Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Jeffrey Wiener Date

Evaluating Agreement Among Observers or Methods of Measurement for Quantitative Data

By

Jeffrey Wiener
Doctor of Philosophy

Biostatistics

_____

Michael J. Haber, Ph.D.
Advisor

_____

Ying Guo, Ph.D.
Committee Member

_____

Mary E. Kelley, Ph.D.
Committee Member

_____

Andrzej S. Kosinski, Ph.D.
Committee Member

Accepted:

_____

Lisa A. Tedesco, Ph.D.
Dean of the Graduate School

_____

Date

Evaluating Agreement Among Observers or Methods of Measurement for Quantitative Data

By

Jeffrey Wiener
M.S., Emory University, 2004
B.A., University of Rochester, 1998

Advisor: Michael J. Haber, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the Graduate School of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2009

Abstract

Evaluating Agreement Among Observers or Methods of Measurement for Quantitative Data

By Jeffrey Wiener

Agreement measures are used to compare measurements of a specific variable made by different observers or methods, and to evaluate whether a substantial difference exists between these sources of measurement. Assessing agreement is applicable to method comparison or observer reliability studies in both the biomedical and psychosocial sciences. Frequently, a reference method or gold standard exists which is considered to be the most accurate of those available.

First, we explore and evaluate multiple unscaled measures of agreement between quantitative measurements by two observers with and without replications. Two scaled coefficients of agreement based on a general disagreement function which makes no distributional assumptions are described, one for the case of no applicable reference method, and the second for the case where one observer is considered a reference. We develop methods of inference for these coefficients, evaluating them against previously developed methods, and also define the asymptotic distribution of the coefficients and assess the robustness of the estimation methods. Next, we extend the described coefficients of agreement to the case where a set of two or more observers are selected at random from a pool of potential observers.

Finally, we model agreement using a disagreement function as our outcome variable. The effects of subject-specific covariates are examined. We apply these methods to a behavioral intervention study on medication adherence in HIV-positive children and to a carotid stenosis screening method comparison study.

Evaluating Agreement Among Observers or Methods of Measurement for Quantitative Data

By

Jeffrey Wiener
M.S., Emory University, 2004
B.A., University of Rochester, 1998

Advisor: Michael J. Haber, Ph.D.

A dissertation submitted to the Faculty of the Graduate School of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2009

Acknowledgments

Completing this dissertation has been a long, often frustrating, but rewarding journey. My sincerest thanks go to my advisor, Dr. Michael Haber, for his patience and passion for the subject matter. His willingness to persevere and guide me through the many changes in the direction of this research was invaluable.

I'd also like to thank my committee members, Dr Ying Guo, Dr. Mary Kelley, and Dr. Andrzej Kosinski. Their enthusiasm to meet with me and share their experience many times over the years has definitely helped shape this research. I wish to thank Dr. Marc Bulterys, Dr. Denise Jamieson, and Dr. Lillian Lin at the Centers for Disease Control and Prevention for their wonderful mentoring throughout my career. I look forward to collaborating with them on great projects in the future. I am also grateful to my colleagues Dr. Sherri Pals, Dr. Carol Lin, Dr. Renee Stein, and Mary Jo Earp for their terrific advice and friendship. Special thanks go to Dr. Ken Dominguez for allowing me to contribute to the Pediatric Impact project, and use this data to demonstrate these methods.

Lastly, I am forever grateful for the unwavering love, support, and encouragement of my parents, Ronald and Amelia Wiener. Their dedication to academics throughout my life has most certainly led me down this path. I am so thankful to be able to share this achievement with them.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Agreement measures are used to compare measurements of a specific variable made by different observers or methods, and to evaluate whether a substantial difference exists between these sources of measurement. Assessing agreement is applicable to method comparison or observer reliability studies in both the biomedical and psychosocial sciences, where performance and consistency of instruments or assays can be evaluated. Since measurement errors exist for all methods of measurement, it is necessary to evaluate how reliable a method is before recommending it for use in the field. Frequently, a reference method of measurement known as a gold standard exists which is considered to be the most proven and accurate of those available. One can then evaluate the validity of an alternative method by assessing its agreement with the reference method. We use the term observer agreement to represent agreement between measurements made by different persons, or by different methods of measurement.

An extensive body of research exists on assessing agreement in categorical measurements, most notably with the development of the Kappa statistic by Cohen (1960) and the weighted kappa statistic (Cohen, 1968). We focus our investigation on the case where measurements are continuous. Bland and Altman (1986) developed a widely used graphical approach to examine agreement between continuous measurements. Multiple articles by Lin (1989, 1992, 1997, 2000), Lin and Torbeck (1998), and Lin et al.

(2002, 2007) constructed several numerical measures for evaluating agreement between continuous measurements. These numerical measures can be unscaled, measuring absolute differences between measurements, or scaled to attain values only between –1 and 1 for ease of interpretation. We concentrate our research on evaluating scaled and unscaled numerical measures and using them as a basis for developing measures for use in the case where agreement between observers is evaluated, with or without replications, and where one observer may or may not be considered a reference.

Specifically, we plan to focus on the following topics:

I. Describing a general approach to evaluating agreement between two fixed observers with replicated measurements, and developing methods of inference for this approach.

II. Developing methods to extend a general approach to evaluating agreement between two observers with replicated measurements to the case where observers are selected randomly.

III. Developing general models for estimating unscaled measures of agreement and modeling them as a function of covariates.

Chapter 2 concentrates on describing existing statistical methods for evaluating agreement between two observers, with no replicated measurements. Here, we describe previously developed measures for assessing agreement between a single observer and a second observer which may or may not be considered a reference. Through simulation studies, we evaluate multiple ways of computing point and variance estimates for these measures for use in more advanced methods developed in later chapters.

In Chapter 3, we focus on developing methods of inference for a general approach to evaluating agreement between two observers with replicated measurements, thereby addressing topic I. We accomplish this by comparing multiple interval estimates for these measures, and describing the behavior of these measures through simulation studies and application to a real data set.

Chapter 4 focuses on extending the general approach to evaluating agreement developed in Chapter 3, to the case where observers are selected at random from a pool of potential observers. Methods of inference are developed and evaluated through a simulation study.

Chapter 5 focuses on addressing topic III, by developing statistical models for estimating measures of agreement in data with a single measurement, and describing its behavior with real data sets. We also propose a model which expresses observer agreement as a function of covariates.

The data sets to be used in this dissertation to illustrate the methods developed and evaluated are as follows:

Systolic Blood Pressure Data (Bland and Altman (1999)):

Systolic blood pressure was measured on 85 subjects by two 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.

Carotid Stenosis Data (Barnhart and Williamson (2001)):

This data was collected in a carotid stenosis screening study funded by the National Institutes of Health (NIH). The goal of the study was to compare two different

methods using magnetic resonance angiography (MRA) for non-invasive screening of carotid artery stenosis with the current gold standard, an invasive intra-arterial angiogram (IA). Stenosis measurements using both methods and the gold standard were conducted on 55 patients separately on the left and right carotid arteries by each of three radiologists.

Pediatric Impact Adherence Data (Lee et al. (2006)):

The Pediatric Impact study is a behavioral intervention to improve adherence to antiretroviral medications in HIV-positive children ages 5-12, funded by the Centers for Disease Control and Prevention (CDC). The study collected multiple measures of adherence over the course of the intervention. Our interest is in baseline 1-month adherence measures collected before the start of the intervention. At baseline, the caregiver for each child and a member of the clinic's care team were separately asked to estimate the percent of prescribed medication doses taken over the past month. These two adherence measures can then be compared to the percent of medication doses taken over the same month as measured by the Medication Event Monitoring System (MEMS) caps, considered to be the gold standard for measuring adherence.

Chapter 2

Approaches for Evaluating Agreement Between Two Observers

2.1 Introduction and notation

In this chapter we review measures of agreement between quantitative measurements by two observers without replications. The values of the measurements for the two observers are denoted by (*X*) and by (*Y*). The data therefore consist of *n* pairs of measurements (*X*,*Y*) where *n* is the number of subjects evaluated by the observers. The differences between the measurements in a pair are denoted by $D = Y - X$. In some cases, one of the two observers may be considered a reference method, or "gold standard". In these cases, (*X*) will denote measurements using the reference method. We assume *X* and *Y* have continuous distributions, with means μ_x and μ_y, variances σ_x^2 and σ_y^2, and covariance σ_{xy}. Even when *X* is a gold standard, we assume that it is measured with error.

Bland and Altman (1986) describe a graphical method for evaluating the agreement between *X* and *Y*. This approach involves constructing a Bland and Altman plot by plotting the difference D against the mean of each measurement pair, $(X + Y)/2$. One can assess agreement between the two observers by determining an interval which includes a given (high) proportion of the difference *D*. Although this approach is quite helpful in uncovering systematic biases in the data and spotting outliers, it can be difficult

to make a firm decision on whether the level of agreement is suitable enough to validate an alternative method.

We devote this chapter to examining the numerical agreement measures developed and described by Lin (1989, 1992, 1997, 2000), Lin and Torbeck (1998), Lin et al. (2002), and Barnhart, Haber, and Lin (2007), as they are most appropriate for extending to replicated measurements explored in later chapters.

2.2 Existing methods

2.2.1 Mean Squared Deviation (MSD)

The mean squared deviation (MSD) is defined as

$$MSD = \varepsilon^2 = E(D^2). \tag{2.1}$$

This statistic evaluates the squared deviation from the identity line, and can be expressed in terms of the distribution moments as

$$\varepsilon^2 = (\mu_y - \mu_x)^2 + \sigma_y^2 + \sigma_x^2 - 2\sigma_{yx}. \tag{2.2}$$

Using this definition, we can most obviously estimate the MSD by plugging in the common sample estimates for means and variances where

$$e_A^2 = (\bar{y} - \bar{x})^2 + s_y^2 + s_x^2 - 2s_{yx}. \tag{2.3}$$

Lin (2000) demonstrates that when X and Y are normally distributed, $W = \ln(e_A^2)$ has an asymptotic normal distribution with a mean of $w = \ln(e^2)$, and, as indicated in Lin et al. (2002), a variance of

$$\sigma_{A1}^2 = \frac{2[1-(\mu_x-\mu_y)^4/\varepsilon^4]}{n-2}. \tag{2.4}$$

The denominator of this variance estimate is altered from the previous article to reduce bias when the sample size is small, as Lin (2000) uses

$$\sigma_{A2}^2 = \frac{2[1-(\mu_x-\mu_y)^4/\varepsilon^4]}{n-1}. \tag{2.5}$$

Both variance estimates will be evaluated through a simulation study.

An alternative point estimate suggested by Lin et al. (2002) to reduce bias when estimating w, is

$$e_B^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-x_i)^2 . \tag{2.6}$$

It is not clear why the use of $n-1$ as the denominator for this estimate is appropriate, as a more intuitive estimate would simply divide by n:

$$e_C^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i-x_i)^2 . \tag{2.7}$$

Hutson et al. (1998) also uses the equation in (2.7) to estimate MSD when developing a more complex measure based on this estimate.

2.2.2 Intraclass Correlation Coefficient (ICC)

The intraclass correlation coefficient (ICC) has traditionally been used to evaluate agreement between continuous measurements. The ICC was first defined by Galton (1889) as a correlation between measurements of the same class, and was later defined by Fisher (1925) as the ratio of between sample variance and total (between + within

sample) variance under an analysis of variance (ANOVA) model. It is commonly used in the psychosocial sciences to measure observer reliability under classical test theory (Lord and Novick, 1968).

When comparing agreement between replications by a single observer, one can define the ICC using a one-way random effects model (Fleiss, 1986), defined as:

$$Y_{ik} = \mu + \alpha_i + \varepsilon_{ik} \tag{2.8}$$

where $\alpha_i \sim N(0, \sigma_\alpha^2)$ and $\varepsilon_{ijk} \sim N(0, \sigma_\varepsilon^2)$. Here, *i* represents the subject, *k* represents the replication, and ε_{ik} represents the measurement error. Each observer is assumed to take *K* measurements on each subject, and *K*=1 indicates no replication. The ICC for this model is

$$ICC_1 = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\varepsilon^2} \quad . \tag{2.9}$$

When comparing replicated measurements by the same observer, this coefficient is known as the reliability coefficient.

For the case where two or more observers are being compared, the ICC can be defined using a two-way random effects model or mixed model (Fleiss, 1986; McGraw and Wong, 1996), identified as:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ik} \tag{2.10}$$

where $j = 1,...,J$ represents the fixed or randomly selected observer, $\beta_j \sim N(0, \sigma_\beta^2)$, and the rest of the notation is the same as for (2.8). The ICC for agreement for this model is

$$ICC_2 = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\varepsilon^2} \quad . \tag{2.11}$$

If the observer is treated as a fixed effect in this model, σ_β^2 is defined as

$\sigma_\beta^2 = \sum_{j=1}^{J} \beta_j^2 /(J-1)$.

When it is appropriate to model observer-subject interaction, the two-way model can be defined as:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ik} \tag{2.12}$$

where $\gamma_{ij} \sim N(0, \sigma_\gamma^2)$, and *k* once again represents the replicated measurement by observer *j* on subject *i*. Now when the observer is considered a random effect, the ICC will be calculated as

$$ICC_3 = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2} \quad . \tag{2.13}$$

If the observer is treated as a fixed effect, σ_β^2 is defined as $\sigma_\beta^2 = \sum_{j=1}^{J} \beta_j^2 /(J-1)$, and the ICC is calculated as

$$ICC_3 = \frac{\sigma_\alpha^2 - \sigma_\gamma^2 /(J-1)}{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2} \quad . \tag{2.14}$$

The ICC is defined under the assumption that variances of measurements are the same over multiple observers. This assumption is not always reasonable, and if not met, can underestimate agreement. The ICC is also quite sensitive to between-subject variation. Haber and Barnhart (2006) present the concept of observer relational agreement, which derives the ICC's without making the restrictive ANOVA assumptions. We only consider the ICC's developed to measure observer agreement, as defined in McGraw and Wong (1996).

2.2.3 Concordance Correlation Coefficient (CCC)

The concordance correlation coefficient (CCC), denoted by ρ_c, was introduced by Lin (1989) for fixed observers, and is computed by standardizing the MSD as

$$\rho_c = 1 - \frac{\varepsilon^2}{\varepsilon^2 \mid \rho = 0} = \frac{2\sigma_{xy}}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}, \tag{2.15}$$

where ρ is the Pearson correlation coefficient. The CCC scales the MSD along the 45° degree line, therefore measuring the degree of variation from this line, effectively converting the MSD into a correlation coefficient. It ranges from -1 to 1, with a value of 1 indicating perfect agreement, a value of 0 indicating no agreement, and a value of -1 indicating perfect reverse agreement.

The CCC can be expressed as the product of an accuracy component and a precision component. The accuracy component measures how close the best fit line is to the 45° line, and the precision component measures how close the data points are to the 45° line. The precision component is equivalent to the Pearson correlation coefficient (ρ), and the accuracy component is defined as

$$\chi_a = \frac{2}{\frac{\sigma_y}{\sigma_x} + \frac{\sigma_x}{\sigma_y} + \frac{(\mu_y - \mu_x)^2}{\sigma_y \sigma_x}}. \tag{2.16}$$

Then $\rho_c = \rho \cdot \chi_a$. The CCC can be estimated by plugging in common sample estimates for means, variances, and the Pearson correlation coefficient as

$$r_c = \frac{2rs_y s_x}{s_y^2 + s_x^2 + (\bar{y} - \bar{x})^2}. \tag{2.17}$$

Lin et al. (2002) recommends the CCC over the ICC when assessing agreement between continuous variables, since the ICC does not have meaningful components of accuracy and precision. However, similar to the ICC, the CCC is also sensitive to between-subject variation. Barnhart and Williamson (2001) developed a generalized estimating equations (GEE) approach to model CCC, which can effectively adjust the agreement measure for covariates. If there are no replicated observations, Carrasco and Jover (2003) showed that the ICC defined in (2.11) is equal to the CCC even when the ANOVA model assumptions are not correct.

To yield the best normal approximation, one uses Fisher's Z-transformation to define a measure as a function of the CCC estimate in (2.17),

$$Z_{XY} = \frac{1}{2}\log(\frac{1+r_c}{1-r_c}) \quad . \tag{2.18}$$

Lin (1989) shows this measure to have an asymptotic normal distribution with a mean of $\log[(1+\rho_c)/(1-\rho_c)]/2$ and a variance of

$$\sigma_Z^2 = \frac{1}{n-2}[\frac{(1-\rho^2)\rho_c^2}{(1-\rho_c^2)\rho^2} + \frac{2\upsilon^2(1-\rho_c)\rho_c^3}{(1-\rho_c^2)^2\rho} - \frac{\upsilon^4\rho_c^4}{2(1-\rho_c^2)^2\rho^2}] \tag{2.19}$$

where

$$\upsilon^2 = \frac{(\mu_y - \mu_x)^2}{\sigma_y\sigma_x} \quad , \tag{2.20}$$

assuming that r_c is the sample concordance correlation coefficient of paired samples from a bivariate normal distribution.

2.2.4 Total Deviation Index (TDI)

The total deviation index (TDI) developed by Lin (2000) is defined as the cutpoint, κ, where a set proportion, π, of the absolute values of *D* fall below this cutpoint.

The TDI can be estimated simply by taking the $(100 \cdot \pi)^{th}$ percentile of the absolute values of *D*. If we assume the distribution of *D* to be normal, Lin (2000) shows that the TDI can be computed as the inverse of a cumulative noncentral chi-squared distribution, but contends that inference based on this estimate is intractable. Lin et al. (2002) suggests approximating the TDI with the following estimate:

$$\kappa_{\pi} = \Phi^{-1}(1 - \frac{1-\pi}{2}) \mid \varepsilon \mid \tag{2.21}$$

where $\Phi^{-1}(\cdot)$ is the inverse cumulative normal distribution and ε is estimated by replacement with an appropriate estimate of MSD described in section 2.2.1. Inference on this estimate can then be performed using the method proposed for the MSD. Lin lists a series of boundaries for μ^2/σ^2 (computed for *D*) based on differing values of π under which this approximation is valid. These boundaries seem somewhat limiting, and will restrict how often this estimate can be used. Choudhary and Nagaraja (2007) proposed an exact test for inference on the TDI for data with a small sample size and a bootstrap test for data with a moderate sample size.

2.2.5 Coverage Probability (CP)

The Coverage Probability (CP), proposed by Lin et al. (2002), does the reverse of the TDI and computes a value of π for a given value of κ. Therefore, the CP is represented as

$$CP_{\kappa} = P(|Y - X| < \kappa). \tag{2.22}$$

Again assuming the distribution of D to be normal, Lin suggests estimating the CP with

$$p_k = \Phi[\frac{\kappa - \hat{\mu}_d}{s_d}] - \Phi[\frac{-\kappa - \hat{\mu}_d}{s_d}], \tag{2.23}$$

where $\hat{\mu}_d = \bar{y} - \bar{x}$ and $s_d^2 = \frac{n}{n-3}(s_y^2 + s_x^2 - 2s_{yx})$. The variance of the estimate in (2.21) can then be computed using

$$\sigma_p^2 = \frac{1}{n-3}\{[\Phi(\frac{-\kappa - \mu_d}{\sigma_d}) - \Phi(\frac{\kappa - \mu_d}{\sigma_d})]^2$$

$$+ \frac{1}{2}[\frac{\kappa - \mu_d}{\sigma_d}\Phi(\frac{\kappa - \mu_d}{\sigma_d}) + \frac{\kappa + \mu_d}{\sigma_d}\Phi(\frac{-\kappa - \mu_d}{\sigma_d})]^2\} + O(\frac{1}{n^2}), \tag{2.24}$$

which is estimated by plugging in $\hat{\mu}_d$ and s_d^2 for μ_d and σ_d^2, respectively. The logit transformation of (2.23) is recommended for inference, since CP is a probability bounded by 0 and 1.

The CP can also be estimated by computing the simple relative frequency of values of $|D|$ in the sample falling below a specified κ. This method removes the assumption of normality, which may restrict usage of (2.23) in practice.

2.3 Evaluation of existing methods

We performed a simulation study by generating 1,000 samples of sizes 25, 50, and 100 from each of two bivariate normal distributions and a scenario where both *X* and *Y* were exponentially distributed. Estimates of MSD were computed for (2.3), (2.6), and

(2.7) and averaged over all simulations. Variance estimates using both (2.4) and (2.5) were computed and used to compute standard normal 95% confidence intervals, which were then used to compute overall coverage probabilities.

The results of these simulations are presented in Tables 2.1 and 2.2. The MSD estimates are comparable with larger sample sizes, but the bias is lowest using e_C^2 when n = 25, suggesting this estimate performs best with small sample sizes. Using e_C^2 as our estimate, the coverage probabilities are closer to 0.95 using the variance estimate in (2.5) (especially evident with the latter two distributions), implying this is the best combination to use for inference on the MSD.

Using the same simulations used to evaluate estimates of the MSD, we calculated averages over simulations for both the simple percentile estimate and Lin's normal approximation estimate (2.21) for the TDI over three different values of π, and used a large sample percentile for comparison. The results of these simulations are presented in Table 2.3. The percentile estimate performs well for both normal and non-normal data, but is less accurate in small samples with larger values of π. The normal approximation estimate performs well for normal data and with better precision than the percentile estimate, but is not appropriate when the data is not normal. This will limit its usage if the normality assumption is violated.

Figure 2.1 shows the range of TDI percentile estimates over all values of π for each of the three simulated distributions where n = 100. For each simulation, the TDI estimate increases gradually over increasing values of π, with a sharp spike upward after π = 0.90. This indicates that the TDI estimates for values of π greater than 0.90 are highly variable.

Once again, the same three simulated distributions were used to compute estimates of CP for differing values of κ using Lin's normal distribution estimate for CP (2.21) and the simple relative frequency estimate, and averaged over all 1,000 simulations. Variances for each estimate were also calculated, using the equation in (2.22) for Lin's estimate and the binomial variance formula for the relative frequency estimate. The results of the simulations are presented in Tables 2.4 and 2.5. When comparing against a large sample relative frequency, the relative frequency estimate performs well over all distributions and values of n and κ. The estimate in (2.21) performs quite poorly for small sample sizes, even with normally distributed data, and as expected, is not useful for non-normal data. Using coverage probabilities computed from standard normal 95% confidence intervals, we conclude that both the variance in (2.22) and the binomial variance are too small when n is small, but are appropriate with larger sample sizes. There also seem to be problems computing variances when the value of κ is high since the coverage probabilities are often very low.

Figure 2.2 plots relative frequency estimates of CP across multiple values of κ for each of the three simulated distributions when $n = 100$. CP increases more slowly with higher values of κ when the correlation between X and Y is lower, which is appropriate given that we expect larger values of D. This demonstrates that the strategy for choosing κ will depend greatly on the data itself, and prior calculations of correlation or agreement may be necessary.

2.4 Discussion

Each of the agreement measures examined in this chapter may be useful depending on the needs of a study.

As described by Lin et al. (2002), the asymptotic power (*i.e.* the asymptotic statistical power to conclude that good agreement exists using a given measure) of CCC is inferior to that of MSD and TDI, and is dependent on between-subjects variation, but can be much easier to interpret and compare across studies given the range is always the same. CP and TDI are preferable over CCC for making statistical inferences using a measure of agreement, given their higher power. The CCC also can be separated into measures of accuracy and precision, which are helpful in determining the reasons for lack of agreement. This measure has been featured quite heavily in recent statistical literature, with numerous extensions of it beyond the single-pair agreement case.

The MSD is more difficult to interpret, given it depends heavily on the range of the data, but will be more useful in developing more general measures later in this dissertation since one can easily compare two MSD's measured on the same scale. Two similar alternative measures to the MSD were introduced in Haber and Barnhart (2008). These include the mean absolute difference (MAD), defined as $MAD = E \mid Y - X \mid$, and the mean relative difference (MRD), defined as $MRD = E(\mid Y - X \mid / X)$. The TDI is intuitive and much easier to interpret, but inferential methods will be difficult with data that is not normally distributed given the difficulty in working theoretically with percentiles. The CP, although intuitively clear, will require early examination of the data to select an appropriate κ, and estimation is much less accurate for small sample sizes and non-normal distributions.

Table 2.1: Simulation results for MSD estimates based on 1000 samples.

Distribution	n	MSD	e^2_A Mean	e^2_A Bias	e^2_A MSE	e^2_B Mean	e^2_B Bias	e^2_B MSE	e^2_C Mean	e^2_C Bias	e^2_C MSE
Bivariate Normal	25	0.842	0.870	0.028	0.061	0.873	0.031	0.061	0.838	-0.004	0.056
mean = (0.15,0) ρ = 0.60	50		0.848	0.006	0.028	0.849	0.007	0.028	0.832	-0.010	0.027
variance = (1.15, 1/1.15)	100		0.845	0.003	0.015	0.845	0.003	0.015	0.837	-0.005	0.014
Bivariate Normal	25	0.142	0.147	0.005	0.002	0.148	0.006	0.002	0.142	0.0003	0.002
mean = (0.15,0) ρ = 0.95	50		0.143	0.001	0.001	0.143	0.001	0.001	0.141	-0.001	0.001
variance = (1.15, 1/1.15)	100		0.143	0.001	0.0003	0.143	0.001	0.0004	0.141	-0.001	0.0003
X ~ Exp(0.15)	25	0.131	0.134	0.003	0.013	0.135	0.004	0.014	0.130	-0.002	0.013
Y|X ~ Exp(x + 0.15)	50		0.132	0.001	0.007	0.133	0.001	0.006	0.130	-0.001	0.006
	100		0.135	0.004	0.004	0.135	0.004	0.004	0.134	0.003	0.004

Table 2.2: Simulation results for MSD coverage probabilities based on 1000 samples. Coverage probabilities computed for 95% confidence intervals using indicated s^2 estimate of σ^2_W.

Distribution	n	σ^2_{A1}	σ^2_{A2}	e^2_A Var(w)	e^2_A Cover. Prob. 1	e^2_A Cover. Prob. 2	e^2_B Var(w)	e^2_B Cover. Prob. 1	e^2_B Cover. Prob. 2	e^2_C Var(w)	e^2_C Cover. Prob. 1	e^2_C Cover. Prob. 2
Bivariate Normal	25	0.087	0.083	0.083	0.954	0.951	0.084	0.954	0.951	0.084	0.955	0.944
mean = (0.15,0) ρ = 0.60	50	0.042	0.041	0.041	0.947	0.947	0.041	0.947	0.947	0.041	0.943	0.941
variance = (1.15, 1/1.15)	100	0.020	0.020	0.021	0.949	0.949	0.021	0.949	0.949	0.021	0.946	0.944
Bivariate Normal	25	0.085	0.081	0.081	0.951	0.949	0.081	0.952	0.949	0.081	0.754	0.944
mean = (0.15,0) ρ = 0.95	50	0.041	0.040	0.039	0.949	0.947	0.039	0.949	0.947	0.039	0.677	0.943
variance = (1.15, 1/1.15)	100	0.020	0.020	0.020	0.951	0.951	0.020	0.954	0.952	0.020	0.634	0.948
X ~ Exp(0.15)	25	0.084	0.081	0.416	0.604	0.598	0.416	0.609	0.601	0.416	0.528	0.578
Y|X ~ Exp(x + 0.15)	50	0.040	0.040	0.243	0.549	0.545	0.243	0.553	0.547	0.243	0.457	0.532
	100	0.020	0.020	0.149	0.513	0.511	0.149	0.514	0.513	0.149	0.421	0.508

Table 2.3: Simulation results for TDI_π based on 1,000 samples.

		$\pi = 0.80$			$\pi = 0.85$			$\pi = 0.90$		
Distribution	n	Large Sample TDI_π percentile*	Mean (se) TDI_π percentile	Mean (se) κ_π	Large Sample TDI_π percentile	Mean (se) TDI_π percentile	Mean (se) κ_π	Large Sample TDI_π percentile	Mean (se) TDI_π percentile	Mean (se) κ_π
Bivariate Normal	25	1.173	1.144	1.161	1.321	1.272	1.305	1.507	1.433	1.491
mean = (0.15,0) $\rho = 0.60$			(0.200)	(0.164)		(0.217)	(0.185)		(0.234)	(0.211)
variance = (1.15, 1/1.15)	50		1.154	1.163		1.292	1.307		1.466	1.493
			(0.142)	(0.116)		(0.153)	(0.131)		(0.177)	(0.149)
	100		1.160	1.169		1.298	1.313		1.480	1.501
			(0.102)	(0.084)		(0.115)	(0.095)		(0.129)	(0.108)
Bivariate Normal	25	0.485	0.472	0.479	0.544	0.526	0.538	0.620	0.591	0.615
mean = (0.15,0) $\rho = 0.95$			(0.082)	(0.067)		(0.089)	(0.075)		(0.097)	(0.086)
variance = (1.15, 1/1.15)	50		0.475	0.478		0.532	0.537		0.602	0.614
			(0.059)	(0.047)		(0.064)	(0.052)		(0.070)	(0.060)
	100		0.480	0.481		0.537	0.540		0.609	0.617
			(0.042)	(0.034)		(0.045)	(0.038)		(0.053)	(0.044)
$X \sim Exp(0.15)$	25	0.336	0.326	0.434	0.409	0.394	0.488	0.524	0.493	0.558
$Y\|X \sim Exp(x + 0.15)$			(0.095)	(0.155)		(0.117)	(0.174)		(0.154)	(0.199)
	50		0.330	0.446		0.398	0.502		0.501	0.573
			(0.068)	(0.119)		(0.083)	(0.134)		(0.113)	(0.153)
	100		0.333	0.459		0.405	0.516		0.515	0.589
			(0.050)	(0.094)		(0.061)	(0.105)		(0.085)	(0.120)

* Large sample percentiles based on simulated sample of n=100,000.

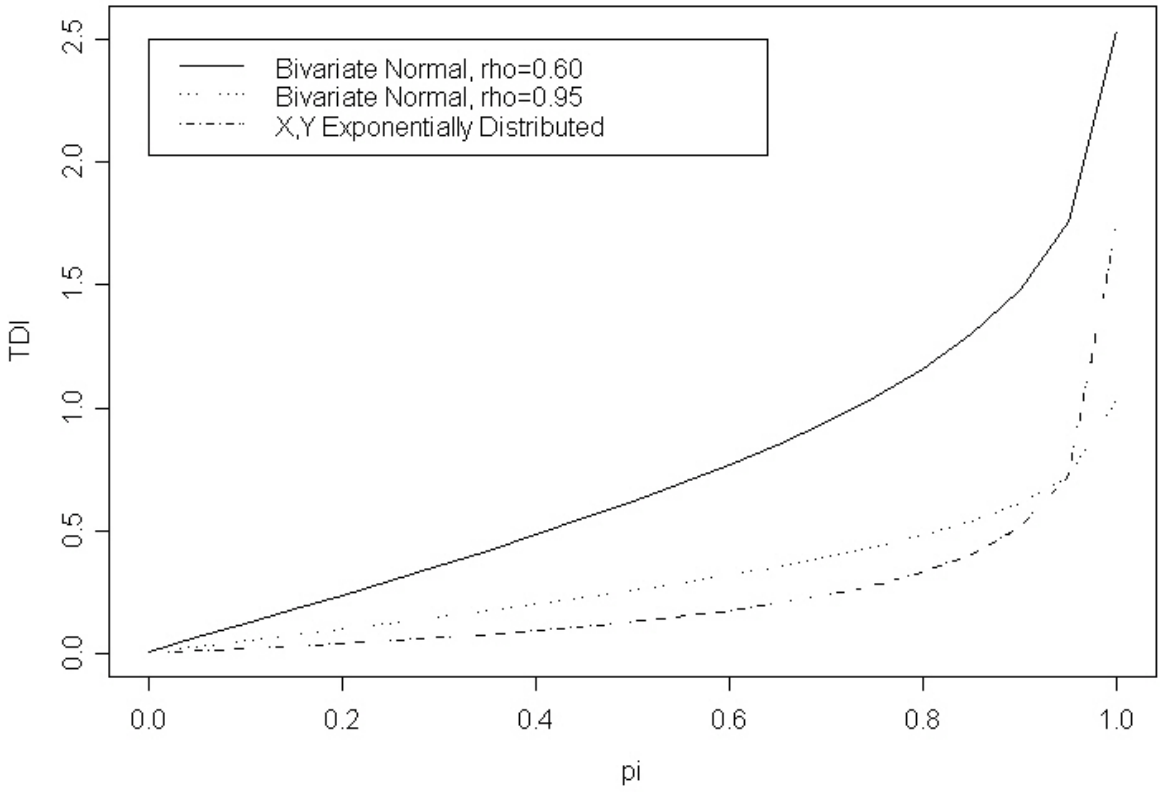

Figure 2.1: TDI percentile estimates averaged over 1000 simulations with n=100, for multiple values of π.

Table 2.4: Simulation results for CP_κ based on 1,000 samples.

		$\kappa = 0.25$			$\kappa = 0.50$			$\kappa = 1.00$		
Distribution	n	Large Sample CP_κ rel. freq.	Mean CP_κ rel. freq.	Mean p_κ	Large Sample CP_κ rel. freq.	Mean CP_κ rel. freq.	Mean p_κ	Large Sample CP_κ rel. freq.	Mean CP_κ rel. freq.	Mean p_κ
Bivariate Normal	25	0.213	0.215	0.205	0.415	0.412	0.395	0.725	0.727	0.696
mean = (0.15,0) $\rho = 0.60$	50		0.216	0.211		0.417	0.406		0.726	0.712
variance = (1.15, 1/1.15)	100		0.215	0.213		0.414	0.411		0.727	0.718
Bivariate Normal	25	0.491	0.488	0.468	0.813	0.815	0.787	0.993	0.992	0.985
mean = (0.15,0) $\rho = 0.95$	50		0.492	0.482		0.816	0.804		0.993	0.990
variance = (1.15, 1/1.15)	100		0.492	0.486		0.815	0.809		0.993	0.991
$X \sim Exp(0.15)$	25	0.714	0.719	0.538	0.892	0.894	0.834	0.975	0.977	0.981
$Y\|X \sim Exp(x + 0.15)$	50		0.717	0.528		0.895	0.836		0.976	0.987
	100		0.716	0.516		0.893	0.831		0.975	0.989

Table 2.5: Simulation results for CP_κ coverage probabilities for 95% confidence intervals based on 1,000 samples.

		$\kappa = 0.25$		$\kappa = 0.50$		$\kappa = 1.00$	
Distribution	n	Cover. Prob. $CP_\kappa \pm 1.96$ $(CP_\kappa^* (1- CP_\kappa))/n$	Cover. Prob. $p_\kappa \pm 1.96$ $\hat{\sigma}^2_\kappa$	Cover. Prob. $CP_\kappa \pm 1.96$ $(CP_\kappa^* (1- CP_\kappa))/n$	Cover. Prob. $p_\kappa \pm 1.96$ $\hat{\sigma}^2_\kappa$	Cover. Prob. $CP_\kappa \pm 1.96$ $(CP_\kappa^* (1- CP_\kappa))/n$	Cover. Prob. $p_\kappa \pm 1.96$ $\hat{\sigma}^2_\kappa$
Bivariate Normal	25	0.914	0.843	0.905	0.959	0.916	0.967
mean = (0.15,0) $\rho = 0.60$	50	0.927	0.933	0.941	0.969	0.932	0.975
variance = (1.15, 1/1.15)	100	0.950	0.973	0.942	0.982	0.950	0.980
Bivariate Normal	25	0.930	0.945	0.844	0.975	0.185	0.995
mean = (0.15,0) $\rho = 0.95$	50	0.954	0.961	0.910	0.991	0.312	0.999
variance = (1.15, 1/1.15)	100	0.950	0.973	0.917	0.987	0.514	1.000
$X \sim Exp(0.15)$	25	0.934	0.445	0.936	0.702	0.452	0.386
$Y\|X \sim Exp(x + 0.15)$	50	0.952	0.246	0.907	0.624	0.689	0.339
	100	0.947	0.059	0.911	0.547	0.914	0.271

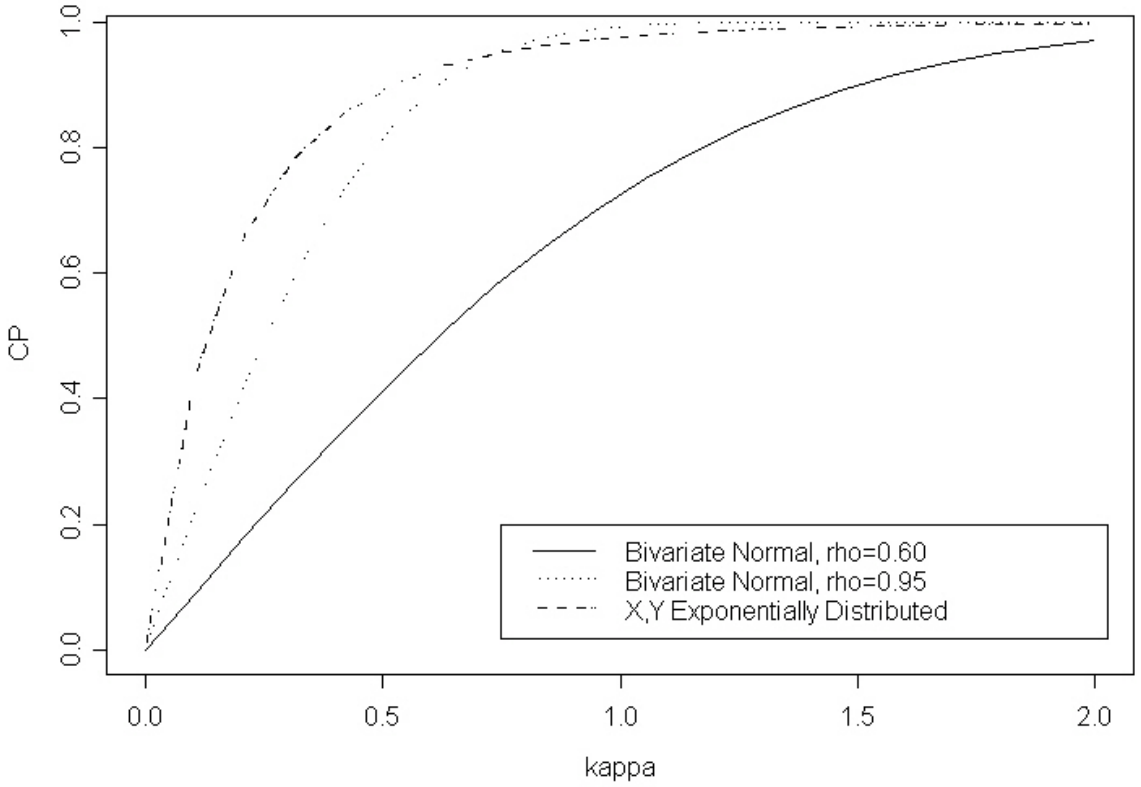

Figure 2.2: CP relative frequency estimates averaged over 1000 simulations with *n*=100, for multiple values of κ.

Chapter 3

A General Approach for Evaluating Agreement Between Two Observers with Replicated Measurements

3.1 Introduction and notation

The previous chapter described several unscaled coefficients of agreement, such as the MSD and the TDI, and several scaled coefficients of agreement, such as different versions of the ICC and the CCC. Here, we seek to describe a general approach to developing a scaled measure of agreement which can be used on agreement data recorded by two observers with replications. The ICC's are usually determined by an ANOVA model which makes restrictive assumptions, such as equal error variance for all observers compared. The CCC compares the MSD between observers to that expected under "chance agreement", or independence between observers. Haber and Barnhart (2006) show that independence or a lack of correlation between observers does not always imply a lack of agreement. This chapter concentrates on the general inter-observer coefficient developed by Haber and Barnhart (2008), which is not bound by the assumption of normality, and can be used when one observer is considered a reference, or when neither observer is considered a reference.

We once again denote the two observers by (X) and (Y), with (X) referring to the reference method if it exists. The data will consist of multiple replications of both (X)

and (Y) for n subjects, with replications denoted as $X_1, X_2, X_3,$ etc. and pairs of measurements comparing any two unmatched replications as (X, X').

3.2 Coefficients of Individual Agreement

The objective is to describe a coefficient of agreement which lies close to one when the two methods are in good agreement, and lies close to zero when the two methods are in poor agreement. Barnhart et al. (2007) and Haber and Barnhart (2007) developed two such methods based on the concept of a disagreement function. This concept is closely linked to the concept of individual bioequivalence in bioequivalence studies (Anderson and Hauck, 1990). One coefficient applies to the case of no applicable reference method, and the second to the case when one observer is considered a reference.

The disagreement function is defined as $G(X,Y)$, where X and Y are measurements made by two observers on the same subject, and it must satisfy the following two conditions:

1) $G(X,Y) \geq 0$

2) $G(X,Y)$ increases as the disagreement between X and Y increases.

We denote by $G(X,X')$ the disagreement between two replicated measurements made by the same observer on the same subject. Previously described unscaled measures of agreement such as the MSD and TDI qualify as appropriate disagreement functions.

The proposed coefficients of agreement seek to compare the disagreement between measurements made by the same observer with the disagreement between

measurements made by different observers. Two different observers with a disagreement function similar to that between replicated measurements by the same observer should have good agreement, and a coefficient of agreement close to one.

For the case that neither of the two observers are considered a reference, the coefficient of individual agreement is defined as

$$\psi^N = \frac{[G(X,X') + G(Y,Y')]/2}{G(X,Y)}. \tag{3.1}$$

Here, at least two replications for each observer are necessary to define ψ^N.

For the case that one observer is considered a reference, the coefficient of individual agreement is defined as

$$\psi^R = \frac{G(X,X')}{G(X,Y)}. \tag{3.2}$$

Replicated measurements are only required for the reference observer to define ψ^R. The values of ψ^N and ψ^R should usually fall between 0 and 1, although values above 1 are possible in certain cases. The coefficient is interpreted as representing poor agreement between observers when it is closest to 0, and agreement gets better as the coefficient increases. A value close to or exceeding 1 indicates that agreement between observers is very good, since the disagreement between observers is similar to the disagreement between replicated measurements by the same observer.

To estimate ψ^N and ψ^R, one must only plug in appropriate estimates for the chosen disagreement function, such as those defined for MSD and TDI in Chapter 2. For example, to estimate ψ^N using the MSD as the disagreement function one can use

$$\hat{\psi}^N = \frac{[\hat{MSD}(X, X') + \hat{MSD}(Y, Y')]/2}{\hat{MSD}(X, Y)} \quad , \tag{3.3}$$

where $\hat{MSD}$ can be calculated using the estimate, e^2, defined in (2.7).

As discussed in Chapter 2, the log estimate of the MSD, $w = \ln(\hat{MSD}(X, Y))$, has an asymptotic normal distribution with the variance estimate specified in (2.5), assuming an underlying bivariate normal distribution. The log of the MSD estimates in (3.3) will be jointly asymptotically normal, and their variance estimates taken from the previous chapter can be denoted as:

$$Var(\ln(\hat{MSD}(X, Y))) = \frac{2[1 - (\mu_x - \mu_y)^4 / MSD(X, Y)^4]}{n-1} \quad , \tag{3.4}$$

$$Var(\ln(\hat{MSD}(X, X'))) = \frac{2[1 - (\mu_x - \mu_{x'})^4 / MSD(X, X')^4]}{n-1} \quad .$$

The last expression reduces to $Var(\ln(\hat{MSD}(X, X'))) = \frac{2[1/MSD(X, X')^4]}{n-1}$ since $\mu_x = \mu_{x'}$.

Using the delta method, we can extend the asymptotic properties of the estimate in (3.3) to show that $\ln(\hat{\psi}^N)$ is also asymptotically normal with a mean of $\ln(\psi^N)$.

Theorem 3.1. As $n \to \infty$, $\sqrt{n}(\ln(\hat{\psi}^N) - \ln(\psi^N))$ is asymptotically normal with mean 0 and a variance of

$$Var(\ln(\hat{\psi}^N)) = \left(\frac{MSD(X, X')}{MSD(X, X') + MSD(Y, Y')}\right)^2 [Var(\ln(\hat{MSD}(X, X'))]$$
$$+ \left(\frac{MSD(Y, Y')}{MSD(X, X') + MSD(Y, Y')}\right)^2 [Var(\ln(\hat{MSD}(Y, Y'))] + Var(\ln(\hat{MSD}(X, Y)))$$

$$+ 2(MSD(X,X'))(MSD(Y,Y'))(\frac{1}{MSD(X,X')+MSD(Y,Y')})^2 Cov(\ln(M\hat{S}D(X,X')),\ln(M\hat{S}D(Y,Y')))$$
$$-(\frac{2MSD(X,X')}{MSD(X,X')+MSD(Y,Y')})[Cov(\ln(M\hat{S}D(X,Y)),\ln(M\hat{S}D(X,X')))]$$
$$-(\frac{2MSD(Y,Y')}{MSD(X,X')+MSD(Y,Y')})Cov(\ln(M\hat{S}D(X,Y)),\ln(M\hat{S}D(Y,Y')))] \quad (3.5)$$

Theorem 3.2. As $n \to \infty$, $\sqrt{n}(\ln(\hat{\psi}^R) - \ln(\psi^R))$ is asymptotically normal with mean 0 and a variance of

$$Var(\ln(\hat{\psi}^R)) = Var(\ln(M\hat{S}D(X,X'))) + Var(\ln(M\hat{S}D(X,Y)))$$
$$- 2Cov(\ln(M\hat{S}D(X,X')),\ln(M\hat{S}D(X,Y))) \quad . \quad (3.6)$$

***Proof*:**

The asymptotic distribution of $\ln(\hat{\psi}^R)$ where MSD is used as the disagreement function,

$$\hat{\psi}^R = \frac{M\hat{S}D(X,X')}{M\hat{S}D(X,Y)} \quad , \quad (3.7)$$

is much simpler to specify. Once again using the delta method, $\ln(\hat{\psi}^R)$ is asymptotically normal with mean of $\ln(\psi^R)$ and the variance specified in (3.6).

To estimate the variances in (3.5) and (3.6), the estimated variances $Var(\ln(M\hat{S}D(X,Y)))$, $Var(\ln(M\hat{S}D(X,X')))$, and $Var(\ln(M\hat{S}D(Y,Y')))$ can be calculated using the equations in (3.4). The covariance $Cov(\ln(M\hat{S}D(X,X')),\ln(M\hat{S}D(Y,Y')))$ is assumed to be 0, since we can assume intraobserver disagreement is independent between observers. The covariance $Cov(\ln(M\hat{S}D(X,Y)),\ln(M\hat{S}D(X,X')))$ is more complicated to compute. We use a technique similar to that used by Hutson et al. (1998) when defining

the asymptotic distribution of a measure of relative agreement. First, we define our estimates in the form:

$$\ln(\hat{MSD}(X,Y)) = g(z_1,z_2,z_3,z_4,z_5) = \ln(z_1 + z_2 - 2z_4)$$

$$\ln(\hat{MSD}(X,X')) = g(z_1,z_2,z_3,z_4,z_5) = \ln(z_2 + z_3 - 2z_5) \quad ,$$

where $\mathbf{z} = (z_1,z_2,z_3,z_4,z_5) = (\frac{1}{n}\sum Y_i^2, \frac{1}{n}\sum X_i^2, \frac{1}{n}\sum X_i'^2, \frac{1}{n}\sum X_i Y_i, \frac{1}{n}\sum X_i' Y_i)$.

We use the theory of functions of asymptotically normal vectors (Serfling (1980)), to define $E(\mathbf{z}) = (\mu'_{200}, \mu'_{020}, \mu'_{002}, \mu_{110}, \mu'_{101})$, where μ'_{rst} represents the trivariate moment $\mu'_{rst} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y^r x^s x'^t dF$. The partial derivatives of our estimates with respect to **z** are

$$\mathbf{d} = (\frac{\partial[\ln(\hat{MSD}(X,Y)]}{\partial z_1}\Big|\mathbf{z} = E(\mathbf{z}),...,\frac{\partial[\ln(\hat{MSD}(X,Y)]}{\partial z_5}\Big|\mathbf{z} = E(\mathbf{z}))$$

$$= (\frac{1}{\mu'_{200} + \mu'_{020} - 2\mu'_{110}}, \frac{1}{\mu'_{200} + \mu'_{020} - 2\mu'_{110}}, 0, \frac{-2}{\mu'_{200} + \mu'_{020} - 2\mu'_{110}}, 0)$$

and $$\mathbf{h} = (\frac{\partial[\ln(\hat{MSD}(X,X')]}{\partial z_1}\Big|\mathbf{z} = E(\mathbf{z}),...,\frac{\partial[\ln(\hat{MSD}(X,X')]}{\partial z_5}\Big|\mathbf{z} = E(\mathbf{z}))$$

$$= (0, \frac{1}{\mu'_{020} + \mu'_{002} - 2\mu'_{101}}, \frac{1}{\mu'_{020} + \mu'_{002} - 2\mu'_{101}}, 0, \frac{-2}{\mu'_{020} + \mu'_{002} - 2\mu'_{101}}) .$$

We can define the variance-covariance matrix of **z** as

$$\Sigma = \begin{bmatrix} \mu'_{400} - \mu'^2_{200} & \mu'_{220} - \mu'_{200}\mu'_{020} & \mu'_{202} - \mu'_{200}\mu'_{002} & \mu'_{310} - \mu'_{200}\mu'_{110} & \mu'_{301} - \mu'_{200}\mu'_{101} \\ & \mu'_{040} - \mu'^2_{020} & \mu'_{022} - \mu'_{020}\mu'_{002} & \mu'_{130} - \mu'_{110}\mu'_{020} & \mu'_{121} - \mu'_{020}\mu'_{101} \\ & & \mu'_{004} - \mu'^2_{002} & \mu'_{112} - \mu'_{110}\mu'_{002} & \mu'_{103} - \mu'_{002}\mu'_{101} \\ & & & \mu'_{220} - \mu'^2_{110} & \mu'_{211} - \mu'_{110}\mu'_{101} \\ & & & & \mu'_{202} - \mu'^2_{101} \end{bmatrix}$$

Now, the covariance can be calculated as

$$Cov(\ln(\hat{MSD}(X,Y)),\ln(\hat{MSD}(X,X'))) = \mathbf{d}\Sigma\mathbf{h}' = d_1(h_1\sigma_{11} + h_2\sigma_{21} + h_3\sigma_{31} + h_4\sigma_{41} + h_5\sigma_{51})$$
$$+ d_2(h_1\sigma_{12} + h_2\sigma_{22} + h_3\sigma_{32} + h_4\sigma_{42} + h_5\sigma_{52})$$
$$+ d_3(h_1\sigma_{13} + h_2\sigma_{23} + h_3\sigma_{33} + h_4\sigma_{43} + h_5\sigma_{53})$$
$$+ d_4(h_1\sigma_{14} + h_2\sigma_{24} + h_3\sigma_{34} + h_4\sigma_{44} + h_5\sigma_{54})$$
$$+ d_5(h_1\sigma_{15} + h_2\sigma_{25} + h_3\sigma_{35} + h_4\sigma_{45} + h_5\sigma_{55}) \quad ,$$

where σ_{ij} is the corresponding element of Σ.

3.3 Methods of Inference for the Coefficients of Individual Agreement using the MSD

We now introduce three methods for estimating the standard errors of the estimated $\psi's$.

3.3.1 Method A (Existing method 1) – assuming independence between estimated mean square errors

The most commonly used disagreement function is the mean squared deviation (MSD), defined in (2.1). Using $G = MSD$, we will estimate the standard errors of the estimates of both ψ^N and ψ^R using multiple approaches.

In the first approach, developed by Haber (personal communication), we do not make any distributional assumptions about X_i and Y_i, except that the first two moments exist. We define K_1 as the number of replications for *X*, and K_2 as the number of replications for *Y*. Define

$$T_i = (\overline{X}_i - \overline{Y}_i)^2 \Big/ 2 \quad , \tag{3.8}$$

$$U_{i1} = \sum_k (X_{ik} - \overline{X}_i)^2 \Big/ (K_1 - 1) \quad , \tag{3.9}$$

and

$$U_{i2} = \sum_k (Y_{ik} - \overline{Y}_i)^2 \Big/ (K_2 - 1) \quad , \tag{3.10}$$

where U_{i1} and U_{i2} are the estimated mean square errors for X_i and Y_i, respectively. We assume that U_{i1} and U_{i2} are independent, and that each of the subject-specific means over replications, $\overline{X}_i$ and $\overline{Y}_i$, are independent of the estimated mean square errors. If $\overline{T} = (\sum_i T_i)/N$ and $\overline{U}_j = (\sum_i U_{ij})/N$ where j = 1,2, then $\overline{T}$, $\overline{U}_1$, and $\overline{U}_2$ are all independent.

Given the definition of MSD in (2.2), we can use estimates defined by Haber *et al* (2005) to define the two coefficients of agreement as

$$\hat{\psi}^N = \frac{\overline{U}_1 + \overline{U}_2}{2\overline{T} + (1 - 1/K_1)\overline{U}_1 + (1 - 1/K_2)\overline{U}_2} \tag{3.11}$$

and

$$\hat{\psi}^R = \frac{2 \cdot \overline{U}_1}{2\overline{T} + (1 - 1/K_1)\overline{U}_1 + (1 - 1/K_2)\overline{U}_2} . \tag{3.12}$$

To estimate the variances of $\hat{\psi}^N$ and $\hat{\psi}^R$, we need to approximate the variance of a ratio as

$$Var(\frac{A}{B}) \approx (A/B)^2 [Var(A)/A^2 + Var(B)/B^2 - 2Cov(A,B)/AB] . \tag{3.13}$$

For $\hat{\psi}^N$, $A = \overline{U}_1 + \overline{U}_2$, $B = 2\overline{T} + (1 - 1/K_1)\overline{U}_1 + (1 - 1/K_2)\overline{U}_2$,

$$Var(A) = S^2(U_1) + S^2(U_2) \Big/ N$$

$$Var(B) = \{4S^2(T) + [(K_1 - 1)/K_1]^2 S^2(U_1) + [(K_2 - 1)/K_2]^2 S^2(U_2)\} \Big/ N, \text{ and}$$

$$Cov(A, B) = \{[(K_1 - 1)/K_1]^2 S^2(U_1) + [(K_2 - 1)/K_2]^2 S^2(U_2)\} \Big/ N.$$

For $\hat{\psi}^R$, $A = 2 \cdot \overline{U}_1$, $B = 2\overline{T} + (1 - 1/K_1)\overline{U}_1 + (1 - 1/K_2)\overline{U}_2$,

$$Var(A) = 4[S^2(U_1)] \Big/ N$$

$$Var(B) = \{4S^2(T) + [(K_1 - 1)/K_1]^2 S^2(U_1) + [(K_2 - 1)/K_2]^2 S^2(U_2)\} \Big/ N,$$

and $Cov(A, B) = 2(K_1 - 1)S^2(U_1) \Big/ K_1 \cdot N$.

We can now substitute the appropriate expressions into (3.13) to estimate the two variances. The sampling variance of a statistic is defined here as $S^2(\cdot)$.

3.3.2 Method B (Existing method 2) – general method using subject-specific estimates

The next approach to estimate the standard errors of ψ^N and ψ^R, generalizes the previous approach so it is not restricted to the case where the MSD is the disagreement function. This approach is similar to the "U-statistics" method described by King and Chinchilli (2001a) and was proposed by Barnhart et al. (2007). It can use any disagreement function. This method of estimation does not require the same number of replications per subject, as with Method A.

Here, we define $G^{(1)} = G(X, X')$, $G^{(2)} = G(Y, Y')$, and $G^{(3)} = G(X, Y)$. If the subject-specific estimates of the disagreement function are defined as $\hat{G}_i$, then $\overline{G} = (\sum_{i=1}^{N} \hat{G}_i) / N$.

Now, the estimates of ψ^N and ψ^R are defined as

$$\hat{\psi}^N = \frac{[\overline{G}(X, X') + \overline{G}(Y, Y')]/2}{\overline{G}(X, Y)} \tag{3.14}$$

and

$$\hat{\psi}^R = \frac{\overline{G}(X, X')}{\overline{G}(X, Y)} \quad . \tag{3.15}$$

We once again use the expression in (3.13) to approximate the variance of a ratio. The appropriate substitutions for $\hat{\psi}^N$ and $\hat{\psi}^R$ are as follows.

For $\hat{\psi}^N$, $A = (\overline{G}^{(1)} + \overline{G}^{(2)})/2$, $B = \overline{G}^{(3)}$

$Var(A) = [S^2(G^{(1)}) + S^2(G^{(2)}) + 2Cov(G^{(1)}, G^{(2)})]/4N$, $Var(B) = S^2(G^{(3)})/N$,

and $Cov(A, B) = [Cov(G^{(1)}, G^{(3)}) + Cov(G^{(2)}, G^{(3)})]/2N$.

For $\hat{\psi}^R$, $A = \overline{G}^{(1)}$, $B = \overline{G}^{(3)}$, $Var(A) = S^2(G^{(1)})/N$,

$Var(B) = S^2(G^{(3)})/N$, and $Cov(A, B) = Cov(G^{(1)}, G^{(3)})/N$.

The expressions for the above sampling variances and covariances are given in Barnhart et al. (2007).

3.3.3 Method C (New method) – estimation and inference using variance components

For our last approach to estimate the standard errors of $\hat{\psi}^N$ and $\hat{\psi}^R$, we can use a 2-way random effects model to estimate these coefficients with variance components, assuming the data are normally distributed.

We define the model as

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \tag{3.16}$$

where i = subject , j = observer , and k = replication. The appropriate variance components are define as $Var(\alpha_i) = \sigma_\alpha^2$, $Var(\beta_i) = \sigma_\beta^2$, $Var(\gamma_{ij}) = \sigma_\gamma^2$, and $Var(\varepsilon_{ijk}) = \sigma_{\varepsilon_j}^2$. The overall error variance is defined as $\sigma_\varepsilon^2 =$ mean of $\sigma_{\varepsilon_1}^2,...,\sigma_{\varepsilon_J}^2$.

Since we are comparing two methods, we can define $X = Y_1$ and $Y = Y_2$.

Now, $MSD\ (X, X') = E(Y_{i1k} - Y_{i1k'})^2 = E(\varepsilon_{i1k} - \varepsilon_{i1k'})^2 = 2\sigma_{\varepsilon_1}^2$,

$MSD(Y, Y') = E(Y_{i2k} - Y_{i2k'})^2 = E(\varepsilon_{i2k} - \varepsilon_{i2k'})^2 = 2\sigma_{\varepsilon_2}^2$,

and $MSD(X, Y) = E(Y_{i1k} - Y_{i2k'})^2 = E(\beta_1 - \beta_2)^2 + E(\gamma_{i1} - \gamma_{i2})^2 + E(\varepsilon_{i1k} - \varepsilon_{i2k'})^2$

$= 2\sigma_\beta^2 + 2\sigma_\gamma^2 + \sigma_{\varepsilon_1}^2 + \sigma_{\varepsilon_2}^2$.

Substituting in the appropriate variance components expressions into the definitions of ψ^N and ψ^R allows us to compute the same estimates from Method A and Method B using the random effects model as

$$\hat{\psi}^N = \frac{(2\hat{\sigma}_{\varepsilon_1}^2 + 2\hat{\sigma}_{\varepsilon_2}^2)/2}{2\hat{\sigma}_\beta^2 + 2\hat{\sigma}_\gamma^2 + \hat{\sigma}_{\varepsilon_1}^2 + \hat{\sigma}_{\varepsilon_2}^2} = \frac{\hat{\sigma}_\varepsilon^2}{\hat{\sigma}_\beta^2 + \hat{\sigma}_\gamma^2 + \hat{\sigma}_\varepsilon^2} \tag{3.17}$$

and

$$\hat{\psi}^R = \frac{2\hat{\sigma}_{\varepsilon_1}^2}{2\hat{\sigma}_\beta^2 + 2\hat{\sigma}_\gamma^2 + \hat{\sigma}_{\varepsilon_1}^2 + \hat{\sigma}_{\varepsilon_2}^2} = \frac{\hat{\sigma}_{\varepsilon_1}^2}{\hat{\sigma}_\beta^2 + \hat{\sigma}_\gamma^2 + \hat{\sigma}_\varepsilon^2}. \tag{3.18}$$

To approximate the variance of $\hat{\psi}^N$, we can use the delta method. The partial derivatives of the three variance components are defined as

$$\frac{\partial \psi^N}{\partial \sigma_\beta^2} = \frac{-\sigma_\varepsilon^2}{(\sigma_\beta^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2)^2} = \frac{-\psi^N}{\sigma_\beta^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2},$$

$$\frac{\partial \psi^N}{\partial \sigma_\gamma^2} = \frac{-\sigma_\varepsilon^2}{(\sigma_\beta^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2)^2} = \frac{-\psi^N}{\sigma_\beta^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2},$$

and $\frac{\partial \psi^N}{\partial \sigma_\varepsilon^2} = \frac{(\sigma_\beta^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2) - \sigma_\varepsilon^2}{(\sigma_\beta^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2)^2} = \frac{1-\psi^N}{\sigma_\beta^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2}.$

The expression for the variance of $\hat{\psi}^N$ is then

$$\begin{aligned} Var(\hat{\psi}^N) &= (\frac{-\psi^N}{\sigma_\beta^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2})^2 Var(\sigma_\beta^2) + (\frac{-\psi^N}{\sigma_\beta^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2})^2 Var(\sigma_\gamma^2) + (\frac{1-\psi^N}{\sigma_\beta^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2})^2 Var(\sigma_\varepsilon^2) \\ &+ 2(\frac{-\psi^N}{\sigma_\beta^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2})^2 \operatorname{cov}(\sigma_\beta^2, \sigma_\gamma^2) + 2(\frac{-\psi^N}{\sigma_\beta^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2})(\frac{1-\psi^N}{\sigma_\beta^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2}) \operatorname{cov}(\sigma_\beta^2, \sigma_\varepsilon^2) \\ &+ 2(\frac{-\psi^N}{\sigma_\beta^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2})(\frac{1-\psi^N}{\sigma_\beta^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2}) \operatorname{cov}(\sigma_\gamma^2, \sigma_\varepsilon^2) \\ &= \frac{\psi^{N^2}[Var(\sigma_\beta^2) + Var(\sigma_\gamma^2) + 2\operatorname{cov}(\sigma_\beta^2, \sigma_\gamma^2)] + (1-\psi^N)^2 Var(\sigma_\varepsilon^2) - 2\psi^N(1-\psi^N)[\operatorname{cov}(\sigma_\beta^2, \sigma_\varepsilon^2) + \operatorname{cov}(\sigma_\gamma^2, \sigma_\varepsilon^2)]}{(\sigma_\beta^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2)^2} \end{aligned} \tag{3.19}$$

The variances and covariances for the variance components of random effects can be determined using the observed inverse Fisher information matrix, as shown in Searle (1992), where

$$\operatorname{var}\begin{bmatrix} \tilde{\sigma}_\varepsilon \\ \tilde{\sigma}_\alpha \\ \tilde{\sigma}_\beta \\ \tilde{\sigma}_\gamma \end{bmatrix} \approx 2\begin{bmatrix} t_{\varepsilon\varepsilon} & t_{\alpha\alpha}/bn & t_{\beta\beta}/an & t_{\gamma\gamma}/n \\ & t_{\alpha\alpha} & abn^2/\theta_4^2 & t_{\alpha\alpha}/b \\ symmetric & & t_{\beta\beta} & t_{\beta\beta}/a \\ & & & t_{\gamma\gamma} \end{bmatrix}^{-1}, \tag{3.20}$$

$$t_{\alpha\alpha} = b^2 n^2 (\frac{a-1}{\theta_{11}^2} + \frac{1}{\theta_4^2}),\ t_{\beta\beta} = a^2 n^2 (\frac{b-1}{\theta_{12}^2} + \frac{1}{\theta_4^2}),$$

$$t_{\gamma\gamma} = n^2 (\frac{(a-1)(b-1)}{\theta_1^2} + \frac{(a-1)}{\theta_{11}^2} + \frac{(b-1)}{\theta_{12}^2} + \frac{1}{\theta_4^2}),\ t_{\varepsilon\varepsilon} = \frac{ab(n-1)}{\theta_0^2} + \frac{t_{\gamma\gamma}}{n^2}),$$

$$\theta_0 = \sigma_\varepsilon^2,\ \theta_{11} = \sigma_\varepsilon^2 + n\sigma_\gamma^2 + bn\sigma_\alpha^2,\ \theta_1 = \sigma_\varepsilon^2 + n\sigma_\gamma^2,\ \theta_{12} = \sigma_\varepsilon^2 + n\sigma_\gamma^2 + an\sigma_\beta^2,$$

and $\theta_4 = \theta_{11} + \theta_{12} - \theta_1$.

The variance components and variance-covariance matrix needed to calculate (3.19) and (3.20) can easily be computed using SAS's PROC MIXED. The estimation for ψ^R is more difficult, but can be achieved by fitting two models, one with all observations in the data set and one with only the observations using the reference method.

3.4 Simulation study – performance and comparison of estimates

To examine the behavior of the estimated standard errors developed in the previous section, we conducted a simulation study. Data was simulated from 3 different scenarios using a simple latent class model. Initial values (or true values), denoted *T*, were drawn from either a standard normal or exponential distribution ($T \sim N(0,1)$ or $T \sim Exp(1)$). The *T* values were then used to generate values for *X* and *Y* using the conditional distributions $X \mid t \sim N(a+bt, (e+ft)^2)$, and $Y \mid t \sim N(c+dt, (g+ht)^2)$. Two replications were generated each for *X* and *Y*. Three different sample sizes were used with each model.

The results of 1,000 simulations for ψ^N are presented in Table 3.1, and the results of 1,000 simulations for ψ^R are presented in Table 3.2. For each table, the appropriate MSD values are estimated using (2.7), and then used to estimate the coefficients for each simulated data set. It is easy to show (Haber and Barnhart, 2008) that the true values for the MSD's can be calculated from $MSD(X, X') = 2e^2 + 4ef\mu_T + 2f^2(\mu_T^2 + \sigma_T^2)$,

$MSD(Y, Y') = 2g^2 + 4gh\mu_T + 2h^2(\mu_T^2 + \sigma_T^2)$, and

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

Standard errors were estimated using Method A, Method B, and Method C, and simple bootstrap resampling. The asymptotic variance described in (3.5) and (3.6) was also included in order to test the validity of this estimate. Coverage probabilities are computed using each estimated standard error. Two-sided 95% confidence intervals were used to compute the coverage probabilities.

Overall, the bias and MSE were low for each scenario. A large-sample estimate for the standard error was used to approximate the true value. For both ψ^N and ψ^R, Method A tends to overestimate the standard error and Method B tends to underestimate the standard error. Method B was very close to that estimated by the bootstrap. Method C for inference on ψ^N was more variable than the other two methods, performing best with non-normal data. Coverage probabilities were consistently closer to 95% when using Method A and with larger values of n. The asymptotic standard errors for $\ln(\hat{\psi}^N)$ and $\ln(\hat{\psi}^R)$ perform similarly to Method B and the bootstrap when drawing initial values from a normal distribution, but the coverage probabilities are lower when values are drawn for an exponential distribution. This was expected given the performance of the

variance estimates for log(MSD) shown in Table 2.2. The performance of the asymptotic standard errors improves consistently with the increase of sample size.

3.5 Sample size estimation

Estimating the necessary sample size to assess agreement using ψ^N or ψ^R could be of interest to investigators planning a study to compare two observers. We derived an expression for this sample size using a 2-sided confidence interval, and Method B for estimating the standard error. Method B was chosen since it can use any disagreement function and performed well in the simulation study. This method for estimating sample size assumes that the number of replications per subject is fixed.

For ψ^N the expression for the width of a 2-sided confidence interval is $width = [\hat{\psi}^N + (z_{1-\alpha/2}\sqrt{Var(\hat{\psi}^N)})] - [\hat{\psi}^N - (z_{1-\alpha/2}\sqrt{Var(\hat{\psi}^N)})]$. Solving for N gives the following expression:

$$width = 2z_{1-\alpha/2}\sqrt{Var(\hat{\psi}^N)}$$

$$\frac{(width)^2}{4z_{1-\alpha/2}^2} = \frac{1}{N}\left[\frac{\overline{G}^{(1)} + \overline{G}^{(2)}/2}{\overline{G}^{(3)}}\right]^2\left[\frac{S^2(G^{(1)}) + S^2(G^{(2)}) + 2Cov(G^{(1)}, G^{(2)})}{4(\overline{G}^{(1)} + \overline{G}^{(2)}/2)^2}\right.$$

$$\left. + \frac{S^2(G^{(3)})}{\overline{G}^{(3)2}} - \frac{Cov(G^{(1)}, G^{(3)}) + Cov(G^{(2)}, G^{(3)})}{(\overline{G}^{(1)} + \overline{G}^{(2)}/2)\cdot\overline{G}^{(3)}}\right]$$

$$N = \frac{4z_{1-\alpha/2}^2}{(width)^2}\left[\frac{\overline{G}^{(1)} + \overline{G}^{(2)}/2}{\overline{G}^{(3)}}\right]^2\left[\frac{S^2(G^{(1)}) + S^2(G^{(2)}) + 2Cov(G^{(1)}, G^{(2)})}{4(\overline{G}^{(1)} + \overline{G}^{(2)}/2)^2}\right.$$

$$+\frac{S^2(G^{(3)})}{\overline{G}^{(3)2}}-\frac{Cov(G^{(1)},G^{(3)})+Cov(G^{(2)},G^{(3)})}{(\overline{G}^{(1)}+\overline{G}^{(2)}/2)\cdot\overline{G}^{(3)}}] \tag{3.21}$$

The corresponding expression for ψ^R is

$$N=\frac{4z^2_{1-\alpha/2}}{(width)^2}[\frac{\overline{G}^{(1)}}{\overline{G}^{(3)}}]^2[\frac{S^2(G^{(1)})}{(\overline{G}^{(1)})^2}+\frac{S^2(G^{(3)})}{\overline{G}^{(3)2}}-\frac{2Cov(G^{(1)},G^{(3)})}{\overline{G}^{(1)}\cdot\overline{G}^{(3)}}] \tag{3.22}$$

3.6 Examples

3.6.1 Bland and Altman Blood Pressure (SBP) Data

To demonstrate the methods developed in this chapter, we use the systolic blood pressure data set from Bland and Altman described earlier. Table 3.3 gives estimates of ψ^N and ψ^R, accompanied by standard error estimates using both Method A, Method B, and Method C with 2-sided 95% confidence intervals. The human observers are used as the reference observer in calculating ψ^R.

The agreement between the two human observers is very high, at $\psi^N = 1.449$. The standard error estimates are both small, indicating agreement between the human observers was consistently high.

Both ψ^N and ψ^R are very low when comparing the semi-automatic blood pressure monitor to the human observers, with all coefficients below 0.2. The ψ^R estimates are lower then the ψ^N estimates for both observers, showing that agreement decreases when a reference observer is assigned. This happened because the within-method error of the

reference method is smaller than that of the other method. All confidence intervals do not contain 1, indicating significantly poor agreement between the human observers and the monitor. Standard error estimates using Method A and Method B are very close to one another, and lead to the same conclusions, but those using Method C for ψ^N are somewhat higher leading to decreased precision.

3.6.2 Carotid Stenosis Data

Table 3.4 demonstrates the same methods applied to the previously described carotid stenosis data. Here, the angiogram is assumed to be the reference method when computing ψ^R, since it was used as the gold standard for comparing the two MRA methods. Results are presented separately for the left and right carotid arteries.

Once again, the ψ^R estimates are consistently lower than those for ψ^N, showing lower agreement when one method is assumed to be a reference. Agreement is highest between the two MRA methods. Overall, the MRA-2D method agrees better with the established gold standard. However, since the confidence intervals for MRA-2D and MRA-3D comparisons against the angiogram overlap substantially, we may not have enough evidence to show that one definitely performs better than the other.

3.7 Robustness of estimates and standard errors

In the simulation study, estimates of ψ^N and ψ^R were shown to have very low bias. Outlying measurements may increase this bias, and also lead to inflated estimates

of the standard errors. To examine the effects of outliers on these coefficients, we consider two scenarios.

In scenario 1, the outlying measurement is abnormally high or low for a single replicated measurement on a subject. Here, the true value of the measurement for this subject is not extreme, only the single measurement, meaning one observer has given a highly dissimilar value for this subject compared to other observers or a gold standard. In scenario 2, the true value of a subject's measurement is abnormally high or low. For this scenario, observer agreement can still be high, since extreme values should be observed for all measurements on this subject.

We adapted the previous simulation study to examine the effects of outliers, by repeating the same simulations for ψ^N and ψ^R and replacing one measurement with an abnormally high value. For scenario 1, an extreme value was only included in the first replicate of *Y*, by adding 100 to the simulated measurement. For scenario 2, an extreme value was included for all replicates of *X* and *Y* on a single subject, by adding a random value from a normal distribution with μ=100.

The results of these simulations are presented in Table 3.5 for ψ^N, and in Table 3.6 for ψ^R. For scenario 1, the estimates are more biased for ψ^N than in previous simulations, and are extremely biased for ψ^R. The poorer estimates for ψ^R are due to *Y* values only being used to estimate the denominator of the coefficient. The coverage probabilities indicate that standard errors estimated using Methods A and B are heavily affected by the outlying measurement. Standard errors estimated using Method C or bootstrap resampling are more resistant to the effects of the outlying measurement. For

scenario 2, little effect is seen by including an outlier in all measurements on a subject, as bias and coverage probabilities are very similar to those presented in Tables 3.1 and 3.2.

These simulation results show that care must be taken when analyzing data with outlying measurements when only one of the measurements is abnormally high or low. Outliers of this type lead to large biases in estimation, since differences between replicates and observers will be very high. Using bootstrap resampling to estimate the standard error is preferred if these outliers are included in the analysis, or Method C can be used for ψ^N since estimation of the variance components is more resistant to outliers than estimation methods using subject-specific estimates. The estimates perform very well when outlying subjects are present, since the differences between replicates and observers will remain low.

Since the MSD is based on the mean of squared differences between measurements, single outlying measurements will inflate the estimated disagreement function. When such measurements are found to exist, one can consider replacing the MSD with a disagreement function based on ranks such as the TDI or the median of absolute differences. In this case the standard errors can still be estimated with Method B, since this method is not restricted to use with the MSD. One may also extend the MSD by replacing the squared distance function with alternative distance functions described in King and Chinchilli (2001b).

3.8 Extension to $J \geq 2$ observers

If we assume there are *J* fixed observers, the expressions for ψ^N and ψ^R in (3.1) and (3.2) can easily be modified. Now, we denote all observations by *Y*, where Y_{ijk} is the *k*-th replicate observation on subject *i* by observer *j* where $j = 1,...,J$. The expanded expressions are

$$\psi^N = \frac{[\sum_{j=1}^{J} G(Y_j, Y_j')]/J}{[2\sum_{j=1}^{J-1}\sum_{j'=j+1}^{J} G(Y_j, Y_{j'})]/(J(J-1))]} \quad , \qquad (3.23)$$

and

$$\psi^R = \frac{G(Y_J, Y_J')}{\sum_{j=1}^{J-1} G(Y_j, Y_J)/(J-1)} \quad . \qquad (3.24)$$

For ψ^R, observer *J* is considered the reference observer.

For inference, we can modify Method A for *J* fixed observers. We redefine T_i and U_{ij} as defined in (3.8), (3.9), and (3.10) as $T_i = \sum_{j=1}^{J-1}\sum_{j'=j+1}^{J}(Y_{ij\bullet} - Y_{ij'\bullet})^2 \Big/ 2$, $T_{iR} = \sum_{j=1}^{J-1}(Y_{ij\bullet} - Y_{iJ\bullet})^2 \Big/ 2$, and $U_{ij} = \sum_{k=1}^{K_j}(Y_{ijk} - Y_{ij\bullet})^2 \Big/ (K_j - 1)$, where $K_1,...,K_J$ are the number of replicates taken by each of the *J* observers and a ($\bullet$) denotes the arithmetic mean with respect to the corresponding index. The estimates in (3.11) and (3.12) are extended to $\hat{\psi}^N = \frac{U_{\bullet\bullet}}{T_\bullet + \sum_{j=1}^{J}(1 - 1/K_j)U_{\bullet j}}$ and $\hat{\psi}^R = \frac{2U_{\bullet J}}{T_{\bullet R} + \sum_{j=1}^{J}(1 - 1/K_j)U_{\bullet j}}$.

Using the approximation in (3.13), for ψ^N we now have:

$$A = U_{\bullet\bullet}, \quad B = T_\bullet + \sum_{j=1}^{J}(1 - 1/K_j)U_{\bullet j} \ ,$$

$$Var(A) = \sum_{j=1}^{J} S^2(U_j) \Big/ N \quad , \quad Var(B) = S^2(T) + \sum_{j=1}^{J} [(K_j - 1)/K_j]^2 S^2(U_j) \Big/ N \quad , \text{ and}$$

$$Cov(A,B) = \sum_{j=1}^{J} [(K_j - 1)/K_j]^2 S^2(U_j) \Big/ N \quad .$$

For ψ^R we now have:

$$A = 2U_{\bullet J} \, , \quad B = T_{\bullet} + \sum_{j=1}^{J} (1 - 1/K_j) U_{\bullet j} \quad ,$$

$$Var(A) = 4S^2(U_J)/N, \quad Var(B) = S^2(T) + \sum_{j=1}^{J} [(K_j - 1)/K_j]^2 S^2(U_j) \Big/ N \quad , \text{ and}$$

$$Cov(A,B) = 2(K_J - 1)S^2(U_J)/K_J \cdot N \, .$$

Method B is more difficult to modify for more than 2 observers, since this would entail defining a *G* variable for each of *J* functions for comparing replicated observations, and for each of $2/(J(J-1))$ functions for comparing observations between observers.

3.9 Discussion

We presented multiple approaches for inference on ψ^N and ψ^R, two inter-observer agreement coefficients evaluated using a disagreement function. These coefficients are quite flexible since one can specify the disagreement function depending on how they wish to interpret the data.

Three methods were presented for deriving the standard error of estimates of ψ^N and ψ^R. Overall, Method B seems preferable, since one does not have to specify a

specific disagreement function to use the estimate. Method A performed well in simulations, but would need to be developed for other disagreement functions for it to be a versatile method of inference. It also needs several independence assumptions in order for the method to be valid, including that the estimated mean square errors are independent of each other and of the subject-specific means over replications.

Using variance components from a random effects model is a convenient method for estimating the standard error of $\hat{\psi}^N$, due to software being readily available for fitting these models and computing an accompanying variance-covariance matrix. However, more assumptions need to be made when using such a model. The model parameters are assumed to be mutually independent, and the random method parameters are assumed to be normally distributed. This may make this approach too restrictive compared to the more versatile one developed here.

Table 3.1: Simulation results for ψ^N estimates based on 1000 samples. Coverage probabilities computed for 2-sided 95% confidence intervals using the indicated s.e.

		True				$\hat{\psi}^N$		
Distribution of T	**n**	**MSD(X,X')**	**MSD(Y,Y')**	**MSD(X,Y)**	ψ^N	**Mean**	**Bias**	**MSE**
Normal:								
a=0 b=1.1 c=0.5 d=1.5	50	4.420	9.000	7.120	0.942	0.962	0.019	0.033
e=1 f=1.1 g=1.5 h=1.5	100					0.949	0.007	0.020
	200					0.951	0.008	0.009
a=0 b=1.1 c=1.5 d=2.5	50	4.420	25.00	18.92	0.777	0.793	0.016	0.033
e=1 f=1.1 g=2.5 h=2.5	100					0.779	0.002	0.016
	200					0.778	0.001	0.009
Exponential:								
a=0 b=1.1 c=0.5 d=1.5	50	11.24	22.50	17.84	0.946	0.946	0.000	0.033
e=1 f=1.1 g=1.5 h=1.5	100					0.944	- 0.001	0.018
	200					0.951	0.005	0.010

		True*	$s.e._A(\hat{\psi}^N)$		$s.e._B(\hat{\psi}^N)$		$s.e._C(\hat{\psi}^N)$		$s.e.(\ln(\hat{\psi}^N_{Asymptotic}))$	$s.e.(\hat{\psi}^N_{Bootstrap})$	
Distribution of T	**n**	$s.e.(\psi^N)$	**Mean**	**Coverage Prob.**	**Mean**	**Coverage Prob.**	**Mean**	**Coverage Prob.**	**Coverage Prob.**	**Mean**	**Coverage Prob.**
Normal:											
a=0 b=1.1 c=0.5 d=1.5	50	0.186	0.257	0.980	0.161	0.887	0.122	0.801	0.889	0.161	0.888
e=1 f=1.1 g=1.5 h=1.5	100	0.134	0.189	0.985	0.124	0.891	0.096	0.840	0.892	0.123	0.895
	200	0.096	0.137	0.990	0.091	0.919	0.070	0.847	0.912	0.091	0.918
a=0 b=1.1 c=1.5 d=2.5	50	0.181	0.238	0.976	0.158	0.883	0.205	0.974	0.890	0.159	0.892
e=1 f=1.1 g=2.5 h=2.5	100	0.131	0.175	0.973	0.121	0.920	0.190	0.996	0.919	0.120	0.921
	200	0.094	0.128	0.980	0.089	0.932	0.188	0.999	0.929	0.089	0.931
Exponential:											
a=0 b=1.1 c=0.5 d=1.5	50	0.179	0.239	0.978	0.149	0.857	0.203	0.974	0.802	0.149	0.864
e=1 f=1.1 g=1.5 h=1.5	100	0.132	0.181	0.988	0.119	0.900	0.132	0.951	0.824	0.118	0.897
	200	0.097	0.136	0.990	0.090	0.906	0.107	0.969	0.857	0.089	0.905

* Estimated from 100,000 simulations

Table 3.2: Simulation results for ψ^R estimates based on 1000 samples. Coverage probabilities computed for 2-sided 95% confidence intervals using the indicated s.e.

		True				$\hat{\psi}^R$		
Distribution of T	**n**	**MSD(X,X')**	**MSD(Y,Y')**	**MSD(X,Y)**	ψ^R	**Mean**	**Bias**	**MSE**
Normal:								
a=0 b=1.1 c=0.5 d=1.5	50	4.420	9.000	7.120	0.621	0.640	0.020	0.041
e=1 f=1.1 g=1.5 h=1.5	100					0.630	0.010	0.021
	200					0.628	0.007	0.011
a=0 b=1.1 c=1.5 d=2.5	50	4.420	25.00	18.92	0.234	0.245	0.012	0.008
e=1 f=1.1 g=2.5 h=2.5	100					0.236	0.003	0.004
	200					0.233	- 0.001	0.002
Exponential:								
a=0 b=1.1 c=0.5 d=1.5	50	11.24	22.50	17.84	0.630	0.630	0.004	0.038
e=1 f=1.1 g=1.5 h=1.5	100					0.633	0.003	0.023
	200					0.636	0.006	0.012

		True*	$s.e._A(\hat{\psi}^R)$		$s.e._B(\hat{\psi}^R)$		$s.e.(\ln(\hat{\psi}^R_{Asymptotic}))$	$s.e.(\hat{\psi}^R_{Bootstrap})$	
Distribution of T	**n**	$s.e.(\psi^R)$	**Mean**	**Coverage Prob.**	**Mean**	**Coverage Prob.**	**Coverage Prob.**	**Mean**	**Coverage Prob.**
Normal:									
a=0 b=1.1 c=0.5 d=1.5	50	0.204	0.212	0.921	0.179	0.887	0.889	0.182	0.894
e=1 f=1.1 g=1.5 h=1.5	100	0.146	0.157	0.944	0.134	0.907	0.910	0.135	0.906
	200	0.103	0.113	0.962	0.098	0.918	0.922	0.098	0.922
a=0 b=1.1 c=1.5 d=2.5	50	0.088	0.091	0.918	0.076	0.878	0.888	0.079	0.896
e=1 f=1.1 g=2.5 h=2.5	100	0.061	0.065	0.944	0.056	0.909	0.911	0.057	0.914
	200	0.043	0.047	0.944	0.040	0.912	0.915	0.040	0.918
Exponential:									
a=0 b=1.1 c=0.5 d=1.5	50	0.203	0.197	0.906	0.166	0.872	0.826	0.168	0.882
e=1 f=1.1 g=1.5 h=1.5	100	0.146	0.151	0.930	0.129	0.891	0.847	0.129	0.896
	200	0.105	0.115	0.954	0.099	0.919	0.880	0.099	0.921

* Estimated from 100,000 simulations

Table 3.3: Estimation of ψ^N and ψ^R for Bland and Altman SBP data.

Comparison	MSD(X,X')	MSD(Y,Y')	MSD(X,Y)
Observer1 vs. Machine	74.8	166	679
Observer2 vs. Machine	76.0	166	676
Observer1vs. Observer2	74.8	76.0	52.0

Comparison	$\hat{\psi}^N$	$s.e._A(\hat{\psi}^N)$	95% CI (2-sided)	$s.e._B(\hat{\psi}^N)$	95% CI (2-sided)	$s.e._C(\hat{\psi}^N)$	95% CI (2-sided)	$s.e.(\hat{\psi}^N_{Bootstrap})$	95% CI (2-sided)
Observer1 vs. Machine	0.178	0.053	(0.075, 0.281)	0.047	(0.086, 0.270)	0.092	(-0.003, 0.358)	0.050	(0.079, 0.276)
Observer2 vs. Machine	0.179	0.053	(0.074, 0.284)	0.048	(0.084, 0.274)	0.094	(-0.006, 0.364)	0.052	(0.077, 0.281)
Observer1vs. Observer2	1.449	0.099	(1.256, 1.642)	0.010	(1.429, 1.468)	- *	-	0.010	(1.430, 1.468)

Comparison	$\hat{\psi}^R$	$s.e._A(\hat{\psi}^R)$	95% CI (2-sided)	$s.e._B(\hat{\psi}^R)$	95% CI (2-sided)	$s.e.(\hat{\psi}^N_{Bootstrap})$	95% CI (2-sided)
Observer1 vs. Machine	0.110	0.031	(0.049, 0.172)	0.033	(0.046, 0.174)	0.037	(0.038, 0.182)
Observer2 vs. Machine	0.112	0.032	(0.050, 0.175)	0.034	(0.046, 0.179)	0.039	(0.037, 0.188)

* Random effects model cannot be fit (G matrix is not positive definite)

Table 3.4: Estimation of ψ^N and ψ^R for carotid stenosis data.

Comparison	MSD(X,X')	MSD(Y,Y')	MSD(X,Y)	$\hat{\psi}^R$	$s.e._A(\hat{\psi}^R)$	95% CI (2-sided)	$s.e._B(\hat{\psi}^R)$	95% CI (2-sided)	$s.e.(\hat{\psi}^R_{Bootstrap})$	95% CI (2-sided)
Left Side:										
Angiogram vs. MRA-2D	279	1153	1211	0.231	0.103	(0.028, 0.433)	0.100	(0.035, 0.427)	0.095	(0.044, 0.418)
Angiogram vs. MRA-3D	279	1040	1461	0.191	0.087	(0.021, 0.361)	0.084	(0.026, 0.357)	0.081	(0.032, 0.351)
MRA-2D vs. MRA-3D	1153	1040	1245	-						
Right Side:										
Angiogram vs. MRA-2D	176	1137	959	0.183	0.051	(0.084, 0.283)	0.051	(0.083, 0.284)	0.053	(0.079, 0.288)
Angiogram vs. MRA-3D	176	1100	1093	0.161	0.050	(0.063, 0.260)	0.051	(0.060, 0.262)	0.054	(0.056, 0.266)
MRA-2D vs. MRA-3D	1137	1100	1219	-						

Comparison	$\hat{\psi}^N$	$s.e._A(\hat{\psi}^N)$	95% CI (2-sided)	$s.e._B(\hat{\psi}^N)$	95% CI (2-sided)	$s.e._C(\hat{\psi}^N)$	95% CI (2-sided)	$s.e.(\hat{\psi}^N_{Bootstrap})$	95% CI (2-sided)
Left Side:									
Angiogram vs. MRA-2D	0.592	0.138	(0.322, 0.861)	0.124	(0.348, 0.835)	0.075	(0.446, 0.738)	0.124	(0.349, 0.835)
Angiogram vs. MRA-3D	0.452	0.114	(0.228, 0.676)	0.107	(0.242, 0.661)	0.077	(0.300, 0.603)	0.104	(0.248, 0.655)
MRA-2D vs. MRA-3D	0.881	0.133	(0.621, 1.141)	0.099	(0.688, 1.075)	0.078	(0.728, 1.035)	0.099	(0.688, 1.075)
Right Side:									
Angiogram vs. MRA-2D	0.684	0.121	(0.447, 0.922)	0.063	(0.562, 0.807)	0.160	(0.370, 0.998)	0.063	(0.560, 0.809)
Angiogram vs. MRA-3D	0.584	0.138	(0.314, 0.854)	0.109	(0.370, 0.798)	0.102	(0.384, 0.784)	0.109	(0.371, 0.797)
MRA-2D vs. MRA-3D	0.917	0.131	(0.661, 1.174)	0.096	(0.729, 1.106)	0.076	(0.769, 1.066)	0.096	(0.729, 1.106)

Table 3.5: Simulation results for ψ^N estimates based on 1000 samples. Coverage probabilities computed for 2-sided 95% confidence intervals using the indicated s.e. In outlier scenario 1, one additional outlying observation is added to the first replicate of Y. In outlier scenario 2, one additional outlying observation is added to all replicates of X and Y for the same subject.

Outlier Scenario 1 Distribution of T	n	True ψ^N	$\hat{\psi}^N$ Mean	Bias	$s.e._A(\hat{\psi}^N)$ Coverage Prob.	$s.e._B(\hat{\psi}^N)$ Coverage Prob.	$s.e._C(\hat{\psi}^N)$ Coverage Prob.	$s.e.(\hat{\psi}^N_{Bootstrap})$ Coverage Prob.
Normal:								
a=0 b=1.1 c=0.5 d=1.5	50	0.942	0.987	0.045	1.000	0.244	0.773	0.996
e=1 f=1.1 g=1.5 h=1.5	100		0.983	0.041	1.000	0.416	0.798	0.994
	200		0.983	0.040	1.000	0.546	0.802	0.985
a=0 b=1.1 c=1.5 d=2.5	50	0.777	0.941	0.163	1.000	0.185	0.833	0.865
e=1 f=1.1 g=2.5 h=2.5	100		0.917	0.140	1.000	0.341	0.847	0.806
	200		0.891	0.113	1.000	0.562	0.856	0.789

Outlier Scenario 2 Distribution of T	n	True ψ^N	$\hat{\psi}^N$ Mean	Bias	$s.e._A(\hat{\psi}^N)$ Coverage Prob.	$s.e._B(\hat{\psi}^N)$ Coverage Prob.	$s.e._C(\hat{\psi}^N)$ Coverage Prob.	$s.e.(\hat{\psi}^N_{Bootstrap})$ Coverage Prob.
Normal:								
a=0 b=1.1 c=0.5 d=1.5	50	0.942	0.952	0.009	0.979	0.880	0.803	0.889
e=1 f=1.1 g=1.5 h=1.5	100		0.941	-0.001	0.985	0.980	0.838	0.913
	200		0.943	0.001	0.988	0.919	0.851	0.920
a=0 b=1.1 c=1.5 d=2.5	50	0.777	0.790	0.012	0.967	0.868	0.979	0.880
e=1 f=1.1 g=2.5 h=2.5	100		0.783	0.006	0.977	0.901	0.998	0.911
	200		0.780	0.003	0.984	0.932	0.998	0.929

Table 3.6: Simulation results for ψ^R estimates based on 1000 samples. Coverage probabilities computed for 2-sided 95% confidence intervals using the indicated s.e. In outlier scenario 1, one additional outlying observation is added to the first replicate of Y. In outlier scenario 2, one additional outlying observation is added to all replicates of X and Y for the same subject.

Outlier Scenario 1 Distribution of T	n	True ψ^R	$\hat{\psi}^R$ Mean	Bias	$s.e._A(\hat{\psi}^R)$ Coverage Prob.	$s.e._B(\hat{\psi}^R)$ Coverage Prob.	$s.e.(\hat{\psi}^R_{Bootstrap})$ Coverage Prob.
Normal:							
a=0 b=1.1 c=0.5 d=1.5	50	0.621	0.041	-0.580	0.000	0.000	0.588
e=1 f=1.1 g=1.5 h=1.5	100		0.077	-0.544	0.001	0.001	0.551
	200		0.136	-0.485	0.000	0.001	0.584
a=0 b=1.1 c=1.5 d=2.5	50	0.234	0.037	-0.197	0.002	0.002	0.630
e=1 f=1.1 g=2.5 h=2.5	100		0.063	-0.171	0.008	0.040	0.607
	200		0.098	-0.136	0.066	0.267	0.666

Outlier Scenario 2 Distribution of T	n	True ψ^R	$\hat{\psi}^R$ Mean	Bias	$s.e._A(\hat{\psi}^R)$ Coverage Prob.	$s.e._B(\hat{\psi}^R)$ Coverage Prob.	$s.e.(\hat{\psi}^R_{Bootstrap})$ Coverage Prob.
Normal:							
a=0 b=1.1 c=0.5 d=1.5	50	0.621	0.639	0.018	0.911	0.880	0.890
e=1 f=1.1 g=1.5 h=1.5	100		0.626	0.005	0.943	0.913	0.912
	200		0.626	0.005	0.962	0.929	0.928
a=0 b=1.1 c=1.5 d=2.5	50	0.234	0.244	0.010	0.922	0.885	0.896
e=1 f=1.1 g=2.5 h=2.5	100		0.238	0.005	0.946	0.914	0.922
	200		0.236	0.003	0.946	0.922	0.923

Chapter 4

A General Approach for Evaluating Agreement Between Observers Selected at Random

4.1 Introduction and notation

Chapter 3 described a general approach for evaluating agreement in the form of the coefficients of individual agreement, ψ^N and ψ^R. The methods developed apply to the case when two or more observers make measurements on a sample of randomly selected subjects with replications, and one observer may or may not be considered a reference observer. We assumed that the observers were a fixed set. Now, we seek to extend the previously described methods to be used where a set of two or more observers make measurements on a sample of subjects, and these observers are selected at random from a pool of potential observers. This situation is common in practice, where a large group of medical staff are trained to administer and interpret a diagnostic test, but only a small subset will actually make the measurement on a given subject.

We assume that there are *J* observers selected at random from a larger population, and *I* subjects. We previously denoted measurements made by two observers as (*X*) and (*Y*). Since we consider the case where we have two or more observers, we denote all observations by *Y*, where Y_{ijk} is the *k*-th replicate observation on subject *i* by observer *j* where $j = 1,...,J$. If one observer in the set is considered a reference observer, observer *J* will be the reference observer and measurements by observers $1,\ldots,J-1$ will be compared to measurements by a fixed observer *J*.

4.2 Coefficients of Individual Agreement for random observers

We use the expressions of ψ^N and ψ^R defined in (3.23) and (3.24), where $J \geq 2$. Using pairwise disagreement functions we have

$$\psi^N = \frac{[\sum_{j=1}^{J} G(Y_j, Y_j')]/J}{[2\sum_{j=1}^{J-1}\sum_{j'=j+1}^{J} G(Y_j, Y_{j'})]/(J(J-1))] } \quad , \qquad (4.1)$$

and

$$\psi^R = \frac{G(Y_J, Y_J')}{\sum_{j=1}^{J-1} G(Y_j, Y_J)/(J-1)} \quad . \qquad (4.2)$$

The disagreement function, *G*, can once again be defined by one of many chosen measures of pairwise agreement, although we will primarily use the MSD for deriving estimators.

Using the framework described by Haber et al. (2005), we now assume the model $Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}$, where $E(\varepsilon_{ijk} \mid ij) = 0$ and $Var(\varepsilon_{ijk} \mid ij) = \sigma_{ij}^2$. The model represents the combination of the true value of the measurement by observer *j* on subject *i* (μ_{ij}), and the intraobserver variability (σ_{ij}^2). If we let $\tau_i^2 = E_j((\mu_{ij} - \mu_{i*})^2)$, this can be described as the interobserver variability for subject *i*, or $\tau_i^2 = Var_j(\mu_{ij} \mid i)$. Then $\tau_*^2 = E(\tau_i^2)$, which is the expected interobserver variability. A $(*)$ denotes the expectation associated with the corresponding index, and a $(\bullet)$ denotes the arithmetic mean with respect to the corresponding index.

Assuming the MSD as our disagreement function, the subject-specific disagreement function between observers can be defined as

$$G_i(Y_j, Y_{j'}) = E_j[(Y_{ij\bullet} - Y_{ij'\bullet})^2 \mid i] = [2\sum_{j=1}^{J-1}\sum_{j'=j+1}^{J}(\mu_{ij} - \mu_{ij'})^2/(J(J-1))] + 2E_j[Var(Y_{ijk} \mid i)]$$

$$= 2\sum_{j=1}^{J}(\mu_{ij} - \mu_{i\bullet})^2/(J-1) + 2\sigma_{i*}^2 = 2\tau_i^2 + 2\sigma_{i*}^2 \quad .$$

The overall disagreement function between observers can be determined by taking the expectation over all subjects:

$$G(Y_j, Y_{j'}) = 2\tau_*^2 + 2\sigma_{**}^2 \quad .$$

Following a similar path to get the overall disagreement function between replicated observations for observers, the subject-specific function is

$$G_i(Y_j, Y_j') = E_j[(Y_{ijk} - Y_{ijk'})^2 \mid i] = 2E_j[Var(Y_{ijk} \mid i)] = 2\sigma_{i*}^2 \quad ,$$

and the overall function is $G(Y_j, Y_j') = 2\sigma_{**}^2$. Now ψ^N can be defined for random observers in terms of the model:

$$\psi^N = \frac{\sigma_{**}^2}{\tau_*^2 + \sigma_{**}^2} \quad . \qquad (4.3)$$

For comparison, when the observers are fixed, Haber et al. (2005) shows that one just needs to use the arithmetic mean of the intraobserver variability over fixed observers instead of the expectation:

$$\psi^N = \frac{\sigma_{*\bullet}^2}{\tau_*^2 + \sigma_{*\bullet}^2} \quad . \qquad (4.4)$$

To develop a corresponding expression for ψ^R, we will need to define the expected interobserver variability and intraobserver variability when observer J is considered a fixed reference observer and $J - 1$ observers are selected at random. We will call these terms, τ_{*R}^2 and σ_{**R}^2. The interobserver variability can be defined as

$$\tau_{iR}^2 = E_j((\mu_{ij} - \mu_{iJ})^2) \quad ,$$

where $\tau_{*R}^2 = E(\tau_{iR}^2)$. We assume that the variance within observers is constant for each of the $J-1$ randomly selected observers, $\sigma_{*j}^2 \equiv \sigma^2$. The intraobserver variability can be defined by taking the average of the variance for the randomly selected observers and the variance for the fixed reference observer:

$$\sigma_{i*R}^2 = (\sigma^2 + \sigma_{iJ}^2)/2 \quad , \text{ where } \sigma_{**R}^2 = E_i(\sigma_{i*R}^2) \quad .$$

Using the MSD, the subject-specific disagreement function between observers will be defined as

$$G_i(Y_j, Y_J) = E_j[(Y_{ij\bullet} - Y_{iJ\bullet})^2 \mid i] = E((\mu_{ij} - \mu_{iJ})^2) + 2E_j[Var(Y_{ijk} \mid i)] = 2\tau_{iR}^2 + 2\sigma_{i*R}^2 .$$

Taking the expectation will give the overall disagreement between observers,

$$G(Y_j, Y_J) = 2\tau_{*R}^2 + 2\sigma_{**R}^2 \quad .$$

For the disagreement function between replicated observations for the reference observer we use

$$G_i(Y_J, Y_J') = E_j[(Y_{iJk} - Y_{iJk'})^2 \mid i] = 2E_j[Var(Y_{iJk} \mid i)] = 2\sigma_{iJ}^2$$

and $G(Y_J, Y_J') = 2\sigma_{*J}^2$. In terms of the model, ψ^R can now be defined as

$$\psi^R = \frac{\sigma_{*J}^2}{\tau_{*R}^2 + \sigma_{**R}^2} \tag{4.5}$$

for random observers, and as

$$\psi^R = \frac{\sigma_{*J}^2}{\tau_{*R}^2 + \sigma_{*\bullet R}^2} \tag{4.6}$$

for fixed observers.

4.3 Estimation and inference

4.3.1 Estimation and inference for ψ^N

To estimate ψ^N when $J \geq 2$ observers are selected at random, we derive an estimator using the MSD as the disagreement function. Following the structure of Haber et al. (2005) for fixed observers, we construct the following two variables:

$$V_i = \sum_{j=1}^{J-1} \sum_{j'=j+1}^{J} (Y_{ij\bullet} - Y_{ij'\bullet})^2 \Big/ 2 \tag{4.7}$$

$$U_{ij} = \sum_{k=1}^{K} (Y_{ijk} - Y_{ij\bullet})^2 \Big/ (K-1) \quad . \tag{4.8}$$

Here, V_i is the observed interobserver variability for subject *i*. For the fixed observers case, $E_Y(V_i) = \tau_i^2 + \sigma_{i\bullet}^2 / K$, $E(V_\bullet) = \tau_*^2 + \sigma_{*\bullet}^2 / K$, $E_Y(U_{ij}) = \sigma_{ij}^2$, and $E(U_{\bullet\bullet}) = \sigma_{*\bullet}^2$. Using $\hat{\tau}_*^2 = V_\bullet - U_{\bullet\bullet} / K$, we can estimate ψ^N in this case from the constructed variables as

$$\hat{\psi}^N = \frac{U_{\bullet\bullet}}{V_\bullet + (1 - 1/K) U_{\bullet\bullet}} \quad . \tag{4.9}$$

When $J = 2$, the estimator reduces to that derived in (3.7).

Haber et al. (2005) goes on to show that ψ^N can be estimated with (4.9) in the case of random observers as well as fixed observers. This is true because

$$E_Y(V_i) = \sum_j (\mu_{ij} - \mu_{i\bullet})^2 \Big/ (J-1) + \sigma_{i\bullet}^2 / K$$

for a fixed subject and a fixed sample of observers. If we just fix the subject,

$$E_j[E_Y(V_i)] = Var_j(\mu_{ij}) + \sigma_{i*}^2 / K = \tau_i^2 + \sigma_{i*}^2 / K \, .$$

If we take a third expectation, this time over subjects, we have $E(V_\bullet) = \tau_*^2 + \sigma_{**}^2 / K$. Therefore, we can use (4.9) in both cases.

We assume that the $U_{\bullet j}$'s are independent from one another, and that V_i is independent of each of the $U_{\bullet j}$'s. To estimate the variance, we once again use the variance approximation for a ratio in (3.13), by plugging in

$A = U_{\bullet\bullet}$, $B = V_\bullet + (1 - 1/K)U_{\bullet\bullet}$,

$Var(A) = S^2(U)/N$, $Var(B) = S^2(V) + [(K-1)/K]^2 S^2(U)/N$, and

$Cov(A,B) = [(K-1)/K]^2 S^2(U)/N$.

The method can also be extended to the case where each observer takes a different amount of replicates. Here, $K_1,...,K_J$, and the U_{ij}'s can be written as

$$U_{ij} = \sum_{k=1}^{K_J} (Y_{ijk} - Y_{ij\bullet})^2 \Big/ (K_J - 1) \quad . \tag{4.10}$$

The same ψ^N estimate, (4.9), can then be used. The K_j's can be included in the expressions for $Var(B)$ and $Cov(A,B)$ by taking the average number of replications per observer.

4.3.2 Estimation and inference for ψ^R

To estimate ψ^R with random observers is slightly more complicated than ψ^N, but can still follow the same framework used in the previous section. Now U_{ij} can be

estimated in the same way, but V_i must be modified for the situation where the fixed observer *J* is considered a reference:

$$U_{ij} = \sum_{k=1}^{K} (Y_{ijk} - Y_{ij\bullet})^2 \Big/ (K-1) \tag{4.11}$$

$$V_{iR} = \sum_{j=1}^{J-1} (Y_{ij\bullet} - Y_{iJ\bullet})^2 \Big/ 2 \quad . \tag{4.12}$$

V_{iR} is the observed interobserver variability for subject *i* between the reference observer *J* and observers 1,…,*J* – 1. For the fixed observers case, $E_Y(V_{iR}) = \tau_{iR}^2 + \sigma_{i\bullet R}^2 / K$, $E(V_{\bullet R}) = \tau_{*R}^2 + \sigma_{*\bullet R}^2 / K$, and $E_Y(U_{ij}) = \sigma_{ij}^2$. We need to use

$$\hat{\tau}_{*R}^2 = V_{\bullet R}/2 - \frac{\sum_{j=1}^{J-1} U_{\bullet j}}{\left[\frac{J-1}{} + U_{\bullet J}\right]} \Big/ 2K$$

to estimate ψ^R in this case from the constructed variables as

$$\hat{\psi}^R = \frac{U_{\bullet J}}{V_{\bullet R}/2 - \frac{\sum_{j=1}^{J-1} U_{\bullet j}}{\left[\frac{J-1}{} + U_{\bullet J}\right]} \Big/ 2K + \frac{\sum_{j=1}^{J-1} U_{\bullet j}}{\left[\frac{J-1}{} + U_{\bullet J}\right]} \Big/ 2} = \frac{2U_{\bullet J}}{V_{\bullet R} + (1 - 1/K)\left[\frac{\sum_{j=1}^{J-1} U_{\bullet j}}{J-1} + U_{\bullet J}\right]} . \tag{4.13}$$

When *J* = 2, the estimator reduces to that derived in (3.8).

We go on to show that ψ^R can also be estimated with (4.13) in the case of random observers as well as fixed observers. This is true because

$$E_Y(V_{iR}) = \sum_j (\mu_{ij} - \mu_{iJ})^2 \Big/ (J-1) + \sigma_{i\bullet R}^2 / K$$

for a fixed subject and a fixed sample of observers. If we just fix the subject,

$$E_j[E_Y(V_i)] = \tau_{iR}^2 + \sigma_{i*R}^2 / K \, .$$

If we take a third expectation over subjects, we have $E(V_{\bullet R}) = \tau^2_{*R} + \sigma^2_{**R} / K$. Therefore, we can use (4.13) in both cases.

We again assume that the $U_{\bullet j}'s$ are independent from one another, and that V_{iR} is independent of each of the $U_{\bullet j}'s$. To estimate the variance using the approximation in (3.13), we plug in

$$A = 2U_{\bullet J}, \quad B = V_{\bullet R} + (1 - 1/K)[\frac{\sum_{j=1}^{J-1} U_{\bullet j}}{J-1} + U_{\bullet J}],$$

$$Var(A) = 4S^2(U_J)/N,$$

$$Var(B) = \{S^2(V_R) + \{[(K-1)/K]^2 S^2(U_{random})\} + [(K-1)/K]^2 S^2(U_J)\}/N, \text{ and}$$

$$Cov(A,B) = 2(K-1)S^2(U_J)/K \cdot N \quad ,$$

where U_{random} refers to all observations made by random observers not considered a reference.

The method can again be extended to the case where each observer takes a different amount of replicates. Here, $K_1,...,K_J$, and the $U_{ij}'s$ can be written as in (4.10). The K_j's can be included in the expressions for $Var(B)$ and $Cov(A,B)$ by taking the average number of replications per random observer.

4.3.3 Estimation and inference using variance components

Method C from the previous chapter defines a 2-way random effects model to estimate ψ^N and ψ^R. Since this method treats observers as a random effect, it is

appropriate to use it in the case where *J* observers are selected at random. It can also accommodate an unequal amount of replicates per observer.

The model for ψ^N is exactly the same as specified in section 3.3.3 where the variance of the estimator was approximated by using the delta method. One can use the appropriate variance components in the estimate in (3.13).

Estimating ψ^R is also similar. The estimation is accomplished by fitting one random effects model for all observations with the fixed reference observer *J*, and another using all observations in the data set. The estimated ψ^R will be

$$\hat{\psi}^R = \frac{\hat{\sigma}^2_{\varepsilon_J}}{\hat{\sigma}^2_\beta + \hat{\sigma}^2_\gamma + \hat{\sigma}^2_\varepsilon} \quad . \tag{4.15}$$

The estimation of $\sigma^2_{\varepsilon_J}$ can be accomplished by taking the estimated error variance component from the model restricted to measurements by observer *J*. Inference can be conducted by using the simple bootstrap percentile method to compute standard errors and 95% confidence intervals.

4.4 Simulation study – performance and comparison of estimates

A simulation study was conducted to examine the behavior of the ψ^N and ψ^R estimates derived in this chapter and their estimated standard errors, for the case where a set of *J* observers are selected at random from a pool of potential observers. Data was simulated using a two-step approach. First, using the simple latent class model described in section 3.4, values for Y_{ijk} were simulated for each of 100 observers, each

with slightly different parameters for the conditional distributions. Initial values were drawn from either a standard normal or exponential distribution ($T \sim N(0,1)$ or $T \sim Exp(1)$). The T values were then used to generate values for Y_{ijk} using the conditional distribution $Y_{ijk} \mid t \sim N(a+bt,(e+ft)^2)$, where b varied over a set range of values for the 100 observers. Next, a sample of 3 observers was selected from the pool of 100, and 3 replications for each observer were simulated and used to compute estimates of ψ^N. In the case where one observer is considered to be a reference, observer 100 was used as the fixed reference for every simulation. To compute estimates of ψ^R, only 2 observers ($J-1$) were selected at random. One-thousand simulations were run to demonstrate the methods described in sections 4.3.1, 4.3.2, and 4.3.3. Biases, MSE's, and coverage probabilities were computed to assess the performance of the estimates and standard errors. The true values for ψ^N and ψ^R were computed for each simulation using the formulae from (Haber and Barnhart, 2008) expressed in section 3.4.

The simulation results are presented in Table 4.1. Overall, the bias and MSE were low for simulations where the underlying distribution was standard normal. Bias increased when using an exponential distribution to simulate the initial data. Both the inference method based on Method A and the method using a random effects model performed well, with coverage probabilities close to 0.95. The random effects model consistently produced lower standard errors than the modified Method A. Coverage probabilities using simple bootstrap resampling to estimate the standard error were somewhat lower than with the other methods, especially when interobserver variability was high. The standard errors for $\hat{\psi}^R$ always higher than those for $\hat{\psi}^N$.

4.5 Examples

4.5.1 Bland and Altman Blood Pressure (SBP) Data

The systolic blood pressure data set used in the previous chapter is now used to demonstrate the methods developed for random observers. Table 4.2 gives overall estimates of ψ^N and ψ^R for all observers, rather than the pairwise estimates presented in Chapter 3, and they are estimated the same way for fixed and random observers. The semi-automatic blood pressure monitor measurements are now used as the reference, to allow for an overall estimate of ψ^R comparing human observers versus the machine. The estimates of ψ^N and ψ^R are low, with both coefficients below 0.25, indicating poor agreement between the 3 observers with and without one observer being considered a reference.

Standard error estimates for $\hat{\psi}^N$ and $\hat{\psi}^R$ are computed separately assuming both fixed observers using the method described in Section 3.8 and random observers using the modified Method A described in this chapter, and also for random observers using a random effects model. The standard error is higher for $\hat{\psi}^N$ when observers are assumed to be fixed, and the standard error is higher for $\hat{\psi}^R$ when observers are assumed to be random. Ninety-five percent confidence intervals using all methods do not contain 1, indicating the coefficients are significantly lower than the null.

4.6.2 Carotid Stenosis Data

Table 4.3 demonstrates the methods for random observers in the same way using the previously analyzed carotid stenosis data. The angiogram is now considered the reference when estimating ψ^R, to allow for overall estimates of agreement between the MRA methods and the angiogram. Results are presented separately for the left and right carotid arteries.

Once again, the standard error is consistently higher for $\hat{\psi}^N$ when observers are assumed to be fixed, and the standard error is consistently higher for $\hat{\psi}^R$ when observers are assumed to be random. The random effects model gives more precise interval estimates than the modified Method A. Estimates of ψ^N are higher than those for ψ^R for both the left and right arteries, indicating that agreement is better when all three observers are compared to one another rather than two observers compared against a reference. This is due to the high agreement between MRA methods which does not contribute to the estimate of ψ^R.

4.6 Discussion

We derived two approaches each to allow inference on the two coefficients of individual agreement described in chapter 3 to be estimated when a set of 2 or more observers is selected randomly. Both methods derived here are based on the MSD as a disagreement function, and can be used to estimate pairwise agreement between two observers or overall agreement between a group of $J > 2$ observers.

Method A described in Chapter 3 was modified for the random observers case, given it was more straightforward to adapt from the pairwise case to more than 2

observers than Method B, which used pairwise subject-specific estimates. The main drawback here is that Method A requires the use of MSD for the disagreement function. This method performed quite well in simulations though, and would be useful in practice.

Estimating the coefficients of individual agreement using variance components from a random effects model is an alternative method which also easily estimates overall agreement in a group of $J > 2$ randomly selected observers. The same drawbacks exist as discussed in Section 3.9, since the model parameters are once again assumed to be mutually independent, and the random method parameters are assumed to be normally distributed.

Table 4.1: Simulation results for ψ^N and ψ^R estimates assuming random observers based on 1000 samples. Coverage probabilities computed for 2-sided 95% confidence intervals using the indicated s.e.

Distribution of T	n	*True (Mean)* ψ^N	*True (Mean)* ψ^R	$\hat{\psi}^N$ Mean	$\hat{\psi}^N$ Bias	$\hat{\psi}^N$ MSE	$\hat{\psi}^R$ Mean	$\hat{\psi}^R$ Bias	$\hat{\psi}^R$ MSE
Normal:									
a=0 b=range(0.2 – 2)	50	0.876	0.780	0.878	0.002	0.015	0.778	-0.002	0.036
e=1 f=1.1	100	0.875	0.784	0.875	0.001	0.011	0.785	0.001	0.024
	200	0.874	0.783	0.873	-0.0002	0.008	0.784	0.001	0.018
a=0 b=range(0.5 – 5)	50	0.579	0.409	0.583	0.003	0.046	0.411	0.002	0.046
e=1 f=1.1	100	0.592	0.426	0.595	0.003	0.043	0.428	0.002	0.044
	200	0.576	0.415	0.575	-0.001	0.040	0.416	0.001	0.039
Exponential:									
a=0 b=range(0.5 – 5)	50	0.588	0.411	0.645	0.057	0.045	0.471	0.060	0.048
e=1 f=1.1	100	0.589	0.432	0.643	0.054	0.039	0.486	0.054	0.047
	200	0.592	0.431	0.643	0.051	0.039	0.486	0.055	0.045

		Method A				Random Effects Model (Method C)		Bootstrap			
		$s.e.(\hat{\psi}^N)$		$s.e.(\hat{\psi}^R)$		$s.e.(\hat{\psi}^N)$		$s.e.(\hat{\psi}^N)$		$s.e.(\hat{\psi}^R)$	
Distribution of T	n	Mean	Coverage Prob.	Mean	Coverage Prob.	Mean	Coverage Prob.	Mean	Coverage Prob.	Mean	Coverage Prob.
Normal:											
a=0 b=range(0.2 – 2)	50	0.106	0.980	0.199	0.979	0.079	0.951	0.098	0.965	0.186	0.962
e=1 f=1.1	100	0.078	0.988	0.146	0.977	0.064	0.943	0.071	0.965	0.133	0.959
	200	0.056	0.980	0.106	0.985	0.045	0.947	0.051	0.973	0.095	0.962
a=0 b=range(0.5 – 5)	50	0.083	0.955	0.148	0.965	0.079	0.942	0.074	0.915	0.089	0.914
e=1 f=1.1	100	0.062	0.958	0.110	0.976	0.058	0.941	0.052	0.917	0.066	0.918
	200	0.044	0.966	0.078	0.966	0.042	0.941	0.039	0.926	0.057	0.927
Exponential:											
a=0 b=range(0.5 – 5)	50	0.105	0.870	0.165	0.953	0.099	0.867	0.086	0.857	0.122	0.921
e=1 f=1.1	100	0.080	0.913	0.125	0.966	0.075	0.910	0.064	0.900	0.098	0.912
	200	0.059	0.946	0.091	0.963	0.058	0.933	0.052	0.921	0.079	0.914

Table 4.2: Estimation of ψ^N and ψ^R for Bland and Altman SBP data, assuming random observers. (Y_1=Observer1, Y_2=Observer2, Y_3=Machine)

MSD(Y_1,Y'_1)	MSD(Y_2,Y'_2)	MSD(Y_3,Y'_3)	MSD(Y_1,Y_2)	MSD(Y_1,Y_3)	MSD(Y_2,Y_3)	$\hat{\psi}^N$	$\hat{\psi}^R$
74.8	76.0	166	52.0	679	676	0.225	0.245

	Method A				Random Effects Model (Method C)	
Observers	$s.e.(\hat{\psi}^N)$	**95% CI (2-sided)**	$s.e.(\hat{\psi}^R)$	**95% CI (2-sided)**	$s.e.(\hat{\psi}^N)$	**95% CI (2-sided)**
Fixed	0.062	(0.104, 0.346)	0.076	(0.096, 0.395)	-	-
Random	0.058	(0.111, 0.340)	0.081	(0.087, 0.403)	0.080	(0.068, 0.383)

Table 4.3: Estimation of ψ^N and ψ^R for carotid stenosis data, assuming random observers. (Y_1=MRA-2D, Y_2=MRA-3D, Y_3=Angiogram)

	MSD(Y_1,Y'_1)	MSD(Y_2,Y'_2)	MSD(Y_3,Y'_3)	MSD(Y_1,Y_2)	MSD(Y_1,Y_3)	MSD(Y_2,Y_3)	$\hat{\psi}^N$	$\hat{\psi}^R$
Left Side	1153	1040	279	1245	1211	1461	0.632	0.209
Right Side	1137	1100	176	1219	959	1093	0.738	0.172

		Method A				Random Effects Model (Method C)	
	Observers	$s.e.(\hat{\psi}^N)$	**95% CI (2-sided)**	$s.e.(\hat{\psi}^R)$	**95% CI (2-sided)**	$s.e.(\hat{\psi}^N)$	**95% CI (2-sided)**
Left Side	Fixed	0.108	(0.421, 0.842)	0.072	(0.067, 0.351)	-	-
	Random	0.090	(0.455, 0.808)	0.093	(0.026, 0.392)	0.060	(0.514, 0.749)
Right Side	Fixed	0.108	(0.526, 0.949)	0.034	(0.104, 0.239)	-	-
	Random	0.089	(0.563, 0.912)	0.048	(0.077, 0.266)	0.077	(0.587, 0.889)

Chapter 5

Modeling Measures of Agreement

5.1 Introduction and notation

In this chapter, we model the measures for assessing agreement between two observers previously described in Chapter 2, as a function of additional variables measured in the study. This can be very important to study investigators since observer agreement will often differ between subjects based on subject-specific characteristics. Knowledge of how these characteristics affect agreement can help determine in which types of subjects a method performs poorly compared to a reference method.

Barnhart and Williamson (2001) used generalized estimating equations (GEE) to model the CCC for comparing two observers as a function of covariates. The CCC is dependent on between-subject variability though, and CCC estimates can decrease since this variability is often similar within subjects with similar covariate values. One should verify that the between-subject variability is consistent across different ranges of the covariates when using this modeling approach.

Choudhary (2007) described a Bayesian semiparametric approach for modeling agreement between two methods. This method models the TDI using tolerance bands, which estimate the range of differences between observers in a specified proportion, π, of the population as a function of a covariate. The mean function of differences between observers is modeled using penalized spline regression.

We seek to describe and demonstrate several models using a disagreement function, *G*, as our outcome variable. We are primarily interested in modeling the MSD, as it has been the focus of our work. As we will consider agreement for two observers, more than two observers, and cases where one observer can be considered a reference, we will continue to use the notation of chapter 4 where Y_{ijk} is the observation on subject *i* by observer *j*. Observer *J* will be considered the reference observer if there exists one. In the case of *H* subject-specific covariates, they will be denoted by $z_1,...,z_H$.

5.2 Pediatric Impact adherence data

To demonstrate modeling methods, we have chosen to analyze data from the Pediatric Impact study, a behavioral intervention to improve medication adherence in HIV-positive children. Medication adherence is evaluated at baseline over the past 1-month, and is measured as the percent of prescribed medication taken over the month. There are 3 observers: the child's caregiver, the child's clinic care team, and the electronic MEMS (Medication Event Monitoring System), denoted as Y_1, Y_2, and Y_3 respectively. The MEMS will be considered the reference observer where we consider one to exist.

Figure 5.1 shows boxplots of the raw data for each of the three observers. As is typical for adherence data, all observers are skewed, the caregiver and care team being heavily skewed with most of their data in the upper range. Table 5.1 gives some basic summary statistics for the three observers, and gives the estimated MSD's. Observer agreement is worst between the caregiver and MEMS. Not all subjects were evaluated by

each observer, so pairwise disagreement functions in all analyses can only be evaluated for those with available measurements from both observers in the pair.

Subject-specific covariates collected with the data include: caregiver relationship to child, caregiver primary language, child's gender, if HIV-status has been disclosed to the child, yes/no side effects from HIV medications, age, ever had an undetectable viral load, and viral load closest to the assessment.

5.3 Models

5.3.1 Two observers with covariates – least squares method

We restrict the first constructed model to the case where $J = 2$ observers and we have H subject-specific covariates. The outcome variable will be $G_i = G(Y_{ij}, Y_{ij'})$, which is the observed disagreement function, G, for subject i. The natural logarithm of G is preferred for the modeling outcome, since when G = MSD, the distribution of squared differences is naturally skewed. An alternative to the natural logarithm would be to model the mean absolute differences, which could be easier to model without a transformation if the data are normally distributed. Also, since the Pediatric Impact data has numerous cases of $G_i = 0$, the natural logarithm can only be used if set to a trivially small amount. We are using the MSD as our disagreement function, so $G_i = (Y_{ij} - Y_{ij'})^2$.

Least squares is the simplest method to fit the model, denoted as:

$$\log(G_i) = \mu + \sum_{h=1}^{H} \beta_h z_{ih} + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma^2). \qquad (5.1)$$

The results of the pairwise models for the data set are presented in Table 5.2. Most covariates were not statistically significant in the models. R-squared values ranged from 0.153 – 0.050. Whether or not HIV status had been disclosed was the strongest predictor of disagreement between the caregiver and the care team (p=0.006). Viral load was close to significantly associated with higher disagreement between the MEMS and caregivers (p=0.07). This is due to low medication adherence being a strong predictor of high viral load. If a low viral load is known by an observer, they would be more likely to predict better adherence, and more likely to agree with the electronic monitor.

Scatterplots of the outcome variables by log viral load are presented in Figure 5.2. Since the distribution of MSD's is very skewed and the log transformation normalizes the outcome, observations with high disagreement (which are correlated with high viral loads) are less influential in the model. An examination of model residuals showed no violations of the constant variance or normality assumptions for the error terms. If the log transformation had not been used, error terms from the resulting models did not appear to be normally distributed, which justifies the choice of the log outcome.

5.3.2 More than two observers, no reference – mixed model

Suppose we have $J \geq 3$ observers, where none of them are regarded as a reference observer. Here, the outcome is modified form the 2-observer case to the mean disagreement function of one observer compared to the other $J-1$ observers,

$G_{ij} = \sum_{j'=1, j \neq j'}^{J} G_i(Y_j, Y_{j'})/(J-1)$, and we model it using the following mixed model once again using the log transformation to account for the squared differences:

$$\log(G_{ij}) = \mu + \alpha_i + \beta_{Oj} + \sum_{h=1}^{H} \beta_h z_{ih} + \varepsilon_{ij} \quad . \qquad (5.2)$$

The observer effect in the model, β_{Oj}, represents the difference between observer j's measurements and the mean over all observers. This observer effect can be treated as either fixed or random, depending on the observer selection process. We can also include interaction terms between the observer effect and covariate effects to describe how covariates affect disagreement differently for different observers. This model can easily be fit using the SAS procedure MIXED:

```
proc mixed method=reml;
class id observer;
model logmeandisagree = observer/ chisq outpm=out residual;
random id;
run;
```

Once again using MSD as our disagreement function, the results from this mixed model using the Pediatric Impact data are presented in Table 5.3. Observer is treated as a fixed effect, since caregivers and care teams are not selected at random for a child. Viral load is used as a single covariate, but it is not statistically significant (p=0.09). No other covariates yielded p-values lesser than 0.1, so were not included in the model. The parameters for caregivers and the care team are both negative compared with the MEMS, indicating that they have higher agreement with the other two observers. This is consistent with the previously summarized data, since the caregiver and the care team are more likely to agree with each other than with the MEMS. Studentized residuals appeared slightly skewed, indicating a possibility of non-normal error terms.

5.3.3 More than two observers with a reference method – mixed model

Now suppose we have $J \geq 3$ observers, where observer J is considered a reference observer. We can define the outcome for each of the $J-1$ non-reference observers as $G_{ij} = G_i(Y_j, Y_J)$, and fit the same mixed models in (5.2). The observer effect in the model, β_{Oj}, will now correspond to the disagreement between the indicated observer and the reference observer.

The results from this model using the Pediatric Impact data with the MEMS as the reference method and viral load as a single covariate are presented in Table 5.4. Observers are once again treated as a fixed effect. Modeling the log MSD does not reveal a statistically significant difference between the caregiver and the care team for agreement with the reference method (p=0.15). This indicates that the disagreement is similar between the caregiver and the MEMS compare to between the care team and the MEMS. Studentized residuals were once again somewhat skewed, indicating a possibility of non-normal error terms.

5.3.4 Penalized spline regression models

The last model we consider for the disagreement function is a semiparametric regression model with $G_i = G(Y_j, Y_{j'})$ as the outcome. This model is defined as

$$\log(G_i) = \mu + f(z_i, \beta, \gamma) + \varepsilon_i \quad , \tag{5.3}$$

where the mean function f is modeled nonparametrically with penalized spline regression (Ruppert et al. (2003)). This function is defined as a pth degree spline model:

$$f(z_i, \beta, \gamma) = \beta_0 + \beta_1 z_i + \ldots + \beta_p z_i^p + \sum_{k=1}^{K} \gamma_k (z_i - c_k)_+^p \quad , \tag{5.4}$$

where K is the number of knots, $c_1,...,c_K$ are the knot locations, β is the vector of regression coefficients for a subject-specific covariate z_i, and γ is the vector of coefficients of the truncated polynomial basis functions $(z_i - c_1)^p,...,(z_i - c_K)^p$. This model does not assume that the covariate effects are linear.

We want to fit this semiparametric model using the Pediatric Impact data to better describe the potential nonlinear relationships between viral load and pairwise disagreement depicted in Figure 5.2. We first fit models for the pairwise outcomes using log viral load as the only nonparametric predictor in the *p*th degree spline model, and the intercept as the only parametric parameter. The results of fitting these models to the data are presented in Figure 5.3, where we see a nonlinear relationship in every plot. Disagreement increases slowly (or sometimes decreases) for lower values of log viral load, but then increases rapidly for high values. The knot locations are chosen as the $\left(\frac{k+1}{K+2}\right)$th sample location of the unique z_i's, where $k = 1,...,K$ and $K = \max(\frac{n}{4},20)$ (Wand et al. (2005)). Log viral load is fit in each model with K = 18 knots. We explored adding additional covariates available in the dataset as parametric parameters in addition to the described nonparametric relationship with log viral load. The only covariate which added an additional significant effect was disclosure of HIV-status to the child which was related to the log disagreement between caregiver and care team. The parameter was estimated as 1.861 with a p-value of 0.0189.

The model defined in (5.3) can also be used for the case of more than two observers, by using the outcome variable defined in Section 5.3.2, $G_{ij} = \sum_{j'=1, j\neq j'}^{J} G_i(Y_j, Y_{j'})/(J-1)$. Now, the model is parameterized the same as (5.2) with

the addition of the mean function f, fit through penalized splines. Agreement for the caregiver and care team was once again found to be higher than the MEMS ($p=0.01$ and $p<0.001$ respectively). This was expected since the parametric part of the model was fit the same as that described in Table 5.3.

5.4 Carotid stenosis data

Disagreement in the carotid stenosis dataset previously used to demonstrate coefficients of agreement is also modeled with a list of possible covariates including: age in years, gender, diabetes, peripheral vascular disease, and previous anticoagulant therapy.

Table 5.5 fits a mixed model without a reference method using MSD as the disagreement function. The observer effect is not significant when modeling log(MSD). Negative parameters for both MRA methods indicate better agreement with each other than the angiogram. Subjects with diabetes had significantly higher disagreement than those without diabetes ($p=0.026$).

Table 5.6 fits a mixed model where the angiogram is considered to be a reference method. Again, the observer effect is not significant. For this model, that indicates that agreement between MRA-2D and the angiogram is not significantly greater than agreement between MRA-3D and the angriogram. Once again, measurements from those subjects with diabetes had significantly higher disagreement with the reference ($p=0.046$).

The effect of the continuous age variable on the pairwise log(MSD)'s is modeled nonparametrically in a semiparametric penalized spline regression model. The results of fitting these models to the data are presented in Figure 5.4. Disagreement increases linearly with age in all cases. When fitting an observer effect in the semiparametric model similar to the mixed model with no reference observer, we find the same result as with Table 5.5. Neither MRA method had significantly better agreement with other methods (both p=0.9).

5.5 Discussion

We described four different methods for modeling a disagreement function for scenarios with 2 or more observers, and where one of the observers may or may not be considered a reference. Each of the models is capable of assessing the effect of multiple covariates on observer agreement. The described semiparametric model is more flexible in describing the effects of covariates since it is not fully parametric and the covariate effects do not have to be linear.

The performance of the four models was examined through analysis on two data sets, which revealed similar conclusions for the effects of observers as previously examined comparison measures. Some covariate effects were found to influence agreement between different observers. This information is important, as it can describe the conditions that would cause an observer to give a less reliable measurement.

The models are defined for the case where all covariates of interest are subject-specific. Although not present in our datasets, some covariates can be observation-

specific, where they are measured separately with each replicated measurement by an observer. There are multiple ways to handle these covariates in the models we describe. The mixed models can incorporate replications and an interaction with the observer effect and the observation-specific covariate. It may also be possible to create a combined covariate incorporating information from both observers contributing to the pairwise disagreement function.

The described models all use a log transformed disagreement function as an outcome. It may also be possible to model the scaled coefficients, ψ^N and ψ^R although the distributions of the subject-specific coefficients are much more unpredictable than a disagreement function. Modeling ψ^N and ψ^R involves defining the coefficients, $\psi^N = \frac{[G(X,X') + G(Y,Y')]/2}{G(X,Y)}$ and $\psi^R = \frac{G(X,X')}{G(X,Y)}$ separately for each subject. These prove to be quite variable, often having values well above 1, and would need suitable transformations to prevent extreme outliers from overly influencing the results.

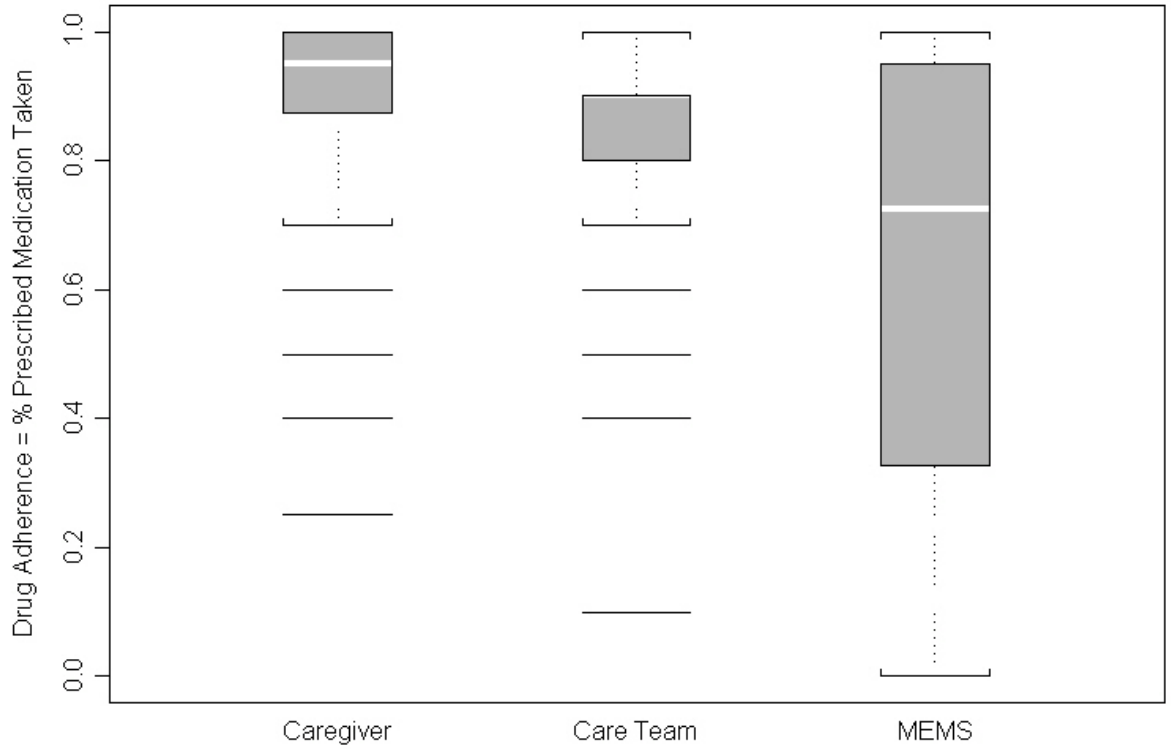

Figure 5.1: Box Plots for three observers – Pediatric Impact data set.

Table 5.1: Summary statistics and estimated MSD's – Pediatric Impact dataset.

Observer	Y	n	Mean	Median	Std. Dev.	Range
Caregiver	Y_1	160	0.898	0.950	0.144	(0.25 – 1.0)
Care Team	Y_2	119	0.829	0.900	0.160	(0.10 – 1.0)
MEMS	Y_3	109	0.633	0.724	0.340	(0.0 – 1.0)
	$MSD(Y_1,Y_2) = 0.026$ **$MSD(Y_1,Y_3) = 0.175$** **$MSD(Y_2,Y_3) = 0.134$**					

Table 5.2: Least squares method – Pediatric Impact dataset results.

	$\log(MSD(Y_1,Y_2))$			**$\log(MSD(Y_1,Y_3))$**			**$\log(MSD(Y_2,Y_3))$**		
	Estimate	SE	p-value	Estimate	SE	p-value	Estimate	SE	p-value
Intercept	-8.659	1.384	< 0.001	-5.443	1.702	0.002	-4.335	1.310	0.002
log10 viral load	0.069	0.217	0.751	0.472	0.256	0.070	0.219	0.192	0.261
Side effects (yes/no)	0.777	0.953	0.418	-0.184	1.258	0.884	-0.836	0.894	0.355
Caregiver language is English (yes/no)	1.312	1.307	0.319	-0.055	1.561	0.972	0.236	1.264	0.852
HIV disclosed (yes/no)	2.272	0.802	0.006	-0.741	0.958	0.443	-0.582	0.731	0.430

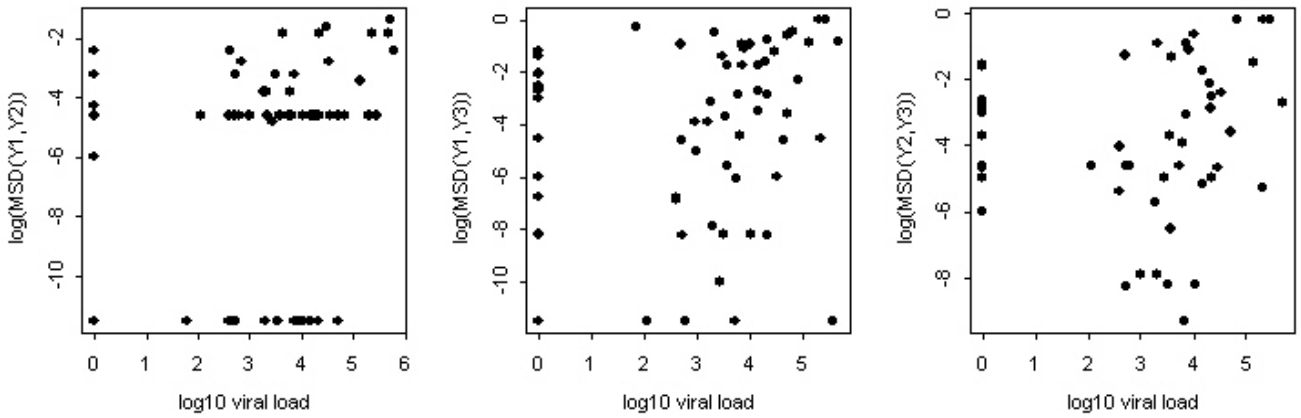

Figure 5.2: Scatterplots for pairwise log(MSD)'s by viral load – Pediatric Impact dataset.

Table 5.3: Mixed model, no reference observer – Pediatric Impact dataset results.

	Estimate	SE	p-value	95% CI
log (MSD_{ij}):				
intercept	-4.340	0.530	<0.001	(-5.405 , -3.274)
observer:				
Caregiver	-0.223	0.082	0.008	(-0.386 , -0.060)
Care Team	-0.301	0.082	<0.001	(-0.464 , -0.138)
MEMS	0	-	-	-
log10 viral load	0.270	0.158	0.090	(-0.043 , 0.582)

Table 5.4: Mixed model, with reference observer – Pediatric Impact dataset results.

	Estimate	SE	p-value	95% CI
log (MSD_{ij}):				
intercept	-4.704	0.669	<0.001	(-6.041 , -3.367)
observer:				
Caregiver	-0.526	0.356	0.146	(-1.241 , 0.189)
Care Team	0	-	-	-
log10 viral load	0.291	0.184	0.120	(-0.079 , 0.660)

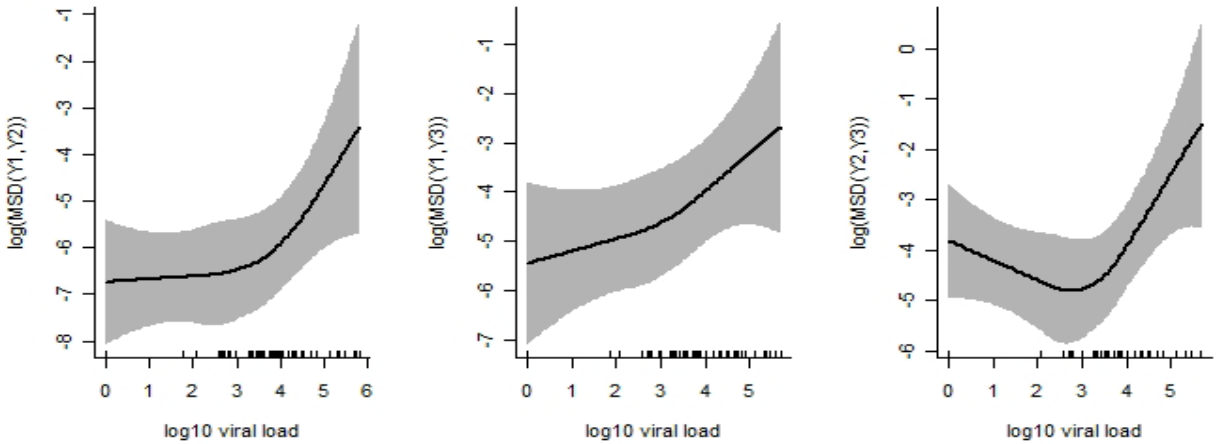

Figure 5.3: Semiparametric fit with shaded standard error bands for pairwise log(MSD)'s by viral load using truncated polynomial basis functions – Pediatric Impact dataset.

Table 5.5: Mixed model, no reference observer – carotid stenosis dataset results.

	Estimate	SE	p-value	95% CI
log (MSD_{ij}):				
intercept	3.32	0.686	<0.001	(1.95 , 4.70)
observer:				
MRA-2D	-0.032	0.571	0.955	(-1.16 , 1.09)
MRA-3D	-0.051	0.571	0.929	(-1.18 , 1.07)
Angiogram	0	-	-	-
diabetes (yes/no)	3.16	1.410	0.026	(0.388 , 5.94)

Table 5.6: Mixed model, with reference observer – carotid stenosis dataset results.

	Estimate	SE	p-value	95% CI
log (MSD_{ij}):				
intercept	-0.944	1.34	0.483	(-3.63 , 1.74)
observer:				
MRA-2D	0.256	0.70	0.715	(-1.13 , 1.64)
MRA-3D	0	-	-	-
diabetes (yes/no)	2.87	1.42	0.046	(0.058 , 5.68)
previous anticoagulant tx	3.87	1.42	0.007	(1.06 , 6.69)

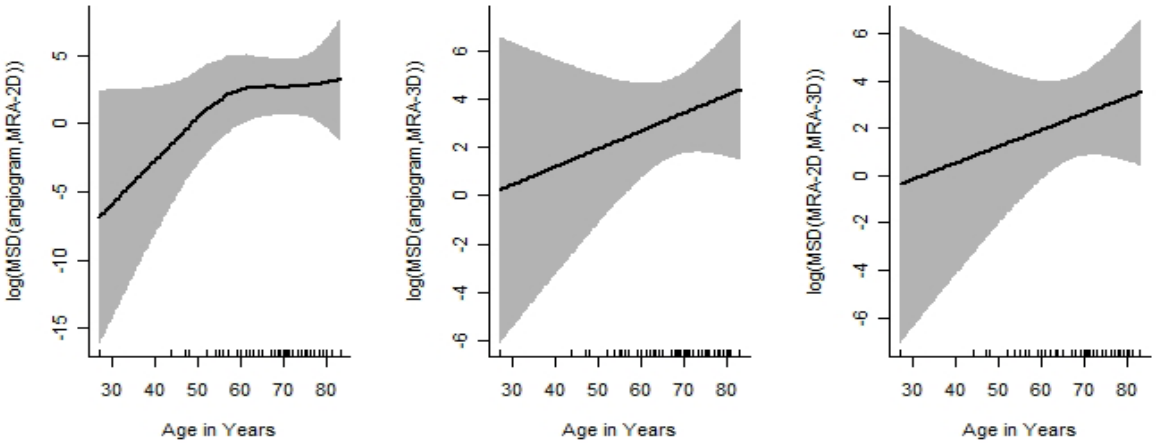

Figure 5.4: Semiparametric fit with shaded standard error bands for pairwise log(MSD)'s by age using truncated polynomial basis functions – carotid stenosis dataset.

Chapter 6

Summary

This research was focused on assessing agreement between observers or methods of measurement where measurements are continuous. A general approach to evaluating agreement between two observers with replicated measurements, known as a coefficient of individual agreement, was described and demonstrated on actual data sets. Several methods for inference were examined, both for the case when observers were considered fixed and when observers were selected randomly. Our research offers multiple options for analysis, depending on the nature of the data set, which performed well enough in simulations to justify their use in practice.

Future areas of research on this topic include development of more flexible methods of inference on ψ^N and ψ^R when observers are selected randomly. Our methods here rely solely on using the mean squared deviation to measure agreement. The MSD, while applicable in most cases, may be difficult to apply when the range of possible values for a method of measurement is limited.

Multiple useful models for describing agreement as a function of covariates were also described. These models used an unscaled disagreement function as their outcome. It may be possible in future research to develop models which actually use the scaled coefficients, ψ^N and ψ^R, as outcomes in a model describing covariate effects.

Bibliography

Anderson, S. and Hauck, W.W. (1990). Consideration of individual bioequivalence. *Journal of Pharmacokinetics and Biopharmaceutics*. **18**:259-273.

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

Barnhart H.X., Haber M., and Kosinski, A.S. (2007). Assessing individual agreement. *Journal of Biopharmaceutical Statistics*. **17**(4): 721-738.

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

Bland, J. M. and Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. **I**, 307-310.

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

Carrasco, J. L. and Jover, L. (2003). Estimating the generalized concordance correlation coefficient through variance components. *Biometrics*. **59**, 849-858.

Choudhary, P. K. and Nagaraja, H. N. (2005). Selecting the instrument closest to a gold standard. *Journal of Statistical Planning and Inference*. **129**, 229-237.

Choudhary, P.K. and Ng, H. K. T. (2006). Assessment of agreement under nonstandard conditions using regression models for mean and variance. *Biometrics*. **62**, 288-296.

Choudhary, P.K. (2007). Semiparametric regression for assessing agreement using tolerance bands. *Computational Statistics and Data Analysis*. **51**, 6229-6241.

Choudhary, P.K. and Nagaraja, H.N. (2007). Tests for assessment of agreement using probability criteria. *Journal of Statistical Planning and Inference*. **137**:279-290.

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

Dunn, O. J. and Clark, V. (1971). Comparison of tests of the equality of dependent correlation coefficients. *Journal of the American Statistical Association*. **66**, 904-908.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver and Reed.

Fisher, R.A. (1935). *The Design of Experiments*. New York: Hafner.

Galton, F. (1889). Family likeness in stature. *Proceedings of the Royal Society*. **40**:42-73.

Haber M, Barnhart HX. (2006). Coefficients of agreement for fixed observers. *Statistical Methods in Medical Research*. **15**: 1-17.

Haber M, Barnhart HX. (2007). A general approach to evaluating agreement between two observers or methods of measurement from quantitative data with replicated measurements. *Statistical Methods in Medical Research*. 1-19.

Haber M, Barnhart HX, Song J, Gruden J. (2005). Observer variability: a new approach in evaluating interobserver agreement. *Journal of Data Science*. **3**:69-83.

Hawkins DM (2002). Diagnostics for conformity of paired quantitative measurements. *Statistics in Medicine*. **21**: 1913-1935.

Hutson, A. D., Wilson, D. C. and Geiser, E. A. (1998). Measuring relative agreement: Echocardiographer versus computer. *Journal of Agricultural, Biological, and Environmental Statistics*. **3**, 163-174.

King, T. S., Chinchilli, V. M. (2001a). A generalized concordance coefficient for continuous and categorical data. *Statistics in Medicine*. **20**, 2131-2147.

King, T. S., Chinchilli, V. M. (2001b). Robust estimators of the concordance correlation coefficient. *Journal of Biopharmaceutical Statistics*. **11**:83-105.

Lee SS, Wiener J, Earp MJ, Simoni J, Demas P, Roa J, New M. Assessment of adherence to antiretroviral medications in children with HIV using the Medication Event Monitoring System. *Presentation at the 2006 NIMH/IAPAC International Conference on HIV Treatment Adherence, Jersey City, NJ*.

Lin, L. I. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*. **45**, 255-268.

Lin, L. I. (1992). Assay validation using the concordance correlation coefficient. *Biometrics*. **48**, 599-604.

Lin, L. I. (1997). Rejoinder to the letter to the editor by Atkinson and Nevill. *Biometrics*. **53**, 777-778.

Lin, L. I. (2000). Total deviation index for measuring individual agreement: with application in lab performance and bioequivalence. *Statistics in Medicine*. **19**, 255-270.

Lin, L. I., 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**, 257-270.

Lin, L. I., and Torbeck, L. D. (1998). Coefficient of accuracy and concordance correlation coefficient: new statistics for method comparison. *PDA Journal of Pharmaceutical Science and Technology*. **52**, 55-59.

Lord, F. M. and Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Reading, Massachusetts: Addison-Wesley.

McGraw, K. O. and Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*. **1**, 30-46.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. New York: Cambridge University Press.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: Wiley.

St. Laurent, R. T. (1998). Evaluating agreement with a gold standard in method comparison studies. *Biometrics*. **54**, 537-545.

Wand, M.P., Coull, B.A., French, J.L., Ganguli, B., Kammann, E.E., Staudenmayer, J. and Zanobetti, A. (2005). SemiPar 1.0. R package. http://cran.r-project.org