Evaluation of Agreement between Measurement Methods from Data with Matched Repeated Measurements via the Coefficient of Individual Agreement

Michael Haber (1), Jingjing Gao (1) and Huiman X Barnhart (2)

(1) Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA, U.S.A

(2) Department of Biostatistics and Bioinformatics, Duke Clinical Research Institute, Duke University, Durham, NC, U.S.A

Correspondence author: Dr. Michael J. Haber, Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta GA 30322, U.S.A. Tel: (404)727-7698. e-mail: mhaber@sph.emory.edu

Revised February 12, 2009

Summary

We propose a simple method for evaluating agreement between methods of measurement when the measured variable is continuous and the data consists of matched repeated observations made with the same method under different conditions. The conditions may represent different time points, raters, laboratories, treatments, etc. Our approach allows the values of the measured variable and the magnitude of disagreement to vary across the conditions. The coefficient of individual agreement (CIA), which is based on the comparison of the between and within-methods mean squared deviation (MSD) is used to quantify the magnitude of agreement between measurement methods. The new approach is illustrated via two examples from studies designed to compare (a) methods of evaluating carotid stenosis and (b) methods of measuring percent body fat.

Key words: coefficient of individual agreement; method comparisons; repeated measurements.

1. Introduction

In studies designed to assess the agreement between methods of measurement, multiple observation are often made with each method on the same subject. These observations can be considered as *replicated* measurements if the observations with the same method on the same subject are conditionally independent and identically distributed. In this case it is assumed that the subject's true value of the measured quantity remains unchanged across the measurements made by the same method. On the other hand, agreement studies may be designed such that multiple matched observation with two (or more) methods are conducted on each subject under specific 'conditions' where the subject's true value may change across conditions. The observations are then considered as *matched repeated* measurements. The 'conditions' may correspond to different time points, raters, laboratories, devices, treatments, etc. For example, in a study designed to compare imaging methods for assessing carotid stenosis (Barnhart and Williamson, 2007) the same three raters used each of the imaging methods to determine the carotid stenosis of each patient. Here the three raters correspond to three 'conditions' under which measurements have been made. Chinchilli et al. (1996), Choudhary (2008), and King et al. (2007a,b) analyzed data from a study in which percentage body fat was estimated using two methods: (1) skinfold calipers, and (2) dual energy x-ray absorptiometry (DEXA), on adolescent girls. Measurements were taken in an initial visit at age 12 years and in subsequent visits which occurred every six months. In this case the 'condition' is the girl's age.

The focus of this article is on evaluation of agreement between methods of measurements from matched repeated observations. We assume that all the measurements are made on the same interval scale, hence we can evaluate the extent of agreement between methods via the differences between measurements made on the same subject with different methods. In addition, we assume that a subject's true value may change across the levels of the variable corresponding to the conditions, and that the magnitude of agreement between methods may vary across conditions. We are interested in (a) assessment of

condition-specific agreement between measurement methods, (b) investigating the effect of the condition on the magnitude of agreement between methods, and (c) if we conclude that agreement between methods remains unchanged across conditions then we also may be interested in an overall measure of the extent of agreement. We are not interested in the agreement between measurements taken under different conditions as the true value of the measured variable on a subject may vary across the conditions. In the carotid stenosis example, the main interest is in comparing the imaging methods when used by the same rater. We do not investigate the agreement between the raters in this example. In the body fat example, one is mainly interested in the agreement between the skinfold calipers and DEXA measured on the same girl in the same visit.

As stated in a recent review paper by Barnhart et al. (2007a), future research is needed on assessing agreement with repeated measurements because previous works on this topic have been limited to scaled agreement indices using the concordance correlation coefficient (CCC) (Chinchilli et al. (1996) and King et al. (2007a,b)), unscaled agreement indices using the total deviation index (TDI) (Choudhary, 2008), and limits of agreement (LOA) (Bland and Altman, (2007)). In this work we focus on an alternative scaled index for assessing agreement, the *coefficient of individual agreement* (CIA), that may be preferable to the CCC because it does not depend on the between-subject variability, as elaborated by Barnhart et al. (2007a,b). The CIA has been introduced by Barnhart et al. (2007c), and Haber and Barnhart (2008), and has been applied to data with *replicated* measurements only. In this work we will show how to estimate the CIA from data with *matched repeated* measurements across conditions when there are no replications at each condition. If there are replications at each condition, we can accomplish goals (a) – (c) by applying the methods described in Barnhart et al. (2007c) and Haber and Barnhart (2008). However, in this work we assume that there is a single observation for each method$\times$condition combination, so that our previous methods (Barnhart et al. (2007c) and Haber and Barnhart (2008)) cannot be used. In general, the CIA compares the disagreement between methods to the disagreement between replicated measurements made by the same method on the same study subject. The agreement between methods is considered acceptable if the variability between observations made with *different*

methods on the same subject is not much larger than the variability between observations with the *same* method on this subject. Hence, good individual agreement implies that replacing one method by another or using the methods interchangeably does not substantially increase the within-subject variability. The reciprocal of the CIA is interpreted as the relative increase in the variability of the measurements made on the same subject if the methods were used interchangeably. In our previous papers (Barnhart et al. (2007c) and Haber and Barnhart (2008)) we suggested that the CIA should be at least 0.8 in order to claim 'good' agreement. This means that using the measurement methods interchangeably does not increase the variability of measurements made on the same subject by more than 25%.

The CCC and CIA are *scaled* agreement indices attaining values in the intervals [-1,1] and [0,1], respectively. The CCC is based on the comparison of the between-methods and the between-subjects variability Hence it depends on the heterogeneity of the population with respect to the measured variable (Atkinson and Nevill (1997), Barnhart et al. (2007b)) and therefore comparisons of CCCs from different studies may not be valid. The CIA, on the other hand, uses the within-methods variability, σ_e^2, as a benchmark to which between-methods variability is compared. In our opinion, the latter is a more appropriate comparison as the within-methods disagreement is related to the performance of the measurement methods, while the between-subjects variability does not reflect any aspect of the measurement process and may vary between populations or samples. A detailed comparison of the two types of scaled agreement coefficients can be found in Barnhart et al. (2007b). Alternatively, one may use an *unscaled* measure of agreement, such as the total deviation index (Choudhary (2008), Lin et al. (2002)). Using an unscaled agreement index requires setting acceptable bound that may not be easy in practice. A thorough review of different approaches, including CCC, CIA and TDI, to evaluation of agreement between observers or measurement methods can be found in Barnhart et al. (2007a).

The key concept in the CIA is the use of the variability between readings of the same method on the same subject as a reference for assessing the disagreement between

different methods. First, one must make sure that this within-method (error) variability, σ_e^2, is 'reasonably small'. Barnhart et al (2007b) suggested to compute the repeatability coefficient (Bland and Altman (1999)), $1.96\sqrt{2\sigma_e^2}$, and check whether it is less than or equal to an acceptable value within which the difference between two readings by the same method should lie for 95% of the subjects. Second, as illustrated in our previous papers (Barnhart et al. (2007c) and Haber and Barnhart (2008)), the within-method variability can be estimated if there are true replications. Those papers did not address the issue of estimating σ_e^2 when there are no replications. The main purpose this paper is to use the repeated measurements in order to estimate σ_e^2, and thus to estimate CIA, by fitting a reasonable model using matched repeated measures in the absence of replications.

In our previous papers (Barnhart et al. (2007c) and Haber and Barnhart (2008)) we considered two situations: (1) one of the methods of measurement is considered a reference, or gold standard, to which the other method is compared, and (2) none of the methods is considered as a reference. In this work we focus on the second situation. We assume that the magnitude of agreement is measured by the mean squared deviation (MSD), defined as the mean of the squared difference between two readings made on the same subject under the same condition. For the sake of simplicity, we first present the new statistical techniques in the context of assessing the agreement between *two* measurement methods and later show how this approach can be extended to the case of multiple methods. The models and methods for the case where the 'conditions' correspond to the levels of a categorical factor, such as raters or laboratories, are described and illustrated in Section 2. In section 3 we consider the case where the factor representing the 'conditions' is continuous, such as time, age or temperature. Section 4 presents generalizations to the case of more than two measurement methods.

2. Conditions correspond to the levels of a categorical factor

In this Section we consider the case where each of N subjects is evaluated by two measurement methods under the same K ($K \geq 2$) conditions. As stated in the introduction, the 'conditions' may correspond to different time points, laboratories, raters,

treatments, etc. We assume that the observed variable is continuous and that the true value of this variable on a given subject may change from one condition to another. We denote the measurements with the two methods by Y_1 and Y_2. The disagreement between the methods is quantified by the mean squared deviation (MSD), defined as:

$$MSD(Y_1, Y_2) = E(Y_1 - Y_2)^2,$$

where the expectation is over all the study subjects. The coefficients of individual agreement (see Barnhart et al. (2007c) and Haber and Barnhart (2008)) compare $MSD(Y_1, Y_2)$ to the MSD of two replicated observations made with same method under the same conditions. Therefore we denote by $MSD(Y_j, Y_j')$ the mean squared deviation between two (hypothetical) replicated observations made with method j ($j = 1,2$) under the same condition. For the case where none of the methods is considered as a reference, the coefficient of individual agreement is defined as:

$$\psi = \frac{[MSD(Y_1, Y_1') + MSD(Y_2, Y_2')]/2}{MSD(Y_1, Y_2)}. \qquad (1)$$

In our previous papers (Barnhart et al. (2007c) and Haber and Barnhart (2008)) this coefficient was denoted by ψ^N.

Since the data considered here do not include replicated observations, Y_j and Y_j', made with same method on the same subject under the same condition, we cannot apply the approach of Barnhart et al. (2007c) and Haber and Barnhart (2008), who used the replication variances for estimation of $MSD(Y_j, Y_j')$ ($j = 1,2$). Instead, we propose to estimate $MSD(Y_j, Y_j')$ from a simple linear model. Denote by Y_{ijk} the observations with the j-th method on the i-th subject under the k-th conditions. In order to estimate these MSD's, we use the following mixed ANOVA model:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + e_{ijk} \quad (i = 1,...,n,\ j = 1,2,\ k = 1,...,m)$$

The α's are the subjects' random effects while the β's and γ's are the fixed effects of the methods and the conditions, respectively. We assume that the random main effects, interactions and errors are independent and normally distributed with mean 0 and

$Var(\alpha_i) = \sigma_\alpha^2$, $Var[(\alpha\beta)_{ij}] = \sigma_{\alpha\beta}^2$, $Var[(\alpha\gamma)_{ik}] = \sigma_{\alpha\gamma}^2$, $Var(e_{ijk}) = \sigma_e^2$. Regarding the fixed effects, we make the common assumption that the sum of the coefficients over every index is zero, i.e., $\sum_j \beta_j = \sum_k \gamma_k = \sum_j (\beta\gamma)_{jk} = \sum_k (\beta\gamma)_{jk} = 0$.

It is important to note that this model allows the measurements Y_{ijk} for the same subject-method combination (i, j) to vary across the m conditions. If we consider two (hypothetical) replicated observations, Y_j and Y_j', that could be made by method j on the same subject under the same condition then:

$$MSD(Y_j, Y_j') = E(Y_{ijk} - Y_{ijk}')^2 = 2\sigma_e^2 \quad (j = 1,2)$$

From the above model it is evident that the disagreement between the two observers may depend on the condition. The $MSD(Y_1, Y_2)$ for the k-th condition can be obtained from the parameters of our model as follows:

$$MSD_k(Y_1, Y_2) = E(Y_{i1k} - Y_{i2k})^2 = [(\beta_1 - \beta_2) + ((\beta\gamma)_{1k} - (\beta\gamma)_{2k})]^2 + 2\sigma_{\alpha\beta}^2 + 2\sigma_e^2.$$

Using the definition (1) we now can obtain the coefficient of individual agreement under the k-th condition as:

$$\psi_k = \frac{MSD(Y_j, Y_j')}{MSD_k(Y_1, Y_2)} = \frac{2\sigma_e^2}{[(\beta_1 - \beta_2) + ((\beta\gamma)_{1k} - (\beta\gamma)_{2k})]^2 + 2\sigma_{\alpha\beta}^2 + 2\sigma_e^2}$$

Estimation and testing

Fitting the mixed model that we use to estimate the coefficients of individual agreement can be done via standard statistical software packages. We used SAS proc MIXED for this purpose. It may also be of interest to test the hypotheses of homogeneous agreement, $\psi_1 = ... = \psi_m$, which is equivalent to $(\beta\gamma)_{j1} = ... = (\beta\gamma)_{jm}$ for $j = 1,2$. If this hypothesis is supported by the data then the common value of all the condition-specific ψ's can be estimated by fitting the simpler form of the mixed model which does not include the method-by-condition interaction terms $(\beta\gamma)$. Confidence intervals for the

estimated coefficients can be computed using the delta method or the nonparametric bootstrap.

Example 1

We now illustrate the method using data from a carotid stenosis screening study. The goal of the study was to compare magnetic resonance angiography (MRA) for noninvasive screening of carotid artery stenosis with invasive intra-arterial angiogram (IA). Two MRA methods were considered: two-dimensional time of flight (MRA-2D) and three-dimensional time of flight (MRA-3D). Each of three raters determined the percent of carotid stenosis using each of the three imaging methods. Thus, a total of nine observations were made on each study subject. Our analysis is based on the 55 study subjects for whom all 9 readings were available. Percent stenosis was measured in both the left and right carotid artery of each subject. We will use here only the data from the left arteries. For more information on the study, including graphical displays of agreement between methods and between raters, the reader is referred to Barnhart and Williamson (2001). The stenosis data can be copied from: www.sph.emory.edu/observeragreement/

Barnhart et al. (2007c) used this data to estimate the coefficients of individual agreement between the three methods where the raters were consider as independent replications. Here we re-estimate the coefficients under the more realistic assumption that each rater has her/his own effect on the observed measurements. Thus, we consider the raters as 'conditions'.

Table 1 presents rater-specific estimates of the CIA's for the left artery data, along with their delta-method-based 95% confidence intervals, for all three pairs of methods. The table also presents the overall estimate of ψ under the assumption that the coefficients for the three raters are equal. The overall estimates can be interpreted as pooled (or summary) estimates of the coefficients across the three raters under the assumption that the disagreement between methods is homogeneous. These pooled estimates are not very

meaningful unless the differences between methods are indeed homogeneous across raters. In Table 1, whenever the upper limit of a CI exceeded 1, it was set to 1.000.

As we stated in the Introduction, it is important to check the repeatability coefficient $1.96\sqrt{2\sigma_e^2}$ for each of the methods. In the context of the present example, this coefficient is a 95% upper bound for the absolute difference of two readings made by the same rater with the same imaging method. The coefficient should be relatively small, so that we feel comfortable when using the *intra*-method variability as a reference to which we compare the *inter*-method variability. The repeatability coefficients corresponding to the three comparisons in Table 1 are 51.5, 49.0 and 63.0 percent, respectively, which are likely to be higher than acceptable values for the absolute difference of two measurements of carotid stenosis performed with the same method on the same patient. Hence, from a practical point of view the estimates in Table 1 are likely to overestimate the actual magnitude of individual agreement.

From Table 1 we can learn that the agreement between the IA method and each of the MRA methods, which was the focus of the original study, is very poor. The comparison of the two MRA methods produces higher estimates of CIA's, in the range 0.81-0.86. However, since we saw in the previous paragraph that these estimates are likely to be inflated due to an unacceptable repeatability coefficient, one may doubt whether the agreement between the two MRA methods is indeed reasonably good.

3. Conditions correspond to a continuous factor

In this Section we assume that matched repeated measurements are performed under conditions that correspond to the values of a continuous variable. The most common situation involves measurements made at different time points, hence we will refer to the variable defining the repeated measurement as 'time' and assume that the subjects' true values are a linear function of time.

Suppose that pairs of observations $(Y_{i1}(t), Y_{i2}(t))$ were made with two methods of measurement on subject i at each of $m_i \geq 2$ different time points, t. (These time points do not have to be the same for all subjects). As we did in Section 2, we begin by fitting a linear mixed model to the observed measurements:

$$Y_{ij}(t) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \gamma t + \delta_i t + \eta_j t + e_{ij}(t) \quad (i = 1,...,n,\ j = 1,2)$$

As before, the random effects $\{\alpha_i\}, \{(\alpha\beta)_{ij}\}, \{\delta_i\}, \{e_{ij}(t)\}$ are independent with zero means and: $Var(\alpha_i) = \sigma_\alpha^2$, $Var[(\alpha\beta)_{ij}] = \sigma_{\alpha\beta}^2$, $Var[(\delta_i] = \sigma_\delta^2$, $Var(e_{ij}(t)) = \sigma_e^2$. For the fixed effects we set $\beta_1 + \beta_2 = \eta_1 + \eta_2 = 0$. The mean squared differences are as follows:

$$MSD(Y_j, Y_j') = E_i(Y_{ij}(t) - Y_{ij}'(t))^2 = 2\sigma_e^2 \quad (j = 1,2)$$

$$MSD(Y_1, Y_2 \mid t) = E_i(Y_{i1}(t) - Y_{i2}(t))^2 = [(\beta_1 - \beta_2) + (\eta_1 - \eta_2)t]^2 + 2\sigma_{\alpha\beta}^2 + 2\sigma_e^2$$

W now can obtain the CIA as a function of time as follows:

$$\psi(t) = \frac{MSD(Y_j, Y_j')}{MSD(Y_1, Y_2 \mid t))} = \frac{2\sigma_e^2}{[(\beta_1 - \beta_2) + (\eta_1 - \eta_2)t]^2 + 2\sigma_{\alpha\beta}^2 + 2\sigma_e^2}.$$

Proc MIXED in SAS can again be used to estimate the parameters in the mixed model and provide an estimate of the function $\psi(t)$. The hypotheses $\eta_1 = \eta_2$ can be tested in order to check whether the CIA does not change significantly over time.

Example 2

In the Young Women Health Study (Lloyd et al., (1993)) percentage body fat was estimated using skinfold calipers and dual energy x-ray absorptiometry (DEXA) on a cohort of adolescent girls. Skinfold caliper and DEXA measurements were made in an initial visit, at age 12 years, and in eight subsequent visits, which occurred every six months. Agreement between the two methods of measurements has been evaluated via the concordance correlation coefficient (CCC) (Chinchilli et al., (1996), King et al., (2007a,b)) and via the total deviation index (TDI) (Choudhary, 2008). Here we estimate the coefficients of individual agreement, using observation from 651 visits of 91 girls. We will use a girl's actual age as the 'condition' (t) since the visits did not occur exactly at ages 12.0, 12.5, 13.0 etc.

Fitting the model to this data yields the following estimates: $\hat{\sigma}_{\alpha}^2 = 6.8553$, $\hat{\sigma}_{\alpha\beta}^2 = 2.4709$, $\hat{\sigma}_{\delta}^2 = 0.01987$, $\hat{\sigma}_{e}^2 = 3.0566$, $\hat{\beta}_1 = -9.3808$, $\hat{\gamma} = -0.2546$, $\hat{\eta}_1 = 0.6074$. The t statistic for the hypothesis $\eta_1 = 0$ is 14.9, hence the data do not support the hypothesis of a time-independent CIA. The repeatability coefficient is 4.8, which can be considered an acceptable 95% bounds for the within-methods error.

Using the above estimates we can write the estimated function $\psi(t)$:

$$\hat{\psi}(t) = \frac{6.1132}{(-18.7616 + 1.2149t)^2 + 11.0550}.$$

Figure 1 displays the estimated coefficients along with their delta-method-based CI's for 12-16 years old girls, which is the range of ages in the data. We see that agreement between the two methods improves with age up to 15.5 years. As stated in the introduction, we suggested that agreement be considered 'acceptable' only if the relevant coefficient of individual agreement exceeds 0.8 (Barnhart et al. (2007c), Haber and Barnhart (2008)). Since the estimates of the CIA remain below 0.6 and their upper CI's remain below 0.8, we conclude that the agreement between the DEXA and the skinfold calipers is not acceptable for girls aged 12-16 years. For comparison, Chinchilly et al. (1996) reported an estimated CCC of 0.42 for this data (their method does not assume that agreement may change with age). King et al. (2007a,b) used only the data from the first three visits of each girl and reported values in the range 0.48-0.67 for their weighted repeated measurements CCC. Choudhary (2008), who analyzed the full dataset using a tolerance interval approach, concluded that 'the agreement between the methods appears best around age 15-17', and that 'on the whole, the agreement between the skinfold and DEXA methods does not seem good enough to justify their interchangeable use'. These conclusion are similar to ours.

4. The case of more than two methods of measurement

When there are more than two measurement methods, the overall coefficients of individual agreement can be obtained from the pairwise MSD's as shown in Barnhart et al. (2007c). Denote the observations made with $J \geq 3$ methods $Y_1, Y_2, ..., Y_J$. When the

conditions correspond to the levels (k) of a categorical factor, an overall coefficient of individual agreement for the $k-$th condition is:

$$\psi_k = \frac{Mean_{1\le j\le J}[MSD(Y_j, Y_j^{'})]}{Mean_{1\le j<j'\le J}[MSD_k(Y_j, Y_{j'})]}$$

Where $MSD(Y_j, Y_j^{'})$ is the mean squared deviation between two replicated observations made by method j under the same condition and $MSD_k(Y_j, Y_{j'})$ is the mean squared deviation between measurements by methods j and j' under the $k-$th condition.

5. Discussion

We presented a simple method for assessing agreement between two or more methods of measurement based on repeated measurements matched on a factor whose levels are considered as conditions. We advocate the use of the coefficient of individual agreement rather than the concordance correlation coefficient, as the latter depends on the between-subjects heterogeneity (Atkinson and Nevill (1997), Barnhart et al. (2007b), Haber and Barnhart (2008)). Our approach allows the true values of the measured variable and the magnitude of disagreement to vary across conditions or over time.

We use the terms 'methods' and 'conditions' broadly here. For example, in the carotid stenosis study (Example 1) we considered the imaging methods as 'methods' and the human raters as 'conditions' because we were interested is the agreement between the imaging methods based on readings by the same rater. Alternatively, we could treat the raters as 'methods' and the imaging methods as 'conditions' and assess the agreement between raters when they are using the same imaging method.

We used SAS Proc MIXED, which assumes that all the measurements are normally distributed, for the analyses of the data in Examples 1 and 2. The SAS codes are available from the first author. It is important to note that the CIA's can be estimated using the method of moments from the various ANOVA mean squares without making the normality assumption. We also wrote R programs for the analysis of the carotid

stenosis and the body fat data. These programs are available at XXX and can be used by readers who do not have SAS.

The coefficients of individual agreement can also be defined and estimated when the observations are binary (Haber et al. (2007)). The methods introduced in this work can also be applied to repeated binary data, for example by using generalized linear mixed models.

Acknowledgement

We wish to thank Dr. Tonya King for providing us the body fat data. We also thank a reviewer for helpful comments. This research was supported by NIH grant R01-MH070028 and was partially supported by grant UL1 RR024128 (for Huiman X. Barnhart).

References

Atkinson, G. and Nevill, A. (1997). Comment on the use of concordance correlation to assess the agreement between variables. *Biometrics* **53**, 775-778.

Barnhart, H. X., Haber, M., and Lin, L. (2007a). An overview on assessing agreement with continuous measurements. *Journal of Biopharmaceutical Statistics* **17**, 529-569.

Barnhart, H. X., Haber, M., Lokhnygina, Y., and Kosinski A. S. (2007b). Comparison of concordance correlation coefficient and coefficient of individual agreement in assessing agreement. *Journal of Biopharmaceutical Statistics* **17**, 721-738.

Barnhart, H. X., Kosinski, A. S., and Haber, M. (2007c). Assessing individual agreement. *Journal of Biopharmaceutical Statistics* 17**,** 697-719.

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

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

Bland, J. M., and Altman, D. G. (2007). Agreement between methods of measurement with multiple observations per individual. *Journal of Biopharmaceutical Statistics* **17,** 571-582.

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

Choudhary, P. K. (2008). A tolerance interval approach for assessment of agreement in method comparison studies with repeated measurements. *Journal of Statistical Planning and inference* **138**, 1102-1115.

Haber, M. and Barnhart, H. X. (2008). A general approach to evaluating agreement between two observers or methods of measurement. *Statistical Methods in Medical Reseach* **17**, 151-169.

Haber, M., Gao, J., and Barnhart, H. X. (2007). Assessing observer agreement in studies involving replicated binary data. *Journal of Biopharmaceutical Statistics* **17**, 757-766.

King, T. S., Chinchilli, V. M., and Carrasco, J. L. (2007a). A repeated measures concordance correlation coefficient. *Statistics in Medicine* 16, 3096-3113.

King, T. S., Chinchilli, V. M., Carrasco, J. L., and Wang, K. (2007b). A class of repeated measures concordance correlation coefficients. *Journal of Biopharmaceutical Statistics* **17,** 653-672.

Lin, L., Hedayat, A. S., Sinha, B., and Yang, M. (2002). Statistical methods in assessing agreement: models, issues and tools. *Journal of the American Statistical Association* **97,** 257-70.

Lloyd, T., Andon, M. B., Rollings, M., Martel, J. K., Landis, J. R., Demers, L. M., Eggli, D. F., Kieselhorst, K., and Kulin, H. E. (1993). Calcium supplementation and bone mineral density in adolescent girls. *Journal of the American Medical Association* **270**, 841-844.

Table 1: Estimated coefficients of individual agreement for carotid stenosis (left artery) data with confidence intervals based on the delta method.

Comparison 1: Y_1 = IA, Y_2 = MRA-2D

	$\hat{\psi}$	95% CI
Rater 1	0.547	(0.373, 0.722)
Rater 2	0.555	(0.383, 0.727)
Rater 3	0.588	(0.435, 0.741)
Overall*	0.581	(0.430, 0.733)

* Assuming no differences among raters: *p*-value for $\psi_1 = \psi_2 = \psi_3$ is 0.09.

Comparison 2: Y_1 = IA, Y_2 = MRA-3D

	$\hat{\psi}$	95% CI
Rater 1	0.415	(0.265, 0.565)
Rater 2	0.432	(0.284, 0.580)
Rater 3	0.441	(0.303, 0.580)
Overall*	0.427	(0.294, 0.559)

* Assuming no differences among raters: *p*-value for $\psi_1 = \psi_2 = \psi_3$ is 0.66.

Comparison 3: Y_1 = MRA-2D, Y_2 = MRA-3D

	$\hat{\psi}$	95% CI
Rater 1	0.861	(0.692, 1.000)
Rater 2	0.866	(0.707, 1.000)
Rater 3	0.815	(0.640, 0.989)
Overall*	0.852	(0.696, 1.000)

* Assuming no differences among raters: *p*-value for $\psi_1 = \psi_2 = \psi_3$ is 0.46.

Figure 1: Estimated coefficients of individual agreement for body fat data with 95% confidence intervals based on the delta method