A General Approach to Evaluating Agreement between Two Observers or Methods of Measurement

Michael Haber
Department of Biostatistics, Rollins School of Public Health
Emory University, Atlanta, GA

and

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

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

Revised 3/22/06

Abstract

We present a general approach to the definition and estimation of coefficients for evaluating agreement between two fixed methods of measurements or human observers. The measured variable is assumed to be continuous, but no other distributional assumptions are made. We introduce the term 'disagreement function' for the function of the observations that is used to quantify the extent of disagreement between the two measurements made on the same subject. The proposed agreement coefficients compare the disagreement between measurements made by different methods on the same subject to the corresponding disagreement between measurements made by the same method. We propose agreement coefficients for two practical situations involving two methods that have a measurement error: (a) comparison of a new method to a gold standard (or a reference method), and (b) comparison of two methods where neither method is considered a gold standard. We consider three disagreement functions based on the differences between two measurements: (a) the mean squared difference (MSD), (b) the mean absolute difference (MAD), and (c) the complement of the coverage probability (CP), which is the probability that the absolute difference does not exceed a predetermined threshold. We then derive nonparametric estimates for the various agreement coefficients. Our approach is illustrated using data from a study comparing systolic blood pressure measurements by a human observer and an automatic monitor. The performance of the new estimates is assessed via stochastic simulations. .

1. Introduction

Frequently one is interested in comparing two methods of measurement that were applied to each of N study subjects. In this paper, the term 'method' may correspond to a measurements device or to a human observer. Denoting the corresponding measurements X and Y, we will say that the two methods are in agreement if they produce the same value on each subject. In reality, even under 'perfect' agreement the values of X and Y on the same subject will usually differ because of measurement errors and other factors. Therefore, it is important to define and estimate coefficients that quantify the extent to which the two methods agree with each other.

We consider two practical scenarios for the comparison of two methods that have a measurement error. In the first scenario, one is interested in comparing the methods without considering either of them as a reference (or a 'gold standard'). This may be of interest, for example, when due to logistic considerations one plans to use method X on some of the subjects and method Y on the remaining subjects. In this case it is important that the two methods produce similar values when applied to the same subject so that they can be used *interchangeably*. In the second scenario, method X is a reference method that has been used in the past and was found to produce reliable measurements, and method Y is a new method that may be less expensive or less invasive. In this case one may consider *replacing* X by Y, and must make sure that the two methods produce similar measurements on the same subject. We will refer to these two scenarios as assessing agreement with or without a reference method.

Our objective is to define coefficients of agreement for each of the two scenarios such that if the coefficient is close to one (or exceeds one) we conclude that the methods are in good agreement, while a value close to zero indicates poor agreement. Over the years, several coefficients of agreement have been proposed

and used, including different versions of the intraclass correlation coefficient (ICC) and the concordance correlation coefficient (CCC). A comparative summary of these coefficients is presented in a recent paper by Haber and Barnhart (2006)[1]. In the present paper we propose a general approach that can be used in defining and estimating coefficients of agreement corresponding to different measures of disagreement. The new coefficients generalize the ψ coefficient presented in our earlier paper.

In order to define a coefficient of agreement, we first have to decide how we quantify the agreement (or lack thereof) between the two methods. For this, we define the concept of a *disagreement function* $G(X,Y)$. This function has to satisfy:

a. $G(X,Y) \geq 0$

b. $G(X,Y) = G(Y,X)$

c. $G(X,Y)$ increases as the disagreement between X and Y (according to a specific criterion) increases.

Once we have decided which disagreement function to use, it is necessary to standardize it in order to obtain a coefficient so that values close to zero indicate poor agreement and values close to (or above) one indicate good agreement. In other words, we need to compare the observed value of $G(X,Y)$ to a reference value that represents either the best or the worst case scenario. The ICC's and the CCC consider the worst case scenario as independence, or agreement by chance. Therefore these coefficients compare the observed value of a disagreement function to the value expected under independence between the two methods[2,3] In our opinion[1], this approach is unjustified because (a) agreement and correlation are different concepts, and (b) two observations made on the same subject are always expected be positively correlated because both are positively correlated with the true value of the measured variable on this subject.

Since we could not find an adequate definition for the worst case scenario we decided to move to the opposite side of the agreement scale and define the best case scenario as the case where the disagreement between observations made by different methods is similar to the disagreement between observations made by the same method. In other words, replacing or interchanging one method by the other does not substantially increase the disagreement between measurements made on the same subject. *Therefore, the coefficients of agreement we propose compare the disagreement between measurements made by different methods to the disagreement between replicated measurements made by the same method*.

We denote by $G(X,X')$ the disagreement between two replicated measurements made by the method *X*, and define $G(Y,Y')$ in an analogous way for method *Y*. Using the above-mentioned principle, we define the coefficient of agreement in the case that none of the methods is considered as a reference as:

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

The numerator is the average disagreement between two replicated measurements made by the same method, and the denominator is the disagreement between *X* and *Y*. The coefficient of agreement for the case where method *X* is considered a reference method is defined as:

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

Here we compare the disagreement between measurements made by the two methods with the disagreement between two replicated measurements made by the reference method. Note that both ψ_G^N and ψ_G^R are nonnegative but they may attain values greater than one, indicating that the between-methods disagreement is lower than the within-methods disagreement.

The comparison of $G(X,Y)$ to $G(X,X')$ when *X* is considered a reference method has been used in the past to evaluate *individual bioequivalence* between a standard (reference) and a new formulation of a drug[4] In that case, the focus was

on comparing the bioavailabilities of the two formulations. The comparison of $G(X,Y)$ to the mean of $G(X,X')$ and $G(Y,Y')$ where $G = E(X-Y)^2$ was introduced in a recent paper by Barnhart et al.[5] to assess the agreement between two or more measurement methods when there is no reference method. They called the coefficients (1) and (2) with $G(X,Y) = E(X-Y)^2$ 'coefficients of individual agreement' (CIA) as they were based on the idea used in assessing individual bioequivalence.

It is important to realize that in these coefficients the within-method disagreement is used as a reference to which we compare the between-methods disagreement. Hence, before using the ψ's one must check that the corresponding within-method disagreement, i.e., $G(X,X')$ for ψ_G^R and both $G(X,X')$ and $G(Y,Y')$ for ψ_G^N, are acceptable.

As with other coefficients, a common practical question is 'how good is good', i.e. what values of the coefficients are large enough to be considered as indicating 'good agreement'. To allow a better interpretation of the new coefficients we suggest to look at their reciprocals, $1/\psi_G^R$ and $1/\psi_G^N$. These quantities represent the relative increases in disagreement due to replacing X by Y or due to interchanging X and Y. In order to claim 'good' agreement, we believe that the disagreement between the two methods should not exceed the disagreement between replicated observations from the same method by more than 25%, i.e., $\psi \geq 0.8$.

The disagreement functions can also be defined at the subject level. Denoting $G_i(X,Y)$, $G_i(X,X')$, $G_i(Y,Y')$ the values of the disagreement functions for subject i, we can define the following *subject-specific* agreement coefficients:

$$\psi_{G,i}^N = \frac{[G_i(X,X') + G_i(Y,Y')]/2}{G_i(X,Y)}\ , \qquad (3)$$

$$\psi_{G,i}^{R} = \frac{G_i(X,X')}{G_i(X,Y)}. \qquad (4)$$

Note that $G_i(X,Y)$ may be zero. Since in this case there is 'good agreement', as measured by the disagreement function *G* for this subject, we define the disagreements coefficients (3) and (4) as one regardless of the values of $G_i(X,X')$ and $G_i(Y,Y')$.

Data: Since the new coefficients involve the evaluation of disagreement between measurements made with the same method on the same subject, we have to use data with *replicated measurements*. When we compare two methods (without a reference) we will assume that at least two measurements are made using each method on each subject. For the case of comparing a new method to a reference we only need replicated measurements on the reference method. The number of replications does not have to be fixed. We denote by $X_{i1},...,X_{iK_i}$ the K_i measurements made on subject *i* using method *X*. Similarly, we denote by $Y_{i1},...,Y_{iL_i}$ the L_i measurements made on subject *i* using method *Y*. We assume that $K_i \geq 2$, $L_i \geq 2$ and that these are *unmatched replications,* i.e., if one permutes the order of the subscripts for one of the methods without changing the order for the other method, then there is no change in the resulting coefficients or their estimates. We do not make any assumptions about the distributions of the measurements.

In this work we will consider two disagreement functions related to agreement measures introduced by Lin et al.[6], namely the *mean squared difference* (MSD), defined as the expectation of the squared deviation between two measurements, and the complement of the *coverage probability* (CP), which is the probability that two measurements are within a specific distance from each other. In addition, we consider a new disagreement function, *the mean absolute difference* (MAD), as the absolute difference is more meaningful than the squared difference.

2. The Mean Squared Deviation (MSD)

In this Section we define $G(X,Y) = MSD(X,Y) = E(X-Y)^2$. We first consider the disagreements for a particular subject, i:

$G_i(X,Y) = E[(X_{ik} - Y_{il})^2 \mid i)]$,

$G_i(X,X') = E[(X_{ik} - X_{ik'})^2 \mid i)]$ for $k < k'$

$G_i(Y,Y') = E[(Y_{il} - Y_{il'})^2 \mid i)]$ for $l < l'$

We then obtain the overall disagreement functions as $G(X,Y) = E_i[G_i(X,Y)]$, $G(X,X') = E_i[G_i(X,X')]$, $G(Y,Y') = E_i[G_i(Y,Y')]$, where E_i stands for the expectation over all study subjects. If we assume conditional independence of X and Y given the subjects' characteristics, then the disagreement functions can also be written in terms of the conditional moments of X and Y:

$G(X,Y) = E_i\{[E(X_{ik} \mid i) - E(Y_{il} \mid i)]^2 + Var(X_{ik} \mid i) + Var(Y_{il} \mid i)\}$,

$G(X,X') = 2E_i[Var(X_{ik} \mid i)]$, $G(Y,Y') = 2E_i[Var(Y_{il} \mid i)]$.

The overall agreement coefficients (1) and (2) and the subject-specific coefficients (3) and (4) are now obtained by substituting these disagreement functions into the corresponding equations.

The overall coefficients (1) and (2) with $G = MSD$ are identical to the CIA^N and CIA^R coefficients in Barnhart et al.[5] when there is no reference method and when method X is considered as reference, respectively. In addition, ψ^N_{MSD} is the same as ψ in Haber and Barnhart[1]. However, ψ in the earlier paper compares $MSD(X,Y)$ to its expected value when $E(X_i) \equiv E(Y_i)$ while the current ψ^N_{MSD} compares $MSD(X,Y)$ to the mean *MSD* of replicated measurements by the same method. These approaches lead to the same coefficient for $G = MSD$, but they may result in different coefficients for other selections of the disagreement function.

2.1 Estimation

We begin with the subject-specific coefficients. For subject i, we estimate

$\hat{G}_i(X,Y) = Mean_{k,l}(X_{ik} - Y_{il})^2$,

$\hat{G}_i(X,X') = Mean_{k<k'}(X_{ik} - X_{ik'})^2 = 2 \cdot MSW(X_i)$,

$\hat{G}_i(Y,Y') = Mean_{l<l'}(Y_{il} - Y_{il'})^2 = 2 \cdot MSW(Y_i)$,

where *MSW* are the within-subject mean squares:

$MSW(X_i) = \sum_k (X_{ik} - \overline{X}_i)^2 / (K_i - 1)$, $MSW(Y_i) = \sum_l (Y_{il} - \overline{Y}_i)^2 / (L_i - 1)$.

We then substitute these estimates of the $\hat{G}_i$'s into (3) or (4) to obtain the estimated subject-specific coefficients of agreement. Based on the note underneath equation (4), we set the estimated coefficients to one whenever $\hat{G}_i(X,Y) = 0$.

To estimate the overall coefficients of agreement we estimate the overall disagreement function as $\hat{G}(X,Y) = Mean_i[\hat{G}_i(X,Y)]$. Likewise, $\hat{G}(X,X')$ and $\hat{G}(Y,Y')$ are estimated as the means over all subjects of the corresponding subject-specific disagreements. The overall agreement coefficient is estimated by substituting the estimated *G*'s into (1) or (2).

2.2 An example

To illustrate the coefficients introduced in this paper we use a data set from Bland and Altman[7]. Systolic blood pressure (SBP) was measured on 85 subjects by two experienced human observers using a sphygmomanometer and by a semi-automatic blood pressure monitor. Three replications were made in quick succession with each of the three methods on each subject. We first assessed the agreement between the two human observers. The agreement was excellent $\hat{\psi}_G^N = 1.44$, hence we will focus here on the agreement between the first human observer, which we label *X,* and the monitor, which we label *Y*. Our estimated

G's were $\hat{G}(X, X') = 74.745$, $\hat{G}(Y, Y') = 166.282$, $\hat{G}(X, Y) = 677.448$. Hence, if we compare the two methods without considering either of them as a reference then the estimated agreement is $\hat{\psi}_G^N = 0.178$. However, it could be more appropriate to consider the experienced human observer as a reference and the monitor as a new method that is tested against this reference. Then we estimate the agreement as $\hat{\psi}_G^R = 0.110$. Either way, the agreement is poor. This is not surprising as the mean reading of the monitor is about 16 points higher than that of the human observer. These estimates are identical to those obtained by Barnhart et al.[5] using an ANOVA model. We used the bootstrap method to obtain standard errors and confidence intervals (CI) for these coefficients. The 95% CI's based on the bootstrap percentiles for ψ_G^N and ψ_G^R were (0.110, 0.306) and (0.067, 0.207), respectively. We also calculated the CI's based on the normal approximation to the distribution of the estimates: (0.078, 0.278) and (0.037,0.183) respectively. Thus, the endpoints of the percentile-based CI's exceed those of the normal CI's, which suggests that the distributions of the estimates are skewed. The CI's based on the normal approximation to the distribution of the logarithms of the estimates are (0.105, 0.303) and (0.061, 0.198) respectively, quite close to the percentage-based CI's.

We can learn more about the agreement between X and Y by plotting the subject-specific coefficients of agreement as functions of our best estimates of the subjects' true SBP. When we did not consider either method as a reference, we plotted the estimates of $\psi_{G,i}^N$ against $(\overline{X}_i + \overline{Y}_i)/2$, where $\overline{X}_i$ and $\overline{Y}_i$ are the means over the replicated observations (Figure 1a). When the human observer (X) was considered a reference, we plotted the estimates of $\psi_{G,i}^R$ against $\overline{X}_i$ (Figure 1b). Both coefficients tend to increase as the SBP increases Thus, the agreement between the two methods is better for individuals with higher SBP compared to those with lower blood pressure. The plots also confirm the skewed distribution of the estimates.

2.3 Simulations

We conducted a simulation study to investigate the bias and precision of the estimates of ψ_G^N and ψ_G^R for $G = MSD$, using a simple latent class model to generate the replicated conditionally independent measurements by the methods X and Y. The true value, T, was assumed normal with mean μ_T and standard deviation σ_T. Let $T = t_i$ denote the true value for subject i. We then generated K replicated observations by method X from $(X_{ik} \mid t_i) \sim N(\mu_{X|t_i}, \sigma^2_{X|t_i})$ and L replicated observations by method Y from $(Y_{il} \mid t_i) \sim N(\mu_{Y|t_i}, \sigma^2_{Y|t_i})$. For simplicity we assumed that the number of replication by each method is the same for all subjects, though the number of replications by X and Y may differ. We assumed further that the above conditional means and standard deviations 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 it is easy to show that the true value of the three G functions for $G = MSD$ are:

$$G(X,Y) = (a-c)^2 + e^2 + g^2 + 2[(a-c)(b-d) + ef + gh)]\mu_T + [(b-d)^2 + f^2 + h^2](\mu_T^2 + \sigma_T^2)$$
$$G(X,X') = 2e^2 + 4ef\mu_T + 2f^2(\mu_T^2 + \sigma_T^2)$$
$$G(Y,Y') = 2g^2 + 4gh\mu_T + 2h^2(\mu_T^2 + \sigma_T^2).$$

The true values of ψ^N and ψ^R are now obtained from (1) and (2).

Our purpose was to simulate data that are similar to the SBP data used in the above example. We considered the mean (over the three replications) readings of the second human observer (whose observations were not used in our example) in the Bland and Altman[7] data as the 'true values' (T) of SBP. The sample moments of these 'true values' will be used as the moments of T; they are $\mu_T = 127.32$ and $\sigma_T = 30.49$. We then obtained the coefficients in the equations defining the conditional moments of X and Y given T by regressing the observed values of these conditional moments on the 'true values'. The coefficients are: $a = -1.03$, $b = 1.01$, $c = 34.33$, $d = 0.85$, $e = 1.91$, $f = 0.03$, $g = 3.62$,

$h = 0.03$. For this choice of the parameters in our simulation model, which will be referred to as 'case 1', the true values of the agreement coefficients are $\psi_G^N = 0.266$, $\psi_G^R = 0.199$. We also conducted simulations with larger true values of the agreement coefficients by changing the values of c and d while leaving the other simulation parameters unchanged. For 'case 2' we used $c = 13, d = 0.95$, which yield $\psi_G^N = 0.670$, $\psi_G^R = 0.502$, and for 'case 3' we used $c = 5, d = 0.98$, which yield $\psi_G^N = 0.940$, $\psi_G^R = 0.704$. (By bringing *c* closer to 0 and *d* closer to 1 we increase the agreement between Y and the true value. This increases the agreement between X and Y because X agrees very well with the true value.) Thus, the three cases correspond to poor, fair and good agreement, respectively.

We used three sample sizes, $n = 50$, 100 and 200. For estimating ψ_G^N we used two combinations of (K, L), namely (3,3) and (2,2), For estimating ψ_G^R we used four combinations of (K, L), namely (3,3), (2,2), (3,1) and (2,1). This allows to assess the importance of using replicated observations on the new method, *Y*, since ψ_G^R can be estimated from data with replicated observations on the standard method, *X*, only.

Table 1 presents the bias and root mean square error (RMSE) of the estimates for all the combinations of (n, K, L) for the three cases defined earlier. Each table entry is based on a set of 200 simulations. We see that the bias is usually positive but very small. The RMSE varies with the sample size and the number of replications. For $n = 50$, using fewer than three replications with either X or Y may result in imprecise estimates. For $n \geq 100$ the precision is acceptable even with the minimum number of replications ($K = L = 2$ for ψ_G^N and $K = 2, L = 1$ for ψ_G^R).

2.4 Effect of between-subjects variability on the agreement coefficients

The latent class model introduced earlier can also be used to investigate the behavior of the agreement coefficients in various situations. One issue that is commonly raised in connection with coefficients of agreement is their dependence on the between-subjects variability of the quantity being measured. Therefore we explore the dependence of the agreement coefficients on the variation of the 'true' SBP in the population, i.e. on σ_T^2. Let us first consider a special case of the latent class model where the conditional ('within subject') variances of the measurements do not depend on the true value t, i.e., $f = h = 0$. In this case $G(X, X')$ and $G(Y, Y')$ do not depend on σ_T^2 and $G(X, Y)$ is a monotonically increasing function of σ_T^2. Therefore both ψ_G^N and ψ_G^R are decreasing functions of σ_T^2. In the general latent class model presented in Section 2.3 all the three *G*'s increase with σ_T^2. We used the values of the eight coefficients ($a, b, ..., h$) determining the conditional moments of *X* and *Y* given *t* as estimated from the data (Section 2.3) to examine the dependence of the agreement coefficients on σ_T^2. For comparison, we also calculated the CCC for each case. The results are presented in Figure 2. We see that the new agreement coefficients decrease as the between-subjects variance increases. However, the rate of decrease is modest. On the other hand, the CCC increases at a very fast rate with σ_T. Also, for $\sigma_T \geq 20$ the CCC is considerably higher than the new psi coefficients. We repeated the calculations of the psi's with higher values of *f* and *h*, to allow more variability of the between-subject variance as the subject's true value t changes. However, the rate of decrease in the new agreement coefficients with σ_T did not change. In our opinion, the decrease of the agreement coefficients as the between-subject variance increases looks reasonable, as it is more difficult for observers to demonstrate good agreement across a wider range of the subjects' true values.

3. The Mean Absolute Deviation (MAD)

In this section we consider the disagreement function $G(X,Y) = E\,|\,X - Y\,|$. The MAD is easier to interpret as compared to the MSD, as it is more readily related to the actual observations. In general, one would expect the ψ_{MAD}'s to be close to the square roots of the corresponding ψ_{MSD}'s. Estimation of the agreement coefficients based on the MAD is very similar to the estimation of the coefficients based on the MSD. We begin with the estimation of the subject-specific *G*'s:

$\hat{G}_i(X,Y) = Mean_{k,l}\,|\,X_{ik} - Y_{il}\,|$,

$\hat{G}_i(X,X') = Mean_{k<k'}\,|\,X_{ik} - X_{ik'}\,|$,

$\hat{G}_i(Y,Y') = Mean_{l<l'}\,|\,Y_{il} - Y_{il'}\,|$.

We then calculate the sample means of the estimated G_i's to obtain the overall estimates of the disagreement functions and of the agreement coefficients. We can also estimate the subject-specific coefficients of agreement.

3.1 SBP example

For the SBP data discussed in Section 2.2, the estimated coefficients with $G = MAD$ are $\hat{\psi}_G^N = 0.426$ and $\hat{\psi}_G^R = 0.363$. The bootstrap percentile-based CI's are (0.348, 0.517) and (0.282, 0.459), respectively, while the bootstrap CI's based on the normal approximation are (0.341, 0.512) and (0.278, 0.448), respectively. The differences between the two kinds of confidence intervals for the same parameter are smaller than they were for $G = MSD$. In other words, the distribution of the ψ_{MAD}'s is more symmetric that that of the ψ_{MAD}'s. This is also evident from the plots of the subject-specific ψ_{MAD}'s (not included) which show that the distribution of these quantities is less skewed.

3.2 Simulations

We used the same simulation model and parameters as in Section 2.3 The true values of the *MAD*'s were calculated as means of folded normal variables. We now illustrate the calculation of $MAD(X,Y)$. Under our simulation model the difference $D = X - Y$ has a normal distribution and $|D|$ has a folded normal distribution. Then $MAD(X,Y)$ is the mean of this distribution:

$$MAD(X,Y) = E|D| = \sqrt{2/\pi}\cdot\sigma_D \cdot \exp(-\mu_D^2/2\sigma_D^2) - \mu_D \cdot [1 - 2\Phi(\mu_D/\sigma_D)],$$

where μ_D and σ_D are the mean and standard deviation of D and Φ is the standard normal distribution function[8]. The first two moments of *D* were obtained from the first two moments of (X,Y), which were calculated from the parameters of the simulation model as follows:

$E(X) = a + b\mu_T$, $E(Y) = c + d\mu_T$,

$Var(X) = e^2 + 2ef\mu_T + f^2\mu_T^2 + (b^2 + f^2)\sigma_T^2$,

$Var(Y) = g^2 + 2gh\mu_T + h^2\mu_T^2 + (d^2 + h^2)\sigma_T^2$,

$Cov(X,Y) = bd\sigma_T^2$.

The $MAD(X,X')$ and $MAD(Y,Y')$ were calculated in a similar way.

Table 2 presents the bias and RMSE of the estimates of ψ_G^N and ψ_G^R for $G = MAD$. The bias is usually negative, while the bias for the MSD-based estimates was usually positive. Comparing the RMSE's of the MSD and MAD-based estimates, we see that for case 1, where the true agreement is low, the RMSE's of the MAD-based estimates are slightly higher than the corresponding RMSE's of the MSD-based estimates. On the other hand, for cases 2 and 3 where the true agreement is moderate or high, the MAD-based estimates have considerably smaller RMSE's as compared to the MSD-based estimates.

4. Coverage Probabilities (CP)

The coverage probability for a given value $c > 0$ was defined by Lin et al.[6] as $CP(c) = P(|X - Y| < c)$. We define the CP-based disagreement function as $G(X, Y, c) = P(|X - Y| \geq c) = 1 - CP(c)$. As before, we will not make any assumptions on the distributions of X and Y and estimate the overall agreement coefficients from the subject-specific estimates. The subject-specific disagreement functions are estimated from the proportions of pairwise observations that are separated by c units or more:

$\hat{G}_i(X, Y, c) = [\#\text{ pairs } (X_{ik}, Y_{il}) \text{ so that } |X_{ik} - Y_{il}| \geq c\,] / K_i L_i$,

$\hat{G}_i(X, X', c) = 2[\#\text{ pairs } (X_{ik}, X_{ik'}),\ k < k', \text{ so that } |X_{ik} - X_{ik'}| \geq c\,] / K_i(K_i - 1)$,

and $\hat{G}_i(Y, Y', c)$ is obtained in an analogous way. The overall disagreement functions and agreement coefficients are calculated from the estimates of the subject-specific disagreement functions in the same way as in Sections 2 and 3.

4.1 SBP example

For the SBP data we used $c = 10$, which means that differences of less than 10 points between two measurements on the same subject are considered 'acceptable'. Our estimates were $\hat{\psi}^N = 0.446$, with a bootstrap percentile-based CI (0.357, 0.547), and $\hat{\psi}^R = 0.406$, with a bootstrap percentile-based CI (0.304, 0.521). The CI's based on the normal assumption were very close to the percentile-based CI's.

The main drawback of the CP-based agreement coefficients is their dependence on the choice of c. In the SBP example, using $c = 5$ rather than $c = 10$ increases the estimated ψ^N and ψ^R to 0.670 and 0.615, respectively.

5. Discussion

We presented a general approach to defining agreement coefficients for the comparison of two fixed observers or methods of measurement. The approach is based on the concept of 'disagreement function', which allows the user to specify her/his criterion for assessing the agreement between observations on the same subject. We used two approaches to standardize the value of the disagreement function in such a way that the ensuing coefficient will be close to or above one when the magnitude of the disagreement is good. When neither of the measurement methods can be consider as a 'gold standard' or a 'reference', we compare the disagreement between the methods to the average disagreement between replicated readings from the same method. On the other hand, when one of the methods has been in use for a while and can be considered a gold standard (reference) then the between-methods disagreement is compared to the disagreement between replicated readings from the reference method.

In the absence of any parametric assumptions, the estimation of the new coefficients requires replicated observations by the same method on the same subject. (When one of the methods is considered a gold standard then replicated observation by the new method are not required). Ideally, the methods or observers should be blind to their previous assessment(s) of the same subject and the true value of the measured variable on a subject should not change between replicated evaluations. We realize that it some cases it may be difficult to obtain replicated observations that satisfy these conditions. The new coefficients are useful, for example, when the measurements are made by automatic devices on blood samples or x-ray slides obtained from each subject. When the methods are human observers it is important to make sure that each observer makes her/his measurements in a random order, so that she/he is unlikely to recall the previous measurements(s) on a given subject.

In most cases the disagreement function can be computed for each subject. Then one can estimate the agreement coefficient for each subject and plot the estimated

coefficients against the subjects' estimated true values or against other subject-specific covariates. These plots may shed more light on the factors that affect agreement and help in identifying outlying observations. They may also be used as a supplement to the traditional Bland-Altman plots[9] which display the difference between measurements of two methods on the same subject as function of the subjects' estimated true value.

In this paper we did not make any assumptions regarding the distributions of the measurements. Thus, we presented non parametric estimates only. We also avoided some of the other assumptions commonly made when evaluating observer agreement. For example, our approach allows the error variances to differ across subjects, which is a quite common phenomenon in reality.

It is interesting to compare the MSD-based coefficient ψ^N with the CCC. Both coefficients have the MSD in the denominator. The new coefficient compares the MSD with its value under the assumption that the two methods are interchangeable, while the CCC used the expected MSD under independence as a yardstick. One important difference between these coefficients is related to the fast increase in the CCC when the between-subject heterogeneity increases (Figure 2). The new coefficient, on the other hand, is less sensitive to the sample heterogeneity and, at least in the examples we explored, decreases when there is more heterogeneity. Intuitively one would expect agreement to decrease when the measured variable exhibits more variability because the increased variability requires the methods to agree across a wider range of values of the measured variable. For further discussion of agreement and heterogeneity the reader is referred to Atkinson & Navel[10].

We considered three disagreement functions, based on the MSD, MAD and CP. We saw that the CP-based coefficients depend on the threshold used to define 'good agreement', so we actually have to deal with a family of coefficients corresponding to a given range of threshold values. Comparing the MSD and

MAD-based approaches, we prefer the latter as the MAD is expressed in the same units as the measured variable. In our simulation studies the MAD-based estimates were more precise than the corresponding MSD-based estimates except when the true agreement was very small.

The approach introduced in this article can be generalized in various directions. The MSD-based coefficients were generalized to situations where several new methods are compared to a common gold standard or several methods are compared without a gold standard [5]. It will be important to develop multiple-methods coefficients based on other disagreement functions. We considered only 'continuous' variables in this work, and we are currently exploring similar coefficients for binary and categorical outcome variables. Finally, throughout this work we assumed that the replications from the two methods are independent of each other (unmatched replications). We found many examples where this assumption does not hold (for example, repeated measurements are performed at a set of fixed time points or under fixed conditions). We plan to adapt our coefficients to incorporate matched replications and account for the variability associated with the factor that underlies the repeated measurements.

Acknowledgement

This research was supported by NIMH grant 1 R01 MH070028.

References

1. Haber M, Barnhart HX. Coefficients of agreement for fixed observers. *Statistical Methods in Medical Research* 2006 (to appear)
2. Zegers FE. A family of chance-corrected association coefficients for metric scales. *Psychometrika* 1986; **51**: 559-62.
3. Lin L. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 1989; **45:** 255-68.
4. Schall R, Luus HG. On population and individual bioequivalence. *Stat. Med.* 1993; **12**: 1109-24.
5. Barnhart HX, Kosinski AS, Haber M. Assessing individual agreement. (submitted).
6. Lin L, Hedayat AS, Sinha B, Yang M. Statistical methods in assessing agreement: models, issues and tools. *J. Amer. Stat. Assoc.* 2002; **97**; 257-70.
7. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Statistical Methods in Medical Research* 1999; **8**: 135-60.
8. Read CB. Folded Distributions. *Encyclopedia of Statistical Sciences* (edited by S Kotz and NL Johnson) Vol 3, pp 160-1. Wiley & Sons, New York, 1983.
9. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurements. *Lancet* 1986; **i**: 307-10.
10. Atkinson G, Nevill A. Comment on the use of concordance correlation to assess the agreement between variables. *Biometrics* 1997; **53**: 775-778.

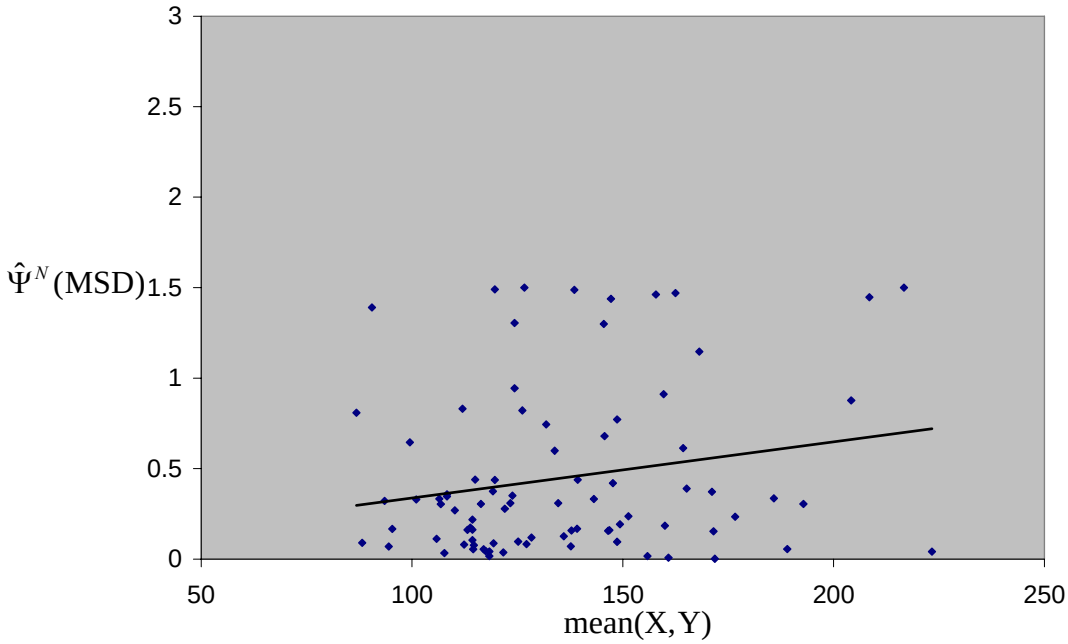
Table 1: Bias and root mean square error (RMSE) of estimates of ψ^N and ψ^R using the mean squared deviation (MSD) as the disagreement function

			Case 1 $\psi^N = 0.266, \psi^R = 0.199$				Case 2 $\psi^N = 0.670, \psi^R = 0.502$				Case 3 $\psi^N = 0.940, \psi^R = 0.704$			
			$\hat{\psi}^N$		$\hat{\psi}^R$		$\hat{\psi}^N$		$\hat{\psi}^R$		$\hat{\psi}^N$		$\hat{\psi}^R$	
N	K	L	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
50	3	3	0.002	0.042	0.001	0.038	0.011	0.074	0.007	0.082	0.006	0.082	0.014	0.109
	2	2	0.009	0.048	0.005	0.047	0.011	0.099	0.011	0.111	0.011	0.130	0.018	0.148
	3	1	–	–	0.008	0.043	–	–	0.011	0.101	–	–	0.030	0.138
	2	1	–	–	0.003	0.054	–	–	0.018	0.127	–	–	0.045	0.195
100	3	3	0.002	0.028	0.004	0.024	-0.001	0.051	-0.001	0.056	0.002	0.055	0.008	0.074
	2	2	0.003	0.036	0.004	0.034	0.003	0.078	0.007	0.078	0.003	0.091	-0.001	0.098
	3	1	–	–	0.001	0.031	–	–	0.013	0.076	–	–	-0.002	0.091
	2	1	–	–	0.008	0.038	–	–	0.000	0.088	–	–	0.014	0.113
200	3	3	0.004	0.020	0.004	0.019	0.000	0.036	-0.002	0.038	-0.002	0.038	0.006	0.054
	2	2	0.001	0.024	0.000	0.023	-0.001	0.052	0.004	0.058	0.002	0.062	-0.002	0.075
	3	1	–	–	-0.001	0.021	–	–	0.003	0.050	–	–	-0.003	0.066
	2	1	–	–	-0.004	0.025	–	–	0.000	0.060	–	–	-0.006	0.088

Table 2: Bias and root mean square error (RMSE) of estimates of ψ^N and ψ^R using the mean absolute deviation (MAD) as the disagreement function

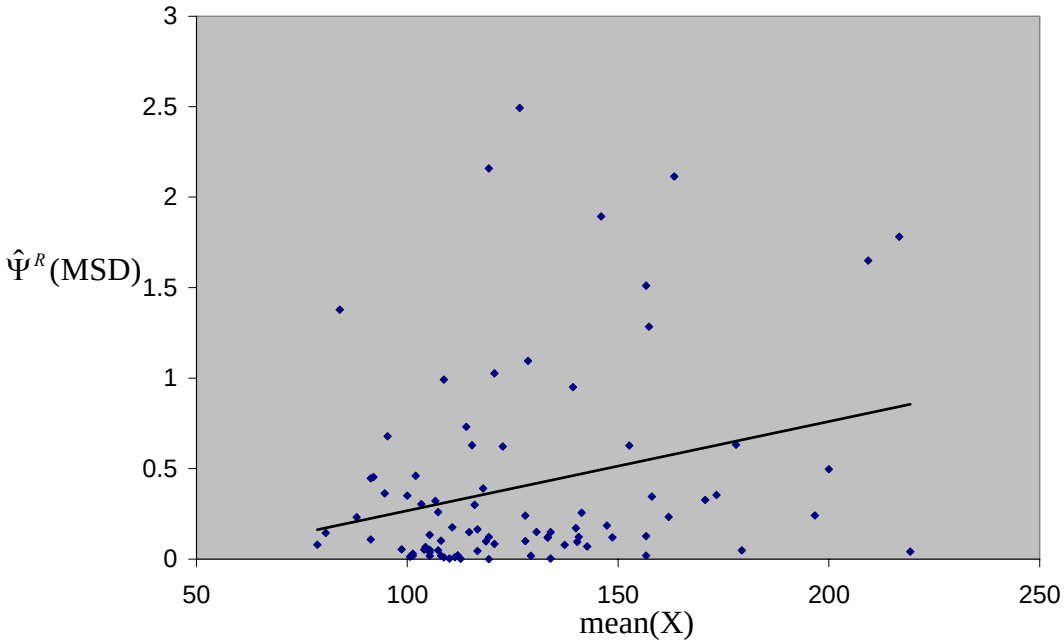
			Case 1 $\psi^N = 0.476, \psi^R = 0.415$		Case 2 $\psi^N = 0.804, \ \psi^R = 0.702$		Case 3 $\psi^N = 0.962, \ \psi^R = 0.839$	
			$\hat{\psi}^N$	$\hat{\psi}^R$	$\hat{\psi}^N$	$\hat{\psi}^R$	$\hat{\psi}^N$	$\hat{\psi}^R$
N	K	L	Bias RMSE	Bias RMSE	Bias RMSE	Bias RMSE	Bias RMSE	Bias RMSE
50	3	3	-0.005 0.036	-0.004 0.039	-0.009 0.051	-0.006 0.056	-0.010 0.046	-0.015 0.064
	2	2	-0.010 0.048	-0.012 0.051	-0.013 0.069	-0.002 0.074	-0.002 0.079	-0.007 0.096
	3	1	– –	-0.005 0.045	– –	-0.014 0.074	– –	0.004 0.085
	2	1	– –	-0.010 0.062	– –	-0.012 0.092	– –	0.006 0.104
100	3	3	-0.010 0.027	-0.011 0.030	-0.013 0.037	-0.015 0.043	-0.003 0.032	-0.008 0.049
	2	2	-0.010 0.034	-0.008 0.039	-0.012 0.048	-0.011 0.060	-0.003 0.049	0.002 0.065
	3	1	– –	-0.009 0.032	– –	-0.011 0.053	– –	-0.003 0.062
	2	1	– –	-0.012 0.038	– –	-0.014 0.067	– –	-0.010 0.074
200	3	3	-0.007 0.018	-0.007 0.019	-0.015 0.027	-0.017 0.033	-0.008 0.024	-0.006 0.033
	2	2	-0.012 0.024	-0.011 0.026	-0.016 0.038	-0.017 0.045	-0.007 0.037	-0.008 0.047
	3	1	– –	-0.014 0.024	– –	-0.018 0.039	– –	-0.011 0.043
	2	1	– –	-0.012 0.029	– –	-0.014 0.046	– –	-0.002 0.058

Figure 1a: $\hat{\Psi}^N$ based on MSD as a function of the mean of X, Y



Least squares line coefficients: α = 0.02770, β = 0.0031;
p-value for testing $H_0 : \beta = 0$ is 0.064

Figure 1b: $\hat{\Psi}^R$ based on MSD as a function of the mean of X



Least squares line coefficients: α = -0.2256, β = 0.049;
p-value for testing $H_0 : \beta = 0$ is 0.012

Figure 2. The ψ coefficients and the CCC as functions of the population standard deviation σ_T