Statistics and Its Interface Volume 7 (2014) 113–120

# Marginal analysis of measurement agreement among multiple raters with non-ignorable missing ratings

Zhen Chen*,† and Yunlong Xie†

In diagnostic medicine, several measurements have been developed to evaluate the agreements among raters when the data are complete. In practice, raters may not be able to give definitive ratings to some participants because symptoms may not be clear-cut. Simply removing subjects with missing ratings may produce biased estimates and result in loss of efficiency. In this article, we propose a within-cluster resampling (WCR) procedure and a marginal approach to handle non-ignorable missing data in measurement agreement data. Simulation studies show that both WCR and marginal approach provide unbiased estimates and have coverage probabilities close to the nominal level. The proposed methods are applied to a data set from the Physician Reliability Study in diagnosing endometriosis.

Keywords and phrases: Fleiss $\kappa$, Scott $\pi$, Within-cluster resampling, Marginal approach.

## 1. INTRODUCTION

Endometriosis is a gynecological disorder in women that occurs when cells from the lining of the uterus grow in other areas of the uterus. The cause of endometriosis is unknown and a gold standard of diagnosing and staging does not exist. In order to better categorize the consistency and reliability in diagnosing endometriosis, we conducted the Physician Reliability Study (PRS) in collaboration with investigators at the University of Utah [1]. In the PRS, 12 physicians in obstetric and gynecology (OB/GYN) separately reviewed participants' clinical information (digital intra-uterus image taken during laparoscopy, surgeon's notes, magnetic resonance imaging, and histopathology reports) and assessed presence and staging of endometriosis. Among these physicians, 4 are international experts, 4 are local experts and the others are residents. Each physician conducted the review in a sequence of four settings, with each successive setting having an additional piece of clinical information. In this article, we focus on the outcome of the absence/presence of endometriosis, and to better reflect real clinical situation, restrict to the 8 physicians (4 local experts and 4 residents) who are practicing at the same medical center (Utah). Only data from the first setting, where the physicians reviewed the digital images only, were used.

An important scientific aim of the PRS is to estimate the agreement parameter among the physicians in terms of diagnosing endometriosis. To this end, one can apply the commonly used Fleiss kappa [2] to estimate the inter-rater agreement. Briefly speaking, Fleiss kappa is a chance-corrected measure of agreement among more than 2 raters and extends Scott's $\pi$ [3] that measures the agreement between two raters. Both Scott's $\pi$ and Fleiss kappa make the assumption that each rater (physician) returns a positive rating with a common probability. The Scott's $\pi$ and Fleiss kappa approach are also equivalent to the common correlation model in Donner and Eliasziw (1992) [4]. However, the PRS data pose a challenge in using the aforementioned approach directly given that some physicians only returned an "Indeterminant" diagnose for some participants. From the perspective of endometriosis diagnosis, these "Indeterminant" ratings are missing data. Among the 148 participants with digital images in setting 1, 88 (59%) participants have 1 or more of their 8 ratings missing. Section 4 provides more detailed information on the missingness.

With missing ratings present, a naive approach is to compute Fleiss kappa based on participants with complete data and remove those with 1 or more missing ratings from analysis. For the PRS, this practice leaves 60 participants who have all 8 ratings. Clearly, this naive approach is not desirable. First, throwing out participants with missing ratings will result in a loss of efficiency since the available sample size is reduced. Second and more importantly, by restricting to those with complete ratings, one implicitly assumes that the missing ratings are ignorable in the senses that the underlying missing mechanism is not related to the observed agreement. However, missing ratings in PRS could be non-ignorable. This can happen, for example, if missing ratings occur when the physicians are diagnosing a participant without clear-cut symptoms (hence giving an "Indeterminant" rating). In this situation, these physicians' ratings will resemble the outcomes of tossing a fair coin, resulting in low agreement. On the other hand, if the symptom is clear-cut,

*Corresponding author.
†The authors gratefully thank the Intramural Research Program of the National Institutes of health, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development.

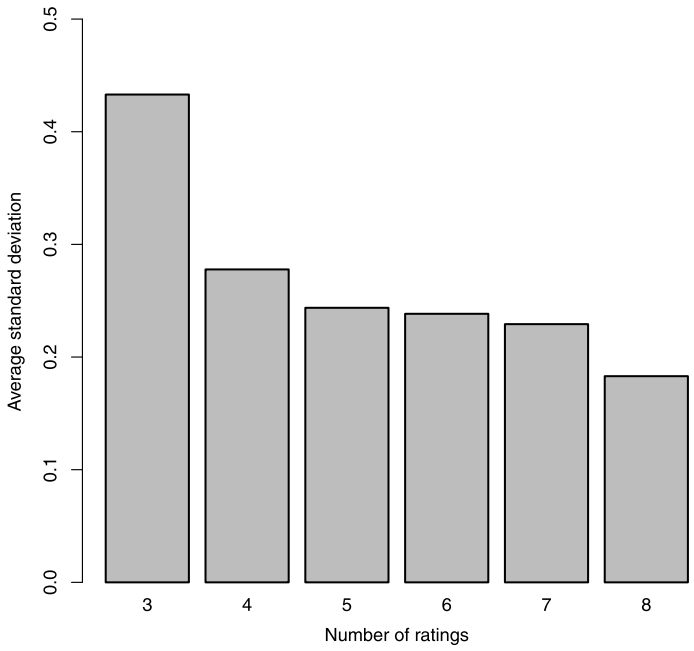


*Figure 1. Relationship between average standard deviation among non-missing ratings and the number of non-missing ratings. Since there is only one participant having only 2 ratings, we remove that participant in this figure.*

we would expect a high agreement on ratings. The PRS data seem to suggest such a relationship between agreement and missing ratings; see Figure 1 and Section 4 for more detailed discussions.

As the raters in PRS are similar in training and practice in the same medical center, and as the missing ratings are more a feature of participants rather than of raters, we make the assumption that the multiple ratings of the same participant are exchangeable. We can then treat the missing data in PRS as an informative cluster size problem [5–7]. The literature on informative cluster size have long established that the bias will be resulted if cluster size is not modeled appropriately.

In this paper, we propose approaches that handle informative cluster size in agreement analysis. We first introduce a resampling procedure similar to the within-cluster resampling of Hoffman et al. (2001) [6]. In this procedure, we randomly select 2 ratings from each participant and form a 2-rater data. We then compute Scott's $\pi$ measure for this 2-rater data set. This procedure is then repeated many times to obtain our estimates. Since this resampling procedure is computation intensive, we also introduce a marginal approach similar to Lorenz et al. (2011) [8]. In essence, we inversely reweight functionals in the Scott's $\pi$ formula and hence account for the informative cluster size in measurement agreement data. See Section 2.3 for details on how the reweightings are implemented. It's expected that these two proposed approaches produce similar results. In Section 2, we introduce notations, present the naive method, and propose the within-cluster resampling and the marginal approaches. We conduct simulation studies to evaluate the operating characteristics of the proposed methods in Section 4. In Section 4, we apply our proposed methods to the PRS data and summarize in Section 5.

## 2. METHODS

### 2.1 Notations and setup

Suppose there are $N$ participants rated by $J$ raters who are exchangeable. Let $i = 1, \ldots, N$ denote participants and $j = 1, \ldots, J$ denote raters. Let $Y_{ij} = 0/1$ denote the binary rating from the $j$-th rater for the $i$-th participant. Although the case of ordinal $Y_{ij}$ can be similarly considered, we focus on the binary rating situation for brevity in presentation. When $J = 2$, there are several different chance-corrected agreement measures that can be estimated when no missing data are present; for example Scott's $\pi$ (1955) [3] and Cohen's kappa (1960) [9]. Compared to Cohen's kappa, Scott's $\pi$ makes the assumption that the probability of a positive rating (i.e., prevalence) is the same for both raters. This assumption also leads to the widely used common correlation model [4]. Since the raters are exchangeable hence have a common prevalence, we focus on Scott's $\pi$ and its extensions. At population level, let $Y_1$ and $Y_2$ be random variables corresponding to the ratings from the two raters and $P_{kl} = P(Y_1 = k, Y_2 = l)$ for $l, k = 0, 1$. Further let $P_m$ be the common probability of a positive rating, i.e. $P_m = P(Y_l = 1), l = 0, 1$. As a chance-corrected measure of agreement, Scott's $\pi$ is defined as

$$\gamma_\pi = \frac{p_a - p_e}{1 - p_e},$$

where $p_a = P_{00} + P_{11}$ is the total agreement and $p_e = P_m^2 + (1 - P_m)^2$ is the agreement due to chance. Let $EX$ denotes the expectation of the random variable $X$. It is easy to show that $p_a = 1 - EZ$, where $Z = (Y_1 + Y_2)(2 - Y_1 - Y_2)$ and that $p_e = (EY_1)^2 + (1 - EY_1)^2$. Define $\omega_1 = EY_1$ and $\omega_2 = EZ$. Then we have

$$\gamma_\pi = \frac{p_a - p_e}{1 - p_e} = \frac{1 - \omega_2 - \omega_1^2 - (1 - \omega_1)^2}{1 - \omega_1^2 - (1 - \omega_1)^2}. \tag{1}$$

With a sample of ratings $(Y_{i1}, Y_{i2})$, $i = 1, \ldots, N$, Scott's $\pi$ can be estimated by the familiar formula (e.g., Donner and Eliasziw 1992 [4])

$$\hat{\gamma}_\pi = \frac{\hat{p}_a - \hat{p}_e}{1 - \hat{p}_e},$$

where $\hat{p}_a = 1 - \frac{1}{N}\sum_{i=1}^N s_i(2 - s_i)$, with $s_i = Y_{i1} + Y_{i2}$ being the number of positive ratings, and $\hat{p}_e = \hat{\pi}^2 + (1 - \hat{\pi})^2$, with $\hat{\pi} = \frac{1}{2N}\sum_{i=1}^N s_i$. For variance of $\hat{\gamma}_\pi$, Gwet (2008) [10] suggested:

*Table 1. A simple illustration of the WCR procedure for agreement data*

| Raw data | | | | | $q$-th WCR Pseudo data | | |
|---|---|---|---|---|---|---|---|
| participant | Rater | | | | participant | Pseudo rater | |
| | 1 | 2 | 3 | 4 | | 1 | 2 |
| 1 | $y_{11}$ | $\mathbf{y_{12}}$ | $\mathbf{y_{13}}$ | $y_{14}$ | 1 | $x^q_{11} = y_{12}$ | $x^q_{12} = y_{13}$ |
| 2 | $\mathbf{y_{21}}$ | $y_{22}$ | NA | $\mathbf{y_{24}}$ | 2 | $x^q_{21} = y_{21}$ | $x^q_{22} = y_{24}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $i$ | $\mathbf{y_{i1}}$ | NA | NA | $\mathbf{y_{i4}}$ | $i$ | $x^q_{i1} = y_{i1}$ | $x^q_{i2} = y_{i4}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $N$ | $y_{N1}$ | $y_{N2}$ | $\mathbf{y_{N3}}$ | $\mathbf{y_{N4}}$ | $N$ | $x^q_{N1} = y_{N3}$ | $x^q_{N2} = y_{N4}$ |

$$\hat{\text{var}}(\hat{\gamma}_\pi) = \frac{1}{N(N-1)} \sum_{i=1}^{N} (\hat{\gamma}^*_{\pi i} - \hat{\gamma}_\pi)^2, \tag{2}$$

where $\hat{\gamma}^*_{\pi i} = \hat{\gamma}_{\pi i} - 2(1-\hat{\gamma}_\pi)\frac{p_{e|i}-\tilde{p}_e}{1-\tilde{p}_e}$, $\hat{\gamma}_{\pi i} = \frac{p_{a|i}-\tilde{p}_e}{1-\tilde{p}_e}$, $\tilde{p}_e = \frac{1}{N^2J^2}[(\sum_{i=1}^N (J-s_i))^2 + (\sum_{i=1}^N s_i)^2]$, $p_{a|i} = \frac{1}{J(J-1)}[(J-s_i)(J-s_i-1)+s_i(s_i-1)]$ and $p_{e|i} = \frac{1}{NJ^2}[(J-s_i)\sum_{i=1}^N (J-s_i) + s_i \sum_{i=1}^N s_i]$. This variance estimate works in the presence of both high and low agreement.

## 2.2 Within-cluster resampling

When missing data are present, let $Y_{ij} = \text{NA}$ when rater $j$ did not give a definitive rating for participant $i$. In the general case of $J > 2$, the naive method is to simply ignore all the participants with any $Y_{ij} = \text{NA}$ for $j = 1, \ldots, J$ and compute Fleiss kappa (and its variance) based on the participants with no missingness. For the reasons mentioned in the Introduction, this approach is not desirable. Based on the assumption that the multiple ratings given to the same participant by the $J$ raters are exchangeable, we treat the missing data problem as one with informative cluster size, and consequently propose the within-cluster resampling (WCR) approach of Hoffman, Sen and Weinberg (2001) [6] to measurement agreement data with missing ratings. In short, WCR constructs a large number of datasets consisting of a single observation (two observations for association measures) from each and every participant (cluster) by random sampling (with replacement). Since the resulted dataset is free of repeated measurement, standard statistical techniques can be applied directly. Moreover, since resampling is done with replacement, observations from participants (clusters) with a smaller number of observations are weighted upward while those from clusters with a larger number of observations weighted downward. Here we propose to use WCR for estimating agreement when ratings are missing non-ignorably. Given that we are dealing with association measures, we only consider those participants with at least two definitive ratings.

Table 1 shows a simple illustration of the WCR procedure: in the left panel, we list the raw data that consist of $N$ participants, each rated by 4 raters. An "NA" is used to denote that a missing rating is recorded. To apply within-cluster resampling, we randomly select 2 ratings (bolded) from each participant. The resulted new data are collected in the right panel of Table 1 where each of the $N$ participants now has 2 ratings. Since these two ratings may not be from the same two raters for each participant, we term the corresponding raters "pseudo-raters". This blending of raters is one of the reasons that we focused on Scott's $\pi$ instead of Cohen's kappa and its multi-rater extensions, since the latters are not compatible with such a blending of raters. This "new pseudo" data can then be used to estimate Scott's $\pi$, $\hat{\gamma}_\pi$. Repeating this process a large number (say $Q$) times will produce $Q$ $\hat{\gamma}_\pi$'s, and a summarizing step will then be used to produce the WCR estimator of measurement agreement for the multi-rater data. To be more specific, let $\hat{\gamma}_\pi(q)$ be the Scott's $\pi$ computed from the $q$-th round of WCR; $\hat{\gamma}_\pi(q)$ can be obtained as in Section 2.1. The WCR estimator is then $\bar{\gamma}_\pi = Q^{-1}\sum_{q=1}^Q \hat{\gamma}_\pi(q)$. Following the argument of Hoffman, Sen and Weinberg (2001) [6], we have the limiting distribution of the WCR estimator: $\sqrt{N}(\bar{\gamma}_\pi - \gamma_\pi) \rightsquigarrow \text{N}(0, \Sigma_{WCR})$, where

$$\begin{aligned}\hat{\Sigma}_{WCR} &= \hat{\text{var}}(\sqrt{N}(\bar{\gamma}_\pi - \gamma_\pi)) \\ &= N\left[Q^{-1}\sum_{q=1}^{Q} \hat{\text{var}}(\hat{\gamma}_\pi(q)) - Q^{-1}\sum_{q=1}^{Q}(\hat{\gamma}_\pi(q) - \bar{\gamma}_\pi)^2\right],\end{aligned}$$

and $\hat{\text{var}}(\hat{\gamma}_\pi(q))$ is given in (2). As such, our WCR procedure converts the task of estimating the agreement from a dataset with multiple raters into one that involves repeatedly computing Scott's $\pi$ from a dataset with 2 raters. By random resampling with replacement, this procedure makes use of all data available while at the same time accounts for non-ignorable missingness by reweighting.

## 2.3 Marginal approach

Although straightforward, the WCR procedure can be computationally intensive. Williamson, Datta and Satten (2003) [7] proposed a marginal approach for clustered data with informative cluster size and showed that the marginal approach is asymptotically equivalent to a within-cluster resampling method. Specifically for estimating an agreement

measure for multiple raters, the marginal approach would inversely reweight each of the functionals in the Scott's $\pi$ formula with the cluster size and thus is equivalent to the within-cluster resampling method. Lorenz, Datta and Harkema (2011) [8] used this marginal approach to estimate correlations. Here, we propose a marginal approach to estimate measurements of agreement.

We consider ratings $(Y_{i1}, \ldots, Y_{iJ_i})$ for subject $i$ where $i = 1, \ldots, N$, and $J_i \in \{2, \ldots, J\}$. To introduce our marginal approach for these agreement data, we first define $F(y_1, y_2)$ to be the distribution function of a randomly chosen pair of ratings for a randomly chosen participant. This definition can be specified in the same spirit as in Lorenz, Datta and Harkema (2011) [8], as $(Y_{i1}, \ldots, Y_{iJ_i})$ can be used to create $\binom{J_i}{2}$ pairs of correlated data within a participant. Since $F(y_1, y_2)$ corresponds to a marginal bivariate distribution of the original ratings and implicitly defines a joint distribution $F(y_1, z)$, where $Z$ is defined in Section 2.1, we can use (1) to define a population-level agreement measure. We then define our estimator for $\gamma_\pi$ by replacing the population quantities in (1) with their sample counterparts. More specifically, the sample quantity corresponding to $\omega_1$ is $W_1 = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{J_i}\sum_{j=1}^{J_i} Y_{ij}$ and that to $\omega_2$ is $W_2 = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\binom{J_i}{2}}\sum_{j=1}^{J_i-1}\sum_{j'=j+1}^{J_i} s_{i,jj'}(2 - s_{i,jj'})$, where $s_{i,jj'} = Y_{ij} + Y_{ij'}$ for $j = 1, \ldots, J_i - 1, j' = j + 1, \ldots, J_i$. Heuristically speaking, the overall average of all the rating data $(W_1)$ is used to estimate $\omega_1 = EY_1$ since all the marginal means are the same. To obtain $W_2$, we first create a participant-specific average of $(Y_{ij} + Y_{ij'})(2 - Y_{ij} - Y_{ij'})$ in the form of $\frac{1}{\binom{J_i}{2}}\sum_{j=1}^{J_i-1}\sum_{j'=j+1}^{J_i} s_{i,jj'}(2 - s_{i,jj'})$. The combination formula $\binom{J_i}{2}$ and the double summation come from the realization that the rating data $(Y_{i1}, \ldots, Y_{iJ_i})$ can form $\binom{J_i}{2}$ pairs $(Y_{i1}, Y_{i2})$, $(Y_{i1}, Y_{i3})$, ..., $(Y_{i,J_i-1}, Y_{iJ_i})$ that can then be used to estimate $\omega_2$. As a result, the point estimator by our marginal approach can be expressed as

$$\hat{\kappa}_{mar} = \frac{1 - W_2 - W_1^2 - (1 - W_1)^2}{1 - W_1^2 - (1 - W_1)^2} = g(W_1, W_2).$$

Note that from (1) $\gamma_\pi = g(\omega_1, \omega_2)$. Given that we have independent data between participants, we can apply the central limit theorem and multivariate delta method to obtain a variance estimator of $\hat{\kappa}_{mar}$. We omit the details of those derivations to save space.

## 3. SIMULATION

In this section, we conduct simulation to evaluate the performance of the methods introduced in Section 2 in terms of biases and coverage probabilities. We first generate measurement agreement data for $J$ raters and then introduce various missing data mechanisms. The naive, WCR and marginal approaches are then applied to the data with missing ratings. The idea is to see which of the three approaches produces estimates that are close to the true values and have good coverage probabilities.

*Table 2. Point estimates of agreement and coverage probabilities when there is no missingness*

| | True Kappa | 0.2 | 0.5 | 0.8 |
|---|---|---|---|---|
| Naive | Point est | 0.196 | 0.495 | 0.795 |
| | Coverage | 93.3% | 95.3% | 95.5% |
| Proposed | Point est | 0.197 | 0.493 | 0.795 |
| | Coverage | 94.3% | 94.1% | 94.3% |
| Marginal | Point est | 0.196 | 0.495 | 0.795 |
| | Coverage | 93.3% | 95.3% | 95.5% |

### 3.1 Generate rating data

We consider three levels of agreement in terms of Fleiss kappa: high ($\kappa = 0.8$), median ($\kappa = 0.5$) and low ($\kappa = 0.2$). Given $\kappa$, we find a set of values of $p$ and $p_j$ where
(i) the marginal probabilities for all raters are the same: $P(Y_j = 1) = p$ for $j = 1, \ldots, J$, and
(ii) $Y_1 \sim$ Bernoulli$(p)$ and $Y_j|Y_1 = 1 \sim$ Bernoulli$(p_j)$ for some pre-specified $p_j \equiv P(Y_j = 1|Y_1 = 1)$ for $j = 2, \ldots, J$.
In (i), the common $p$ is needed since we are operating under the assumption of a common prevalence; see the discussion in Section 2.2; and by (ii), we assume that $Y_{j_1}$ and $Y_{j_2}$ are conditional independent given $Y_1$ for $j_1 \neq j_2 \geq 2$. To choose the set of $(p, p_j)$, we make use of the Fleiss Kappa formula as presented in equation (4a) of Warrens (2010) [14] and the following identities:

$$P(Y_1 = 1, Y_j = 1) = pp_j, j = 2, \ldots, J$$
$$P(Y_1 = 0, Y_j = 0) = 1 - 2p + pp_j, j = 2, \ldots, J$$
$$P(Y_{j_1} = 1, Y_{j_2} = 1) = p_{j_1}p_{j_2}p + \frac{p^2}{1-p}(1 - p_{j_1})(1 - p_{j_2}),$$
for $j_1, j_2 = 2, \ldots, J$
and
$$P(Y_{j_1} = 0, Y_{j_2} = 0)$$
$$= (1 - p_{j_1})(1 - p_{j_2})p + \frac{(1 - 2p + pp_{j_1})(1 - 2p + pp_{j_2})}{1 - p},$$
for $j_1, j_2 = 2, \ldots, J$.

Given $p$ and $p_j$'s, we can then simulate data $Y_{i1}, Y_{i2}, \ldots, Y_{iJ}$. More specifically, for each $i$, we generate $Y_{i1} \sim$ Bernoulli$(p)$, $Y_{ij}|Y_{i1} = 1 \sim$ Bernoulli$(p_j)$, $Y_{ij}|Y_{i1} = 0 \sim$ Bernoulli$(P(Y_j = 1|Y_1 = 0))$ for $j = 2, \ldots, J$. The probability $P(Y_j = 1|Y_1 = 0)$ can be determined as:

$$P(Y_j = 1|Y_1 = 0) = \frac{p(1 - p_j)}{1 - p}.$$

### 3.2 Introduce missing data

For completeness, we applied the three approaches to the simulated data before introducing missingness. As shown in Table 2, all point estimates of agreement are nearly unbiased and all coverage probabilities are close to 0.95. To introduce

*Table 3. Percentages of participants with missing ratings in different scenarios. In scenario 1, the missingness is completely at random and the percentage of the missing rate are the same for different values of true Fleiss kappa. For this reason, we list the case of $\kappa = 0.2$ only. In scenario 3, $a = -4$ is used for all $b$ and $\kappa$ values*

| Scenario | Parameter | | Number of missing ratings per participant | | | | | | | Overall |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | |
| | $\kappa$ | $q_1$ | | | | | | | | |
| | 0.2 | 0.1 | 53.0% | 35.5% | 10.0% | 1.5% | 0.1% | 0.0% | 0.0% | 7.5% |
| 1 | 0.2 | 0.2 | 26.2% | 39.5% | 24.5% | 8.2% | 1.5% | 0.2% | 0.0% | 15.0% |
| | 0.2 | 0.3 | 11.8% | 30.3% | 32.3% | 18.5% | 5.9% | 1.1% | 0.1% | 22.5% |
| | True $\kappa$ | $q_2$ | | | | | | | | |
| | 0.2 | 0.1 | 65.1% | 28.4% | 5.8% | 0.6% | 0.0% | 0.0% | 0.0% | 5.3% |
| | 0.2 | 0.2 | 41.9% | 37.6% | 16.1% | 3.9% | 0.5% | 0.0% | 0.0% | 10.5% |
| | 0.2 | 0.3 | 26.8% | 36.1% | 24.5% | 9.9% | 2.3% | 0.3% | 0.0% | 15.7% |
| | 0.5 | 0.1 | 68.2% | 25.4% | 5.7% | 0.7% | 0.0% | 0.0% | 0.0% | 4.9% |
| 2 | 0.5 | 0.2 | 47.7% | 32.6% | 14.9% | 4.2% | 0.6% | 0.1% | 0.0% | 9.7% |
| | 0.5 | 0.3 | 34.6% | 30.7% | 21.5% | 10.0% | 2.8% | 0.4% | 0.0% | 14.6% |
| | 0.8 | 0.1 | 83.1% | 13.4% | 3.0% | 0.4% | 0.0% | 0.0% | 0.0% | 2.6% |
| | 0.8 | 0.2 | 72.3% | 16.8% | 8.0% | 2.4% | 0.4% | 0.0% | 0.0% | 5.3% |
| | 0.8 | 0.3 | 65.6% | 15.5% | 11.2% | 5.7% | 1.7% | 0.3% | 0.0% | 7.9% |
| | True $\kappa$ | $b$ | | | | | | | | |
| | 0.2 | 13 | 60.2% | 18.7% | 11.9% | 6.4% | 2.3% | 0.5% | 0.0% | 9.2% |
| | 0.2 | 15 | 50.0% | 14.6% | 12.5% | 10.6% | 7.9% | 3.6% | 0.8% | 15.7% |
| | 0.5 | 13 | 80.2% | 11.8% | 5.1% | 2.1% | 0.6% | 0.1% | 0.0% | 3.9% |
| 3 | 0.5 | 15 | 72.7% | 11.9% | 7.4% | 4.5% | 2.4% | 1.0% | 0.2% | 7.0% |
| | 0.8 | 13 | 95.6% | 3.4% | 0.8% | 0.2% | 0.0% | 0.0% | 0.0% | 0.7% |
| | 0.8 | 15 | 92.9% | 4.5% | 1.6% | 0.7% | 0.3% | 0.1% | 0.0% | 1.4% |

missingness, let $R_{ij} = I(Y_{ij} = \text{NA})$ be the missing indicator. For each participant $i$, we randomly select $J-2$ ratings and assign missingness according to these mechanisms:

1. $P(R_{ij} = 1) = q_1$ for some pre-specified $q_1$;
2. $P(R_{ij} = 1|Y_{ij} = 1) = q_2$ for some pre-specified $q_2$;
3. $P(R_{ij} = 1) = a + b\text{var}(Y_{i1}, \ldots, Y_{iJ})$ for some pre-specified $a$ and $b$.

Scenario 1 corresponds to ignorable missingness since the probability of missing does not depend on the observed or unobserved agreement. In contrast, scenarios 2 and 3 correspond to nonignorable missingness since the probability of missing depends on the prevalence (scenario 2) or the agreement through the variability among the $J$ ratings (scenario 3). In Table 3, we present the average percentage of participants with missing ratings based on the 1,000 simulated data sets for different values of $q_1$, $q_2$, and $b$ ($a = -4$ in scenario 3). More specifically, in the middle columns of the table, we list the average percentage of the participants who have $0, 1, \ldots, 6$ ratings missing respectively for each scenario; in the rightmost column, we list the average percentage of the missing ratings per simulated data set. Since the missing data mechanism is free of agreement in scenario 1, the percent of participants with missingness is the same for different kappa values. For this reason, we only present the missing percentages when the true $\kappa = 0.2$ in scenario 1. In all three scenarios, for any given true agreement level, the degree of missingness, both in terms of the distribution of number of missing ratings per participant and in overall missing ratings, increases as the design parameter ($q_1$, $q_2$, and $b$ for scenarios 1, 2 and 3 respectively) increases. For example, in scenario 1, when $q_1$ increases from 0.1 to 0.2, the percentage of participants with no missing rating decreases from 53.0% to 26.2%, while the percentage of participants with one missing rating increases from 35.5% to 39.5%.

## 3.3 Simulation results

For each generated data, we apply the naive, WCR and marginal approach to obtain the estimated agreement measure. In implementing WCR, we choose to use $Q = 10{,}000$ iterations. We summarize these estimates with their coverage probabilities in Table 4 with missingness subject to scenarios 1, 2 and 3, respectively. For ignorable missingness (scenario 1), all three methods work well, with no discernable differences in point estimates and coverage probabilities. This is expected as the missingness does not depend on the ratings or the agreement. However, when the missingness is non-ignorable (scenarios 2 and 3) the proposed resampling and marginal methods perform better than the naive method. More specifically, for scenario 2, when the missing probability and the true agreement are both high, the estimates by the naive method are biased and have low coverage probabilities. For example, when $q_2 = 0.3$ and true Fleiss $\kappa = 0.8$, the point estimate by naive method is 0.603 with coverage probability 0.425. For scenario 3, when the true Fleiss $\kappa$ is 0.5, the naive method works poorly since

*Table 4. Estimated measurements of agreement and the coverage probabilities based on 1,000 simulated data sets with missingness according to scenarios 1, 2 and 3*

| | | Estimate | | | Coverage | | |
|---|---|---|---|---|---|---|---|
| True $\kappa$ | $q_1/q_2/b$[1] | Naïve | WCR | Marginal | Naïve | WCR | Marginal |
| Scenario 1 | | | | | | | |
| 0.2 | 0.1 | 0.197 | 0.196 | 0.197 | 0.930 | 0.934 | 0.953 |
| | 0.2 | 0.193 | 0.197 | 0.199 | 0.916 | 0.931 | 0.943 |
| | 0.3 | 0.180 | 0.195 | 0.197 | 0.862 | 0.932 | 0.943 |
| 0.5 | 0.1 | 0.498 | 0.500 | 0.501 | 0.941 | 0.958 | 0.959 |
| | 0.2 | 0.493 | 0.498 | 0.499 | 0.944 | 0.947 | 0.951 |
| | 0.3 | 0.482 | 0.499 | 0.499 | 0.923 | 0.937 | 0.937 |
| 0.8 | 0.1 | 0.797 | 0.798 | 0.798 | 0.951 | 0.941 | 0.944 |
| | 0.2 | 0.795 | 0.799 | 0.799 | 0.933 | 0.951 | 0.953 |
| | 0.3 | 0.790 | 0.798 | 0.798 | 0.930 | 0.954 | 0.957 |
| Scenario 2 | | | | | | | |
| 0.2 | 0.1 | 0.208 | 0.197 | 0.199 | 0.954 | 0.933 | 0.944 |
| | 0.2 | 0.212 | 0.201 | 0.202 | 0.931 | 0.935 | 0.937 |
| | 0.3 | 0.204 | 0.200 | 0.201 | 0.953 | 0.958 | 0.961 |
| 0.5 | 0.1 | 0.509 | 0.500 | 0.501 | 0.953 | 0.954 | 0.956 |
| | 0.2 | 0.494 | 0.498 | 0.499 | 0.959 | 0.952 | 0.957 |
| | 0.3 | 0.444 | 0.497 | 0.497 | 0.868 | 0.943 | 0.943 |
| 0.8 | 0.1 | 0.772 | 0.797 | 0.797 | 0.905 | 0.946 | 0.951 |
| | 0.2 | 0.716 | 0.796 | 0.797 | 0.747 | 0.962 | 0.963 |
| | 0.3 | 0.603 | 0.793 | 0.793 | **0.425** | 0.959 | 0.957 |
| Scenario 3 | | | | | | | |
| 0.2 | 13 | 0.277 | 0.195 | 0.197 | 0.706 | 0.929 | 0.941 |
| | 15 | 0.295 | 0.195 | 0.196 | 0.713 | 0.932 | 0.943 |
| 0.5 | 13 | 0.599 | 0.500 | 0.501 | **0.208** | 0.954 | 0.956 |
| | 15 | 0.642 | 0.499 | 0.499 | **0.058** | 0.948 | 0.945 |
| 0.8 | 13 | 0.826 | 0.798 | 0.798 | 0.774 | 0.940 | 0.941 |
| | 15 | 0.841 | 0.799 | 0.799 | 0.537 | 0.934 | 0.934 |

[1] Parameters used to generate missing data probability in the three scenarios. See Section 3.2.

*Table 5. Distribution of number of missing ratings in PRS data*

| | Number of Missing | | | | | | | | | Overall |
|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | |
| $n$ | 60 | 25 | 28 | 20 | 9 | 4 | 1 | 1 | 0 | 210 |
| % | 40.82% | 17.01% | 19.05% | 13.61% | 6.12% | 2.72% | 0.68% | 0.68% | 0% | 17.74% |

the estimates are seriously biased and the coverage probabilities extremely low. In contrast, the WCR and marginal method both work well for scenarios 2 and 3: the estimates by both methods are close to the true values and the coverage probabilities close to 0.95. In summary, the proposed resampling method and the marginal method are asymptotically equivalent and robust with respect to different types of missingness. In addition, both perform better than the naive approach when missingness is not ignorable and as good as the naive one when missingness is ignorable.

## 4. ANALYSIS OF THE PRS DATA

We conducted the PRS with the aim to understand how physicians agree with each other in diagnosing endometriosis. More detailed descriptions of the PRS are provided elsewhere (Schliep et al. 2012). [1] In this paper, we sought to examine how the 8 physicians who are practising at the University of Utah medical center agree with each other (inter-rater agreement) when they review the intra-uterus digital images of the participants (setting 1). In the PRS, the physicians were expected to determine whether a participant has endometriosis after reviewing the clinical information. However, they were also allowed to assign an "indeterminant" answer when they felt unsure whether endometriosis is present. For this reason, missing ratings occured in some participants. Among the 148 participants with digital images at setting 1, only 60 have all 8 ratings. The majority of those with missing ratings have either one or two ratings missing (Table 5). Only one par-

ticipant has seven or more missing ratings. She was removed from the analysis because the minimum number of available ratings has to be 2 for the proposed methods to work.

There is a reason to think that the missing ratings may be related to the agreement so that the missingness is non-ignorable. When the symptoms of endometriosis are clear-cut, physicians are less likely to give "indeterminant" and more likely to give similar ratings, resulting in a higher agreement for participants with fewer non-missing ratings. The PRS data as plotted in Figure 1 seems to confirm this negative relationship between agreement (as approximated by the variability among the non-missing ratings) and number of available ratings. Among those with 4 or fewer ratings, the average variability (standard deviation of the ratings) is 0.36 while among those with 5 or more ratings, it is 0.22. If missingness is ignorable, then the two groups, fewer/more ratings, should have about the same variability.

The naive approach only uses participants with all 8 ratings available. As a result, we use the 60 qualified participants to estimate the Fleiss kappa in R (function "kappam.fleiss" in package "irr"; Gamer et al. 2012 [12]). The obtained point estimate is 0.586 with a 95% confidence interval $(0.466, 0.705)$. This indicates a moderate inter-rater agreement in diagnosing endometriosis among the eight physicians (Landis and Koch, 1977 [13]).

In applying the WCR approach to the PRS study, we use data from 147 participants with at least 2 ratings. Using $Q = 10{,}000$, we obtain a point estimate of 0.536, with 95% confidence interval (0.454, 0.617). The marginal approach produces very similar results, as expected, with a point estimate 0.537, with 95% confidence interval (0.455, 0.619). This suggests that the naive approach can inflate the agreement by about 10%. In addition, the 95% confidence interval of the estimate from the naive approach is wider than those from the proposed approaches (0.24 versus 0.17 in width), suggesting that the proposed approaches are more efficient.

## 5. DISCUSSIONS

In this article, we proposed new methods to estimate agreement measures for multi-rater data when missing ratings are present. In the WCR procedure, we randomly select 2 ratings from each participant, form a 2-rater data and then compute Scott's $\pi$ measure for this 2-rater data set. We repeat this step many times to obtain our estimates. The estimate obtained by this procedure is asymptotically equivalent to a marginal approach and has smaller biases and closer to correct coverage probabilities than the naive method which simply ignores all the participants with missing ratings. In the PRS, the estimates of agreement obtained by WCR and the marginal approach are smaller than the estimate obtained by the naive method, suggesting that the estimates from the naive method might be biased. Moreover, the estimates from the proposed methods appear to be more efficient than those from the naive approach. This is due to the fact that the proposed approaches make use of all available data while the naive approach only uses a subset.

Although we only considered binary rating in this manuscript, the extension to categorical ratings are similar and straightforward. To obtain the estimate, we just compute Scott's $\pi$ for categorical ratings for each resampled 2-rater data set and then conduct a summary step if WCR is to be used, or reweight each of the functionals in the agreement formula for categorical ratings if the original approach is to be used. This will also work when weighted kappas are to be estimated.

We have focused on Scott's $\pi$ and Fleiss kappa in this manuscript. These agreement measures make the common prevalence assumption, which enables us to propose the WCR and marginal approach. The extension to Cohen's kappa, which assumes heterogeneous prevalence among raters, warrants further research.

In this manuscript, we have made the assumption that the multiple ratings given to the same participant in PRS are exchangeable so that the cluster-weighted estimation approach can be applicable. The exchangeability assumption is reasonable in PRS data because raters in PRS are similar to each other and because missing diagnoses in PRS occur when symptoms of endometriosis are not clear-cut so that missing diagnoses arise as a result of participant-specific rather than rater-specific characteristics. While reasonable for the PRS, the exchangeability assumption may not be appropriate for other situations where raters are inherently different and missing data are more a feature of the raters than the participants. As an example of this rater-specific missingness, consider a case where one rater tends to disagree with the other raters who are generally in agreement, and where the probability of missingness is associated with the disagreement. In this example, one cannot reasonably assume that the ratings are exchangeable and therefore cannot treat the problem as one with informative cluster size. One appropriate approach would be to treat it as a non-ignorable missing data problem and to specify models with explicit assumptions on the missing data mechanism.

## ACKNOWLEDGEMENTS

We thank the Center for Information Technology, the National Institutes of Health, for providing access to the high performance computational capabilities of the Biowulf cluster. This research was supported by the Intramural Research Program of the National Institutes of Health, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development.

*Received 11 February 2013*

## REFERENCES


[1] Schliep, K., Stanford, J. B., Zhang, B. et al. on behalf of The ENDO Study Working Group (2012). Inter- and intra-rater reliability in the diagnosis and staging of endometriosis: The ENDO Study. *Obstetrics and Gynecology* **120**(1) 104–112.

[2] Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin* **76**(5) 378–382.

[3] Scott, W. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly* **19**(3) 321–325.

[4] Donner, A. and Eliasziw, M. (1992). A goodness-of-fit approach to inference procedures for the kappa statistic: Confidence interval construction, significance-testing and sample size estimation. *Statistics in Medicine* **11** 1511–1519.

[5] Dunson, D. B., Chen, Z. and Harry, J. (2003). Bayesian joint models of cluster size and subunit-specific outcomes. *Biometrics* **59**(3) 521–530. MR2004257

[6] Hoffman, E. B., Sen, P. K. and Weinberg, C. R. (2001). Within-cluster resampling. *Biometrika* **88**(4) 1121–1134. MR1872223

[7] Williamson, J. M., Datta, S. and Satten, G. A. (2003). Marginal analyses of clustered data when cluster size is informative. *Biometrics* **59** 36–42. MR1978471

[8] Lorenz, D. J., Datta, S. and Harkema, S. J. (2011). Marginal association measures for clustered data. *Statistics in Medicine* **30**(27) 3181–3191. MR2861468

[9] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* **20**(1) 37–46.

[10] Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology* **61** 29–48. MR2522110

[11] SAS Institute, Inc. *Base SAS 9.2 Procedures Guide: Statistical Procedures*. 3rd edition. Cary, NC: SAS Institute Inc, 2010, p. 98.

[12] Gamer, M., Lemon, J., Fellows, I. and Singh P. Various Coefficients of Interrater Reliability and Agreement. http://cran.r-project.org/web/packages/irr/irr.pdf. (2012, accessed 15 September 2012).

[13] Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* **33**(1) 159–174.

[14] Warrens, M. J. (2010). Inequalities between multi-rater kappas. *Advances in Data Analysis and Classification* **4** 271–286. MR2748691



Zhen Chen
Biostatistics and Bioinformatics Branch
Division of Intramural Population Health Research
Eunice Kennedy Shriver National Institute of
Child Health and Human Development
National Institutes of Health
Bethesda, MD 20892
USA
E-mail address: chenzhe@mail.nih.gov

Yunlong Xie
Biostatistics and Bioinformatics Branch
Division of Intramural Population Health Research
Eunice Kennedy Shriver National Institute of
Child Health and Human Development
National Institutes of Health
Bethesda, MD 20892
USA
E-mail address: yunlong.xie@nih.gov