BIOMETRICS 57, 931–940
September 2001

Modeling Concordance Correlation via GEE to Evaluate Reproducibility

Huiman X. Barnhart

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

and

John M. Williamson

Division of HIV/AIDS Prevention: Surveillance and Epidemiology (MS E-48),
National Centers for HIV, STD, and TB Prevention, Centers for Disease Control and Prevention,
1600 Clifton Road NE, Atlanta, GA 30333, U.S.A.
email: jow5@cdc.gov

SUMMARY. Clinical studies are often concerned with assessing whether different raters/methods produce similar values for measuring a quantitative variable. Use of the concordance correlation coefficient as a measure of reproducibility has gained popularity in practice since its introduction by Lin (1989, *Biometrics* **45,** 255–268). Lin's method is applicable for studies evaluating two raters/two methods without replications. Chinchilli et al. (1996, *Biometrics* **52,** 341–353) extended Lin's approach to repeated measures designs by using a weighted concordance correlation coefficient. However, the existing methods cannot easily accommodate covariate adjustment, especially when one needs to model agreement. In this article, we propose a generalized estimating equations (GEE) approach to model the concordance correlation coefficient via three sets of estimating equations. The proposed approach is flexible in that (1) it can accommodate more than two correlated readings and test for the equality of dependent concordant correlation estimates; (2) it can incorporate covariates predictive of the marginal distribution; (3) it can be used to identify covariates predictive of concordance correlation; and (4) it requires minimal distribution assumptions. A simulation study is conducted to evaluate the asymptotic properties of the proposed approach. The method is illustrated with data from two biomedical studies.

KEY WORDS: Agreement; Correlated data; Reliability; Reproducibility.

1. Introduction

Accurate and precise measurement is an important component of any proper study design. Before a method, a rater, or an instrument is adopted for use in measuring a variable of interest, a reliability or a validity study is often conducted in clinical or experimental settings. These reliability studies are often concerned with assessing whether different methods/raters produce similar values. The kappa statistic (Cohen, 1960) and the weighted kappa statistic (Cohen, 1968) are the most popular indices for measuring agreement for discrete outcomes. Traditionally, intraclass correlation (Fleiss, 1986; Quan and Shih, 1996) and within-subject coefficient of variation (Lee, Koh, and Ong, 1989) have been used as indices to evaluate reproducibility. These indices can be estimated by using random-effects models. As elaborated by Lin (1989, 1992), the concordance correlation coefficient is more appropriate for measuring agreement when the variable of interest is continuous. The advantage of this index is that it includes components of both precision and accuracy. Unlike the concordance correlation coefficient, neither the intraclass correlation coefficient nor the within-subject coefficient of variation contain measurement of accuracy. In general, the correlation coefficient measures the precision component and the ratio of the correlation coefficient and concordance correlation coefficient measures the accuracy component. Several authors (Krippendorff, 1970; Fleiss and Cohen, 1973; Donner and Koval, 1980; Robieson, 1999) noted that the concordance correlation coefficient (or intraclass correlation coefficient in special cases) computed from ordinal scaled data is equivalent to the weighted kappa when integer scores are used.

Chinchilli et al. (1996) extended Lin's approach to repeated measures designs by using a weighted concordance correlation coefficient. However, these methods are difficult to use when one would like to model the agreement measure as a function of covariates. Barlow (1996) showed that the estimates of agreement using the kappa coefficient may be inflated if one fails to account for confounding in the marginal distribution. Similar issues may arise when one uses the concordance correlation coefficient for measuring agreement in the case of a continuous variable. Several authors (Barlow, 1996; Molenberghs, Fitzmaurice, and Lipsitz, 1996; Shoukri and Mian, 1996) have proposed methods to construct inference on kappa while adjusting for covariates through the marginal distribution. However, as noted by Klar, Lipsitz, and Ibrahim (2000) and Gonin et al. (2000), the adjustment in the marginal distribution only produces an overall/summary agreement measure. There are practical needs to compare agreement for multiple or stratified samples, and investigators may wish to "determine and evaluate the strength of agreement taking into account patient-specific (age and gender) as well as rater specific (whether board certified in dermatology) characteristics" (Gonin et al., 2000, p. 1). This suggests modeling agreement as a function of these covariates.

Often, reliability studies will be devised to compare agreement between various methods or instruments used to assess subjects. In addition, a reliability study may be used to evaluate the agreement between a new method and a widely acceptable proven method, a gold standard, which is also measured with error. In these situations, numerous assessments or ratings will be made on the same subject. These measurements will tend to be positively correlated and this correlation must be taken into account to conduct valid inference. Here we focus on the analysis of reliability studies that compare various methods or instruments with correlated continuous measurements assessed on the same subject with adjustment for covariates.

Several authors (Lipsitz and Fitzmaurice, 1996; Molenberghs et al., 1996) have used generalized estimating equations (GEE) to construct inference on an overall kappa statistic while adjusting for covariates in the marginal distribution. More recently, several approaches have been proposed using GEE equations to model kappa or weighted kappa as a function of covariates (Gonin et al., 2000; Klar et al., 2000; Williamson et al., 2000). In this article, we propose a GEE approach to model both the concordance correlation coefficient and the marginal distribution while adjusting for covariates. The proposed approach is flexible in the sense that (1) it can accommodate more than two correlated readings; (2) it can easily incorporate covariates predictive of the marginal distribution; (3) it can be used to identify covariates predictive of concordance correlation; and (4) it requires minimal distributional assumptions. In Section 2, we present the proposed GEE approach for modeling the concordance correlation coefficient. A simulation study is conducted to evaluate the asymptotic properties of the proposed approach under small and moderate sample sizes in Section 3. In Section 4, we illustrate the proposed method using two examples. The first example compares three readings using a mercury sphygmomanometer and one reading from an electronic digital instrument in measuring blood pressure. The second example compares two new methods and a gold standard method in evaluating carotid stenosis from a carotid stenosis screening study.

2. Modeling Concordance Correlation via GEE

Suppose that J readings are taken on N subjects in a study. These J readings may come from a combination of several raters and several measuring instruments. Let $\mathbf{Y}_i$ ($i = 1, \ldots, N$) be the $J \times 1$ vector that contains the J readings and let the $J \times p$ matrix $\mathbf{X}_i$ denote the corresponding p covariates for the ith subject. For simplicity, the first column of $\mathbf{X}_i$ is a vector of all ones representing an intercept term. The covariates can either be subject-specific and/or reading-specific variables. Examples of subject-specific covariates are age, gender, and race. The indicator variables for raters or instruments are examples of reading-specific covariates.

For any two readings Y_j and $Y_{j'}$ ($1 \le j, j' \le J, j \ne j'$), Lin (1989) used the expected squared difference, $\mathrm{E}[(Y_j - Y_{j'})^2]$, scaled between -1 and 1 to define the concordance correlation coefficient as

$$\rho_{cjj'} = 1 - \frac{\mathrm{E}\left[(Y_j - Y_{j'})^2\right]}{\sigma_j^2 + \sigma_{j'}^2 + (\mu_j - \mu_{j'})^2} = \frac{2\sigma_{jj'}}{\sigma_j^2 + \sigma_{j'}^2 + (\mu_j - \mu_{j'})^2},$$

where $\mu_j = \mathrm{E}(Y_j)$, $\mu_{j'} = \mathrm{E}(Y_{j'})$, $\sigma_j^2 = \mathrm{var}(Y_j)$, $\sigma_{j'}^2 = \mathrm{var}(Y_{j'})$, $\sigma_{jj'} = \mathrm{cov}(Y_j, Y_{j'}) = \sigma_j\sigma_{j'}\rho_{jj'}$, and $\rho_{jj'}$ is the usual correlation between Y_j and $Y_{j'}$. Note that $\rho_{jj'} = \rho_{cjj'}$ if and only if $\mu_j = \mu_{j'}$ and $\sigma_j^2 = \sigma_{j'}^2$. In general, we have $-1 \le -|\rho_{jj'}| \le \rho_{cjj'} \le |\rho_{jj'}| \le 1$ (Lin, 1989).

In order to model concordance correlation, one needs to estimate the means and the variances of Y_j and $Y_{j'}$. Three sets of estimating equations are proposed. Let $\boldsymbol{\beta}$ be a $p \times 1$ marginal parameter vector. In the first set of equations, we model the marginal mean of $\mathbf{Y}$ by $\mathrm{E}(\mathbf{Y}_i) = \boldsymbol{\mu}_i = \mathbf{X}_i\boldsymbol{\beta}$, and the parameter estimates of $\boldsymbol{\beta}$ are obtained by GEE as

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

where $\mathbf{D}_i = \partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta}$ and $\mathbf{V}_i$ is the working covariance matrix for $\mathbf{Y}_i$ (Zeger and Liang, 1986). In the second set of equations, we estimate the variance of $\mathbf{Y}$. Note that the variances are smaller when one accounts for covariates in the above marginal model. For simplicity, we assume that the variance does not depend on covariates. Let $\mathbf{Y}_i^2 = (Y_{i1}^2, Y_{i2}^2, \ldots, Y_{iJ}^2)'$ and $\boldsymbol{\delta}_i^2 = \mathrm{E}(\mathbf{Y}_i^2)$. Then $\boldsymbol{\delta}_i^2 = \boldsymbol{\sigma}^2 + \boldsymbol{\mu}_i^2$, where $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_J^2)'$ with $\sigma_j^2 = \mathrm{var}(Y_{ij})$ and $\boldsymbol{\mu}_i^2 = (\mu_{i1}^2, \ldots, \mu_{iJ}^2)'$. We solve the following second set of estimating equations to obtain estimates for $\boldsymbol{\sigma}^2$:

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

where $\mathbf{F}_i = \partial\boldsymbol{\delta}_i^2/\partial\boldsymbol{\sigma}^2$ and $\mathbf{H}_i$ is the working covariance matrix for $\mathbf{Y_i}^2$. In practice, we can use a diagonal matrix for $\mathbf{H}_i$ with the jth component as $\mathrm{var}(Y_{ij}^2)$. If $\mathbf{Y}_i$ is normally distributed, then we have $\mathrm{var}(Y_{ij}^2) = 2\sigma_j^4 + 4\mu_{ij}^2\sigma_j^2$. We use this expression as the diagonal component of $\mathbf{H}_i$ even if $\mathbf{Y}_i$ is not normally distributed.

Note that $\mathrm{E}(Y_jY_{j'}) = \mathrm{cov}(Y_j, Y_{j'}) + \mu_j\mu_{j'} = \rho_{cjj'}(\sigma_j^2 + \sigma_{j'}^2 + (\mu_j - \mu_{j'})^2)/2 + \mu_j\mu_{j'}$. This implies that the concordance correlation can be obtained once the means and the variances are estimated. Therefore, we use the $J(J-1)/2$ pairwise products $Y_jY_{j'}$ to model the concordance correlation in a third set of estimating equations. Let $\mathbf{U}_i = (Y_{i1}Y_{i2}, Y_{i1}Y_{i3}, \ldots, Y_{i(J-1)}Y_{iJ})'$ be the $J(J-1)/2 \times 1$ vector of the pairwise products of readings for subject i and denote $\boldsymbol{\theta}_i = \mathrm{E}(\mathbf{U}_i)$. Then $\boldsymbol{\theta}_i$ is a function of $\boldsymbol{\sigma}^2$, $\boldsymbol{\mu}_i$, and the concordance correlation coefficients (see the expression for $\mathrm{E}(Y_j Y_{j'})$). We use Fisher's Z-transformation to model the concordance correlation coefficient because it ranges between -1 and 1 as

$$\frac{1}{2}\log\frac{1+\rho_{cijj'}}{1-\rho_{cijj'}} = \mathbf{Q}_{ijj'}\boldsymbol{\alpha},$$

where $\rho_{cijj'}$ is the concordance correlation between the jth and the j'th readings for subject i and $\mathbf{Q}_{ijj'}$ includes a subset of covariates in $\mathbf{X}_i$ as well as indicator variables for pairs (j, j'). For example, if we believe that the concordance correlations among the first three readings are similar, we may pool the information together by using an indicator variable taking value one if the (j, j')th pair is equal to (1,2), (1,3), or (2,3) and taking the value zero otherwise. We solve for $\boldsymbol{\alpha}$ using the following third set of estimating equations:

$$\sum_{i=1}^{N} \mathbf{C}_i'\mathbf{W}_i^{-1}\left(\mathbf{U}_i - \boldsymbol{\theta}_i\left(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2\right)\right) = \mathbf{0}, \qquad (3)$$

where $\mathbf{C}_i = \partial\boldsymbol{\theta}_i/\partial\boldsymbol{\alpha}$ and $\mathbf{W}_i$ is the working covariance matrix for $\mathbf{U}_i$. In practice, we use the diagonal matrix for $\mathbf{W}_i$ with the (j, j')th component as $\mathrm{var}(Y_{ij}Y_{ij'})$. If $\mathbf{Y}_i$ is distributed as a multivariate normal random variable, then $\mathrm{var}(Y_{ij}Y_{ij'}) = \sigma_j^2\sigma_{j'}^2 + \mu_{ij}^2\sigma_{j'}^2 + \mu_{ij'}^2\sigma_j^2 + \sigma_{jj'}^2 + 2\mu_{ij}\mu_{ij'}\sigma_{jj'}$, where $\sigma_{jj'} = (\rho_{cijj'}(\sigma_j^2+\sigma_{j'}^2+(\mu_{ij}-\mu_{ij'})^2))/2$. We use this expression for a diagonal element in $\mathbf{W}_i$ even if $\mathbf{Y}_i$ is not normally distributed.

Our main interest here is to obtain the parameter estimates of $\boldsymbol{\alpha}$ and then the estimates for the concordance correlation coefficients. When there are no covariates, the above estimation process requires no iterations and the resulting concordance correlation estimate is exactly equivalent to Lin's coefficient. We obtain $\hat{\boldsymbol{\alpha}}$ by solving three sets of estimating equations. Specifically, we obtain parameter estimates for $\boldsymbol{\beta}$ ($\hat{\boldsymbol{\beta}}$) by using a modified Fisher-scoring iterative procedure, $\hat{\boldsymbol{\beta}}_{l+1} = \hat{\boldsymbol{\beta}}_l - \{\Sigma_{i=1}^N \hat{\mathbf{D}}_i'\hat{\mathbf{V}}_i^{-1}\hat{\mathbf{D}}_i\}^{-1}\{\Sigma_{i=1}^N \hat{\mathbf{D}}_i'\hat{\mathbf{V}}_i^{-1}(\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i(\hat{\boldsymbol{\beta}}_l)\}$, for the first set of equations. Thus, we have estimates for $\boldsymbol{\mu}_i$, $\hat{\boldsymbol{\mu}}_i$. Replacing $\boldsymbol{\mu}_i$ with $\hat{\boldsymbol{\mu}}_i$ in the second set of equations, we obtain parameter estimates for $\boldsymbol{\sigma}^2$ ($\hat{\boldsymbol{\sigma}^2}$). Finally, we replace $\boldsymbol{\mu}_i$ and $\boldsymbol{\sigma}^2$ with $\hat{\boldsymbol{\mu}}_i$ and $\hat{\boldsymbol{\sigma}}^2$ in the third set of equations and obtain the parameter estimates for $\boldsymbol{\alpha}$, $\hat{\boldsymbol{\alpha}}$. We again used the modified Fisher-scoring iterative procedure for solving the second and third equations. Following similar arguments used for the usual generalized estimating equations (Liang, Zeger, and Qaqish, 1992), we can easily show that the parameter estimates are consistent provided that the three models are correctly specified. This is true whether or not the working correlation matrices in the three sets of equations are correctly specified.

In order to perform inference on the concordance correlations, one needs to obtain estimates for the standard error of $\hat{\boldsymbol{\alpha}}$. Note that $\hat{\boldsymbol{\alpha}}$ is obtained in the third set of equations with estimated $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$. Therefore, we will need to incorporate the uncertainty in estimating $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$ in order to obtain the estimates for the standard error of $\hat{\boldsymbol{\alpha}}$. Following Prentice (1988), we can show that the joint asymptotic distribution of $N^{1/2}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})', (\hat{\boldsymbol{\sigma}}^2 - \boldsymbol{\sigma}^2)', (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})']'$ is Gaussian with mean zero and variance matrix as N times

$$\mathbf{B} = \boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Psi}'^{-1} = \boldsymbol{\Psi}^{-1}\begin{pmatrix} \Lambda_{11} & \Lambda_{12} & \Lambda_{13} \\ \Lambda_{21} & \Lambda_{22} & \Lambda_{23} \\ \Lambda_{31} & \Lambda_{32} & \Lambda_{33} \end{pmatrix}\boldsymbol{\Psi}'^{-1},$$

where

$$\boldsymbol{\Psi} = \begin{pmatrix} \sum_{i=1}^{N}\mathbf{D}_i'\mathbf{V}_i^{-1}\mathbf{D}_i & 0 & 0 \\ \sum_{i=1}^{N}\mathbf{F}_i'\mathbf{H}_i^{-1}\mathbf{L}_i & \sum_{i=1}^{N}\mathbf{F}_i'\mathbf{H}_i^{-1}\mathbf{F}_i & 0 \\ \sum_{i=1}^{N}\mathbf{C}_i'\mathbf{W}_i^{-1}\mathbf{E}_i & \sum_{i=1}^{N}\mathbf{C}_i'\mathbf{W}_i^{-1}\mathbf{G}_i & \sum_{i=1}^{N}\mathbf{C}_i'\mathbf{W}_i^{-1}\mathbf{C}_i \end{pmatrix}$$

with $\mathbf{L}_i = \partial\boldsymbol{\delta}_i^2/\partial\boldsymbol{\beta}, \mathbf{E}_i = \partial\boldsymbol{\theta}_i/\partial\boldsymbol{\beta}, \mathbf{G}_i = \partial\boldsymbol{\theta}_i/\partial\boldsymbol{\sigma}^2$, and

$$\Lambda_{11} = \sum_{i=1}^{N}\mathbf{D}_i'\mathbf{V}_i^{-1}\mathrm{var}(\mathbf{Y}_i)\mathbf{V}_i^{-1}\mathbf{D}_i,$$
$$\Lambda_{12} = \sum_{i=1}^{N}\mathbf{D}_i'\mathbf{V}_i^{-1}\mathrm{cov}(\mathbf{Y}_i, \mathbf{Y}_i^2)\mathbf{H}_i^{-1}\mathbf{F}_i,$$
$$\Lambda_{13} = \sum_{i=1}^{N}\mathbf{D}_i'\mathbf{V}_i^{-1}\mathrm{cov}(\mathbf{Y}_i, \mathbf{U}_i)\mathbf{W}_i^{-1}\mathbf{C}_i,$$
$$\Lambda_{22} = \sum_{i=1}^{N}\mathbf{F}_i'\mathbf{H}_i^{-1}\mathrm{var}(\mathbf{Y}_i^2)\mathbf{H}_i^{-1}\mathbf{F}_i,$$
$$\Lambda_{23} = \sum_{i=1}^{N}\mathbf{F}_i'\mathbf{H}_i^{-1}\mathrm{cov}(\mathbf{Y}_i^2, \mathbf{U}_i)\mathbf{W}_i^{-1}\mathbf{C}_i,$$
$$\Lambda_{33} = \sum_{i=1}^{N}\mathbf{C}_i'\mathbf{W}_i^{-1}\mathrm{var}(\mathbf{U}_i)\mathbf{W}_i^{-1}\mathbf{C}_i,$$
$$\Lambda_{21} = \Lambda_{12}',$$
$$\Lambda_{31} = \Lambda_{13}',$$
$$\Lambda_{32} = \Lambda_{23}'.$$

We obtain an estimate for $\mathbf{B}$ by replacing the parameters by their corresponding estimates that are the solutions to the estimating equations and estimate the covariances in $\boldsymbol{\Lambda}$ by the consistent moment estimates,

$$\mathrm{var}(\mathbf{Y}_i) = (\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i)(\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i)',$$
$$\mathrm{cov}(\mathbf{Y}_i, \mathbf{Y}_i^2) = (\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i)(\mathbf{Y}_i^2 - \hat{\boldsymbol{\delta}}_i)',$$
$$\mathrm{cov}(\mathbf{Y}_i, \mathbf{U}_i) = (\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i)(\mathbf{U}_i - \hat{\boldsymbol{\theta}}_i)',$$
$$\mathrm{var}(\mathbf{Y}_i^2) = (\mathbf{Y}_i^2 - \hat{\boldsymbol{\delta}}_i^2)(\mathbf{Y}_i^2 - \hat{\boldsymbol{\delta}}_i^2)',$$
$$\mathrm{cov}(\mathbf{Y}_i^2, \mathbf{U}_i) = (\mathbf{Y}_i^2 - \hat{\boldsymbol{\delta}}_i^2)(\mathbf{U}_i - \hat{\boldsymbol{\theta}}_i)',$$
$$\mathrm{var}(\mathbf{U}_i) = (\mathbf{U}_i - \hat{\boldsymbol{\theta}}_i)(\mathbf{U}_i - \hat{\boldsymbol{\theta}}_i)'.$$

We refer to $\hat{\mathbf{B}}$ as the empirically corrected variance estimate of $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\delta}}^2$, and $\hat{\boldsymbol{\theta}}$. The estimate for the variance of $\hat{\boldsymbol{\alpha}}$ is obtained by taking the appropriate rows and columns in $\hat{\mathbf{B}}$. Finally, once we have $\hat{\boldsymbol{\alpha}}$ and the corresponding 95% confidence intervals, the inverse of Fisher's Z-transformation is used to obtained the estimates for the concordance correlation coefficients and their corresponding confidence intervals. We note that one

Table 1
Results of first set of simulations based on 500 data sets[a]

True value		ρ_{c1}[b]		ρ_{c2}[b]		ρ_{c12}[b]		Type I error rate (%)
ρ	ρ_{c12}	Mean (SD)	Mean of est. SE (95% cover.)	Mean (SD)	Mean of est. SE (95% cover.)	Mean (SD)	Mean of est. SE (95% cover.)	
				$N = 26$				
0.5	0.471	0.434 (0.182)	0.157 (90%)	0.460 (0.153)	0.138 (92%)	0.428 (0.134)	0.117 (92%)	9.2
0.7	0.660	0.644 (0.145)	0.124 (93%)	0.669 (0.119)	0.105 (92%)	0.618 (0.112)	0.099 (91%)	9.6
0.9[c]	0.848	0.872 (0.099)	0.081 (90%)	0.882 (0.057)	0.048 (93%)	0.824 (0.072)	0.059 (91%)	7.6
				$N = 50$				
0.5	0.471	0.461 (0.109)	0.105 (93%)	0.476 (0.106)	0.102 (94%)	0.441 (0.092)	0.084 (91%)	5.2
0.7	0.660	0.675 (0.082)	0.078 (94%)	0.682 (0.080)	0.073 (92%)	0.640 (0.070)	0.066 (92%)	6.8
0.9[d]	0.848	0.895 (0.049)	0.044 (95%)	0.893 (0.033)	0.030 (95%)	0.840 (0.039)	0.035 (94%)	4.0
				$N = 100$				
0.5	0.471	0.481 (0.078)	0.073 (92%)	0.486 (0.074)	0.072 (94%)	0.455 (0.062)	0.058 (92%)	7.0
0.7	0.660	0.690 (0.056)	0.053 (93%)	0.696 (0.053)	0.051 (94%)	0.654 (0.047)	0.045 (94%)	5.6
0.9	0.848	0.897 (0.029)	0.028 (96%)	0.897 (0.021)	0.021 (97%)	0.845 (0.024)	0.023 (95%)	4.0

[a] Data are generated from a six-dimensional multivariate normal distribution with equal means, equal correlations ($\rho = \rho_{c1} = \rho_{c2}$), and unequal variances (1.0 and 2.0) within methods 1 and 2, respectively.
[b] ρ_{c1}, ρ_{c2}, and ρ_{c12} are concordance correlation coefficients within methods 1 and 2 and between methods 1 and 2, respectively.
[c] Convergence occurs in 486 data sets.
[d] Convergence occurs in 499 data sets.

may obtain the same estimates of concordance correlation coefficients as Lin's method when there are no covariates. However, our empirically corrected standard errors may be larger than Lin's because we don't make normality assumptions concerning $\text{var}(Y_{ij}^2)$ and $\text{var}(Y_{ij}Y_{ij'})$ when computing the proposed standard errors.

Noting that $\text{E}(Y_jY_{j'}) = \rho_{jj'}\sigma_j\sigma_{j'} + \mu_j\mu_{j'}$, it requires minimal effort to model the correlation coefficient as a function of covariates by the same three sets of estimating equations. The estimates of the correlation coefficient can provide insight on whether the poor agreement is due to precision or due to accuracy because the correlation coefficient (ρ) measures the precision and $\chi_a = \rho_c/\rho$ measures the accuracy.

3. Simulations

We conducted analyses using simulated data to assess the performance of our method. We examined the bias of the concordance correlation estimates and determined how well the empirically corrected standard error estimate performs with small to moderate sample sizes with two sets of simulations. We assumed that each subject was assessed by three raters with two methods for both sets of simulations. Let $\mathbf{Y}_i$ ($i = 1, \ldots, N$) be the 6×1 vector denoting the six readings on subject i from the three raters using methods 1 and 2 (where the first three readings are from method 1). We generated continuous responses, $\mathbf{Y}_i$, from a multivariate normal distribution.

For the first set of simulations, we assumed a common mean for the normal random variables, generating each reading with differing variances for each of the two assessment methods, i.e., $\mathbf{Y}_i \sim \text{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ with $\mu_{ij} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$ for $j = 1, \ldots, 6$ and $i = 1, \ldots, N$. The first covariate, x_{1i}, is a binary subject-specific variable taking a value of zero for $i = 1, \ldots, N/2$ and one otherwise. The second covariate, x_{2i}, is a continuous subject-specific variable generated as a $\text{U}(-1, 1)$ random variable. The values of the parameters generating the data are $\beta_0 = 0.0$ and $\beta_1 = \beta_2 = 1.0$. We assume that the variances of the readings are 1.0 and 2.0 for methods 1 and 2, respectively. The correlation coefficients in the covariance matrix $\boldsymbol{\Sigma}$ are assumed to be the same and equal to ρ. Thus, the concordance correlation coefficients within each method are the same and equal to the correlation coefficient ρ. Let ρ_{c1} and ρ_{c2} be the concordance correlation coefficients for methods 1 and 2, respectively. Then we have $\rho_{c1} = \rho_{c2} = \rho$. However, the concordance correlation coefficient (ρ_{c12}) measuring the agreement between the two methods is smaller than the correlation coefficient because we assumed differing variances for the two methods. We analyzed 500 data sets of size 26, 50, and 100 (N) for the following values of the common correlation coefficient: 0.5, 0.7, and 0.9. A test of H_0: $\rho_{c1} = \rho_{c2}$ is performed to check the Type I error rate.

Table 1 summarizes the first set of simulations. The estimate of the concordance correlation coefficient is biased downward with the bias decreasing as sample size increases. The average of the 500 empirically corrected standard error estimates of $\hat{\rho}_c$ (computed as $((\partial\rho_c/\partial\hat{\alpha})^2\hat{\text{var}}(\hat{\alpha}))^{1/2}$ using the delta method) tends to be smaller than the empirical standard deviation of $\hat{\rho}_c$ (calculated from the 500 $\hat{\rho}_c$ values) for small sample sizes ($N = 26, 50$). This results in smaller 95% coverage for ρ_c. The estimated Type I error rate for comparing the equality of the concordance correlation coefficients for the two methods is near 5%, although the test may be somewhat liberal for the smallest sample size ($N = 26$).

In a second set of simulations, we examine estimation of the concordance correlation coefficient for the situation when the marginal distribution differs for the two methods but the

Table 2
Results from the second set of simulations based on 500 *data sets*[a]

True value		ρ_{c1}[b]		ρ_{c2}[b]		ρ_{c12}[b]		
ρ	ρ_{c12}	Mean (SD)	Mean of est. SE (95% cover.)	Mean (SD)	Mean of est. SE (95% cover.)	Mean (SD)	Mean of est. SE (95% cover.)	Type I error rate (%)
				N = 26, β_3 = 0.25				
0.5	0.485	0.429(0.185)	0.156 (90%)	0.434 (0.168)	0.158 (93%)	0.428 (0.141)	0.126 (91%)	8.4
0.7	0.679	0.651 (0.150)	0.119 (91%)	0.642 (0.162)	0.130 (91%)	0.630 (0.131)	0.109 (90%)	9.6
0.9[c]	0.873	0.874 (0.102)	0.069 (89%)	0.877 (0.096)	0.070 (91%)	0.842 (0.102)	0.069 (89%)	5.7
				N = 26, β_3 = 0.5				
0.5	0.444	0.428 (0.192)	0.159 (90%)	0.433 (0.197)	0.169 (89%)	0.388 (0.153)	0.125 (87%)	7.8
0.7[c]	0.622	0.639 (0.144)	0.123 (91%)	0.628 (0.170)	0.142 (91%)	0.562 (0.139)	0.116 (90%)	6.0
0.9[c]	0.800	0.876 (0.078)	0.071 (90%)	0.872 (0.099)	0.077 (88%)	0.754 (0.123)	0.086 (89%)	5.5
				N = 50, β_3 = 0.25				
0.5	0.485	0.468 (0.117)	0.107 (93%)	0.475 (0.119)	0.109 (92%)	0.460 (0.094)	0.088 (93%)	10.0
0.7	0.679	0.683 (0.078)	0.075 (93%)	0.681 (0.086)	0.077 (93%)	0.661 (0.073)	0.068 (93%)	6.4
0.9[c]	0.873	0.895 (0.040)	0.037 (92%)	0.895 (0.043)	0.039 (94%)	0.862 (0.058)	0.039 (93%)	2.6
				N = 50, β_3 = 0.5				
0.5	0.444	0.468 (0.114)	0.105 (92%)	0.471 (0.133)	0.112 (91%)	0.419 (0.097)	0.087 (92%)	8.8
0.7	0.622	0.680 (0.086)	0.077 (94%)	0.674 (0.092)	0.084 (94%)	0.598 (0.084)	0.078 (92%)	8.0
0.9[c]	0.800	0.893 (0.041)	0.037 (93%)	0.891 (0.046)	0.041 (90%)	0.782 (0.059)	0.054 (93%)	4.1
				N = 100, β_3 = 0.25				
0.5	0.485	0.489 (0.077)	0.073 (92%)	0.486 (0.077)	0.075 (94%)	0.473 (0.063)	0.061 (94%)	4.4
0.7	0.679	0.685 (0.058)	0.053 (92%)	0.687 (0.058)	0.054 (93%)	0.666 (0.051)	0.048 (94%)	5.6
0.9	0.873	0.896 (0.024)	0.024 (94%)	0.897 (0.027)	0.025 (93%)	0.867 (0.027)	0.026 (94%)	5.6
				N = 100, β_3 = 0.5				
0.5	0.444	0.487 (0.072)	0.073 (95%)	0.484 (0.082)	0.079 (93%)	0.433 (0.063)	0.063 (94%)	8.4
0.7	0.622	0.691 (0.055)	0.052 (94%)	0.693 (0.058)	0.055 (96%)	0.614 (0.057)	0.054 (94%)	6.4
0.9	0.800	0.896 (0.025)	0.024 (93%)	0.898 (0.028)	0.026 (95%)	0.790 (0.039)	0.036 (93%)	5.0

[a] Data are generated from a six-dimensional multivariate normal distribution with unequal means for methods 1 and 2, equal correlations ($\rho = \rho_{c1} = \rho_{c2}$), and equal variances (= 1.0).
[b] ρ_{c1}, ρ_{c2}, and ρ_{c12} are concordance correlation coefficients within methods 1 and 2 and between methods 1 and 2, respectively.
[c] Convergence occurs in 475, 498, 496, or 493 data sets.

variance of the ratings is constant. We generated multivariate normally distributed ratings as follows: $\mathbf{Y}_i \sim \mathrm{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ with $\mu_{ij} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3ij}$ for $j = 1, \ldots, 6$ and $i = 1, \ldots, N$. The covariates x_{1i} and x_{2i} and the parameters β_0, β_1, and β_2 are the same as in the first set of simulations. The third covariate, x_{3ij}, denotes the assessment method and takes a value of zero for $j = 1, 2, 3$ (first method) and a value of two for $j = 4, 5, 6$ (second method). We generated data with $\beta_3 = 0.25$ and 0.5, thereby assuming that the raters tend to assess subjects higher with the second method than with the first method. We assume that $\mathrm{var}(Y_{ij}) = 1.0$ and $\mathrm{corr}(Y_{ij}, Y_{ij'}) = \rho$ for $i = 1, \ldots, N$ and $j, j' = 1, \ldots, 6$ with $j \neq j'$. Again, the concordance correlation coefficients (ρ_{c1}, ρ_{c2}) for each method are the same and are equal to ρ. However, the concordance correlation coefficient between the two methods (ρ_{c12}) is smaller than ρ because the marginal means are different for the two methods. We analyzed 500 data sets for each value of $\rho = 0.5, 0.7, 0.9$; $\beta_3 = 0.25, 0.5$; and sample size $N = 26, 50$, and 100. We also test H_0: $\rho_{c1} = \rho_{c2}$ to assess the Type I error rate.

Table 2 summarizes the second set of simulations. Differing marginal distributions for the readings on an individual (different β_3 values) do not appreciably affect the bias of the concordance correlation coefficients $(\rho_{c1}, \rho_{c2}, \rho_{c12})$. The mean of the 500 empirically corrected standard error estimates performs fairly well in comparison with the empirical standard deviation of $\hat{\rho}_c$ (calculated from the 500 ρ_c values) for moderate sample size but tends to be smaller for small sample sizes. This results in smaller 95% coverage for ρ_c. The Type I error rate tends to be somewhat liberal for $\rho = 0.5$ and somewhat conservative for larger values (e.g., $\rho = 0.9$).

In summary, the proposed GEE method for estimating concordance correlation coefficient tends to underestimate the true concordance correlation coefficient in small samples, although this bias is appreciably smaller in larger samples. The coverage is good for the largest sample size ($N = 100$) but is smaller than 95% for smaller sample sizes. This may be due to the fact that the empirically corrected standard error is smaller than the standard deviation from the empirical distribution. One may consider multiplying the variance esti-

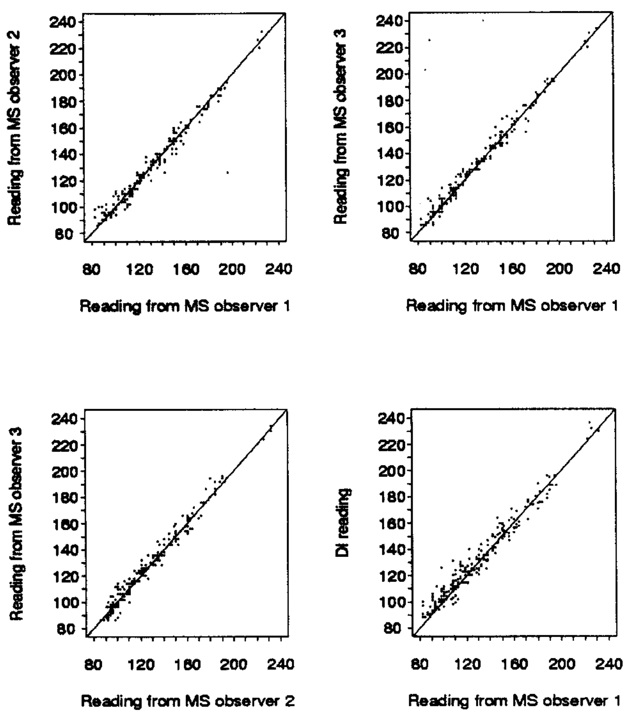

Figure 1. Systolic blood pressure data (mm Hg).

mate errors by a factor of $N/(N-1)$ or even $N/(N-2)$ with a small sample size.

4. Examples

We illustrate the proposed method by analyzing data from two biomedical studies. The first example is from a study evaluating inexpensive electronic digital instruments (DI) for measuring blood pressure in field study settings. It is easier for field personnel to use an electronic digital instrument than a mercury sphygmomanometer (MS) in measuring adult blood pressure. The original analysis for this study has been reported elsewhere using Lin's method (Torun et al., 1998). We use this data set for illustrative purpose only. In this study, systolic and diastolic blood pressures were measured by three observers using the MS method as well as by the DI method for each subject. No covariates are available in this data set. A total of 228 adult subjects had eight readings, four for systolic blood pressure (SBP) and four for diastolic blood pressure (DBP). The ranges of SBP and DBP among the 228 subjects are 82–236 and 50–148 mm Hg, respectively.

We plotted the SBP readings from MS observer i against the SBP readings from MS observer j to examine agreement within MS observers (Figure 1). We also plotted the SBP readings from MS observer 1 against the SBP readings using the DI method to explore the agreement between the MS method and the DI method (Figure 1). These plots show that the points are clustered around the 45° line with small variation. Thus, we expect to see high precision and accuracy from these data. Plots of the DBP readings show similar findings (figure not shown). We computed all possible pairwise concordance correlation coefficients and their corresponding 95% confidence intervals (CI) using both Lin's method and the proposed method (Table 3).

We used indicator variables for the three observers as covariates in the marginal model for the proposed method (the DI method is treated as the reference group here). Using $j = 1, 2, 3, 4$ to index the three MS readings from the three observers and one reading from the DI method, we model the concordance correlation coefficients as follows:

$$\frac{1}{2}\log\frac{1+\rho_{cijj'}}{1-\rho_{cijj'}} = \alpha_1 Z_{12} + \alpha_2 Z_{13} + \alpha_3 Z_{23} + \alpha_4 Z_{14} + \alpha_5 Z_{24} + \alpha_6 Z_{34},$$

where $Z_{kl}(1 \leq k, l \leq 4, k \neq l)$ are the indicator variables for pair (j, j'), i.e., $Z_{kl} = 1$ if $(k, l) = (j, j')$ and $Z_{kl} = 0$ otherwise. The above models are used to analyze the SBP and the DBP data separately. Once $\hat{\alpha}$ is obtained from the third set of estimating equations, the estimated pairwise concordance correlation coefficients as well as their 95% CIs are computed using the inverse of Fisher's Z-transformation. Note that the parameter estimates from the proposed method are identical to the ones obtained using Lin's method. The 95% confidence intervals using the proposed method are slightly wider than the corresponding ones using Lin's method. This is because we use the empirically corrected standard error estimates, which do not require the normality assumption.

The estimates for the three concordance correlation coefficients within observers are similar (ranging from 0.987 to 0.989 for the SBP readings and from 0.961 to 0.971 for the DBP readings) and the estimates for the three concordance correlation coefficients between the MS and DI methods are also similar (ranging from 0.969 to 0.977 for the SBP readings and from 0.954 to 0.957 for the DBP readings). Therefore, it is useful to summarize the reproducibility by two numbers, one of which is the within-observer reproducibility and the other is the between-method reproducibility. We simplify the above model for the concordance correlation coefficients by using two parameters,

$$\frac{1}{2}\log\frac{1+\rho_{cijj'}}{1-\rho_{cijj'}} = \alpha_1 Z_1 + \alpha_2 Z_2,$$

where $Z_1 = 1$ if (j, j') equals $(1,2), (1,3)$, or $(2,3)$ and zero otherwise and $Z_2 = 1$ if (j, j') equals $(1,4), (2,4)$, or $(3,4)$ and zero otherwise. The results of this simpler model are presented under pooled estimates in Table 3. The results indicate that the DI method has excellent reproducibility compared with the MS method (the concordance correlation coefficients are 0.973 with 95% CI 0.964–0.980 for SBP and 0.951 with 95% CI 0.934–0.964 for DBP) and that it is highly reproducible for observers using the MS method (the concordance correlation coefficients are 0.988 with 95% CI 0.984–0.991 for SBP and 0.965 with 95% CI 0.934–0.964 for DBP). These findings are similar to the ones reported by Torun et al. (1998).

We fit the same model using the correlation coefficient as measure of association to examine the components of precision and accuracy. For the SBP readings, we found that the correlation coefficients are 0.988 and 0.973, respectively, for the readings within MS observers and the readings between the MS and DI methods. This implies that the accuracy estimates are 1.0 and 0.996, respectively. For the DBP readings, we found that the correlation coefficients are 0.969 and 0.955, respectively, for the readings within MS observers and the readings between the MS and DI methods. This produces an

Table 3
Estimated concordance correlation coefficients from blood pressure data

	Lin's method		Proposed method	
	$\hat{\rho}_c$	95% CI	$\hat{\rho}_c$	95% CI
Systolic Blood Pressure				
Pairwise estimates				
MS observer 1 vs. observer 2	0.988	(0.984, 0.991)	0.988	(0.984, 0.991)
MS observer 1 vs. observer 3	0.989	(0.986, 0.992)	0.989	(0.985, 0.992)
MS observer 2 vs. observer 3	0.987	(0.983, 0.990)	0.987	(0.983, 0.990)
MS observer 1 vs. DI	0.973	(0.965, 0.979)	0.973	(0.964, 0.980)
MS observer 2 vs. DI	0.969	(0.959, 0.976)	0.969	(0.957, 0.977)
MS observer 3 vs. DI	0.977	(0.971, 0.983)	0.977	(0.970, 0.983)
Pooled estimates				
Within MS observers	—	—	0.988	(0.984, 0.991)
Between MS and DI	—	—	0.973	(0.964, 0.980)
Diastolic Blood Pressure				
Pairwise estimates				
MS observer 1 vs. observer 2	0.961	(0.949, 0.969)	0.961	(0.945, 0.972)
MS observer 1 vs. observer 3	0.971	(0.962, 0.978)	0.971	(0.958, 0.980)
MS observer 2 vs. observer 3	0.965	(0.955, 0.973)	0.965	(0.951, 0.975)
MS observer 1 vs. DI	0.947	(0.931, 0.959)	0.947	(0.926, 0.962)
MS observer 2 vs. DI	0.954	(0.940, 0.965)	0.954	(0.935, 0.967)
MS observer 3 vs. DI	0.957	(0.944, 0.966)	0.957	(0.940, 0.969)
Pooled estimates				
Within MS observers	—	—	0.965	(0.953, 0.974)
Between MS and DI	—	—	0.951	(0.934, 0.964)

accuracy estimate of 0.996 for both the readings within MS observers and the readings between MS and DI methods. These findings are consistent with the plots shown in Figure 1.

We also analyzed data from a carotid stenosis screening study. The goal of the study was to determine the suitability of magnetic resonance angiography (MRA) for the noninvasive screening of carotid artery stenosis compared with the current gold standard (invasive intra-arterial angiogram). Several methods use MRA technology and we consider the following two MRA methods: two-dimensional (2D) time of flight and three-dimensional (3D) time of flight. The raters assess both the left and right carotid arteries even though a patient may only have a problem in one of the arteries. Three raters used three methods (2D MRA, 3D MAR, and angiogram) to assess both arteries. Therefore, a total of 18 readings are available for each patient with nine readings from the left artery and nine readings from the right artery. These 18 readings are likely to be correlated because the readings are assessed from the same patient. There are 55 patients with all 18 readings from this study. The ranges of the carotid stenosis readings are 0–100%. To examine the data visually, we produced a total of 18 plots comparing pairwise readings from the same rater using different methods and comparing pairwise readings from different raters using the same method. For illustrative purposes, we present four representative plots in Figure 2 (three plots are from the readings of rater 1 using three different methods and one plot is from the readings of rater 2 and rater 3 using the angiogram method). These plots show that the readings have large variation around the 45° line. We fit a regression line for each of the plots. The estimated slope coefficients are 0.628, 0.789, 0.675, and 0.819, respectively. Thus, we expect moderate precision and relatively good accuracy from this example.

We considered the covariates age, gender, hypertension, diabetes, coronary artery disease, peripheral vascular disease, previous carotid endarterectomy, and previous anticoagulant therapy in the marginal and concordance modeling. Because the sample size is small for this example and our main interest is in reproducibility of the two new methods (2D MRA and 3D MRA), we begin with the model for the concordance correlation coefficients by including parameters corresponding to within-observer reproducibility, between-method reproducibility, and parameters for the available subject-specific covariates. We centered the age variable at its mean of 67 years. The results from the final model are presented in Table 4 for both the marginal model and the concordance correlation model.

From Table 4, we conclude that raters 2 and 3 have a tendency to assess higher stenosis percentage than rater 1. Raters using the 2D MRA and the 3D MRA methods tend to assess higher stenosis readings than using the angiogram method. Carotid stenosis is smaller for patients who are female, diabetic, or have previous anticoagulant therapy than patients who are male, nondiabetic, or without previous anticoagulant therapy. Patients with peripheral vascular disease have higher stenosis than their counterparts. From the concordance correlation model, we conclude that reproducibility is moderate, with the highest reproducibility between raters using the

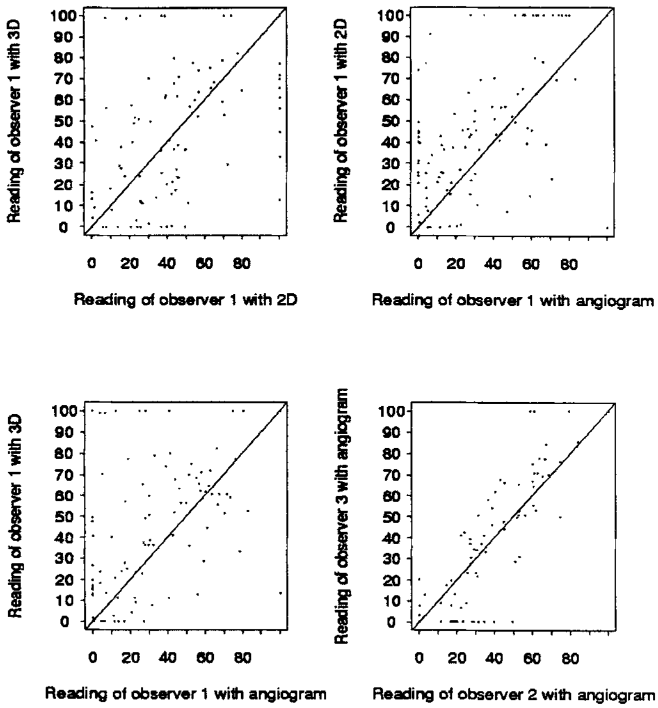

Figure 2. Stenosis readings from the carotid screening study.

angiogram ($\rho_c = 0.711$) and the smallest reproducibility between raters using the 2D MRA method compared with raters using the angiogram ($\rho_c = 0.534$). The reproducibility is higher for older patients (p-value $= 0.004$ based on the Wald test for the hypothesis $\alpha_{age} = 0$). Note that the 95% confidence intervals for the concordance correlation coefficients are wide due to the small sample size and moderate precision. Furthermore, these estimates may be slightly biased downward based on our simulation study with small sample sizes. Therefore, we should be cautious with the interpretations of these results and one may need to confirm these results with a future reproducibility study including a larger number of patients.

We also fit the same model using the correlation coefficient as a measure of association, and the results are presented in Table 4. The estimates of ρ range from 0.578 (between 3D MRA and angiogram) to 0.732 (within angiogram raters). This implies that the precision of the readings is moderate. Because the correlation coefficients are not very different from the concordance correlation coefficients, we obtain high estimates for accuracy. However, the confidence intervals for accuracy estimates are wide (not shown).

5. Discussion

We have proposed a GEE approach to model the concordance correlation coefficient to evaluate reproducibility. The proposed approach makes minimal distributional assumptions and takes into account the correlation between measurements made on the same subject when conducting inference. The proposed approach enables the data analyst to model differing concordance correlation coefficients simultaneously while adjusting for covariates. We illustrated the proposed method with the analyses of data from two biomedical studies.

Three sets of estimating equations may be necessary for modeling the concordance correlation coefficient. The second set of estimating equations may not be needed if moment estimates for the variances are used in place of $\boldsymbol{\sigma}^2$ in the third set of equations. We found that the standard errors for $\hat{\alpha}$ are not consistently estimated if only the first and the third sets of estimating equations are used because the uncertainty in estimating $\boldsymbol{\sigma}^2$ was not taken into account. Note that the last two estimating equations may be be combined into one estimating equation by using the responses $Y_{ij}Y_{ij'}$, $j \le j'$. The reason that we used two sets of estimating equations is to make the distinction between the parameters $\boldsymbol{\sigma}^2$ and $\boldsymbol{\alpha}$, where $\boldsymbol{\alpha}$ (ρ_c) is the main interest.

Subjects may not have an equal number of measurements in the presence of missing data. The GEE estimates may be biased if the missing data mechanism is not completely at random (Liang and Zeger, 1986). If the probability of missing is known (e.g., the missingness is by design), then one can modify the proposed GEE equations with weighted GEE equations (Robins, Rotnitzky, and Zhao, 1995). Further investigation is needed if one needs to estimate the missing data mechanism.

As discussed in the Introduction, intraclass correlation (Fleiss, 1986; Quan and Shih, 1996) and within-subject coefficient of variation (Lee et al., 1989) have been used traditionally as indices to evaluate reproducibility. These indices can be estimated by using random-effects models. However, these models require full distributional assumptions. Even though these models can make covariate adjustments in the marginal mean, it is not clear how one can model the agreement measure with covariates.

As indicated by Atkinson and Nevill (1997), any type of correlation coefficient is highly dependent on the measurement range. Lin and Chinchilli (1997) recommended that one should always report the range of the data and judge agreement of different measurement methods over a similar analytical range. These cautionary notes should be kept in mind when one uses the proposed method in practice. Also, King and Chinchilli (2000) have proposed a general index, the generalized concordance correlation coefficient, for evaluating agreement for continuous and categorical data. This index is a generalization of the concordance correlation coefficient by applying alternative functions of distance between readings other than squared distance. King and Chinchilli also introduced a stratified concordance correlation coefficient that adjusts for categorical covariates in the marginal mean and an extended concordance correlation coefficient that measures agreement among more than two responses.

The computer programs used in this article are available from the authors.

Acknowledgements

This research was supported in part by the National Institutes of Health (NIH) grant R29 HL58014 (HXB). 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 Drs Peterman and Haber (supported by the NIH grant R01 NS 30928) for providing the data set from the carotid

Table 4
Analyses of the carotid stenosis screening study

	Estimate	SE	P-value
Marginal parameter			
Intercept	47.382	6.015	<0.0001
Indicator for 2D MRA	11.074	1.723	<0.0001
Indicator for 3D MRA	11.863	2.240	<0.0001
Indicator for rater 2	3.880	1.006	0.0001
Indicator for rater 3	1.679	1.067	0.116
Age in years	0.411	0.154	0.007
Indicator for female gender	−12.455	3.731	0.0008
Diabetes (1=yes, 0=no)	−11.749	4.952	0.018
Peripheral vascular disease (1 = yes, 0 = no)	14.400	4.294	0.0008
Previous anticoagulant therapy (1 = yes, 0 = no)	−13.305	6.188	0.032

Concordance Correlation Model and Correlation Coefficient Model

	$\hat{\rho}_c$	95% CI	$\hat{\rho}$	$\hat{\rho}_c/\hat{\rho}$
Within 2D MRA raters	0.623	(0.369, 0.789)	0.639	0.971
Within 3D MRA raters	0.569	(0.302, 0.754)	0.590	0.973
Within angiogram raters	0.711	(0.447, 0.861)	0.732	0.965
Between 2D MRA and angiogram	0.625	(0.408, 0.775)	0.669	0.933
Between 3D MRA and angiogram	0.534	(0.311, 0.701)	0.578	0.924
Between 2D MRA and 3D MRA	0.567	(0.328, 0.738)	0.588	0.965
One-year increase in age	0.019	(0.006, 0.032)	0.021	—

screening study. We thank the associate editor and two referees for their helpful and constructive comments, which lead to a greatly improved manuscript.

Résumé

Les études cliniques sont souvent concernées par l'évaluation de différentes raters/methods qui donnent des valeurs semblables lors du mesurage d'une variable quantitative. L'usage du coefficient de concordance de corrélation comme une mesure de reproductibilité a gagné en popularité dans la pratique depuis son introduction par Lyn (1989). La méthode de Lin est applicable pour des études évaluant two raters /two method sans répétition. Chinchilli et al (1996) ont étendu l'approche de Lin pour des mesures répétées en utilisant un coefficient de concordance de corrélation pondéré. Cependant les méthodes existantes n'autorisent pas facilement un ajustement de covariable. Dans ce papier nous proposons une approche généralisée d'estimation d'équation (GEE) pour modéliser le coefficient de concordance de corrélation via trois jeux d'équations. L'approche proposée est flexible dans le fait qu'elle autorise plus de deux relevés corrélés et teste l'égalité des estimateurs des coefficients de concordance de corrélation. Elle autorise l'incorporation de covariables prédictives de la distribution marginale. Elle peut être utilisée pour identifier des covariables prédictives de la corrélation de concordance; elle demande un minimum de condition sur les distributions. Une étude de simulations est conduite pour évaluer les propriétés asymptotiques de l'approche proposée. La méthode est illustrée avec des données de deux études biomédicales.

References

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

Barlow, W. (1996). Measurement of interrater agreement with adjustment for covariates. *Biometrics* **52,** 695–702.

Chinchilli, V. M., Martel, J. K., Kumanyika, S., and Lloyd, T. (1996). A weighted concordance correlation coefficient for repeated measurement designs. *Biometrics* **52,** 341–353.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* **20,** 37–46.

Cohen, J. (1968). Weighted kappa: Normal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin* **70,** 213–220.

Donner, A. and Koval, J. (1980). The estimation of intraclass correlation in the analysis of family data. *Biometrics* **36,** 19–25.

Fleiss, J. L. (1986). *The Design and Analysis of Clinical Experiments.* New York: Wiley.

Fleiss, J. L. and Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement* **33,** 613–619.

Gonin, R., Lipsitz, S. R., Fitzmaurice, G. M., and Molenberghs, G. (2000). Regression modeling of weighted κ by using generalized estimating equations. *Applied Statistics* **49,** 1–18.

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

Klar, N., Lipsitz, S. R., and Ibrahim, J. G. (2000). An estimating equations approach for modeling kappa. *Biometrical Journal* **42,** 45–58.

Krippendorff, K. (1970). Bivariate agreement coefficients for reliability of data. In *Social Methodology,* B. F. Borgatta and G. W. Bohrnstedt (eds), 139–150. San Francisco: Jossey-Bass.

Lee, J., Koh, D., and Ong, C. N. (1989). Statistical evaluation of agreement between two methods for measuring a quantitative variable. *Computers in Biology and Medicine* **19,** 61–70.

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

Liang, K. Y, Zeger, S. L., and Qaqish, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society, Series B* **54,** 3–24.

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. and Chinchilli, V. (1997). Rejoinder to the letter to the editor from Atkinson and Nevill. *Biometrics* **53,** 777–778.

Lipsitz, S. R. and Fitzmaurice, G. (1996). Estimating equations for measures of association between repeated binary responses. *Biometrics* **52,** 903–912.

Molenberghs, G., Fitzmaurice, G. M., and Lipsitz, S. R. (1996). Efficient estimation of the intraclass correlation for a binary trait. *Journal of Agricultural, Biological, and Environmental Statistics* **1,** 78–96.

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics* **44,** 1033–1048.

Quan, H. and Shih, W. J. (1996). Assessing reproducibility by the within-subject coefficient of variation with random effects models. *Biometrics* **52,** 1195–1203.

Robieson, W. (1999). On weighted kappa and concordance correlation coefficient. Ph.D. thesis, University of Illinois in Chicago/Graduate College/Mathematics.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of American Statistical Association* **90,** 106–121.

Shoukri, M. M. and Mian, I. U. H. (1996). Maximum likelihood estimation of the kappa coefficient from bivariate logistic regression. *Statistics in Medicine* **15,** 1409–1419.

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.

Williamson, J. M., Manatunga, A. K., and Lipsitz, S. R. (2000). Modeling kappa for measuring dependent categorical agreement data. *Biostatistics* **1,** 191–202.

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

Received May 2000. *Revised December* 2001.
Accepted December 2001.