STATISTICS IN MEDICINE
Statist. Med. 2005; **24**:1371–1384
Published online 30 November 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2006

Assessing intra, inter and total agreement with replicated readings

Huiman X. Barnhart[1,*,†], Jingli Song[2,‡] and Michael J. Haber[3,§]

[1]*Department of Biostatistics and Bioinformatics, Duke Clinical Research Institute, Duke University, P.O. Box 17969, Durham, NC 27715, U.S.A.*
[2]*Eli Lilly and Company, Lilly Corporate Center, DC 6134, Indianapolis, IN 46285, U.S.A.*
[3]*Department of Biostatistics, The Rollins School of Public Health, Emory University, Atlanta, GA 30322, U.S.A.*

SUMMARY

In clinical studies, assessing agreement of multiple readings on the same subject plays an important role in the evaluation of continuous measurement scale. The multiple readings within a subject may be replicated readings by using the same method or/and readings by using several methods (e.g. different technologies or several raters). The traditional agreement data for a given subject often consist of either replicated readings from only one method or multiple readings from several methods where only one reading is taken from each of these methods. In the first case, only intra-method agreement can be evaluated. In the second case, traditional agreement indices such as intra-class correlation (ICC) or concordance correlation coefficient (CCC) is often reported as inter-method agreement. We argue that these indices are in fact measures of total agreement that contains both inter and intra agreement. Only if there are replicated readings from several methods for a given subject, then one can assess intra, inter and total agreement simultaneously. In this paper, we present new inter-method agreement index, inter-CCC, and total agreement index, total-CCC, for agreement data with replicated readings from several methods where the ICCs within methods are used to assess intra-method agreement for each of the several methods. The relationship of the total-CCC with the inter-CCC and the ICCs is investigated. We propose a generalized estimating equations approach for estimation and inference. Simulation studies are conducted to assess the performance of the proposed approach and data from a carotid stenosis screening study is used for illustration. Copyright © 2004 John Wiley & Sons, Ltd.

KEY WORDS: agreement; reliability; concordance correlation coefficient; intraclass correlation; generalized estimating equations

*Correspondence to: Huiman X. Barnhart, Department of Biostatistics and Bioinformatics, Duke Clinical Research Institute, Duke University, P.O. Box 17969, Durham, NC 27715, U.S.A.
†E-mail: huiman.barnhart@duke.edu
‡E-mail: songji@lilly.com
§E-mail: mhaber@sph.emory.edu

Contract/grant sponsor: Emory University Quadrangle Fund
Contract/grant sponsor: NIH; contract/grant number: R0I MH070028-01A1

Received March 2004
Accepted July 2004

1. INTRODUCTION

In clinical studies, continuous scales are often taken by several methods (e.g. different technologies or different raters). Before using the measurement scale in practice, one often needs to assess agreement of multiple readings taken by several methods. If there is an observed disagreement, one often wants to know whether the disagreement is due to random error within a method or due to true differences attributed by the different methods. If the disagreement is due to the random error within a particular method, this method may not be used in practice. If disagreement is due to the true difference among the methods, the methods will need to be modified for improvement. Therefore, assessing agreement often leads to assessing both intra-method agreement and inter-method agreement, where intra-method agreement measures consistency of readings taken by the same method and the inter-method agreement measures consistency of true readings attributed by the methods. A true reading here may be interpreted as the mean value of infinite replicated readings produced by the method on the same subject. Note that we define the inter-method agreement based on the true readings, not on the observed readings, because we believe that the inter-method agreement should not be obscured by the random error within method. An observed reading is the sum of true reading by the method and a random error within the method. Traditionally, agreement data for a given subject often consist of replicated readings from only one method or multiple readings from several methods where only one reading is taken from each of the several methods. In the first case, only intra-method agreement can be evaluated. In the second case, one cannot evaluate intra-method agreement nor inter-method agreement because one cannot estimate the random variation within a method and the method's true reading when there are no replications within a method. Because observed multiple readings contain variation due to both true differences attributed by the methods and random error within the methods, any agreement measure based on the observed readings are in fact total agreement that contains both intra- and inter-method agreement. Therefore, the popular agreement indices based on the observed readings such as concordance correlation coefficient (CCC) [1] or different versions of intra-class correlation coefficient (ICC) [2] are in fact the measure of total agreement although they are often reported as inter-method agreement. Most of the ICC indices make assumptions that the means and/or variances of readings by different methods are equal (the exchangeable assumption). This assumption may not be realistic because one method may produce consistently higher value than other method and one method may have larger variability than other method. We will focus on CCC-type indices except in assessing intra agreement.

To assess intra, inter and total agreement simultaneously, we advocate an agreement study design where each of the different methods produce replicated readings on each subject. With replicated readings, we can use intra-ICC to evaluate intra-method agreement for each method. A new index, inter-CCC, is proposed in Section 2 to assess the inter-method agreement based on the true readings. The total agreement is evaluated with another proposed index, total-CCC, based on observed readings (see Section 2). We also investigate the relationship of the total-CCC with intra-ICCs and inter-CCC. In Section 3, we propose a generalized estimating equations (GEE) approach for estimation and inferences about these indices. We conduct simulation studies to assess the performance of a GEE approach in Section 4. In Section 5, we use data from a carotid stenosis screening study to illustrate the use of these indices. In Section 6, we extend the proposed

method to the case when covariates are involved. A brief discussion is presented in Section 7.

2. METHOD

Suppose that there are N randomly selected subjects and J fixed methods, where jth method produces $K_{ij}(K_{ij} \geqslant 2)$ replicated measurements for the ith subject. Here, we allow the number of replications to be different by methods and subjects. Thus, without making any distributional assumption, it is reasonable to assume that conditional on subject i and method j the replicated measurements Y_{ijk}, $k = 1, \ldots, K_{ij}$, are independently and identically distributed (iid). Let $E_k(Y_{ijk}|ij) = \mu_{ij}$ and $\mathrm{Var}_k(Y_{ijk}|ij) = \sigma_{ij}^2$ be the conditional mean and variance of Y_{ijk}. We use $E_k(X)$ and $\mathrm{Var}_k(X)$, respectively, to denote the expectation and variance with respect to the random variable X associated with index k. Similarly, we use index ij in the conditioning argument to denote conditioning on subject i and method j. Here μ_{ij} is a random variable because subjects are considered random. The parameter μ_{ij} can be viewed as the true reading of the jth method on the ith subject. This reading can be interpreted as the mean value of the jth method produced on the ith subject if the jth method could produce infinitely many replicated readings on the ith subject. Let $Y_{ijk} = \mu_{ij} + e_{ijk}$. We use the following notation and assumptions: $E_i(\mu_{ij}|j) = \mu_{*j}$, $\mathrm{Var}_i(\mu_{ij}|j) = \delta_j^2$, $\mathrm{Corr}_i(\mu_{ij}, \mu_{ij'}|jj') = \rho_{jj'}^{\mu}$, $E_k(e_{ijk}|ij) = 0$, $\mathrm{Var}_k(e_{ijk}|ij) = \sigma_{ij}^2$, $E_i(\sigma_{ij}^2|j) = \sigma_{*j}^2$, e_{ijk} are mutually independent, μ_{ij} and e_{ijk} are mutually independent. Here, we use $*$ in place of index i to denote the expectation with respect to the random variable associated with index i.

2.1. Using ICCs to assess intra-method agreement

For a given method, say method j, we have $E_i(y_{ijk}|j) = \mu_{*j}$ and $\mathrm{Var}_i(y_{ijk}|j) = \delta_j^2 + \sigma_{*j}^2$ for any k. Thus, the interchangeability assumption is met for intraclass correlation coefficient under the one-way ANOVA model [3] for the jth method. The jth ICC can be written as

$$\rho_j^I = \frac{\mathrm{Cov}(y_{ijk}, y_{ijk'})}{\sqrt{\mathrm{Var}(y_{ijk})\mathrm{Var}(y_{ijk'})}} = \frac{\delta_j^2}{\delta_j^2 + \sigma_{*j}^2} \tag{1}$$

The ICCs, ρ_j^I's, can be used to assess intra-method agreement.

2.2. Using inter-CCC to assess inter-method agreement

Lin [1] proposed a concordance correlation coefficient (CCC) to evaluate the agreement between two fixed methods. Barnhart *et al.* [4] extended the CCC to overall CCC (OCCC) for evaluating agreement among multiple methods each taking one reading for each subject. In this paper, we use the abbreviation CCC for concordance correlation coefficient, regardless of whether there are two or more than two methods. As stated in the introduction, these indices are defined at the level of observed readings and thus are, in fact, measures of total (inter+intra) method agreement. Because we are interested in an inter-method agreement index based on the difference of the methods' true readings, we use the true inter-method variability, $\tau_i^2 = \sum_{j=1}^{J}(\mu_{ij} - \mu_{i\bullet})^2/(J-1)$, where $\mu_{i\bullet} = \sum_{j=1}^{J}\mu_{ij}/J$, to measure the true differences among methods. The inter-CCC is defined similar to the CCC and OCCC with Y_{ij}'s

replaced by μ_{ij}'s

$$\begin{aligned}\rho_c(\mu) &= 1 - \frac{E(\tau_i^2)}{E(\tau_i^2|\boldsymbol{\mu}_j's \text{ are independent})} \\ &= 1 - \frac{E(\sum_{j=1}^{J}(\mu_{ij} - \mu_{i\bullet})^2/(J-1))}{E(\sum_{j=1}^{J}(\mu_{ij} - \mu_{i\bullet})^2/(J-1)|\boldsymbol{\mu}_j's \text{ are independent})} \\ &= 1 - \frac{E(\sum_{j=1}^{J-1}\sum_{j'>j}(\mu_{ij} - \mu_{ij'})^2/J(J-1))}{E(\sum_{j=1}^{J-1}\sum_{j'>j}(\mu_{ij} - \mu_{ij'})^2/J(J-1)|\boldsymbol{\mu}_j's \text{ are independent})}\end{aligned}$$

where $\boldsymbol{\mu}_j$ is the random vector with the ith component as μ_{ij}. The proposed inter-CCC has the following properties similar to OCCC:

- $-1 \leqslant \rho_c(\mu) \leqslant 1$, $\rho_c(\mu) = 1$ indicates a perfect inter-method agreement without concerning any internal error of a method.
- $\rho_c(\mu)$ can be expressed as a function of μ_{*j}'s, δ_j's and $\rho^{\mu}_{jj'}$.

$$\rho_c(\mu) = \frac{2\sum_{j=1}^{J-1}\sum_{j'>j}\rho^{\mu}_{jj'}\delta_j\delta_{j'}}{(J-1)\sum_{j=1}^{J}\delta_j^2 + \sum_{j=1}^{J-1}\sum_{j'>j}(\mu_{*j} - \mu_{*j'})^2} \tag{2}$$

- $\rho_c(\mu)$ is a weighted average of pairwise inter-CCCs.

$$\rho_c(\mu) = \frac{\sum_{j=1}^{J-1}\sum_{j'>J}\xi_{jj'}\rho_{c(jj')}(\mu)}{\sum_{j=1}^{J-1}\sum_{j'>J}\xi_{jj'}} \tag{3}$$

  where $\xi_{jj'} = \delta_j^2 + \delta_{j'}^2 + (\mu_{*j} - \mu_{*j'})^2$ is the weight for the pair (j, j'), $\rho_{c(jj')}(\mu) = 2\rho^{\mu}_{jj'}\delta_j\delta_{j'}/[\delta_j^2 + \delta_{j'}^2 + (\mu_{*j} - \mu_{*j'})^2]$ is the inter-CCC between method j and method j'.
- $\rho_c(\mu) = 1$ if and only if $\rho^{\mu}_{jj'} = \rho^{\mu}$, $\delta_j^2 = \delta_*^2$ and $\mu_{*j} = \mu_{*\bullet}$, for $j = 1, \ldots, J$.
- If we assume $\rho^{\mu}_{jj'} = \rho^{\mu}$ for all j and j', i.e. the same precision for all the pairs of methods, then the inter-CCC can be expressed as a product of precision and overall accuracy, $\rho_c(\mu) = \rho^{\mu}\chi^a_{\mu}$, where ρ^{μ} is the precision and $\chi^a_{\mu} = (2\sum_{j=1}^{J-1}\sum_{j'>j}\delta_j\delta_{j'})/[(J-1)\sum_{j=1}^{J}\delta_j^2 + \sum_{j=1}^{J-1}\sum_{j'>j}(\mu_{*j} - \mu_{*j'})^2]$ is the overall accuracy.
- Equation (2) does not contain the term σ^2_{*j}, indicating that it is not affected by the random error within method.

2.3. *Using total-CCC to assess total agreement*

As mentioned in the previous sub-section, the CCC and the OCCC measure the total agreement between two methods and multiple methods, respectively, when $K_{ij} = 1$. In this sub-section, we extend the index to total-CCC for the general case of $K_{ij} \geqslant 2$. For the purpose of definition that is consistent with CCC and OCCC, the total-CCC is based on a set of J observed readings where no two readings are from the same method. This is also due to the intent that only a single reading from one of the methods (instead of the average of readings from several methods) will be used in practice if the agreement among methods is shown to be high. Because the replicated readings made by the same method are iid, we can use J observed

readings from a random sample of size J where the jth reading is randomly sampled from the K_{ij} readings made by method j. Specifically, let Y_{ijo} denote a randomly selected reading from the K_{ij} readings, $Y_{ij1},\ldots,Y_{ijK_{ij}}$, made by method j on subject i. Then $(Y_{ijo},\ j=1,\ldots,J)$ forms a random sample of J readings for subject i. Let $Y_{i\bullet o}$ be the arithmetic mean of all Y_{ijo}s, $j=1,\ldots,J$. We define total-CCC as

$$\rho_c(Y)=1-\frac{E_i[\sum_{j=1}^{J}(Y_{ijo}-Y_{i\bullet o})^2/(J-1)]}{E_i[\sum_{j=1}^{J}(Y_{ijo}-Y_{i\bullet o})^2/(J-1)|Y_{i1o},\ldots,Y_{iJo}\text{ are uncorrelated}]}$$

Note $E_i(Y_{ijo}|j)=E_i(E_k(Y_{ijk}|jk))=\mu_{*j}$, $\mathrm{Var}_i(Y_{ijo}|j)=\mathrm{Var}_i(E_k(Y_{ijk}|ij))+E_i(\mathrm{Var}_k(Y_{ijk}|ij))=\delta_j^2+\sigma_{*j}^2$ and $\mathrm{Cov}_i(Y_{ijo},Y_{ij'o})=\mathrm{Cov}_i(\mu_{ij}+e_{ijk},\mu_{ij'}+e_{ij'k})=\rho_{jj'}^{\mu}\delta_j\delta_{j'}$. We can rewrite the total-CCC as a function of the μ_{*j}s, σ_{*j}^2s, δ_j^2s and $\rho_{jj'}^{\mu}$s

$$\rho_c(Y)=\frac{2\sum_{j=1}^{J-1}\sum_{j<j'}\rho_{jj'}^{\mu}\delta_j\delta_{j'}}{(J-1)\sum_{j=1}^{J}\delta_j^2+\sum_{j=1}^{J-1}\sum_{j<j'}(\mu_{*j}-\mu_{*j'})^2+(J-1)\sum_{j=1}^{J}\sigma_{*j}^2} \tag{4}$$

The total-CCC reduces to the CCC if $K_{ij}=1$ and $J=2$. This index is also the OCCC if $K_{ij}=1$ and $J>2$. Note that the total-CCC depends only on the distribution of Y_{ijk}. It does not depend on the random sampling of the readings nor on the number of replications (K_{ij}).

The difference between the total-CCC in (4) and the inter-CCC in (2) is that equation (4) contains the term σ_{*j}^2. This is because $\rho_c(Y)$ is defined at the level of observed readings. The total-CCC contains both between method and within method variability. Therefore, the total-CCC is a function of the inter-CCC and the ICCs. We investigate the relationship among the three indices in the next sub-section.

2.4. *Relationship among total-CCC, inter-CCC and ICCs*

Rewriting formula (4), the total-CCC is related to the inter-CCC and the ICCs as

$$\begin{aligned}\frac{1}{\rho_c(Y)}&=\frac{(J-1)\sum_{j=1}^{J}\delta_j^2+\sum_{j=1}^{J-1}\sum_{j<j'}(\mu_{*j}-\mu_{*j'})^2]}{2\sum_{j=1}^{J-1}\sum_{j<j'}\rho_{jj'}^{\mu}\delta_j\delta_{j'}}+\frac{(J-1)\sum_{j=1}^{J}\sigma_{*j}^2}{2\sum_{j=1}^{J-1}\sum_{j<j'}\rho_{jj'}^{\mu}\delta_j\delta_{j'}}\\&=\frac{1}{\rho_c(\mu)}+\frac{1}{J}\sum_{j=1}^{J}\omega_j(1-\rho_j^I)\end{aligned}$$

where $\omega_j=(\sigma_{*j}^2+\delta_j^2)/[\sum_{j=1}^{J-1}\sum_{j<j'}\rho_{jj'}^{\mu}\delta_j\delta_{j'}/(J(J-1)/2)]$. This implies

- $\rho_c(\mu)$ is always greater than or equal to $\rho_c(Y)$, assuming that the correlation between a pair of methods' readings is positive. In other words, there is more agreement between methods' true readings than between methods' observed readings.
- The weight ω_j is the ratio of the variance of the jth method to the average of all the between-method's variances.
- The bigger the $\rho_c(\mu)$, the bigger the $\rho_c(Y)$; the bigger the ρ_j^Is, the bigger the $\rho_c(Y)$. In other words, high total agreement among methods implies both high inter-method and the high intra-method agreement.

3. ESTIMATION AND INFERENCE

Let $\boldsymbol{\theta}=(\boldsymbol{\mu}_*,\boldsymbol{\sigma}_*^2,\boldsymbol{\delta}^2,\boldsymbol{\rho}^\mu)'$ be the vector of parameters with $\boldsymbol{\mu}_*=(\mu_{*1},\ldots,\mu_{*J})'$, $\boldsymbol{\delta}^2=(\delta_1^2,\ldots,\delta_J^2)'$, $\boldsymbol{\sigma}_*^2=(\sigma_{*1}^2,\ldots,\sigma_{*J}^2)'$, and $\boldsymbol{\rho}^\mu=(\rho_{12}^\mu,\ldots,\rho_{(J-1)J}^\mu)$. Equations (1), (2) and (4) show that ρ_j^I, $\rho_c(\mu)$, and $\rho_c(Y)$ are functions of $\boldsymbol{\theta}$. Therefore, estimation of these indices can be obtained by plugging in the estimate of $\boldsymbol{\theta}$ into these equations:

$$\hat{\rho}_j^I=\frac{\hat{\delta}_j^2}{\hat{\delta}_j^2+\hat{\sigma}_{*j}^2} \tag{5}$$

$$\hat{\rho}_c(\mu)=\frac{2\sum_{j=1}^{J-1}\sum_{j'>j}\hat{\rho}_{jj'}^\mu\hat{\delta}_j\hat{\delta}_{j'}}{(J-1)\sum_{j=1}^{J}\hat{\delta}_j^2+\sum_{j=1}^{J-1}\sum_{j'>j}(\hat{\mu}_{*j}-\hat{\mu}_{*j'})^2} \tag{6}$$

$$\hat{\rho}_c(Y)=\frac{2\sum_{j=1}^{J-1}\sum_{j'>j}\hat{\rho}_{jj'}^\mu\hat{\delta}_j\hat{\delta}_{j'}}{(J-1)\sum_{j=1}^{J}\hat{\delta}_j^2+(J-1)\sum_{j=1}^{J}\hat{\sigma}_{*j}^2+\sum_{j=1}^{J-1}\sum_{j'>j}(\hat{\mu}_{*j}-\hat{\mu}_{*j'})^2} \tag{7}$$

If we have $\hat{\boldsymbol{\theta}}$ and the estimated covariance matrix of $\hat{\boldsymbol{\theta}}$, then we will use delta method to perform inference on $\hat{\rho}_j^I$, $\hat{\rho}_c(\mu)$, and $\hat{\rho}_c(Y)$. Below we propose a GEE approach [5–7] for obtaining the estimates of $\boldsymbol{\theta}$ and covariance matrix of $\hat{\boldsymbol{\theta}}$.

Following the notations and assumptions as Section 2.1, we can estimate the parameters $\boldsymbol{\mu}_*=(\mu_{*1},\ldots,\mu_{*J})'$, $\boldsymbol{\delta}^2=(\delta_1^2,\ldots,\delta_J^2)'$, $\boldsymbol{\sigma}_*^2=(\sigma_{*1}^2,\ldots,\sigma_{*J}^2)'$ and $\boldsymbol{\rho}^\mu=(\rho_{12}^\mu,\ldots,\rho_{(J-1)J}^\mu)$ through a series of estimating equations.

- In the first set of estimating equations, we estimate $\boldsymbol{\mu}_*$ by modelling the marginal mean of $\mathbf{Y}_{i\bullet}=(Y_{i1\bullet},\ldots,Y_{iJ\bullet})'$, where $Y_{ij\bullet}=\sum_{k=1}^{K_{ij}}Y_{ijk}/K_{ij}$, $j=1,\ldots,J$, $i=1,\ldots,N$. We use $\bullet$ in the place of index j to represent the average of all the J methods.

$$\sum_{i=1}^{N}\mathbf{D}_{i1}'\mathbf{V}_{i1}^{-1}(\mathbf{Y}_{i\bullet}-\boldsymbol{\mu}_*)=\mathbf{0}$$

  where $\mathbf{D}_{i1}=\partial\boldsymbol{\mu}_*/\partial\boldsymbol{\mu}_*=I_J$ and $\mathbf{V}_{i1}$ is the working covariance matrix for $\mathbf{Y}_i$. Here I_J is a $J\times J$ identity matrix.
- In the second set of estimating equations, we estimate $\boldsymbol{\sigma}_*^2$ by

$$\sum_{i=1}^{N}\mathbf{D}_{i2}'\mathbf{V}_{i2}^{-1}(\mathbf{U}_i-\boldsymbol{\sigma}_*^2)=\mathbf{0}$$

  where $\mathbf{U}_i=(\sum_{k=1}^{K_{ij}}(Y_{i1k}-Y_{i1\bullet})^2/(K_{i1}-1),\ldots,(Y_{iJk}-Y_{iJ\bullet})^2/(K_{iJ}-1))'$, $\mathbf{D}_{i2}=\partial\boldsymbol{\sigma}_*^2/\partial\boldsymbol{\sigma}_*^2=I_J$, and $\mathbf{V}_{i2}$ is the working covariance matrix for $\mathbf{U}_i$.
- The third set of estimating equations is to estimate $\boldsymbol{\delta}^2$ by modelling the marginal mean of $\mathbf{W}_i$.

$$\sum_{i=1}^{N}\mathbf{D}_{i3}'\mathbf{V}_{i3}^{-1}(\mathbf{W}_i-\mathbf{g}(\boldsymbol{\mu}_*,\boldsymbol{\delta}^2))=\mathbf{0}$$

where $W_{ij} = (Y_{ij\bullet}^2 - U_{ij}/K_{ij})$ and $\mathbf{W}_i = (W_{i1}, \ldots, W_{iJ})'$, $\mathbf{g}(\boldsymbol{\mu}_*, \boldsymbol{\delta}^2) = E(\mathbf{W}_i) = \boldsymbol{\mu}_*^2 + \boldsymbol{\delta}^2$, $\mathbf{D}_{i3} = \partial \mathbf{g}(\boldsymbol{\delta}^2, \boldsymbol{\sigma}_*^2, \boldsymbol{\mu}_*)/\partial \boldsymbol{\delta}^2 = I_J$, and $\mathbf{V}_{i3}$ is the working covariance matrix for $\mathbf{W}i$.

- At last, we estimate $\boldsymbol{\rho}^{\mu}$ by the cross products $Y_{ij\bullet}Y_{ij'\bullet}(j' > j)$. Let $\mathbf{Z}_i = (Y_{i1\bullet}Y_{i2\bullet}, Y_{i1\bullet}Y_{i3\bullet}, \ldots, Y_{i(J-1)\bullet}Y_{iJ\bullet})'$ and note that $E(Y_{ij\bullet}Y_{ij'\bullet}) = \mu_{*j}\mu_{*j'} + \rho_{jj'}^{\mu}\delta_j\delta_{j'}$. We use Fisher's Z-transformation to model the correlation of pairwise true readings of methods as

$$\frac{1}{2}\log\frac{1+\rho_{jj'}^{\mu}}{1-\rho_{jj'}^{\mu}} = \mathbf{Q}_{ijj'}\boldsymbol{\alpha}$$

where $\mathbf{Q}_{ijj'}$ is an indicator variable for pairs (j, j') when no other covariates are involved. We will discuss the case with covariates in Section 6. There are two reasons to model the $\rho_{jj'}^{\mu}$ through Fisher's Z-transformation. First, $\boldsymbol{\alpha}$ ranges from $-\infty$ to ∞ so that it can effectively control the boundary problem. Second, according to Lin [1], this transformation provides better normality and stability. We estimate $\boldsymbol{\alpha}$ by the following estimating equations:

$$\sum_{i=1}^{N} \mathbf{D}_{i4}'\mathbf{V}_{i4}^{-1}(\mathbf{Z}_i - \mathbf{h}(\boldsymbol{\alpha}, \boldsymbol{\delta}^2, \boldsymbol{\sigma}_*^2, \boldsymbol{\mu}_*)) = \mathbf{0}$$

where $\mathbf{D}_{i4} = \partial \mathbf{h}(\alpha, \boldsymbol{\delta}^2, \boldsymbol{\sigma}_*^2, \boldsymbol{\mu}_*)/\partial \boldsymbol{\alpha}$, $\mathbf{h}(\boldsymbol{\alpha}, \boldsymbol{\delta}^2, \boldsymbol{\sigma}_*^2, \boldsymbol{\mu}_*) = E(\mathbf{Z}_i)$ and $\mathbf{V}_{i4}$ is the working covariance matrix for $\mathbf{Z}_i$.

The estimating process begins by obtaining $\hat{\boldsymbol{\mu}}_*$ and $\hat{\boldsymbol{\sigma}}_*^2$ in the first two estimating equations via a modified Fisher-scoring algorithm. We then plug $\hat{\boldsymbol{\mu}}_*$ and $\hat{\boldsymbol{\sigma}}_*^2$ in for $\boldsymbol{\mu}_*$ and $\boldsymbol{\sigma}_*^2$ in the third estimating equation and solve for $\boldsymbol{\delta}^2$. Next, we plug in $\hat{\boldsymbol{\mu}}_*$, $\hat{\boldsymbol{\sigma}}_*^2$ and $\hat{\boldsymbol{\delta}}^2$ into the fourth estimating equation and solve for $\boldsymbol{\alpha}$. Estimate for $\boldsymbol{\rho}^{\mu}$ is then obtained by using the inverse of the Fisher's Z-transformation. If there are no covariates involved in the study and the independent working covariance matrices are used in each set of estimating equations, the estimates of $\boldsymbol{\mu}_*$, $\boldsymbol{\delta}^2$, and $\boldsymbol{\sigma}_*^2$ can be expressed in closed forms and they turn out to be the moment estimates. The estimate of $\boldsymbol{\alpha}$ can be solved by using a Fisher-scoring numerical method. In our simulation studies (Section 4), we set these working covariance matrices, $\mathbf{V}_{i1}$, $\mathbf{V}_{i2}$, $\mathbf{V}_{i3}$ and $\mathbf{V}_{i4}$, as the sample covariance matrices of $\mathbf{Y}_{i\bullet}$, $\mathbf{U}_i$, $\mathbf{W}_i$ and $\mathbf{Z}_i$, respectively. Following Barnhart and Williamson [8] and Prentice [9], we can similarly obtain an empirically corrected covariance matrix for the estimated parameters $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\mu}}_*, \hat{\boldsymbol{\delta}}^2, \hat{\boldsymbol{\sigma}}_*^2, \hat{\boldsymbol{\alpha}})'$.

If the data are normally distributed, we can obtain estimates for $\boldsymbol{\theta}$ by using MIXED procedure in the popular commercial software SAS [10]. The following syntax may be used:

```
proc mixed;
 class id method;
 model Y=method/noint s;
 random method/G subject=id type=un V;
 repeated /R group=method;
 run;
```

where the solution in the `model` statement provides the estimate for $\boldsymbol{\mu}_*$, the G matrix provides the estimates for $\boldsymbol{\delta}^2$ and $\boldsymbol{\rho}^{\mu}$, and the R matrix provides the estimates for $\boldsymbol{\sigma}_*^2$. Note that the MIXED procedure does not provide estimate for the covariance of $\hat{\boldsymbol{\theta}}$ and thus one would not

be able to carry out inferences. An alternative approach is to find the estimate for covariance of $\hat{\boldsymbol{\theta}}$ via bootstrap approach where sampling with replacement is taken at subject level.

4. SIMULATION STUDY

Simulation studies were conducted to evaluate the performance of the GEE approach for estimation and inference of the proposed indices. Simulations were performed for small ($n=25$), moderate ($n=50$) and large sample sizes ($n=100$, and 400) with two settings of parameters. In both settings, we assume that there are three methods and that each method produces $K_{ij}=3$ replicated readings on each subject. We first generate $(\mu_{i1},\mu_{i2},\mu_{i3})'$ from a three-dimensional (3D) multivariate normal distribution with mean $\boldsymbol{\mu}_*=(\mu_{*1},\mu_{*2},\mu_{*3})'$ and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\mu}}$ with δ_j^2 and $\rho_{jj'}^{\mu}\delta_j\delta_{j'}$ in the diagonal and off diagonal entries, respectively. Second, we generate $(e_{ij1},e_{ij2},e_{ij3})'$ from a 3D multivariate normal distribution with mean $(0,0,0)'$ and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{e}j}=\sigma_{*j}^2 I_3$, $j=1,2,3$. Then the data for the ith subject is $(y_{i12},y_{i13},y_{i21},y_{i22},y_{i23},y_{i31},y_{i32},y_{i33})'$ with $y_{ijk}=\mu_{ij}+e_{ijk}, j=1,2,3$ and $k=1,2,3$. Simulation results are based on 1000 simulated data sets.

The first parameter setting assumes that there is a small mean change among the true readings of each method and a high correlation among pairs of the true readings. We also assume low intra-method variability. This parameter setting implies high inter, moderate intra and total agreement as shown in Table I. Specifically, we used $\boldsymbol{\mu}_*=(0,0.1,0.2)'$, $\boldsymbol{\sigma}_*^2=(1.0,1.1,1.2)'$, $\boldsymbol{\delta}^2=(4.0,4.1,4.2)'$ and $\boldsymbol{\rho}^{\boldsymbol{\mu}}=(0.96,0.97,0.98)'$. We evaluated the performance of the GEE approach for this parameter setting based on 1000 simulated data sets. Table I shows the results of the GEE approach. These results indicate that for small sample size, there exists a minor bias on the point estimation and this method tends to underestimate the true parameter. For large sample sizes, the point estimation tends to be unbiased, indicating the consistency of the point estimates of these indices. The 95 per cent coverage is very close to 95 per cent for the large sample sizes.

The second parameter setting is designed for a large within method variability compared with the variability of the true readings of the methods. We used moderate correlation coefficients among the 'true' readings of all the three methods. Specifically, we set $\boldsymbol{\mu}_*=(1.0,1.2,1.4)'$, $\boldsymbol{\sigma}_*^2=(2.0,3.0,4.0)'$, $\boldsymbol{\delta}^2=(2.0,3.0,4.0)'$ and $\boldsymbol{\rho}^{\boldsymbol{\mu}}=(0.5,0.6,0.7)'$. This parameter setting implies moderate to low inter, intra and total agreement (see Table II). We present the simulation results of the GEE in Table II. From this table, we observe a similar trend in the results as shown in the first simulation study.

5. EXAMPLE

We use data from a carotid stenosis screening study conducted at Emory University from 1994 to 1996 for illustration. The data was originally analysed in previous papers [4, 8]. The purpose of the carotid stenosis screening study is to determine the suitability of the two magnetic resonance angiography (MRA) method and the intra-arterial angiogram (IA) method. The two methods using the MRA technology are two-dimensional (2D) time of flight and 3D time of flight. For each method, three raters using all three methods measured the carotid

Table I. Results of the first set of simulations based on 1000 data sets.

Sample size	Index	True value	Mean estimation	Empirical std.err	Std dev.	95 per cent Coverage
25	$\rho_c(\mu)$	0.968	0.965	0.030	0.031	0.943
	ρ_1^I	0.800	0.780	0.060	0.068	0.915
	ρ_2^I	0.788	0.765	0.063	0.074	0.901
	ρ_3^I	0.777	0.754	0.065	0.071	0.912
	$\rho_c(Y)$	0.763	0.741	0.055	0.061	0.912
50	$\rho_c(\mu)$	0.968	0.966	0.020	0.021	0.937
	ρ_1^I	0.800	0.792	0.042	0.043	0.944
	ρ_2^I	0.788	0.782	0.044	0.046	0.925
	ρ_3^I	0.777	0.754	0.046	0.071	0.912
	$\rho_c(Y)$	0.763	0.755	0.039	0.041	0.933
100	$\rho_c(\mu)$	0.968	0.966	0.014	0.015	0.935
	ρ_1^I	0.800	0.796	0.030	0.032	0.924
	ρ_2^I	0.788	0.783	0.031	0.034	0.934
	ρ_3^I	0.777	0.775	0.032	0.034	0.938
	$\rho_c(Y)$	0.763	0.759	0.028	0.031	0.928
400	$\rho_c(\mu)$	0.968	0.967	0.007	0.007	0.950
	ρ_1^I	0.800	0.799	0.015	0.015	0.953
	ρ_2^I	0.788	0.787	0.016	0.016	0.951
	ρ_3^I	0.777	0.777	0.016	0.017	0.951
	$\rho_c(Y)$	0.763	0.762	0.014	0.014	0.949

stenosis of both the left and right arteries for 55 subjects. One would like to estimate intra-, inter-method and total agreement between the three methods. Because there are also raters involved, one may also like to find out intra-, inter- and total agreement between the three raters. However, because there is no replications made by the raters, we can only assess total agreement between the three raters. We make the assumption that the three rater's readings using the same method are replications of the method. This is supported by previous analyses. According to the results in Reference [4], the accuracy component of the OCCC among the three raters using the same method is very high ($\geqslant 0.96$). Therefore, it is reasonable to treat readings from the three raters of each method as replications of each method. This assumption is also supported by scatter plots (not shown) where readings by the three raters using the same method scatter evenly around the 45° line. When we treat the readings from the three raters using a particular method as the replications for this method, the total agreement between the three raters using a particular method is in fact the intra-method agreement. Therefore, we only assess intra-, inter-method and total agreement between the three methods here.

Table III show the points estimates using the GEE and mixed model approaches and 95 per cent CI based on the GEE approach for left and right arteries separately. Estimates based on GEE and MIXED procedure are very similar indicating that the data are probably normally distributed. Therefore, our discussion of results are based on the GEE approach here where standard errors are the by-product of the approach.

Table II. Results of the second set of simulations based on 1000 data sets.

Sample size	Index	True value	Mean estimation	Empirical std.err	Std dev.	95 per cent Coverage
25	$\rho_c(\mu)$	0.586	0.567	0.134	0.142	0.904
	ρ_1^I	0.500	0.473	0.108	0.121	0.906
	ρ_2^I	0.500	0.466	0.108	0.125	0.878
	ρ_3^I	0.500	0.470	0.107	0.115	0.908
	$\rho_c(Y)$	0.295	0.277	0.079	0.086	0.882
50	$\rho_c(\mu)$	0.586	0.577	0.096	0.100	0.925
	ρ_1^I	0.500	0.489	0.078	0.081	0.942
	ρ_2^I	0.500	0.490	0.078	0.084	0.919
	ρ_3^I	0.500	0.487	0.079	0.082	0.924
	$\rho_c(Y)$	0.295	0.288	0.060	0.061	0.922
100	$\rho_c(\mu)$	0.586	0.580	0.069	0.75	0.929
	ρ_1^I	0.500	0.495	0.056	0.060	0.926
	ρ_2^I	0.500	0.493	0.057	0.057	0.946
	ρ_3^I	0.500	0.496	0.057	0.057	0.941
	$\rho_c(Y)$	0.295	0.292	0.043	0.047	0.919
400	$\rho_c(\mu)$	0.586	0.583	0.035	0.035	0.946
	ρ_1^I	0.500	0.498	0.029	0.028	0.961
	ρ_2^I	0.500	0.498	0.029	0.029	0.953
	ρ_3^I	0.500	0.498	0.029	0.030	0.947
	$\rho_c(Y)$	0.295	0.293	0.022	0.022	0.944

Table III. Results for carotid stenosis data.

	Left artery			Right artery		
	Estimates		95 per cent CI	Estimates		95 per cent CI
Index	GEE	MIXED	GEE	GEE	MIXED	GEE
Intra-method agreement						
IA	0.882	0.884	(0.782, 0.982)	0.915	0.916	(0.866, 0.964)
2D	0.621	0.626	(0.456, 0.786)	0.604	0.610	(0.443, 0.765)
3D	0.614	0.647	(0.443, 0.785)	0.616	0.621	(0.453, 0.779)
Inter-method agreement						
Among 3 methods	0.763	0.758	(0.608, 0.918)	0.848	0.847	(0.736, 0.960)
IA vs 2D	0.755	0.754	(0.553, 0.957)	0.846	0.845	(0.746, 0.946)
IA vs 3D	0.624	0.614	(0.381, 0.867)	0.765	0.764	(0.573, 0.957)
2D vs 3D	0.925	0.919	(0.796, 1.000)	0.943	0.939	(0.814, 1.000)
Total agreement						
Among 3 methods	0.533	0.539	(0.386, 0.680)	0.594	0.597	(0.459, 0.729)
IA vs 2D	0.557	0.559	(0.383, 0.731)	0.634	0.636	(0.497, 0.771)
IA vs 3D	0.464	0.468	(0.266, 0.662)	0.575	0.577	(0.398, 0.751)
2D vs 3D	0.573	0.587	(0.434, 0.712)	0.576	0.579	(0.437, 0.715)

Overall, the total agreement among the three methods is poor (0.533 and 0.594 for left and right arteries, respectively). This is due to moderate inter-method agreement (0.763 and 0.848 for left and right arteries, respectively) and fair intra-method agreement (about 0.61–0.62) using the MRA technology. The intra-method agreement using the IA method is high (0.882 and 0.915 for left and right arteries, respectively). By examining the pairwise inter-method agreement, we note that the moderate inter-method agreement is due to the moderate agreement between the methods using different technology (IA vs MRA 2D or IA vs MRA 3D) (pairwise inter-method agreement ranging from 0.624 to 0.846). The pair inter-method agreement between the MRA 2D and MRA 3D method is actually high (0.925 and 0.943 for left and right arteries, respectively) indicating that there is not major true differences between the methods using the MRA technology. If one needs to choose one of the two MRA methods, the MRA 2D may be preferred because it may be easier to produce and to examine 2D images than the 3D images. In summary, we observe disagreement among the two MRA methods and the IA method. This is mainly due to disagreement among raters using the MRA methods. Therefore, training is needed for raters who use the MRA methods in order to improve the agreement between the MRA methods and the IA methods. There is no difference among the two MRA methods.

We observe that the inter-method agreement of the two methods using the same MRA technology is higher than the inter-method agreement of the two methods using different technology. It is of interest to test whether the difference is significant. We extend our approach to incorporate covariate impact on agreement in the next section which allows us to carry out such a test.

6. AN EXTENSION

If an agreement study also contains information on covariates, such as age, gender, rater experience, etc. then it may be of interest to investigate the direction of association between the covariates and the agreement index so that we can better understand the impact of the covariates on the measurement process. The proposed GEE approach can be easily extended to incorporate covariates. Let $\mathbf{Y}_i$ $(i=1,\ldots,N)$ be the vector of readings for subject i with corresponding matrix X_i for p covariates. The covariates can either be subject-specific or method-specific or both. First, we model the marginal mean of $\mathbf{Y}_i$ by $E(\mathbf{Y}_i)=\boldsymbol{\mu}_i(\beta)=\mathbf{X}_i\boldsymbol{\beta}$. The parameter estimates of $\boldsymbol{\beta}$ are obtained by GEE as

$$\sum_{i=1}^{N}\mathbf{D}_{i1}'\mathbf{V}_{i1}^{-1}(\mathbf{Y}_{i\bullet}-\boldsymbol{\mu}_i(\boldsymbol{\beta}))=\mathbf{0}$$

where $\mathbf{D}_{i1}=\partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta}$ and $\mathbf{V}_{i1}$ is the working covariance matrix. In the second and third sets of estimating equations, we estimate $\boldsymbol{\sigma}_*^2$ and $\boldsymbol{\delta}^2$. For simplicity, we assume that $\boldsymbol{\sigma}_*^2$ and $\boldsymbol{\delta}^2$ do not depend on covariates. The second and third sets of estimating equations are as follows:

$$\sum_{i=1}^{N}\mathbf{D}_{i2}'\mathbf{V}_{i2}^{-1}(\mathbf{U}_i-\boldsymbol{\delta}^2)=\mathbf{0}$$

$$\sum_{i=1}^{N}\mathbf{D}_{i3}'\mathbf{V}_{i3}^{-1}(\mathbf{W}_i-\mathbf{g}(\boldsymbol{\mu}(\boldsymbol{\beta}),\boldsymbol{\delta}^2))=\mathbf{0}$$

The fourth set of estimating equations is to model the marginal mean of $\mathbf{Z}_i = (Y_{i1\bullet}Y_{i2\bullet}, Y_{i1\bullet}Y_{i3\bullet}, \ldots, Y_{i(J-1)\bullet}Y_{iJ\bullet})'$ as a function of covariates to evaluate the covariates' impact on agreement. Note that $E(\mathbf{Z}_{ijj'}) = \mu_{*j}\mu_{*j'} + \rho^{\mu}_{jj'}\delta_j\delta_{j'} = \mu_{*j}\mu_{*j'} + (\frac{1}{2})\xi_{jj'}\rho_{c(jj')}(\mu)$, where $\xi_{jj'} = \delta_j^2 + \delta_{j'}^2 + (\mu_j - \mu_{j'})^2$. We use Fisher's Z-transformation to model the pairwise inter-CCCs, $\rho_{c(jj')}(\mu)$, as: $(1/2)\log(1 + \rho_{c(jj')}(\mu))/(1 - \rho_{c(jj')}(\mu)) = \mathbf{Q}_{ijj'}\boldsymbol{\alpha}$. where $\mathbf{Q}_{ijj'}$ includes a subset of covariates of $\mathbf{X}_i$ as well as indicator variables for pair (j, j'). Therefore, the fourth set of estimating equations is to obtain the parameter estimates of $\boldsymbol{\alpha}$ by modelling the marginal mean of $\mathbf{Z}_i$

$$\sum_{i=1}^{N} \mathbf{D}'_{i4}\mathbf{V}_{i4}^{-1}(\mathbf{Z}_i - \mathbf{h}(\boldsymbol{\beta}, \delta^2, \boldsymbol{\alpha})) = \mathbf{0}$$

We apply the above extension to the carotid stenosis example to test whether the inter-method agreement between two methods (MRA 2D vs MRA 3D) using the MRA technology is different from the inter-method agreement between two methods using different technology (IA vs MRA 2D and IA vs MRA 3D). We used the following design matrices for $\mathbf{X}$ and $\mathbf{Q}$. $\mathbf{X}$ is a 3×3 design matrix formed by intercept term and two indicator variables for the MRA 2D and MRA 3D methods, respectively. Let $\boldsymbol{\mu}_* = (\mu_{*1}, \mu_{*2}, \mu_{*3})'$ be the means of IA, MRA 2D, and MRA 3D readings, respectively, then we have

$$\boldsymbol{\mu}_* = \mathbf{X}\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_0 + \beta_1 \\ \beta_0 + \beta_2 \end{pmatrix}$$

$\mathbf{Q}$ is a matrix formed by intercept term and two indicator variables for method pairs of (IA, MRA 2D) and (IA, MRA 3D) respectively. We then have

$$\begin{pmatrix} \frac{1}{2}\log\frac{1+\rho_{c12}(\mu)}{1-\rho_{c12}(\mu)} \\ \frac{1}{2}\log\frac{1+\rho_{c13}(\mu)}{1-\rho_{c13}(\mu)} \\ \frac{1}{2}\log\frac{1+\rho_{c23}(\mu)}{1-\rho_{c23}(\mu)} \end{pmatrix} = \mathbf{Q}\boldsymbol{\alpha} = \begin{pmatrix} \alpha_0 + \alpha_1 \\ \alpha_0 + \alpha_2 \\ \alpha_0 \end{pmatrix}$$

To test if the inter-method agreement of MRA 2D vs MRA 3D is significantly different from that of IA vs MRA 2D and from that of IA vs MRA 3D, we test $H_0 : \alpha_1 = 0$ and $H_0 : \alpha_2 = 0$, respectively.

The GEE estimates of $\boldsymbol{\alpha}$ are $(1.62, -0.64, -0.89)'$ and $(1.76, -0.52, -0.75)'$ for left and right arteries, respectively. The p-value for testing $H_0 : \alpha_1 = 0$ is 0.143 and 0.222 for the left and right artery, respectively, indicating that the inter-CCC between methods of MRA 2D and MRA 3D is not significantly higher than the inter-CCC between method IA and MAR 2D. The p-value for testing $H_0 : \alpha_2 = 0$ is 0.042 for the left artery and 0.101 for the right artery. Therefore, the inter-method agreement of the two MRA methods is significantly higher than that of the IA and MRA 3D methods for the left artery.

7. DISCUSSION

In this paper, we have proposed indices for assessing the intra, inter and the total agreement with replicated readings produced by several methods, where the methods can be either different technologies, instruments, or human observers. We evaluated the inter-method agreement based on the method's true readings. We used the ICC of a one-way ANOVA model to assess the intra-method agreement and we developed the total agreement index based on the observed readings. Furthermore, we investigated the relationship between the total agreement index and intra, inter agreement indices. These indices help to identify whether the disagreement is due to either intra-method variability or inter-method variability or both. Therefore, these indices provide directions on how to improve agreement among multiple methods. We proposed the GEE approach for estimation and inference of these indices. We evaluated the performance of GEE through simulation studies and we used the data from a carotid stenosis study to illustrate the use of the proposed methods. We also presented an extension of evaluating the covariates' effect on agreement. We note that the mixed model approach can be used to obtain estimates from SAS software if the data is normal. However, inference using the mixed model approach will require bootstrapping and it is not obvious how to incorporate covariates effects on the correlation parameters. The GEE approach is attractive due to its easy implementation when there are covariates.

We used CCC-type indices to assess inter-method and total agreement because these indices do not require exchangeable assumption that is needed in ICC-type indices based on classical test theory [11]. Extending the classical test theory, generalizability theory (GT) [12, 13] has been developed to assess measurements in education and psychology. Although the coefficients defined in the context of GT also do not require the exchangeable assumption, they depend on explicit specification of ANOVA models. It is possible to extend the definitions of these coefficients using the GT concepts without the ANOVA model assumption and then compare these coefficients to the CCC-type indices. This is beyond the scope of this paper and it is a topic for future research. For a special case where there are J fixed methods and no replications, if we further assume that the readings follow a two way ANOVA model without interaction where subject is treated as random effect and method is treated as fixed effect, then the CCC (total agreement) is the same as one version of ICC (case 3A in Reference [14]) and the coefficient of dependability under GT.

Dunn and Roberts [15] developed a modelling approach for method comparison data. Their approach differs from ours in two aspects. First, Dunn and Roberts [15] made the assumption that the true value from one method is linearly related to the true value of the other method at the subject level. This assumption will impose restriction of $\rho^{\mu}_{jj'} = 1$ in our context. Second, Dunn and Roberts used the ratio of the precision to compare inter-method agreement. Based on the definition, this ratio is in fact an index of partial total agreement. The ratio is a measure of total agreement because it is defined at the level of observed readings (noting that intra-method variabilities are used in the formula of this ratio). It is only a partial measure of total agreement because only the scale shift (β parameter in their paper), no location shift (μ parameter in their paper) between the true values is taken into account in the definition.

Our research focuses on the fixed methods only. In practice, the methods may be considered as a random sample from a large population. This will be a topic for future investigation. Haber *et al.* [16] have developed a new index, *coefficient of inter-observer variability* (CIV), for evaluating observer agreement. The main difference between CIV and CCC-type indices

is that the two coefficients are scaled differently. The CIV does not depend on the between-subject variability, and it does not depend on whether the raters are considered fixed or random. It would be of future interest to study the relationship between the CIV and the CCC-type indices.

To design an agreement study with replications, one will need to decide on the optimal choices of number of subjects and number of replications. This is the topic for our future research.

ACKNOWLEDGEMENTS

This research was supported in part by Emory University Quadrangle Fund and by NIH Grant R01 MH070028-01A1.

REFERENCES

1. Lin LI. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 1989; **45**:225–268.
2. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin* 1979; **86**:420–428.
3. Bartko JJ. The intraclass correlation coefficient as a measure of reliability. *Psychological Reports* 1966; **19**:3–11.
4. Barnhart HX, Haber MJ, Song J. Overall concordance correlation coefficient for evaluating agreement among multiple observers. *Biometrics* 2002; **58**:1020–1027.
5. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**:13–22.
6. Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 1986; **42**: 121–130.
7. Liang KY, Zeger SL, Qaqish B. Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society*, *Series B*, *Methodological* 1992; **54**:3–24.
8. Barnhart HX, Williamson JM. Modeling concordance correlation via GEE to evaluate reproducibility. *Biometrics* 2001; **57**:931–940.
9. Prentice RL. Correlated binary regression with covariates specific to each binary observation. *Biometrics* 1988; **44**:1033–1048.
10. SAS Institute. *SAS/STAT User's Guide*, *Version 8* (4th edn). SAS Publishing: Cary, NC, 2000.
11. Lord FM, Novick MR. *Statistical Theories of Mental Test Scores*. Addison-Wesley Publishing Company: Reading, MA, 1968.
12. Cronbach LJ, Gleser GC, Nanda H, Rajaratnam N. *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. Wiley: New York, 1972.
13. Brennan RL. *Generalizability Theory*. Springer: New York, 2001.
14. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychological Methods* 1996; **1**:30–46.
15. Dunn G, Roberts C. Modelling method comparison data. *Statistical Methods in Medical Research* 1999; **8**: 161–179.
16. Haber MJ, Barnhart HX, Song J. Assessing agreement among observers via interobserver variability. *Journal of Data Science* 2005; **3**: to appear.

