Research Article

Statistics in Medicine

Received 8 March 2011, Accepted 5 December 2011 Published online 13 March 2012 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5323

A new permutation-based method for assessing agreement between two observers making replicated quantitative readings

Yi Pan,[a] Michael Haber,[a*†] Jingjing Gao[b] and Huiman X. Barnhart[c]

The coefficient of individual equivalence is a permutation-based measure of agreement between two observers making replicated readings on each subject. It compares the observed disagreement between the observers to the expected disagreement under individual equivalence. Individual equivalence of observers requires that for every study subject, the conditional distributions of the readings of the observers given the subject's characteristics are identical. Therefore, under individual equivalence it does not matter which observer is making a particular reading on a given subject. We introduce both nonparametric and parametric methods to estimate the coefficient as well as its standard error. We compare the new coefficient with the coefficient of individual agreement and with the concordance correlation coefficient. We also evaluate the performance of the estimates of the new coefficient via simulations and illustrate this new approach using data from a study comparing two noninvasive techniques for measuring carotid stenosis to an invasive gold standard.

Keywords: agreement; coefficient of individual equivalence; quantitative measurements

1. Introduction

One reason for the existence of so many indices for assessing agreement between observers or methods of measurement is the lack of a good criterion for 'reasonable' or 'acceptable' agreement. In this article, we propose such a criterion. For simplicity, suppose that there are only two observers, whose readings are represented by the random variables X and Y. Further, let $f_X(u)$ and $f_Y(u)$ denote the respective probability mass functions (for categorical observations) or the probability density functions (for quantitative observations). Hawkins [1] called two observers *equivalent* if for each study subject the conditional distributions of X and Y, given the subject's characteristics, are identical: that is, $f_X(u|i) = f_Y(u|i)$ for every u and every subject $i = 1, \ldots N$. If observers are equivalent, then one can argue that their agreement is at least 'acceptable' because, from a statistical point of view, *it does not matter whether the next observation on a given subject will be made by observer X or by observer Y*. In other words, the two observers can be used interchangeably, or one observer can be replaced by the other at any time during the study. We will call this property *individual equivalence*.

We assume that the magnitude of disagreement between two readings, x and y, on the same subject is quantified via a disagreement function $G(x, y)$. The most commonly used disagreement function is the mean squared deviation (MSD), $G(x, y) = (x - y)^2$, but other disagreement functions can be used

[a] *Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA 30322, USA*
[b] *Center for Comprehensive Informatics, Emory University, Atlanta, GA 30322, USA*
[c] *Department of Biostatistics and Bioinformatics and Duke Clinical Research Institute, Duke University, PO Box 17969, Durham, NC 27715, USA*
**Correspondence to: Michael Haber, Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA 30322, USA.*
[†] *E-mail: mhaber@sph.emory.edu*

[2]. *The objective of this article is to introduce a method to determine whether the disagreement between observers is in line with or is substantially larger than 'acceptable' disagreement, which we define as the disagreement that can be expected if the observers are equivalent.* Therefore, we define a new *coefficient of individual equivalence* (CIE) on the basis of comparing the observed value of the disagreement function with its expected value under the hypothesis of individual equivalence. The expected disagreement under equivalence can be viewed as *disagreement by chance*, that is, the magnitude of the disagreement function that would be observed if, in fact, all the readings on the same subject were made by the same observer and later someone randomly assigned the letter X to some of the observations and the letter Y to the remaining observations. This concept of disagreement by chance is very different from the concept of agreement by chance, which is often used to obtain scaled agreement measures (e.g., kappa or the concordance correlation coefficient (CCC)). Agreement by chance is defined as the expected agreement under independence of the observers.

In order to assess the magnitude of departure from individual equivalence, replicated readings of at least one observer on each subject should be available. Suppose that for each subject there are K replicated readings by observer X and L replicated readings by observer Y ($K, L \geqslant 1$ and $K + L \geqslant 3$). Under the hypothesis of individual equivalence, all the C_K^{K+L} allocations of the $K + L$ readings, so that K readings are assigned to observer X and the remaining L readings are assigned to observer Y, are equally likely. Therefore, the expected disagreement under equivalence will be estimated as the mean of the values of the disagreement function over all the possible C_K^{K+L} allocations. The equivalence between the original definition of individual equivalence and the permutation-based approach is shown in Appendix A.

Pan *et al.* [3] introduced and discussed the CIE when two observers make binary readings on each subject. In the present work, we focus on the properties of the new approach when the observations are quantitative. In addition to the nonparametric approach, which is similar to Pan *et al.* [3], we introduce a parametric approach for estimation and inference. We present a motivating example where the CIE can be applied to assess agreement between quantitative replicated measurements in Section 2. We introduce the CIE, which compares the observed and expected disagreement, as well as its nonparametric and parametric estimation, in Section 3. We discuss an adjusted CIE, as well as its estimation, in Section 4. We present simulation results in Section 5, whereas in Section 6, we apply the new concepts and methods to the motivating example. We discuss comparisons between the CIE, the coefficient of individual agreement (CIA), and the CCC in Section 7. A discussion in Section 8 concludes this article.

2. A motivating example

We will use a carotid stenosis screening study as an example. This study was designed to determine the suitability of magnetic resonance angiography (MRA) for noninvasive screening of carotid artery stenosis compared with invasive intra-arterial angiogram (IA). The main interest is in comparing two MRA techniques, two-dimensional (MRA-2D) and three-dimensional (MRA-3D) MRA time of flight, with the IA which is considered as the 'gold standard'. In this example, the three screening methods are considered as the 'observers'. Readings were made by each of three raters using each of the three methods to assess carotid stenosis on each of the 55 patients. For this illustration, the three readings made by different raters are considered as replications. Separate readings were made on the left and right carotid arteries. However, in this example our interest is restricted to the left side. For more details on this study, the reader is referred to Barnhart and Williamson [4].

3. Definition and estimation of the coefficient of individual equivalence

3.1. Definition of the coefficient of individual equivalence

The CIE compares the observed disagreement with the expected disagreement under individual equivalence. Let X, Y be two observers making quantitative observations on N subjects, $i = 1, \ldots, N$. For subject i, let $G_i(X, Y)$ denote the disagreement between X and Y, $G_i(X, X')$ the disagreement between two replicated measurements made by the observer X, and $G_i(Y, Y')$ is analogously defined for observer Y. We also define the overall disagreement functions $G(X, X') = E_i[G_i(X, X')]$, $G(Y, Y') = E_i[G_i(Y, Y')]$, $G(X, Y) = E_i[G_i(X, Y)]$, where E_i denotes the mean over all N subjects. As stated in the introduction, we assume that observers X and Y make K and L replicated observations,

respectively, on each subject, where $K \geqslant 1$, $L \geqslant 1$ and $K + L \geqslant 3$. Then the CIE is defined by

$$\text{CIE} = \frac{E_i(G_i(X,Y) \text{ under individual equivalence})}{E_i(G_i(X,Y))} = \frac{E_i(G_i^E)}{G(X,Y)} = \frac{G^E}{G(X,Y)}, \tag{1}$$

where G_i^E is defined as the expected value of $G_i(X,Y)$ under the assumption that all the C_K^{K+L} assignments of K $X's$ and L $Y's$ to the $K+L$ observations made on subject i are equally likely. In Appendix A, we prove that the above assumption is equivalent to the assumption of individual equivalence $f_X(u|i) = f_Y(u|i)$ for every u and i [1].

Evaluation of G^E can be simplified as follows. Denote the $K+L$ measurements on subject i as $Z_{i1}, Z_{i2}, Z_{i3}, \ldots, Z_{i(K+L)}$. Then G_i^E, the mean of $G_i(X,Y)$ over all C_K^{K+L} assignments, is equal to the mean of all C_2^{K+L} terms $G(Z_{ih}, Z_{ih'})$, $1 \leqslant h < h' \leqslant K+L$. The proof can be found in Pan *et al*. [3].

Therefore,

$$\begin{aligned} G_i^E &= \frac{\sum_{h<h'} G(Z_{ih}, Z_{ih'})}{C_2^{K+L}} \\ &= \frac{C_2^K G_i(X,X') + C_2^L G_i(Y,Y') + K \cdot L \cdot G_i(X,Y)}{C_2^{K+L}} \end{aligned} \tag{2}$$

and $G^E = E_i[G_i^E]$. Thus, from (1) and (2), it is evident that 'true' CIE can be written as

$$\text{CIE} = \frac{C_2^K G(X,X') + C_2^L G(Y,Y')}{C_2^{K+L} G(X,Y)} + \frac{K \cdot L}{C_2^{K+L}}. \tag{3}$$

3.2. *Nonparametric estimation of coefficient of individual equivalence*

The CIE can be estimated non-perimetrically as follows. Denote the observed values for subject i by $X_i = (X_{i1}, \ldots, X_{iK})$ and $Y_i = (Y_{i1}, \ldots, Y_{iL})$. Then the estimated disagreement between observer X and Y for subject i is $\hat{G}_i(X,Y) = \text{mean}_{k,l}[G(X_{ik}, Y_{il})]$ (the mean over all $K \cdot L$ pairs of an observation from X and an observation from Y). The estimated overall disagreement between observer X and Y is $\hat{G}(X,Y) = \text{mean}_i[\hat{G}_i(X,Y)]$. The estimated disagreement between two observations of observer X on subject i is $\hat{G}_i(X,X') = \text{mean}_{k<k'}[G(X_{ik}, X_{ik'}]$, and the overall disagreement is $\hat{G}(X,X') = \text{mean}_i[\hat{G}_i(X,X')]$. $\hat{G}(Y,Y')$ is obtained in a similar way.

Thus, from (3),

$$\widehat{\text{CIE}} = \frac{C_2^K \hat{G}(X,X') + C_2^L \hat{G}(Y,Y')}{C_2^{K+L} \hat{G}(X,Y)} + \frac{K \cdot L}{C_2^{K+L}}. \tag{4}$$

The nonparametric method can be used with any disagreement function G.

3.3. *Estimation of coefficient of individual equivalence with $G(X,Y) = E(X-Y)^2$ using linear mixed effects models for normally distributed observations*

When X and Y are normally distributed and $G(X,Y) = E(X-Y)^2$, a parametric approach to estimation and inference on CIE can be based on a two-way mixed linear model with subject effect α_i, observer effect β_j, and the interaction between subject and observer γ_{ij}. The residual term, ε_{ijk}, represents the within-observer replication variability. When there are $J \geqslant 2$ observers ($j = 1, \cdots, J$), we model the kth replication of observer j on subject i as

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \tag{5}$$

where the subject effect is random and the observer effect is fixed. We assume that the random effects' distributions are $\alpha_i \sim N(0, \sigma_\alpha^2)$, $\gamma_{ij} \sim N(0, \sigma_\gamma^2)$, and $\varepsilon_{ijk} \sim N(0, \sigma_{\varepsilon_j}^2)$. We also assume that α_i, γ_{ij}, and ε_{ijk} are independent. Note that each observer may have a different within-observer variance σ_{ε_j}. Because the observer is treated as a fixed effect, define σ_β^2 as $\sigma_\beta^2 = \frac{1}{J-1}\sum_{j=1}^J (\beta_j - \bar{\beta})^2$.

Suppose now that there are only two observers, X and Y, and denote $X = Y_1$ and $Y = Y_2$. Let $G(X,Y) = \mathrm{MSD}(X,Y) = E(X-Y)^2$. It is easy to see that

$$\begin{aligned}\mathrm{MSD}(X,X') &= E(Y_{i1k} - Y_{i1k'})^2 = E(\varepsilon_{i1k} - \varepsilon_{i1k'})^2 = 2\sigma^2_{\varepsilon_1}\\ \mathrm{MSD}(Y,Y') &= E(Y_{i2k} - Y_{i2k'})^2 = E(\varepsilon_{i2k} - \varepsilon_{i2k'})^2 = 2\sigma^2_{\varepsilon_2}\\ \mathrm{MSD}(X,Y) &= E(Y_{i1k} - Y_{i2k'})^2 = E(\beta_1 - \beta_2)^2 + E(\gamma_{i1} - \gamma_{i2})^2 + E(\varepsilon_{i1k} - \varepsilon_{i2k'})^2\\ &= 2\sigma^2_\beta + 2\sigma^2_\gamma + \sigma^2_{\varepsilon_1} + \sigma^2_{\varepsilon_2}.\end{aligned}$$

As we have seen, the expression of CIE can be simplified as (3). When $G = \mathrm{MSD}$ is used as the disagreement function, $G(X,X')$, $G(Y,Y')$, and $G(X,Y)$ can be written in terms of the parameters of the mixed model as shown before. Hence,

$$\mathrm{CIE} = \frac{C_2^K\sigma^2_{\varepsilon_1} + C_2^L\sigma^2_{\varepsilon_2}}{C_2^{K+L}\cdot(\sigma^2_\beta + \sigma^2_\gamma + 1/2\sigma^2_{\varepsilon_1} + 1/2\sigma^2_{\varepsilon_2})} + \frac{K\cdot L}{C_2^{K+L}}.$$

The CIE can be estimated by replacing the parameters with their restricted maximum likelihood estimates. In the special case $K = L$, the estimate of CIE can be simplified as

$$\widehat{\mathrm{CIE}} = \frac{K-1}{2K-1}\cdot\frac{\hat\sigma^2_\varepsilon}{\hat\sigma^2_\beta + \hat\sigma^2_\gamma + \hat\sigma^2_\varepsilon} + \frac{K}{(2K-1)}.$$

where $\hat\sigma^2_\varepsilon = (\hat\sigma^2_{\varepsilon_1} + \hat\sigma^2_{\varepsilon_2})/2$.

3.4. Minimum and maximum values of coefficient of individual equivalence

3.4.1. Minimum value of coefficient of individual equivalence. In order to derive a coefficient ranging from zero to one, we need to look at the minimum of CIE and make an adjustment if necessary. In (2), if $G(X,X') = G(Y,Y') = 0$ while keeping $G(X,Y) > 0$, CIE achieves its minimum as follows:

$$\begin{aligned}\mathrm{CIE}_{\min} &= \frac{K\cdot L}{C_2^{K+L}}\\ &= \frac{2K\cdot L}{(K+L)(K+L-1)}.\end{aligned} \tag{6}$$

3.4.2. Maximum value of coefficient of individual equivalence. The CIE can be written as shown in (3). Note that acceptable within-observer disagreement is required when considering the use of CIE to assess inter-observer agreement. We only need to consider the situation where for each observer the within-observer disagreement is less than the between-observer disagreement, that is, $G(X,X') \leqslant G(X,Y)$ and $G(Y,Y') \leqslant G(X,Y)$. If $G(X,X') > G(X,Y)$ or $G(Y,Y') > G(X,Y)$, then the inter-observer agreement is also acceptable because inter-observer disagreement is smaller than the acceptable intra-observer agreement so that no further evaluation for inter-observer agreement is needed. Under these assumptions, CIE achieves its maximum when $G(X,X') = G(X,Y)$ and $G(Y,Y') = G(X,Y)$.

$$\begin{aligned}\mathrm{CIE}_{\max} &= \frac{C_2^K + C_2^L + K\cdot L}{C_2^{K+L}} = \frac{\frac{K(K-1)}{2} + \frac{L(L-1)}{2} + K\cdot L}{\frac{(K+L)(K+L-1)}{2}}\\ &= \frac{K(K-1) + L(L-1) + 2K\cdot L}{(K+L)(K+L-1)} = \frac{(K+L)^2 - (K+L)}{(K+L)(K+L-1)} = 1.\end{aligned}$$

Therefore, the maximum value of CIE for continuous measurement under reasonable assumptions is 1.

4. Adjusted coefficient of individual equivalence

4.1. Definition

As we see, the minimum of CIE is greater than 0 and when both K and L are greater than 1, $\mathrm{CIE}_{\min}$ is greater than 0.5. To force the index to range from 0 to 1, we define an adjusted CIE. The adjusted CIE is

denoted as CIEA and defined as follows:

$$\text{CIEA} = \frac{\text{CIE} - \text{CIE}_{\min}}{1 - \text{CIE}_{\min}}, \tag{7}$$

where $\text{CIE}_{\min}$ is defined in (6).The estimate of CIEA can be easily obtained from $\widehat{\text{CIE}}$ because CIEA is a linear function of CIE.

4.2. *Large sample distribution and standard error of* $\widehat{\text{CIEA}}$

Let us first consider the nonparametric estimator of CIE. Because both estimators of G^E and $G(X,Y)$ are means over all N subjects, the asymptotic normality of the estimate of CIE can be established by applying the multivariate central limit theorem and multivariate delta method when the sample size is large enough.

Let $A = \hat{G}^E = \frac{1}{N}\sum_{i=1}^{N}\hat{G}_i^E$ and $B = \hat{G}_i(X,Y) = \frac{1}{N}\sum_{i=1}^{N}\hat{G}_i(X,Y)$. Then $\widehat{\text{CIE}} = A/B$. The sample variance of A is $S^2(A) = \frac{1}{N-1}\sum_{i=1}^{N}(\hat{G}_i^E - \hat{G}^E)^2$ and then $\widehat{\text{Var}}(A) = S^2(A)/N$. Similarly, $\widehat{\text{Var}}(B) = S^2(B)/N$, where $S^2(B) = \frac{1}{N-1}\sum_{i=1}^{N}(\hat{G}_i(X,Y) - \hat{G}(X,Y))^2$. Also, $\widehat{\text{Cov}}(A,B) = \left[\sum_{i=1}^{N}(\hat{G}_i^E - \hat{G}^E)(\hat{G}_i(X,Y) - \hat{G}(X,Y))\right]/N(N-1)$. Finally,

$$\widehat{\text{Var}}(\widehat{\text{CIE}}) = \widehat{\text{Var}}\left(\frac{A}{B}\right) \approx \frac{A^2}{B^2}\left[\frac{\widehat{\text{Var}}(A)}{A^2} + \frac{\widehat{\text{Var}}(B)}{B^2} - \frac{\widehat{\text{Cov}}(A,B)}{A\cdot B}\right].$$

For the parametric estimation of CIE from a linear mixed model, bootstrap standard errors (SEs) and normalized confidence interval based on bootstrap SEs can be used.

As CIEA is just a linear transformation of CIE, the above asymptotic properties hold for $\widehat{\text{CIEA}}$. The standard error of $\widehat{\text{CIEA}}$ for both nonparametric method and linear mixed model estimation is obtained as follows:

$$\hat{\text{SE}}\left(\widehat{\text{CIEA}}\right) = \frac{1}{(1-\text{CIE}_{\min})}\sqrt{\widehat{\text{Var}}\left(\widehat{\text{CIE}}\right)}.$$

4.3. *Interpretation of adjusted coefficient of individual equivalence*

As with every new agreement coefficient, it is necessary to specify criteria for 'good' or 'acceptable' agreement in a given application. Intuitively, CIE compares the observed disagreement with disagreement under individual equivalence, and CIEA is the adjusted CIE using the minimum value of CIE. We suggest that the true CIEA value should be greater than or equal to 0.8 in order to have good agreement. This implies that the lower confidence limit of CIEA needs to be greater than or equal to 0.8 when analyzing data in order to claim good agreement.

5. Simulation studies

Simulation studies were conducted to evaluate the performance of both nonparametric and parametric approaches for estimation and inference on the CIEA. For nonparametric estimation, simulations were performed for small ($n = 50$), moderate ($n = 100$), and large sample sizes ($n = 200$). For the parametric method, only sample size $n = 100$ was included. In all settings, we looked at balanced number of replications for both raters ($K = L = 3$) and unbalanced scenarios ($K = 1$, $L = 2$) and ($K = 2$, $L = 3$). The ($K = 1$, $L = 2$) scenario corresponds to the case where X is a perfect gold standard that does not require replications. The disagreement function $G = \text{MSD}$ was used in all the simulations.

In the first set of simulations, we assumed that the true value $T_i \sim N(\mu_T, \sigma_T^2)$. Furthermore, we assumed the conditional means and standard deviations of the observers' readings given subjects' true values are linear functions of t: $\mu_{X|t} = a + bt, \mu_{Y|t} = c + dt, \sigma_{X|t} = e + ft, \sigma_{Y|t} = g + ht$. Then for subject i, K replicated measurements of observer X were generated from $N(\mu_{X|t_i}, \sigma^2_{X|t_i})$, and L

replicated measurements of observer Y were generated from $N(\mu_{Y|t_i}, \sigma^2_{Y|t_i})$. The three MSD functions can be expressed as [2]:

$$\begin{aligned}\mathrm{MSD}(X,Y) &= (a-c)^2 + e^2 + g^2 + 2[(a-c)(b-d) + ef + gh]\mu_T \\ &\quad + [(b-d)^2 + f^2 + h^2](\mu_T^2 + \sigma_T^2) \\ \mathrm{MSD}(X,X') &= 2e^2 + 4ef\mu_T + 2f^2(\mu_T^2 + \sigma_T^2) \\ \mathrm{MSD}(Y,Y') &= 2g^2 + 4gh\mu_T + 2h^2(\mu_T^2 + \sigma_T^2).\end{aligned}$$

Data from a carotid stenosis study (see Section 2) were used to investigate the behavior of the new coefficient. The distribution of T was defined using the sample moments from the data, $\mu_T = 43.29, \sigma_T = 29.87$. We used two sets of parameters for the within-observer means and standard deviations in order to explore the effect of difference in variances of X and Y given T: (1) $b = 1, d = 1, e = g = 1.5, f = h = 0.3$; (2) $b = 1, d = 1, e = 1.5, g = 1, f = h = 0.3$. To accommodate good, moderate, and poor agreement, respectively, we used $a = 0$ and let $c = 3.8, 16.3$, and 28.1. To investigate the performance of the nonparametric and parametric approaches when the distribution of the true value is skewed, we conducted a second set of simulation as the true value T followed an exponential distribution $T_i \sim \mathrm{EXP}(\lambda_T)$ with the mean and standard deviation parameter as 29.87. The number of simulations when we have equal within-observer variances ($e = g$) was 1000. For scenarios with unequal variances, the number of simulations was reduced to 100 because we had to fit the analysis of variance mixed model with heteroscedastic error terms. The number of bootstrap samples in estimating the standard error for the parametric approach was 100.

Tables I–IV present the true values, biases, standard errors, root mean squared errors (RMSE), and the coverage probabilities of the estimates of CIEA for all the combinations of (N, K, L) for the poor, moderate, and good agreement cases defined earlier when T was normally distributed. We considered nonparametric estimation with equal variance, nonparametric estimation with unequal variance, parametric estimation with equal variance, and parametric estimation with unequal variance. As shown in Tables I–IV, the bias was minimal for all combinations, and it decreased when the sample size increased. We present both standard errors based on simulations of CIEA and the mean of estimated standard errors. The similarity between those two standard errors confirms the robustness of our standard error estimates. For moderate and large sample sizes, the coverage probabilities were very close to the nominal 95% level.

Furthermore, we noticed that when equal variances were used, the true values of CIEA were not affected by the choice of K and L. When the variance parameters were not the same, the true values of CIEA varied in a limited amount. In addition, we used RMSE for comparing the efficiency of the nonparametric and parametric estimates. For all the scenarios we considered, the RMSE of the parametric estimates was consistently smaller than that of the nonparametric method. This indicates that when the data were generated from normal distributions, linear mixed models produced more efficient estimates than the nonparametric method did. We obtained similar simulation results when T was exponentially distributed (results not shown), but the coverage probabilities under all scenarios were smaller than those when T was normally distributed.

6. Carotid stenosis example

The carotid stenosis example, introduced in Section 2, compares two MRA techniques, two-dimensional (MRA-2D) and three-dimensional (MRA-3D) MRA time of flight, with the IA, which was considered as the 'gold standard'. We treated three readings from three raters as the replicated observations on 55 patients.

The estimates, standard errors, and 95% confidence intervals of the CIEA by nonparametric and parametric estimation methods are presented in Table V for all the pairwise comparisons of the three screening methods. For the nonparametric method, the estimated CIEA of MRA-2D and the gold standard IA was 0.592 with the 95% CI (0.348, 0.835), which indicated moderate agreement. Similarly, a moderate agreement was obtained when comparing MRA-3D with IA with CIEA = 0.452 (95% = (0.242, 0.661)). Even the comparison between MRA-2D and MRA-3D with CIEA = 0.881 and 95% CI (0.688, 1.000) failed to show good agreement because the 95% lower limit was below 0.8. The parametric approach produced compatible estimates of CIEA.

Table I. Nonparametric simulation results: estimation of CIEA with $(K, L) = (1, 2)$, (2,3), and (3,3) when T is normal and within-observer variances are equal ($e = g = 1.5$).

Sample size	K	L	c	True	Bias	SE*	SE†	RMSE‡	CP§
50	1	2	0.0	1.000	0.04	0.300	0.262	0.264	0.884
100	1	2	0.0	1.000	0.015	0.205	0.195	0.195	0.924
200	1	2	0.0	1.000	0.006	0.146	0.141	0.141	0.931
50	2	3	0.0	1.000	0.008	0.141	0.125	0.125	0.894
100	2	3	0.0	1.000	0.008	0.100	0.093	0.094	0.906
200	2	3	0.0	1.000	0.002	0.072	0.068	0.068	0.928
50	3	3	0.0	1.000	0.000	0.099	0.092	0.092	0.905
100	3	3	0.0	1.000	0.004	0.068	0.068	0.068	0.934
200	3	3	0.0	1.000	0.004	0.051	0.05	0.05	0.936
50	1	2	3.8	0.976	0.038	0.293	0.255	0.258	0.879
100	1	2	3.8	0.976	0.014	0.2	0.19	0.191	0.926
200	1	2	3.8	0.976	0.005	0.143	0.138	0.138	0.931
50	2	3	3.8	0.976	0.008	0.139	0.124	0.124	0.896
100	2	3	3.8	0.976	0.007	0.099	0.092	0.093	0.909
200	2	3	3.8	0.976	0.002	0.071	0.067	0.067	0.931
50	3	3	3.8	0.976	0.000	0.099	0.092	0.092	0.899
100	3	3	3.8	0.976	0.004	0.069	0.068	0.068	0.93
200	3	3	3.8	0.976	0.004	0.051	0.05	0.05	0.932
50	1	2	16.3	0.686	0.017	0.204	0.185	0.186	0.904
100	1	2	16.3	0.686	0.005	0.142	0.136	0.136	0.922
200	1	2	16.3	0.686	0.000	0.101	0.098	0.098	0.93
50	2	3	16.3	0.686	0.003	0.111	0.104	0.104	0.914
100	2	3	16.3	0.686	0.003	0.079	0.075	0.076	0.932
200	2	3	16.3	0.686	0.000	0.057	0.054	0.054	0.925
50	3	3	16.3	0.686	−0.002	0.088	0.084	0.084	0.922
100	3	3	16.3	0.686	0.001	0.062	0.062	0.062	0.934
200	3	3	16.3	0.686	0.001	0.046	0.044	0.044	0.933
50	1	2	28.1	0.424	0.006	0.126	0.117	0.117	0.899
100	1	2	28.1	0.424	0.001	0.089	0.086	0.086	0.921
200	1	2	28.1	0.424	−0.001	0.063	0.061	0.061	0.931
50	2	3	28.1	0.424	0.001	0.076	0.073	0.073	0.91
100	2	3	28.1	0.424	0.001	0.055	0.053	0.053	0.923
200	2	3	28.1	0.424	−0.001	0.039	0.038	0.038	0.925
50	3	3	28.1	0.424	−0.001	0.064	0.063	0.063	0.93
100	3	3	28.1	0.424	0.000	0.046	0.046	0.046	0.928
200	3	3	28.1	0.424	0.000	0.034	0.033	0.033	0.936

*Standard errors based on simulations of CIE.
†Mean of estimated standard errors calculated from SE estimator.
‡Root mean squared error.
§Coverage probability of 95% confidence interval.

7. Comparison of adjusted coefficient of individual equivalence with coefficients of individual agreement and the concordance correlation coefficient

7.1. Equivalence between adjusted coefficient of individual equivalence and coefficient of individual agreement when $K = L$

Coefficients of individual agreement were proposed by Barnhart, Haber, and colleagues [2, 5–9] and can be used to assess agreement for both quantitative and categorical measurements. The CIAs with a specific disagreement function G are defined as:

$$\psi^N = \frac{[G(X, X') + G(Y, Y')]/2}{G(X, Y)}, \tag{8}$$

Table II. Nonparametric simulation results: estimation of CIEA when $(K, L) = (1, 2)$, (2,3), and (3,3) when T is normal and within-observer variances are unequal $(e = 1.5, g = 1)$.

Sample size	K	L	c	True	Bias	SE*	SE†	RMSE‡	CP§
50	1	2	3.8	0.951	0.038	0.290	0.252	0.255	0.878
100	1	2	3.8	0.951	0.014	0.198	0.188	0.188	0.926
200	1	2	3.8	0.951	0.005	0.142	0.136	0.136	0.931
50	2	3	3.8	0.963	0.008	0.138	0.123	0.124	0.894
100	2	3	3.8	0.963	0.007	0.098	0.092	0.092	0.909
200	2	3	3.8	0.963	0.002	0.070	0.067	0.067	0.934
50	3	3	3.8	0.975	0.000	0.100	0.092	0.092	0.9
100	3	3	3.8	0.975	0.004	0.069	0.068	0.068	0.928
200	3	3	3.8	0.975	0.004	0.051	0.05	0.05	0.933
50	1	2	16.3	0.663	0.017	0.201	0.181	0.182	0.901
100	1	2	16.3	0.663	0.005	0.140	0.134	0.134	0.921
200	1	2	16.3	0.663	0.000	0.099	0.096	0.096	0.931
50	2	3	16.3	0.672	0.003	0.110	0.102	0.103	0.914
100	2	3	16.3	0.672	0.003	0.078	0.075	0.075	0.93
200	2	3	16.3	0.672	−0.001	0.056	0.054	0.054	0.924
50	3	3	16.3	0.681	−0.002	0.088	0.084	0.084	0.921
100	3	3	16.3	0.681	0.001	0.062	0.062	0.062	0.934
200	3	3	16.3	0.681	0.001	0.046	0.044	0.044	0.933
50	1	2	28.1	0.407	0.006	0.123	0.114	0.114	0.901
100	1	2	28.1	0.407	0.001	0.087	0.083	0.083	0.921
200	1	2	28.1	0.407	−0.001	0.061	0.06	0.06	0.931
50	2	3	28.1	0.412	0.001	0.075	0.072	0.072	0.905
100	2	3	28.1	0.412	0.001	0.054	0.052	0.052	0.923
200	2	3	28.1	0.412	−0.001	0.039	0.037	0.037	0.923
50	3	3	28.1	0.418	−0.001	0.064	0.063	0.063	0.931
100	3	3	28.1	0.418	0.000	0.046	0.045	0.045	0.928
200	3	3	28.1	0.418	0.000	0.034	0.033	0.033	0.934

*Standard errors based on simulations of CIE.
†Mean of estimated standard errors calculated from SE estimator.
‡Root mean squared error.
§Coverage probability of 95% confidence interval.

Table III. Parametric simulation results: estimation of CIEA when $(K, L) = (1, 2)$, (2, 3), and (3, 3) when T is normal and within-observer variances are equal $(e = g = 1.5)$.

Sample size	K	L	c	True	Bias	SE*	SE†	RMSE‡	CP§
100	1	2	0	1.000	−0.08	0.101	0.093	0.122	0.946
100	2	3	0	1.000	−0.037	0.054	0.047	0.06	0.964
100	3	3	0	1.000	−0.028	0.040	0.038	0.048	0.977
100	1	2	3.8	0.975	−0.078	0.102	0.095	0.122	0.936
100	2	3	3.8	0.975	−0.036	0.056	0.050	0.062	0.953
100	3	3	3.8	0.975	−0.028	0.042	0.041	0.05	0.964
100	1	2	16.3	0.686	−0.050	0.088	0.086	0.100	0.899
100	2	3	16.3	0.686	−0.022	0.059	0.057	0.061	0.910
100	3	3	16.3	0.686	−0.018	0.051	0.050	0.053	0.933
100	1	2	28.1	0.424	−0.029	0.062	0.061	0.067	0.889
100	2	3	28.1	0.424	−0.012	0.046	0.044	0.046	0.898
100	3	3	28.1	0.424	−0.010	0.041	0.040	0.041	0.918

*Standard errors based on simulations of CIE.
†Mean of estimated standard errors calculated from bootstrap SE estimator.
‡Root mean squared error.
§coverage probability of 95% confidence interval.

Table IV. Parametric simulation results: estimation of CIEA when $(K, L) = (1, 2)$, $(2, 3)$, and $(3, 3)$ when T is normal and within-observer variances are unequal ($e = 1.5, g = 1$).

Sample size	K	L	c	True	Bias	SE*	SE†	RMSE‡	CP§
100	1	2	3.8	0.951	−0.006	0.206	0.171	0.171	0.920
100	2	3	3.8	0.963	−0.031	0.061	0.068	0.075	0.940
100	3	3	3.8	0.975	−0.036	0.041	0.046	0.058	0.960
100	1	2	16.3	0.663	−0.005	0.135	0.118	0.118	0.940
100	2	3	16.3	0.672	−0.022	0.064	0.061	0.065	0.900
100	3	3	16.3	0.681	−0.026	0.051	0.051	0.058	0.880
100	1	2	28.1	0.407	−0.004	0.083	0.074	0.074	0.940
100	2	3	28.1	0.412	−0.010	0.050	0.046	0.047	0.900
100	3	3	28.1	0.418	−0.013	0.044	0.040	0.042	0.900

*Standard errors based on simulations of CIE.
†Mean of estimated standard errors calculated from bootstrap SE estimator.
‡Root mean squared error.
§Coverage probability of 95% confidence interval.

Table V. Estimation of agreement in the carotid stenosis study using CIEA.

Methods compared	Estimate	SE of CIEA	95% CI for CIEA
Nonparametric estimation			
(IA, MRA-2D)	0.592	0.124	(0.348, 0.835)
(IA, MRA-3D)	0.452	0.107	(0.242, 0.661)
(MRA-2D, MRA-3D)	0.881	0.099	(0.688, 1.000)
Parametric estimation			
(IA, MRA-2D)	0.585	0.133	(0.373, 0.874)
(IA, MRA-3D)	0.447	0.100	(0.290, 0.671)
(MRA-2D, MRA-3D)	0.875	0.091	(0.634, 0.998)

$$\psi^R = \frac{G(X, X')}{G(X, Y)}. \tag{9}$$

These coefficients are close to 1 if the disagreement between observers is not much larger than the disagreement within observers. ψ^N is used when neither of the observers is considered as a reference; ψ^R is used when observer X is the reference. In this article, only ψ^N is considered and is denoted by CIA.

When $K = L$, the mean over i of G_i^E is $[C_2^K(G(X, X') + G(Y, Y')) + K^2 G(X, Y)]/C_2^{2K}$. Compared with the numerator of ψ^N: $[G(X, X') + G(Y, Y')]/2 = W$, the numerator of CIE is $[2C_2^K W + K^2 G(X, Y)]/C_2^{2K}$. Therefore,

$$\text{CIE} = [K(K-1)CIA + K^2]/C_2^{2K}.$$

Note that when $K = L$, $\text{CIE}_{\min} = K/(2K-1)$, and it is straightforward to show that CIEA = CIA. Similarly, we establish the equality between $\widehat{\text{CIEA}}$ and $\widehat{\text{CIA}}$. However, CIEA does not necessarily equal CIA when K and L are unequal.

7.2. Comparison of adjusted coefficient of individual equivalence with the concordance correlation coefficient

The CCC is commonly used for assessing agreement for continuous outcomes. It was first published by Lin [10] for the simplest case where there are two raters and each makes one reading per subject. Lin's CCC is defined as follows: assume that the observations (X, Y) are from a bivariate distribution with

mean vector (μ_x, μ_y) and variance–covariance matrix $\begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$. Lin's CCC is defined as

$$\begin{aligned} \text{CCC}_{\text{Lin}} &= 1 - \frac{E(X-Y)^2}{E[(X-Y)^2|\rho=0]} \\ &= \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2+(\mu_x-\mu_y)^2}, \end{aligned}$$

where ρ is the Pearson correlation coefficient between two observers.

Following the introduction of the CIEA for quantitative measurements, it is of interest to compare this coefficient with CCC, which shares the same denominator as CIE with $G = \text{MSD}$. Because we have replicated observations from each observer, we compare the CIEA with the total CCC [11] defined as

$$\text{CCC}_{\text{total}} = 1 - \frac{E(X_{ik}-Y_{il})^2}{E_I(X_{ik}-Y_{il})^2},$$

where E_I is the expectation given independence of X, Y. Barnhart *et al.* [12] compared the total CCC and CIA when the between-subject variability is increased and found that the CCC is inflated when the between-subject variability is large. We will follow the same strategy to compare the CIEA with the total CCC.

Let us use the simple latent class model introduced in Section 5. Obviously, all three MSDs increase with the between-subject variability σ_T^2. We are going to explore the dependence of total CCC, CIA, and CIEA on σ_T^2. In Section 5, we defined the distribution of T using the sample moments from the stenosis data, $\mu_T = 43.29$, $\sigma_T = 29.87$. To investigate the dependence of the coefficients on the between-subject variability, we now keep μ_T fixed at 43.29 but let σ_T vary from 0 to 60. The parameters defining the conditional means and standard deviations of the observers' measurements given T are $b = d = 1, e = 1.5, g = 1$, and $f = h = 0.3$. We keep $a = 0$ and let $c = 3.8, 16.3, 28.1$ to account for good, moderate, and poor agreement, respectively. From the previous section, we know that when $K = L$, CIA and CIEA are identical. Therefore, we consider $K = L = 3$ and $K = 2, L = 3$ as the two scenarios in our comparison.

In Figure 1, the total CCC, CIA, and CIEA were plotted while varying σ_T. With $K = L = 3$, we considered good, moderate, and poor agreement (cases (a), (b), and (c) in Figure 1, respectively). As we mentioned earlier when $K = L$, CIEA and CIA coincide. We also consider case (d) where $K = 2, L = 3$ for moderate agreement. In (a), where we had good agreement, both CIEA and CIA were constant with the increment in σ_T, whereas total CCC increased rapidly with σ_T. Similarly, when we had moderate and poor agreement (cases (b) and (c)), CIEA and CIA increased at a modest rate with σ_T, whereas total CCC kept increasing very rapidly. Furthermore, when $\sigma_T \geqslant 20$, total CCC was much higher than CIEA and CIA. When K and L are not equal, a similar trend was observed except that CIA was a little bit higher than CIEA (Figure 1(d)). Therefore, we believe that CCC is inflated when the between-subject variability is large, whereas CIEA and CIA are more stable.

8. Discussion

In this paper, we introduced the CIE for replicated quantitative measurements. CIE compares the observed disagreement with its expected value under individual equivalence, that is, when the probability density function of the readings of both observers on the same subject are identical. We introduced and validated both nonparametric and parametric estimation methods through a simulation study. We concluded that nonparametric approach always gives us robust estimation. The parametric method also works well and gives us smaller RMSE when the true values are normally distributed. We applied the new methods to the carotid stenosis data introduced in Section 2.

As a scaled index, CIE has the advantage of judging the degree of agreement on the basis of a standardized value. For other scaled agreement coefficients, such as introclass correlation coefficients (ICCs) and the concordance correlation coefficient (CCC), the values of the coefficients are not comparable across different populations, and sometimes artificially high or low agreement values may be obtained because of the dependence of those indices on the population heterogeneity. Both CIA and CIE are fundamentally different from ICCs and CCC. The CIA uses the within-subject, rather than between-subject, variability as the scaling factor. The CIE, on the other hand, introduces the idea of disagreement by chance under individual equivalence while sharing the same denominator as CIA for the observed disagreement.

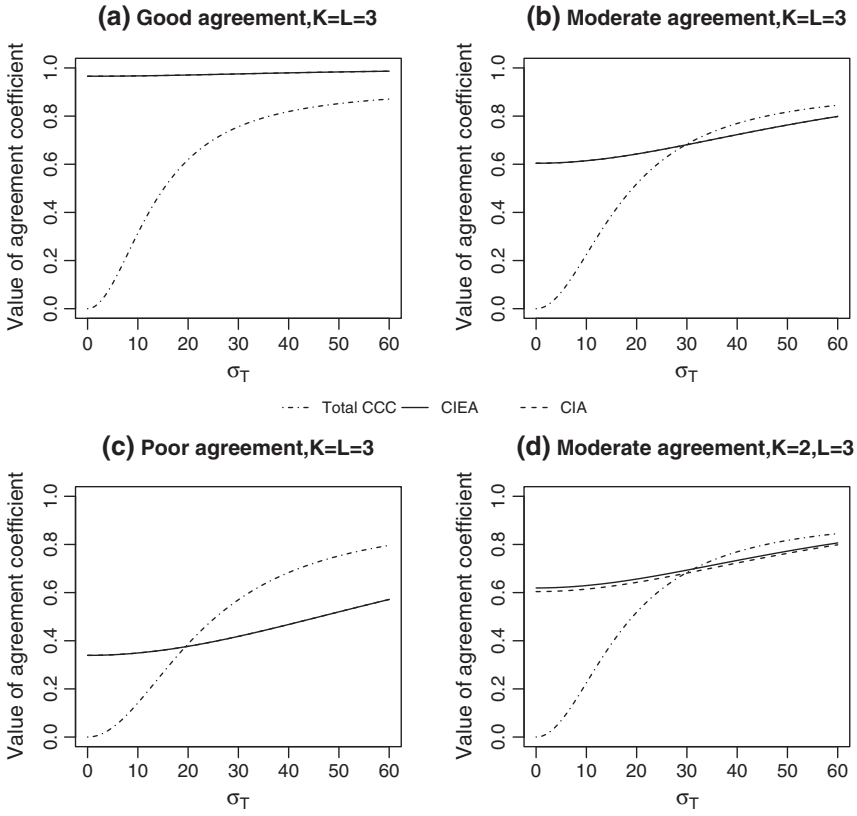

Figure 1. CCC, CIEA, and CIA with varying between-subject variability σ_T^2.

The CIA with MSD as the disagreement functions compares the observed disagreement with the expected when $E(X_i|T) = E(Y_i|T)$ for every subject i [2]. The CIE compares the observed disagreement with the expected when X and Y have the same conditional distribution (given T) for every subject, which is a stronger requirement. When we compare the definition of CIA with the expression for CIE involving the between-observer and within-observer disagreement (1), we see that the approaches differ by the weights that are assigned to the within-observer disagreements, $G(X, X')$ and $G(Y, Y')$. Basically, in CIA equal weights are assigned to $G(X, X')$ and $G(Y, Y')$, whereas in CIE, where we apply the permutation-based method, C_2^K and C_2^L are used as weights for $G(X, X')$ and $G(Y, Y')$, respectively. The comparison between CIEA and CIA and the relation among CIEA, CIA, and CCC have been discussed. We found that CIEA = CIA when $K = L$. However, unlike CIA, CIEA can be used when $K = 1$ or $L = 1$ such as the scenario when we do not have replicated measurements on a gold standard that is measured without error.

In general, we do not expect either $G(X, X')$ or $G(Y, Y')$ to exceed $G(X, Y)$, and thus CIE and CIEA are generally less than or equal to 1. However, in practice it is possible for the estimated value of $G(X, X')$ or $G(Y, Y')$ (or both) to exceed the estimated value of $G(X, Y)$. Therefore, in practice it is possible to have estimated CIE or CIEA greater than 1. In this case, we may set the estimates of CIE and CIEA to one. When the estimates of CIEA range from 0 to 1, it is a good idea to consider the square root arcsin transformation which can be used to stabilize the standard error and improve the normality. Through simulations (results not shown), we found that this transformation improved the coverage probabilities for moderate and poor agreement. However, for good agreement the estimated CIEA sometimes exceeded 1, and we redefined the estimate to 1 in these cases. This caused the coverage probabilities to be much lower than 0.95 in the 'good agreement' scenarios. Thus, we do not think that one should consider the square root arcsin transformation.

Furthermore, in this study we used the conditional version of CIEA given K and L. The dependence of CIEA on K and L results from the permutation-based approach. We observed that the dependency of CIEA on the number of replications seems to be a function of the ratio K/L. CIEA does not depend on K, L when the within-observer variances of both observers are the same (Figure 2(a)). In general, there is a difference about 0.02 in CIEA when the ratio of K/L increases from 0.1 to 1.0 when the variances

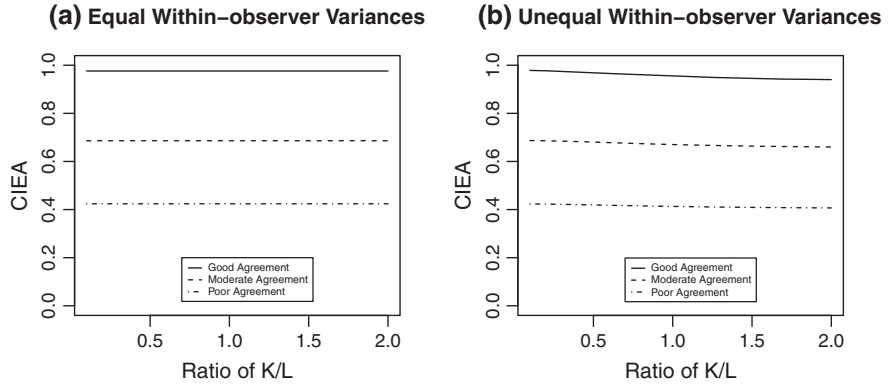

Figure 2. Dependence of CIEA on the ratio of K/L when $L = 1000$.

are different. As shown in Figure 2(b) ($L = 1000$), for all the good, moderate, and poor agreement scenarios, CIEA stays approximately constant across different values of K/L.

If the continuous data are dichotomized, for example at the mean, and the agreement is assessed on the basis of the dichotomized binary data, the CIEA [3] for the binary data may be smaller or larger than the original CIEA for continuous data. Specifically, the CIEA for binary data is similar to the CIEA for continuous data if the original CIEA for continuous data is high or moderate (results not shown).

Sample size calculations are essential in agreement studies because it is important to determine the number of subjects and the number of replications needed in order to achieve a desired precision when estimating CIEA. Gao(2010) [13] obtained the SE of CIA in terms of N, K, and L. One can get a similar expression for the SE of CIEA, so that the corresponding sample size of interest can be calculated. In addition, the definition and nonparametric estimation of CIE can be easily extended when the number of replications on each subject by the same observer is not fixed.

Future research on the CIE may involve (a) extension to more than two observers; (b) data with repeated observations by the same observer on the same subject where the subjects' true value may change, for example over time or under different conditions; and (c) data with missing observations.

Acknowledgements

The authors thank the two anonymous reviewers and the editor for suggestions that helped to improve this paper.

References

1. Hawkins DM. Diagnostics for conformity of paired quantitative measurements. *Statistics in Medicine* 2002; **21**:1913–1935.
2. Haber M, Barnhart HX. A general approach to evaluating agreement between two observers or methods of measurement. *Statistical Methods in Medical Rsearch* 2008; **17**:151–169.
3. Pan Y, Haber M, Barnhart HX. A new permutation-based method for assessing agreement between two observers making replicated binary readings. *Statistics in Medicine* 2011; **30**:839–853.
4. Barnhart HX, Williamson JM. Modeling concordance correlation via GEE to evaluate reproducibility. *Biometrics* 2001; **57**:931–940.
5. Haber M, Gao J, Barnhart HX. Assessing observer agreement in studies involving replicated binary observations. *Journal of Biopharmaceutical Statistics* 2007; **17**:757–766.
6. Haber M, Barnhart HX. Coefficients of agreement for fixed observers. *Statistical Methods in Medical Research* 2006; **15**:255–271.
7. Barnhart HX, Haber M, Kosinski AS. Assessing individual agreement. *Journal of Biopharmaceutical Statistics* 2007; **17**:697–719.
8. Pan Y, Gao J, Haber M, Barnhart HX. Estimation of coefficients of individual agreement (CIAs) for quantitative and binary data using SAS and R. *Computer Methods and Programs in Biomedicine* 2010; **98**:214–219.
9. Haber M, Gao J, Barnhart H. Evaluation of agreement between measurement methods from data with matched repeated measurements via the coefficient of individual agreement. *Journal of Data Science* 2010; **8**:457–469.
10. Lin LI. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 1989; **45**:255–268.
11. Barnhart HX, Song J, Haber M. Assessing intra, inter, and total agreement with replicated measurements. *Statistics in Medicine* 2005; **24**:1371–1384.

12. Barnhart HX, Haber M, Lokhnygina Y, Kosinski AS. Comparison of concordance correlation coefficient and coefficient of individual agreement in assessing agreement. *Journal of Biopharmaceutical Statistics* 2007; **17**:721–738.
13. Gao J. Assessing Observer Agreement for Categorical Observations. *PhD Dissertation*, Department of Biostatistics and Bioinformatics, Emory University, 2010.

APPENDIX A.

We now show that the definition of individual equivalence, $f_X(u|i) = f_Y(u|i)$ for all u, i, is equivalent to the equality of the probabilities of all the assignments of K $X's$ and L $Y's$ to the $K + L$ observations on subject i. It is clear that if the conditional distributions are identical then all the assignments are equally likely. To show the converse, let $x_1, \cdots, x_K, y_1, \cdots, y_L$ be a possible outcome on subject i (we omit the index i to simplify the notation). The likelihood (conditioning on i) of this outcome is

$$L_1 = f_X(x_1) \cdot f_X(x_2) \cdots f_X(x_k) \cdot f_Y(y_1) \cdot f_Y(y_2) \cdots f_Y(y_L).$$

Now suppose the x_1 is assigned to Y and y_1 is assigned to X, making no changes with the remaining observations. The likelihood of the new outcome is

$$L_2 = f_Y(x_1) \cdot f_X(x_2) \cdots f_X(x_k) \cdot f_X(y_1) \cdot f_Y(y_2) \cdots f_Y(y_L).$$

Under our assumption $L_1 = L_2$, hence

$$f_X(x_1) \cdot f_Y(y_1) = f_Y(x_1) \cdot f_X(y_1).$$

If we integrate over y_1, we get $f_X(x_1) = f_Y(x_1)$. Because this holds for any possible value x_1, we have shown that f_X and f_Y are identical when conditioning on i.