Estimating the Diagnostic Accuracy from Multiple Raters Based on a Bivariate Random Effects Model

Hiroyuki Saeki*1,*3, Toshiro Tango*2 and Jinfang Wang*3

*1Development Department, FUJIFILM RI Pharma Co., Ltd., 2-14-1 Kyobashi, Chuo-ku, Tokyo, 104-0031, Japan
*2Center for Medical Statistics, 2-9-6 Higashi Shinbashi, Minato-ku, Tokyo, 105-0021, Japan
*3Graduate School of Science, Chiba University, 1-33 Yayoi-cho, Inage-ku, Chiba, 263-8522, Japan
e-mail:hiroyuki.saeki@fujifilm.com

In clinical investigations designed to demonstrate the efficacy of a diagnostic procedure, the procedure is usually evaluated by multiple independent raters. Although the sensitivity and specificity may be estimated by considering consensus evaluations to treat results from multiple raters as if there were a single rater, raters are not considered independent in consensus evaluations. Typically, estimation methods are based on an “average rater” or a “majority rater” to account for multiple raters. In this paper, we propose a method for summarizing sensitivities and specificities evaluated from multiple independent raters based on a bivariate random effects model (BVRM) to account between-rater variance and correlation between sensitivity and specificity.

In addition, we propose methods to draw joint confidence regions of sensitivity and specificity based on the BVRM. Simulation results show that the differences in the biases between the proposed method and the average rater method are small and that the empirical coverage probabilities of the proposed joint confidence regions are close to the nominal level. The proposed methods are illustrated using data from florbetapir F 18 positron emission tomographic imaging to predict the presence of β-amyloid in the brains of subjects with Alzheimer’s disease.

Key words: Bivariate random effects model; Multiple raters; Sensitivity and specificity; Joint confidence region.

1. Introduction

Sensitivity and specificity are crucial measures for evaluating the performance of a diagnostic procedure. Sensitivity is generally defined as the probability of obtaining a positive result from a diagnostic procedure for a diseased subject. Specificity is the probability of obtaining a negative result from a diagnostic procedure for a non-diseased subject. In clinical investigations designed to evaluate the efficacy of a diagnostic procedure, results from the diagnostic procedure are
often evaluated by multiple independent raters (Lehr and Kashanian, 2009). Although the sensitivity and specificity may be estimated by considering consensus evaluations that treat multiple results from multiple raters as a single result from a single rater, this method is not recommended for a primary evaluation (FDA, 2004; Obuchowski and Lieber, 2008; CHMP, 2009). Consensus evaluations may be highly biased by non-independent evaluations. For example, senior or persuasive raters may affect the evaluations of junior or passive raters. In recent years, results from multiple raters have been summarized using an average of raters (“average rater”) or a majority of raters (“majority rater”) to estimate sensitivity and specificity. These methods have been used to evaluate the efficacy of diagnostic procedures for regulatory purposes (FDA, 2011; FDA, 2012). Kunz (2015) investigated the performance of these methods based on confidence intervals using simulation studies. However, there are few adequate methods for summarizing sensitivities and specificities that take into account the covariance between raters and the correlation between sensitivity and specificity.

Several statistical methods for meta-analysis of data from diagnostic studies have been proposed and recommended to generate a summary receiver operating characteristic (SROC) curve to represent the performance of a diagnostic procedure. The hierarchical summary receiver operating characteristic (HSROC) model is a method for meta-analysis of diagnostic studies. This model can account for variability between studies based on thresholds and accuracies (Rutter and Gatsonis, 2001). A bivariate random effects model (BVRM) is an alternative method for meta-analysis of diagnostic studies, which accounts not only for the between-study variability but also for the negative correlation between sensitivity and specificity because of the trade-off between the two measures as the test threshold varies (Reitsma et al., 2005). Harbord et al. (2007) showed that these two models are closely related and, in common situations, identical. Moreover, Jackson et al. (2010) proposed a method for estimating the variance of random effects in multivariate meta-analysis by extending the non-iterative method of moments suggested by DerSimonian and Laird (1986). These methods for meta-analysis of diagnostic studies can be used to estimate SROC, sensitivity, and specificity from independent diagnostic studies. However, multiple raters independently evaluate the results arising from the same subjects to whom a diagnostic procedure is applied; therefore, attention must be paid to correlations of sensitivities or specificities between raters and within subjects. Therefore, the methods for meta-analysis of diagnostic studies cannot be directly applied to the summarization of the sensitivities and specificities of a diagnostic procedure from multiple raters.

This article proposes methods for summarizing the diagnostic accuracies evaluated from multiple independent raters based on the BVRM by generalizing the methods for meta-analysis of diagnostic studies. The rest of this article is structured as follows. Section 2 describes the data structure and the proposed BVRM. Section 3 presents the details of the methods to summarize sensitivities and specificities, to construct the joint confidence regions, and to generate the SROC
curve evaluated from multiple raters based on the BVRM. In Section 4, biases and mean square errors (MSEs) of summarized sensitivities and specificities based on the BVRM are compared with those based on the average rater method and the majority rater method using Monte Carlo simulations. The coverage probabilities of confidence regions based on the BVRM, the average rater method, and the majority rater method are also compared. In Section 5, the methods are applied to data from florbetapir F 18 positron emission tomographic (PET) imaging study to predict the presence of \( \beta \)-amyloid in the brains of subjects with Alzheimer’s disease (AD) (Clark et al., 2011; FDA, 2012).

2. Data Structure and Models

2.1 Data structure

Consider a clinical investigation designed to evaluate a diagnostic procedure for diseased (\( D \)) and non-diseased (\( \bar{D} \)) subjects. The diagnostic procedure is performed on each diseased subject (\( i = 1, \ldots, n \)) and each non-diseased subject (\( j = 1, \ldots, m \)) and independently evaluated by each rater (\( r = 1, \ldots, R \)). Each rater assigns a value of 1 or 0 to each subject, where 1 evaluates the subject as “positive” (+) and 0 evaluates the subject as “negative” (−). Let the number of raters be \( R = 3 \). The observed probabilities and their corresponding observations are classified in a \( 2 \times 8 \) contingency table using the framework of Saeki and Tango (2011) (Table 1). For example, \( P_{a(++)} \) denotes the observed probability that raters 1 and 2 positively score a diseased subject, while rater 3 negatively scores the diseased subject, and \( x_{a(++)} \) is the corresponding observed frequency. Similarly, \( P_{b(---)} \) denotes the observed probability that raters 1 and 2 negatively score a non-diseased subject, while rater 3 positively scores the non-diseased subject, and \( x_{b(---)} \) is the corresponding observed frequency. We define the sensitivity (\( p_{ar} \)) and specificity (\( p_{br} \)) for rater \( r = 1 \) as follows:

\[
\begin{align*}
p_{a1} & = P( + \mid D, r = 1 ) = P_{a(++)} + P_{a(+-)} + P_{a(-+)} + P_{a(--)} , \\
p_{b1} & = P( - \mid \bar{D}, r = 1 ) = P_{b(---)} + P_{b(+-)} + P_{b(-+)} + P_{b(++)} .
\end{align*}
\]

The plugin estimators for the sensitivity and specificity for rater 1 are given by

\[
\begin{align*}
\hat{p}_{a1} & = \frac{ x_{a(++)} + x_{a(+-)} + x_{a(-+)} + x_{a(--)} }{ n } = x_{a1} / n , \\
\hat{p}_{b1} & = \frac{ x_{b(---)} + x_{b(+-)} + x_{b(-+)} + x_{b(++)} }{ m } = x_{b1} / m ,
\end{align*}
\]

where \( x_{a1} \) is the number of positive scores for the diseased subjects and \( x_{b1} \) is the number of negative scores for the non-diseased subjects given by rater 1. These definitions can be generalized to \( R \) (\( \geq 3 \)) raters in a straightforward manner. The resulting observed probabilities and their corresponding observations can be classified as a \( 2 \times 2^R \) contingency table similar to Table 1.

2.2 A bivariate random effects model

We propose a bivariate random effects model (BVRM) for summarizing the \( R \) sensitivities \( p_{ar} \) and the \( R \) specificities \( p_{br} \) based on results from \( R \) independent raters.

Jpn J Biomet Vol. 37, No. 1, 2016
We shall use an idea employed by Reitsma et al. (2005): let us define \( \mu_{ar} = \log\{p_{ar}/(1 - p_{ar})\} \) as the logit-transformed sensitivity of rater \( r \) and \( \mu_{br} = \log\{p_{br}/(1 - p_{br})\} \) as the logit-transformed specificity of rater \( r \). The logit transformation is used to transform the unit interval \((0,1)\) to continuous values in the range \((-\infty, \infty)\). We shall assume that the logit sensitivities \( \mu_{ar} \) of the multiple raters are normally distributed around some common mean value, \( \theta_a \), with a between-rater variance of \( \tau_a^2 \). A similar random effect assumption is adopted for the specificities of the raters, where \( \theta_b \) is used to denote the mean value of logit specificity and \( \tau_b^2 \) is the variance in logit specificity between raters. In addition, we incorporate the possibility of correlation between (logit) sensitivity and specificity within a rater. Combining the two normal models leads to the following bivariate normal model:

\[
\begin{pmatrix}
\mu_{ar} \\
\mu_{br}
\end{pmatrix}
\sim N
\begin{pmatrix}
\theta_a \\
\theta_b
\end{pmatrix},
\begin{pmatrix}
\tau_a^2 & \kappa\tau_a\tau_b \\
\kappa\tau_a\tau_b & \tau_b^2
\end{pmatrix},
\]

where \( \kappa \) is the correlation coefficient between \( \mu_{ar} \) and \( \mu_{br} \).

Furthermore, we extend this bivariate model by incorporating the precisions with which sensitivities and specificities have been measured by multiple raters. Here, we have to pay attention to correlations between raters working with identical subjects because multiple raters evaluate all the same subjects. If we treat the variances of estimates for the logit sensitivities \( \hat{\mu}_{ar} \) by multiple raters as fixed quantities, we can obtain the following:

\[
\begin{pmatrix}
\hat{\mu}_{a1} \\
\vdots \\
\hat{\mu}_{ar} \\
\hat{\mu}_{aR}
\end{pmatrix}
\sim N
\begin{pmatrix}
\mu_{ar} \\
\vdots \\
\mu_{ar} \\
\mu_{aR}
\end{pmatrix},
\begin{pmatrix}
\hat{\sigma}_{ar}^2 \\
\vdots \\
\hat{\sigma}_{ar}^2 \\
\hat{\sigma}_{aR}^2
\end{pmatrix},
\]

where \( \hat{\sigma}_{ar}^2 \) is the variance of \( \hat{\mu}_{ar} \) and \( \rho_{ar} \) is a correlation coefficient between \( \hat{\mu}_{ar} \) and \( \hat{\mu}_{ar'} \) (\( r \neq r' \)). Moreover, the variances of estimates for the logit specificities \( \hat{\mu}_{br} \) are modeled similarly. Here, we indicate that \( y_a = (\mu_{a1}, \cdots, \mu_{aR})^T \), \( y_b = (\mu_{b1}, \cdots, \mu_{bR})^T \), and \( \theta_a, \theta_b, \tau_{ar} \) are vectors of \( \theta_a \) and \( \theta_b \) in \( R \) dimensions, respectively, and

\[\text{Jpn J Biomet Vol. 37, No. 1, 2016}\]
Estimating the Diagnostic Accuracy from Multiple Raters

\[
A = \begin{pmatrix}
\sigma_{a1}^2 + \tau_a^2 & \rho_{a1r} \sigma_{a1} \sigma_{ar} & \ldots & \rho_{a1r} \sigma_{a1} \sigma_{aR} \\
\vdots & \ddots & \ddots & \vdots \\
\rho_{ar} \sigma_{ar} \sigma_{a1} & \ldots & \sigma_{ar}^2 + \tau_a^2 & \rho_{ar} \sigma_{ar} \sigma_{aR} \\
\ldots & \ldots & \ldots & \ldots \\
\rho_{a1r} \sigma_{a1} \sigma_{aR} & \ldots & \rho_{a1r} \sigma_{a1} \sigma_{aR} & \sigma_{aR}^2 + \tau_a^2
\end{pmatrix},
\]

\[
B = \begin{pmatrix}
\sigma_{b1}^2 + \tau_b^2 & \rho_{b1r} \sigma_{b1} \sigma_{br} & \ldots & \rho_{b1r} \sigma_{b1} \sigma_{bR} \\
\vdots & \ddots & \ddots & \vdots \\
\rho_{br} \sigma_{br} \sigma_{b1} & \ldots & \sigma_{br}^2 + \tau_b^2 & \rho_{br} \sigma_{br} \sigma_{bR} \\
\ldots & \ldots & \ldots & \ldots \\
\rho_{b1r} \sigma_{b1} \sigma_{bR} & \ldots & \rho_{b1r} \sigma_{b1} \sigma_{bR} & \sigma_{bR}^2 + \tau_b^2
\end{pmatrix},
\]

\[
T = \begin{pmatrix}
\kappa \tau_a \tau_b & \ldots & 0 & \ldots & 0 \\
\vdots & \ddots & \ddots & \vdots & \vdots \\
\kappa \tau_a \tau_b & \ldots & 0 & \kappa \tau_a \tau_b
\end{pmatrix},
\]

Then, taking all raters into account, the final model based on formulas (5) and (6) becomes as follows:

\[
\begin{pmatrix}
y_a \\
y_b
\end{pmatrix} \sim N\left(\begin{pmatrix}
\theta_{aR} \\
\theta_{bR}
\end{pmatrix}, \begin{pmatrix}
A & T \\
T & B
\end{pmatrix}\right).
\]

(7)

3. Estimation of the diagnostic accuracy

3.1 Summarizing sensitivity and specificity based on the BVRM

This section proposes an overall method for estimating sensitivities and specificities based on results from multiple raters using the BVRM described in Section 2.

The plugin estimates of the sensitivity \( \hat{\rho}_{ar} \) and the specificity \( \hat{\rho}_{br} \) for rater \( r \) are given by

\[
\hat{\rho}_{ar} = \frac{x_{ar}}{n}, \quad \hat{\rho}_{br} = \frac{x_{br}}{m},
\]

(8)

where \( x_{ar} \) and \( x_{br} \) are the numbers of positive and negative judgments in diseased and non-diseased subjects for rater \( r \), respectively [formulas (3) and (4)].

The estimators of the variances \( \sigma_{ar}^2 \) and \( \sigma_{br}^2 \) are denoted by \( \hat{s}_{ar}^2 \) and \( \hat{s}_{br}^2 \), respectively. These estimators are based on estimates of the logit sensitivity \( \hat{\mu}_{ar} \) and logit specificity \( \hat{\mu}_{br} \) for rater \( r \) and are defined as

\[
\hat{s}_{ar}^2 = 1/\{n\hat{\rho}_{ar}(1-\hat{\rho}_{ar})\}, \quad \hat{s}_{br}^2 = 1/\{m\hat{\rho}_{br}(1-\hat{\rho}_{br})\}.
\]

(9)

Furthermore, we derive the estimators of the covariances \( \hat{s}_{arr} \) and \( \hat{s}_{br} \) for \( \rho_{arr} \sigma_{ar} \sigma_{ar} \) and \( \rho_{br} \sigma_{br} \sigma_{br} \)
\[ \hat{\theta} = \left( X'\hat{\Sigma}^{-1}X \right)^{-1} X'\hat{\Sigma}^{-1}Y = \begin{pmatrix} \hat{\theta}_a \\ \hat{\theta}_b \end{pmatrix}, \]

where

\[ X = \begin{pmatrix} 1_R & 0_R \\ 0_R & 1_R \end{pmatrix}, \quad \hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_a & \hat{\Sigma}_{ab} \\ \hat{\Sigma}_{ba} & \hat{\Sigma}_b \end{pmatrix}, \quad Y = \begin{pmatrix} \hat{y}_a \\ \hat{y}_b \end{pmatrix}, \]

\[ \hat{\Sigma}_a = \begin{pmatrix} \hat{\sigma}_{a1}^2 + \hat{\tau}_a^2 & \cdots & \hat{\sigma}_{a1R} \\ \vdots & \ddots & \vdots \\ \hat{\sigma}_{aR1} & \cdots & \hat{\sigma}_{aRR} + \hat{\tau}_a^2 \end{pmatrix}, \]

\[ \hat{\Sigma}_b = \begin{pmatrix} \hat{\sigma}_{b1}^2 + \hat{\tau}_b^2 & \cdots & \hat{\sigma}_{b1R} \\ \vdots & \ddots & \vdots \\ \hat{\sigma}_{bR1} & \cdots & \hat{\sigma}_{bRR} + \hat{\tau}_b^2 \end{pmatrix}, \]

and \( 1_R \) and \( 0_R \) are vectors of ones and zeros in \( R \) dimensions, respectively, and \( \hat{\Sigma}_l \) is an \( R \times R \) diagonal matrix with diagonal elements \( \hat{\tau}_{al} \), which estimates \( \kappa \tau_a \tau_b \). Moreover, the covariance matrix of \( \hat{\theta} \) is estimated as follows (for details, see Appendix A2):

\[ \sqrt{\text{Var}(\hat{\theta})} = (X'\hat{\Sigma}^{-1}X)^{-1}. \]

Now, we must estimate the variances and covariance of the random effects, namely \( \hat{\sigma}_a^2, \hat{\tau}_b^2 \). 

Jpn J Biomet Vol. 37, No. 1, 2016
and $\kappa_\tau a_\tau b$. These estimators can be constructed using the method of moments as proposed by Jackson et al. (2010) (for details, see Appendix A3). Based on these estimators, the variance-covariance matrix $\Sigma_{DL}$, which expresses the variability between raters, and the correlation between a logit sensitivity and logit specificity can be written as follows:

$$\Sigma_{DL} = \begin{pmatrix} \hat{\tau}_a^2 & \hat{\tau}_{ab} \\ \hat{\tau}_{ab} & \hat{\tau}_b^2 \end{pmatrix}. \quad (13)$$

However, $\Sigma_{DL}$ may not be positive semi-definite. In such cases, we can use $\Sigma_{DL} +$ based on the spectrum decomposition, namely

$$\Sigma_{DL} + = \max(0, \lambda_1) e_1 e_1^t + \max(0, \lambda_2) e_2 e_2^t, \quad (14)$$

where $\lambda_1$ and $\lambda_2$ are the eigenvalues of $\Sigma_{DL}$, and $e_1$ and $e_2$ are the corresponding normalized eigenvectors, respectively. Finally, the summarized sensitivity $\hat{p}_a$ and specificity $\hat{p}_b$ are calculated using the inverse-logit transformation as follows:

$$\hat{p}_a = \frac{\exp(\hat{\theta}_a)}{1 + \exp(\hat{\theta}_a)}, \quad \hat{p}_b = \frac{\exp(\hat{\theta}_b)}{1 + \exp(\hat{\theta}_b)}. \quad (15)$$

### 3.2 Construction of a confidence region

As sensitivity and specificity may be highly correlated, separate confidence intervals for the summarized sensitivity ($p_a$) and specificity ($p_b$) may not be adequate. Hence, we suggest constructing a joint confidence region for the parameter vector based on the BVRM. Bantis et al. (2014) showed a confidence rectangle and an egg-shaped confidence region based on a confidence ellipse as joint confidence regions.

The $100(1 - \alpha)$% confidence rectangle is depicted on the basis of a $100(1 - \alpha/2)$% joint confidence interval for sensitivity and 1 – specificity in ROC space, that is $(1 - p_{\text{low}}^a, 1 - p_{\text{up}}^a) \times (p_{\text{low}}^b, p_{\text{up}}^b)$. The $100(1 - \alpha/2)$% joint confidence interval for the sensitivity is given by

$$p_{\text{low}}^a = \frac{\exp(\hat{\theta}_a - z_{(1-\alpha/4)} \sqrt{\{\text{Var}(\hat{\theta})\}_{11}})}{1 + \exp(\hat{\theta}_a - z_{(1-\alpha/4)} \sqrt{\{\text{Var}(\hat{\theta})\}_{11}}),} \quad (16)$$

where $z_{(1-\alpha/4)}$ is the upper $(\alpha/4)^{th}$ percentile of the standard normal distribution and $\{\text{Var}(\hat{\theta})\}_{11}$ denotes the $(1,1)$ element of the covariance matrix $\text{Var}(\hat{\theta})$. The $100(1 - \alpha/2)$% joint confidence interval for 1 – specificity is similarly calculated using a formula paralleling that of (16).

To account for the correlation between sensitivity and specificity, a $100(1 - \alpha)$% confidence ellipse can be created in the logit ROC space as follows:

$$(\hat{\theta} - \theta)^t \{\text{Var}(\hat{\theta})\}^{-1} (\hat{\theta} - \theta) = \chi^2_{1-\alpha}(2), \quad (17)$$
where \( \hat{\theta} \) is defined by formula (11), \( \hat{\text{Var}}(\hat{\theta}) \) is defined by formula (12), and \( \chi_2^2(1 - \alpha) \) is the upper 100(1 − \( \alpha \))% point of the \( \chi^2 \) distribution with two degrees of freedom. The confidence ellipse depicted in logit ROC space can be back-transformed to conventional ROC space to give an egg-shaped confidence region for \( p_a \) and \( p_b \).

3.3 SROC curve

Harbord et al. (2007) showed that BVRM proposed by Reitsuma et al. (2005) and hierarchical summary receiver operating characteristic (HSROC) model (Rutter and Gatsonis, 2001) for meta-analysis of diagnostic accuracy studies are very closely related and are in fact identical in common situations, particularly in the absence of study-level covariates; they are different parameterizations of the same model. We can reparametrize the BVRM to construct a logit SROC curve using the HSROC model by defining

\[
\hat{\theta}_a = \Lambda \exp(-\beta/2) - \exp(-\beta)\hat{\theta}_b,
\]

where \( \beta = \log(\hat{\tau}_b/\hat{\tau}_a) \), and \( \Lambda = (\hat{\tau}_b/\hat{\tau}_a)^{1/2}\hat{\theta}_a + (\hat{\tau}_a/\hat{\tau}_b)^{1/2}\hat{\theta}_b \). An SROC curve is derived by the back-transforming formula (18) to obtain

\[
p_a = \frac{\exp(\Lambda/\gamma) \{(1 - p_b)/p_b\}^{\gamma^2}}{1 + \exp(\Lambda/\gamma) \{(1 - p_b)/p_b\}^{\gamma^2}},
\]

where \( \gamma = \exp(\beta/2) \).

4. Simulation studies

The performances of the summarized sensitivity, specificity and joint confidence regions based on the BVRM were assessed using Monte Carlo simulation studies. Comparisons were made with the performances based on the average rater method and the majority rater method introduced by Kunz (2015) (Appendix A4). The pre-specified parameters were the number of raters (\( R = 3, 5, 7 \)), number of subjects (\( n = m = 25, 50, 100 \)), true sensitivity (\( p_a = 0.8, 0.9 \)), true specificity (\( p_b = 0.8, 0.9 \)), correlation coefficients between sensitivity and specificity in the BVRM (\( \kappa = 0, -0.3 \)) and correlation coefficients between raters (\( \rho_{arr} = \rho_{brr} = 0, 0.3, 0.6 \)). We generated an \( (n + m) \times 2R \) random matrix from a multivariate normal distribution with zero mean vector and the correlation matrix from \( \kappa, \rho_{arr} \) and \( \rho_{brr} \) using the R function "mvrnorm" in the package MASS. The upper-left \( n \times R \) sub-matrix between the elements (1, 1) and (n, R) of the random matrix contained the simulation data for the sensitivities from \( R \) raters, and the lower-right \( m \times R \) sub-matrix between the elements (n + 1, R + 1) and (n + m, 2R) of the random matrix contained the simulation data for the specificities from \( R \) raters. If each element of the matrix for the sensitivities exceeded \( z_{(1 - p_a)} \), the outcome of the element was set to one; otherwise, it was set to zero. Similarly, if each element of the matrix for the specificities exceeded \( z_{(1 - p_b)} \), the outcome of the element was set to one; otherwise, it was set to zero. Here, \( z_{(1 - p)} \) is the 100p percentile of the standard normal distribution. In our simulation study, 10,000 data sets were generated.
generated for each configuration. In addition, there were many singular matrices as \( \tilde{V} \) is affected by high correlations between raters within subjects, especially when the sample sizes \( n \) and \( m \) are small and the number of raters is 5 or 7. For this reason, we set the sample sizes 50 and 100 for 5 raters, and the sample size 100 for 7 raters in the simulation studies.

### 4.1 Bias and MSE

The biases and MSEs of the summarized sensitivities and specificities, as estimated by the average rater method, majority rater method, and BVRM, were evaluated using Monte Carlo simulations. Tables 2 and 3 present the biases and MSEs for \( p_a = p_b = 0.8 \) and \( p_a = p_b = 0.9 \), respectively. The biases of the average rater method were close to zero in all scenarios. The biases of the BVRM were larger than those of the average rater method; however, this difference decreased when the sample size increased to 50 and 100. On the other hand, the biases of the majority rater method were far from zero in all scenarios. MSEs of the average rater method and BVRM were also close to zero in contrast to those of the majority rater method in all scenarios.

### 4.2 Coverage probability

We evaluated the coverage probability of the joint confidence regions based on the BVRM in comparison with that of the confidence regions based on the average rater and majority rater methods. Table 4 and 5 show the empirical coverage probabilities of the confidence rectangles based on the average rater method, majority rater method and BVRM, and the confidence ellipse based on the BVRM. Both the confidence rectangle and the confidence ellipse based on the BVRM performed generally well, when the number of raters was three. However, the coverage probabilities of the confidence regions based on the BVRM decreased, when the number of raters was 5 or 7 and the correlations between raters within subjects were high. The coverage probabilities of the confidence rectangles based on the average rater method were close to the nominal level, when the sample sizes were large. However, the coverage probabilities of the confidence rectangles based on the average method were low, when the correlations between raters within subjects were high, the results being similar to those of the BVRM. On the other hand, the coverage probabilities of the confidence rectangles based on the majority rater method were significantly under the nominal level in almost all cases.

### 5. An example

Clark et al. (2011) reported a clinical investigation to determine if florbetapir F 18 positron emission tomographic (PET) imaging performed antemortem accurately predicts the presence of \( \beta \)-amyloid in the brain at autopsy. In this study, 152 subjects [age, mean (range), years: 78.1 (38–103)] were enrolled, and 35 of those subjects [age, mean (range), years: 79.3 (47–103)] were approaching the end of life. The 35 subjects who received PET imaging 12 months or less prior to death underwent postmortem evaluations (brain autopsies). Three nuclear medicine physicians independently evaluated the PET images. Florbetapir-PET images were visually assessed using...
Table 2. Biases and MSEs of the summarized sensitivities and specificities based on the average rater method (A), majority rater method (M), and BVRM (B) with $p_a = p_b = 0.8$ based on 10,000 replicates.

| $R$ | $n = m$ | $\kappa$ | $\rho_{\text{rez}} = \rho_{\text{rez}}$ | Bias | MSE |
|-----|---------|---------|----------------|-------|------|
|     |         |         |                 | A     | M    | B   | A     | M    | B   |
| 3   | 25      | 0       | 0               | Se    | -0.0003 | 0.0959 | -0.0035 | 0.0021 | 0.0130 | 0.0021 |
|     |         |         |                 | Sp    | -0.0010 | 0.0947 | -0.0042 | 0.0021 | 0.0127 | 0.0021 |
|     |         |         |                 | $0.3$ Se | -0.0004 | 0.0580 | -0.0067 | 0.0028 | 0.0083 | 0.0028 |
|     |         |         |                 | Sp    | -0.0001 | 0.0589 | -0.0061 | 0.0029 | 0.0084 | 0.0028 |
|     |         |         |                 | $0.6$ Se | -0.0034 | 0.0249 | -0.0138 | 0.0039 | 0.0067 | 0.0040 |
|     |         |         |                 | Sp    | -0.0038 | 0.0242 | -0.0136 | 0.0038 | 0.0066 | 0.0040 |
|     |         |         |                 | $-0.3$ Se | 0.0005 | 0.0963 | -0.0029 | 0.0021 | 0.0129 | 0.0021 |
|     |         |         |                 | Sp    | -0.0002 | 0.0959 | -0.0032 | 0.0021 | 0.0129 | 0.0021 |
|     |         |         |                 | $0.3$ Se | -0.0008 | 0.0584 | -0.0067 | 0.0028 | 0.0083 | 0.0029 |
|     |         |         |                 | Sp    | -0.0011 | 0.0579 | -0.0071 | 0.0029 | 0.0084 | 0.0029 |
|     |         |         |                 | $0.6$ Se | -0.0028 | 0.0258 | -0.0126 | 0.0038 | 0.0066 | 0.0039 |
|     |         |         |                 | Sp    | -0.0035 | 0.0253 | -0.0135 | 0.0039 | 0.0068 | 0.0039 |
|     |         |         |                 | $50$ Se | -0.0002 | 0.0959 | -0.0010 | 0.0011 | 0.0110 | 0.0011 |
|     |         |         |                 | Sp    | 0.0002  | 0.0964 | -0.0005 | 0.0011 | 0.0112 | 0.0011 |
|     |         |         |                 | $0.3$ Se | -0.0002 | 0.0587 | -0.0029 | 0.0014 | 0.0058 | 0.0014 |
|     |         |         |                 | Sp    | -0.0002 | 0.0587 | -0.0028 | 0.0014 | 0.0058 | 0.0014 |
|     |         |         |                 | $0.6$ Se | -0.0012 | 0.0288 | -0.0068 | 0.0018 | 0.0036 | 0.0018 |
|     |         |         |                 | Sp    | -0.0004 | 0.0290 | -0.0059 | 0.0019 | 0.0037 | 0.0018 |
|     |         |         |                 | $-0.3$ Se | 0.0003 | 0.0960 | -0.0005 | 0.0011 | 0.0111 | 0.0011 |
|     |         |         |                 | Sp    | -0.0004 | 0.0956 | -0.0012 | 0.0011 | 0.0110 | 0.0011 |
|     |         |         |                 | $0.3$ Se | -0.0001 | 0.0583 | -0.0027 | 0.0014 | 0.0058 | 0.0014 |
|     |         |         |                 | Sp    | -0.0001 | 0.0586 | -0.0028 | 0.0014 | 0.0059 | 0.0014 |
|     |         |         |                 | $0.6$ Se | -0.0009 | 0.0286 | -0.0063 | 0.0018 | 0.0036 | 0.0018 |
|     |         |         |                 | Sp    | -0.0005 | 0.0290 | -0.0059 | 0.0018 | 0.0037 | 0.0018 |
|     |         |         |                 | $100$ Se | 0.0003  | 0.0964 | -0.0001 | 0.0005 | 0.0102 | 0.0005 |
|     |         |         |                 | Sp    | -0.0005 | 0.0953 | -0.0008 | 0.0005 | 0.0100 | 0.0005 |
|     |         |         |                 | $0.3$ Se | 0.0001  | 0.0590 | -0.0011 | 0.0007 | 0.0047 | 0.0007 |
|     |         |         |                 | Sp    | 0.0001  | 0.0587 | -0.0010 | 0.0007 | 0.0047 | 0.0007 |
|     |         |         |                 | $0.6$ Se | -0.0001 | 0.0300 | -0.0026 | 0.0009 | 0.0023 | 0.0009 |
|     |         |         |                 | Sp    | -0.0008 | 0.0289 | -0.0034 | 0.0010 | 0.0023 | 0.0009 |
|     |         |         |                 | $-0.3$ Se | 0.0001  | 0.0961 | -0.0002 | 0.0005 | 0.0101 | 0.0005 |
|     |         |         |                 | Sp    | -0.0004 | 0.0953 | -0.0006 | 0.0005 | 0.0100 | 0.0005 |
|     |         |         |                 | $0.3$ Se | -0.0002 | 0.0585 | -0.0013 | 0.0007 | 0.0046 | 0.0007 |
|     |         |         |                 | Sp    | 0.0001  | 0.0588 | -0.0012 | 0.0007 | 0.0047 | 0.0007 |
|     |         |         |                 | $0.6$ Se | -0.0001 | 0.0298 | -0.0027 | 0.0009 | 0.0023 | 0.0009 |
|     |         |         |                 | Sp    | -0.0002 | 0.0298 | -0.0029 | 0.0009 | 0.0023 | 0.0009 |

Se: Sensitivity; Sp: Specificity.

a semi-quantitative score ranging from 0 (no amyloid) to 4 (high levels of cortical amyloid) and compared with the immunohistochemistry (IHC) measure of brain $\beta$-amyloid as the gold standard. Fortunately, we can use the results of the semi-quantitative score from the three raters in this study because florbetapir F 18 injection has been approved by the U.S. Food and Drug
Table 2. (continued) Biases and MSEs of the summarized sensitivities and specificities based on the average rater method (A), majority rater method (M), and BVRM (B) with $p_a = p_b = 0.8$ based on 10,000 replicates.

| R | n = m | $\kappa$ | $\rho_{arr'} = \rho_{brr'}$ | Bias | MSE |
|---|---|---|---|---|---|
| 5 | 50 | 0 | 0 | Se -0.0001 0.1425 -0.0020 0.0006 0.0214 0.0006 | Sp -0.0001 0.1422 -0.0021 0.0007 0.0213 0.0006 |
| 0.3 | 0 | Se -0.0003 0.0835 -0.0051 0.0010 0.0091 0.0011 | Sp -0.0007 0.0836 -0.0056 0.0011 0.0091 0.0011 |
| 0.6 | 0 | Se -0.0025 0.0389 -0.0097 0.0016 0.0043 0.0017 | Sp -0.0026 0.0380 -0.0099 0.0016 0.0043 0.0017 |
| -0.3 | 0 | Se -0.0004 0.1424 -0.0024 0.0006 0.0214 0.0007 | Sp 0.0003 0.1421 -0.0018 0.0006 0.0213 0.0007 |
| 0.3 | 0 | Se -0.0002 0.0838 -0.0050 0.0010 0.0091 0.0011 | Sp -0.0002 0.0833 -0.0050 0.0011 0.0090 0.0011 |
| 0.6 | 0 | Se -0.0026 0.0398 -0.0098 0.0015 0.0043 0.0017 | Sp -0.0027 0.0384 -0.0099 0.0016 0.0043 0.0017 |
| 0 | 0 | Se 0.0002 0.1422 -0.0007 0.0003 0.0208 0.0003 | Sp -0.0002 0.1419 -0.0011 0.0003 0.0207 0.0003 |
| 0.3 | 0 | Se -0.0001 0.0829 -0.0027 0.0005 0.0079 0.0005 | Sp -0.0002 0.0827 -0.0028 0.0005 0.0079 0.0005 |
| 0.6 | 0 | Se -0.0004 0.0399 -0.0050 0.0008 0.0029 0.0008 | Sp -0.0007 0.0397 -0.0054 0.0008 0.0029 0.0008 |
| -0.3 | 0 | Se 0.0006 0.1422 -0.0003 0.0003 0.0208 0.0003 | Sp -0.0002 0.1417 -0.0011 0.0003 0.0206 0.0003 |
| 0.3 | 0 | Se -0.0002 0.0826 -0.0029 0.0005 0.0078 0.0005 | Sp -0.0001 0.0827 -0.0028 0.0005 0.0079 0.0005 |
| 0.6 | 0 | Se 0.0003 0.0409 -0.0044 0.0008 0.0031 0.0008 | Sp -0.0005 0.0397 -0.0052 0.0008 0.0030 0.0008 |
| 7 | 100 | 0 | 0 | Se 0.0001 0.1667 -0.0012 0.0002 0.0281 0.0002 | Sp -0.0001 0.1665 -0.0013 0.0002 0.0281 0.0002 |
| 0.3 | 0 | Se -0.0001 0.0963 -0.0034 0.0004 0.0102 0.0005 | Sp -0.0003 0.0962 -0.0035 0.0004 0.0102 0.0005 |
| 0.6 | 0 | Se -0.0016 0.0451 -0.0065 0.0008 0.0034 0.0008 | Sp -0.0003 0.0462 -0.0053 0.0007 0.0035 0.0008 |
| -0.3 | 0 | Se 0.0001 0.1666 -0.0012 0.0002 0.0281 0.0002 | Sp -0.0003 0.1662 -0.0015 0.0002 0.0279 0.0002 |
| 0.3 | 0 | Se -0.0001 0.0962 -0.0033 0.0005 0.0102 0.0005 | Sp -0.0005 0.0960 -0.0038 0.0005 0.0102 0.0005 |
| 0.6 | 0 | Se -0.0016 0.0441 -0.0066 0.0007 0.0033 0.0008 | Sp -0.0009 0.0454 -0.0058 0.0007 0.0034 0.0008 |

Se: Sensitivity; Sp: Specificity.

Administration (FDA) as a radioactive diagnostic agent for PET imaging of the brain to estimate $\beta$-amyloid neuritic plaque density in adult patients with cognitive impairment who are being evaluated for Alzheimer’s disease (AD) and other causes of cognitive decline. Information about the agent has been disclosed on the FDA website (2012). Here, we calculate the summarized
Table 3. Biases and MSEs of the summarized sensitivities and specificities based on the average rater method (A), majority rater method (M), and BVRM (B) with \( p_a = p_b = 0.9 \) based on 10,000 replicates.

| \( R \) | \( n = m \) | \( \kappa \) | \( \rho_{\text{her}} = \rho_{\text{her}}' \) | Bias | MSE |
|---|---|---|---|---|---|
| | | | A | M | B |
| 3 | 25 | 0 | 0 | Se | 0.0006 | 0.0725 | -0.0139 | 0.0012 | 0.0063 | 0.0013 |
| | | | 0.3 | Se | 0.0013 | 0.0511 | -0.0154 | 0.0015 | 0.0045 | 0.0016 |
| | | | 0.6 | Se | 0.0038 | 0.0308 | -0.0166 | 0.0021 | 0.0039 | 0.0022 |
| | | | -0.3 | Se | -0.0002 | 0.0718 | -0.0147 | 0.0012 | 0.0063 | 0.0013 |
| | | | 0.3 | Se | 0.0014 | 0.0513 | -0.0153 | 0.0015 | 0.0045 | 0.0016 |
| | | | 0.6 | Se | 0.0044 | 0.0316 | -0.0159 | 0.0021 | 0.0039 | 0.0022 |
| | | | -0.3 | Se | -0.0002 | 0.0725 | -0.0146 | 0.0012 | 0.0063 | 0.0013 |
| | | | 0.3 | Se | 0.0011 | 0.0511 | -0.0153 | 0.0015 | 0.0046 | 0.0016 |
| | | | 0.6 | Se | 0.0038 | 0.0307 | -0.0164 | 0.0021 | 0.0039 | 0.0022 |
| | | | 50 | 0 | Se | 0.0001 | 0.0723 | -0.0022 | 0.0006 | 0.0058 | 0.0006 |
| | | | 0.3 | Se | -0.0007 | 0.0488 | -0.0045 | 0.0008 | 0.0034 | 0.0008 |
| | | | 0.6 | Se | -0.0012 | 0.0259 | -0.0076 | 0.0010 | 0.0021 | 0.0011 |
| | | | -0.3 | Se | -0.0002 | 0.0717 | -0.0024 | 0.0006 | 0.0057 | 0.0006 |
| | | | 0.3 | Se | -0.0004 | 0.0489 | -0.0043 | 0.0008 | 0.0034 | 0.0008 |
| | | | 0.6 | Se | -0.0013 | 0.0259 | -0.0075 | 0.0010 | 0.0021 | 0.0011 |
| | | | -0.3 | Se | -0.0001 | 0.0721 | -0.0003 | 0.0006 | 0.0055 | 0.0005 |
| | | | 0.3 | Se | -0.0005 | 0.0718 | -0.0009 | 0.0003 | 0.0054 | 0.0003 |
| | | | 0.6 | Se | -0.0001 | 0.0489 | -0.0014 | 0.0004 | 0.0029 | 0.0004 |
| | | | -0.3 | Se | -0.0002 | 0.0723 | -0.0003 | 0.0003 | 0.0055 | 0.0003 |
| | | | 0.3 | Se | -0.0002 | 0.0484 | -0.0017 | 0.0004 | 0.0028 | 0.0004 |
| | | | 0.6 | Se | -0.0006 | 0.0260 | -0.0038 | 0.0005 | 0.0014 | 0.0005 |
| | | | -0.3 | Se | -0.0004 | 0.0266 | -0.0037 | 0.0005 | 0.0014 | 0.0005 |

Se: Sensitivity; Sp: Specificity.

sensitivity and specificity of the florbetapir-PET data from the 35 subjects suspected of having AD. We dichotomize the results using the definitions of FDA (2012), grouping scores from 2 to 4 as “positive” and scores from 0 to 1 as “negative.” Moreover, diagnosis based on results from the IHC measure is used as the gold standard test; when IHC is equal to or more than 1%, the presence of a \( \beta \)-amyloid burden is concluded; when IHC is less than 1%, the absence...
Table 3. (continued) Biases and MSEs of the summarized sensitivities and specificities based on the average rater method (A), majority rater method (M), and BVRM (B) with $p_a = p_b = 0.9$ based on 10,000 replicates.

| $R$ | $n = m$ | $\kappa$ | $\rho_{\text{arr}} = \rho_{\text{brr}}$ | Bias | MSE |
|-----|---------|---------|-------------------|-------|-------|
| 5   | 50      | 0       | 0                 | A     | M     | B     | A     | M     | B     |
|     |         |         |                   | -0.0005 | 0.0917 | -0.0039 | 0.0003 | 0.0086 | 0.0004 |
|     |         |         |                   | 0.0006 | 0.0919 | -0.0039 | 0.0004 | 0.0086 | 0.0004 |
|     |         |         |                   | 0.0010 | 0.0667 | -0.0062 | 0.0005 | 0.0051 | 0.0006 |
|     |         |         |                   | 0.0012 | 0.0663 | -0.0064 | 0.0005 | 0.0051 | 0.0006 |
|     |         |         |                   | 0.0024 | 0.0374 | -0.0096 | 0.0008 | 0.0027 | 0.0010 |
|     |         |         |                   | 0.0020 | 0.0376 | -0.0096 | 0.0009 | 0.0028 | 0.0010 |
|     |         |         |                   | -0.0007 | 0.0918 | -0.0039 | 0.0003 | 0.0086 | 0.0004 |
|     |         |         |                   | -0.0008 | 0.0917 | -0.0040 | 0.0004 | 0.0086 | 0.0004 |
|     |         |         |                   | -0.0011 | 0.0666 | -0.0063 | 0.0005 | 0.0051 | 0.0006 |
|     |         |         |                   | -0.0010 | 0.0669 | -0.0060 | 0.0005 | 0.0051 | 0.0006 |
|     |         |         |                   | -0.0015 | 0.0383 | -0.0086 | 0.0008 | 0.0028 | 0.0010 |
|     |         |         |                   | -0.0027 | 0.0369 | -0.0097 | 0.0009 | 0.0027 | 0.0010 |
|     | 100     | 0       | 0                 | 0.0001 | 0.0915 | -0.0013 | 0.0002 | 0.0085 | 0.0002 |
|     |         |         |                   | 0.0003 | 0.0915 | -0.0016 | 0.0002 | 0.0084 | 0.0002 |
|     |         |         |                   | 0.0001 | 0.0651 | -0.0031 | 0.0003 | 0.0046 | 0.0003 |
|     |         |         |                   | 0.0002 | 0.0649 | -0.0033 | 0.0003 | 0.0046 | 0.0003 |
|     |         |         |                   | 0.0011 | 0.0351 | -0.0059 | 0.0004 | 0.0019 | 0.0005 |
|     |         |         |                   | 0.0012 | 0.0347 | -0.0059 | 0.0004 | 0.0018 | 0.0005 |
|     |         |         |                   | 0.0002 | 0.0915 | -0.0012 | 0.0002 | 0.0084 | 0.0002 |
|     |         |         |                   | 0.0002 | 0.0914 | -0.0016 | 0.0002 | 0.0084 | 0.0002 |
|     |         |         |                   | 0.0003 | 0.0647 | -0.0033 | 0.0003 | 0.0045 | 0.0003 |
|     |         |         |                   | 0.0003 | 0.0650 | -0.0033 | 0.0003 | 0.0046 | 0.0003 |
|     |         |         |                   | 0.0005 | 0.0355 | -0.0053 | 0.0004 | 0.0019 | 0.0005 |
|     |         |         |                   | 0.0010 | 0.0352 | -0.0058 | 0.0004 | 0.0019 | 0.0005 |
|     | 7       | 100     | 0                 | 0.0001 | 0.0973 | -0.0018 | 0.0001 | 0.0095 | 0.0001 |
|     |         |         |                   | 0.0001 | 0.0973 | -0.0018 | 0.0001 | 0.0095 | 0.0001 |
|     |         |         |                   | -0.0003 | 0.0730 | -0.0039 | 0.0002 | 0.0056 | 0.0003 |
|     |         |         |                   | -0.0005 | 0.0734 | -0.0041 | 0.0002 | 0.0056 | 0.0003 |
|     |         |         |                   | -0.0008 | 0.0416 | -0.0055 | 0.0004 | 0.0023 | 0.0005 |
|     |         |         |                   | -0.0008 | 0.0413 | -0.0055 | 0.0004 | 0.0023 | 0.0005 |
|     |         |         |                   | -0.0004 | 0.0733 | -0.0039 | 0.0002 | 0.0056 | 0.0003 |
|     |         |         |                   | -0.0002 | 0.0734 | -0.0037 | 0.0002 | 0.0057 | 0.0003 |
|     |         |         |                   | -0.0014 | 0.0410 | -0.0062 | 0.0004 | 0.0023 | 0.0005 |
|     |         |         |                   | -0.0007 | 0.0413 | -0.0055 | 0.0004 | 0.0023 | 0.0005 |

Se: Sensitivity; Sp: Specificity.

of a β-amyloid burden is concluded. We generated a 2×8 contingency table from the individual data to show the relationship between the florbetapir-PET and the gold standard (Table 6). The estimated sensitivities and specificities by the three raters are $[0.900 \ (18/20), 1 \ (15/15)]$, $[0.550 \ (11/20), 1 \ (15/15)]$ and $[0.850 \ (17/20), 0.800 \ (12/15)]$, respectively. Correlation coefficients
Table 4. Coverage probabilities of the 95% joint confidence regions based on the average rater method (A), majority rater method (M), and BVRM (B) with $p_a = p_b = 0.8$ based on 10,000 replicates.

| $R$ | $n = m$ | $\kappa$ | $\rho_{arr'} = \rho_{brr'}$ | Coverage Prob. (%) | Rectangle | Ellipse |
|-----|---------|----------|---------------------|-------------------|-----------|---------|
|     |         |          |                     | Rectangle         | A         | M       | B       |
| 3   | 25      | 0        | 0                   | 92.12             | 56.51     | 96.94   | 96.46   |
|     |         |          | 0.3                 | 90.91             | 78.02     | 96.14   | 95.37   |
|     |         |          | 0.6                 | 89.35             | 87.93     | 94.58   | 94.24   |
|     |         |          | -0.3                | 92.25             | 56.66     | 97.00   | 96.69   |
|     |         |          | 0.3                 | 90.76             | 77.93     | 95.98   | 95.41   |
|     |         |          | 0.6                 | 89.47             | 87.67     | 94.74   | 94.24   |
| 50  | 0       | 0        | 0                   | 93.28             | 17.87     | 96.92   | 96.39   |
|     |         |          | 0.3                 | 93.76             | 52.76     | 96.25   | 96.08   |
|     |         |          | 0.6                 | 92.50             | 76.67     | 94.85   | 94.74   |
|     |         |          | -0.3                | 93.52             | 17.69     | 96.71   | 96.73   |
|     |         |          | 0.3                 | 93.39             | 53.27     | 96.21   | 96.12   |
|     |         |          | 0.6                 | 93.11             | 76.56     | 95.14   | 95.25   |
| 100 | 0       | 0        | 0                   | 94.25             | 5.57      | 96.62   | 96.39   |
|     |         |          | 0.3                 | 93.63             | 43.86     | 95.95   | 95.75   |
|     |         |          | 0.6                 | 93.25             | 79.59     | 95.07   | 94.94   |
|     |         |          | -0.3                | 94.69             | 5.86      | 97.39   | 96.92   |
|     |         |          | 0.3                 | 93.96             | 44.64     | 96.23   | 96.03   |
|     |         |          | 0.6                 | 93.80             | 78.55     | 95.51   | 95.35   |
| 5   | 50      | 0        | 0                   | 93.62             | 0.39      | 95.97   | 95.87   |
|     |         |          | 0.3                 | 93.20             | 28.28     | 93.92   | 93.47   |
|     |         |          | 0.6                 | 92.82             | 70.23     | 92.80   | 92.35   |
|     |         |          | -0.3                | 94.02             | 0.47      | 96.40   | 95.93   |
|     |         |          | 0.3                 | 92.86             | 28.42     | 94.51   | 94.28   |
|     |         |          | 0.6                 | 93.11             | 68.81     | 93.35   | 93.50   |
| 100 | 0       | 0        | 0                   | 94.36             | 0.02      | 96.30   | 96.11   |
|     |         |          | 0.3                 | 93.79             | 15.01     | 94.93   | 94.88   |
|     |         |          | 0.6                 | 93.60             | 69.02     | 94.18   | 94.02   |
|     |         |          | -0.3                | 94.27             | 0.00      | 96.79   | 96.55   |
|     |         |          | 0.3                 | 94.10             | 15.50     | 94.76   | 94.79   |
|     |         |          | 0.6                 | 93.05             | 67.62     | 94.01   | 94.02   |
| 7   | 100     | 0        | 0                   | 94.60             | 0.00      | 96.12   | 96.07   |
|     |         |          | 0.3                 | 94.18             | 5.49      | 93.96   | 93.68   |
|     |         |          | 0.6                 | 93.99             | 61.64     | 93.27   | 93.18   |
|     |         |          | -0.3                | 94.62             | 0.00      | 96.18   | 95.81   |
|     |         |          | 0.3                 | 94.49             | 5.38      | 93.91   | 93.63   |
|     |         |          | 0.6                 | 93.81             | 62.04     | 93.48   | 93.29   |

Rectangle: confidence rectangle; Ellipse: confidence ellipse.

between the raters within subjects with $\beta$-amyloid are $\hat{\rho}_{a12} = 0.37$, $\hat{\rho}_{a13} = 0.79$, and $\hat{\rho}_{a23} = 0.46$, respectively. On the other hand, correlation coefficients between the raters within subjects without $\beta$-amyloid are not be able to estimate because results from the rater1 and 2 are all negative.
Table 5. Coverage probabilities of the 95% joint confidence regions based on the average rater method (A), majority rater method (M), and BVRM (B) with \( p_a = p_b = 0.9 \) based on 10,000 replicates.

| \( R \) | \( n = m \) | \( \kappa \) | \( \rho_{arr'} = \rho_{brr'} \) | \( \text{Coverage Prob. (%)} \) | \( \text{Rectangle} \) | \( \text{Ellipse} \) |
|---|---|---|---|---|---|---|
| 3 | 25 | 0 | 0 | 0 | 89.77 | 25.56 | 95.64 | 95.97 |
| | | | | | 85.77 | 50.10 | 95.21 | 95.40 |
| | | | | | 79.96 | 65.92 | 93.80 | 94.05 |
| | | | | | 0.6 | 90.14 | 25.92 | 95.76 | 95.77 |
| | | | | | 0.3 | 85.99 | 49.63 | 95.42 | 95.44 |
| | | | | | 0.6 | 78.85 | 64.84 | 93.90 | 94.14 |
| | 50 | 0 | 0 | 0 | 93.27 | 16.93 | 97.15 | 96.85 |
| | | | | | 0.3 | 90.89 | 53.65 | 96.26 | 95.99 |
| | | | | | 0.6 | 87.66 | 77.63 | 94.35 | 93.81 |
| | | | | | -0.3 | 92.97 | 16.90 | 97.12 | 96.53 |
| | | | | | 0.3 | 90.91 | 52.80 | 96.11 | 95.81 |
| | | | | | 0.6 | 88.00 | 77.99 | 94.34 | 93.92 |
| | 100 | 0 | 0 | 0 | 93.53 | 0.45 | 97.10 | 96.81 |
| | | | | | 0.3 | 93.09 | 16.19 | 96.30 | 96.07 |
| | | | | | 0.6 | 91.67 | 56.51 | 94.75 | 94.42 |
| | | | | | -0.3 | 94.14 | 0.32 | 97.27 | 97.17 |
| | | | | | 0.3 | 93.06 | 17.10 | 96.24 | 95.71 |
| | | | | | 0.6 | 92.12 | 56.77 | 94.43 | 94.23 |
| 5 | 50 | 0 | 0 | 0 | 93.54 | 0.49 | 95.87 | 95.84 |
| | | | | | 0.3 | 91.65 | 25.29 | 92.90 | 92.62 |
| | | | | | 0.6 | 88.70 | 65.31 | 91.39 | 90.81 |
| | | | | | -0.3 | 93.32 | 0.45 | 95.78 | 95.66 |
| | | | | | 0.3 | 91.59 | 24.58 | 93.32 | 92.94 |
| | | | | | 0.6 | 88.53 | 65.80 | 90.96 | 90.13 |
| | 100 | 0 | 0 | 0 | 94.36 | 0.00 | 96.39 | 96.32 |
| | | | | | 0.3 | 92.53 | 1.72 | 93.86 | 93.98 |
| | | | | | 0.6 | 92.15 | 40.87 | 92.09 | 91.79 |
| | | | | | -0.3 | 93.82 | 0.00 | 96.45 | 96.26 |
| | | | | | 0.3 | 93.78 | 1.92 | 94.53 | 94.41 |
| | | | | | 0.6 | 91.02 | 38.44 | 92.47 | 92.07 |
| 7 | 100 | 0 | 0 | 0 | 94.03 | 0.00 | 95.69 | 95.57 |
| | | | | | 0.3 | 93.22 | 0.27 | 92.47 | 92.47 |
| | | | | | 0.6 | 90.19 | 28.99 | 90.34 | 90.14 |
| | | | | | -0.3 | 94.21 | 0.00 | 95.85 | 96.02 |
| | | | | | 0.3 | 93.08 | 0.32 | 92.56 | 92.09 |
| | | | | | 0.6 | 91.24 | 29.27 | 91.74 | 91.39 |

Rectangle: confidence rectangle; Ellipse: confidence ellipse.

As the empirical specificities from the raters 1 and 2 are 1, we add 1/8 to each cell in Table 6 to compute the values based on the BVRM. Observed logit sensitivities and logit specificities are \( \hat{\mu}_{a1} = 2.001, \hat{\mu}_{a2} = 0.191, \hat{\mu}_{a3} = 1.609 \) (Shapiro-Wilk test: \( p = 0.396 \)), and \( \hat{\mu}_{b1} = 3.434, \hat{\mu}_{b2} = 3.434, \hat{\mu}_{b3} = 1.273 \) (Shapiro-Wilk test: \( p < 0.001 \)), respectively. The correlation coefficient Jpn J Biomet Vol. 37, No. 1, 2016
Table 6. Diagnostic results by three nuclear medicine physicians for detection of β-amyloid in the brain using florbetapir F 18 PET imaging (Clark et al., 2011; FDA, 2012).

| β-amyloid burden | Judgment of (Rater1, Rater2, Rater3) |
|------------------|-------------------------------------|
| Presence         | (+ + +) (++) (+ +) (+) (− +) (+) (− −) (+) (− −) (−) (− −) Total |
|                  | 11 0 6 0 1 0 0 2 0 20               |
| Absence          | 0 0 0 0 0 0 0 3 12 15               |

between logit sensitivities and logit specificities is \( \hat{\kappa} = -0.311 \). Although the normality of the distribution of the logit specificities are not supported by Shapiro-Wilk test, we compute the summarized sensitivities, specificities, and joint confidence regions based on the BVRM, average rater, and majority rater methods to be \([0.749 (0.411, 0.927), 0.857 (0.568, 0.965)], [0.767 (0.625, 0.909), 0.933 (0.874, 0.993)], and [0.850 (0.666, 1), 1 (1, 1)], respectively. In addition, the SROC curves (a straight line in logit ROC space), 95% confidence rectangles, confidence ellipse, and egg-shaped confidence region based on the BVRM are shown in Figure 1.

6. Discussion

When we attempt to summarize sensitivities and specificities evaluated from multiple raters, we must pay attention not only to negative correlation between sensitivity and specificity but also to correlations between the raters within subjects when multiple raters independently evaluate the same results. However, a statistical method for meta-analysis of diagnostic studies based on BVRM proposed by Reitsma et al. (2005) cannot account for correlations between raters within subjects. Hence, we have extended the method for meta-analysis based on BVRM by incorporating such correlations. In addition, Harbord et al. (2007) showed that methods for meta-analysis
based on BVRM proposed by Reitsma et al. (2005) and the HSROC model proposed by Rutter and Gatsonis (2001) are closely related and, in common situations, identical. The SROC curve is the most common approach for representing the performance of a diagnostic procedure. We generated the SROC curve using estimators calculated by our proposed method based on the BVRM. Moreover, as sensitivity and specificity may be highly correlated, separate confidence intervals for estimates of sensitivity and specificity may not be adequate. Therefore, we proposed a method for constructing a joint confidence region for sensitivity and specificity based on the BVRM.

In Section 4, we evaluated the performance of the summarized sensitivities, specificities, and joint confidence regions based on the BVRM, average rater, and majority rater methods in Monte Carlo simulation studies. The summarized sensitivity and specificity based on the average rater method were shown to have biases closer to zero compared with those based on the BVRM or majority rater method in all scenarios. However, when the sample size was large \((n = m = 50, 100)\), the biases of the summarized sensitivity and specificity based on the BVRM were very small. MSEs of the summarized sensitivity and specificity based on the average rater method and BVRM were close to each other. On the other hand, the biases and MSEs of estimators based on the majority rater method were larger than those based on the average rater method or the BVRM. The coverage probabilities of the joint confidence regions based on the BVRM were close to the nominal level, representing an improvement over those based on the average rater or majority rater methods, when the number of raters was three. This is due to the fact that the BVRM properly takes into consideration of the relevant variances and covariances. Especially, the BVRM indicated good performances even if the correlation between sensitivity and specificity \(\kappa\), and those between raters within subjects \(\rho_{arr}^{r}\) and \(\rho_{brr}^{r}\) were equal to zero. However, the confidence regions based on the BVRM may be affected by multicollinearity when the number of raters is large and the correlations between raters within subjects are high. Therefore, we recommend the summarized sensitivity and specificity based on the BVRM for estimating the diagnostic accuracy from multiple raters when there are adequate numbers of raters and subjects. In addition, the use of the summarized sensitivity and specificity based on the majority rater method is not recommended on the basis of the results from the simulation studies.

In Section 5, a real example showed that the summarized sensitivity and specificity based on the BVRM were lower than those based on the average rater method. On the other hand, the summarized sensitivity and specificity based on the majority rater method were considerably different from those based on other methods. These results suggest that the data from multiple raters in clinical investigations designed to demonstrate the efficacy of a diagnostic procedure should adequately deal with the structure of the data from multiple raters. Furthermore, depictions of SROC and the joint confidence region may expedite understanding by and com-
communication with clinicians rather than the traditional approach of describing the sensitivity and specificity of each rater individually.

Our proposed method based on the BVRM requires the addition of $1/2R$ to each cell if the plugin estimators $\hat{p}_{ar}$ or $\hat{p}_{br}$ are zero or one. This will not affect the results too strongly if the sample size is sufficiently large. Finally, although the focus of the paper is on methodologies for summarizing the information from multiple raters, direct modeling of individual subject data from multiple raters may also sometimes merit consideration. In addition, we note that for a situation without a gold standard, our proposed method is not adequate to summarize sensitivities and specificities of a diagnostic procedure. These topics are left for future research.

Acknowledgements

The authors thank the editor and anonymous referees their constructive comments that substantially improved the quality of the article. This study was partially supported by JSPS KAKENHI Grant Number 25330034.

REFERENCES

Bantis, L. E., Nakas, C. T. and Reiser, B. (2014). Construction of confidence regions in the ROC space after the estimation of the optimal Youden index-based cut-off point. *Biometrics* **70**, 212–223.

CHMP. (2009). Appendix 1 to the guideline on clinical evaluation of diagnostic agents (CPMP/EWP/1119/98 REV. 1) on imaging agents (Doc. Ref. EMEA/CHMP/EWP/321180/2008). Available from: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003581.pdf [18 February 2016].

Clark, C. M., Schneider, J. A., Bedell, B. J., Beach, T. G., Bilker, W. B., Mintun, M. A., Pontecorvo, M. J., Hefti, F., Carpenter, A. P., Ffletter, M. L., Krautkramer, M. J., Kung, H. F., Coleman, R. E., Doraisswamy, P. M., Fleisher, A. S., Sabbagh, M. N., Sadowsky, C. H., Reiman, E. M., Zehntner, S. P. and Skovronsky, D. M. (2011). Use of florbetapir-PET for imaging beta-amyloid pathology. *JAMA* **305**, 275–283.

DerSimonian, R. and Laird, N. (1986). Meta-Analysis in Clinical Trials. *Controlled Clinical Trials* **7**, 177–188.

FDA. (2004). Guidance for industry. Developing medical imaging drug and biological products. Part 3: design, analysis, and interpretation of clinical studies. Available from: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm071604.pdf [18 February 2016].

FDA. (2011). U.S. Food and Drug Administration, Center for Drug Evaluation and Research. Gadovist NDA 201277Orig1s000 MEDICAL REVIEW(S). Available from: Jpn J Biomet Vol. 37, No. 1, 2016
FDA. (2012). U.S. Food and Drug Administration, Center for Drug Evaluation and Research. Amyvid NDA 202008Orig1s000 MEDICAL REVIEW(S). Available from: http://www.accessdata.fda.gov/drugsatfda_docs/nda/2012/202008Orig1s000MedR.pdf [18 February 2016].

Harbord, R. M., Deeks, J. J., Egger, M., Whiting, P. and Sterne, J. A. C. (2007). A unification of models for meta-analysis of diagnostic accuracy studies. Biostatistics 8, 239–251.

Jackson, D., White, I. R. and Thompson S. G. (2010). Extending DerSimonian and Laird’s methodology to perform multivariate random effects meta-analyses. Statistics in Medicine 29, 1282–1297.

Kunz, M. (2015). A comparison of analysis procedures for correlated binary data in dedicated multi-rater imaging trials. Pharmaceutical Statistics 14, 34–43.

Lehr, R. G. and Kashanian, F. K. (2009). Three persistent issues in analysis of clinical trials involving diagnostic contrast agents. Drug Information Journal 43, 525–532.

Obuchowski, N. A. and Lieber, M. L. (2008). Statistics and methodology. Skeletal Radiology 37, 393–396.

Reitsma, J. B., Glas, A. S., Rutjes, A. W. S., Scholten, R. J. P. M., Bossuyt, P. M. and Zwinderman, A. H. (2005). Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. Journal of Clinical Epidemiology 58, 982–990.

Rutter, C. M. and Gatsonis, C. A. (2001). A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Statistics in Medicine 20, 2865–2884.

Saeki, H. and Tango, T. (2011). Non-inferiority test and confidence interval for the difference in correlated proportions in diagnostic procedures based on multiple raters. Statistics in Medicine 30, 3313–3327.

Schwenke, C. and Busse, R. (2007). Analysis of differences in proportions from clustered data with multiple measurements in diagnostic studies. Methods of Information in Medicine 46, 548–552.

Appendix

A1. Estimation of the covariance matrix of the logit-transformed proportions using the multivariate delta method

The estimators of the variances and covariance components, \( \hat{s}_{ar}^2, \hat{s}_{br}^2, \hat{s}_{arr} \) and \( \hat{s}_{brr} (r \neq r') \), in Section 3 are constructed from the covariance matrix of the logit-transformed proportions using the following multivariate delta method. For ease of explanation, we consider probability \( p \) when
\[ r = 1 \text{ and } r' = 2, \text{ based on Table 1. The binomial variances of } \hat{p}_1 \text{ and } \hat{p}_2 \text{ are given by} \]
\[ \text{Var}(\hat{p}_1) = \frac{p_1(1 - p_1)}{n}, \quad \text{Var}(\hat{p}_2) = \frac{p_2(1 - p_2)}{n}. \]

Moreover, we derive the covariance of \( \hat{p}_1 \) and \( \hat{p}_2 \) as follows:
\[
\text{Cov}(\hat{p}_1, \hat{p}_2) = E[(\hat{p}_1 - p_1)(\hat{p}_2 - p_2)] \\
= E[\{(\hat{q}_{++} + \hat{q}_{-+}) - (q_{++} + q_{-+})\} \{(\hat{q}_{++} + \hat{q}_{-+}) - (q_{++} + q_{-+})\}] \\
= \text{Var}(\hat{q}_{++}) + \text{Cov}(\hat{q}_{++}, \hat{q}_{-+}) + \text{Cov}(\hat{q}_{++}, \hat{q}_{-+}) + \text{Cov}(\hat{q}_{++}, \hat{q}_{-+}) \\
= \frac{1}{n}(q_{++}q_{-+} - q_{++}q_{-+}),
\]

where, \( q_{++} = P_{++} + P_{-+}, q_{-+} = P_{+-} + P_{+}, q_{-+} = P_{+} + P_{-} \) and \( q_{-+} = 1 - q_{++} - q_{-+} - q_{-+} \). Therefore, we derive the covariance matrix of the logit-transformed proportion using the multivariate delta method as follows:
\[
\text{Var}(h(\hat{p})) \approx \nabla h(p)'^t \cdot \text{Cov}(\hat{p}) \cdot \nabla h(p) \\
= \left( \begin{array}{cc}
1/\{np_1(1 - p_1)\} & (q_{++}q_{-+} - q_{++}q_{-+})/\{np_1p_2(1 - p_1)(1 - p_2)\} \\
1/\{np_2(1 - p_2)\} & 1/\{np_2(1 - p_2)\}
\end{array} \right),
\]

where \( \hat{p} = (\hat{p}_1, \hat{p}_2)' \), \( p = (p_1, p_2)' \), \( h(\cdot) = \logit(\cdot) \), and
\[
\nabla h(p) = \left( \begin{array}{cc}
1/\{p_1(1 - p_1)\} & 0 \\
1/\{p_2(1 - p_2)\} & 1/\{p_2(1 - p_2)\}
\end{array} \right), \\
\text{Cov}(\hat{p}) = \left( \begin{array}{cc}
\text{Var}(\hat{p}_1) & \text{Cov}(\hat{p}_1, \hat{p}_2) \\
\text{Cov}(\hat{p}_1, \hat{p}_2) & \text{Var}(\hat{p}_2)
\end{array} \right) \\
= \left( \begin{array}{cc}
p_1(1 - p_1)/n & (q_{++}q_{-+} - q_{++}q_{-+})/n \\
p_2(1 - p_2)/n & p_2(1 - p_2)/n
\end{array} \right).
\]

**A2. Estimation of the covariance matrix of the mean logit sensitivity and logit specificity**

The covariance matrix of the estimated mean logit sensitivity and logit specificity based on the BVRM in Section 3 is calculated as follows:
\[
\widehat{\text{Var}}(\hat{\theta}) = \text{Var}\{(X'\hat{\Gamma}^{-1}X)^{-1}(X'\hat{\Gamma}^{-1}Y)\} \\
= C^{-1} \text{Var}(X'\hat{\Gamma}^{-1}Y)(C^{-1})'^t \\
= C^{-1}(X'\hat{\Gamma}^{-1}) \text{Var}(Y)(X'\hat{\Gamma}^{-1})'^t(C^{-1})' \\
= C^{-1}(X'\hat{\Gamma}^{-1})(\hat{\Gamma}^{-1})'Y(C^{-1})' \\
= C^{-1}C'Y(C^{-1})'X(C^{-1})' \\
= C^{-1}C'(C^{-1})'X(C^{-1})^{-1}. 
\]

Jpn J Biomet Vol. 37, No. 1, 2016
where \( C = X^t \hat{V}^{-1} X \).

**A3. Estimation of the variances and covariances of random effects**

The estimators of the variances and covariance of the random effects, \( \hat{\tau}_a^2 \), \( \hat{\tau}_b^2 \), and \( \kappa \hat{\tau}_a \hat{\tau}_b \), can be calculated using the method of moments proposed by Jackson et al. (2010). These estimators take the following forms,

\[
\hat{\tau}_a^2 = \frac{Q_a - (R - 1)}{\sum_{r=1}^{R} W_{ar} - \sum_{r=1}^{R} \frac{W_{ar}^2}{\sum_{r=1}^{R} W_{ar}}}, \\
\hat{\tau}_b^2 = \frac{Q_b - (R - 1)}{\sum_{r=1}^{R} W_{br} - \sum_{r=1}^{R} \frac{W_{br}^2}{\sum_{r=1}^{R} W_{br}}}, \\
\hat{\tau}_{ab} = \frac{Q_{ab}}{\sum_{r=1}^{R} \sqrt{W_{ar}W_{br}} - \sum_{r=1}^{R} \frac{W_{ar}W_{br}}{\sum_{r=1}^{R} \sqrt{W_{ar}W_{br}}}},
\]

where the matrix \( W \) and other quantities are given by

\[
W = X^t K^{-1}, \quad W_{(1,r)} = w_{ar}, \quad W_{(2,(R+r))} = w_{br},
\]

\[
K = \begin{pmatrix} S_a^* & O_R \\ O_R & S_b^* \end{pmatrix},
\]

\[
S_a^* = \begin{pmatrix} \hat{s}_{a1}^2 & \cdots & \hat{s}_{a1R} \\ \cdots & \cdots & \cdots \\ \hat{s}_{ar}^2 & \cdots & \hat{s}_{arR} \\ \cdots & \cdots & \cdots \\ \hat{s}_{aR}^2 & \cdots & \hat{s}_{aR} 
\end{pmatrix}, \quad S_b^* = \begin{pmatrix} \hat{s}_{b1}^2 & \cdots & \hat{s}_{b1R} \\ \cdots & \cdots & \cdots \\ \hat{s}_{br}^2 & \cdots & \hat{s}_{brR} \\ \cdots & \cdots & \cdots \\ \hat{s}_{bR}^2 & \cdots & \hat{s}_{bR} 
\end{pmatrix},
\]

\[
\hat{\theta}_{a(F)} = \frac{\sum_{r=1}^{R} w_{ar} \hat{\mu}_{ar}}{\sum_{r=1}^{R} w_{ar}}, \quad \hat{\theta}_{b(F)} = \frac{\sum_{r=1}^{R} w_{br} \hat{\mu}_{br}}{\sum_{r=1}^{R} w_{br}},
\]

\[
\hat{\theta}_{a(C)} = \frac{\sum_{r=1}^{R} \sqrt{w_{ar}w_{br}} \hat{\mu}_{ar}}{\sum_{r=1}^{R} \sqrt{w_{ar}w_{br}}}, \quad \hat{\theta}_{b(C)} = \frac{\sum_{r=1}^{R} \sqrt{w_{ar}w_{br}} \hat{\mu}_{br}}{\sum_{r=1}^{R} \sqrt{w_{ar}w_{br}}},
\]

\[
Q_a = \sum_{r=1}^{R} w_{ar}(\hat{\mu}_{ar} - \hat{\theta}_{a(F)})^2, \quad Q_b = \sum_{r=1}^{R} w_{br}(\hat{\mu}_{br} - \hat{\theta}_{b(F)})^2,
\]

\[
Q_{ab} = \sum_{r=1}^{R} \sqrt{w_{ar}w_{br}}(\hat{\mu}_{ar} - \hat{\theta}_{a(C)})(\hat{\mu}_{br} - \hat{\theta}_{b(C)}).
\]

**A4. Calculation of the summary statistics and the confidence intervals investigated by Kunz (2015)**

Kunz (2015) investigated methods of summarization, namely the average rater and majority rater methods for proportions from multiple raters. Because there is no distinction between sensitivity and specificity, we shall use notations of sensitivity here. Each rater assigns subjects as either “positive” or “negative.” The random variable \( u_{ir} \) gives the assessment of subject \( i \) (\( i = 1, \cdots, n \)) for rater \( r \) (\( r = 1, \cdots, R \)), which is defined as follows:

\( \text{Jpn J Biomet Vol. 37, No. 1, 2016} \)
\[ u_{ir} = \begin{cases} 1 & \text{the assessment of rater } r \text{ for subject } i \text{ is "positive"} \\ 0 & \text{otherwise.} \end{cases} \]

The estimator, variance, and covariance based on the average rater method can be calculated using the method proposed by Schwenke and Busse (2007) as follows:

\[
\hat{A}_r = \frac{\sum_{i=1}^{n} u_{ir}}{n}, \\
\hat{A}_R = \frac{\sum_{r=1}^{R} \hat{A}_r}{R}, \\
\text{Var}( \hat{A}_R ) = \frac{1}{R^2} \left\{ \sum_{r=1}^{R} \text{Var}( \hat{A}_r ) + 2 \sum_{r=1, r<r'}^{R} \text{Cov}( \hat{A}_r, \hat{A}_{r'} ) \right\},
\]

where,

\[
\text{Var}( \hat{A}_r ) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (u_{ir} - \hat{A}_r)^2, \\
\text{Cov}( \hat{A}_r, \hat{A}_{r'} ) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (u_{ir} - \hat{A}_r)(u_{ir'} - \hat{A}_{r'}). 
\]

The estimator and variance based on the majority rater method can be calculated as follows:

\[
\hat{M}_R = \text{median}(u_{1R}, \cdots, u_{nR}), \\
\hat{M}_R = \frac{1}{n} \sum_{i=1}^{n} M_{iR}, \\
\text{Var}( \hat{M}_R ) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (\hat{M}_R - \hat{M}_R)^2.
\]

The confidence interval is obtained using a simple normal approximation:

\[
\hat{a}_{\text{low}} = \hat{a} - z_{(1-\alpha/2)} \sqrt{\text{Var}(\hat{a})}, \\
\hat{a}_{\text{up}} = \hat{a} + z_{(1-\alpha/2)} \sqrt{\text{Var}(\hat{a})}. \tag{A.1}
\]

Plugging into formula (A.1), e.g., for \( \hat{a} \), the average rater method \( \hat{A}_R \) yields the lower and upper bounds of an asymptotic two-sided 100(\( \alpha/2 \))% confidence interval for \( \hat{A}_R \). In a similar manner, the two-sided confidence interval can be obtained for \( \hat{M}_R \).