

4-26-2009

# Interval Estimation for the Difference in Paired Areas under the ROC Curves in the Absence of a Gold Standard Test

Hsin-Neng Hsieh
*National Taiwan University*

Hsiu-Yuan Su
*National Taiwan University*

Xiao-Hua Zhou
*University of Washington*, azhou@u.washington.edu

# Interval Estimation for the Difference in Paired Areas under the ROC Curves in the Absence of a Gold Standard Test

Hsin-Neng Hsieh[1], Hsiu-Yuan Su[1] and Xiao-Hua Zhou[2,3]

[1]Division of Biometry, Institute of Agronomy,
National Taiwan University, Taipei, Taiwan
[2]HSR&D VA Puget Sound Health Care System, Seattle, Washington 98101, U.S.A.
[3]Department of Biostatistics, University of Washington,
Box 357232, Seattle, Washington 98195, U.S.A.

Correspondent: Xiao-Hua Zhou, PhD, Department of Biostatistics, University of Washington,
Box 357232, Seattle, Washington 98195, U.S.A.

E-mail: azhou@u.washington.edu

**Summary:** Receiver operating characteristic (ROC) curves can be used to assess the accuracy of tests measured on ordinal or continuous scales. The most commonly used measure for the overall diagnostic accuracy of diagnostic tests is the area under the ROC curve (AUC). A gold standard test on the true disease status is required to estimate the AUC. However, a gold standard test may sometimes be too expensive or infeasible. Therefore, in many medical research studies, the true disease status of the subjects may remain unknown. Under the normality assumption on test results from each disease group of subjects, using the expectation-maximization (EM) algorithm in conjunction with a bootstrap method, we propose a maximum likelihood based procedure for construction of confidence intervals for the difference in paired areas under ROC curves in the absence of a gold standard test. Simulation results show that the proposed interval estimation procedure yields satisfactory coverage probabilities and interval lengths. The proposed method is illustrated with two examples.

**Key words and phrases:** Area under the ROC curve, EM algorithm, bootstrap method, gold standard test, maximum likelihood estimation.

## 1. Introduction

One of the primary objectives in any diagnostic test evaluation study is to compare the diagnostic accuracy of the new diagnostic procedure with that of a current procedure, and one of the common measures for the overall diagnostic accuracy is the area under the receiver operating characteristic (ROC) curve (AUC) [1, 2]. The difference in the AUCs can be used as a measure for comparison of diagnostic accuracy between two diagnostic tests.

When the gold standard (GS) test on the disease status is available, several methods have been proposed to compare the difference in the AUCs. The non-parametric method presented in DeLong *et al.* [3] and the maximum likelihood estimation under the normal assumption provided by McClish [4] are the two most commonly used methods for this problem. Most recently, Li *et al.* [5] proposed an exact interval estimation based on the concept of a generalized pivotal quantity and showed that their method outperforms both the nonparametric method and the maximum likelihood method in their intensive simulation study.

However, a GS test may not always exist or may be too expensive or infeasible. Therefore, in many diagnostic accuracy studies, an imperfect GS test is used to evaluate the accuracy of tests instead, which can result in biased estimates of diagnostic accuracy. The statistical inferences for ROC analysis without the GS test remain relatively unexplored. Henkelman *et al.* [7] considered the estimation problem of ROC curves of continuous-scale tests in the absence of a GS test and showed that ROC curves of two or more continuous-scale tests could be estimated in the absence of a GS test under the multivariate normality assumption on test results of a diseased and non-diseased subjects. Beiden *et al.* [8] also proposed maximum likelihood (ML) estimates of the ROC curves of continuous-scale tests using the EM algorithm.

For binary and ordinal scale test data, some methods were proposed for estimating sensitivity and specificity of the two diagnostic correlated tests in the absence of a GS test. For example, Enøe *et al.* [9], Dendukuri *et al.* [10], and Georgiadis *et al.* [11] applied Bayesian modeling to solving binary scale diagnostic testing problems,

and Zhou *et al.* [12] developed a nonparametric maximum likelihood method for estimating ROC curves and AUCs of ordinal-scale tests in the absence of a GS test.

Choi *et al.* [13] proposed a Bayesian method for construction of the difference between AUCs of the two correlated tests under the no-gold-standard (NGS) situation, and their method is based on the assumption that observed data come from a mixture of two bivariate normal distributions. Branscum *et al.* [14] proposed another Bayesian approach for ROC curve estimation, based on mixtures of Polya trees, which allows more flexibility, especially if the underlying distributions of test results are multimodal. Although the Bayesian methods appear to perform well in a limited simulation study, the methods still require a carefully chosen prior for the model parameters. Branscum *et al.* [14] cautioned the use of noninformative priors in Bayesian analysis of NGS diagnostic testing problems and advocated the use of real and informative prior in such the Bayesian analysis. In addition, the Bayesian methods may be sensitive to the bivariate parametric distributional assumption on test results, as noted in Choi *et al.* [13].

All the proposed methods above, except Choi's method, for dealing with the absence of a GS test focus on point estimation of ROC curves, not on interval estimation. In this paper, we focus on interval estimation for the difference in paired AUCs under the NGS situation. Using the EM algorithm in conjunction with the bootstrap method, we propose a new likelihood-based procedure for the construction of confidence intervals for the difference in paired AUCs under the NGS case. We present the proposed methods in Section 2, and carry out the simulation in Section 3 to compare the performance of the proposed method with existing methods. In Section 4, we illustrate the use of the proposed method with two published data sets. Finally, we conclude the article with some discussion and final remarks in Section 5.

## 2. The Proposed Methods

Let $T_1$ and $T_2$ be test results of two diagnostic tests on the same patient whose disease status is denoted by $D$. If the patient is diseased, then $D = 1$; and if the patient is non-diseased, then $D = 0$. We denote the results of the two tests

on a diseased patient by $X_1$ and $X_2$, respectively, and those on a non-diseased patient by $Y_1$ and $Y_2$, respectively. Furthermore, let $P(X_j > c) = S_{X,j}(c)$ and $P(Y_j > c) = S_{Y,j}(c)$ be the true positive and false positive fractions at a threshold $c$ for diagnostic test $j$, respectively. For diagnostic test $j$, an ROC curve plots $\{S_{Y,j}(c), S_{X,j}(c)\}$ for all possible values of threshold $c$. We can also write the ROC curve as a function of $t = S_{Y,j}(c)$, given by $\mathrm{ROC}_j(t) = S_{X,j}(S_{Y,j}^{-1}(t))$, where $S_{Y,j}^{-1}(t)$ is the inverse function of $S_{Y,j}(t)$. The AUC for diagnostic test $j$ is $A_j = \int_0^1 \mathrm{ROC}_j(t)dt$, which can be shown to be $A_j = P(X_j \geq Y_j)$.

Assume that two test results of a diseased subject, $X_1$ and $X_2$, follow a bivariate normal distribution,

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N_2(\boldsymbol{\mu}_D, \boldsymbol{\Sigma}_D),$$

and that test results of a non-diseased subject, $Y_1$ and $Y_2$, also follow a bivariate normal distribution,

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \sim N_2(\boldsymbol{\mu}_{\bar{D}}, \boldsymbol{\Sigma}_{\bar{D}}).$$

Here

$$\boldsymbol{\mu}_D = \begin{bmatrix} \mu_{1D} \\ \mu_{2D} \end{bmatrix}, \ \boldsymbol{\Sigma}_D = \begin{bmatrix} \sigma_{1D}^2 & \rho_D \\ \rho_D & \sigma_{2D}^2 \end{bmatrix}, \ \boldsymbol{\mu}_{\bar{D}} = \begin{bmatrix} \mu_{1\bar{D}} \\ \mu_{2\bar{D}} \end{bmatrix} \text{ and } \boldsymbol{\Sigma}_{\bar{D}} = \begin{bmatrix} \sigma_{1\bar{D}}^2 & \rho_{\bar{D}} \\ \rho_{\bar{D}} & \sigma_{2\bar{D}}^2 \end{bmatrix}.$$

The vector of parameters in this setting is given by

$$\boldsymbol{\theta}' = (p, \mu_{1D}, \mu_{2D}, \mu_{1\bar{D}}, \mu_{2\bar{D}}, \sigma_{1D}^2, \sigma_{2D}^2, \sigma_{1\bar{D}}^2, \sigma_{2\bar{D}}^2, \rho_D, \rho_{\bar{D}}),$$

where $p = P(D = 1)$. It is worth noting that under our model, the conditionally independent assumption is a special case with $\rho_D = \rho_{\bar{D}} = 0$.

The AUC for diagnostic test $j$ under the above setting can be further expressed as $A_j = \Phi(\eta_j)$,
where

$$\eta_j = \frac{\mu_{jD} - \mu_{j\bar{D}}}{\sqrt{\sigma_{jD}^2 + \sigma_{j\bar{D}}^2}}, \text{ for } j = 1, 2,$$

and $\Phi(\cdot)$ is the standard normal distribution function.

If a GS test on the true disease status exists, then $X$ and $Y$ are available. Thus, the ML estimate of $\boldsymbol{\theta}$ can be easily derived, and the interval estimation for $\Delta$ can be obtained using a bootstrap method. We summarize this estimation method in Appendix A. However, if a GS test is not available, then $X$ and $Y$ are missing, and the derivation is not so straightforward. We propose ML-based interval estimation for $\Delta$ using the EM algorithm and bootstrap method under the NGS situation.

**2.1 EM algorithm**

The EM algorithm is a general purpose algorithm to iteratively compute the ML estimates when observed data can be viewed as incomplete data. Let $t_{ji}$ be the observed result of the $j^{th}$ test on the $i^{th}$ subject, $D_i$ be the unobserved disease status of the $i^{th}$ subject, and $p = P(D_i = 1)$. Let $\mathbf{t}_i = (t_{1i}, t_{2i})$, $\mathbf{t} = (\mathbf{t}_1, \ldots, \mathbf{t}_n)$, and $\mathbf{D} = (D_1, \ldots, D_n)$. Recall that $\mathbf{X}$ and $\mathbf{Y}$ follow $N_2(\boldsymbol{\mu}_D, \boldsymbol{\Sigma}_D)$ and $N_2(\boldsymbol{\mu}_{\bar{D}}, \boldsymbol{\Sigma}_{\bar{D}})$, respectively. If $\mathbf{D}$ had been observed, then the complete data log-likelihood function would be given as follows:

$$l^c(\boldsymbol{\theta}|\mathbf{t}, \mathbf{D}) = \sum_{i=1}^{n} [D_i \log(p f_{\mathbf{X}}(\mathbf{t}_i)) + (1 - D_i) \log((1-p) f_{\mathbf{Y}}(\mathbf{t}_i))],$$

where $f_{\mathbf{X}}(\mathbf{t})$ is the density function of $N_2(\boldsymbol{\mu}_D, \boldsymbol{\Sigma}_D)$, and $f_{\mathbf{Y}}(\mathbf{t})$ is the density function of $N_2(\boldsymbol{\mu}_{\bar{D}}, \boldsymbol{\Sigma}_{\bar{D}})$. Let $\boldsymbol{\theta}^{(m)}$ denote the estimate of $\boldsymbol{\theta}$ after the $m^{th}$ iteration of the EM algorithm. The following E-step and M-step are used to find $\boldsymbol{\theta}^{(m+1)}$, an updated estimate of $\boldsymbol{\theta}$.

• **E-step:** The E-step computes the conditional expectation of $l^c(\boldsymbol{\theta})$ under the observed data $\mathbf{t}$ and the current parameter estimate, $\boldsymbol{\theta} = \boldsymbol{\theta}^{(m)}$. That is,

$$E(l^c(\boldsymbol{\theta})|\mathbf{t}, \boldsymbol{\theta} = \boldsymbol{\theta}^m) = \sum_{i=1}^{n} P(D_i = 1|\mathbf{t}_i, \boldsymbol{\theta}^{(m)}) \log(p) f_{\mathbf{X}}(\mathbf{t}_i) + P(D_i = 0|\mathbf{t}_i, \boldsymbol{\theta}^{(m)}) \log(1-p) f_{\mathbf{Y}}(\mathbf{t}_i).$$

If we define $z_{id}^{(m)}$ as

$$z_{id}^{(m)} = P(D_i = d|\mathbf{t}_i, p^{(m)}, \mu_{1D}^{(m)}, \mu_{2D}^{(m)}, \mu_{1\bar{D}}^{(m)}, \mu_{2\bar{D}}^{(m)}, \sigma_{1D}^{2(m)}, \sigma_{2D}^{2(m)}, \rho_D^{(m)}, \sigma_{1\bar{D}}^{2(m)}, \sigma_{2\bar{D}}^{2(m)}, \rho_{\bar{D}}^{(m)}),$$

we can show that

$$z_{i1}^{(m)} = \frac{p^{(m)} f_{\mathbf{X}}^{(m)}(\mathbf{t}_i)}{p^{(m)} f_{\mathbf{X}}^{(m)}(\mathbf{t}_i) + (1-p)^{(m)} f_{\mathbf{Y}}^{(m)}(\mathbf{t}_i)}, \tag{1}$$

$$z_{i0}^{(m)} = \frac{(1-p)^{(m)} f_{\mathbf{Y}}^{(m)}(\mathbf{t}_i)}{p^{(m)} f_{\mathbf{X}}^{(m)}(\mathbf{t}_i) + (1-p)^{(m)} f_{\mathbf{Y}}^{(m)}(\mathbf{t}_i)}, \tag{2}$$

and

$$E(l^c(\boldsymbol{\theta})|\mathbf{t}, \boldsymbol{\theta} = \boldsymbol{\theta}^m) = \sum_{i=1}^{n} z_{i1}^{(m)} \log(p f_{\mathbf{X}}(\mathbf{t}_i)) + z_{i0}^{(m)} \log((1-p) f_{\mathbf{Y}}(\mathbf{t}_i)). \tag{3}$$

• **M-step:** The M-step finds the updated estimate $\boldsymbol{\theta}^{(m+1)}$ for $\boldsymbol{\theta}$ by maximizing $E(l^c(\boldsymbol{\theta})|\mathbf{t}, \boldsymbol{\theta} = \boldsymbol{\theta}^m)$ with respect to $\boldsymbol{\theta}$. The elements of $\boldsymbol{\theta}^{(m+1)}$ are summarized in Appendix B.

The convergent value of $\boldsymbol{\theta}^{(m+1)}$ in the EM algorithm is the ML estimate of $\boldsymbol{\theta}$. Finally, plugging the ML estimate of $\boldsymbol{\theta}$ into $\Delta = A_1 - A_2$, we obtain the ML estimate of $\Delta$, $\hat{\Delta}$.

### 2.2 Bootstrap method

Due to the complicated variance form of $\hat{\Delta}$, we use a bootstrap method to obtain its variance estimate. Then, $\hat{\Delta}$ and its variance estimate are used to construct the confidence interval of the difference in paired AUCs in the absence of a GS test. An equal-tailed $100(1-\alpha)\%$ bootstrap confidence interval for $\Delta = A_1 - A_2$ can be obtained from the following procedure.

**Step 1:** Set initial values for $p$, $\boldsymbol{\mu}_D$, $\boldsymbol{\Sigma}_D$ and $\boldsymbol{\mu}_{\bar{D}}$, $\boldsymbol{\Sigma}_{\bar{D}}$.

**Step 2:** Use the EM algorithm to obtain $\hat{\Delta}$, based on the observed data, $\mathbf{t} = (\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_n)$, .

**Step 3:** Generate $B$ bootstrap samples, $\mathbf{t}^\star = (\mathbf{t}_1^\star, \mathbf{t}_2^\star, \ldots, \mathbf{t}_n^\star)$, from the observed data, $\mathbf{t}$, without replacement, such that each bootstrap sample has a size $n$, where $B = 200$.

**Step 4:** Use the EM algorithm to estimate $\Delta = A_1 - A_2$ for each bootstrap sample. Then, from these B bootstrap estimates of $\Delta$, we can form the sample variance estimate for the variance of $\widehat{\Delta}$, denoted by $\widehat{var}(\hat{\Delta}_{boot})$.

**Step 5:** Use the resulting $\hat{\Delta}$ in Step 2 and $\widehat{var}(\hat{\Delta}_{boot})$ in Step 4 to construct $(1-\alpha)100\%$ confidence interval for $\Delta$ as follows:

$$(\hat{\Delta} - z_{1-\alpha/2}\sqrt{\widehat{var}(\hat{\Delta}_{boot})}, \hat{\Delta} + z_{\alpha/2}\sqrt{\widehat{var}(\hat{\Delta}_{boot})}).$$

## 3. Simulation Studies

Three simulation studies were conducted. The first simulation would evaluate how much efficiency the proposed ML method might lose if the GS information was used in estimation. We compared the coverage probabilities of our proposed ML-based method under the NGS with those of the ML-based method under existence of a GS test. The second simulation would assess the relative performance of our method in comparison of the existing method under the NGS case. We compared the performance of the proposed ML-based method under the NGS with Choi's method, which also does not require the existence of a GS test. The third simulation study would assess the performance of our method for non-normal data. We assessed performance of the proposed ML-based method when test data were skewed.

### 3.1 Simulation study I

We chose the same simulation parameters as those in Li *et al.* [5]. First, we generated a random sample of two test results of $n_1$ diseased subjects, $\mathbf{X}_1, ..., \mathbf{X}_{n_1}$, from $N_2(\boldsymbol{\mu}_D, \boldsymbol{\Sigma}_D)$ and a random sample of two test results of $n - n_1$ non-diseased subjects, $\mathbf{Y}_1, ..., \mathbf{Y}_{n-n_1}$, from $N_2(\boldsymbol{\mu}_{\bar{D}}, \boldsymbol{\Sigma}_{\bar{D}})$. Without loss of generality, the mean and variance vectors of $\mathbf{Y}_{i'}$ - $(\mu_{1\bar{D}}, \mu_{2\bar{D}})$ and $(\sigma^2_{1\bar{D}}, \sigma^2_{2\bar{D}})$ - were fixed at $(0, 0)$ and $(1, 1)$, respectively, and the variance vector of $\mathbf{X}_i$ - $(\sigma^2_{1D}, \sigma^2_{2D})$- was fixed at $(1, 1)$. The true disease prevalence $p$ was chosen to be 0.1, 0.3, or 0.5. We varied the sample size $(n)$ to be 40, 70, 100, 150, 200, or 500. Data were generated under the following three conditions for the ROC curves of the two tests: (i) both diagnostic

tests had large AUCs; (ii) one had large AUC, while the other had a small AUC; and (iii) both had small AUC. Specifically, for achieving conditions (i), (ii), and (iii), we chose the mean vector of $\mathbf{X}_i$, $(\mu_{1D}, \mu_{2D})$, to be $(2.326, 1.812)$, $(2.326, 0.545)$ and $(0.742, 0.545)$, respectively, leading to the value of $\Delta = A_1 - A_2$ being 0.05 in (i), 0.30 in (ii), and 0.05 in (iii). Finally, the correlation coefficients between $X_1$ and $X_2$ and between $Y_1$ and $Y_2$ were chosen. Since $\sigma^2_{1D}, \sigma^2_{2D}, \sigma^2_{1\bar{D}}$ and $\sigma^2_{2\bar{D}}$ were all fixed at 1, $\rho_D$ and $\rho_{\bar{D}}$ were the conditional correlations, respectively, which were set to be 0.5 (medium correlations between two diagnostic tests) or 0.99 (high correlations between two diagnostic tests).

For each specified parameter combination, the data were generated 5,000 times independently. We applied the proposed method to each simulated data set to obtain the 95% confidence interval of $\Delta$. The actual coverage probability was computed by the proportion of the 5,000 simulated confidence intervals that covered $\Delta$, and the expected interval length was computed by the average of the 5,000 confidence intervals.

We display the simulation results in Tables I, II, and III.

Insert Tables I, II and III here

From the results in Tables I, II, and III, we drew the following conclusions.

(1) Under both the GS and NGS situations, the proposed ML-based intervals had the empirical coverage probabilities that were close to the nominal confidence level 95% for most cases and were slightly liberal for some of the smaller sample sizes. In addition, the proposed method performed better when the disease prevalence was 0.5 than when the disease prevalence was 0.1 or 0.3. Also, the empirical coverage probabilities of the proposed method performed better for high correlations between paired tests ($\rho_D = 0.99$ and $\rho_{\bar{D}} = 0.99$) than for low correlation cases.

(2) The coverage probabilities of the proposed ML-based interval under the NGS case were slightly higher than those of the ML-based method under the GS

case. However, the expected interval lengths under the NGS case were larger than those under the GS case and could be much larger when the correlations between two paired tests were 0.5. These results are consistent with current knowledge in the statistical literature about estimation of diagnostic tests in the NGS case.

### 3.2 Simulation study II

To compare the relative performance of the proposed ML-based method against the existing method of Choi's *et al.* [13], we chose the same simulation parameters as in Choi *et al.* [13]. We report the simulation results in Table IV. Note that in Table IV, we directly cited the simulation results for Choi's method from the original paper.

Insert Table IV here

Table IV shows that the proposed ML-based method under the NGS case had smaller bias and slightly better coverage probability than Choi's method under the NGS case.

### 3.3 Simulation study III

To investigate robustness of the proposed ML-based method under the NGS case when data do not follow a normal distribution, we generated data from a bivariate skewed exponential distribution with the difference of the AUCs of 0.05 ($\Delta = 0.05$) using the method described in Liu *et al.*[17]. We display the simulation results in Table V.

Insert Table V here

From Table V, we conclude that when $p = 0.5$ and the correlations between the two diagnostic tests were high, the coverage probabilities of the proposed ML-based method under both the GS and NGS cases were close to the nominal confidence

level of 95%. When $p$ was less than 0.5, coverage probabilities of the proposed ML-based method under the both GS and NGS cases were also close to the nominal level when the sample size was greater than 100, but could be much less than the nominal level when the sample size was small. The expected interval length of the proposed ML-based method under the NGS case was much larger than that of the ML-based method under the GS case.

## 4. Numerical Examples

### 4.1. The study of pancreatic cancer serum biomarkers

We first consider the pancreatic cancer data from a case-control study, reported in Wieand *et al.* [18]. This case-control study included 90 cases with pancreatic cancer and 51 controls who did not have cancer but had pancreatitis. Serum samples from each patient were assayed for CA-125 (a cancer antigen) and CA-19-9 (a carbohydrate antigen), both of which were measured on a continuous positive scale. Although each patient had a disease status based on a GS test, we first treated the disease status of each patient as unknown and used the proposed method to calculate the difference of areas under the ROC curves of CA-125 and CA-19-9 and construct a 95% confidence interval without the use of the GS information. Then, using the GS information on the disease status for each patient, we also derived the confidence interval for the difference of areas under the ROC curves of CA-125 and CA-19-9.

Because the pancreatic cancer data were skewed, we took the log transformation of the data to make the data approximately meet the bivariate normality assumption. The estimated prevalence, $p$, of pancreatic cancer under the NGS was 0.69. Note that the sample proportion of pancreatic cancer was 0.64. Under the NGS case, the resulting 95% confidence interval for the difference between the AUCs of two biomarkers was $(-0.181, 0.324)$. On the other hand, using the GS information, we obtained the resulting 95% confidence interval of $(0.080, 0.295)$. It was noteworthy that the AUC estimates of CA-19-9 and CA-125 under the NGS situation were 0.83 and 0.66, respectively, while the corresponding AUC estimates under the GS

case were 0.88 and 0.68, respectively. Since the confidence interval excluded zero in the GS and contained zero in the NGS, the use of the GS information in this example would lead to a different conclusion from the one obtained without the use of the GS on the disease status. We also noted that the 95% confidence interval of the difference in AUCs in the GS case was much narrower than in the NGS case.

### 4.2. The study of accuracy of magnetic resonance angiography (MRA) readings by two readers

This example, presented in Masaryk *et al.* [19], was a study on atherosclerosis of the carotid arteries. Although each patient in this data set had the GS information available, due to potential error in the gold standard procedure, we used this data set to contrast the results observed without using the GS information with the results observed with using the GS information. In the study, each of two radiologists assessed 65 carotid arteries (left and right) in 36 patients using MRA. Thirty three patients had MRA test results from the left artery, and 32 patients from the right artery. We compared the accuracy of readings between these two radiologists, based on the AUC. In this study, we only used the data of the left artery to estimate the difference between the two corresponding AUCs. Because the values of the data range from $-122$ to 100, we added 150 to each observation to make all values positive. Since the data were skewed, based on visual assessment of the data values, we chose the log transformation to make the data have approximately bivariate normal distributions. Without using the GS information on the disease status, we obtained the estimated prevalence of left artery disease, $p$, to be 0.40. Note that the sample proportion of disease was 0.36. In the NGS case, the resulting 95% confidence interval was $(-0.112, 0.134)$. Using the GS information on the disease status, the 95% confidence interval was $(-0.027, 0.005)$. The AUC estimates of two readers without using the GS information were 0.93 and 0.95, respectively, while the corresponding AUC estimates, when using the GS information, were 0.90 and 0.94, respectively. Although the confidence interval derived for the NGS case is wider than that for the GS case, both the intervals included zero. Hence, there was no strong evidence to indicate that the accuracy of MRA readings obtained from

the two readers was significantly different.

## 5. Discussion and Final Remarks

Under the normality assumption on the diagnostic test results from each diseased group of subjects, and using the EM algorithm in conjunction with the bootstrap method, we proposed a procedure for the construction of confidence intervals for the difference in paired AUCs without the existence of a GS test on the true disease status of a patient. The proposed methods performed well for finite sample sizes in our simulation studies. An R program for computation of the proposed ML-based method is available from the authors upon request.

Our method is based on the percentile bootstrap method. Obuchowski and Lieber [20] assessed the adequacy of various bootstrap confidence intervals for the AUC when test results were continuous and when sample sizes were small, and they found that bootstrap percentile $t$ confidence interval is preferable. To see whether the use of the bootstrap $t$ method could improve the performance of our method, we conducted one additional simulation study. In this simulation study, we generated test results of a diseased and non-diseased subject from a bivariate normal distribution with $\sigma_D = 0.5$ and $\sigma_{\bar{D}} = 0.99$, respectively, and the other parameter estimates shown in Table VI.

Insert Table VI here

From Table VI, we observe that the empirical coverage probabilities of the bootstrap percentile $t$ confidence interval and the proposed bootstrap confidence interval are both close to the nominal confidence level 95% in both the GS and NGS cases. In general, these two bootstrap methods appear to have similar performance in both the GS and NGS cases.

Hui and Zhou [21] reviewed the statistical methods for estimating the diagnostic accuracy of one or more new tests in the absence of a GS test. They pointed out that most of these methods are based on mixture models and assume the conditional independence that the two diagnostic tests are independent, conditional on the

true disease status. Our newly proposed method does not require the conditionally independent assumption. However, we do need to assume the bivariate normality assumption on distributions of test results of a diseased and non-diseased subject.

One may be concerned that the EM algorithm used in this study may not always lead to the global ML estimates. To overcome this problem, Zhou *et al.* [12] suggested randomly perturbing the starting points, or recomputing the ML estimates based on a set of plausible initial values. Thus, we used different starting points for parameters, and found that the parameter estimates always converged to the same values.

When test results are skewed, the performance of the empirical coverage probability of the proposed ML-based method is still robust, unlike the existing interval estimation method of Choi *et al.*, which is sensitive to the departure from the bivariate normality assumption.

## Acknowledgments

We would like to thank anonymous reviewers and an Associate Editor for their constructive comments. We would like to thank Vicki Ding for her helpful comments on our manuscript. Zhou's work was supported by grants from NACC (U01 AG16976) and from the National Institute of Health (R01 EB005829). Xiao-Hua Zhou, Ph.D, is presently a Core Investigator, Research Career Scientist (RCS OS-196), and Biostatistics Unit Director at the Northwest HSR&D Center of Excellence, Department of Veterans Affairs Medical Center, Seattle, WA. The views expressed in this article are those of the authors and do not necessarily represent the views of the Department of Veterans Affairs.

## References


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

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

3. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* 1988; **44**: 837-845.

4. McClish DK. Analyzing a portion of the ROC curve. *Medical Decision Making* 1989; **9**: 190-195.

5. Li CR, Liao CT, Liu JP. On the exact interval estimation for the difference in paired areas under the ROC curves. *Statistics in Medicine* 2008; **27**: 224-242.

6. Weerahandi S. Generalized confidence intervals. *Journal of the American Statistical Association* 1993; **88**: 899-905.

7. Henkelman RM, Kay I, Bronskill MJ. Receiver operator characteristic (ROC) analysis without truth. *Medical Decision Making* 1990; **10**: 24-29.

8. Beiden SV, Campbell G, Meier KL, Wagner RF. On the problem of ROC analysis without truth: the EM algorithm and the information matrix. in *Proceedings of SPIE* 2000; **3981**: 126-134.

9. Enøe C, Georgiadis MP, Johnson WO. Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown. *Preventive Veterinary Medicine* 2000; **45**: 61-81.

10. Dendukuri N, Joseph L. Bayesian approaches to modeling the conditional dependence between multiple dianostic tests. *Biometrics* 2001; **57**: 158-167.

11. Georgiadis MP, Johnson WO, Gardner IA, Singh R. Correlation-adjusted estimation of sensitivity and specificity of two diagnostic tests. *Applied Statistics* 2003; **52**: 63-76.

12. Zhou XH, Castelluccio P, Zhou C. Nonparametric estimation of ROC curves in the absence of a gold standard. *Biometrics* 2005; **61**: 600-609.

13. Choi YK, Johnson WO, Collins MT, Gardner IA. Bayesian inference for receiver operating characteristic curves in the absence of a gold standard. *Journal of Agricultural, Biological, and Environmental Statistics* 2006; **11**: 210-229.

14. Branscum AJ, Johnson WO, Hanson TE, Gardner IA. Bayesian semiparametric ROC curve estimation and disease diagnosis. *Statistics in Medicine* 2008; **27**: 2474-2496.

15. Metz CE. Basic principles of ROC analysis. *Seminars in Nuclear Medicine* 1978; **8**: 283-298.

16. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; **143**: 29-36.

17. Liu JP, Ma MC, Wu CY, Tai JY. Tests of equivalence and non-inferiority for diagnostic accuracy based on the paired areas under ROC curves. *Statistics in Medicine* 2006; **25**: 1219-1238.

18. Wieand S, Gail MH, James BR, James KL. A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika* 1989; **76**: 585-592.

19. Masaryk AM, Ross JS, DiCello MC, Modic MT, Paranandi L, Masaryk TJ. 3DFT MR angiography of the carotid bifurcation: Potential and limitations as a screening examination. *Radiology* 1991; **179**: 797-804.

20. Obuchowski N, Lieber ML. Confidence intervals for the receiver operating characteristic area in studies with small samples. *Acad. Radiol* 1998; **5**: 561-571.

21. Hui SL, Zhou XH. Evaluation of diagnostic tests without gold standards. *Statistical Methods in Medical Research* 1998; **7**: 354-370.

## Appendix A.

Using the bootstrap method, we next derive confidence intervals for $\Delta$ when a GS test on the disease status is available for each patient.

Suppose that $\mathbf{X}_1, \ldots, \mathbf{X}_{n_1}$ and $\mathbf{Y}_1, \ldots, \mathbf{Y}_{n-n_1}$ are two random samples from $N_2(\boldsymbol{\mu}_D, \boldsymbol{\Sigma}_D)$ and $N_2(\boldsymbol{\mu}_{\bar{D}}, \boldsymbol{\Sigma}_{\bar{D}})$, respectively. Then we obtain the following estimators for $(\boldsymbol{\mu}_D, \boldsymbol{\Sigma}_D)$ and $(\boldsymbol{\mu}_{\bar{D}}, \boldsymbol{\Sigma}_{\bar{D}})$:

$$\begin{aligned}(\hat{\boldsymbol{\mu}}_D,\ \hat{\boldsymbol{\Sigma}}_D) &= \left( \frac{1}{n_1}\sum_{i=1}^{n_1} \mathbf{X}_i,\ \frac{1}{n_1-1}\sum_{i=1}^{n_1}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^{\mathrm{T}} \right) \\ &= \left( \begin{bmatrix} \bar{X}_1 \\ \bar{X}_2 \end{bmatrix},\ \frac{1}{n_1-1} \begin{bmatrix} SX_1 & SX_{12} \\ SX_{12} & SX_2 \end{bmatrix} \right)\end{aligned}$$

and

$$\begin{aligned}(\hat{\boldsymbol{\mu}}_{\bar{D}},\ \hat{\boldsymbol{\Sigma}}_{\bar{D}}) &= \left( \frac{1}{n-n_1}\sum_{i'=1}^{n-n_1} \mathbf{Y}_{i'},\ \frac{1}{(n-n_1)-1}\sum_{i'=1}^{n-n_1}(\mathbf{Y}_{i'} - \bar{\mathbf{Y}})(\mathbf{Y}_{i'} - \bar{\mathbf{Y}})^{\mathrm{T}} \right) \\ &= \left( \begin{bmatrix} \bar{Y}_1 \\ \bar{Y}_2 \end{bmatrix},\ \frac{1}{(n-n_1)-1} \begin{bmatrix} SY_1 & SY_{12} \\ SY_{12} & SY_2. \end{bmatrix} \right)\end{aligned}$$

From these estimates, we obtain the ML estimate, $\widehat{\Delta}$, of $\Delta$, the difference in the paired AUCs of the two tests. We use the following procedure to obtain a two-sided $100(1-\alpha)\%$ bootstrap confidence interval for $\Delta$.

**Step 1:** Compute the ML estimate of $\Delta$, $\widehat{\Delta}$, based on the observed data, $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n_1})$ and $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{n-n_1})$.

**Step 2:** Generate $B$ bootstrap random samples, $\mathbf{x}^\star = (\mathbf{x}_1^\star, \mathbf{x}_2^\star, \ldots, \mathbf{x}_{n_1}^\star)$ and $\mathbf{y}^\star = (\mathbf{y}_1^\star, \mathbf{y}_2^\star, \ldots, \mathbf{y}_{n-n_1}^\star)$, with a size of $n$, by sampling with replacement from the observed data, $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n_1})$ and $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{n-n_1})$, where $B = 200$.

**Step 3:** Estimate $\Delta = A_1 - A_2$ for each of the $B$ bootstrap random samples. Then we can compute the sample variance of these $B$ estimates and denote it by $\widehat{var}(\hat{\Delta}_{boot})$.

**Step 4:** Use the resulting $\hat{\Delta}$ in Step 1 and $\widehat{var}(\hat{\Delta}_{boot})$ in Step 3 to construct the $(1-\alpha)100\%$ confidence interval for $\Delta$ as follows:

$$(\hat{\Delta} - z_{1-\alpha/2}\sqrt{\widehat{var}(\hat{\Delta}_{boot})}, \hat{\Delta} + z_{\alpha/2}\sqrt{\widehat{var}(\hat{\Delta}_{boot})}).$$

## Appendix B. The estimators in M-step.

$$\hat{p}^{(m+1)} = \frac{1}{n}\sum_{i=1}^{n} z_{i1}^{(m)},$$

$$\hat{\mu}_{1D}^{(m+1)} = \frac{\sum_{i=1}^{n} z_{i1}^{(m)} t_{i1}}{\sum_{i=1}^{n} z_{i1}^{(m)}},$$

$$\hat{\mu}_{2D}^{(m+1)} = \frac{\sum_{i=1}^{n} z_{i1}^{(m)} t_{i2}}{\sum_{i=1}^{n} z_{i1}^{(m)}},$$

$$\hat{\sigma}_{1D}^{2(m+1)} = \frac{\sum_{i=1}^{n} z_{i1}^{(m)} (t_{i1} - \hat{\mu}_{1D})^2}{\sum_{i=1}^{n} z_{i1}^{(m)}},$$

$$\hat{\sigma}_{2D}^{2(m+1)} = \frac{\sum_{i=1}^{n} z_{i1}^{(m)} (t_{i2} - \hat{\mu}_{2D})^2}{\sum_{i=1}^{n} z_{i1}^{(m)}},$$

$$\hat{\rho}_{D}^{(m+1)} = \frac{\sum_{i=1}^{n} z_{i1}^{(m)} (t_{i1} - \hat{\mu}_{1D})(t_{i2} - \hat{\mu}_{2D})}{\sum_{i=1}^{n} z_{i1}^{(m)}}.$$

Also,

$$\hat{\mu}_{1\bar{D}}^{(m+1)} = \frac{\sum_{i=1}^{n} z_{i0}^{(m)} t_{i1}}{\sum_{i=1}^{n} z_{i0}^{(m)}},$$

$$\hat{\mu}_{2\bar{D}}^{(m+1)} = \frac{\sum_{i=1}^{n} z_{i0}^{(m)} t_{i2}}{\sum_{i=1}^{n} z_{i0}^{(m)}},$$

$$\hat{\sigma}_{1\bar{D}}^{2(m+1)} = \frac{\sum_{i=1}^{n} z_{i0}^{(m)} (t_{i1} - \hat{\mu}_{1\bar{D}})^2}{\sum_{i=1}^{n} z_{i0}^{(m)}},$$

$$\hat{\sigma}_{2\bar{D}}^{2(m+1)} = \frac{\sum_{i=1}^{n} z_{i0}^{(m)} (t_{i2} - \hat{\mu}_{2\bar{D}})^2}{\sum_{i=1}^{n} z_{i0}^{(m)}},$$

$$\hat{\rho}_{\bar{D}}^{(m+1)} = \frac{\sum_{i=1}^{n} z_{i0}^{(m)} (t_{i1} - \hat{\mu}_{1\bar{D}})(t_{i2} - \hat{\mu}_{2\bar{D}})}{\sum_{i=1}^{n} z_{i0}^{(m)}}.$$

**Table I.** The coverage probabilities (CP) and expected lengths (EL) of the 95% confidence interval for the difference in paired areas under the ROC curves when the true disease prevalence ($p$) = 0.1, 0.3 and 0.5, $\Delta = 0.05$, in condition (i).

| | | | $p = 0.1$ | | | | $p = 0.3$ | | | | $p = 0.5$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | GS | | NGS | | GS | | NGS | | GS | | NGS | |
| $\rho_D$ | $\rho_{\bar{D}}$ | $n$ | CP | EL | CP | EL | CP | EL | CP | EL | CP | EL | CP | EL |
| 0.50 | 0.50 | 40 | 0.869 | 0.217 | 0.927 | 0.494 | 0.960 | 0.176 | 0.963 | 0.451 | 0.965 | 0.168 | 0.967 | 0.430 |
| | | 70 | 0.936 | 0.189 | 0.952 | 0.489 | 0.962 | 0.136 | 0.972 | 0.429 | 0.967 | 0.127 | 0.976 | 0.396 |
| | | 100 | 0.944 | 0.164 | 0.960 | 0.488 | 0.963 | 0.116 | 0.977 | 0.401 | 0.968 | 0.107 | 0.987 | 0.365 |
| | | 150 | 0.947 | 0.135 | 0.970 | 0.466 | 0.965 | 0.094 | 0.979 | 0.345 | 0.969 | 0.088 | 0.982 | 0.363 |
| | | 200 | 0.957 | 0.122 | 0.974 | 0.444 | 0.968 | 0.083 | 0.983 | 0.316 | 0.970 | 0.076 | 0.983 | 0.289 |
| | | 500 | 0.958 | 0.079 | 0.978 | 0.332 | 0.971 | 0.053 | 0.981 | 0.190 | 0.972 | 0.048 | 0.981 | 0.165 |
| 0.50 | 0.99 | 40 | 0.832 | 0.208 | 0.931 | 0.219 | 0.931 | 0.151 | 0.932 | 0.201 | 0.943 | 0.125 | 0.945 | 0.180 |
| | | 70 | 0.906 | 0.170 | 0.940 | 0.195 | 0.943 | 0.117 | 0.943 | 0.148 | 0.945 | 0.095 | 0.946 | 0.130 |
| | | 100 | 0.940 | 0.153 | 0.945 | 0.188 | 0.944 | 0.100 | 0.952 | 0.120 | 0.946 | 0.080 | 0.947 | 0.106 |
| | | 150 | 0.945 | 0.129 | 0.947 | 0.161 | 0.946 | 0.082 | 0.953 | 0.095 | 0.948 | 0.067 | 0.947 | 0.083 |
| | | 200 | 0.948 | 0.113 | 0.954 | 0.135 | 0.949 | 0.071 | 0.947 | 0.079 | 0.950 | 0.058 | 0.951 | 0.070 |
| | | 500 | 0.950 | 0.076 | 0.951 | 0.081 | 0.951 | 0.045 | 0.955 | 0.049 | 0.952 | 0.037 | 0.959 | 0.043 |
| 0.99 | 0.99 | 40 | 0.881 | 0.102 | 0.889 | 0.106 | 0.895 | 0.072 | 0.898 | 0.084 | 0.908 | 0.066 | 0.908 | 0.079 |
| | | 70 | 0.916 | 0.080 | 0.917 | 0.091 | 0.932 | 0.056 | 0.936 | 0.066 | 0.934 | 0.052 | 0.938 | 0.059 |
| | | 100 | 0.918 | 0.069 | 0.920 | 0.084 | 0.935 | 0.047 | 0.937 | 0.055 | 0.945 | 0.044 | 0.948 | 0.050 |
| | | 150 | 0.942 | 0.058 | 0.945 | 0.071 | 0.945 | 0.039 | 0.947 | 0.044 | 0.950 | 0.036 | 0.950 | 0.040 |
| | | 200 | 0.945 | 0.051 | 0.949 | 0.062 | 0.948 | 0.034 | 0.950 | 0.038 | 0.950 | 0.032 | 0.951 | 0.034 |
| | | 500 | 0.950 | 0.033 | 0.953 | 0.038 | 0.951 | 0.022 | 0.954 | 0.023 | 0.953 | 0.020 | 0.954 | 0.021 |

*Note*: the value of $\Delta(= A_1 - A_2)$ is fixed at 0.05(=0.95-0.90) in condition (i). $A_1, A_2$: the paired areas under the ROC curves. $n$: the sample size of subjects. $\rho_D$: the correlation between the paired test results of diseased subjects. $\rho_{\bar{D}}$: the correlation between the paired test results of non-diseased subjects.

**Table II.** The coverage probabilities (CP) and expected lengths (EL) of the 95% confidence interval for the difference in paired areas under the ROC curves when the true disease prevalence ($p$) = 0.1, 0.3 and 0.5, $\Delta = 0.30$, in condition (ii).

| | | | $p = 0.1$ | | | | $p = 0.3$ | | | | $p = 0.5$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | GS | | NGS | | GS | | NGS | | GS | | NGS | |
| $\rho_D$ | $\rho_{\bar{D}}$ | $n$ | CP | EL | CP | EL | CP | EL | CP | EL | CP | EL | CP | EL |
| 0.50 | 0.50 | 40 | 0.907 | 0.438 | 0.918 | 0.619 | 0.951 | 0.320 | 0.944 | 0.613 | 0.952 | 0.298 | 0.954 | 0.589 |
| | | 70 | 0.935 | 0.352 | 0.940 | 0.608 | 0.953 | 0.245 | 0.963 | 0.553 | 0.956 | 0.227 | 0.966 | 0.530 |
| | | 100 | 0.947 | 0.302 | 0.949 | 0.607 | 0.956 | 0.206 | 0.967 | 0.513 | 0.959 | 0.190 | 0.971 | 0.490 |
| | | 150 | 0.953 | 0.250 | 0.956 | 0.578 | 0.958 | 0.169 | 0.973 | 0.453 | 0.960 | 0.155 | 0.977 | 0.427 |
| | | 200 | 0.955 | 0.219 | 0.963 | 0.555 | 0.959 | 0.147 | 0.974 | 0.404 | 0.962 | 0.135 | 0.977 | 0.377 |
| | | 500 | 0.961 | 0.141 | 0.967 | 0.420 | 0.961 | 0.093 | 0.975 | 0.243 | 0.964 | 0.085 | 0.977 | 0.215 |
| 0.50 | 0.99 | 40 | 0.901 | 0.426 | 0.930 | 0.473 | 0.939 | 0.299 | 0.940 | 0.359 | 0.943 | 0.268 | 0.944 | 0.326 |
| | | 70 | 0.931 | 0.342 | 0.932 | 0.391 | 0.941 | 0.229 | 0.944 | 0.268 | 0.944 | 0.204 | 0.945 | 0.246 |
| | | 100 | 0.944 | 0.293 | 0.946 | 0.352 | 0.945 | 0.193 | 0.947 | 0.223 | 0.949 | 0.172 | 0.956 | 0.205 |
| | | 150 | 0.945 | 0.244 | 0.947 | 0.285 | 0.947 | 0.158 | 0.949 | 0.181 | 0.951 | 0.140 | 0.956 | 0.166 |
| | | 200 | 0.949 | 0.214 | 0.956 | 0.245 | 0.950 | 0.138 | 0.957 | 0.156 | 0.953 | 0.122 | 0.960 | 0.143 |
| | | 500 | 0.952 | 0.137 | 0.960 | 0.154 | 0.959 | 0.087 | 0.965 | 0.097 | 0.964 | 0.077 | 0.969 | 0.088 |
| 0.99 | 0.99 | 40 | 0.941 | 0.370 | 0.944 | 0.384 | 0.944 | 0.257 | 0.944 | 0.281 | 0.945 | 0.237 | 0.945 | 0.254 |
| | | 70 | 0.943 | 0.284 | 0.946 | 0.304 | 0.945 | 0.194 | 0.947 | 0.205 | 0.946 | 0.180 | 0.947 | 0.187 |
| | | 100 | 0.945 | 0.241 | 0.950 | 0.268 | 0.946 | 0.163 | 0.950 | 0.169 | 0.948 | 0.149 | 0.953 | 0.154 |
| | | 150 | 0.946 | 0.199 | 0.951 | 0.216 | 0.947 | 0.133 | 0.953 | 0.136 | 0.948 | 0.122 | 0.954 | 0.125 |
| | | 200 | 0.950 | 0.173 | 0.954 | 0.184 | 0.950 | 0.115 | 0.957 | 0.117 | 0.953 | 0.106 | 0.958 | 0.107 |
| | | 500 | 0.952 | 0.111 | 0.958 | 0.113 | 0.952 | 0.073 | 0.958 | 0.073 | 0.954 | 0.067 | 0.960 | 0.067 |

*Note*: the value of $\Delta(= A_1 - A_2)$ is fixed at 0.30(=0.95-0.65) in condition (ii). $A_1, A_2$: the paired areas under the ROC curves. $n$: the sample size of subjects. $\rho_D$: the correlation between the paired test results of diseased subjects. $\rho_{\bar{D}}$: the correlation between the paired test results of non-diseased subjects.

**Table III.** The coverage probabilities (CP) and expected lengths (EL) of the 95% confidence interval for the difference in paired areas under the ROC curves when the true disease prevalence ($p$) $= 0.1, 0.3$ and $0.5$, $\Delta = 0.05$, in condition (iii).

| | | | $p = 0.1$ | | | | $p = 0.3$ | | | | $p = 0.5$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | GS | | NGS | | GS | | NGS | | GS | | NGS | |
| $\rho_D$ | $\rho_{\bar{D}}$ | $n$ | CP | EL | CP | EL | CP | EL | CP | EL | CP | EL | CP | EL |
| 0.50 | 0.50 | 40 | 0.916 | 0.452 | 0.962 | 0.793 | 0.968 | 0.342 | 0.977 | 0.837 | 0.971 | 0.319 | 0.978 | 0.841 |
| | | 70 | 0.965 | 0.368 | 0.966 | 0.793 | 0.968 | 0.263 | 0.977 | 0.820 | 0.972 | 0.244 | 0.978 | 0.814 |
| | | 100 | 0.966 | 0.319 | 0.968 | 0.787 | 0.969 | 0.222 | 0.978 | 0.806 | 0.973 | 0.205 | 0.979 | 0.802 |
| | | 150 | 0.968 | 0.267 | 0.969 | 0.772 | 0.970 | 0.182 | 0.978 | 0.789 | 0.974 | 0.168 | 0.980 | 0.785 |
| | | 200 | 0.969 | 0.234 | 0.971 | 0.768 | 0.972 | 0.159 | 0.979 | 0.770 | 0.976 | 0.145 | 0.980 | 0.769 |
| | | 500 | 0.972 | 0.152 | 0.974 | 0.768 | 0.975 | 0.101 | 0.979 | 0.763 | 0.977 | 0.092 | 0.981 | 0.717 |
| 0.50 | 0.99 | 40 | 0.901 | 0.408 | 0.953 | 0.485 | 0.954 | 0.283 | 0.955 | 0.383 | 0.963 | 0.229 | 0.964 | 0.307 |
| | | 70 | 0.958 | 0.341 | 0.959 | 0.435 | 0.962 | 0.219 | 0.966 | 0.276 | 0.965 | 0.175 | 0.966 | 0.213 |
| | | 100 | 0.958 | 0.297 | 0.965 | 0.400 | 0.963 | 0.185 | 0.966 | 0.219 | 0.966 | 0.147 | 0.967 | 0.170 |
| | | 150 | 0.964 | 0.251 | 0.966 | 0.330 | 0.966 | 0.152 | 0.966 | 0.169 | 0.967 | 0.121 | 0.967 | 0.133 |
| | | 200 | 0.966 | 0.221 | 0.969 | 0.277 | 0.967 | 0.132 | 0.967 | 0.143 | 0.967 | 0.104 | 0.968 | 0.113 |
| | | 500 | 0.967 | 0.144 | 0.967 | 0.275 | 0.967 | 0.085 | 0.968 | 0.143 | 0.968 | 0.066 | 0.968 | 0.070 |
| 0.99 | 0.99 | 40 | 0.871 | 0.078 | 0.946 | 0.113 | 0.952 | 0.059 | 0.956 | 0.112 | 0.952 | 0.054 | 0.958 | 0.114 |
| | | 70 | 0.921 | 0.063 | 0.948 | 0.109 | 0.953 | 0.044 | 0.957 | 0.107 | 0.953 | 0.041 | 0.958 | 0.106 |
| | | 100 | 0.944 | 0.053 | 0.949 | 0.108 | 0.956 | 0.037 | 0.957 | 0.104 | 0.957 | 0.034 | 0.959 | 0.102 |
| | | 150 | 0.954 | 0.045 | 0.954 | 0.106 | 0.957 | 0.030 | 0.957 | 0.102 | 0.958 | 0.028 | 0.959 | 0.098 |
| | | 200 | 0.954 | 0.039 | 0.956 | 0.105 | 0.958 | 0.026 | 0.959 | 0.097 | 0.963 | 0.024 | 0.960 | 0.095 |
| | | 500 | 0.956 | 0.025 | 0.958 | 0.105 | 0.959 | 0.017 | 0.960 | 0.088 | 0.967 | 0.015 | 0.967 | 0.082 |

*Note*: the value of $\Delta(= A_1 - A_2)$ is fixed at 0.05(=0.70-0.65) in condition (iii). $A_1, A_2$: the paired areas under the ROC curves. $n$: the sample size of subjects. $\rho_D$: the correlation between the paired test results of diseased subjects. $\rho_{\bar{D}}$: the correlation between the paired test results of non-diseased subjects.

**Table IV.** The simulation comparison of difference in AUCs of the Choi's method and proposed method.

| | | | | | Choi's Method | | | | | | Proposed ML-based Method | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | GS | | | NGS | | | GS | | | NGS | | |
| $\sigma_D^2$ | $\mu_D$ | $\rho_D$ | $\rho_{\bar{D}}$ | True | $\hat{\Delta}$ | CP | EL | $\hat{\Delta}$ | CP | EL | $\hat{\Delta}$ | CP | EL | $\hat{\Delta}$ | CP | EL |
| (2,2) | (4,3) | 0.0 | 0.5 | 0.031 | 0.031 | 0.97 | 0.053 | 0.034 | 0.94 | 0.063 | 0.031 | 0.94 | 0.048 | 0.031 | 0.94 | 0.061 |
| (2,2) | (4,3) | 0.5 | 0.5 | 0.031 | 0.032 | 0.95 | 0.051 | 0.033 | 0.95 | 0.065 | 0.031 | 0.94 | 0.046 | 0.031 | 0.95 | 0.065 |
| (2,2) | (3,2) | 0.5 | 0.5 | 0.083 | 0.084 | 0.93 | 0.098 | 0.091 | 0.93 | 0.171 | 0.082 | 0.94 | 0.093 | 0.083 | 0.97 | 0.192 |
| (2,2) | (3,2) | 0.9 | 0.9 | 0.083 | 0.082 | 0.95 | 0.084 | 0.079 | 0.96 | 0.105 | 0.083 | 0.94 | 0.081 | 0.084 | 0.95 | 0.115 |

Note: we fixed $\mu_{\bar{D}} = (0, 0)$ and $\sigma_{\bar{D}}^2 = (1, 1)$; $\hat{\Delta}(= \hat{A}_1 - \hat{A}_2)$

**Table V.** The coverage probabilities (CP) and expected lengths (EL) of the 95% confidence interval for skewed data when the true disease prevalence $(p) = 0.1, 0.3$ and 0.5.

| | | | | | Proposed ML-based Method | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | GS | | NGS | |
| $\Delta$ | $p$ | $\rho_D$ | $\rho_{\bar{D}}$ | $n$ | CP | EL | CP | EL |
| 0.05 | 0.1 | 0.99 | 0.99 | 40 | 0.911 | 0.095 | 0.932 | 0.149 |
| | | | | 70 | 0.939 | 0.075 | 0.937 | 0.138 |
| | | | | 100 | 0.940 | 0.064 | 0.945 | 0.132 |
| | | | | 150 | 0.942 | 0.053 | 0.947 | 0.122 |
| | | | | 200 | 0.948 | 0.046 | 0.949 | 0.110 |
| | | | | 500 | 0.951 | 0.030 | 0.954 | 0.054 |
| 0.05 | 0.3 | 0.99 | 0.99 | 40 | 0.925 | 0.075 | 0.933 | 0.131 |
| | | | | 70 | 0.940 | 0.058 | 0.947 | 0.120 |
| | | | | 100 | 0.942 | 0.049 | 0.947 | 0.115 |
| | | | | 150 | 0.943 | 0.040 | 0.950 | 0.107 |
| | | | | 200 | 0.950 | 0.035 | 0.951 | 0.099 |
| | | | | 500 | 0.953 | 0.022 | 0.956 | 0.066 |
| 0.05 | 0.5 | 0.99 | 0.99 | 40 | 0.939 | 0.072 | 0.946 | 0.123 |
| | | | | 70 | 0.941 | 0.056 | 0.948 | 0.109 |
| | | | | 100 | 0.943 | 0.047 | 0.950 | 0.099 |
| | | | | 150 | 0.944 | 0.039 | 0.950 | 0.086 |
| | | | | 200 | 0.951 | 0.031 | 0.952 | 0.076 |
| | | | | 500 | 0.964 | 0.022 | 0.967 | 0.041 |

*Note*: $\Delta$: the difference in paired areas under the ROC curves. $n$: the sample size of subjects. $\rho_D$: the correlation between the paired test results of diseased subjects. $\rho_{\bar{D}}$: the correlation between the paired test results of non-diseased subjects.

**Table VI.** The coverage probabilities (CP) and expected lengths (EL) of the 95% bootstrap percentile $t$ confidence interval and proposed confidence interval for the difference in paired areas under the ROC curves when the true disease prevalence $(p) = 0.1, 0.3$ and $0.5$, $\rho_D = 0.50$, $\rho_{\bar{D}} = 0.99$, $\Delta = 0.05$, in condition (i).

| | | | | Bootstrap Percentile $t$ CI | | | | Proposed ML-based Method CI | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | GS | | NGS | | GS | | NGS | |
| $p$ | $\rho_D$ | $\rho_{\bar{D}}$ | $n$ | CP | EL | CP | EL | CP | EL | CP | EL |
| 0.1 | 0.50 | 0.99 | 40 | 0.846 | 0.198 | 0.905 | 0.204 | 0.832 | 0.208 | 0.931 | 0.219 |
| | | | 70 | 0.907 | 0.164 | 0.909 | 0.188 | 0.906 | 0.170 | 0.940 | 0.195 |
| | | | 100 | 0.941 | 0.150 | 0.945 | 0.183 | 0.940 | 0.153 | 0.945 | 0.188 |
| | | | 150 | 0.946 | 0.127 | 0.947 | 0.156 | 0.945 | 0.129 | 0.947 | 0.161 |
| | | | 200 | 0.948 | 0.110 | 0.949 | 0.133 | 0.948 | 0.113 | 0.954 | 0.135 |
| | | | 500 | 0.949 | 0.074 | 0.949 | 0.079 | 0.950 | 0.076 | 0.952 | 0.081 |
| 0.3 | 0.50 | 0.99 | 40 | 0.932 | 0.147 | 0.932 | 0.192 | 0.931 | 0.151 | 0.933 | 0.201 |
| | | | 70 | 0.944 | 0.115 | 0.952 | 0.146 | 0.943 | 0.117 | 0.943 | 0.148 |
| | | | 100 | 0.946 | 0.097 | 0.952 | 0.117 | 0.944 | 0.100 | 0.952 | 0.120 |
| | | | 150 | 0.948 | 0.080 | 0.953 | 0.093 | 0.946 | 0.082 | 0.953 | 0.095 |
| | | | 200 | 0.950 | 0.069 | 0.954 | 0.078 | 0.949 | 0.071 | 0.947 | 0.079 |
| | | | 500 | 0.952 | 0.045 | 0.958 | 0.048 | 0.951 | 0.045 | 0.955 | 0.049 |
| 0.5 | 0.50 | 0.99 | 40 | 0.944 | 0.122 | 0.945 | 0.172 | 0.943 | 0.125 | 0.945 | 0.180 |
| | | | 70 | 0.946 | 0.093 | 0.958 | 0.126 | 0.945 | 0.095 | 0.946 | 0.130 |
| | | | 100 | 0.951 | 0.078 | 0.961 | 0.102 | 0.946 | 0.080 | 0.947 | 0.106 |
| | | | 150 | 0.952 | 0.066 | 0.962 | 0.079 | 0.948 | 0.067 | 0.947 | 0.083 |
| | | | 200 | 0.954 | 0.057 | 0.963 | 0.067 | 0.950 | 0.058 | 0.952 | 0.070 |
| | | | 500 | 0.955 | 0.036 | 0.964 | 0.042 | 0.952 | 0.037 | 0.959 | 0.043 |

*Note*: the value of $\Delta(= A_1 - A_2)$ is fixed at 0.05(=0.95-0.90) in condition (i). $A_1, A_2$: the paired areas under the ROC curves. $n$: the sample size of subjects. $\rho_D$: the correlation between the paired test results of diseased subjects. $\rho_{\bar{D}}$: the correlation between the paired test results of non-diseased subjects.