Published in final edited form as:
Comput Methods Programs Biomed. 2010 May ; 98(2): 214–219. doi:10.1016/j.cmpb.2009.12.002.

Estimation of Coefficients of Individual Agreement (CIA's) for Quantitative and Binary Data using SAS and R

Yi Pan,
Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, 1518 Clifton Road NE, Atlanta, GA, 30322

Jingjing Gao,
Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, 1518 Clifton Road NE, Atlanta, GA, 30322

Michael Haber, and
Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, 1518 Clifton Road NE, Atlanta, GA, 30322

Huiman X. Barnhart
Department of Biostatistics and Bioinformatics and Duke Clinical Research Institute, Duke University, PO Box 17969, Durham, NC 27715

Abstract

The coefficients of individual agreement (CIA's), which are based on the ratio of the intra-and inter-observer disagreement, provide a general approach for evaluating agreement between two fixed methods of measurements or human observers. In this paper, programs in both SAS and R are presented for estimation of the CIA's between two observers with quantitative or binary measurements. A detailed illustration of the computations, macro variable definitions, input and output for the SAS and R programs are also included in the text. The programs provide estimations of CIA's, their standard errors as well as confidence intervals, for the cases with or without a reference method. Data from a carotid stenosis screening study is used as an example of quantitative measurements. Data from a study involving the evaluation of mammograms by ten radiologists is used to illustrate a binary data example.

Keywords

Agreement; Coefficient of individual agreement; Macro; Mean Squared Deviation

1. Introduction

In medical and other related sciences, many statistical approaches have been proposed for assessing agreement among observers or measurement methods. In a recent review paper,

Name, full address, and telephone number of the author for correspondence: Michael Haber, Ph.D, Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, 1518 Clifton Road NE, Atlanta, GA, 30322, mhaber@sph.emory.edu, tel: (404) 727-7698, fax: (404) 727-1370.

Barnhart et al. [1] classified the existing methods for evaluation agreement as follows: (1) descriptive tools, such as descriptive statistics and plots, (2) unscaled agreement indices, such as mean squared deviation (MSD), coverage probability (CP) and total deviation index (TDI) [2,3,4], and (3) scaled agreement indices.

Among the scaled agreement indices, the intraclass correlation coefficient (ICC) [5,6] and the concordance correlation coefficient (CCC) [7,8,9,10] are the most popular. Under certain conditions, the CCC is equivalent to one version of the ICC. Specifically, if the ANOVA model assumptions are satisfied, the CCC reduces to the agreement ICC defined by this ANOVA model [1,6,9].

The CCC is based on comparing the mean squared deviation (MSD) [4] to its value under independence. However, independence and disagreement are two different concepts [11]. Furthermore, the CCC depends on the between-subject variability. Atkinson and Nevill [12] pointed out that an increase in the between-subject variability results in a larger value of CCC even if the individual differences between measurements by the two methods remain the same. Barnhart et al. [13] also showed that the CCC depends on the between-subject variability due to the fact that it is scaled relative to the maximum disagreement defined as the expected MSD under independence. The ICC and the CCC are originally defined for quantitative data. However, these coefficients have been shown to be equivalent to the weighted Kappa for categorical data [14,4]. In addition, Shoukri [15] and King and Chinchilli [8] also defined ICC and CCC for qualitative data.

Haber, Barnhart and colleagues [16,17,18] introduced the coefficients of individual agreement (CIA's), which are scaled relative to an acceptable disagreement, with the goal of establishing interchangeability of observers. An 'acceptable disagreement' requires that the differences between measurements of different observers are similar to the differences between replicated measurements of the same observer. The concept of individual agreement is derived from the idea of individual bioequivalence in bioequivalence studies [17,19,20,21]. Similar agreement indices have been proposed by Haber et al. [22] and Shao and Zhong [23]. The CIA's compare differences between measurements from different observers to the differences of replicated measurements of the same observer. Therefore, they require replications which allow us to estimate the within-observer variability. The numbers of replications can be different across subjects and observers.

It is recommended to assess the intra-observer agreement before using CIA's because if the intra-observer agreement is not as good as expected, then any conclusion of inter-observer agreement may not be reliable. To confirm the reasonable intra-observer agreement, repeatability coefficient, introduced by Bland and Altman [24], can be calculated for each observer. Let X_1 and X_2 be two readings made by the same observer on the same subject.

The repeatability coefficient, defined as $c=1.96\sqrt{2\sigma_w^2}$, where σ_w^2 is the within-observer variance, satisfies $P(|X_1 - X_2| \leq c)=0.95$. Estimation of σ_w^2 can be done using the MSE from a one-way ANOVA where subject is the factor.

Previous papers used different computational approaches in estimating CIA and making inference for continuous and categorical data. In this paper, we present a unified non-parametric approach for estimation of CIA's with and without a reference for both continuous and categorical data by using a SAS macro and an R function. The programs also provide estimates for the standard errors of the estimated CIA's, as well as confidence intervals. Computational methods and theory, as well as the estimation of CIA's and the standard errors, are introduced in the next section. The details of the SAS and R programs

are described in section 3. Two examples are included to illustrate the estimation of CIA's in section 4. A brief summary follows in section 5.

2. Method

2.1 General Definition

Haber and Barnhart [16] considered the CIA's for the case of two observers, a continuous measured variable and a general disagreement function. Denote the readings of the two observers by *X* and *Y*. A disagreement function on *G(X,Y)* must satisfy (a) *G(X,Y)* ≥ 0, and (b) *G(X,Y)* increases as the disagreement between *X* and *Y* increases, according to a specific criterion. The agreement between *X* and *Y* is 'acceptable' if between and within observer disagreement function are similar, i.e., if *G(X,Y)* ≈ *G*(*X*, *X*') and *G(X,Y)* ≈ *G*(*Y*,*Y*'), where *G*(*X*, *X*') is the disagreement between two replicated readings made by observer *X* and *G*(*Y*,*Y*') is similarly defined for observer *Y*. Therefore, the estimation of the new coefficients requires replicated observations made by the same observer on the same subject. It is implicitly assumed that the extent of intra-observer agreement is acceptable for both *X* and *Y*.

When neither *X* or *Y* is a 'reference' observer the CIA with a specific disagreement function *G* is defined as:

$$\psi^N = \frac{[G(X, X') + G(Y, Y')]/2}{G(X, Y)}. \quad (1)$$

When *X* is a 'reference' (gold standard) and *Y* is a 'new' observer, the CIA is defined as

$$\psi^R = \frac{G(X, Y')}{G(X, Y)}. \quad (2)$$

When the mean squared deviation (MSD) is used as the *G* function, the coefficient ψ^N varies between 0 and 1 [16], while ψ^R may exceed 1. For both coefficients, a value close to unity or above unity indicates an acceptable agreement. Haber and Barnhart [16] and Haber et al. [18] suggested that ψ ≥ 0.8 indicates acceptable agreement. Alternatively, one can consider agreement as acceptable if the confidence interval for the CIA includes 1.

Hereafter, we will use the most common disagreement function, *G(X ,Y) = MSD(X,Y) E(X −Y)*^{2}, where *MSD* is the mean squared deviation. CIA's with continuous observations and different disagreement functions have been discussed in Haber and Barnhart [16]. If *MSD* is used as the G function and the observations are continuous, we note that ψ^R is related to the FDA's individual bioequivalence criteria [21], defined as

$$IBC = \frac{E(Y_{iT} - Y_{iR})^2 - E(Y_{iR} - Y_{iR'})^2}{E(Y_{iR} - Y_{iR'})^2},$$

with the following relationship (Barnhart et al. [17])

$$\psi^R = \frac{2}{2 + IBC},$$

when we set X as Y_R(reference drug) and Y as Y_T (test drug).

When the observations are binary, we have: $G(X,Y) = E(X - Y)^2 = P(X = 0,Y = 1)+ P(X = 1,Y = 0) = P(X \neq Y)$The binary case has been discussed in Haber et al. [18].

Suppose that both observers evaluate the same N study subjects, indexed by $i = 1,\ldots,N$. Let X_{ik} denote the k -th replicated observation of X ($k = 1,\ldots, K_i$) and Y_{il} denote the l -th replicated observation of Y($l = 1,\ldots, L_i$) on subject i. Note that we allow different numbers of replications across subjects and methods. In general, let $\boldsymbol{Z} = (X,Y)$' with $Z_{ijm} = \mu_{ij} + \varepsilon_{ijm_{ij}}$, $i = 1, \ldots,n$; $j = X,Y;m_{ij} = 1,\ldots,M_{ij}$(= K_i if $j = X$, or L_i if $j = Y$), where μ_{ij} be further written as $\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$ and ε_{ijm} are independent. Then we have the same model as the one in FDA's guidance [21] with X as Y_R (reference drug) and Y as Y_T (test drug). In general, $G(X,Y)$ involves the parameters from the bivariate distribution of $\boldsymbol{Z}$. Here we use a non-parametric approach, which is equivalent to using the empirical distribution of $\boldsymbol{Z}$ for estimation of $G(X,Y)$. Thus, this model specification is not needed for computation.

2.2 Estimation

In this section, the observations may be continuous or binary. We assume that the numbers of replications, K_i and L_i may vary across subjects.

2.2.1 Estimated ψ/'s—We first consider the disagreements for a particular subject i. Define:

$$\begin{aligned} G_i(X,Y) &= E[(X_{ik} - Y_{il})^2 \,|\, i], \\ G_i(X,X') &= E[(X_{ik} - X_{ik'})^2 \,|\, i] \text{ for } k<k', \\ G_i(Y,Y') &= E[(Y_{il} - Y_{il'})^2 \,|\, i] \text{ for } l<l' \end{aligned} \tag{3}$$

We then obtain the overall disagreement functions as

$$G(X,Y) = E[(G_i(X,Y)], \quad G(X,X') = E[G_i(X,X')], \quad G(Y,Y') = E[G_i(Y,Y')],$$

where E stands for the expectation over all study subjects. We first estimate the disagreement functions for each subject:

$$\begin{aligned} \widehat{G}_i(X,Y) &= \frac{1}{K_i L_i} \sum_{k=1}^{K_i} \sum_{l=1}^{L_i} (X_{ik} - Y_{il})^2, \\ \widehat{G}_i(X,X') &= \frac{2}{K_i(K_i-1)} \sum_{k=1}^{K_i-1} \sum_{k'=k+1}^{K_i} (X_{ik} - X_{ik'})^2, \\ \widehat{G}_i(Y,Y') &= \frac{2}{L_i(L_i-1)} \sum_{l=1}^{L_i-1} \sum_{l'=l+1}^{L_i} (Y_{il} - Y_{il'})^2, \end{aligned} \tag{4}$$

Then, the estimates of the overall disagreement functions are:

$\widehat{G}(X,Y) = \overline{G}(X,Y) = \frac{1}{N}\sum_{i=1}^{N} \widehat{G}_i(X,Y), \widehat{G}(X,X') = \overline{G}(X,X') = \frac{1}{N}\sum_{i=1}^{N} \widehat{G}_i(X,X')$, and

$\widehat{G}(Y,Y') = \overline{G}(Y,Y') = \frac{1}{N}\sum_{i=1}^{N} \widehat{G}_i(Y,Y')$, where $\overline{G} = \frac{1}{N}\sum_{i=1}^{N} \widehat{G}_i$ Finally, the estimated ψ's are obtained by substituting the estimated G's into their definitions in Section 2.1.

2.2.2 Standard Errors of Estimated ψ's—To simplify the notations, let $G^{(1)} = G(X,X')$, $G^{(2)} = G(Y,Y')$ and $G^{(3)} = G(X,Y)$. Then $\hat{\psi}^N = A_1 / B$, where $A_1 = (\bar{G}^{(1)} + \bar{G}^{(2)})/2$ and B = $\bar{G}^{(3)}$. Similarly, $\hat{\psi}^R = A_2 / B$, where $A_2 = \bar{G}^{(1)}$. Now, for p = 1,2,3, denote the sample variances $S^2(G^{(p)})=\frac{1}{N-1}\sum_{i=1}^{N}(\widehat{G}_i^{(p)} - \bar{G}^{(p)})^2$, so that $\hat{Var}(\bar{G}^{(p)}) = S^2(G^{(p)})/N$. In addition, for $1 \le p < q \le 3$, denote the sample covariance of $G^{(p)}$ and $G^{(q)}$ by $Cov(G^{(p)}, G^{(q)})=[\sum_{i=1}^{N}(\widehat{G}_i^{(p)} - \bar{G}^{(p)})(\widehat{G}_i^{(q)} - \bar{G}^{(q)})]/(N-1)$, so that $\hat{Cov}(\bar{G}^{(p)}, \bar{G}^{(q)}) = Cov(G^{(p)}, G^{(q)})/N$.

Using the above notations,

$$\begin{gathered}\widehat{Var}(A_1)=[S^2(G^{(1)})+S^2(G^{(2)})+2Cov(G^{(1)},G^{(2)})]/4N,\\ \widehat{Var}(A_2)=S^2(G^{(1)})/N,\\ \widehat{Var}(B)=S^2(G^{(3)})/N,\\ \widehat{Cov}(A_1,B)=[Cov(G^{(1)},G^{(3)})+Cov(G^{(2)},G^{(3)})]/2N,\\ \widehat{Cov}(A_2,B)=[Cov(G^{(1)},G^{(3)})/N.\end{gathered}$$

Finally, substitute these in the approximation for the variance of a ratio:

$$\widehat{Var}(\widehat{\psi}^N)=\widehat{Var}(\frac{A_1}{B}) \approx \frac{A_1^2}{B^2}[\frac{\widehat{Var}(A_1)}{A_1^2}+\frac{\widehat{Var}(B)}{B^2}-\frac{2\widehat{Cov}(A_1,B)}{A_1B}]. \quad (5)$$

Similarly,

$$\widehat{Var}(\widehat{\psi}^R)=\widehat{Var}(\frac{A_2}{B}) \approx \frac{A_2^2}{B^2}[\frac{\widehat{Var}(A_2)}{A_2^2}+\frac{\widehat{Var}(B)}{B^2}-\frac{2\widehat{Cov}(A_2,B)}{A_2B}]. \quad (6)$$

The standard errors of those two estimators are simply the square roots of the variance estimators.

The estimated standard errors, along with the corresponding confidence intervals, are computed assuming large-sample normality of the estimators. Based on our simulations for the continuous and binary cases under various settings, sample size of 50 subjects provided coverage probabilities of at least 90% and the histograms demonstrated normal curves. This demonstrates that with sample size of at least 50, the normality assumption regarding the distribution of the estimators of CIA's is adequate and hence our estimating method for standard errors is applicable. Our simulations also showed that the standard errors estimated by this approach are close to the corresponding standard deviations of the simulated estimates.

3. SAS and R Programs

The programs in either SAS or R estimate the CIA's and their standard errors for two observers, *X* and *Y*. It can be used for quantitative data with the MSD disagreement function or with binary data. As demonstrated before, the estimation of CIA's requires replicated observations by each observer on each subject to estimate the within-observer disagreement. The coefficient ψ^R can be estimated when observer *Y* does not have replications while observer *X* has replications and is considered as the reference. The coefficient ψ^N can only

be estimated when both *X* and *Y* have replications. The programs deal separately with the cases where there are replications on *Y* and when there are no replications on *Y*. Since each subject may have different numbers of observations across observers, only subjects with more than one observation on both *X* and *Y* are used to estimate ψ^N, while only subjects with more than one observation on *X* and at least one observation on *Y* are used to estimate ψ^R. A warning message is given if less than 10 subjects are used to estimate either ψ^N or ψ^R. Furthermore, from the theoretical aspect, ψ^N cannot be greater than 1, but in reality, $\hat{\psi}^N$ can be greater than 1 due to the randomness of the data. Last but not the least, we assume that any missing data are missing completely at random. Therefore we do not distinguish between cases where observations are missing due to the study design, due to the inability to observe some of the data, or for any other reasons.

3.1 Input Data

In the SAS macro, eight parameters, namely subject identifier, *data_name, measurement, method, observer1, observer2, alpha* and *title* are required for input. Regardless of what the subject identifier is called in the original data, either labeled as id or not, the macro variable *id* specifies the subject identifier that is used by the program. *Data_name* specifies the name of the dataset which contains the original data, while *measurement* is the name of the variable containing the outcome observations and *method* is the name of the variable indicating observers/methods. *Observer1* and *observer2* are the names of the first and second observers, respectively. The names should exactly match those shown in the dataset and they are case sensitive. *Alpha* is one minus the confidence coefficient, and by default alpha is set to be 0.05. *Title* is designed to help the user identify the current analyses and mark the name of the current comparison if several comparisons are involved. The R function only contains *data_name, observer1, observer2 and alpha*. Besides that, *samplesize* which indicates the total number of subjects in the dataset is also included as a parameter. In the R function,"method" is the required name for observer and quotes need to be used if *observer1* and *observer2* are character variables.

For both the SAS macro and the R function, the dataset is required to be in a 'long format', with the identifying number in one column, methods (observer) in one column and measurement in one column. If the original data is in a 'wide format' then it should be transferred to the required 'long format' before adopting the macro. An example of such transformation is included in the SAS macro. As stated above, we allow the numbers of replications to be different across subjects and observers.

3.2 Output

The following output is provided by our programs: (1) *MSD(X,Y), MSD*(*X, X*'), *MSD*(*Y,Y*'); (2) estimate of ψ^N, standard error (SE) of ψ^N and 100(1-α)% CI for ψ^N ; (3) estimate of ψ^R, SE of ψ^R and 100(1-α)% CI for ψ^R. There is no gold standard (reference) when estimating ψ^N, while *observer1* is treated as the reference when calculating ψ^R. In addition, ψ^N is not estimated when there is only one observation on *observer2* for each subject.

4. Examples

Examples of the SAS and R programs input and output are presented in Table 1 and Table 2.

4.1 A Quantitative Example – Carotid Stenosis Data

The carotid stenosis screening study was designed to determine the suitability of magnetic resonance angiography (MRA) for noninvasive screening of carotid artery stenosis, compared to invasive intra-arterial angiogram (IA). The main interest was in comparing two MRA techniques, two-dimensional (MRA-2D) and three-dimensional (MRA-3D) MRA

time of flight, to the IA, which was considered as the "gold standard". In this example, the three screening methods were considered as the "observers". Readings were made by 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 were considered as replications. Separate readings were made on the left and right carotid arteries. However, in this example, our interest was restricted to the left side. For more details on this study, the reader is referred to Barnhart and Williamson [25].

Assessing the agreement between MRA-2D and IA, $\hat{\psi}^N$ was 0.592 with 95% CI (0.348, 0.835), while $\hat{\psi}^R$ was 0.231 with 95% CI (0.035, 0.427), using IA as the reference. The agreement between MRA-3D and IA showed similar results s with $\hat{\psi}^N$ = 0.452 (95% CI: (0.242, 0.661)) and $\hat{\psi}^R$ = 0.191 (95% CI: (0.026, 0.357)) [17]. These results indicate that if the IA method is treated as a reference, one would conclude that MRA-2D and MRA-3D do not have good individual agreement with the IA method. These conclusions also hold when neither of the methods is considered as a gold standard. The SAS and R codes, as well as the output comparing MRA-2D to IA, are shown in Table 1.

4.2 A Binary Example – Data from a Mammography Study

In a mammography study [26], 150 female patients underwent a mammography at the Yale-New Haven Hospital in 1987. Each of ten radiologists read each patient's mammogram and classified it into one of four diagnosis categories: (1) normal, (2) abnormal – probably benign, (3) abnormal – intermediate or (4) abnormal – suggestive of cancer. Four months later, the same films were reviewed again, in a random order, by the same radiologists. We considered the two evaluations as replications. In the present analysis, we considered a radiologist's rating as "positive" if the mammogram was classified into the fourth category, which was abnormal and suggestive of cancer. Otherwise, the rating was considered as "negative". Each of the study participants was followed up for three years, and then a definitive diagnosis was made. The definitive diagnosis was breast cancer if it was histopathologically confirmed within the three years of follow-up. We considered this diagnosis as the patient's "true" breast cancer status. Based on this criterion, 27 of 150 patients (18%) had breast cancer. Ten radiologists were involved. Since the total of sensitivity and specificity was highest for radiologist A, we illustrated the new coefficients in [18] by estimating the agreement between radiologist A and each of the remaining nine radiologists. Radiologist A was considered as the reference in estimating ψ^R.

In the current illustration we focus on the agreement between radiologists A and F. When neither radiologist is considered as a reference, $\hat{\psi}^N$ is 0.762 with 95% CI (0.476, 1.047); while with radiologist A as the reference, $\hat{\psi}^R$ is 0.571 with 95% CI (0.162, 0.981). Thus, the CI for $\hat{\psi}^N$ includes 1 and we can claim that the agreement between radiologists A and F is acceptable when neither of them is considered as the reference. On the other hand, the CI for ψ^R does not include 1, hence agreement is not acceptable when radiologist A is considered as a reference. SAS and R codes as well as outputs are shown in Table 2.

5. Summary

In this paper, we provide computer programs to estimate the coefficients of individual agreement (CIA's), which are based on the ratio of within-observers and between-observers disagreement. Both a SAS macro and an R function were introduced to estimate CIA's along with their SE's and confidence intervals, when one of the observers was considered as a reference and when neither of the observers was a reference. Two examples demonstrated that our programs worked well for both quantitative and binary measurements.

6. Macro Availability and Software Requirement

The CIA programs in SAS and R are available directly from the authors. They can also be found at the following website: http://www.sph.emory.edu/observeragreement/. The programs were written in SAS V9.1.3 and R 2.9.1. The R users need to download, install and load the packages “fSeries” and “reshape” from CRAN mirror. The functions “melt” and “cast” were involved in transferring the data format.

Acknowledgments

This research was partially supported by NIH grants R0l MH070028 and UL1 RR024128.

References

1. Barnhart HX, Haber M, Lin LI. An overview on assessing agreement with continuous measurement. J. Biopharm. Stat 2007;17:529–569. [PubMed: 17613641]
2. Choudhary PK. A tolerance interval approach for assessment of agreement in method comparison studies with repeated measurements. Journal of Statistical Planning and Inference 2008;138:1102–1115.
3. Choudhary PK, Nagaraja HN. Tests for assessment of agreement using probability criteria. Journal of Statistical Planning and Inference 2007;137:279–290.
4. Lin LI, Hedayat AS, Sinha B, Yang M. Statistical methods in assessing agreement: models, issues and tools. Journal of American Statistical Association 2002;97:257–270.
5. Bartko JJ. The intraclass correlation coefficient as a measure of reliability. Psychological Reports 1966;34:3–11. [PubMed: 5942109]
6. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychological Reports 1996;1:30–46.
7. Lin LI. A concordance correlation coefficient to evaluate reproducibility. Biometrics 1989;45:255–268. [PubMed: 2720055]
8. King TS, Chinchilli VM. A generalized concordance correlation coefficient for continuous and categorical data. Stat. Med 2001;20:2131–2147. [PubMed: 11439426]
9. Barnhart HX, Haber M, Song JL. Overall concordance correlation coefficient for evaluating agreement among multiple observers. Biometrics 2002:1020–1027. [PubMed: 12495158]
10. Williamson JM, Crawford SB, Lin HM. Resampling dependent concordance correlation coefficients. J. Biopharm. Stat 2007;17:685–696. [PubMed: 17613648]
11. Haber M, Barnhart HX. Coefficient of agreement for fixed observers. Statistical Methods in Medical Research 2006;15:1–17.
12. Atkinson G, Nevill A. Comment on the use of concordance correlation to assess the agreement between two variables. Biometrics 1997;57:931–940.
13. Barnhart HX, Haber M, Lokhnygina Y, Kosinski AS. Comparison of concordance correlation coefficient and coefficient of individual agreement in assessing agreement. J. Biopharm. Stat 2007;17:721–738. [PubMed: 17613650]
14. Fleiss JL, Cohen J. The equivalence of the weighted kappa and the intraclass correlation coefficient as a measure of reliability. Educational and Psychological Measurements 1973;33:613–619.
15. Shoukri, MM. Measures of interobserver agreement. Chapman & Hall/CRC; 2004.
16. Haber M, Barnhart HX. A general approach to evaluating agreement between two observers or methods of measurement. Statistical Methods in Medical Research 2008;17:151–169. [PubMed: 17698934]
17. Barnhart HX, Haber M, Kosinski AS. Assessing individual agreement. J. Biopharm. Stat 2007;17:697–719. [PubMed: 17613649]
18. Haber M, Gao J, Barnhart HX. Assessing observer agreement in studies involving replicated binary observation. J. Biopharm. Stat 2007;17:757–766. [PubMed: 17613652]

19. Anderson S, Hauck WW. Consideration of individual bioequivalence. Journal of Ppharmacokinetics and Biopharmaceutics 1990;18:259–273.
20. Schall R, Luus HG. On population and individual bioequivalence. Statistics in Medicine 1993;12:1109–1124. [PubMed: 8210816]
21. FDA. Guidance for industry: statistical approaches to establishing bioequivalence, Food and Drug Administration. Center for Drug Evaluation and Research (CDER); 2001 Jan. BP
22. Haber M, Barnhart HX, Song J, Gruden J. Observer variability: a new approach in evaluating interobserver agreement. Journal of Data Science 2005;3:69–83.
23. Shao J, Zhong B. Assessing the agreement between two quantitative assays with repeated measurements. J. Biopharm. Stat 2004;14:201–212. [PubMed: 15027509]
24. Bland JM, Altman DG. Measuring agreement in method comparison studies. Statistical Methods in Medical Research 1999;8:135–160. [PubMed: 10501650]
25. Barnhart HX, Williamson JM. Modeling concordance correlation via GEE to evaluate reproducibility. Biometrics 2001;57:931–940. [PubMed: 11550947]
26. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists interpretation of mammograms. New Eng. J. Med 1994;331:1439–1499.

Table 1

Estimating the agreement between MRA-2D to IA using CIA's in the stenosis screening study (left side).

Description	The three readings by different raters were treated as replications. In method column, "1" denotes IA while "2" denotes MRA-2D. Measurement column contains the corresponding readings from each method.		
Original Dataset	id	method	measurement
	1	1	100.000
	1	1	100.000
	1	1	100.000
	1	2	100.000
	1	2	100.000
	1	2	100.000
	2	1	73.077
	2	1	66.575
	2	1	77.186
	2	2	69.472
	2	2	60.410
	2	2	72.214
SAS code	%include 'D:\CIA\program\CIA_macro.sas' ; %***CIA*** (id=id, data_name=sten_left, measurement=measurement, method=method, observer1=1, observer2=2, alpha=0.05, title=Comparison between IA and MRA-2D);		
R code	sten <- read.csv(file="C:\\ CIA\\data\\sten_left.csv", header=T) out <- CIA(data_name=sten_left, observer1="1", observer2="2", samplesize=55, alpha=0.05)		
Output	Comparison between IA and MRA-2D		
	Estimate of MSD Functions		
	MSD_XX	MSD_YY	MSD_XY
	279.447	1153.48	1210.77
	Estimated Psi_N		
	Est_ Psi_N	SE_Est_ Psi_N	95.0% CI for Psi_N
	0.592	0.124	(0.348,0.835)
	Estimated Psi_R		
	Est_ Psi_R	SE_Est_ Psi_R	95.0% CI for Psi_R
	0.231	0.100	(0.035,0.427)

Table 2

Estimating the agreement between radiologist A and F using CIA's in the mammography study.

Description	The two readings of radiologist A were called x1 and x2 since A was treated as the reference for ψ^R. The measurements of radiologist F were y1 and y2.		
Original Dataset	study_id	rater	reading
	1	A	0
	1	A	0
	1	F	0
	1	F	0
	2	A	0
	2	A	0
	2	F	0
	2	F	0
SAS code	%include 'D:\CIA\program\CIA_macro.sas' ; %***CIA*** (id=id, data_name=mammo, measurement=reading, method=rater, observer1=A, observer2=F, alpha=0.05, title=Comparison between A and F);		
R code	mammo <- read.csv(file="C:\\ CIA\\data\\mammo.csv", header=T) out <- CIA(data_name=mammo, observer1="A", observer2="F", samplesize=150, alpha=0.05)		
Output	Comparison between A and F		
	Estimate of MSD Functions		
	MSD_XX	MSD_YY	MSD_XY
	0.040	0.067	0.070
	Estimated Psi_N		
	Est_ Psi_N	SE_Est_ Psi_N	95.0% CI for Psi_N
	0.762	0.146	(0.476,1.047)
	Estimated Psi_R		
	Est_ Psi_R	SE_Est_ Psi_R	95.0% CI for Psi_R
	0.571	0.209	(0.162,0.981)