

parameters. Simulations were conducted with combinations of $s = 10, 20, 40$ and $n = 2, 3, 5$. This resulted in a total of 99 designs. Table 5 reports the noncentrality parameter, λ_D, based on Equation (9). The range of λ_D for the simulation designs goes from 0.051 to 180.0. Thus, the designs used in the simulation study cover a wide range of parameters.

All simulations were conducted using SAS PROC IML as follows. One of the 99 designs was selected and 2,000 datasets were simulated using the SAS function RANNOR. The boostrap-t bounds were based on 1,999 bootstrap samples, and the GCI bounds were based on 10,000 GPQ values as described in the procedures outlined earlier. The probability based on the inverse noncentral chi-squared and univariate normal distributions in Equations (12) and (14) were computed using the SAS functions CINV and PROBNORM, respectively. The probability density function for the normal distribution, ϕ, was computed using the SAS function PDF with specified distribution NORMAL. The empirical power (or type I error probability of the test under the null hypothesis) was determined by counting the number of times the null hypothesis was rejected.

The simulation results for the stated type I error probability of 0.05 for TDI and different sample sizes are displayed in Fig. 1. The purpose of the simulation is to demonstrate that the type I error probability of the test (i.e., consumer's risk) is at most the stated test level of 0.05. By using the binomial model, if the true test size is 0.05, there is less than a 5% chance that an estimated test size based on 2,000 simulation will be greater than 0.06. If the test maintains the type I error probability, the maximum value in Fig. 1 should be (apart from simulation error [$\cong 0.01$]) the stated test size of 0.05. Some of the values exceed the type I error probability because the specification limits were not exactly equal to the specification limit of 10.

It appears that the bootstrap method has type I error probability greater than the stated level. Also, as the number of subjects increases, for fixed n, type I error probability for GCI approaches the stated level. That is, the type I error probability of GCI increases as the number of subjects increases. The same is true as the number of replicates increases for a fixed number of subjects. Simulation results were also completed for a design with $\kappa_0 = 0.223$ and $\pi_0 = 75\%$. These simulation results were consistent with those reported in this paper.

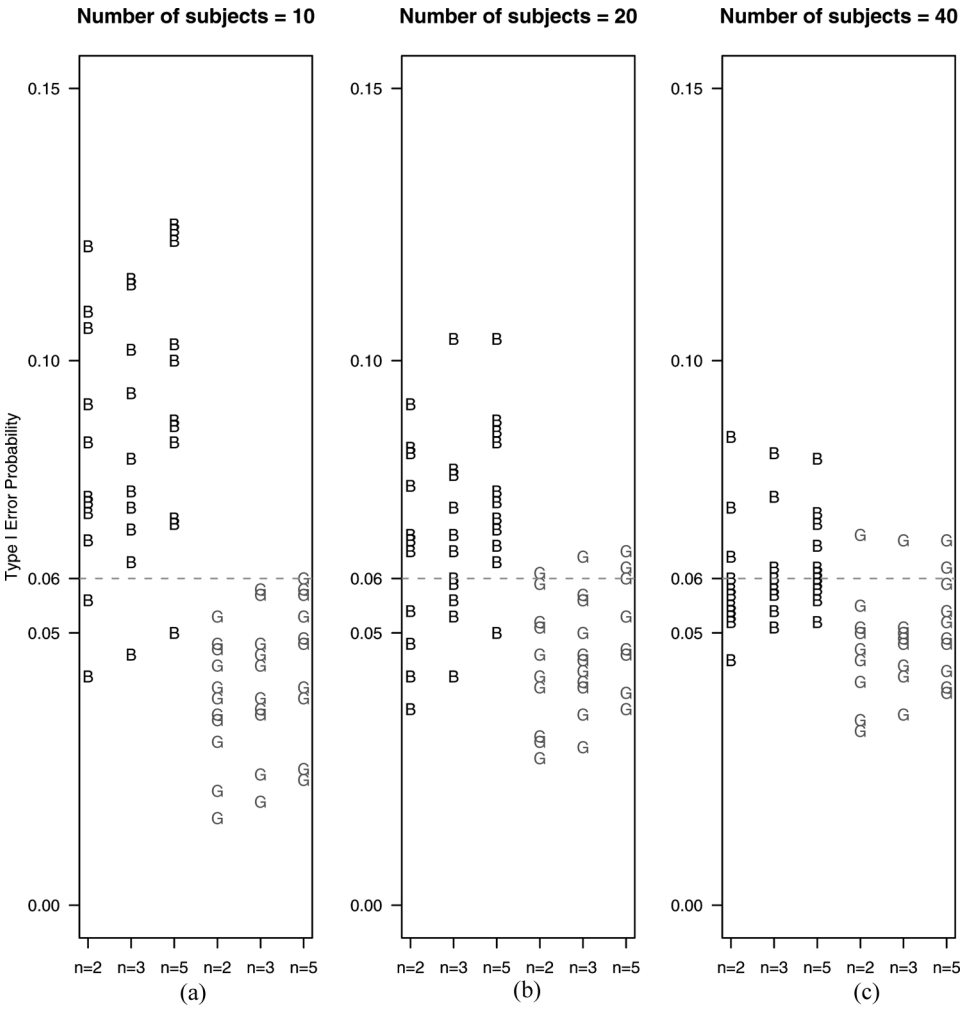

Figure 1 Side-by-side plots of type I error probability of tests by number of subjects and number of replicates. The results marked with "*B*" correspond to the bootstrap method, and "*G*" correspond to the generalized confidence interval method.

6. AN EXAMPLE APPLICATION

We demonstrate the procedures described in this article with a dataset used in Bland and Altman (1986). The goal of the study was to compare two methods of measuring peak expiratory flow rate (PEFR). The data consist of two paired measurements on the same subject made with a large Wright peak flow meter and a mini Wright meter. Paired differences that are less than 101/min are considered of no practical clinical significance. That is, paired difference of less than 101/min would not affect decisions on patient management. Therefore, the large meter can be replaced, or the two meters can be used interchangeably if a large proportion

Table 6 MLE computation for example

Term	Equation	Value in sample
$\tilde{\mu}_D$		5.971
$\tilde{\sigma}^2_D$	Table 3	1354.793
$\tilde{\lambda}_D$	(9)	0.026
$\tilde{\kappa}_{0.90}$	Table 3 and (12)	61.34
$\tilde{\sigma}^2_{\ln \kappa}$	(28)	301.851
$\tilde{\pi}_{10}$	Table 3 and (14)	0.211

(at least 90%) of the PEFR readings taken by the two meters on the same subject are within 101/min.

A serious error would be to declare the mini meter as effective as the large meter when it is not. Thus, it is important to put a cap on the consumer's risk and an equivalence testing setting is preferred. To demonstrate the method proposed in the paper, we test for equivalence using the above equivalence specification for hypotheses on the TDI and the CP.

From Table 6, the absolute difference between the sample means is 5.971, the estimated noncentrality parameter is 0.026, the estimated MLE $\kappa_{0.90}$ is 61.34, and the estimated MLE π_{10} is 21.1%. The estimated mean squares for the data are shown in Table 7.

The equivalence of the meters can be tested using the following hypothesis test on the TDI:

$$H_0 : \kappa_{0.90} \geq 10 \quad \text{vs.} \quad H_a : \kappa_{0.90} < 10. \tag{19}$$

With 10,000 generated GPQ's, a 95% GCI upper bound for $\kappa_{0.90}$ is 85.341/min. With 1,999 bootstrap samples, a 95% bootstrap-t upper bound for $\kappa_{0.90}$ is 79.141/min. Note that the fact the bootstrap bound is a "tighter" bound is likely due to an inflated type I error probability. Since both upper bounds are greater than 10, the meters are not equivalent based on either of the tests.

To demonstrate how to construct lower bounds on CP, consider the equivalence specification that the meters are equivalent if at least 90% of the absolute difference is less than 10. With this specification, we can test the equivalence of the meters using a hypothesis test on the CP:

$$H_0 : \pi_{10} \leq 0.90 \quad \text{vs.} \quad H_a : \pi_{10} > 0.90. \tag{20}$$

With 10,000 generated GPQ's, a 95% lower bound for π_{10} is 15.1%. Table 8 displays some computed upper bounds on TDI values for specified CP values based on 1,999

Table 7 ANOVA based on Table 2 for example

Source of variation	DF	MS
Subjects (I)	16	2209.90
Error (N)	17	629.68

Table 8 Upper bound on TDI for specified values CP

Specified CP	95% bootstrap upper bounds TDI	95% GCI upper bounds TDI
0.111	6.663	7.278
0.121	7.395	7.956
0.131	7.929	8.673
0.141	8.551	9.339
0.151	9.207	9.999
0.161	9.859	10.572
0.171	10.404	11.280
0.181	11.029	11.950
0.191	11.556	12.666
0.201	12.232	13.337
0.211	12.794	14.052

bootstrap samples. For illustration purposes, we include GCI upper bounds for the same CP values. Note that the result obtained with this approach is similar to the GCI approach where we directly computed the bound on CP. Table 8 indicates that a 95% lower bound on π_{10} lies between 16.1% and 17.1%. Using a linear regression line, the estimated 95% lower bound on π_{10} is 16.5%. Since both lower bounds are

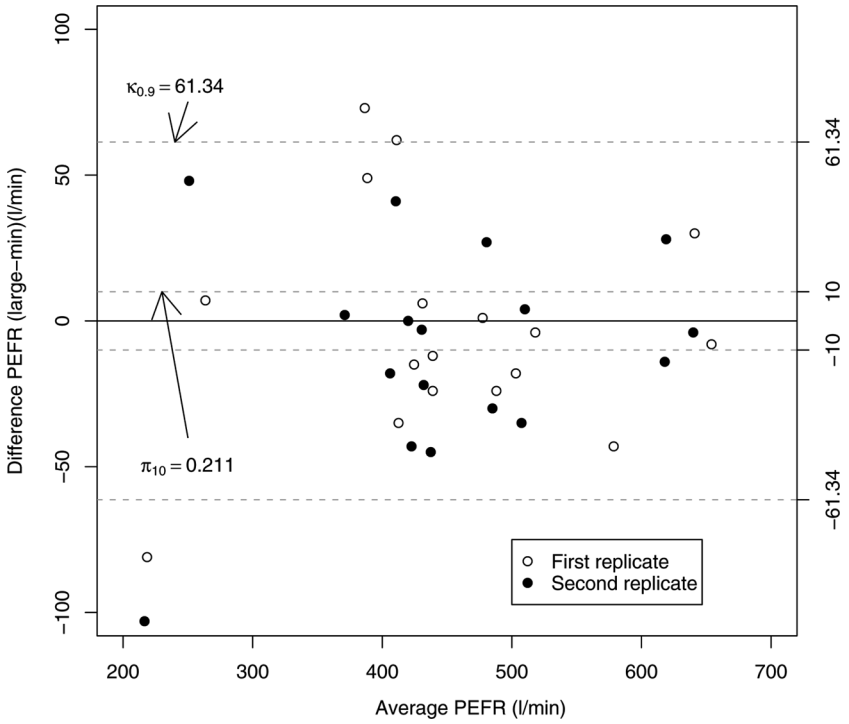

Figure 2 Difference vs. average plot. The TDI and CP values were calculated using the MLEs.

less than 90%, the mini meter is not equivalent to the large meter based on either of the bounds.

Both results indicate that there is high level of disagreement between the mini and large meters. This conclusion is also illustrated by Fig. 2 which displays a scatterplot of paired differences vs. averages of paired PEFR.

7. CONCLUSION

In this paper we propose an equivalence test for assessing individual agreement based on the total deviation index and the coverage probability. The bounds used in the tests were constructed using a bootstrap approach and generalized confidence intervals.

The simulation results suggest that, in general, the bounds constructed using the bootstrap approach do not maintain the stated test size for sample sizes of 10 and 20 subjects. The bootstrap method generally maintains the stated test size for sample size of 40 subjects. On the other hand, the generalized confidence intervals maintain the stated test size for all sample sizes of subjects considered in the study. Thus, the GCI approach is preferred for comparative studies with small samples and provides a good alternative for moderate to large samples.

The GCI approach, at least as presented in this paper, is applicable when repeated measurements for the two methods are paired and the true values of the measured quantities do not change over time. Like the definition of TDI and CP, the proposed methods also depend on the assumption of normality. The robustness of the proposed methods to these assumptions were not studied in our paper. However, it has been our experience that these assumptions are realistic in these applications.

APPENDIX

LARGE SAMPLE PROPERTY OF $\ln \tilde{\kappa}_\pi$

In this section, we derive a result used in the construction of the bootstrap-t confidence intervals. The bootstrap-t requires the maximum likelihood estimate for μ_D, γ_I, and γ_E. From Searle et al. (1992), the maximum likelihood estimates for μ_D, γ_I, γ_E are

$$\tilde{\mu}_D = \overline{D}_{**}, \tag{21}$$

$$\tilde{\gamma}_I = \tfrac{1}{n}\left((1 - 1/s)S_I^2 - S_E^2\right), \quad \text{and} \tag{22}$$

$$\tilde{\gamma}_E = S_E^2, \tag{23}$$

respectively, with large-sample variance–covariance matrix or inverse information matrix

$$\Im^{-1} = \frac{1}{s}\begin{bmatrix} \theta_I & 0 & 0 \\ 0 & 2\left(\frac{\theta_I^2}{n} + \frac{2\theta_E^2}{n(n-1)}\right) & -\frac{2\theta_E^2}{n-1} \\ 0 & -\frac{2\theta_E^2}{n-1} & \frac{2n\theta_E^2}{n-1} \end{bmatrix}, \tag{24}$$

where θ_I and θ_E are expected mean squares in Table 2.

From the multivariate delta method, the random variable

$$\frac{\ln \tilde{\kappa}_{\pi_0} - \ln \kappa_{\pi_0}}{\sqrt{\sigma^2_{\ln \kappa}}} \tag{25}$$

is asymptotically normally distributed with mean 0 and variance 1, where $\tilde{\kappa}_{\pi_0}$ is κ_{π_0} evaluated by replacing the MLE of μ_D, and σ^2_D, and

$$\sigma^2_{\ln \kappa} = G^\top \Im^{-1} G, \quad \text{where} \tag{26}$$

$$G = \begin{bmatrix} \frac{\partial \ln \kappa_{\pi_0}}{\partial \mu_D} \\ \frac{\partial \ln \kappa_{\pi_0}}{\partial \gamma_I} \\ \frac{\partial \ln \kappa_{\pi_0}}{\partial \gamma_E} \end{bmatrix} = \frac{1}{\kappa_\pi} \begin{bmatrix} \frac{\phi(q_u)-\phi(q_l)}{\phi(q_u)+\phi(q_l)} \\ \frac{\sqrt{\sigma^2_D}}{2\gamma_I}\left(\frac{q_u\phi(q_u)-q_l\phi(q_l)}{\phi(q_u)+\phi(q_l)}\right) \\ \frac{\sqrt{\sigma^2_D}}{2\gamma_E}\left(\frac{q_u\phi(q_u)-q_l\phi(q_l)}{\phi(q_u)+\phi(q_l)}\right) \end{bmatrix}, \tag{27}$$

where ϕ is the probability density function of a normal distribution with mean 0 and variance 1, $q_u = (\kappa_\pi - \mu_D)/\sqrt{\sigma^2_D}$, and $q_l = (-\kappa_\pi - \mu_D)/\sqrt{\sigma^2_D}$.

By simple matrix multiplication, we obtain

$$\sigma^2_{\ln \kappa} = \frac{1}{\kappa^2_\pi}\left[\Im^{-1}_{11}\left(\frac{\partial \kappa_{\pi_0}}{\partial \mu_D}\right)^2 + \Im^{-1}_{22}\left(\frac{\partial \kappa_{\pi_0}}{\partial \gamma_I}\right)^2 + \Im^{-1}_{33}\left(\frac{\partial \kappa_{\pi_0}}{\partial \gamma_E}\right)^2 + \Im^{-1}_{23}\left(\frac{\partial \kappa_{\pi_0}}{\partial \gamma_I}\right)\left(\frac{\partial \kappa_{\pi_0}}{\partial \gamma_E}\right)\right], \tag{28}$$

where $\Im^{-1}_{rc}$ is the entry located in the rth row and cth column of the matrix $\Im^{-1}$.

REFERENCES

Bland, J. M., Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet* 1(8476):307–310.

Barhhart, H. X, Haber, M. J., Lin, L. I. (2007). An overview on assessing agreement with continuous measurements. *Journal of Biopharmaceutical Statistics* 17:529–569.

Burdick, R. K., Borror, C. M., Montgomerry, D. C. (2005). Design and analysis of gauge *R&R* studies: making decisions with confidence intervals in random and mixed ANOVA models. ASA-SIAM Series on Statistics and Applied Probability, SIAM, Philadelphia, ASA, Alexandria, VA.

Choudhary, P. K. (2007). A tolerance interval approach for assessment of agreement in method comparison studies with repeated measurements. *Journal of Stastitical Planning and Inference* 138(4):1102–1115.

Choudhary, P. K., Nagaraja, H. N. (2007). Test for assessment of agreement using probability criteria. *Journal of Statistical Planning and Inference* 137:279–290.

Hannig, J., Iyer, H., Patterson, P. (2006). Fiducial generalized confidence intervals. *Journal of the American Statistical Association* 101:254–269.

Lin, L. (2000). Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence. *Statistics in Medicine* 19:255–270.

Lin, L., Hedayat, A. S., Yang, M. (2002). Statistical methods in assessing agreement: models, issues, and tools. *Journal of the American Statistical Association* 97:257–270.

Lin, L., Hedayat, A. S., Wu, W. (2007). A unified approach for assessing agreement for continous and categorical data. *Journal of Biopharmaceutical Statistics* 17:629–652.

Tsui, K. W., Weerahandi, S. (1989). Generalized *p*-Values in significance testing of hypotheses in the presence of nuisance parameters. *Journal of American Statistical Association* 84:602–607.

Weerahandi, S. (1993). Generalized confidence intervals. *Journal of the American Statistical Association* 88:899–905.

Weerahandi, S. (1995). *Exact Statistical Methods for Data Analysis*. New York: Springer-Verlag.