Published in final edited form as:
Comput Stat Data Anal. 2013 December ; 68: . doi:10.1016/j.csda.2013.07.007.

Estimating confidence intervals for the difference in diagnostic accuracy with three ordinal diagnostic categories without a gold standard

Le Kang[a], **Chengjie Xiong**[b], and **Lili Tian**[c,*]

[a]Center for Devices and Radiological Health, US Food and Drug Administration, Silver Spring, MD 20993, United States

[b]Division of Biostatistics, Washington University in St. Louis, St. Louis, MO 63110, United States

[c]Department of Biostatistics, University at Buffalo, Buffalo, NY 14214, United States

Abstract

With three ordinal diagnostic categories, the most commonly used measures for the overall diagnostic accuracy are the volume under the ROC surface (VUS) and partial volume under the ROC surface (PVUS), which are the extensions of the area under the ROC curve (AUC) and partial area under the ROC curve (PAUC), respectively. A gold standard (GS) test on the true disease status is required to estimate the VUS and PVUS. However, oftentimes it may be difficult, inappropriate, or impossible to have a GS because of misclassification error, risk to the subjects or ethical concerns. Therefore, in many medical research studies, the true disease status may remain unobservable. Under the normality assumption, a maximum likelihood (ML) based approach using the expectation–maximization (EM) algorithm for parameter estimation is proposed. Three methods using the concepts of generalized pivot and parametric/nonparametric bootstrap for confidence interval estimation of the difference in paired VUSs and PVUSs without a GS are compared. The coverage probabilities of the investigated approaches are numerically studied. The proposed approaches are then applied to a real data set of 118 subjects from a cohort study in early stage Alzheimer's disease (AD) from the Washington University Knight Alzheimer's Disease Research Center to compare the overall diagnostic accuracy of early stage AD between two different pairs of neuropsychological tests.

Keywords

EM algorithm; Generalized pivot; Gold standard; Parametric bootstrap; Volume under the ROC surface

1. Introduction

Diagnostic testing is an extremely important aspect of medical care. Medical diagnosis involves the classification of patients into two or more categories. These categories may imply the presence or absence of a particular medical condition. Signs, symptoms or clinical tests are used to determine the classification. The evaluation of a diagnostic test procedure involves the estimation of parameters that describe the accuracy of diagnostic test relative to

[*]Correspondence to: Department of Biostatistics, University at Buffalo, 706 Kimball Tower, 3435 Main Street, Buffalo, NY 14214-3000, United States. Tel.: +1 716 829 2715; fax: +1 716 829 2200. ltian@buffalo.edu (L. Tian).

true classification and it is of paramount importance to compare the accuracies of diagnostic tests to decide on the best test for certain disease. For instance, one of the common indices used for overall diagnostic accuracy on the case when subjects are categorized in a binary fashion, i.e., non-diseased and diseased, is AUC (Zhou et al., 2002; Pepe, 2003; Shapiro, 1999). The comparison of the overall diagnostic accuracy between two diagnostic tests is frequently addressed by measuring the difference in the paired AUCs.

In practice, the diagnostic decision is not limited to a binary choice in many situations. For example, a clinical assessment, NPZ-8, of the presence of HIV-related cognitive dysfunction (AIDS Dementia Complex—ADC) would discriminate between patients exhibiting clinical symptoms of ADC (combined stages 1–3), subjects exhibiting minor neurological symptoms (ADC stage 0.5) and neurologically unimpaired individuals (ADC stage 0) (Nakas and Yiannoutsos, 2004). Another example provided by Xiong et al. (2006) concerns mild cognitive impairment (MCI) or early stage Alzheimer's disease (AD) being a transitional stage between the cognitive changes from normal aging and the more severe problems caused by the AD. Thereafter, we refer to the disease status between "non-diseased" and "diseased" as "intermediate", in other words, transitional status.

Given that an independent gold standard (GS) test on the disease status is available, Scurfield (1996) and Xiong et al. (2006) extended binary statistical tools such as the ROC curve and AUC and developed the volume and partial volume under the ROC surface (VUS and PVUS) to summarize the diagnostic accuracy with three ordinal diagnostic categories. Furthermore, Nakas and Yiannoutsos (2004) proposed a nonparametric estimation of a single VUS; Xiong et al. (2007) proposed a large sample approach for comparing several VUSs for normally distributed data. Most recently, Tian et al. (2011) addressed exact confidence interval estimation for the difference in paired VUSs and PVUSs based on the concepts of generalized pivot and showed that their approach generally can provide confidence intervals with reasonable coverage probabilities even at small sample sizes.

Notice that all the aforementioned methods assume the existence of a GS test. In other words, the true disease category is known. For instance, in the diagnosis of early stage AD (Xiong et al., 2006, 2007), the dementia severity of Alzheimer type was staged by the Clinical Dementia Rating (CDR) according to published rules (Morris, 1993), which is considered as a "GS" for evaluating different neuropsychological tests and biomarkers for early stage AD. The resulting diagnosis by clinical assessments such as CDR, although expected to be quite accurate, presumably was not totally free of misclassification errors and thus was not perfect. Such misclassifications are known to produce bias in estimating the diagnostic accuracy of disease markers, e.g., VUS. Further, such bias may prove to be detrimental when it comes to compare the diagnostic accuracy of multiple disease markers. It is therefore important to develop valid statistical methods of diagnostic comparison that do not rely on the existence of a perfect GS.

Some works involving estimating diagnostic accuracy without a GS have been done for binary diagnostic tests. For example, Henkelman et al. (1990) considered the estimation of ROC curves of continuous-scale tests in the absence of a GS test; Beiden et al. (2000) proposed maximum likelihood (ML) estimates of the ROC curves using the EM algorithm; Hsieh et al. (2009) proposed a ML based procedure for construction of confidence intervals for the difference in paired AUCs without a GS; Zhou et al. (2005) also developed a nonparametric ML method for estimating ROC curves in the absence of a GS test.

In this paper, we will focus on interval estimation for the difference in paired VUSs and PVUSs with three ordinal diagnostic categories without a GS by proposing a ML based approach using the EM algorithm in conjunction with the generalized variable approach as well as the parametric/nonparametric bootstrap methods. This paper is organized as follows. We first introduce some preliminaries about VUS and PVUS in Section 2. In Section 3, we will present the proposed methods. The performance of the proposed approaches including their robustness will be assessed by a numerical study in Section 4. In Section 5, our proposed methods will be applied to a real world study of very early stage AD diagnosis. We close with a broader discussion for evaluating diagnostic tests without a GS.

2. Preliminaries

The ROC surface, analogous to the ROC curve, has been proposed to assess the accuracy of tests with three ordinal diagnostic categories. Let Y_1, Y_2 and Y_3 denote the scores resulting from a diagnostic test and let F_1, F_2 and F_3 be the corresponding cumulative distribution functions for non-diseased, intermediate and diseased subjects, respectively. Assume the results of a diagnostic test are measured on continuous scale and higher values indicate greater severity of the disease. Given a pair of threshold values c_1 and $c_3(c_1 < c_3)$, let $\delta_1 = F_1(c_1)$, $\delta_3 = 1 - F_3(c_3)$ be the true classification rates for non-diseased and diseased categories, respectively. Then the probability that a randomly selected subject from an intermediate category has a score between c_1 and c_3 is

$$\delta_2 = F_2(c_3) - F_2(c_1) = F_2[F_3^{-1}(1-\delta_3)] - F_2[F_1^{-1}(\delta_1)]. \quad (1)$$

The triplet $(\delta_1,\delta_2,\delta_3)$, where $\delta_2 = \delta_2(\delta_1,\delta_3)$ is a function of (δ_1,δ_3), would produce an ROC surface in the three-dimensional space for all possible $(c_1, c_3) \in \mathbb{R}^2$. As the ROC curve for a binary diagnosis represents the trade-off between sensitivity and specificity, which are correct classification probabilities for the two categories (non-diseased and diseased), the ROC surface represents the three-way trade-off among the correct classification probabilities for the three categories.

In order to summarize the overall diagnostic accuracy for the diagnostic test, the volume under the ROC surface (VUS) has been considered. It is defined as

$$\text{VUS} = \int_0^1 \int_0^{1-F_3[F_1^{-1}(\delta_1)]} F_2[F_3^{-1}(1-\delta_3)] - F_2[F_1^{-1}(\delta_1)] d\delta_3 d\delta_1. \quad (2)$$

This is a generalization of the AUC for a ROC curve under a binary classification. One could show that VUS is mathematically equivalent to the probability $P(Y_1 < Y_2 < Y_3)$ when Y_1, Y_2 and Y_3 are randomly selected from each diagnostic category, respectively. For a useless test (e.g., when Y_1, Y_2 and Y_3 have identical distributions), VUS is 1 /6. Similar to the PAUC of a ROC curve in which investigators are only interested in a certain lower range of false positive rate, the partial volume under the ROC surface (PVUS) has also been considered to measure the diagnostic accuracy with pre-specified minimum classification rates for the non-diseased and diseased subjects,

$$\text{PVUS} = \int\int_{\mathscr{R}} \left\{ F_2[F_3^{-1}(1-\delta_3)] - F_2[F_1^{-1}(\delta_1)] \right\} d\delta_3 d\delta_1, \quad (3)$$

where $\Re = \left\{ (\delta_1, \delta_3) | \delta_{10} \le \delta_1 \le 1, \delta_{30} \le \delta_3 \le 1 - F_3[F_1^{-1}(\delta_1)] \right\}$ with δ_{10} and δ_{30} being the desired minimum classification rates for non-diseased and diseased categories, respectively. When non-diseased, intermediate and diseased categories can be discriminated perfectly, PVUS reaches its maximum value $\text{PVUS}_{\text{max}} = (1 - \delta_{10})(1 - \delta_{30})$. The better the discriminating ability of the diagnostic test, the closer the value of PVUS to PVUS_{max}. Note that PVUS = VUS if $\delta_{10} = \delta_{30} = 0$.

We now use 1, 2, and 3 to represent the non-diseased, intermediate and diseased categories, respectively. Consider the case with two diagnostic tests *A* and *B*. Let Y_{kA} and Y_{kB} stand for the measurements for a randomly selected subject from the *k*th (*k* = 1, 2, 3) disease category for test *A* and test *B*, respectively. Assume that $(Y_{kA}, Y_{kB})'$ follow a bivariate normal distribution, i.e.,

$$\boldsymbol{Y}_k=\begin{pmatrix} Y_{kA} \\ Y_{kB} \end{pmatrix} \sim N_2(\mu_k, \boldsymbol{\Sigma}_k), k=1,2,3, \quad (4)$$

where

$$\mu_k=\begin{pmatrix} \mu_{kA} \\ \mu_{kB} \end{pmatrix}, \boldsymbol{\Sigma}_k=\begin{pmatrix} \sigma_{kA}^2 & \sigma_{kAB} \\ \sigma_{kAB} & \sigma_{kB}^2 \end{pmatrix}. \quad (5)$$

As in Xiong et al. (2006), the VUS and PVUS for diagnostic test *A* under the above setting can be further expressed as

$$\text{VUS}_A=\int_{-\infty}^{\infty}\phi(a_A s-b_A)\phi(-c_A s+d_A)\Phi(s)ds, \quad (6)$$

$$\text{PVUS}_A=\int_{[\Phi^{-1}(\delta_{10})+b_A]/a_A}^{[d_A-\Phi^{-1}(\delta_{30})]/c_A}[\Phi(a_A s-b_A)\Phi(-c_A s+d_A)-\delta_{10}\Phi(-c_A s+d_A)-\delta_{30}\Phi(a_A s-b_A)+\delta_{10}\delta_{30}]\phi(s)ds, \quad (7)$$

where a_A = σ_{2A}t/σ_{1A}, b_A = (μ_{1A} – μ_{2A}) / σ_{1A}, c_A = σ_{2A}/σ_{3A}, d_A = (μ_{3A} – μ_{2A}) /σ_{3A}, Φ(•) is the standard normal distribution function, and Φ(•) is the standard normal density function. By replacing *A* with *B* in Eqs. (6) and Eqs. (7), we can obtain VUS_B and PVUS_B. For test *A*, the ML estimates of VUS_A and PVUS_A can be obtained by substituting μ_{kA}, σ_{kA} (*k* = 1, 2, 3) in Eqs. (6) and Eqs. (7) with the corresponding sample mean $\bar{y_{kA}}$, and sample standard deviation s_{kA}. Similarly, the ML estimates of VUS_B and PVUS_B can be obtained. To compare the diagnostic accuracy between test *A* and test *B*, ΔVUS and ΔPVUS, defined as follows

$$\Delta\text{VUS}=\text{VUS}_A-\text{VUS}_B, \quad (8)$$

$$\Delta\text{PVUS}=\text{PVUS}_A-\text{PVUS}_B,$$

have to be estimated.

When a GS is available, the ML estimates of ΔVUS and ΔPVUS can be easily obtained. With a GS, the large sample test for comparing several VUSs by Xiong et al. (2007) can be extended to confidence interval estimation; and Tian et al. (2011) addressed the exact confidence interval estimation of ΔVUS and/or ΔPVUS based on the concepts of generalized pivot. Tian et al. (2011) compared the generalized variable (GV) approach with a parametric bootstrap approach and the large sample approach (Xiong et al., 2007) and showed the GV approach usually can provide confidence intervals with better coverage probabilities. Besides the aforementioned methods, another approach for this purpose is to use nonparametric bootstrap resampling method to estimate the corresponding variances of ΔVUS or ΔPVUS instead of using the large sample delta method as in Xiong et al. (2007).

3. The proposed methods

Our goal is to develop interval estimation for ΔVUS and ΔPVUS without a GS. The following proposed approach is based on the EM algorithm in conjunction with the generalized variable (GV) approach as well as parametric bootstrap (PB) and nonparametric bootstrap (NB) methods. We refer to them as EM-GV, EM-PB and approach, respectively.

Let $D = k$ ($k = 1, 2, 3$) indicate the unobserved true disease category for the non-diseased, intermediate and diseased subjects, respectively. We denote the test results of A and B on a non-diseased, intermediate and diseased individual by $Y_{kA}$and Y_{kB} ($k = 1, 2, 3$), respectively. Following Eqs. (4) and Eqs. (5), the vector of unknown parameters in this setting is given by

$$\theta' = (p_1, p_3, \mu_{kA}, \mu_{kB}, \sigma^2_{kA}, \sigma^2_{kB}, \sigma_{kAB}),$$

where $p_1 = P(D = 1)$, $p_3 = P(D = 3)$ denoting the prevalence of non-diseased and diseased populations. Under this model, the conditional independence structure between diagnostic tests given disease status is a special case with $\sigma_{kAB} = 0$ ($k = 1, 2, 3$).

When a GS is not available, we propose to estimate θ using the EM algorithm. After the convergent value of θ is obtained via the EM algorithm, the ML estimates of VUSs and PVUSs, will be obtained by plugging in the ML estimate of θ. Finally, the ML estimate of the difference in paired VUSs and PVUSs, will be obtained. A similar approach has been used by Hsieh et al. (2009) to estimate the difference in paired AUCs without a GS.

3.1. EM algorithm

Let t_{ij} be the observed result of the test j, $j = A, B$ on the ith individual, D_i be the unobserved true disease category associated with ith individual, and $p_1 = P(D_i = 1)$, $p_3 = P(D_i = 3)$. It is easy to see that $p_2 = P(D_i = 2) = 1 – p_1 - p_3$. Let $\boldsymbol{t}_i = (t_{iA}, t_{iB})$, $\boldsymbol{t} = (\boldsymbol{t}_1, \boldsymbol{t}_2, \dots, \boldsymbol{t}_n)$ and $\boldsymbol{D} = (D_1, D_2, \dots, D_n)$.

If $\boldsymbol{D}$ has been observed, the complete data likelihood function would be given as follows,

$$\mathscr{L}(\theta|\boldsymbol{t}, \boldsymbol{D})\prod_{i=1}^{n}[p_1 f_{\boldsymbol{Y}_1}(\boldsymbol{t}_i)]^{I(D_i=1)}[p_2 f_{\boldsymbol{Y}_2}(\boldsymbol{t}_i)]^{I(D_i=2)}[p_3 f_{\boldsymbol{Y}_3}(\boldsymbol{t}_i)]^{I(D_i=3)}. \quad (10)$$

The complete data log-likelihood function would be

$$\ell(\theta|\boldsymbol{t}, \boldsymbol{D}) = \sum_{i=1}^{n}\left\{I(D_i=1)\log[p_1 f_{\boldsymbol{Y}_1}(\boldsymbol{t}_i)] + I(D_i=2)\log[1-p_1-p_3)f_{\boldsymbol{Y}_2}(\boldsymbol{t}_i)] + I(D_i=3)\log[p_3 f_{\boldsymbol{Y}_3}(\boldsymbol{t}_i)]\right\}, \quad (11)$$

where $f_{\boldsymbol{Y}_1}(\boldsymbol{t}_i)$, $f_{\boldsymbol{Y}_2}(\boldsymbol{t}_i)$, $f_{\boldsymbol{Y}_3}(\boldsymbol{t}_i)$ are the density functions of non-diseased, intermediate and diseased category, respectively.

Let $\theta^{(m)}$ denote the estimate of θ after m iteration of EM algorithm. The following E-step and M-step are used to find $\theta^{(m+1)}$, an updated estimate of θ.

E-step: The E-step computes the conditional expectation of the complete data log-likelihood function with observed data $\boldsymbol{t}$ and the current parameter estimate $\theta^{(m)}$,

$$\begin{aligned} Q=&E\ell(\theta|\boldsymbol{t},\theta^{(m)})\\ =&\sum_{i=1}^{n}\left\{P(D_i=1|\boldsymbol{t}_i,\theta^{(m)})\log[p_1 f_{\boldsymbol{Y}_1}(\boldsymbol{t}_i)]+P(D_i=2|\boldsymbol{t}_i,\theta^{(m)})\log[(1-p_1-p_3)f_{\boldsymbol{Y}_2}(\boldsymbol{t}_i)]+P(D_i=3|\boldsymbol{t}_i,\theta^{(m)})\log[p_3 f_{\boldsymbol{Y}_3}(\boldsymbol{t}_i)]\right\}. \end{aligned} \quad (12)$$

Define $q_{i1}^{(m)}=P(D_i=1|\boldsymbol{t}_i\theta^{(m)}), q_{i2}^{(m)}=P(D_i=2|\boldsymbol{t}_i,\theta^{(m)}), q_{i3}^{(m)}=P(D_i=3|\boldsymbol{t}_i,\theta^{(m)})$ one can show that

$$q_{i1}^{(m)}=\frac{p_1^{(m)}f_{\boldsymbol{Y}_1}^{(m)}(\boldsymbol{t}_i)}{p_1^{(m)}f_{\boldsymbol{Y}_1}^{(m)}(\boldsymbol{t}_i)+\left(1-p_1^{(m)}-p_3^{(m)}\right)f_{\boldsymbol{Y}_2}^{(m)}(\boldsymbol{t}_i)+p_3^{(m)}f_{\boldsymbol{Y}_3}^{(m)}(\boldsymbol{t}_i)}, \quad (13)$$

$$q_{i2}^{(m)}=\frac{\left(1-p_1^{(m)}-p_3^{(m)}\right)f_{\boldsymbol{Y}_2}^{(m)}(\boldsymbol{t}_i)}{p_1^{(m)}f_{\boldsymbol{Y}_1}^{(m)}(\boldsymbol{t}_i)+\left(1-p_1^{(m)}-p_3^{(m)}\right)f_{\boldsymbol{Y}_2}^{(m)}(\boldsymbol{t}_i)+p_3^{(m)}f_{\boldsymbol{Y}_3}^{(m)}(\boldsymbol{t}_i)}, \quad (14)$$

$$q_{i3}^{(m)}=\frac{p_3^{(m)}f_{\boldsymbol{Y}_3}^{(m)}(\boldsymbol{t}_i)}{p_1^{(m)}f_{\boldsymbol{Y}_1}^{(m)}(\boldsymbol{t}_i)+\left(1-p_1^{(m)}-p_3^{(m)}\right)f_{\boldsymbol{Y}_2}^{(m)}(\boldsymbol{t}_i)+p_3^{(m)}f_{\boldsymbol{Y}_3}^{(m)}(\boldsymbol{t}_i)}. \quad (15)$$

Therefore,

$$Q=\sum_{i=1}^{n}\left\{q_{i1}^{(m)}\log[p_1 f_{\boldsymbol{Y}_1}(\boldsymbol{t}_i)]+q_{i2}^{(m)}\log[(1-p_1-p_3)f_{\boldsymbol{Y}_2}(\mathbf{t}_i)]+q_{i3}^{(m)}\log[p_3 f_{\boldsymbol{Y}_3}(\mathbf{t}_i)]\right\}. \quad (16)$$

M-step: The *M*-step finds the updated estimate $\theta^{(m+1)}$ by maximizing Q with respect to θ. For instance, setting

$$\begin{cases} \frac{\partial Q}{\partial p_1}=\sum_{i=1}^{n}\left\{q_{i1}^{(m)}\frac{1}{p_1}-q_{i2}^{(m)}\frac{1}{1-p_1-p_3}\right\}\equiv 0\\ \frac{\partial Q}{\partial p_3}=\sum_{i=1}^{n}\left\{q_{i3}^{(m)}\frac{1}{p_3}-q_{i2}^{(m)}\frac{1}{1-p_1-p_3}\right\}\equiv 0 \end{cases} \quad (17)$$

would give us

$$\begin{cases} \frac{1}{p_1}\sum_{i=1}^{n}q_{i1}^{(m)}=\frac{1}{1-p_1-p_3}\sum_{i=1}^{n}\left(1-q_{i1}^{(m)}-q_{i3}^{(m)}\right)\\ \frac{1}{p_3}\sum_{i=1}^{n}q_{i3}^{(m)}=\frac{1}{1-p_1-p_3}\sum_{i=1}^{n}\left(1-q_{i1}^{(m)}-q_{i3}^{(m)}\right), \end{cases} \quad (18)$$

and thus

$$\begin{cases} \widehat{p}_1^{(m+1)}=\frac{1}{n}\sum_{i=1}^{n}q_{i1}^{(m)}\\ \widehat{p}_3^{(m+1)}=\frac{1}{n}\sum_{i=1}^{n}q_{i3}^{(m)}. \end{cases} \quad (19)$$

Similarly, we could get the remaining elements of $\theta^{(m+1)}$.The results are summarized in the Appendix A. The convergent value of $\theta^{(m+1)}$ in the EM algorithm is the ML estimate of θ. Due to the invariance property of ML estimator, plugging the ML estimate $\hat{\theta}$ into ΔVUS and ΔPVUS would give the ML estimates $\widehat{\Delta \text{VUS}}$ and $\widehat{\Delta \text{PVUS}}$.

3.2. Three methods for confidence interval estimation

***3.2.1. Generalized Pivots for* ΔVUS *and* ΔPVUS—**The generalized pivots for μ_k and Σ_k in Eq. (5) are given as (Tian et al., 2011; Lin et al., 2007)

$$R_{\mu_k}=\widehat{\boldsymbol{\mu}}_k-\left(\widehat{\boldsymbol{\Sigma}}_k^{1/2}\mathbf{W}_{k_1}^{-1}\widehat{\boldsymbol{\Sigma}}_k^{1/2}\right)^{1/2}\boldsymbol{Z}_k, R_{\boldsymbol{\Sigma}_k}=n_k\widehat{\boldsymbol{\Sigma}}_k^{1/2}\mathbf{W}_{k_2}^{-1}\widehat{\boldsymbol{\Sigma}}_k^{1/2}\text{for}k=1,2,3,$$

where $\hat{\mu}_k$ and $\hat{\Sigma}_k$ are the ML estimates for μ_k and Σ_k, $\boldsymbol{Z}_k \sim N_2$ ($\mathbf{0}$, $\boldsymbol{I}_2$) with $\boldsymbol{I}_2$ being a 2 by 2 identity matrix and $\boldsymbol{W}_{k_1}$, $\boldsymbol{W}_{k_2} \sim W_2$ ($n_k - 1$, $\boldsymbol{I}_2$) where W_p (m, Σ) denotes a p-dimensional Wishart distribution with degrees of freedom m and scale matrix Σ. Notice that n_k is the sample size for each disease category, which is not obtainable in our case without a GS. Therefore, we have to estimate n_k as well. A naïve estimate for n_k is $\tilde{n}_k = n\hat{p}_k$, where n is the total number of patients and $\hat{p}_k$ is computed from the EM algorithm. However, our preliminary simulations indicate this naïve estimate $\tilde{n}_k$ might not perform well. To account for the randomness brought by no GS test for disease category information, we propose to estimate n_k by $\hat{n}_k$ from a multinomial random variate with the total number of observations being n and the probability for each disease category $\widehat{ps_k}$.

Note that

$$R_{\mu_k}=\begin{pmatrix} R_{\mu_{kA}} \\ R_{\mu_{kB}} \end{pmatrix}, R_{\Sigma_k}=\begin{pmatrix} R_{\sigma_{kA}^2} & R_{\sigma_{kAB}} \\ R_{\sigma_{kAB}} & R_{\sigma_{kB}^2} \end{pmatrix}\text{for}k=1,2,3,$$

the generalized pivots R_{VUS_A} and R_{PVUS_A} for VUS and PVUS for diagnostic test A can be derived as follows,

$$R_{\text{VUS}_A}=\int_{-\infty}^{\infty}\Phi(R_{a_A}s-R_{b_A})\Phi(-R_{c_A}s+R_{d_A})\phi(s)ds, \quad (20)$$

$$R_{\text{PVUS}_A}=\int_{[\Phi^{-1}(\delta_{10})+R_{b_A}]/R_{a_A}}^{[R_{d_A}-\Phi^{-1}(\delta_{30})]/R_{c_A}}[\Phi(R_{a_A}s-R_{b_A})\Phi(-R_{c_A}s+R_{d_A})-\delta_{10}\Phi(-R_{c_A}s+R_{d_A})-\delta_{30}\Phi(R_{a_A}s-R_{b_A})+\delta_{10}\delta_{30}]\phi(s)ds, \quad (21)$$

where $R_{aA} = R_{\sigma 2A}/R_{\sigma 1A}$, $R_{bA} = (R_{\mu 1A} - R_{\mu 2A}) / R_{\sigma 1A}$, $R_{cA} = R_{\sigma 2A} / R_{\sigma 3A}$, $R_{dA} = (R_{\mu 3A} - R_{\mu 2A}) / R_{\sigma 3A}$.

By replacing A with B in Eqs. (20) and Eqs. (21), we can obtain the generalized pivots R_{VUS_B} and $R_{\text{ PVUS}_B}$. Furthermore, the generalized pivots for ΔVUS and ΔPVUS can be defined as

$$R_{\Delta\text{VUS}}=R_{\text{VUS}_A}-R_{\text{VUS}_B}, \quad (22)$$

$$R_{\Delta\text{PVUS}}=R_{\text{PVUS}_A}-R_{\text{PVUS}_B}. \quad (23)$$

See Tian et al. (2011) for details.

For a given data set containing measurements for test *A* and test *B* without a GS, our proposed EM-GV approach for the confidence intervals for ΔVUS and ΔPVUS can be carried out through the following steps:

1. Estimate the vector parameter θ using the EM algorithm to obtain the ML estimates $\hat{\mu}_k$, $\hat{\Sigma}_k$ for k = 1, 2, 3, as well as $\hat{p}_1$ and $\hat{p}_3$.
2. Generate $\hat{n}_k \sim$ *Multinomial* $(n, \hat{p}_k)$.
3. Generate $\boldsymbol{Z}_k \sim N_2\ (\boldsymbol{0}, \boldsymbol{I}_2)$ and $\boldsymbol{W}_{k_1}$, $\boldsymbol{W}_{k_2} \sim W_2\ (\hat{n}_k - 1, \boldsymbol{I}_2)$. Calculate R_{μ_k} and R_{Σ_k}.
4. Compute $R_{\Delta\text{VUS}}$ and $R_{\Delta\text{PVUS}}$ following Eqs. (22)–Eqs. (23).
5. Repeat Steps 2–5 a total 2000 times and obtain an array of $R_{\Delta\text{VUS}}$‘s values and an array of $R_{\Delta\text{PVUS}}$‘s values.

Denote $R_{\Delta\text{VUS}}(\alpha)$ as the 100αth percentile of $R_{\Delta\text{VUS}}$’s. A two-sided 100(1 – α)% confidence interval estimate of ΔVUS is ($R_{\Delta\text{VUS}}$ (α/2), $R_{\Delta\text{VUS}}$ (1 – α/2)). The confidence interval estimation about ΔPVUS can be done in a similar way.

3.2.2. Parametric bootstrap intervals for ΔVUS and ΔPVUS—

Benton and Krishnamoorthy (2002) investigated the performance of the parametric bootstrap (PB) method in the interval estimation of parameters in various statistical problems. They suggested that the PB method is a relatively easy way to obtain good statistical approximation. See also Indurkhya (1994) and Lee (1994) for a good exposition of parametric bootstrap and its application. Our proposed EM-PB method for the confidence intervals for ΔVUS and ΔPVUS can be constructed as follows:

1. Estimate the vector parameter θ using the EM algorithm. Obtain the ML estimates $\hat{\mu}_k$ and $\hat{\Sigma}_k$ for k = 1, 2, 3, as well as $\hat{p}_1$ and $\hat{p}_3$.
2. Generate $\hat{n}_k \sim$ *Multinomial*$(n, \hat{p}_k)$.
3. Generate $\hat{n}_k$’s $\boldsymbol{y}_k \sim N_2(\hat{\mu}_k, \hat{\Sigma}_k)$ for k = 1, 2, 3. Calculate $\bar{\boldsymbol{y}}_k$ and $\boldsymbol{S}_k$, which are the sample mean vector and sample covariance matrix, respectively.
4. Compute $\Delta\widehat{\text{VUS}}$ and $\Delta\widehat{\text{PVUS}}$ following Eqs. (6)–Eqs. (9).
5. Repeat Steps 2–5 a total 2000 times and obtain an array of $\Delta\widehat{\text{VUS}}$‘s values and an array of $\Delta\widehat{\text{PVUS}}$‘s values.

Denote $B_{\Delta\text{VUS}}(\alpha)$ as the 100αth percentile of parametric bootstrap samples $\Delta\widehat{\text{VUS}}$‘s. A two-sided 100(1 – α)% confidence interval estimate of ΔVUS is ($B_{\Delta\text{VUS}}$(α/2), $B_{\Delta\text{VUS}}$ (1 – α/2)). The confidence interval estimation about ΔPVUS can be done also similarly.

3.2.3. Nonparametric bootstrap intervals for ΔVUS and ΔPVUS—

Nonparametric bootstrap resampling is a popular practice to estimate the variance of an estimator and to give the associated confidence intervals (Efron and Tibshirani, 1993; Efron, 1979). Our proposed method for the confidence intervals for ΔVUS and ΔPVUS can be constructed as follows:

1. Estimate the vector parameter θ using the EM algorithm based on the observed data. Obtain the ML estimates $\hat{\mu}_k$ and $\hat{\Sigma}_k$ for k = 1, 2, 3.
2. Compute $\Delta\widehat{\text{VUS}}$ and $\Delta\widehat{\text{PVUS}}$ following Eqs. (6)–Eqs. (9).

3. Generate 2000 bootstrap samples from the observed data with replacement such that each bootstrap sample is of size *n*.
4. Use the EM algorithm to obtain $\Delta\widehat{\text{VUS}}/\Delta\widehat{\text{PVUS}}$ for each bootstrap sample, and from these 2000 bootstrap estimates of $\Delta\widehat{\text{VUS}}/\Delta\widehat{\text{PVUS}}$, we could easily form the bootstrap percentile confidence intervals.

4. Simulation studies

Simulation studies were conducted to compare the coverage probabilities and expected interval lengths of these three approaches for confidence interval estimation without GS, namely, EM-GV, EM-PB and EM-NB. Robustness of the proposed approaches for normal mixture data is also investigated. Data from the bivariate normal distribution $N_2(\mu_k, \Sigma_k)$ (k = 1, 2, 3) for non-diseased, intermediate and diseased categories, respectively, is generated with various means and variance-covariance matrices. Four different configurations of means μ_k and variance-covariance matrices Σ_k are set as follows,

Config	μ_1	μ_2	μ_3	Σ_1	Σ_2	Σ_3
1	(0,0)	(2,3)	(5,4)	$\begin{bmatrix} 0.25 & 0.25 \\ 0.25 & 1 \end{bmatrix}$	$\begin{bmatrix} 0.25 & 0.25 \\ 0.25 & 1 \end{bmatrix}$	$\begin{bmatrix} 0.25 & 0.25 \\ 0.25 & 1 \end{bmatrix}$
2	(0,0)	(2,3)	(5,4)	$\begin{bmatrix} 0.25 & 0.25 \\ 0.25 & 1 \end{bmatrix}$	$\begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 2 \end{bmatrix}$	$\begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix}$
3	(−2.6, −5.7)	(0.3, –0.7	(2.8, 4.0)	$\begin{bmatrix} 4.27 & 3.29 \\ 3.29 & 10.75 \end{bmatrix}$	$\begin{bmatrix} 4.89 & 4.30 \\ 4.30 & 9.59 \end{bmatrix}$	$\begin{bmatrix} 3.15 & 2.14 \\ 2.14 & 5.05 \end{bmatrix}$
4	(−2.6,–1.6	(0.3,–0.8	(2.8,0.6)	$\begin{bmatrix} 4.27 & 0.40 \\ 0.40 & 0.59 \end{bmatrix}$	$\begin{bmatrix} 4.89 & 1.19 \\ 1.19 & 1.20 \end{bmatrix}$	$\begin{bmatrix} 3.15 & 0.55 \\ 0.55 & 0.77 \end{bmatrix}$

The different configurations of means and variance-covariances were chosen to represent possible distributional structures of two biomarkers for each of the three disease categories. The last two configurations come from the real data example.

When a GS is available, there exists a generalized variable approach (Tian et al., 2011) and a large sample approach (Xiong et al., 2007) for interval estimations of ΔVUS and ΔPVUS. Furthermore, Tian et al. (2011) compared the generalized variable (GV) approach with a parametric bootstrap approach and the large sample approach (Xiong et al., 2007) and showed the GV approach usually can provide confidence intervals with satisfactory coverage probabilities. Besides the aforementioned methods, the nonparametric bootstrap method can also be used for the same purpose when there is a GS. However, no simulation study has been done regarding the performance of the nonparametric bootstrap approach. We performed a simulation study for this purpose. The simulation results show the GV outperforms the nonparametric bootstrap method in terms of interval coverage probabilities. For this reason, we will present the results of the GV approach (Tian et al., 2011) for interval estimations of ΔVUS and ΔPVUS when a GS is available for the purpose of assessing the possible efficiency loss due to the missing GS for the three methods: EM-GV, EM-PB and EM-NB.

Table 1 presents the coverage probabilities and expected lengths of confidence intervals for ΔVUS at nominal level .95, obtained by the GV (Tian et al, 2011) in comparison with the EM-GV, M-PB and EM-NB approaches without a GS. The simulation study is based on 2000 random samples. Within each of the 2000 random samples, 2000 $R_{\Delta\text{VUS}}$'s and $B_{\Delta\text{VUS}}$'s, as well as 2000 $\widehat{\Delta\text{VUS}}$'s from nonparametric bootstrap resamples following Section 3.2.3 were calculated to estimate the confidence intervals.

Under Config. 1, the coverage probabilities by EM-GV are generally satisfactory while EM-PB and EM-NB can be slightly liberal for certain scenarios. With Config. 2, EM-GV works reasonably well while EM-PB tends to be quite liberal and EM-NB tends to be slightly conservative. With Config. 3 & 4, we observe the poor coverage probabilities for both EM-GV and EM-PB methods while EM-NB tends to be conservative.

A closer examination of the configurations used in simulation study can explain the findings stated above to a certain extent. Fig. 1–Fig. 4 present true density contours (assuming GS) vs estimated density contours (no GS) for Config. 1–4 respectively. Under Config. 1, the parameter estimate from the EM algorithm is unbiased and accurate and the estimated density contour resembles true density contour. Similarly for Config. 2, the true density contour and estimated density contour are still relatively close. On the other hand, for Config. 3 & 4, as shown in Fig. 3 and 4, due to the fact that the observations from three disease categories are seriously overlapped, it is hard to perfectly separate the three ordinal diagnostic categories and hence it is impossible to get consistent estimates for unknown parameter θ using the EM algorithm. Consequently, the true density contour for Config. 3 & 4 and estimated density contour (Fig. 3 and Fig. 4) are not even close. This phenomenon is even obvious when the sample sizes are large. Thus it is clear that the performances of EM-GV and EM-PB strongly depend on how well these three disease categories can be identified; in other words, it depends on the consistency of ML estimates from the EM algorithm.

Table 2 presents simulation results for ΔPVUS by the EM-GV, EM-PB and EM-NB methods without a GS. The desired minimum classification rates δ_{10} and δ_{30} for non-diseased and diseased categories are set as 0.5, i.e., ΔPVUS is obtained for the region with both the minimum desired specificity and sensitivity for non-disease and diseased categories as 0.5. In general, a similar phenomenon as in Table 1 was observed here. When characterization of the three ordinal diagnostic categories is straightforward as with Config 1 or 2, EM-GV works relatively well. On the contrary, when it is troublesome to distinguish different disease categories, neither EM-GV or EM-PB works, while EM-NB provides somewhat conservative intervals.

Robustness study: To investigate the robustness of the three interval estimation methods without GS, a simulation study was conducted for the mixture of multivariate normal data. Table 3 and Table 4 present simulation results of proposed intervals for ΔVUS and ΔPVUS with normal mixture data, respectively. The simulation configurations are the same as in Table 1 and Table 2. The mixture of bivariate normal random samples was generated as follows,

$$\boldsymbol{Y}_k = \begin{bmatrix} Y_{kA} \\ Y_{kB} \end{bmatrix} \sim 0.9 N_2\left(\boldsymbol{\mu}_k, \frac{1}{1.1}\boldsymbol{\Sigma}_k\right) + 0.1 N_2\left(\boldsymbol{\mu}_k, \frac{2}{1.1}\boldsymbol{\Sigma}_k\right), k=1,2,3.$$

Note that $\boldsymbol{Y}_k$ from such a mixture distribution has mean μ_k and variance Σ_k. The coverage probabilities were similar to the ones with normal data in most cases.

As suggested by one of the referees, we also investigate the robustness of these interval estimation methods with skewed data. The bivariate exponential data was generated via copula. The results were provided in the on-line supplementary document (see Appendix B). We found that all of EM-GV, EM-PB and EM-NB approaches fail to produce confidence intervals with nominal levels for exponential data. This is not surprising, however, because the likelihood function, as well as the generalized variable concept and the parametric bootstrap approach, is based on multivariate normality assumption and hence it is unfair to blindly apply our normal-based method to the skewed data directly without careful data exploration. With a simple log transformation, the coverage probabilities for all three approaches (EM-GV, EM-PB and EM-NB) were significantly improved as shown in Table 5 in the on-line supplementary document (see Appendix B). Especially, the performance of EM-NB is quite satisfactory.

In summary, the investigated EM-GV, EM-PB and EM-NB approaches for confidence interval estimation of ΔVUS and ΔPVUS without a GS are parametric approaches based on multivariate normality. Overall speaking, when the EM algorithm can provide unbiased and accurate ML estimates, all of them have reasonably satisfactory coverage probabilities. When normal assumption is not satisfied, the proposed approaches should be used with caution and potential failures may be expected. We recommend checking the normality assumption before using our proposed methods.

Remark

The proposed methods can easily provide *P*-values for hypothesis testing. For example, the percentage that $R_{\Delta\mathrm{VUS}}$’s, $B_{\Delta\mathrm{VUS}}$’s or bootstrap resampling $\widehat{\Delta\mathrm{VUS}}$‘s from replacement are less than or equal to $\Delta\mathrm{VUS}_0$ is a Monte Carlo estimate of the generalized or parametric/nonparametric bootstrap *P*-value for testing $\Delta\mathrm{VUS} = \Delta\mathrm{VUS}_0$ versus $\Delta\mathrm{VUS} > \Delta\mathrm{VUS}_0$.

5. Data application

In this section, we apply all three investigated interval estimation methods to a data set from the longitudinal cohort study of the Washington University Knight Alzheimer’s Disease Research Center to compare the diagnostic accuracy of early stage AD between two different pairs of neuropsychological tests. Each individual was assessed by experienced clinicians. The severity of dementia was based on the Clinical Dementia Rating (CDR) according to published rules (Morris, 1993). In Xiong et al. (2006) paper, the CDR was considered as the GS. Based on the GS information, 45 people were classified as non-demented (i.e., CDR 0, $D = 1$), 44 individuals were classified to have very mild AD (i.e., CDR 0.5, $D = 2$), and 29 individuals were classified to be mildly demented (i.e., CDR 1, $D = 3$).

After the clinical evaluation, each individual also completed psychometric tests (Xiong et al., 2006). The clinical assessment and psychometric testing were conducted independently by clinicians and psychometricians. We are interested in the comparison of the diagnostic accuracy between different diagnostic tests. We focus on two pairs of factor scores derived from the psychometric tests: the mental control/frontal factor versus the verbal memory/temporal factor, and the mental control/frontal factor versus the visual retention test (10-s exposure).

We first examined the bivariate distribution of the tests in the sample through the Shapiro–Wilk test for multivariate normality and found no significant statistical evidence that these two pairs of psychometric tests deviate from the normal distributions for each of three categories based on the CDR.

The fact that misclassification errors in the CDR are unavoidable makes it reasonable to apply our methods without a GS. Basically, the input data only consist of two columns with measurements from test *A* and test *B* for each individual. Xiong et al. (2006) reported the means and standard deviations of these factor scores as well as their estimated VUSs (0.657 for frontal factor, 0.752 for temporal factor, and 0.587 for visual retention test) using the CDR. They also assessed the pairwise difference on VUSs (95% confidence interval (−0.206, 0.016) for ΔVUS between frontal factor and temporal factor; 95% confidence interval (−0.066, 0.206) for ΔVUS between frontal factor and visual retention) (Xiong et al., 2007). Our proposed EM-GV, EM-PB and EM-NB methods without a GS clinical evaluation were applied to the same factor scores. The estimated difference in paired VUSs was −0.374 with a 95% confidence interval (−0.545, −0.250) from EM-GV, (−0.517, −0.227) from EM-PB and (−0.484,0.270) from EM-NB between frontal factor and temporal factor; the estimated difference in paired VUSs was 0.124 with a 95% confidence interval (0.038, 0.198) from EM-GV, (0.042, 0.209) from EM-PB and (−0.297,0.471) from EM-NB between frontal factor and visual retention.

It is interesting to note that confidence intervals for the difference in paired VUSs from both EM-GV and EM-PB do not contain zero while the ones from EM-NB contain zero. Considering the coverage probabilities presented in the simulation studies, it is better referring to the intervals from EM-NB, although it has been shown that EM-NB in this case might be conservative. Thus, on the safe side, we may not conclude there is a statistically significant difference in diagnostic accuracy between the temporal factor and frontal factor, as well as between the frontal factor and the visual retention test. These results are consistent with the findings of Xiong et al. (2007).

6. Discussion

This paper addresses the problem of confidence interval estimation of the difference between paired VUSs (ΔVUS) and between paired PVUSs (ΔPVUS) without a GS. The combination of the following two components makes this problem unique: (1) diseases with three ordinal diagnostic categories; (2) with no or questionable GS. The investigated EM-GV, EM-PB and EM-NB methods are ML based approaches using the EM algorithm in conjunction with generalized variables and parametric/nonparametric bootstrap approaches. They provide flexibility when patients’ classification information was contaminated or even lost.

Based on our simulation studies, both EM-GV and EM-PB approaches perform reasonably well for finite sample sizes considering the coverage probabilities when the three ordinal diagnostic categories are easy to separate. When it is really difficult to discriminate three categories, the EM-NB approach provides confidence intervals with coverage more close to the nominal level, at the cost of producing the intervals much wider. On the other hand, when the three ordinal diagnostic categories are relatively easy to be characterized, EM-GV provides more accurate intervals. Compared to the EM-NB method, in those scenarios, efficiency can be gained by using the EM-GV method.

We suggest exploratory data visualization to check the separability of three ordinal diagnostic categories before choosing to use either EM-GV, EM-PB or EM-NB method for confidence interval estimation of the difference in diagnostic accuracy without GS. Another method to diagnose category separability is to consider *k*-means clustering. Whether clustering result depends on the initial clusters or not may indicate good separability of different diagnostic categories.

Programs in *R* and *C* for estimating the difference as well as the estimated intervals for the difference in paired VUSs and PVUSs using the EM-GV, EM-PB and EM-NB are available upon request.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This study was partly supported by National Institute on Aging grant P50 AG05681, P01 AG03991, and R01 AG034119 for Dr. Chengjie Xiong. The opinions expressed are those of the authors and not necessarily those of the Editors. The authors thank the referees for helpful discussions and comments.

Appendix A

$$\frac{\partial Q}{\partial \mu_{1A}}=\sum_{i=1}^{n} q_{i1}^{(m)} \begin{pmatrix} t_{iA}-\mu_{1A} \\ t_{iB}-\mu_{1B} \end{pmatrix}' \Sigma_1^{-1} \begin{pmatrix} t_{iA}-\mu_{1A} \\ t_{iB}-\mu_{1B} \end{pmatrix} \equiv 0$$, hereafter denote

$$\Sigma_1^{-1}=\begin{bmatrix} A_{11} & A_{12} \\ A_{12} & A_{22} \end{bmatrix},$$

$$\begin{cases} \frac{\partial Q}{\partial \mu_{1A}}=\sum\limits_{i=1}^{n} q_{i1}^{(m)} \left\{A_{11}(t_{iA}-\mu_{1A})+A_{12}(t_{iB}-\mu_{1B})\right\} \equiv 0 \\ \frac{\partial Q}{\partial \mu_{1B}}=\sum\limits_{i=1}^{n} q_{i1}^{(m)} \left\{A_{12}(t_{iA}-\mu_{1A})+A_{22}(t_{iB}-\mu_{1B})\right\} \equiv 0 \end{cases}$$

and we obtain

$$\begin{cases} \widehat{\mu}_{1A}^{(m+1)}=\frac{\sum\limits_{i=1}^{n} q_{i1}^{(m)} t_{iA}}{\sum\limits_{i=1}^{n} q_{i1}^{(m)}} \\ \widehat{\mu}_{1B}^{(m+1)}=\frac{\sum\limits_{i=1}^{n} q_{i1}^{(m)} t_{iB}}{\sum\limits_{i=1}^{n} q_{i1}^{(m)}}. \end{cases}$$

In the similar way, we could have

$$\widehat{\sigma}_{1A}^{2(m+1)}=\frac{\sum\limits_{i=1}^{n} q_{i1}^{(m)} (t_{iA}-\widehat{\mu}_{1A})^2}{\sum\limits_{i=1}^{n} q_{i1}^{(m)}}, \widehat{\sigma}_{1B}^{2(m+1)}=\frac{\sum\limits_{i=1}^{n} q_{i1}^{(m)} (t_{iB}-\widehat{\mu}_{1B})^2}{\sum\limits_{i=1}^{n} q_{i1}^{(m)}}, \widehat{\sigma}_{1AB}^{2(m+1)}=\frac{\sum\limits_{i=1}^{n} q_{i1}^{(m)} (t_{iA}-\widehat{\mu}_{1A})(t_{iB}-\widehat{\mu}_{1B})}{\sum\limits_{i=1}^{n} q_{i1}^{(m)}}.$$

Also,

$$
\begin{aligned}
&\widehat{\mu}_{3A}^{(m+1)}\\
&=\frac{\sum_{i=1}^{n} q_{i3}^{(m)} t_{iA}}{\sum_{i=1}^{n} q_{i3}^{(m)}}, \widehat{\mu}_{3B}^{(m+1)}\\
&=\frac{\sum_{i=1}^{n} q_{i3}^{(m)} t_{iB}}{\sum_{i=1}^{n} q_{i3}^{(m)}}, \widehat{\sigma}_{3A}^{2(m+1)}\\
&=\frac{\sum_{i=1}^{n} q_{i3}^{(m)} (t_{iA}-\widehat{\mu}_{1A})^2}{\sum_{i=1}^{n} q_{i3}^{(m)}}, \widehat{\sigma}_{3B}^{(m+1)}\\
&=\frac{\sum_{i=1}^{n} q_{i3}^{(m)} (t_{iB}-\widehat{\mu}_{3B})^2}{\sum_{i=1}^{n} q_{i3}^{(m)}}, \widehat{\sigma}_{3AB}^{2(m+1)}\\
&=\frac{\sum_{i=1}^{n} q_{i3}^{(m)} (t_{iA}-\widehat{\mu}_{3A})(t_{iB}-\widehat{\mu}_{3B})}{\sum_{i=1}^{n} q_{i3}^{(m)}} \widehat{\mu}_{2A}^{(m+1)}\\
&=\frac{\sum_{i=1}^{n} \left(1-q_{i1}^{(m)}-q_{i3}^{(m)}\right) t_{iA}}{\sum_{i=1}^{n} \left(1-q_{i1}^{(m)}-q_{i3}^{(m)}\right)}, \widehat{\mu}_{2B}^{(m+1)}\\
&=\frac{\sum_{i=1}^{n} \left(1-q_{i1}^{(m)}-q_{i3}^{(m)}\right) t_{iB}}{\sum_{i=1}^{n} \left(1-q_{i1}^{(m)}-q_{i3}^{(m)}\right)}, \widehat{\sigma}_{2A}^{2(m+1)}\\
&=\frac{\sum_{i=1}^{n} \left(1-q_{i1}^{(m)}-q_{i3}^{(m)}\right) (t_{iA}-\widehat{\mu}_{2A})^2}{\sum_{i=1}^{n} \left(1-q_{i1}^{(m)}-q_{i3}^{(m)}\right)}, \widehat{\sigma}_{2B}^{2(m+1)}\\
&=\frac{\sum_{i=1}^{n} \left(1-q_{i1}^{(m)}-q_{i3}^{(m)}\right) (t_{iB}-\widehat{\mu}_{2B})^2}{\sum_{i=1}^{n} \left(1-q_{i1}^{(m)}-q_{i3}^{(m)}\right)}, \widehat{\sigma}_{2AB}^{2(m+1)}\\
&=\frac{\sum_{i=1}^{n} \left(1-q_{i1}^{(m)}-q_{i3}^{(m)}\right) (t_{iA}-\widehat{\mu}_{2A})(t_{iB}-\widehat{\mu}_{2B})}{\sum_{i=1}^{n} \left(1-q_{i1}^{(m)}-q_{i3}^{(m)}\right)},
\end{aligned}
$$

Appendix B. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.csda.2013.07.007.

References

Beiden SV, Campbell G, Meier KL, Wagner RF. On the problem of ROC analysis without truth: the EM algorithm and the information matrix. Proceedings of the Society of Photo-Optical Instrumentation Engineers (SPIE). 2000; 3981:126–134.

Benton D, Krishnamoorthy K. Performance of the parametric bootstrap method in small sample interval estimates. Advances and Applications in Statistics. 2002; 2:269–285.

Efron B. Bootstrap methods: another look at the jackknife. Annals of Statistics. 1979; 7:1–26.

Efron, B.; Tibshirani, R. An Introduction to the Bootstrap. London: Chapman & Hall; 1993.

Henkelman RM, Kay I, Bronskill MJ. Receiver operator characteristic (ROC) analysis without truth. Medical Decision Making. 1990; 10:24–29. [PubMed: 2325524]

Hsieh HN, Su HY, Zhou XH. Interval estimation for the difference in paired areas under the ROC curves in the absence of a gold standard test. Statistics in Medicine. 2009; 28:3108–3123. [PubMed: 19691022]

Indurkhya A. A parametric bootstrap procedure to estimate the selected mean using censored data. Statistics & Probability Letters. 1994; 21:291–298.

Lee SMS. Optimal choice between parametric and non-parametric bootstrap estimates. Mathematical Proceedings of the Cambridge Philosophical Society. 1994; 115:335–363.

Lin SH, Lee JC, Wang RS. Generalized inferences on the common mean vector of several multivariate normal populations. Journal of Statistical Planning and Inference. 2007; 137:2240–2249.

Morris JC. The clinical dementia rating (CDR): current version and scoring rules. Neurology. 1993; 43:1412–1414.

Nakas CT, Yiannoutsos CT. Ordered multiple-class ROC analysis with continuous measurements. Statistics in Medicine. 2004; 23:3437–3449. [PubMed: 15505886]

Pepe, MS. The Statistical Evaluation of Medical Tests for Classification and Prediction. New York: Oxford University Press; 2003.

Scurfield BK. Multiple-event forced-choice tasks in the theory of signal detectability. Journal of Mathematical Psychology. 1996; 40:253–269. [PubMed: 8979976]

Shapiro DE. The interpretation of diagnostic tests. Statistical Methods in Medical Research. 1999; 8:113–134. [PubMed: 10501649]

Tian L, Xiong CJ, Lai CY, Vexler A. Exact confidence interval estimation for the difference in diagnostic accuracy with three ordinal diagnostic groups. Journal of Statistical Planning and Inference. 2011; 141:549–558. [PubMed: 23538945]

Xiong CJ, van Belle G, Miller JP, Morris JC. Measuring and estimating diagnostic accuracy when there are three ordinal diagnostic groups. Statistics in Medicine. 2006; 25:1251–1273. [PubMed: 16345029]

Xiong CJ, van Belle G, Miller JP, Yan Y, Gao F, Yu K, Morris JC. A parametric comparison of diagnostic accuracy with three ordinal diagnostic groups. Biometrical Journal. 2007; 49:682–693. [PubMed: 17763377]

Zhou XH, Castelluccio P, Zhou C. Nonparametric estimation of ROC curves in the absence of a gold standard. Biometrics. 2005; 61:600–609. [PubMed: 16011710]

Zhou, XH.; Obuchowski, NA.; McClish, DK. Statistical Methods in Diagnostic Medicine. New York: Wiley; 2002.

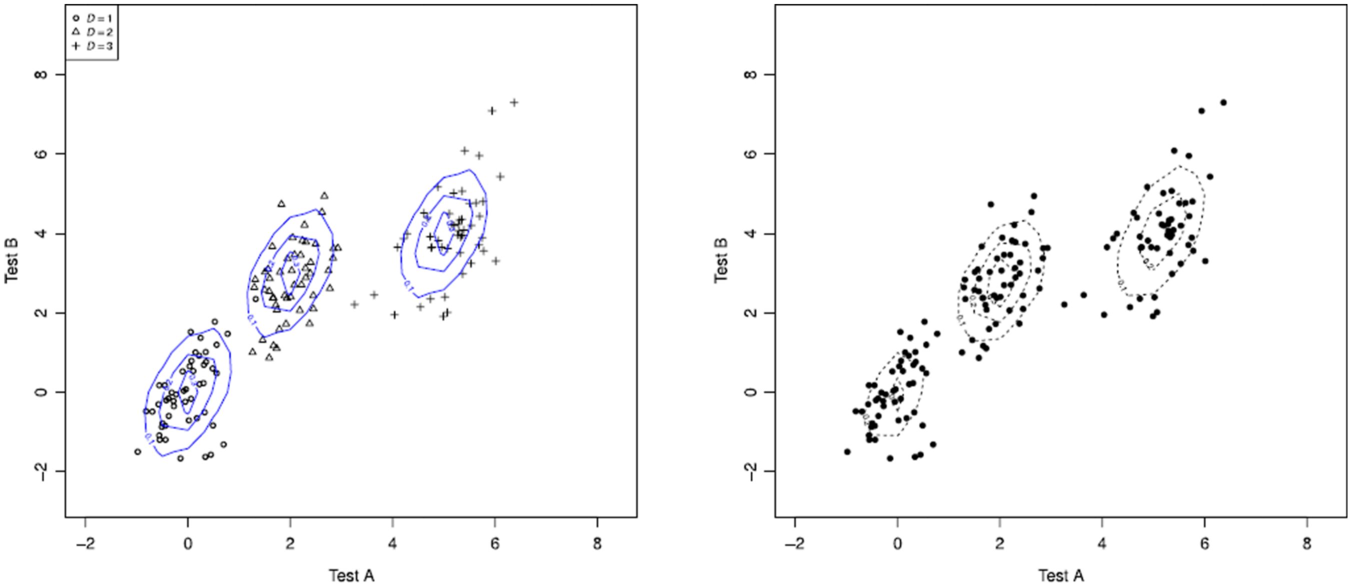

Fig. 1.

Density contours for Config. 1: (a) with GS (left); (b) without GS (right, estimated from EM algorithm).

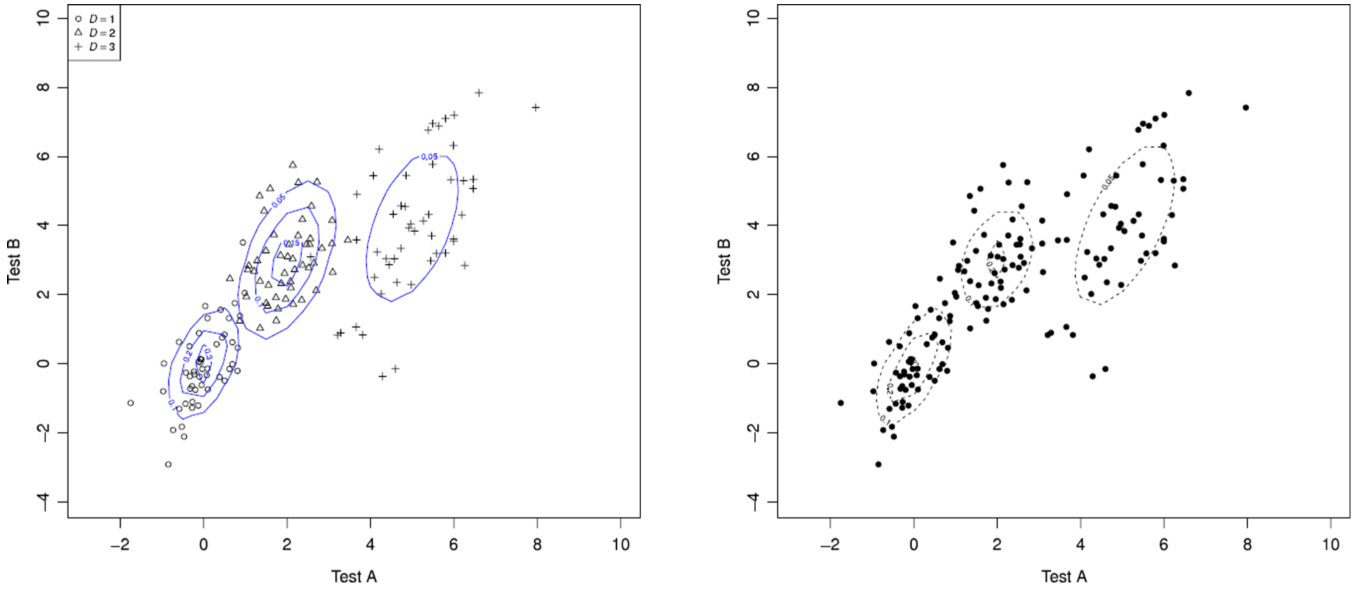

Fig. 2.

Density contours for Config. 2: (a) with GS (left); (b) without GS (right, estimated from EM algorithm).

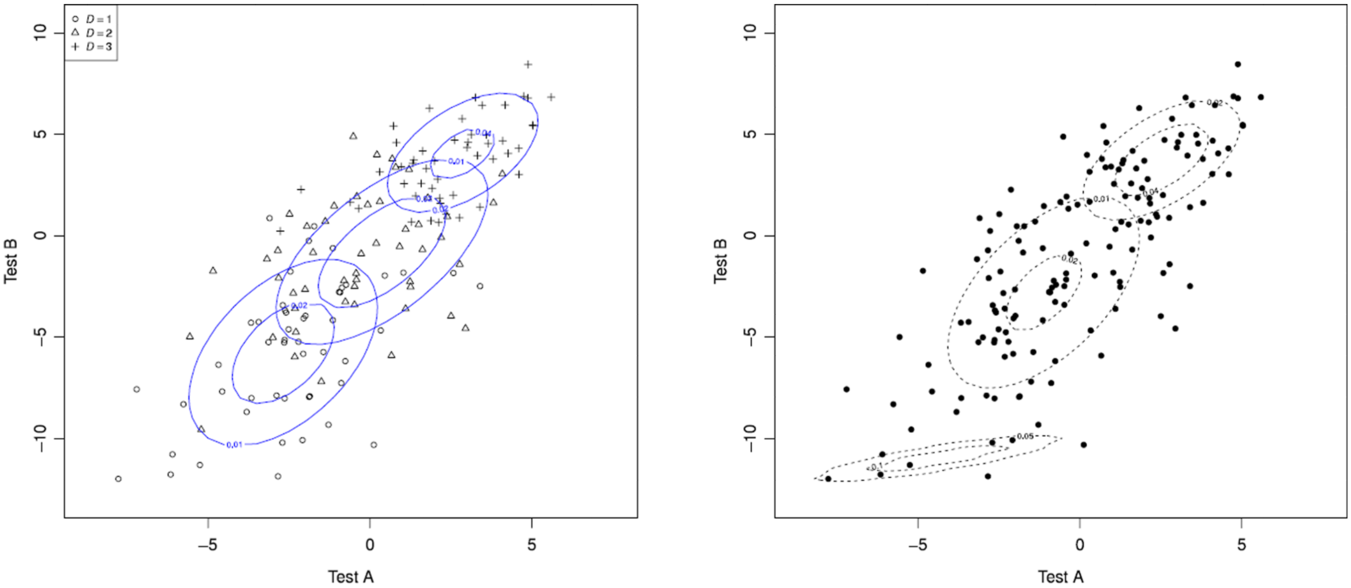

Fig. 3.
Density contours for Config. 3: (a) with GS (left); (b) without GS (right, estimated from EM algorithm).

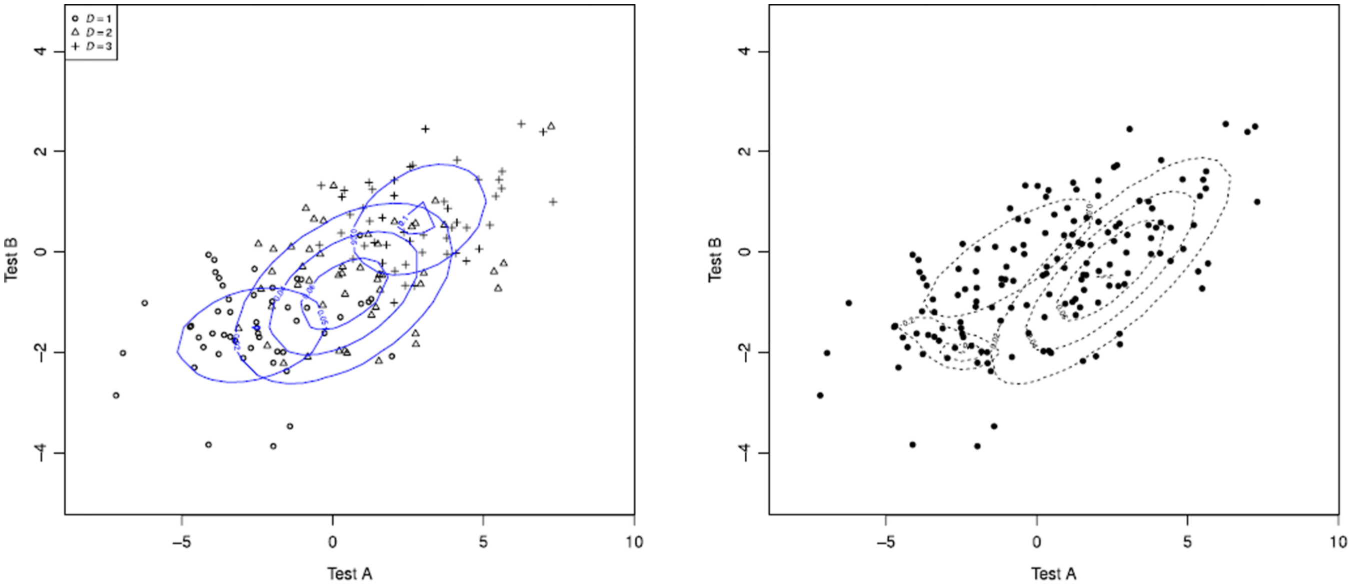

Fig. 4.

Density contours for Config. 4: (a) with GS (left); (b) without GS (right, estimated from EM algorithm).

Table 1

Coverage probabilities (and expected lengths) of proposed 95% confidence intervals for ΔVUS (2000 simulations).

Config	Sample sizes	GV	EM-GV	EM-PB	EM-NB
1	(20, 20, 20)	.947 (0.2869)	.944 (0.3021)	.937 (0.2842)	.934 (0.3148)
	(20, 30, 50)	.954 (0.2601)	.946 (0.2717)	.933 (0.2602)	.940 (0.2814)
	(30, 30, 30)	.952 (0.2333)	.942 (0.2393)	.941 (0.2301)	.943 (0.2467)
	(30, 40, 50)	.953 (0.2084)	.947 (0.2096)	.932 (0.2064)	.942 (0.2191)
	(50, 50, 50)	.944 (0.1792)	.940 (0.1814)	.944 (0.1783)	.945 (0.1869)
2	(20, 20, 20)	.961 (0.3011)	.948 (0.3236)	.911 (0.3154)	.974 (0.4362)
	(20, 30, 50)	.965 (0.2867)	.959 (0.3057)	.903 (0.2851)	.962 (0.4045)
	(30, 30, 30)	.962 (0.2458)	.940 (0.2654)	.912 (0.2530)	.977 (0.3451)
	(30, 40, 50)	.959 (0.2214)	.951 (0.2432)	.898 (0.2289)	.967 (0.3038)
	(50, 50, 50)	.956 (0.1913)	.933 (0.2063)	.917 (0.1925)	.970 (0.2539)
3	(20, 20, 20)	.958 (0.2981)	.779 (0.3737)	.715 (0.3164)	.996 (0.7842)
	(20, 30, 50)	.954 (0.2498)	.726 (0.3226)	.614 (0.2691)	.993 (0.7464)
	(30, 30, 30)	.952 (0.2391)	.698 (0.3053)	.626 (0.2612)	.992 (0.7493)
	(30, 40, 50)	.946 (0.2132)	.658 (0.2679)	.634 (0.2426)	.992 (0.7163)
	(50, 50, 50)	.954 (0.1828)	.626 (0.2267)	.612 (0.2102)	.988 (0.7041)
4	(20, 20, 20)	.957 (0.3545)	.789 (0.4322)	.688 (0.3786)	.989 (0.9673)
	(20, 30, 50)	.958 (0.2976)	.708 (0.3497)	.666 (0.3154)	.991 (0.9318)
	(30, 30, 30)	.958 (0.2867)	.702 (0.3366)	.647 (0.3113)	.990 (0.9328)
	(30, 40, 50)	.952 (0.2556)	.677 (0.2998)	.633 (0.2814)	.993 (0.9013)
	(50, 50, 50)	.962 (0.2219)	.627 (0.2528)	.604 (0.2418)	.987 (0.8735)

See Section 4 parameter configurations.

Table 2

Coverage probabilities (and expected lengths)ofproposed 95% confidence intervals for ΔPVUS(2000simulations).

Config	Sample sizes	GV	EM-GV	EM-PB	EM-NB
1	(20, 20, 20)	.953 (0.1033)	.948 (0.1069)	.936 (0.1056)	.937 (0.1157)
	(20, 30, 50)	.951 (0.0960)	.952 (0.0987)	.928 (0.0962)	.939 (0.1025)
	(30, 30, 30)	.955 (0.0853)	.944 (0.0870)	.939 (0.0867)	.943 (0.0945)
	(30, 40, 50)	.948 (0.0742)	.946 (0.0788)	.930 (0.0765)	.941 (0.0811)
	(50, 50, 50)	.945 (0.0667)	.941 (0.0675)	.941 (0.0663)	.944 (0.0722)
2	(20, 20, 20)	.965 (0.0993)	.939 (0.1049)	.915 (0.1065)	.963 (0.1513)
	(20, 30, 50)	.954 (0.0917)	.944 (0.0999)	.908 (0.0922)	.968 (0.1335)
	(30, 30, 30)	.958 (0.0822)	.938 (0.0842)	.915 (0.0871)	.972 (0.1202)
	(30, 40, 50)	.952 (0.0724)	.948 (0.0779)	.902 (0.0740)	.970 (0.1035)
	(50, 50, 50)	.956 (0.0652)	.936 (0.0659)	.921 (0.0678)	.972 (0.0894)
3	(20, 20, 20)	.956 (0.1049)	.794 (0.1227)	.734 (0.1097)	.994 (0.2674)
	(20, 30, 50)	.956 (0.0914)	.748 (0.1075)	.653 (0.0921)	.987 (0.2552)
	(30, 30, 30)	.956 (0.0861)	.715 (0.1012)	.644 (0.0899)	.991 (0.2569)
	(30, 40, 50)	.947 (0.0789)	.673 (0.0913)	.642 (0.0842)	.989 (0.2462)
	(50, 50, 50)	.950 (0.0672)	.621 (0.0767)	.612 (0.0724)	.983 (0.2413)
4	(20, 20, 20)	.956 (0.1098)	.792 (0.1283)	.708 (0.1167)	.993 (0.3071)
	(20, 30, 50)	.956 (0.0941)	.716 (0.1052)	.678 (0.0969)	.989 (0.2943)
	(30, 30, 30)	.953 (0.0902)	.708 (0.1016)	.666 (0.0961)	.990 (0.2963)
	(30, 40, 50)	.958 (0.0819)	.686 (0.0914)	.643 (0.0873)	.988 (0.2857)
	(50, 50, 50)	.956 (0.0713)	.645 (0.0778)	.615 (0.0755)	.982 (0.2764)

See Section 4 parameter configurations.

Table 3

Coverage probabilities (and expected lengths) of proposed 95% confidence intervals for ΔVUS with mixture normal data (2000 simulations).

Config	Sample sizes	GV	EM-GV	EM-PB	EM-NB
1	(20, 20, 20)	.949 (0.2903)	.945 (0.3113)	.936 (0.2855)	.932 (0.3227)
	(20, 30, 50)	.953 (0.2697)	.947 (0.2809)	.934 (0.2637)	.941 (0.2952)
	(30, 30, 30)	.951 (0.2404)	.943 (0.2442)	.944 (0.2344)	.943 (0.2511)
	(30, 40, 50)	.952 (0.2095)	.946 (0.2123)	.935 (0.2109)	.943 (0.2240)
	(50, 50, 50)	.948 (0.1811)	.941 (0.1928)	.942 (0.1796)	.944 (0.1933)
2	(20, 20, 20)	.962 (0.3112)	.947 (0.3288)	.916 (0.3246)	.973 (0.4467)
	(20, 30, 50)	.964 (0.2973)	.960 (0.3103)	.901 (0.3030)	.966 (0.4185)
	(30, 30, 30)	.960 (0.2584)	.941 (0.2697)	.915 (0.2612)	.973 (0.3573)
	(30, 40, 50)	.957 (0.2330)	.950 (0.2481)	.897 (0.2323)	.968 (0.3190)
	(50, 50, 50)	.959 (0.2001)	.934 (0.2132)	.914 (0.2045)	.971 (0.2666)
3	(20, 20, 20)	.956 (0.2971)	.796 (0.3829)	.692 (0.3201)	.995 (0.7767)
	(20, 30, 50)	.960 (0.2511)	.734 (0.3194)	.642 (0.2768)	.994 (0.7439)
	(30, 30, 30)	.951 (0.2388)	.715 (0.3094)	.624 (0.2693)	.992 (0.7391)
	(30, 40, 50)	.949 (0.2144)	.667 (0.2731)	.642 (0.2411)	.989 (0.7181)
	(50, 50, 50)	.942 (0.1823)	.627 (0.2341)	.603 (0.2158)	.990 (0.6953)
4	(20, 20, 20)	.964 (0.3538)	.782 (0.4336)	.701 (0.3939)	.986 (0.9561)
	(20, 30, 50)	.958 (0.2978)	.716 (0.3613)	.672 (0.3222)	.989 (0.9274)
	(30, 30, 30)	.956 (0.2873)	.713 (0.3518)	.668 (0.3214)	.988 (0.9202)
	(30, 40, 50)	.951 (0.2552)	.686 (0.3095)	.642 (0.2856)	.987 (0.8869)
	(50, 50, 50)	.954 (0.2211)	.641 (0.2631)	.618 (0.2488)	.984 (0.8573)

See Section 4 parameter configurations.

Table 4

Coverage probabilities (and expected lengths) of proposed 95% confidence intervals for ΔPVUS with mixture normal data (2000 simulations).

Config	Sample sizes	GV	EM-GV	EM-PB	EM-NB
1	(20, 20, 20)	.954 (0.1075)	.946 (0.1109)	.938 (0.1121)	.939 (0.1189)
	(20, 30, 50)	.952 (0.0992)	.948 (0.1013)	.930 (0.1014)	.941 (0.1095)
	(30, 30, 30)	.953 (0.0911)	.942 (0.0902)	.938 (0.0899)	.945 (0.1003)
	(30, 40, 50)	.945 (0.0754)	.943 (0.0821)	.931 (0.0802)	.942 (0.0854)
	(50, 50, 50)	.947 (0.0682)	.942 (0.0715)	.939 (0.0698)	.942 (0.0786)
2	(20, 20, 20)	.958 (0.1002)	.941 (0.1103)	.917 (0.1079)	.962 (0.1635)
	(20, 30, 50)	.949 (0.0985)	.943 (0.1012)	.911 (0.0930)	.964 (0.1412)
	(30, 30, 30)	.953 (0.0854)	.936 (0.0861)	.913 (0.0914)	.965 (0.1296)
	(30, 40, 50)	.951 (0.0741)	.945 (0.0768)	.909 (0.0755)	.969 (0.1102)
	(50, 50, 50)	.954 (0.0677)	.938 (0.0664)	.919 (0.0681)	.970 (0.0911)
3	(20, 20, 20)	.952 (0.1045)	.796 (0.1236)	.729 (0.1085)	.989 (0.2656)
	(20, 30, 50)	.957 (0.0921)	.743 (0.1089)	.664 (0.0939)	.985 (0.2533)
	(30, 30, 30)	.951 (0.0859)	.714 (0.1016)	.651 (0.0925)	.987 (0.2533)
	(30, 40, 50)	.959 (0.0784)	.680 (0.0922)	.655 (0.0848)	.988 (0.2459)
	(50, 50, 50)	.944 (0.0671)	.624 (0.0777)	.620 (0.0725)	.980 (0.2378)
4	(20, 20, 20)	.958 (0.1096)	.796 (0.1271)	.714 (0.1182)	.988 (0.3036)
	(20, 30, 50)	.950 (0.0947)	.714 (0.1059)	.682 (0.0978)	.983 (0.2908)
	(30, 30, 30)	.959 (0.0905)	.713 (0.1029)	.670 (0.0977)	.984 (0.2915)
	(30, 40, 50)	.952 (0.0821)	.702 (0.0918)	.645 (0.0870)	.980 (0.2804)
	(50, 50, 50)	.948 (0.0709)	.656 (0.0774)	.618 (0.0756)	.978 (0.2704)

See Section 4 parameter configurations.