BIOMETRICS 58, 1020–1027
December 2002

Overall Concordance Correlation Coefficient for Evaluating Agreement Among Multiple Observers

Huiman X. Barnhart,* Michael Haber, and Jingli Song

Department of Biostatistics, The Rollins School of Public Health of Emory University,
1518 Clifton Road NE, Atlanta, Georgia 30322, U.S.A.
**mail:* hbarnha@sph.emory.edu

SUMMARY. Accurate and precise measurement is an important component of any proper study design. As elaborated by Lin (1989, *Biometrics* **45,** 255–268), the concordance correlation coefficient (CCC) is more appropriate than other indices for measuring agreement when the variable of interest is continuous. However, this agreement index is defined in the context of comparing two fixed observers. In order to use multiple observers in a study involving large numbers of subjects, there is a need to assess agreement among these multiple observers. In this article, we present an overall CCC (OCCC) in terms of the interobserver variability for assessing agreement among multiple fixed observers. The OCCC turns out to be equivalent to the generalized CCC (King and Chinchilli, 2001, *Statistics in Medicine* **20,** 2131–2147; Lin, 1989; Lin, 2000, *Biometrics* **56,** 324–325) when the squared distance function is used. We evaluated the OCCC through generalized estimating equations (Barnhart and Williamson, 2001, *Biometrics* **57,** 931–940) and U-statistics (King and Chinchilli, 2001) for inference. This article offers the following important points. First, it addresses the precision and accuracy indices as components of the OCCC. Second, it clarifies that the OCCC is the weighted average of all pairwise CCCs. Third, it is intuitively defined in terms of interobserver variability. Fourth, the inference approaches of GEE and the U-statistics are compared via simulations for small samples. Fifth, we illustrate the use of the OCCC by two medical examples with the GEE, U-statistics, and bootstrap approaches.

KEY WORDS: Agreement; Overall concordance correlation coefficient; Reproducibility.

1. Introduction

In health care practice, clinical measurements serve as a basis for diagnostic, prognostic, and therapeutic evaluations. As technology continues to advance, new methods/instruments for diagnostic, prognostic, and therapeutic evaluations become available. Before a new method or a new instrument is adopted for use in measuring a variable of interest, one needs to ensure the accuracy and precision of the measurement. Often, a reliability or a validity study involving multiple observers is conducted in clinical or experimental settings. If the outcome variable is continuous, Lin (1989, 1992) stated that the appropriate index for measuring agreement between two observers is the concordance correlation coefficient (CCC). Lin (1989) argued that, even though the agreement is often evaluated by using the Pearson correlation coefficient, the paired t-test, the least square analysis of slope (=1) and intercept (=0), the coefficient of variation, or the intraclass correlation coefficient, none of these can fully assess the desired reproducibility characteristics. The Pearson correlation coefficient only measures precision of a linear relationship, not accuracy. Both the paired t-test and least squares analysis can falsely reject (accept) the hypothesis of high agreement when the residual error is very small (large). The coefficient of variation and the intraclass correlation coefficient often assume that the two readings by two observers are interchangeable. The advantage of CCC is that this index is based on the differences between the observations made by two observers on the same subject, and thus it evaluates the agreement between two readings by measuring the variation from the 45° line through the origin. The CCC has good intuitive interpretation because it includes components of both precision (degree of variation) and accuracy (degree of location or scale shift).

Use of the CCC as a measure of reproducibility has gained popularity in practice since its introduction by Lin (1989). However, this agreement index is defined in the context of comparing two fixed observers. Because the reliability and validity studies often involve more than two observers, especially for studies with large numbers of subjects, there is a need to assess agreement among multiple observers. In this article, we present an overall CCC (OCCC) in terms of the interobserver variability for assessing agreement among multiple fixed observers. The OCCC turns out to be equivalent to the generalized CCC (Lin, 1989, 2000; King and Chinchilli, 2001) when the squared distance function is used. We evaluated the OCCC through generalized estimating equations (Barnhart and Williamson, 2001) and U-statistics (King and Chinchilli, 2001) for inference. In Section 2, we give our definition of this index and examine its properties in comparison with the extended concordance correlation coefficient. We propose an alternative approach, generalized estimating equations (GEE) (Barnhart and Williamson, 2001), to conduct inference on the OCCC in Section 3. Simulation studies are performed in Section 4 to compare the proposed GEE approach with the U-statistics approach by King and Chinchilli (2001). We illustrate the use of the OCCC and compare the GEE, U-statistics, and bootstrap inference approaches by using two medical examples in Section 5. A short discussion is presented in Section 6.

2. Overall Concordance Correlation Coefficient

Suppose that each of J observers assesses each of N subjects (a random sample from a population of interest) with a continuous scale Y. Let $Y_1, \ldots, Y_J$ be the readings from the J fixed observers for a randomly selected subject. For two observers, say the jth and kth observers, the concordance correlation coefficient (CCC) introduced by Lin (1989) is defined by the following scaled expected squared difference $\mathrm{E}\{(Y_j - Y_k)^2\}$,

$$\rho^c_{jk} = 1 - \frac{\mathrm{E}\left\{(Y_j - Y_k)^2\right\}}{\mathrm{E}\left\{(Y_j - Y_k)^2 \mid Y_j, Y_k \text{ are uncorrelated}\right\}} = \frac{2\sigma_{jk}}{\sigma_j^2 + \sigma_k^2 + (\mu_j - \mu_k)^2} = \rho_{jk}\chi^a_{jk},$$

where $\mu_j = \mathrm{E}(Y_j)$, $\mu_k = \mathrm{E}(Y_k)$, $\sigma_j^2 = \mathrm{var}(Y_j)$, $\sigma_k^2 = \mathrm{var}(Y_k)$, and $\sigma_{jk} = \mathrm{cov}(Y_j, Y_k) = \sigma_j\sigma_k\rho_{jk}$. The CCC is a product of two components, precision (ρ_{jk}) and accuracy (χ^a_{jk}), where $\chi^a_{jk} = 2\sigma_j\sigma_k/\{\sigma_j^2 + \sigma_k^2 + (\mu_j - \mu_k)^2\} = \{(v_{jk} + 1/v_{jk} + u^2_{jk})/2\}^{-1}$, with $v_{jk} = \sigma_j/\sigma_k$ representing scale shift and $u_{jk} = (\mu_j - \mu_k)/(\sigma_j\sigma_k)^{1/2}$ representing location shift relative to the scale. Note that $\rho^c_{jk} = \rho_{jk}$ if and only if $\mu_j = \mu_k$ and $\sigma_j^2 = \sigma_k^2$. In general, we have $-1 \leq -|\rho_{jk}| \leq \rho^c_{jk} \leq |\rho_{jk}| \leq 1$ and $\rho^c_{jk} = 1$ if and only if $\mu_j = \mu_k$, $\sigma_j^2 = \sigma_k^2$, and $\rho_{jk} = 1$.

Because it is intuitive to use the squared difference to describe the disagreement between readings from two observers, it is natural to use interobserver sample variability, $V = \Sigma_{j=1}^J (Y_j - Y_\bullet)^2/(J-1)$, where $Y_\bullet$ is the arithmetic mean, in place of the squared difference to describe the variability among readings from multiple observers. We propose an index, the overall concordance correlation coefficient (OCCC), for measuring agreement among multiple observers by scaling the expected interobserver variability to be between -1 and 1. The OCCC is defined as

$$\rho^c_o = 1 - \frac{\mathrm{E}(V)}{\mathrm{E}(V \mid Y_1, \ldots, Y_J \text{ are uncorrelated})}.$$

Note that, when $J = 2$, we have $\rho^c_o = \rho^c_{12}$. Thus, the OCCC is a natural extension of the CCC. We derive relationships between the OCCC and the pairwise CCC and between the OCCC and the moments of the Y_j's. First, note that we can write V as a linear combination of all pairwise squared differences, $V = \Sigma_{j=1}^{J-1}\Sigma_{k=j+1}^{J}(Y_j - Y_k)^2/\{J(J-1)\}$. Thus, we have

$$\rho^c_o = 1 - \mathrm{E}\left\{\sum_{j=1}^{J-1}\sum_{k=j+1}^{J}(Y_j - Y_k)^2\right\} \div \mathrm{E}\left\{\sum_{j=1}^{J-1}\sum_{k=j+1}^{J}(Y_j - Y_k)^2 \,\middle|\, Y_1, \ldots, Y_J \text{ are uncorrelated}\right\} \tag{1}$$

$$= \frac{\displaystyle\sum_{j=1}^{J-1}\sum_{k=j+1}^{J}\xi_{jk}\rho^c_{jk}}{\displaystyle\sum_{j=1}^{J-1}\sum_{k=j+1}^{J}\xi_{jk}}, \tag{2}$$

where $\xi_{jk} = \mathrm{E}\{(Y_j - Y_k)^2 \mid Y_1, \ldots, Y_J \text{ are uncorrelated}\} = (\mu_j - \mu_k)^2 + \sigma_j^2 + \sigma_k^2$. Therefore, the OCCC can be interpreted as a weighted average of all pairwise CCCs with weights ξ_{jk}'s, where higher weights are given to the pairs of observers whose readings have higher variances and larger mean differences. This makes intuitive sense as the pairs of observers are penalized proportionally by the disagreement due to their variances and the squared mean difference. Second, we can also rewrite the OCCC as a function of means, variances, and covariances as

$$\rho^c_o = \frac{2\displaystyle\sum_{j=1}^{J-1}\sum_{k=j+1}^{J}\sigma_{jk}}{(J-1)\displaystyle\sum_{j=1}^{J}\sigma_j^2 + J\sum_{j=1}^{J}(\mu_j - \mu_\bullet)^2} = \frac{2\displaystyle\sum_{j=1}^{J-1}\sum_{k=j+1}^{J}\sigma_{jk}}{(J-1)\displaystyle\sum_{j=1}^{J}\sigma_j^2 + \sum_{j=1}^{J-1}\sum_{k=j+1}^{J}(\mu_j - \mu_k)^2}.$$

If we rewrite the pairwise CCCs as products of precision and accuracy, we have from (2) that

$$\rho^c_o = \frac{\displaystyle\sum_{j=1}^{J-1}\sum_{k=j+1}^{J}\xi_{jk}\rho_{jk}\chi^a_{jk}}{\displaystyle\sum_{j=1}^{J-1}\sum_{k=j+1}^{J}\xi_{jk}}.$$

Furthermore, if we assume $\rho_{jk} = \rho$ for all j and k (i.e., same precision for every pair of observers), then the OCCC can be expressed as a product of precision and overall accuracy, $\rho^c_o = \rho\chi^a$, with

$$\chi^a = \frac{\displaystyle\sum_{j=1}^{J-1}\sum_{k=j+1}^{J}\xi_{jk}\chi^a_{jk}}{\displaystyle\sum_{j=1}^{J-1}\sum_{k=j+1}^{J}\xi_{jk}}.$$

It turns out that the proposed OCCC is the same as the index suggested by Lin (1989) in the section of future studies with correction of typographical errors (Lin, 2000). This index is also equivalent to the extended concordance correlation coefficient proposed independently by King and Chinchilli (2001) when a squared distance function is used. Their index is based on the square difference of all possible pairs, but they did not examine any properties of their index. We note that the OCCC defined in this article offers three important insights. First, it addresses the precision and accuracy indices as components of the OCCC. Second, it clarifies that the OCCC is the weighted average of pairwise CCCs. Third, it is intuitively defined in terms of interobserver variability. Lin (1989) did not provide inference information regarding this index. King and Chinchilli (2001) proposed an inference approach based on U-statistics. However, they did not perform simulation studies for this index nor did they provide an example of application of the index. In the next section, we propose two alternative approaches, generalized estimating equations (GEE) and bootstrap, for inference. We evaluate the OCCC through simulation studies in Section 4 to compare the GEE approach to the U-statistics approach. We then compare the GEE, the U-statistics, and the bootstrap approaches using two medical examples in Section 5.

To better understand the OCCC index, we also derive simple expressions for ρ_o^c in special cases. First, the OCCC is the same as the correlation coefficient if and only if $\mu_j = \mu$, $\sigma_j^2 = \sigma^2$, and $\sigma_{jk} = \sigma^2\rho$ for all j and k. Second, under homogeneity assumptions of $\sigma_j^2 = \sigma^2$ and $\rho_{jk} = \rho$ for all j and k, we have

$$\rho_o^c = \frac{\sigma^2\rho}{\sigma^2 + \displaystyle\sum_{j=1}^{J}\frac{(\mu_j - \mu_\bullet)^2}{J-1}} = \rho\chi^a,$$

where $\chi^a = \sigma^2/\{\sigma^2 + \Sigma_{j=1}^J(\mu_j - \mu_\bullet)^2/(J-1)\}$. In this case, the OCCC is the ratio of the variance multiplied by the correlation to the sum of the variance and the expected interobserver variability $(\Sigma_{j=1}^J(\mu_j - \mu_\bullet)^2/(J-1))$. Third, if we assume a two-way analysis of variance (ANOVA) model without interaction for the observed readings, we can write $Y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}$, where Y_{ij} is the reading from observer j for subject i, $i = 1, \ldots, N$. The common assumptions for the two-way ANOVA model are that (a) α_i's are i.i.d. with mean zero and variance σ_s^2, (b) β_j's are fixed with $\Sigma_{j=1}^J\beta_j = 0$, (c) e_{ij}'s are i.i.d. with mean zero and variance σ_e^2, and (d) α_i and e_{ij} are mutually independent. Under these assumptions, we have $\mathrm{E}\{\Sigma_{j=1}^J(Y_{ij} - Y_{i\bullet})^2\} = \Sigma_{j=1}^J\beta_j^2 + (J-1)\sigma_e^2$ and $\mathrm{E}_i\{\Sigma_{j=1}^J(Y_{ij} - Y_{i\bullet})^2 \mid Y_{ij} \text{ are independent}\} = \Sigma_{j=1}^J\beta_j^2 + (J-1)\sigma_s^2 + (J-1)\sigma_e^2$. Thus, we have

$$\rho_o^c = \frac{\sigma_s^2}{\sigma_s^2 + \displaystyle\sum_{j=1}^{J}\beta_j^2/(J-1) + \sigma_e^2}.$$

In this case, the OCCC is the same as the intraclass correlation coefficient for assessing observer agreement under the same two-way model (see Case 3A in the article by McGraw and Wong (1996)), i.e., it is equal to the proportion of subject variability (σ_s^2) over the sum of subject variability, interobserver variability $(\Sigma_{j=1}^J\beta_j^2/(J-1))$, and error variability. We note that other versions of intraclass correlation coefficients do not include the interobserver variability $(\Sigma_{j=1}^J\beta_j^2/(J-1))$ in the denominator (Bartko, 1966).

3. Estimation and Inference

Let Y_{ij} be the reading from observer j for subject $i, i = 1, \ldots, N$. Estimation for the OCCC can be accomplished by the method of moments. Specifically, the OCCC is estimated by

$$\hat{\rho}_o^c = \frac{\displaystyle\sum_{j=1}^{J-1}\sum_{k=j+1}^{J}\hat{\xi}_{jk}\hat{\rho}_{jk}^c}{\displaystyle\sum_{j=1}^{J-1}\sum_{k=j+1}^{J}\hat{\xi}_{jk}}, \tag{3}$$

where $\hat{\xi}_{jk} = (Y_{\bullet j} - Y_{\bullet k})^2 + S_j^2 + S_k^2$ and $\hat{\rho}_{jk}^c = 2S_{jk}/\{S_j^2 + S_k^2 + (Y_{\bullet j} - Y_{\bullet k})^2\}$, with $Y_{\bullet j}$'s, S_j's, and S_{jk}'s as sample means, variances, and covariances, respectively. The expression for $\hat{\rho}_o^c$ can also be rewritten in terms of sample means, variances, and covariances as

$$\hat{\rho}_o^c = \frac{2\displaystyle\sum_{j=1}^{J-1}\sum_{k=j+1}^{J}S_{jk}}{(J-1)\displaystyle\sum_{j=1}^{J}S_j^2 + J\sum_{j=1}^{J}(Y_{\bullet j} - Y_{\bullet\bullet})^2}. \tag{4}$$

We now turn to the estimation of the standard error of $\hat{\rho}_o^c$. Note that numerous assessments (J readings by the J observers) are made on the same subject and these measurements will tend to be positively correlated. This correlation must be taken into account for valid inference. King and Chinchilli (2001) proposed using U-statistics for inference on OCCC. We propose an alternative approach, the generalized estimating equations approach (Barnhart and Williamson, 2001), for the inference on $\hat{\rho}_o^c$. Following Barnhart and Williamson (2001) for the case of no subject-specific covariates and using the Fisher's Z-transformation $\alpha_{jk} = 0.5\log\{(1 + \rho_{jk}^c)/(1 - \rho_{jk}^c)\}$ for stability, we note that no iterations are needed to solve these GEE equations because there are no covariates. Briefly, the three GEE equations are as follows:

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

$$\sum_{i=1}^{N}\mathbf{F}_i'\mathbf{H}_i^{-1}\left\{\mathbf{Y}_i^2 - \boldsymbol{\delta}_i^2\left(\boldsymbol{\sigma}^2, \boldsymbol{\mu}\right)\right\} = \mathbf{0},$$

$$\sum_{i=1}^{N}\mathbf{C}_i'\mathbf{W}_i^{-1}\left\{\mathbf{U}_i - \boldsymbol{\theta}_i\left(\boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2\right)\right\} = \mathbf{0},$$

where $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iJ})'$, $\mathrm{E}(\mathbf{Y}_i) = \boldsymbol{\mu}_i(\boldsymbol{\mu})$ with $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_J)'$, $\mathbf{D}_i = \partial\boldsymbol{\mu}_i(\boldsymbol{\mu})/\partial\boldsymbol{\mu}$, $\mathbf{Y}_i^2 = (Y_{i1}^2, \ldots, Y_{iJ}^2)'$, $\mathrm{E}(\mathbf{Y}_i^2) = \boldsymbol{\delta}_i^2(\boldsymbol{\sigma}^2, \boldsymbol{\mu})$ with $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_J^2)'$, $\mathbf{H}_i = \partial\boldsymbol{\delta}_i^2(\boldsymbol{\sigma}^2)/\partial\boldsymbol{\sigma}^2$, $\mathbf{U}_i = (Y_{i1}Y_{i2},\ Y_{i1}Y_{i3}, \ldots, Y_{i1}Y_{iJ}, \ldots, Y_{i(J-1)}Y_{iJ})'$, $\mathrm{E}(\mathbf{U}_i) = \boldsymbol{\theta}(\boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ with $\boldsymbol{\alpha} = (\alpha_{11}, \ldots, \alpha_{(J-1)J})'$, $\mathbf{C}_i = \partial\boldsymbol{\theta}_i(\boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2)/\partial\boldsymbol{\alpha}$, and $\mathbf{V}_i, \mathbf{H}_i$, and $\mathbf{W}_i$ are working covariance matrices. If the independent working correlation matrices are used, the corresponding estimates for μ_j, σ_j^2, and ρ_{jk}^c turn out to be the moment estimates. Furthermore, an empirically adjusted covariance matrix for $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\sigma}}^2, \hat{\boldsymbol{\alpha}})'$ can be obtained using the sandwich estimator in Barnhart and Williamson (2001). Use of the sandwich estimator allows one to adjust for the correlation among multiple readings made on the same subject even though this correlation may not be correctly specified (Liang and Zeger, 1986; Zeger and Liang, 1986). A delta method can then be applied to (3) to obtain the standard error for $\hat{\rho}_o^c$. Another approach for inference is to use bootstrap samples (Efron and Tibshirani, 1993). One can take m (e.g., m = 1000) bootstrap cluster samples from $\mathbf{Y}_i$'s $(i = 1, \ldots, N)$, where $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iJ})'$ and the unit of cluster is subject. Cluster sampling allows the adjustment of correlation among multiple readings made on the same subject. Then estimates of ρ_o^c can be computed for each of the m bootstrap samples by using sample means, variances, and covariances as in formula (4). Finally, the standard error or percentile confidence interval can be obtained by using the empirical distribution of the m estimates of ρ_o^c. Both of these inference approaches require assumptions only up to the second moment and no distributional assumption is required. The bootstrap approach is more computationally intensive than the GEE approach. However, because formula (4) is relatively easy to compute, we expect that both approaches are easy to implement in practice. In the next section, we compare the GEE approach with the U-statistics approach (King and Chinchilli, 2001) by simulation studies.

4. Simulations

We examine the bias of the estimated OCCC and determine how well the proposed standard error estimates perform with small sample sizes (100, 50, and 25) with two sets of simulations. In both settings, we assume that there are four observers and the data are generated from a multivariate normal distribution with mean $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3, \mu_4)'$ and covariance matrix $\boldsymbol{\Sigma}$. All simulation results are based on 1000 simulated data sets. The GEE and the U-statistics approaches are used for inference. We have used the bootstrap approach with m = 1000 for the sample sizes of 100 and 50 and the results (not shown) are similar to the GEE and the U-statistics approaches.

In the first set of simulations, we assume homogeneity and use the following true specifications: $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3, \mu_4) = (0.0, 0.2, 0.4, 0.6)$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1.0 & \rho & \rho & \rho \\ \rho & 1.0 & \rho & \rho \\ \rho & \rho & 1.0 & \rho \\ \rho & \rho & \rho & 1.0 \end{pmatrix}.$$

In this setting, we have $\rho_o^c = 3\rho/3.2 < \rho$. The simulation results are presented in Table 1. The approaches based on the GEE and the U-statistics yielded almost identical results. The estimated mean standard error is very close to the empirical standard deviation based on 1000 estimates of ρ_o^c from the 1000 data sets for all sample sizes. The 95% coverage based on the estimated standard error is smaller than expected, probably due to the slight underestimation of the true OCCC.

In the second set of simulations, we assume that there is no observer bias but there are differences in observer variability. The true specifications are $\boldsymbol{\mu} = (0, 0, 0, 0)$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1.0 & \rho & \sqrt{2}\rho & \sqrt{2}\rho \\ \rho & 1.0 & \sqrt{2}\rho & \sqrt{2}\rho \\ \sqrt{2}\rho & \sqrt{2}\rho & 2.0 & 2\rho \\ \sqrt{2}\rho & \sqrt{2}\rho & 2\rho & 2.0 \end{pmatrix}.$$

Table 1

Results of the first set of simulations based on 1000 *data sets*

True	True	Sample				Mean	95%	Adj. 95% coverage		
ρ	ρ_o^c	size	Method	Mean	SD	est. SE	Coverage	$N/(N-1)$[a]	$N/(N-2)$[b]	$N/(N-3)$[c]
0.5	0.469	100	GEE	0.464	0.0517	0.0492	93.8%	95.4%	96.3%	96.9%
			U-stat.			0.0491	93.8%	95.4%	96.3%	96.9%
		50	GEE	0.459	0.0702	0.0679	93.1%	93.9%	94.4%	94.9%
			U-stat.			0.0679	93.1%	93.8%	94.4%	95.0%
		25	GEE	0.449	0.1001	0.0906	89.5%	90.9%	91.5%	93.0%
			U-stat.			0.0904	89.4%	90.6%	91.4%	92.8%
0.7	0.656	100	GEE	0.651	0.0410	0.0398	93.1%	94.1%	95.1%	96.1%
			U-stat.			0.0398	93.2%	94.3%	94.9%	96.1%
		50	GEE	0.646	0.0580	0.0549	92.3%	92.6%	94.0%	95.0%
			U-stat.			0.0550	92.4%	92.7%	93.9%	95.1%
		25	GEE	0.635	0.0841	0.0753	90.4%	91.8%	93.3%	94.0%
			U-stat.			0.0756	90.5%	91.7%	93.2%	94.2%
0.9	0.844	100	GEE	0.840	0.0226	0.0211	92.4%	93.7%	95.2%	96.0%
			U-stat.			0.0216	92.7%	94.7%	95.4%	96.4%
		50	GEE	0.836	0.0315	0.0300	93.9%	94.4%	95.5%	96.4%
			U-stat.			0.0307	94.2%	94.8%	96.1%	96.9%
		25	GEE	0.828	0.0498	0.0419	91.2%	91.8%	93.6%	94.3%
			U-stat.			0.0428	91.6%	92.3%	93.9%	94.5%

[a] The confidence interval is calculated by $\hat{\rho}_c \pm 1.96N/(N-1) \times SE$.

[b] The confidence interval is calculated by $\hat{\rho}_c \pm 1.96N/(N-2) \times SE$.

[c] The confidence interval is calculated by $\hat{\rho}_c \pm 1.96N/(N-3) \times SE$.

Table 2
Results of the second set of simulations based on 1000 data sets

True	True	Sample				Mean	95%	Adj. 95% coverage		
ρ	ρ_o^c	size	Method	Mean	SD	est. SE	Coverage	$N/(N-1)$[a]	$N/(N-2)$[b]	$N/(N-3)$[c]
0.5	0.481	100	GEE	0.475	0.0501	0.0485	93.1%	95.4%	96.3%	96.9%
			U-stat.			0.0484	93.0%	94.5%	95.4%	96.2%
		50	GEE	0.472	0.0691	0.0669	92.5%	93.9%	94.4%	94.9%
			U-stat.			0.0669	92.5%	93.2%	94.7%	95.3%
		25	GEE	0.462	0.1003	0.0886	89.8%	90.9%	91.5%	93.0%
			U-stat.			0.0884	89.9%	91.2%	92.3%	93.4%
0.7	0.673	100	GEE	0.669	0.0375	0.0367	94.6%	95.5%	96.6%	97.0%
			U-stat.			0.0367	94.6%	95.5%	96.6%	97.0%
		50	GEE	0.663	0.0556	0.0509	91.8%	93.1%	94.4%	95.2%
			U-stat.			0.0509	91.6%	93.3%	94.4%	95.5%
		25	GEE	0.651	0.0756	0.0711	91.2%	92.3%	93.2%	93.6%
			U-stat.			0.0713	91.4%	92.2%	93.1%	93.8%
0.9	0.866	100	GEE	0.863	0.0157	0.0157	95.8%	96.4%	96.9%	97.7%
			U-stat.			0.0157	95.8%	96.5%	96.9%	97.5%
		50	GEE	0.860	0.0240	0.0222	92.7%	94.2%	95.1%	95.9%
			U-stat.			0.0222	92.7%	94.0%	94.9%	95.8%
		25	GEE	0.853	0.0369	0.0317	92.2%	95.4%	93.9%	95.4%
			U-stat.			0.0318	92.0%	93.0%	94.0%	95.3%

[a] The confidence interval is calculated by $\hat{\rho}_c \pm 1.96N/(N-1) \times SE$.
[b] The confidence interval is calculated by $\hat{\rho}_c \pm 1.96N/(N-2) \times SE$.
[c] The confidence interval is calculated by $\hat{\rho}_c \pm 1.96N/(N-3) \times SE$.

We have $\rho_o^c = (3 + 4 \times 2^{1/2})\rho/9 < \rho$. The simulation results are presented in Table 2. Again, the two approaches based on the GEE and the U-statistics yielded almost identical results. Similar to the first set of simulations, we observe that the 95% confidence intervals for the true OCCC tend to be smaller than expected. To improve the 95% coverage, we consider multiplying the standard error estimate by a factor of $N/(N-1)$, $N/(N-2)$, or $N/(N-3)$ with a small sample size (see last three columns of Tables 1 and 2). We found that the factor of $N/(N-1)$ may do better for a sample size of 100, the factor of $N/(N-2)$ works well for a sample size of 50, and the factor of $N/(N-3)$ works well for a sample size of 25.

5. Examples

Data from two biomedical studies are used to illustrate the use of the OCCC for measuring overall agreement from multiple observers. The first example is from a study in measuring blood pressure. Three readings from three observers using the mercury sphygmomanometer (MS) and one reading using the inexpensive electronic digital instruments (DI) are available for systolic or diastolic blood pressure. The original analysis for this study has been reported elsewhere (Torun et al., 1998) for pairwise CCCs. A total of 228 adult subjects were evaluated in the study and each subject had eight readings, four for systolic blood pressure (SBP) and four for diastolic blood pressure (DBP). The ranges of SBP and DBP among these subjects are 82–236 mm Hg and 50–148 mm Hg, respectively.

Six pairwise plots can be generated to examine the agreement between any two of the four readings (see Figure 1 for readings of the systolic blood pressure). These plots show that the points are clustered around the 45° line, with small variation. Thus, we expect to see high CCCs and a high OCCC. Plots of the four DBP readings show similar findings (figure not shown). We computed all possible pairwise concordance correlation coefficients and their corresponding 95% confidence intervals (CIs) using both Lin's method, the GEE method (Table 3), and the U-statistics approach. The

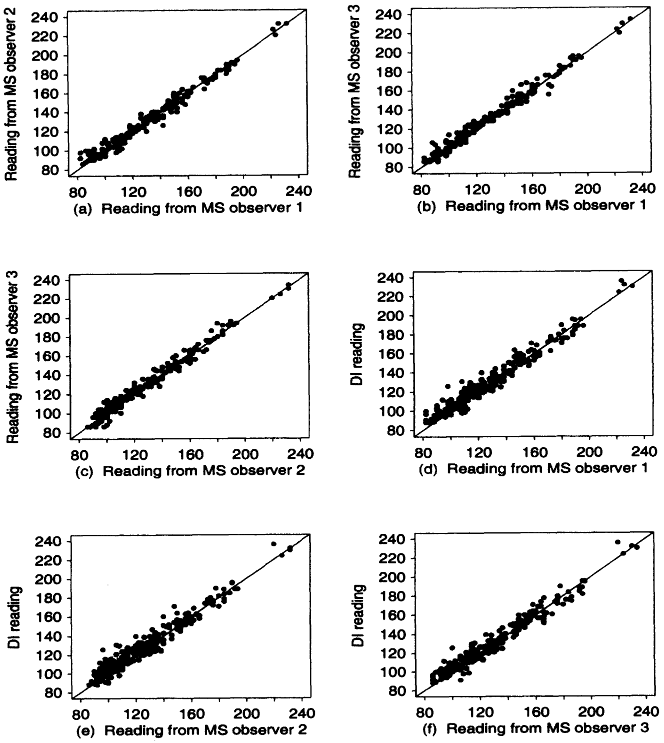

Figure 1. Systolic blood pressure data (mm Hg).

Table 3
Estimated concordance correlation coefficients from blood pressure data

Systolic Blood Pressure						
			95% CI			
Pairwise CCC	$\hat{\rho}^c_{jk}$	Lin (1989)	GEE	U-stat.	ρ	χ^a
MS observer 1 vs. observer 2	0.988	(0.984, 0.991)	(0.984, 0.991)	(0.984, 0.991)	0.988	0.999
MS observer 1 vs. observer 3	0.989	(0.986, 0.992)	(0.985, 0.992)	(0.985, 0.992)	0.990	0.999
MS observer 2 vs. observer 3	0.987	(0.983, 0.990)	(0.983, 0.990)	(0.983, 0.990)	0.988	0.999
MS observer 1 vs. DI	0.973	(0.965, 0.979)	(0.964, 0.980)	(0.964, 0.980)	0.978	0.995
MS observer 2 vs. DI	0.969	(0.959, 0.976)	(0.957, 0.977)	(0.960, 0.977)	0.972	0.996
MS observer 3 vs. DI	0.977	(0.971, 0.983)	(0.970, 0.983)	(0.971, 0.984)	0.979	0.998
Overall CCC	$\hat{\rho}^c_o$	Bootstrap	GEE	U-stat.	ρ	χ^a
Among four readings	0.981	(0.976, 0.986)	(0.976, 0.985)	(0.976, 0.986)	0.983	0.998
Diastolic Blood Pressure						
			95% CI			
Pairwise CCC	$\hat{\rho}^c_{jk}$	Lin (1989)	GEE	U-stat.	ρ	χ^a
MS observer 1 vs. observer 2	0.961	(0.949, 0.969)	(0.945, 0.972)	(0.947, 0.974)	0.965	0.995
MS observer 1 vs. observer 3	0.971	(0.962, 0.978)	(0.958, 0.980)	(0.960, 0.982)	0.976	0.994
MS observer 2 vs. observer 3	0.965	(0.955, 0.973)	(0.951, 0.975)	(0.953, 0.977)	0.967	0.998
MS observer 1 vs. DI	0.947	(0.931, 0.959)	(0.926, 0.962)	(0.929, 0.964)	0.956	0.991
MS observer 2 vs. DI	0.954	(0.940, 0.965)	(0.935, 0.967)	(0.938, 0.967)	0.957	0.997
MS observer 3 vs. DI	0.957	(0.944, 0.966)	(0.940, 0.969)	(0.942, 0.971)	0.957	0.999
Overall CCC	$\hat{\rho}^c_o$	Bootstrap	GEE	U-stat.	ρ	χ^a
Among four readings (GEE)	0.959	(0.947, 0.971)	(0.947, 0.971)	(0.947, 0.971)	0.963	0.996

bootstrap method (using 1000 bootstrap samples) based on standard deviation of the empirical distribution gave similar results (not shown) as the GEE and the U-statistics approaches for the pairwise estimates. The OCCC for measuring the agreement among the four readings is 0.981 and 0.959 for the systolic and diastolic blood pressures, respectively. The three inference approaches based on the bootstrap, the GEE, and the U-statistics produced similar 95% confidence intervals of $\hat{\rho}^c_o$, (0.976, 0.985) and (0.947, 0.971) for the systolic and diastolic blood pressures, respectively. As expected, we found that both the precision (ρ) and accuracy (χ^a) components of the pairwise CCCs and the OCCCs are very high (ranging from 0.956 to 0.999). These high agreement values of OCCC indicate that the blood pressure reading using the DI method is interchangeable with readings by the three observers using the MS method.

The second example is from a carotid stenosis screening study conducted at Emory University from 1994 to 1996. Three observers, each using three different methods (magnetic resonance angiography (MRA) two-dimensional [2D] time of flight, MRA three-dimensional [3D] time of flight, and intra-arterial angiogram [IA]) to assess the stenosis of both left and right carotid arteries. A total of 18 readings is available for 55 patients with 9 readings from the left artery and 9 readings from the right artery. We are interested in estimating both the overall agreement among the three methods and overall agreement among the three observers within each method. For the overall agreement among the three methods, we compare the three average readings (over the three observers) for the methods. The first three graphs, (a)–(c), in Figures 2 and 3 display the pairwise plots of the three average readings for left and right arteries, respectively. Pairwise plots can also be produced among three observers for each method. For illustration, we present only the pairwise plots of the three observers using the IA method (the last three graphs [(d)–(f)] in Figures 2 and 3 for left and right arteries, respectively). The results from the bootstrap, the GEE, and the U-statistics approaches are presented in Table 4. As suggested in the simulation study, we also present the adjusted 95% confidence intervals using the factor of $N/(N-2)$ due to the small sample size ($N = 55$). We find that the agreement among the three methods is slightly higher for the right artery than for the left artery (0.742 versus 0.668, not significant). The observers agree better if they used the IA method ($\hat{\rho}^c_o$=0.882 or 0.915 for left and right arteries, respectively) as compared with using the MRA 2D or the MRA 3D method ($\hat{\rho}^c_o$ is between 0.61 and 0.64). To further understand these moderate OCCC values, we computed the corresponding components of precision (ρ) and accuracy (χ^a) by assuming same pairwise correlations (the data suggest that this assumption is reasonable). We found that, for all the OCCCs, the accuracy values are high (ranging from 0.958 to 0.992). However, the precision

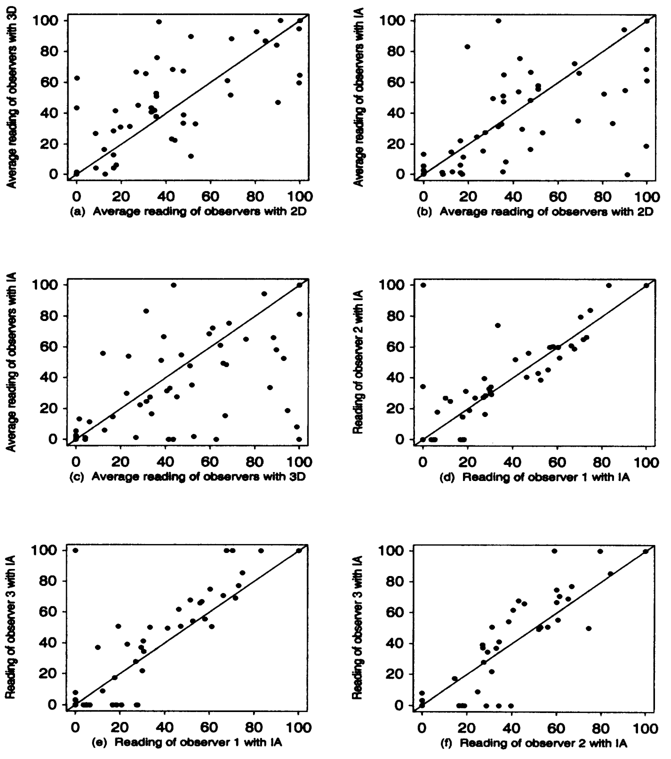

Figure 2. Left carotid artery stenosis data.

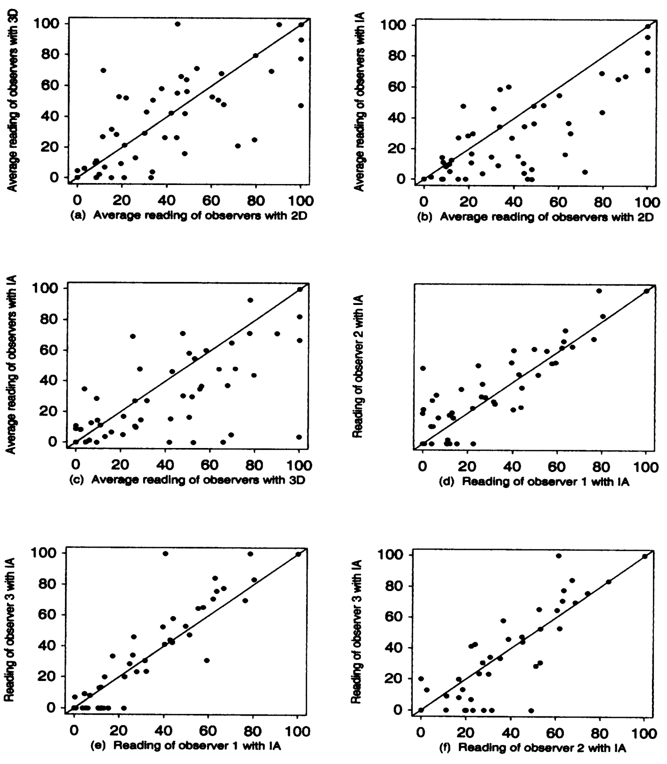

Figure 3. Right carotid artery stenosis data.

values are moderate (ranging from 0.630 to 0.683) except the ones among three raters using the IA method ($\chi^a = 0.891$ and 0.923). This finding agrees with the data presented in Figures 2 and 3, where the points are evenly scattered around the 45° line (high accuracy) but are not tightly scattered (moderate precision). We note that the bootstrap method using the standard deviation of the empirical distribution from 1000 bootstrap samples gave similar results as the the GEE

Table 4

Overall concordance correlation coefficients for the carotid stenosis screening study

		95% CI				
	$\hat{\rho}_o^c$	Bootstrap	GEE	U-stat.	ρ	χ^a
		Left Artery				
Among three methods	0.668	(0.525, 0.811)	(0.522, 0.813)	(0.524, 0.812)	0.683	0.977
(Adjusted 95% CI)[a]		(0.517, 0.818)	(0.517, 0.818)	(0.518, 0.817)		
Among three raters using MRA 2D	0.623	(0.459, 0.787)	(0.460, 0.786)	(0.461, 0.784)	0.640	0.973
(Adjusted 95% CI)		(0.451, 0.794)	(0.454, 0.792)	(0.455, 0.790)		
Among three raters using MRA 3D	0.642	(0.475, 0.809)	(0.477, 0.807)	(0.478, 0.806)	0.650	0.988
(Adjusted 95% CI)		(0.469, 0.815)	(0.471, 0.814)	(0.472, 0.812)		
Among three raters using IA	0.882	(0.782, 0.982)	(0.782, 0.982)	(0.784, 0.980)	0.891	0.989
(Adjusted 95% CI)		(0.776, 0.987)	(0.778, 0.986)	(0.780, 0.983)		
		Right Artery				
Among three methods	0.742	(0.621, 0.863)	(0.623, 0.862)	(0.624, 0.861)	0.773	0.960
(Adjusted 95% CI)		(0.617, 0.868)	(0.619, 0.866)	(0.619, 0.865)		
Among three raters using MRA 2D	0.607	(0.445, 0.769)	(0.451, 0.764)	(0.451, 0.763)	0.634	0.958
(Adjusted 95% CI)		(0.439, 0.775)	(0.445, 0.770)	(0.445, 0.769)		
Among three raters using MRA 3D	0.618	(0.462, 0.774)	(0.457, 0.779)	(0.459, 0.777)	0.630	0.981
(Adjusted 95% CI)		(0.456, 0.780)	(0.451, 0.785)	(0.453, 0.783)		
Among three raters using IA	0.915	(0.866, 0.964)	(0.865, 0.965)	(0.866, 0.964)	0.923	0.992
(Adjusted 95% CI)		(0.861, 0.969)	(0.864, 0.967)	(0.864, 0.966)		

[a] Adjusted by using a factor of $N/(N-2)$.

and the U-statistics methods in terms of 95% confidence intervals.

6. Discussion

We have evaluated the agreement index, OCCC, to measure agreement among multiple fixed observers. Our definition for this index provides two intuitive appeals. First, both precision and accuracy are components of the OCCC. Second, we can interpret the OCCC as a weighted average of the pairwise agreement indices based on the concordance correlation coefficient. A larger weight is placed on poor pairwise agreement in this overall agreement index. The estimation of the OCCC is simple because one only needs to compute the sample means, variances, and covariance. All three inference approaches based on bootstrap, the GEE, and the U-statistics can be used for inference where no distributional assumption is required.

Note that Barnhart and Williamson (2001) showed how easily one can incorporate covariates in modeling pairwise CCCs using the GEE approach. We believe that the proposed GEE approach can be modified similarly to model the covariates' impact on the OCCC. It is not obvious to us how one can incorporate covariates in the approach based on the U-statistics. This can be a potential advantage of the GEE approach over the U-statistics approach.

Acknowledgements

This research was supported by a grant from the Emory University Research Committee. We would like to thank Drs Martorell and Torun (supported by the NIH grant R01 HD 29927 and the Nestle Foundation grant 573917) for providing the blood pressure data set. We also would like to thank Dr Peterman (supported by the NIH grant R01 NS 30928) for providing the data set from the carotid screening study.

Résumé

Une mesure juste et précise est une composante importante de toute étude expérimentale. Tel qu'il fut construit par Lin (1989), le coefficient de corrélation de concordance (CCC) est plus adapté que tout autre indice pour mesurer l'agrément quand les deux variables en cause sont continues. Néanmoins, cet indice d'agrément n'est défini que pour comparer deux observateurs fixés. Pour utiliser plusieurs observateurs dans une étude avec un grand nombre de sujets, il est nécessaire d'estimer l'agrément entre de nombreux observateurs. Dans cet article, nous présentons un CCC global (OCCC) en fonction de la variabilité inter-observateurs pour l'estimation de l'agrément entre plusieurs observateurs fixés. L'OCCC se révèle être équivalent au CCC généralisé (King et Chinchilli, 2001; Lin, 1989, 2000) quand la fonction du carré de la distance est utilisée. Nous évaluons l'OCCC par des GEE équations d'estimation généralisées ou GEE (Barnhart et Williamson, 2001) et des U-statistiques (King et Chinchilli, 2001) pour l'inférence. Cet article développe les aspects importants suivants: 1) il aborde les indices de justesse et de précision comme des composantes de l'OCCC; 2) il éclaircit le fait que l'OCCC soit la moyenne pondérée de toutes les pires de CCC; 3) il est intuitivement défini en fonction de la variabilité inter-observateurs; 4) les approches inférentielles des GEE et des U-statistiques sont comparées par des simulations sur de petits échantillons; 5) l'illustration de l'emploi des OCCC est faite sur deux exemples avec les approches GEE, U-statistiques et bootstrap.

References

Barnhart, H. X. and Williamson, J. M. (2001). Modeling concordance correlation via GEE to evaluate reproducibility. *Biometrics* **57,** 931–940.

Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports* **19,** 3–11.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.

King, T. S. and Chinchilli, V. M. (2001). A generalized concordance correlation coefficient for continuous and categorical data. *Statistics in Medicine* **20,** 2131–2147.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73,** 13–22.

Lin, L. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics* **45,** 255–268.

Lin, L. (1992). Assay validation using the concordance correlation coefficient. *Biometrics* **48,** 599–604.

Lin, L. (2000). A note on the concordance correlation coefficient. *Biometrics* **56,** 324–325.

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

Torun, 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.

Zeger, S. L. and Liang, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42,** 121–130.

Received October 2001. *Revised July* 2002.
Accepted August 2002.