Assessing Individual Agreement

Huiman X. Barnhart and **Andrzej S. Kosinski**

Department of Biostatistics and Bioinformatics

and Duke Clinical Research Institute

Duke University

PO Box 17969

Durham, NC 27715

huiman.barnhart@duke.edu

Tel: 919-668-8403

Fax: 919-668-7049

and

Michael J. Haber

Department of Biostatistics

The Rollins School of Public Health

Emory University

Atlanta, GA 30322

mhaber@sph.emory.edu

Corresponding Author: Huiman X. Barnhart

SUMMARY

Evaluating agreement between measurement methods or between observers is important in method comparison studies and in reliability studies. Often we are interested in whether a new method can replace an existing invasive or expensive method, or whether multiple methods or multiple observers can be used interchangeably. Ideally, interchangeability is established only if individual measurements from different methods are similar to replicated measurements from the same method. This is the concept of individual equivalence. Interchangeability between methods is similar to bioequivalence between drugs in bioequivalence studies. Following the FDA guidelines on individual bioequivalence, we propose to assess individual agreement among multiple methods via individual equivalence using the moment criteria. In the case where there is a reference method, we extend the individual bioequivalence criteria to individual equivalence criteria and propose to use individual equivalence coefficient (IEC) to compare multiple methods to one or multiple references. In the case where there is no reference method available, we propose a new IEC to assess individual agreement between multiple methods. Furthermore, we propose a coefficient of individual agreement (CIA) that links the IEC with two recent agreement indices. A method of moments is used for estimation, where one can utilize output from ANOVA models. The nonparametric and bootstrap approaches are used for inference. Five examples are used for illustration.

KEY WORDS: agreement; method comparison; bioequivalence; individual equivalence; intraclass correlation coefficient; concordance correlation coefficient

1 Introduction

Evaluating agreement between methods or observers is important in method comparison studies and reliability studies. Oftentimes, we are interested in whether the observers can be used interchangeably, or whether a new method that is easy to use can replace an existing standard method that may be expensive or invasive. For example, when coronary artery calcium score is used to evaluate a patient's coronary artery atherosclerosis, it is important that different radiologists produce similar scores so that they can be used interchangeably. In physical therapy, different types of machines, such as manual goniometer and Lamoreux-type electrogoniometer, can be used to measure knee joint angle (Eliaziw et al., 1994) and one is interested in knowing whether the electogoniometer can replace the manual goniometer. In a carotid stenosis screening study (Barnhart and Williamson, 2001), one is interested in knowing whether the two new methods, 2 dimensional flight and 3 dimensional flight, using the technology of magnetic resonance angiography (MRA), can replace the standard invasive procedure, intra-arterial angiogram, in measuring carotid stenosis. In a blood pressure study (Bland and Altman, 1999), one is interested in whether an automatic blood pressure machine can replace human observers.

Traditionally, assessing agreement has been based on indices such as intraclass correlation coefficient (ICC) or concordance correlation coefficient (CCC) (McGraw and Wong, 1996; Carrasco and Jover, 2003, Lin, 1989; Lin, et al. 2002; Barnhart et al. 2005). These indices depend on between-subject variability. As illustrated in Figure 1 and by Atkinson and Nevill (1997), large between-subject variability would imply large value of ICC or CCC even if the individual difference between measurements by the two methods remains the same. Therefore, it is questionable whether the ICC or the CCC are adequate in establishing interchangeability of methods or observers. Ideally, interchangeability is established only if individual measurements from these methods are similar to replicated measurements within

a method. In other words, the individual difference between measurements from different methods is small so that this difference is close to the difference of replicated measurements within a method. This is the concept of individual equivalence. We note that the difference of replicated measurements can be summarized by within-subject variance. Therefore, we are interested in individual agreement through individual equivalence where the degree of individual agreement is defined as closeness between individual measurements relative to the within-subject variability.

Interchangeability between methods here is similar to individual bioequivalence or switchability between a test drug and a reference drug in individual bioequivalence studies. The concept of individual bioequivalence was first introduced by Anderson and Hauck (1990) to establish that the bioavailability of a new formulation is sufficiently close to that of the standard formulation in most individuals. A probability criteria was introduced there. Sheiner (1992) used a moment criteria to define individual bioequivalence. Schall and Luus (1993) extended their ideas and proposed general bioequivalence criteria that included both the probability criteria and the moment criteria as special cases. The Food and Drug Administration (FDA) modified and adopted the moment criteria in the recent FDA guidelines (2001) for establishing individual bioequivalence.

Similar to the FDA guidelines, in this paper, we propose to assess individual agreement between 2 or more methods using the moment criteria. We consider two situations: (1) a reference method exists; and (2) no reference method is available. In the case where there is a reference method, we extend the individual bioequivalence criteria in the FDA guidelines using individual equivalence coefficient (IEC) to compare multiple methods to a reference method. We also extend the individual equivalence criteria to the case with multiple references. In the case where there is no reference method available, we propose a new IEC to assess individual agreement between multiple methods. Furthermore, we propose a coefficient of individual agreement (CIA) that links the IEC with two recent agreement

indices (δ and ψ proposed by Shao and Zhong (2004) and Haber, et al. (2005) respectively), which may be used to assess individual agreement, a concept presented in this paper.

In section 2, we review the individaul bioequivalence criteria in the FDA guidelines and the two recent agreement indices. We present the relationships between these parameters under some assumptions for better understanding. In section 3, we present the new IECs and CIAs for comparison of multiple methods with and without a reference method. A method of moment is used for estimation where one can utilize output from ANOVA models. Nonparametric and Bootstrap approaches are used for inference. Five examples are used for illustration in section 4. We conclude with a discussion in section 5.

2 Review of Individual Bioequivalence and Agreement Indices

2.1 Existence of a Reference

We first introduce the FDA guidelines for assessing individual bioequivalence between two drugs, a test drug T and a reference drug R. Let Y_{iT} and Y_{iR} be the measurements, e.g., logarithm of bioavailability, from the ith subjects after taking test drug T and reference drug R respectively. To establish bioequivalence at the individual level, the individual difference between responses from the test and reference drugs is compared to the difference between two replicated responses from the reference drug. FDA (2001) compared the mean of the squared differences between responses from test and reference drugs to the mean of the squared differences between two responses from the reference drug. The reference-scaled individual bioequivalence criterion is defined as

$$IBC = \frac{E(Y_{iT} - Y_{iR})^2 - E(Y_{iR} - Y_{iR'})^2}{E(Y_{iR} - Y_{iR'})^2)/2} \leq \theta_I$$

where the left hand side defines individual equivalence coefficient (IBC), $Y_{iR'}$ is a replication of Y_{iR} and θ_I is the bioequivalence limit set by the regulatory agency.

The measurement Y_{ij} is often re-written as the sum of true value μ_{ij} and random error ϵ_{ij}, i.e., $Y_{ij} = \mu_{ij} + \epsilon_{ij}, j =$T, R, with the following common assumptions: μ_{ij} and ϵ_{ij} are independent with means $E(\mu_{ij}) = \mu_j$ and $E(\epsilon_{ij}) = 0$, and between-subject and within-subject variances of $Var(\mu_{ij}) = \sigma^2_{Bj}$ and $Var(\epsilon_{ij}) = \sigma^2_{Wj}$, respectively. Under this model, the reference scaled individual equivalence criteria can be re-written by using the population parameters as

$$IBC = \frac{(\mu_T - \mu_R)^2 + \sigma^2_D + \sigma^2_{WT} - \sigma^2_{WR}}{\sigma^2_{WR}} \leq \theta_I$$

where $\sigma^2_{WT} = Var(\epsilon_{iT})$ and $\sigma^2_{WR} = Var(\epsilon_{iR})$ are within-subject variances for T and R respectively, σ^2_D is subject-by-formulation interaction variance component defined as $\sigma^2_D = Var(\mu_{iT} - \mu_{iR}) = (\sigma_{BT} - \sigma_{BR})^2 + 2(1 - \rho_\mu)\sigma_{BT}\sigma_{BR}$ with $\rho_\mu = corr(\mu_{iT}, \mu_{iR})$ and between-subject variabilities $\sigma^2_{BT} = Var(\mu_{iT})$ and $\sigma^2_{BR} = Var(\mu_{iR})$. We note that the following three components (relative to σ^2_{WR}) affect the value of IBC simultaneously: (1) difference of population means $\mu_T - \mu_R$; (2) difference of within-subject variances $\sigma^2_{WT} - \sigma^2_{WR}$; and (3) subject-by-formulation interaction σ^2_D. The between-subject variances σ^2_{BT} and σ^2_{BR} have impact on IBC only through the interaction term σ^2_D.

The FDA guidelines also consider a constant-scaled IBC, which uses a constant $\sigma^2_{W_0}$ in place of σ^2_{WR} in the denominator when $\sigma^2_{WR} < \sigma^2_{W_0}$. We will discuss issues related to constant scaling in the discussion section.

2.2 No Reference

Shao and Zhong (2004) proposed an equivalence criterion for assessing agreement between two methods where none of the methods is considered as a reference. They compared the

conditional mean of individual difference between responses from two methods relative to the conditional variance of the individual difference, conditional on the true value of the subject. Let Y_{ij} be a measurement for subject i from method $j, j = 1, 2$. Shao and Zhong (2004) defined the agreement index δ as

$$\delta = \frac{E(E(Y_{i1} - Y_{i2})^2 | i\text{th true value})}{E(Var(Y_{i1} - Y_{i2}) | i\text{th true value})}.$$

A satisfactory agreement between methods 1 and 2 corresponds to $\delta \leq \delta_0$ where δ_0 is a pre-specified positive constant. Using the above model of $Y_{ij} = \mu_{ij} + \epsilon_{ij}$ with j =1, 2, and further assuming that conditioning on the ith subject's true value corresponds to conditioning on μ_{i1} and μ_{i2}, then $E(E(Y_{i1} - Y_{i2})^2 | i\text{th true value}) = E(\mu_{i1} - \mu_{i2})^2 = (\mu_1 - \mu_2)^2 + \sigma_D^2$ and $E(Var(Y_{i1} - Y_{i2}) | i\text{th true value}) = \sigma_{W1}^2 + \sigma_{W2}^2$. Therefore,

$$\delta = \frac{(\mu_1 - \mu_2)^2 + \sigma_D^2}{\sigma_{W1}^2 + \sigma_{W2}^2}.$$

We see that δ is the ratio of expected individual true difference $E(\mu_{i1} - \mu_{i2})^2$ to the sum of the two within-subject variances. In other words, δ compares the individual sqaured difference between the two methods to the random variability due to replication. Therefore, δ may be considered as an index for assessing individual agreement.

The IBC and δ differ in the following ways: (1) the IBC is developed when one of the methods is a reference, and the δ is developed when neither methods can be considered as a reference; (2) the IBC uses the expected individual squared difference $E(Y_{iT} - Y_{iR})^2$ at the observed level rather than $E(\mu_{i1} - \mu_{i2})^2$ at the true level; (3) the IBC uses subtraction (with a scaling factor) rather than ratio when comparing individual difference to the within-subject variance. Therefore, when comparing individual sqaured difference to the within-subject variance, only the within-subject variance from the reference method is used in IBC, while both within-subject variances are used for this comparison in δ.

Despite the differences between the IBC and δ, they are mathematically related. If we denote the first method as T and the second method as R even though R is not considered

as a reference, both the IBC and δ are functions of $E(\mu_{iT} - \mu_{iR})^2$ and the within-subject variances from both methods. They have the following relationship:

$$IBC = \frac{\sigma^2_{WT} + \sigma^2_{WR}}{\sigma^2_{WR}}(\delta + \frac{\sigma^2_{WT} - \sigma^2_{WR}}{\sigma^2_{WT} + \sigma^2_{WR}}).$$

Under the assumption of equal within-subject variances: $\sigma^2_{WT} = \sigma^2_{WR}$, i.e., $\sigma^2_{W1} = \sigma^2_{W2}$, we have $IBC = 2\delta$.

Haber et al. (2005) proposed an agreement index ψ for assessing agreement between J observers without reference. For comparison purposes, the J observers are treated as J methods here and we pay special attention to the case of $J = 2$ with two methods. Let index $j, j = 1, \ldots, J$ to denote jth method. Using the same model $Y_{ij} = \mu_{ij} + \epsilon_{ij}$, Haber et al. (2005) defined (true) individual inter-method variability for subject i as the between-method variance of the true values μ_{ij}, $\tau_i^2 = \sum_j (\mu_{ij} - \mu_{i\bullet})^2/(J-1)$. They defined an agreement index ψ by comparing the expected individual inter-method variability $\tau_*^2 = E(\tau_i)$ to the average of within-subject variances $\sigma_*^2 = \sum_j \sigma^2_{Wj}/J$, scaled to be between 0 and 1.

$$\psi = \frac{\sigma_*^2}{\tau_*^2 + \sigma_*^2}.$$

We can see that ψ can be used as an index to assess individual agreement because it compares individual difference relative to within-subject variance. To understand how inter-method variability is related to pairwise differences, we note the following relationship

$$\sum_j (\mu_{ij} - \mu_{i\bullet})^2/(J-1) = \sum_{j=1}^{J-1} \sum_{j'=j+1}^{J} (\mu_{ij} - \mu_{ij'})^2/(J(J-1)).$$

For $J = 2$, we have $\tau_i^2 = (\mu_{i1} - \mu_{i2})^2/2$, half of the individual difference at the true value level. In this case, we have $2\tau_*^2 = E(\mu_{i1} - \mu_{i2})^2 = (\mu_1 - \mu_2)^2 + \sigma_D^2$ and $2\sigma_* = \sigma^2_{W1} + \sigma^2_{W2}$. Thus,

$$\psi = \frac{\sigma^2_{W1} + \sigma^2_{W2}}{(\mu_1 - \mu_2)^2 + \sigma_D^2 + \sigma^2_{W1} + \sigma^2_{W2}}.$$

Like index δ, ψ differs from the IBC the same way as δ differs from the IBC. Both ψ and δ are developed when none of the methods are considered as reference, and they are related

by the following relationship:

$$\psi = \frac{1}{\delta + 1}, \quad \text{or} \quad \delta = \frac{1-\psi}{\psi}.$$

We denote the first method as T and the second method as R, even though the second method is not considered as a reference here. The IBC and ψ have the following mathematical relationship:

$$IBC = 2\frac{\tau_*^2 + \sigma_*^2 - \sigma_{WR}^2}{\sigma_{WR}^2} = 2(\frac{\sigma_*^2}{\sigma_{WR}^2}\frac{1}{\psi} - 1)$$

or equivalently

$$\psi = \frac{\sigma_{*\bullet}^2}{\sigma_{WR}^2}\frac{2}{(IBC+2)}.$$

Under the additional assumption of equal within-subject variances: $\sigma_{WT}^2 = \sigma_{WR}^2$, we have

$$IBC = \frac{2(1-\psi)}{\psi}, \quad \text{or} \quad \psi = \frac{2}{(IBC+2)}.$$

The FDA (2001) recommended to use IBC bound of $\theta_I = [(log(1.25))^2 + 0.05]/0.2^2 = 2.494827$ for declaring individual bioequivalence when $IBC \leq \theta_I$. Under the assumption of equal within-subject variances, i.e., $\sigma_{WT}^2 = \sigma_{WR}^2$ or $\sigma_{W1}^2 = \sigma_{W2}^2$, this bound corresponds to $\delta_0 = \theta_I/2 = 1.2474$ and index $\psi \geq 0.445$ for assessing individual agreement. Note that if $\tau_*^2 = \sigma_*^2$, i.e., the true inter-method variability is the same as the average of within-subject variability σ_*^2, e.g., the (expected) true individual squared difference between the test and reference methods is the same as the expected squared difference due to replication, then we have $\psi = 0.5$. Using the FDA's criteria for ψ, i.e., $\psi \geq 0.445$, it implies that the inter-method variability (τ_*^2) is within 125% of the within-subject variance (σ_*^2).

In summary, the IBC can be used to assess individual agreement via individual equivalence between two methods where one of them is considered as a reference. The agreement indices δ and ψ can be used to assess individual agreement between two methods via individual equivalence between two methods where none of them are considered as a reference.

When the within-subject variance based on the reference method is the same as the within-subject variance based on the test method, the IBC and the agreement indices δ and ψ have simple one-to-one relationships and their interpretations complement each other. In practice, there may be more than one test method, e.g, the two new methods in the carotid stenosis screening study, that need to be compared with the reference method. If there is no reference, one can use the agreement index ψ developed for comparison between these multiple methods. The natural questions are (1) how to extend the IBC and ψ to compare multiple methods versus a reference, and (2) whether one can extend IBC to compare multiple methods without reference. These questions are addressed in the next section.

3 Assessing Individual Agreement between Multiple Methods

3.1 Existence of a Reference Method

Suppose that there is a total of J methods with the first $J-1$ methods as new methods and the Jth method as a reference method. For the ith individual, let Y_{ij} be the measurement from the jth method, and Y_{iJk} and $Y_{iJk'}$ be the replicated measurements from the reference method.

Similar to FDA's individual bioequivalence criteria, we propose to assess individual agreement between $J-1$ methods against a reference method by using the individual equivalence coefficient (IEC):

$$IEC^R = \frac{(\sum_{j=1}^{J-1} E(Y_{ij} - Y_{iJ})^2)/(J-1) - E(Y_{iJk} - Y_{iJk'})^2}{E(Y_{iJk} - Y_{iJk'})^2/2}, \tag{1}$$

where superscript R indicates that a reference is utilized. Using the model $Y_{ij} = \mu_{ij} + \epsilon_{ij}$

with the same assumption as in section 2, the above can be re-written as

$$IEC^R = \frac{\sum_{j=1}^{J-1}(\mu_j - \mu_J)^2 + \sum_{j=1}^{J-1}\sigma^2_{D_{jJ}} + \sum_{j=1}^{J-1}\sigma^2_{Wj} - (J-1)\sigma^2_{WJ}}{(J-1)\sigma^2_{WJ}}, \tag{2}$$

where $\sigma^2_{D_{jJ}} = Var(\mu_{ij} - \mu_{iJ})$. We use acronym IEC rather than IBC because it is intended for use in any continuous measurement rather than restricted to bioavailability measures.

We propose the following coefficient of individual agreement (CIA) that is similar to ψ in section 2.2 by treating the Jth method as reference.

$$CIA^R = \psi^R = \frac{E(Y_{iJk} - Y_{iJk'})^2}{\sum_{j=1}^{J-1} E(Y_{ij} - Y_{iJ})^2/(J-1)} = \frac{\sigma^2_{WJ}}{\tau^2_{*R} + \sigma^2_{*R}},$$

where the true inter-method variability is

$$\tau^2_{*R} = E(\frac{\sum_{j=1}^{J-1}(\mu_{ij} - \mu_{iJ})^2}{2(J-1)}) = \frac{1}{2}(\frac{\sum_{j=1}^{J-1}(\mu_j - \mu_J)^2}{J-1} + \frac{\sum_{j=1}^{J-1}\sigma^2_{D_{jJ}}}{J-1}),$$

and the weighted average of within-subject variability σ^2_{*R} is

$$\sigma^2_{*R} = \frac{1}{2}(\frac{\sum_{j=1}^{J-1}\sigma^2_{Wj}}{J-1} + \sigma^2_{WJ}).$$

In practice, the within-subject variance in the reference method is likely to be smaller than the ones from the new methods, i.e., $\sigma^2_{WJ} \leq \sigma^2_{Wj}$, thus, we have $0 \leq \psi^R \leq 1$. Otherwise, ψ^R may be greater than 1. With these new definitions, the relationship between IEC^R and ψ^R is as follows,

$$IEC^R = \frac{2(\tau^2_{*R} + \sigma^2_{*R}) - 2\sigma^2_{WJ}}{\sigma^2_{WJ}} = \frac{2(1-\psi^R)}{\psi^R}), \quad \text{, or} \quad \psi^R = \frac{2}{IEC^R + 2}.$$

If we use IEC^R_{jJ} to denote the IEC value comparing the jth method and the reference, then the overall IEC^R is the average of these pairwise IECs:

$$IEC^R = \sum_{j=1}^{J-1} IEC^R_{jJ}/(J-1). \tag{3}$$

For $J = 2$, IEC^R reduces to IBC. In this case, τ^2_{*R} and and σ^2_{*R} are the same as the τ^2_* and σ^2_* (see section 2.2) defined by Haber et al. (2005) when there is no reference.

In general, we want to have low value of IEC^R and high value of ψ^R to claim satisfactory individual agreement. One may use the FDA recommended boundary of $\theta_I = 2.4948$ or equivalently, to have $\psi^R \geq 0.445$ for good individual agreement. Several factors can contribute to unsatisfactory individual agreement: (1) population means from the test methods are different from the mean from the reference method; (2) within-subject variances from the test methods are different from the within-subject variance from the reference method; (3) inter-method variability is large, which may be caused by the difference in population means or subject-by-method interaction $\sigma^2_{D_R}$, where $\sigma^2_{D_R} = \sum_{j=1}^{J-1} \sigma^2_{D_{jR}}/(J-1)$. Therefore, when reporting estimates on IEC^R and CIA^R, it may be useful to report estimates on $\mu_j, \sigma^2_{Wj}, j = 1, \ldots, J$, $\tau^2_{*R}, \sigma^2_{D_R}$, and σ^2_{*R}.

Estimation and Inference

To estimate IEC^R using equation (2), replicated measurements for each individual by each method are needed in order to estimate σ^2_{Wj}, σ^2_{WR} and $\sigma^2_{D_{jR}}$. In bioequivalence studies, a cross-over design is usually used in order to obtain replications under the assumption of no carry-over effect. However, in agreement studies, replications can normally be obtained with simple parallel design because the methods usually do not have a lasting effect on the individual. With the parallel design, let $Y_{ijk}, i = 1, \ldots, n, j = 1, \ldots, J, k = 1, \ldots, K$ be the observed measurements for individual i, method j and replication k. For simplicity, we assume equal number of replications within each subject and method even though the approach can be applied to unequal number of replications. The method of moment can be used to estimate IEC^R and ψ^R. Specifically, the unbiased estimates for within-subject variances are as follows,

$$\hat{\sigma}^2_{Wj} = MSE_{Wj} = \frac{\sum_{ik}(Y_{ijk} - Y_{ij\bullet})^2}{n * (K-1)}, j = 1, \ldots, J,$$

Note that

$$E[(Y_{ij\bullet} - Y_{iJ\bullet})^2] = E[(\mu_{ij} - \mu_{iJ})^2] + \frac{\sigma^2_{Wj}}{K} + \frac{\sigma^2_{WJ}}{K}.$$

Thus, the unbiased estimates for τ^2_{*R} and σ^2_{*R} are

$$\hat{\tau}^2_{*R} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{J-1}(Y_{ij\bullet} - Y_{iJ\bullet})^2}{2(J-1)n} - \frac{\sum_{j=1}^{J-1} MSE_{Wj}}{2(J-1)K} - \frac{MSE_{WJ}}{2K},$$

$$\hat{\sigma}^2_{*R} = (\frac{\sum_{j=1}^{J-1} MSE_{Wj}}{J-1} + MSE_{WJ})/2.$$

Therefore, we have

$$\hat{IEC}^R = \frac{2(\hat{\tau}^2_{*R} + \hat{\sigma}^2_{*R} - MSE_{WJ})}{MSE_{WJ}}, \quad \hat{\psi}^R = \frac{MSE_{WJ}}{\hat{\tau}^2_{*R} + \hat{\sigma}^2_{*R}}.$$

To see how the subject-by-method interaction affects the inter-method variability τ^2_{*R}, we can also calculate an estimate for $\sigma^2_{D_R}$ as

$$\hat{\sigma}^2_{D_R} = 2\hat{\tau}^2_{*R} - \frac{\sum_j(\hat{\mu}_j - \hat{\mu}_J)^2}{J-1},$$

where $\hat{\mu}_j = Y_{\bullet j\bullet}, \hat{\mu}_J = Y_{\bullet J\bullet}$ are the estimates for the population means.

We note that the above estimate for ψ^R can be re-written as

$$\hat{\psi}^R = \frac{2A_{\bullet J}}{B^R_\bullet} = \frac{2\sum_{i=1}^{n} A_{iJ}}{\sum_{i=1}^{n} B^R_i}$$

where $A_{\bullet J}$ and $B^R_\bullet$ are means of i.i.d. random variables A_{iJ} and B^R_i, respectively, with

$$A_{ij} = \frac{\sum_k (Y_{ijk} - Y_{ij\bullet})^2}{K-1}, i = 1, \ldots, n, j = 1, \ldots, J$$

and

$$B^R_i = \frac{\sum_{j=1}^{J-1}(Y_{ij\bullet} - Y_{iJ\bullet})^2}{J-1} + (1 - \frac{1}{K})\frac{\sum_{j=1}^{J-1} A_{ij}}{J-1} + (1 - \frac{1}{K})A_{iJ}, i = 1, \ldots, n.$$

We use the delta method to estimate the variance of a ratio:

$$Var(\frac{A_{\bullet J}}{B^R_\bullet}) \approx (\frac{A_{\bullet J}}{B^R_\bullet})^2[\frac{Var(A_{\bullet J})}{A^2_{\bullet J}} + \frac{Var(B^R_\bullet)}{(B^R_\bullet)^2} - \frac{2Cov(A_{\bullet J}, B^R_\bullet)}{A_{\bullet J}B^R_\bullet}],$$

where $Var(A_{\bullet J}), Var(B^R_\bullet)$ and $Cov(A_{\bullet J}, B^R_\bullet)$ can be estimated empirically, e.g., $\hat{Var}(A_{\bullet J}) = \hat{Var}(A_{iJ})/n = \sum_{i=1}^{n}(A_{iJ} - A_{\bullet J})^2/(n(n-1)), \hat{Var}(B^R_\bullet) = \hat{Var}(B^R_i)/n = \sum_{i=1}^{n}(B^R_i - B^R_\bullet)^2/(n(n-1))$ and $\hat{Cov}(A_{\bullet J}, B^R_\bullet) = \hat{Cov}(A_{iJ}, B^R_i)/n = \sum_{i=1}^{n}(A_{iJ} - A_{\bullet J})(B^R_i - B^R_\bullet)/(n(n-1))$. Thus, a

non-parametric estimate for the s.e. of $\hat{\psi}^R$ is $s.e.(\hat{\psi}^R) = 2 * \sqrt{\hat{Var}(A_{\bullet J}/B^R_{\bullet})}$. One can then obtain the 95% confidence interval (CI) as $(\hat{\psi}^R - 1.96 * s.e.(\hat{\psi}^R), \hat{\psi}^R + 1.96 * s.e.(\hat{\psi}^R)$.

The bootstrap percentile method can also be used to obtain 95% CI for IEC^R and ψ^R because it is easy and fast to compute these estimates. Specifically, m (say 10,000) samples with replacement can be taken from the n subjects where the sampling unit is subject, not measurement. We then apply the above estimation method and obtain m estimates of IEC^R and ψ^R. The lower 2.5percentiles of the ψ^R estimates are the 95% CI for ψ^R.

The estimates of $\hat{IEC}^R$ and $\hat{\psi}^R$ can also obtained via ANOVA models that require minimum programming. Specifically, the sums of squares from two sets of ANOVA models can be used to compute MSE_{Wj}, MSE_{WJ}, $\hat{\tau}^2_{*R}$, and thus $\hat{IEC}^R$ and $\hat{\psi}^R$. The first set of the J ANOVA models is used to obtain MSE_{Wj}, MSE_{WJ} and the second set of the $(J-1)$ ANOVA models is used to obtain $\hat{\tau}^2_{*R}$. The MSE_{Wj} corresponds to the mean square error terms in the following first set of J one way ANOVA models

$$Y_{ijk} = \mu + \alpha_i + \epsilon_{ijk}, j = 1, \ldots, J.$$

For each method j, we fit the second set of $(J-1)$ two way ANOVA models (without main effect for the method) for measurements made only by the jth method and the Jth method (reference method).

$$Y_{ij'k} = \mu + \alpha_i + \gamma_{ij'} + \epsilon_{ij'k}, j' = j, \quad \text{or} \quad J$$

Let MS_{jJ} and MSE_{jJ} denote the mean squares for the interaction term γ_{ij} and the error term ϵ_{ijk}, respectively. It can be shown that

$$\frac{MS_{jJ} - MSE_{jJ}}{K} = \frac{\sum_i (Y_{ij\bullet} - Y_{iR\bullet})^2}{2n} - \frac{MSE_{Wj} + MSE_{WJ}}{2K}.$$

Thus, we have

$$\hat{\tau}^2_{*R} = \frac{\sum_j (MS_{jJ} - MSE_{jJ})}{K(J-1)}.$$

Based on definition of IEC^R in equation (1), it is not necessary to have replications in the test methods in order to estimate IEC^R. However, replications from the reference method are needed. Let Y_{ij} be the measurement for subject i by method j and Y_{iJk} be the measurement for subject i by reference method with replication k. Then we can estimate σ^2_{WJ} as above and estimate IEC^R and ψ^R by

$$\hat{IEC}^R = \frac{\sum_{j=1}^{J-1} \hat{E}(Y_{ij} - Y_{iJ})^2/(J-1) - 2\hat{\sigma}^2_{WJ}}{\hat{\sigma}^2_{WJ}}, \quad \hat{\psi}^R = \frac{2}{\hat{IEC}^R + 2},$$

where $\hat{E}(Y_{ij} - Y_{iJ})^2 = \sum_k (Y_{ij} - Y_{iJk})^2/K$. Note that we would not be able to obtain estimates for $\sigma^2_{D_R}$ and σ^2_{*R}. We illustrate this approach with example 5 in section 4. The two estimation approaches based on equations (1) and (2) should yield similar results because the common assumptions on model $Y_{ij} = \mu_{ij} + \epsilon_{ij}$ (see section 2) are usually reasonable in practice.

Extension to Multiple References

In practice, there may be multiple reference methods available. For example, in the blood pressure data example from Bland and Altman (1999), the new automatic machine is compared to two human observers, where both human observers are treated as a reference. Suppose that there are J new methods and R multiple references with a total of $J + R$ methods, we extend IEC^R and ψ^R as follows:

$$IEC^R = \frac{(\sum_r \sum_j E(Y_{ij} - Y_{ir})^2)/(JR) - \sum_r E(Y_{irk} - Y_{irk'})^2/R}{\sum_r E(Y_{irk} - Y_{irk'})^2/(2R)}, \quad \psi^R = \frac{2}{IEC^R + 2}.$$

If we use model $Y_{ij} = \mu_{ij} + \epsilon_{ij}$ with the assumptions in section 2.1, we have

$$\begin{aligned} IEC^R &= \frac{\sum_r \sum_j (\mu_j - \mu_r)^2/(JR) + \sum_r \sum_j \sigma^2_{D_{jr}}/(JR) + \sum_j \sigma^2_{Wj}/J - \sum_r \sigma^2_{Wr}/R}{\sum_r \sigma^2_{Wr}/R} \\ &= \frac{2(\tau^2_{*R} + \sigma^2_{*R}) - 2\sum_r \sigma^2_{Wr}/R}{\sum_r \sigma^2_{Wr}/R} == \frac{2(1-\psi^R)}{\psi^R}, \end{aligned}$$

where the inter-method variability τ^2_{*R} and weighted within-subject variability σ^2_{*R} are defined as

$$\tau^2_{*R} = \frac{1}{2}\frac{\sum_r \sum_j E(\mu_{ij} - \mu_{ir})^2}{JR} = \frac{1}{2}\left(\frac{\sum_r \sum_j (\mu_j - \mu_r)^2}{JR} + \frac{\sum_r \sum_j \sigma^2_{D_{jr}}}{JR}\right),$$

$$\sigma^2_{*R} = \frac{1}{2}(\frac{\sum_j \sigma^2_{Wj}}{J} + \frac{\sum_r \sigma^2_{Wr}}{R}).$$

The ψ^R can be re-written as

$$\psi^R = \frac{\sum_r E(Y_{irk} - Y_{irk'})^2}{\sum_r \sum_j E(Y_{ij} - Y_{ir})^2/J} = \frac{\sum_r \sigma^2_{Wr}/R}{\tau^2_{*R} + \sigma^2_{*R}}.$$

Estimation and reference on IEC^R and ψ^R can be carried out similarly as described above for data with replications by both new and reference methods or for data with replications only by the reference methods.

3.2 No Reference Method

If there is a total of J methods and none of them can be considered as a reference, we compare the average of all possible squared individual differences between methods to the average of J within-subject variances from these methods. Specifically, we propose to assess individual agreement between J methods by IEC or equivalentaly CIA with

$$\begin{aligned} IEC^N &= \frac{2(\sum_{j=1}^{J-1}\sum_{j'=j+1}^{J} E(Y_{ij} - Y_{ij'})^2)/(J(J-1)) - \sum_j E(Y_{ijk} - Y_{ijk'})^2/J}{\sum_j E(Y_{ijk} - Y_{ijk'})^2/(2J)} \\ &= \frac{2(\sum_{j=1}^{J-1}\sum_{j'=j+1}^{J}((\mu_j - \mu_{j'})^2 + \sigma^2_{D_{jj'}} + \sigma^2_{Wj} + \sigma^2_{Wj'}))/(J(J-1)) - 2\sigma^2_*}{\sigma^2_*} = \frac{2\tau^2_*}{\sigma^2_*}, \end{aligned}$$

and

$$CIA^N = \psi^N = \frac{\sum_{j=1}^{J} E(Y_{ijk} - Y_{ijk'})^2/2}{\sum_{j=1}^{J-1}\sum_{j'=j+1}^{J} E[(Y_{ij} - Y_{ij'})^2]/(J-1)} = \frac{\sigma^2_*}{\tau^2_* + \sigma^2_*}$$

where τ^2_* and σ^2_* are the true inter-method variability and within-method variability as defined in Haber, et al. (2005),

$$\tau^2_* = \frac{E(\sum_j(\mu_{ij} - \mu_{i\bullet})^2)}{J-1} = \frac{1}{2}(\frac{2\sum_{j=1}^{J-1}\sum_{j'=j+1}^{J}(\mu_j - \mu_{j'})^2}{J(J-1)} + \sigma^2_D), \quad \sigma^2_* = \sum_j \sigma^2_{Wj}/J,$$

with $\sigma^2_D = 2\sum_{j=1}^{J-1}\sum_{j'=j+1}^{J}\sigma^2_{D_{jj'}}/(J(J-1))$ and $\sigma^2_{D_{jj'}} = Var(\mu_{ij} - \mu_{ij'})$. We note that ψ^N is the same as the agreement index ψ defined in Haber et al. (2005). In general, $0 \leq \psi^N \leq 1$

and the relationship between IEC^N and ψ^N is the same as in the case when there is a reference method, i.e.,

$$IEC^N = \frac{2(1-\psi^N)}{\psi^N}, \quad \text{or} \quad \psi^N = \frac{2}{IEC^N + 2}.$$

The interpretation of IEC^N and ψ^N are similar to IEC^R and ψ^R in section 3.1.

Let $IEC^N_{jj'}$ denote the pairwise IEC comparing jth and j'th methods without reference. Then one can show that the overall IEC^N is the weighted average of the pairwise $IEC^N_{jj'}$'s,

$$IEC^N = \frac{2(\sum_{j=1}^{J-1}\sum_{j'=j+1}^{J} w_{jj'} IEC^N_{jj'})}{J(J-1)}, \quad \text{where} \quad w_{jj'} = \frac{(\sigma^2_{Wj} + \sigma^2_{Wj'})/2}{\sum_j \sigma^2_{Wj}/J}. \tag{4}$$

If the within-subject variances are equal, then the IEC^N is the simple average of the pairwise $IEC^N_{jj'}$'s.

Relationship between IEC^N and IEC^R

If the Jth method is treated as a reference, IEC^R and CIA^R defined in section 3.1 are not the same as IEC^N and CIA^N in general when the Jth method is not treated as a reference. Note that

$$\begin{aligned} IEC^N_{jJ} &= \frac{E(Y_{ij} - Y_{iJ})^2 - (\sigma^2_{Wj} + \sigma^2_{WJ})}{(\sigma^2_{Wj} + \sigma^2_{WJ})/2} = \frac{E(Y_{ij} - Y_{iJ})^2 - 2\sigma^2_{WJ} - (\sigma^2_{Wj} - \sigma^2_{WJ})}{(\sigma^2_{Wj} + \sigma^2_{WJ})/2} \\ &= \frac{IEC^R_{jJ}\sigma^2_{WJ} - (\sigma^2_{Wj} - \sigma^2_{WJ})}{(\sigma^2_{Wj} + \sigma^2_{WJ})/2}, \end{aligned}$$

where IEC^R_{jJ} is the IEC comparing methods j and J with the Jth method as a reference. In practice where Jth method is a reference, we may expect that $\sigma^2_{WJ} \le \sigma^2_{Wj}$ which implies that $IEC^N_{jJ} \le IEC^R_{jJ}$. Equality occurs when $\sigma^2_{WJ} = \sigma^2_{Wj}$. Using equations (3) and (4), we find that

$$\begin{aligned} IEC^N &= \frac{2(\sum_{j=1}^{J-2}\sum_{j'=j+1}^{J-1} w_{jj'} IEC^N_{jj'} + \sum_{j=1}^{J-1} w_{jJ} IEC^N_{jJ})}{j(J-1)} \\ &= \frac{2(\sum_{j=1}^{J-2}\sum_{j'=j+1}^{J-1} w_{jj'} IEC^N_{jj'} + \sum_{j=1}^{J-1} w_{jJ} \frac{IEC^R_{jJ}\sigma^2_{WJ} - (\sigma^2_{Wj} - \sigma^2_{WJ})}{(\sigma^2_{Wj} + \sigma^2_{WJ})/2})}{J(J-1)} \end{aligned}$$

$$= \frac{2(\sum_{j=1}^{J-2}\sum_{j'=j+1}^{J-1} w_{jj'} IEC^N_{jj'} + \frac{\sigma^2_{WJ}}{\sum_{j=1}^{J}\sigma^2_{Wj}/J}\sum_{j=1}^{J-1} IEC^R_{jJ} - \frac{J\sum_{j=1}^{J-1}(\sigma^2_{Wj}-\sigma^2_{WJ})}{\sum_{j=1}^{J}\sigma^2_{WJ}})}{J(J-1)}$$

$$= IEC^N_{(J-1)}\frac{J-2}{J} + IEC^R\frac{2\sigma^2_{WJ}}{\sum_{j=1}^{J}\sigma^2_{Wj}} - \frac{2\sum_{j=1}^{J-1}(\sigma^2_{Wj}-\sigma^2_{WJ})}{(J-1)\sum_{j=1}^{J}\sigma^2_{WJ}},$$

where $IEC^N_{(J-1)}$ is the IEC comparing the first $J-1$ methods without a reference. In general, if the Jth method is a reference, we expect that $\sigma^2_{WJ} \leq \sigma^2_{Wj}, j = 1, \ldots, J-1$. This implies that $IEC^N_{jj'} \leq IEC^R_{jJ}$ and thus $IEC^N_{(J-1)} \leq IEC^R$. Therefore, we have that $IEC^N \leq IEC^R$ and $\psi^N \geq \psi^R$. Intuitively, this means that if the within-subject variances are larger than the within-subject variance from the reference, it should be harder to claim satisfactory individual agreement using ψ^R than using ψ^N. We have $\psi^N = \psi^R$ if $\sigma^2_{Wj} = \sigma^2_{WJ}$ for all j.

Estimation and Inference

Let Y_{ijk} be the measurements for the ith subject, by the jth method at the kth replication, $i = 1, \ldots, n, j = 1 \ldots, J, k = 1, \ldots, K$. The method of moment is again used for estimation of IEC^N and ψ^N. As shown in Haber et al. (2005), the unbiased estimates for τ^2_* and σ^2_* are as follows:

$$\hat{\tau}^2_* = \frac{\sum_{ij}(Y_{ij\bullet} - Y_{i\bullet\bullet})^2}{I(J-1)} - \frac{\sum_{ijk}(Y_{ijk} - Y_{ij\bullet})^2}{IJK(K-1)},$$

$$\hat{\sigma}^2_* = \frac{\sum_{ijk}(Y_{ijk} - Y_{ij\bullet})^2}{IJ(K-1)}.$$

Therefore, the estimate for IEC^N and ψ^N are

$$\hat{IEC}^N = \frac{2\hat{\tau}^2_*}{\hat{\sigma}^2_*}, \quad \text{or} \quad \hat{\psi^N} = \frac{\hat{\sigma}^2_*}{\hat{\tau}^2_* + \hat{\sigma}^2_*}.$$

One can also obtain estimate for σ^2_D as

$$\hat{\sigma}^2_D = 2\hat{\tau}^2_* - \frac{2\sum_{j=1}^{J-1}\sum_{j'=j+1}^{J}(Y_{\bullet j\bullet} - Y_{\bullet j'\bullet})}{J(J-1)}.$$

Similar to section 3.1, the above estimate for ψ^N can be re-written as ratio of the means of iid random variables:

$$\hat{\psi}^N = \frac{A_{\bullet\bullet}}{B^N_\bullet} = \frac{\sum_{i=1}^{n} A_{i\bullet}}{\sum_{i=1}^{n} B^N_i}$$

where $A_{i\bullet} = \sum_{J=1}^{J} A_{ij}/J$ with $A_{ij} = \sum_k (Y_{ijk} - Y_{ij\bullet})^2/(K-1)$ and

$$\begin{aligned} B_i^N &= \frac{\sum_{j=1}^{J}(Y_{ij\bullet} - Y_{i\bullet\bullet})^2}{J-1} + (1 - \frac{1}{K})\frac{\sum_{j=1}^{J} A_{ij}}{J} \\ &= \frac{\sum_{j=1}^{J-1}\sum_{j'=j+1}^{J}(Y_{ij\bullet} - Y_{ij'\bullet})^2}{J(J-1)} + (1 - \frac{1}{K})\frac{\sum_{j=1}^{J} A_{ij}}{J}. \end{aligned}$$

Thus, we can obtain nonparametric estimate for the s.e. of $\hat{\psi}^N$ by delta method for the variance of the ratio and then the 95% CI for ψ^N. The percentile bootstrap method can also be used to obtain the 95% CI for ψ^N.

For easy computation, one can also utilize a two way ANOVA model to compute the estimates of IEC^N and ψ^N. If we fit the following two way ANOVA model without main effect for method,

$$Y_{ijk} = \mu + \alpha_i + \gamma_{ij} + \epsilon_{ijk},$$

and let MS and MSE be the mean square terms corresponding to the interaction term γ_{ij} and the error term ϵ_{ijk}, then we have

$$\hat{\tau}_*^2 = \frac{MS - MSE}{K}, \quad \text{and} \quad \hat{\sigma}_*^2 = MSE.$$

Thus,

$$\hat{IEC}^N = \frac{2(MS - MSE)}{K * MSE}, \quad \text{and} \quad \hat{\psi}^N = \frac{K * MSE}{MS + (K-1) * MSE}.$$

4 Examples

Five examples are used to illustrate the proposed concepts and methodology in assessing individual agreement via individual equivalence. The first example compares two machines where one of them may or may not be treated as a reference. The second example compares two radiologists where neither of them is considered as a reference. Example three compares three methods in measuring carotid artery stenosis, where one of the methods is a standard

method. Example four compares two human observers to an automatic machine in measuring blood pressure, where both human observers are treated as references. The last example compares a new digital device to human observers in measuring blood pressure, where no replicated measurements were taken by the new method.

In all examples, we compute estimates and the 95% CIs based on nonparametric and percentile bootstrap (based on 10,000 bootstrap samples) approaches for IEC and CIA for cases of with and without a reference when applicable. For better interpretation and understanding the results, we also provide estimates for population means (μ), within-subject variances (σ_W^2), between-subject variances (σ_B^2), intra-class correlations (ICC) (based on one-way ANOVA model, Barnhart, et al., 2005) for each method, aggregated within-subject variability (σ_*^2), and subject-by-method interaction (σ_D^2). It is useful to compare the magnitudes of τ_*^2 with σ_*^2, and $2\tau_*^2$ with σ_D^2. In the tables, we drop the superscripts R and N, and we label which method is a reference when applicable. Because IEC and CIA are equivalent coefficients, we only display the numbers for CIA in the tables.

Example 1. Manual Goniometer vs Electrogoniometer

Eliasziw et al. (1994) presented data from a study that compared a large universal plastic manual goniometer and a Lamoreux-type electrogoniometer for measuring knee joint angle (in degrees). Twenty-nine individuals (n=29) were measured three consecutive times (K=3) on each goniometer. The estimates for population means, within and between-subject variability as well as intraclass correlations by goniometer, are displayed in the first part of Table 1. The electrogoniometer produced a slightly smaller mean angle than the manual goniometer, and had slightly larger within-subject variance than the manual goniometer. The between-subject variances (53.8 and 51.4) are much larger than the within-subject variances (0.736 and 0.977) and, this fact leads to high values of ICC.

When the manual goniometer is treated as a reference, the moment estimate (95% CI)

of ψ^R is 0.246 (0.132, 0.361). This implies that the electrogoniometer does not have good individual agreement with the manual goniometer. If the manual goniometer is not treated as a reference, the moment estimate (95% CI) of ψ^N is 0.287 (0.149, 0.425). Again, individual agreement between the two goniometers is not very good although Eliasziw et al. reported ICC value of 0.961 based on ANOVA model for assessing inter-method reliability. This high value is largely due to substantial between-subject variability.

Examples 2. Comparison of two radiologiests in calcium scoring

In this example, we are interested in knowing whether two radiologists (J=2) can be used interchangeably when they grade the coronary artery calcium score. Neither of the radiologists is considered as a reference. Two replicated readings (K=2) are obtained from these two radiologists for 12 patients (n=12) (see data in Haber et al., 2005). While there are some differences in mean score and within-subject variability between the two radiologists, the between-subject variabilities are huge which lead to intra ICCs (> 0.99) close to the boundary (Table 2). The point estimate for ψ^N is 0.754 that may indicate good agreement. However, the 95% CI of ψ^N) is (0.298, 1.0) impling that there is not enough information, due to small sample, to claim good individual agreement.

Example 3. Carotid Stenosis Data

This example compares three methods (J=3) in measuring carotid stenosis. The study was designed to compare two new methods, two-dimensional magnetic resonance angiography (MRA-2D) and three-dimensional MRA (MRA-3D), to the standard practice, invasive intra-arterial angiogram (IA) (Barnhart and Williamson, 2001). Clearly the standard method should be viewed as reference although our previous analysis treated IA as another method for illustration. Here we report our results both ways where IA is treated or is not treated as a reference. Three raters used each of the three methods to assess carotid stenosis on each of 55 patients (n=55). For illustration, we assume that readings by the three raters using the

same method are replicates (K=3), the same assumption as we did in the previous analysis (Barnhart, et al, 2005). The readings ranged from 0% to 100% blockage of the artery and the results are displayed separately for left and right arteries (Tables 3 and 4).

For both left and right carotid arteries, the MRA-2D and MRA-3D produced higher mean stenosis and higher within-subject variances than the IA method. Between-subject variances are comparable across the three methods. The intra ICC is higher for the IA method (0.884 for left and 0.916 for right) than for the MRA-2D and MAR-3D methods (0.626 and 0.647 for left, and 0.610 and 0.622 for right). The estimates of CIA as well as the 95% CIs are very different for the two cases where IA is treated or is not treated as reference. If the IA method is treated as a reference, one would conclude that MRA-2D and MRA-3D do not have good individual agreement with the IA method. If the IA method is not treated as a reference, the MRA-2D and MRA-3D have satisfactory individual agreement with the IA method. This difference in conclusion is mainly due to the the fact that there is substantially lower within-subject variability by the IA method than by MRA-2D and MRA-3D methods (e.g. 139.7 vs. 576.7 or 520.2 for left carotid artery). The pairwise comparisons show that MRA-2D and MRA-3D agree well where neither method is treated as a reference. Neither the MRA-2D nor the MRA-3D method has good individual agreement with the IA method where the IA method is a reference.

Example 4. Bland and Altman BP data

Bland and Altman (1999) presented data on systolic blood pressure from a study where two experienced human observers (denoted observers 1 and 2) and a semi-automatic blood pressure monitor (denoted machine) made three quick successive observations (K=3) on 85 individuals (n=85). They used a different subset of the data to illustrate different concepts of their methodology. By checking the original source of the data (see Bland and Altman, 1991), it appears that the semi-automatic blood pressure monitor was developed to replace

human observers, and the human observers should be considered as references. Therefore, we have a situation with two references (R=2 human observers) and one new method (J=1). For comparison purposes, we also report results where the human observers are not treated as references.

The simple statistics in table 5 show that the semi-automatic machine produced higher mean systolic blood pressure and higher within-subject variability than the two human observers. Because there is substantial between-subject variability, the intra ICC have high values for both human observer and the semi-automatic machine. The CIA and the corresponding 95% CIs are substantially less than 0.445, regardless of whether the human observers are treated as references or not. This implies that the semi-automatic blood pressure monitor does not have good individual agreement with human observers, and thus one would not want to replace human observers with the semi-automatic machine. The pairwise comparisons show that the two human observers have excellent individual agreement. In fact, the true difference between the two human observers is estimated to be smaller than the difference due to replication which lead to negative point estimate for τ_*^2 based on our formula. (Negative estimate can happen when different variance components are estimated separately, and we recommend setting $\hat{\tau}_*^2 = 0$ in this case).

Example 5. Digital Blood Pressure Device vs. Human Observer

In a study that investigated whether the digital blood pressure device can replace a human observer in a field study (Torun, et al, 1998), 228 subjects (n=228) were measured by a new digital device once, and then by three human observers. Example 4 shows that the two human observers have excellent individual agreement. This implies that the readings by experienced observers may be treated as replicates from the same experienced observer. For illustration, we extrapolate these results from example 4 to this example and treat the three readings from the three human observers as replicated readings. This allows us to demon-

strate that one can estimate ψ^R when there are replications by the reference method, but no replications by the new method. The point estimate (95% CI) of ψ^N are 0.462 (0.360, 0.578) and 0.729 (0.633, 0.826) for systolic and diastolic blood pressure, respectively. This implies that the digital device has borderline individual agreement with regards to systolic blood pressure, and good individual agreement with respect to diastolic blood pressure. We can interpret CIA^R similarly. For comparison, Barnhart and Williamson (2001) reported pooled CCC as 0.973 and 0.951 for systolic and diastolic blood pressure respectively. These numbers are reflected by a considerable between-subject variability and relatively small within-subject variability by the human observer.

5 Discussion

In this paper, we have proposed two equivalent coefficients, IEC and CIA, for assessing individual agreement between multiple methods for scenarios of existing reference or no reference. The concept of individual agreement provides a quantitative assessment when one wants to replace an existing method with a new method or using the several new methods (or observers) interchangeably. The illustration of five examples show that the concepts have wide applications in variaty of agreement studies.

A simulation study to investigate the bias and mean square error of the proposed estimates of CIA has been reported in a separate paper (Haber and Barnhart, 2007) for the case of two methods. We found that our approach performs consistently well for different combinations of true parameter values and sample sizes with 2 or 3 replications.

As mentioned in the introduction, the CCC increases as the between-subject variability increases evan though the individual difference between any two readings remains the same. We provide a in-depth comparison on the properties of the CCC and the CIA in Barnhart et

al. (2007) where the relationship between CCC and CIA as well as the impact of between-subject variability are presented algebraically and graphically. We also propose there a new CCC for multiple methods where one of them is treated as reference.

We proposed a nonparametric and bootstrap approaches for inference. A generalized estimating equations approach used in Barnhart and Williamson (2001) can be modified for both estimation and inference.

As illustrated in example 3, one should be cautious in interpreting results when the within-subject variances differ greatly between the methods. In this case, one may consider using one of the methods as a reference and look into results from pairwise comparisons.

We used the reference-scaled approach of IBC to define our IEC. If the within-subject variance due to replication is very small, the IEC value would appear to be very large when this within-subject variance is used in the demoninator for scaling. In this case, a constant scaled IEC may be preferred and we can extend our concept accordingly. For example, if the Jth method is a reference and $\sigma_J^2 \leq \sigma_{J0}^2$ where σ_{J0}^2 is the maximum tolerable within-subject variance, one can define constant-scaled IEC as

$$IEC^R = \frac{(\sum_{j=1}^{J-1} E(Y_{ij} - Y_{iJ})^2)/(J-1) - 2\sigma_{J0}^2}{\sigma_{J0}^2}.$$

and the corresponding CIA as

$$CIA^R = \frac{2\sigma_{J0}^2}{\sum_{j=1}^{J-1} E(Y_{ij} - Y_{iJ})^2)/(J-1)}.$$

The individual equivalence bound is based on upper limit of IEC, $\theta_I = 2.4948$ or lower limit of CIA, $CIA_I = 0.445$. It is possible that this criterion is too strict for claiming individual equivalence for some continuous scales. One may choose the boundary based on subject matter. For example, it may be reasonable to conclude individual equivalence if the inter-method variability is within 150% of within-subject variability for systolic blood pressure. This would imply that $\theta_I = 3$ and $CIA_I = 0.4$.

We used moment criteria to assess individual agreement via individual equivalence. In the bioequivalence literature, a probability criterion (Schall and Luus, 1993) was also proposed for establishing individual bioequivalence. This is closely related to the coverage probability and total deviation index approaches in the agreement literature (Lin, et al, 2002). However, the latter approaches only consider the probability of individual difference falling within a boundary, rather than magnitude of this probability relative to the probability of the difference between replications falling within the same boundary. If the boundary is chosen so that this latter probability based on replications is 1, then the coverage probability and total deviation index approaches may be used to assess individual agreement.

Acknowledgements

This research is supported by the National Institutes of Health Grant R01 MH70028

References

1. Atkinson, G. and Nevill, A. (1997). Comment on the use of concordance correlation to assess the agreement between two variables. *Biometrics* 53: 775-777.

2. Anderson, S. and Hauck, W.W. (1990). Consideration of individual bioequivalence. *Journal of Pharmacokinetics and Biophamaceutics* 18: 259-273.

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

4. Barnhart, H.X., Song, J. and Haber, M. (2005). Assessing Assessing intra, inter, and total agreement with Replicated Measurements. *Statistics in Medicine* 24: 1371-1384

5. Barnhart, H.X., Haber, M., Lokhnygina, Y. and Kosinski, A.S. (2007). Comparison of Concordance Correlation Coefficient and Coefficient of Individual Agreement in Assess-

ing Agreement. *Journal of Biopharmaceutical Statistics*, a special issue on agreement, accepted.

6. Bland, J.M. and Altman, D.G. (1991). The analysis of blood pressure data. In O'Brien E, O'Malley K eds. *Blood Pressure Measurement*. Elsevier: Amsterdam, pp 287-314.

7. Bland, J.M. and Altman, D.G. (1999). Measuring agreement in method comparison studies. *Statistics Methods in Medical Research* 8: 135-160.

8. Carrasco, J.L. and Jover, L. (2003). Estimating the generalized concordance correlation coefficient through variance components. *Biometrics* 59: 849-858.

9. Eliasziw, M., Young, S.L., Woodbury, M.G., and Fryday-Field, K. (1994). Statistical methodology for the concurrent assessment of interrater and intrarater reliability: Using goniometric measurements as an example. *Physical Therapy* 74: 777-788.

10. Food and Drug Administration (FDA) (2001). Guidance for industry: Statistical approaches to establishing bioequivalence, Food and Drug Administration, Center for Drug Evaluation and Research (CDER).

11. Haber, M., Barnhart, H.X., Song, J., and Gruden, J. (2005). Interobserver Variability: a New Approach in Evaluating Interobserver Agreement. *Journal of Data Sciences* 3: 69-83.

12. Haber, M. and Barnhart, H.X. (2007). A General Approach to Evaluating Agreement between Two Observers or Methods of Measurement from Quantitative Data with Replicated Measurements, *Statistical Methods in Medical Research*, in press.

13. Lin, L.I. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 45:225-268.

14. Lin, L.I. (2000). A note on the concordance correlation coefficient. *biometrics* 56: 324-325.

15. Lin, L.I., Hedayat, A.S., Sinha, B. and Yang, M. (2002). Statistical methods in assessing agreement: Models, issues and tools. *Journal of American Statistical Association* 97: 257-270.

16. McGraw, K.O. and Wong, S.P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods* 1:30-46.

17. Shall, R. and Luus, H.G. (1993). On population and individual bioequivalence. *Statistics in Medicine* 12: 1109-1124.

18. Shao, J. and Zhong, B. (2004). Assessing the agreement between two quantitative assays with repeated measurements. *Journal of Biopharmaceutical Statistics* 14: 201-212.

19. Sheiner, L. (1992). Bioequivalence revisited. *Statistics in Medicine* 12: 1109-1124.

20. Tourn, B., Grajeda, R., Mendez, H., Flores, R., Martorell, R., and Schroeder, D. (1998). Evaluation of inexpensive digital sphygmomanometers for field studies of blood pressure. *Federation of American Societies of Experimental Biology Journal* 12: 5072.

Figure 1. Dependency of ICC and CCC on between-subject variability

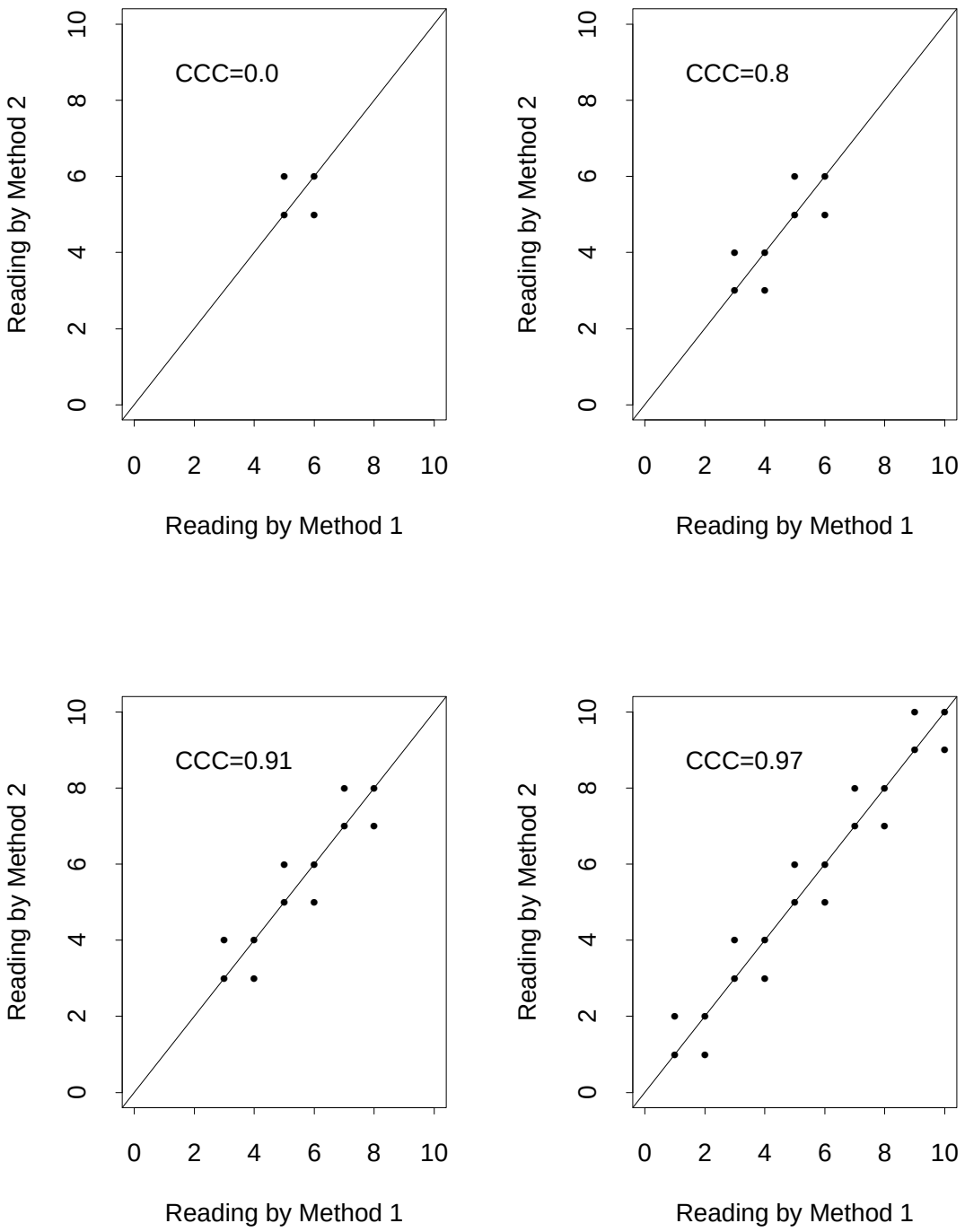

Table 1. Comparison of Manual Goniometer and Electrogoniometer

	Manual Goniometer	Electro-goniometer	
	Estimate	Estimate	Difference
μ	1.437	0.046	1.391
σ_W^2	0.736	0.977	-0.241
σ_B^2	53.8	51.4	2.4
Intra ICC	0.986	0.982	0.004

	Individual Agreement			
	Reference: Manual Goniometer		No reference	
	estimate (95% CI)	95% bCI*	Estimate (95% CI)	95% bCI*
CIA	0.246 (0.132, 0.361)	(0.155, 0.387)	0.287 (0.149, 0.425)	(0.176, 0.442)
τ_*^2	2.130	–	2.130	–
σ_D^2	2.326	–	2.326	–
σ_*^2	0.856	–	0.856	–

* bCI is the confidence interval based on the bootstrap percentile method.

Table 2. Comparison of Two Radiologists on Calcium Scoring

	Radiologist A	Radiologist B	
	Estimate	Estimate	Difference
μ	35.833	36.125	-0.292
σ_W^2	7.667	0.125	7.542
σ_B^2	1025.7	1116.2	-90.5
Intra ICC	0.993	0.999	
	Individual Agreement without Reference		
	Estimate (95% CI)	95% bCI*	
CIA	0.754 (0.298, 1.0)	(0.219, 1.0)	
τ_*^2	1.271	–	
σ_D^2	2.457	–	
σ_*^2	3.896	–	

* bCI is the confidence interval based on the bootstrap percentile method.

Table 3. Comparison of MRA-2D and MRA-3D with IA for Left Carotid Artery

	MRA-2D estimate	MRA-3D estimate	IA estimate
μ	43.7	48.2	38.0
σ_W^2	576.7	520.2	139.7
σ_B^2	966.5	953.7	1061.2
Intra ICC	0.626	0.647	0.884

Individual Agreement

	Reference: IA		No reference	
	estimate (95% CI)	95% bCI*	Estimate (95% CI)	95% bCI*
CIA	0.209 (0.032,0.386)	(0.085,0.420)	0.632 (0.441,0.823)	(0.454,0.828)
τ_*^2	323.8	–	240.5	–
σ_D^2	579.3	–	428.7	–
σ_*^2	344.1	–	412.2	–
	Pairwise: MRA-2D vs. MRA-3D			
CIA	–	–	0.881 (0.688,1.0)	(0.684,1.0)
	Pairwise: MRA-2D vs. IA			
CIA	0.231 (0.035,0.427)	(0.092,0.468)	0.592 (0.348,0.835)	(0.382,0.867)
	Pairwise: MRA-3D vs. IA			
CIA	0.191 (0.026, 0.357)	(0.076, 0.400)	0.452 (0.242,0.661)	(0.273,0.690)

* bCI is the confidence interval based on the bootstrap percentile method.

Table 4. Comparison of MRA-2D and MRA-3D with IA for Right Carotid Artery

	MRA-2D estimate	MRA-3D estimate	IA estimate		
μ	45.9	43.9	33.8		
σ_W^2	568.5	550.0	88.0		
σ_B^2	887.7	903.6	965.2		
Intra ICC	0.610	0.622	0.916		
	Individual Agreement				
	Reference: IA			No reference	
	estimate (95% CI)	95% bCI*		Estimate (95% CI)	95% bCI*
CIA	0.172 (0.078,0.265)	(0.094,0.244)		0.738 (0.587,0.889)	(0.589,0.881)
τ_*^2	189.4	–		143.0	–
σ_D^2	255.1	–		202.3	–
σ_*^2	323.6	–		402.2	–
	Pairwise: MRA-2D vs. MRA-3D				
CIA	–	–		0.917 (0.729,1.0)	(0.734, 1.0))
	Pairwise: MRA-2D vs. IA				
CIA	0.183 (0.083,0.284)	(0.104,0.307)		0.684 (0.562,0.807)	(0.572,0.814)
	Pairwise: MRA-3D vs. IA				
CIA	0.161 (0.060,0.262)	(0.084,0.295)		0.584 (0.370,0.798)	(0.392,0.815)

* bCI is the confidence interval based on the bootstrap percentile method.

Table 5. Comparison of Observers and Automatic Machine in Measuring Blood Pressure.

	Observer 1	Observer 2	Machine		
	estimate	estimate	estimate		
μ	127.4	127.3	143.0		
σ_W^2	37.4	38.0	83.1		
σ_B^2	936.0	917.1	983.2		
Intra ICC	0.962	0.960	0.922		
	Individual Agreement				
	Reference: Observers			No reference	
	estimate (95% CI)	95% bCI*		Estimate (95% CI)	95% bCI*
	Overall results				
CIA	0.111 (0.046,0.177)	(0.064,0.205)		0.225 (0.112,0.339)	(0.139,0.384)
τ_*^2	278.1	–		181.5	–
σ_D^2	311.4	–		199.8	–
σ_*^2	60.4	–		52.8	–
	Pairwise: Observer 1 vs. Observer 2				
CIA	–	–		1.0	–
	Pairwise: Machine vs. observer 1				
CIA	0.110 (0.0460,0.175)	(0.064,0.210)		0.178 (0.086,0.270)	(0.107,0.302)
	Pairwise: Machine vs. observer 2				
CIA	0.112 (0.046,0.178)	(0.065,0.213)		0.179 (0.084,0.274)	(0.107,0.310)

* bCI is the confidence interval based on the bootstrap percentile method.

Table 6. Comparison of Observers and Digital Device in Measuring Blood Pressure.

	Systolic		Diastolic	
	Observer	Digital Device	Observer	Digital Device
	estimate	estimate	estimate	estimate
μ	129.3	133.4	79.0	77.8
σ_W^2	11.4	–	8.4	–
σ_B^2	938.7	–	236.3	–
Intra ICC	0.988	–	0.966	–
	Individual Agreement with Observer as reference			
	Systolic		Diastolic	
	estimate (95% CI)	95% bCI*	Estimate (95% CI)	95% bCI*
CIA	0.462 (0.346,0.578)	(0.356,0.589)	0.729 (0.633,0.826)	(0.588,0.890)

* bCI is the confidence interval based on the bootstrap percentile method.