Comparison of non-parametric confidence intervals for the area under the ROC curve of a continuous-scale diagnostic test

Gengsheng Qin and Lejla Hotilovac

Stat Methods Med Res 2008; 17; 207
DOI: 10.1177/0962280207087173

Statistical Methods in Medical Research 2008; **17**: 207–221

Comparison of non-parametric confidence intervals for the area under the ROC curve of a continuous-scale diagnostic test

Gengsheng Qin, Lejla Hotilovac Department of Mathematics and Statistics, Georgia State University, 30 Pryor Street, Atlanta, GA 30303, USA

The accuracy of a diagnostic test with continuous-scale results is of high importance in clinical medicine. It is often summarised by the area under the ROC curve (AUC). In this article, we discuss and compare nine non-parametric confidence intervals of the AUC for a continuous-scale diagnostic test. Simulation studies are conducted to evaluate the relative performance of the confidence intervals for the AUC in terms of coverage probability and average interval length. A real example is used to illustrate the application of the recommended methods.

1 Introduction

One of the main goals of a diagnostic test is to distinguish the diseased from non-diseased patients. The accuracy of a binary-scale diagnostic test can be measured by its sensitivity and specificity, which are defined as the probabilities of the test correctly identifying the diseased and non-diseased subjects respectively. When the response of a diagnostic test is continuous, we have to choose a cut-off point for the positive result in order to compute the sensitivity and specificity of the test. As the cut-off point changes, specificity and sensitivity vary inversely to each other. The Receiver Operating Characteristic (ROC) curve, denoted by $R(p)$, is a plot of sensitivity against 1-specificity as the cut-off point runs through the whole range of possible test values.

The ROC curve of a diagnostic test best represents the relationship between specificity and 1-sensitivity among all cut-off points of the test.[1,2] It was derived from statistical decision theory and originally developed in the context of electronic signal detection.[3] It has been used in medical imaging and radiology, psychiatry, non-destructive testing and manufacturing inspection systems.[4–6] Recent applications of ROC curve include assessment of the effectiveness of continuous diagnostic markers in distinguishing between diseased and non-diseased individuals.

The area under the ROC curve (AUC), defined as $\delta = \int_0^1 R(p)\mathrm{d}p$, is the most popular global summary measure for a diagnostic test. It indicates the overall performance of a diagnostic test in terms of its accuracy at various diagnostic thresholds used to discriminate disease cases and non-disease cases. Let Y and X be the results of a continuous-scale test for a diseased and a non-diseased subject, respectively. Bamber[7] showed that the

Address for correspondence: Gengsheng Qin, Department of Mathematics and Statistics, Georgia State University, 30 Pryor Street, Atlanta, GA 30303, USA. E-mail: gqin@gsu.edu

Los Angeles, London, New Delhi and Singapore
10.1177/0962280207087173

AUC $\delta = P(Y \geq X)$. It can be interpreted as the probability that, in a randomly selected pair of diseased and non-diseased subjects, the test value of the diseased subject is higher than or equal to that of the non-diseased subject. In a more general context, Wolfe and Hogg[8] recommended the use of AUC as a general measure for the differences between two distributions.

One important statistical problem in the ROC study is the interval estimation of the AUC for a continuous-scale diagnostic test. Parametric methods[9–15] have been proposed for construction of confidence intervals of the AUC. For example, when the test results follow a binormal distribution, the AUC can be explicitly expressed in terms of means and standard deviations. So it can be estimated directly by substituting sample means and standard errors. The variance estimator can also be obtained via the delta method.[9] Metz *et al.*[10] proposed algorithms for fitting binormal ROC curves to continuously-distributed data and estimating the AUC based on the binormal ROC curve parameters, with large-sample standard errors and confidence intervals. In these studies, researchers considered different parametric distributions like exponential distributions,[11] gamma distributions[12] and skew-normal distributions,[13] AUC values between 0.50 (corresponding to a useless diagnostic test) and 1.00 (corresponding to a perfect diagnostic test),[14,15] and sample sizes between 10 and 70 for both diseased and non-diseased subjects.[15] However, parametric methods for the inferences on AUC may be sensitive to the model assumptions and can only provide a limited range of distributional forms for the test results from 'diseased' and 'non-diseased' populations. Therefore, many non-parametric approaches have been proposed for the inference on AUC.[7,9,16–19] Currently, the most popular non-parametric intervals for the AUC include Mann–Whitney statistic-based intervals and DeLong's non-parametric interval.[17]

Whether the AUC is estimated parametrically or non-parametrically, confidence intervals for the AUC are usually obtained by using the normal approximation to the distribution of the estimators. Obuchowski and Lieber[15] performed a simulation study to evaluate the coverage of 95% normal approximation-based intervals, bootstrap percentile, bootstrap-*t*, bootstrap bias-corrected accelerated intervals and confidence intervals with the Student *t* distribution for AUC of moderate (0.80) and high (0.95) accuracy. They found that the asymptotic methods do not provide adequate coverage for small samples; for AUC values of high accuracy, the sample size must be large (more than 200) for the asymptotic methods to be applicable. Instead, they recommended using one of three bootstrap methods (bootstrap percentile, bootstrap-*t* or bootstrap bias-corrected accelerated method) depending on the estimation approach (parametric versus non-parametric) and AUC (moderate versus high). They concluded that there was not a single best alternative for constructing confidence intervals for a single AUC for small samples. Recently, Qin and Zhou[19] proposed an empirical likelihood (EL)-based interval for a single AUC. The EL interval has nice theoretical properties and outperforms the existing normal approximation-based intervals, bootstrap percentile and bootstrap-*t* intervals. In this article, we focus on comparison of non-parametric confidence intervals for the AUC. We will discuss and compare a Mann–Whitney statistic-based interval, a logit transformation-based interval, a non-parametric interval by DeLong *et al.*,[17] an EL-based interval,[19] and five bootstrap-based confidence intervals for the AUC. Extensive simulation studies are conducted to evaluate the relative performance of these confidence intervals in terms of coverage probability and average interval length.

The article is organised as follows. In Section 2, we present nine non-parametric confidence intervals for the area under the ROC curve. In Section 3, we conduct simulation studies to evaluate the relative performance of these confidence intervals. In Section 4, we illustrate the application of the recommended methods in a real data set. Finally, we present discussion and conclusion in Section 5.

2 Non-parametric intervals for the AUC

Let F and G be the distribution functions of X and Y, respectively. For a fixed value of specificity at $(1-p)$, the corresponding sensitivity of the test is $R(p) = 1 - G(F^{-1}(1-p))$, where $F^{-1}(\cdot)$ is the inverse function of $F(\cdot)$. Let $X_1, \dots, X_m$ be the test results of a random sample of non-diseased subjects and $Y_1, \dots, Y_n$ be the test results of a random sample of diseased subjects. Our goal is to construct confidence intervals for the AUC δ.

2.1 Mann–Whitney and logit-transformation-based confidence intervals for the AUC

The simplest non-parametric estimator for the AUC is the well-known Mann–Whitney two-sample rank statistic, defined by

$$\widehat{\delta} = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} I(Y_j \geq X_i).$$

This estimator is an unbiased estimator for the AUC, which equals the trapezoidal area under the empirical ROC curve.[7] Sen[18] shown that

$$\left(\frac{mn}{m+n}\right)^{1/2} \frac{\widehat{\delta} - \delta}{S} \xrightarrow{\mathscr{L}} N(0,1),$$

where

$$S = \left(\frac{mS_{01}^2 + nS_{10}^2}{m+n}\right)^{1/2},$$

$$S_{10}^2 = \frac{1}{(m-1)n^2}\left[\sum_{i=1}^{m}(R_i - i)^2 - m\left(\bar{R} - \frac{m+1}{2}\right)^2\right],$$

$$S_{01}^2 = \frac{1}{(n-1)m^2}\left[\sum_{j=1}^{n}(S_j - j)^2 - n\left(\bar{S} - \frac{n+1}{2}\right)^2\right],$$

$$\bar{R} = \frac{1}{m}\sum_{i=1}^{m} R_i, \qquad \bar{S} = \frac{1}{n}\sum_{j=1}^{n} S_j.$$

Here, R_i is the rank of $X_{(i)}$ (the i-th ordered value among X_i's) in the combined sample of X_i's and Y_j's, S_j is the rank of $Y_{(j)}$ (the j-th ordered value among Y_j's) in the combined sample of X_i's and Y_j's, and $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-th quantile of standard normal distribution.

Using above asymptotic normality, we can construct a confidence interval for the AUC (MW interval):

$$(l_1, u_1) = \widehat{\delta} \mp z_{1-\alpha/2} \left(\frac{(m+n)S^2}{mn} \right)^{1/2}. \tag{1}$$

Although MW interval has asymptotically correct coverage probability, it suffers from low coverage accuracy for high values of AUC (e.g. 0.90, 0.95). AUC values of high accuracy are of interest in some biomedical studies. For example, Rao *et al.*[20] used ROC analysis to evaluate the diagnostic accuracy of serum troponin T (a marker) measurement after myocardial infarction in identifying left ventricular ejection fraction. They reported that the estimated AUC of serum troponin T concentration is 0.9773, and a 95% confidence interval for the AUC is (0.9409, 1.0136).

Since the AUC is restricted to [0,1], Pepe[2] has argued that an asymmetric confidence interval within (0,1) should be preferred. Using a logistic transformation, the lower and upper limits of $(1-\alpha)$-th confidence interval for $\text{logit}(\delta) = \log(\delta/(1-\delta))$ are

$$\text{LL} = \log \frac{\widehat{\delta}}{1-\widehat{\delta}} - z_{1-\alpha/2} \frac{\sqrt{\text{var}(\widehat{\delta})}}{\widehat{\delta}(1-\widehat{\delta})}, \qquad \text{UL} = \log \frac{\widehat{\delta}}{1-\widehat{\delta}} + z_{1-\alpha/2} \frac{\sqrt{\text{var}(\widehat{\delta})}}{\widehat{\delta}(1-\widehat{\delta})},$$

respectively. Here we take $\text{var}(\widehat{\delta}) = (m+n)S^2/mn$. Therefore, the $(1-\alpha)$-th logit-transformation (LT)-based confidence interval for the AUC is

$$(l_2, u_2) = \left(\frac{\exp(LL)}{1+\exp(LL)}, \frac{\exp(UL)}{1+\exp(UL)} \right).$$

The LT interval has good small sample performance, but it has two drawbacks. First, $\text{logit}(\widehat{\delta}) = \log\left(\widehat{\delta}/(1-\widehat{\delta})\right)$ is an unstable estimator for $\text{logit}(\delta)$ when $\widehat{\delta}$ is close to one (i.e. the test has high accuracy). It is possible to have larger variance for an unstable estimator. Our simulation study in this article shows that the LT interval is slightly conservative. It has slightly longer interval length than its competitors such as EL-based interval. Second, the method breaks down when $\widehat{\delta}$ equals one.

2.2 DeLong's non-parametric interval for the AUC

DeLong *et al.*[17] developed a fully non-parametric approach for construction of confidence interval of the AUC based on the theory of U-statistics. Let

$$D_{10}(X_i) = \frac{1}{n}\sum_{j=1}^{n} I(Y_j \geq X_i), \qquad D_{01}(Y_j) = \frac{1}{m}\sum_{i=1}^{m} I(Y_j \geq X_i),$$

$$\mathrm{Var}_D(\widehat{\delta}) = \frac{1}{m(m-1)}\sum_{i=1}^{m}\left(D_{10}(X_i) - \widehat{\delta}\right)^2 + \frac{1}{n(n-1)}\sum_{j=1}^{n}\left(D_{01}(Y_j) - \widehat{\delta}\right)^2.$$

DeLong *et al.*[17] shown that

$$\frac{\widehat{\delta} - \delta}{\mathrm{Var}_D^{1/2}(\widehat{\delta})} \xrightarrow{\mathscr{L}} N(0, 1).$$

Therefore, $(1-\alpha)$-th DeLong's confidence interval for the AUC can be constructed as follows:

$$(l_3, u_3) = \widehat{\delta} \mp z_{1-\alpha/2}\mathrm{Var}_D^{1/2}(\widehat{\delta}).$$

DeLong's interval is simple and easy to use. A computer program written in the SAS language is also available from their website. Their approach has become the standard way for calculating confidence interval for a single AUC.

2.3 Empirical likelihood-based interval for the AUC

For test value Y from a diseased subject, Pepe and Cai[21] defined the placement value as

$$U = 1 - F(Y)$$

It is evident that

$$E(1-U) = E(F(Y)) = P(Y \geq X) = \delta.$$

Based on this relationship between the AUC and the placement value U, Qin and Zhou[19] proposed an empirical likelihood approach for the inference of the AUC. They defined the empirical log-likelihood ratio for the AUC as

$$l(\delta) = -2\log R(\delta) = 2\sum_{j=1}^{n}\log\{1 + \lambda(1 - \widehat{U}_j - \delta)\}, \tag{2}$$

where λ is the solution of

$$\frac{1}{n}\sum_{j=1}^{n}\frac{1 - \widehat{U}_j - \delta}{1 + \lambda(1 - \widehat{U}_j - \delta)} = 0, \tag{3}$$

and $\widehat{U}_j = 1 - \widehat{F}(Y_j)$, $j = 1, 2, \ldots, n$, $\widehat{F}$ is the empirical distribution of F.

Qin and Zhou[19] proved that the limiting distribution of $l(\delta)$ is a scaled chi-square distribution with one degree of freedom. That is,

$$r(\delta)l(\delta) \xrightarrow{\mathscr{L}} \chi_1^2, \tag{4}$$

where the scale constant $r(\delta)$ is

$$r(\delta) = \frac{m}{m+n} \frac{\sum_{j=1}^{n}(1-\widehat{U}_j-\delta)^2}{nS^2}.$$

So, an EL-based confidence intervals for the AUC can be constructed as follows:

$$(l_4, u_4) = \{\delta : r(\widehat{\delta})l(\delta) \leq \chi_1^2(1-\alpha)\}, \tag{5}$$

where $\chi_1^2(1-\alpha)$ is the $(1-\alpha)$-*th* quantile of the chi-square distribution χ_1^2. (l_4, u_4) is an approximate confidence intervals for the AUC with asymptotically correct coverage probability $1-\alpha$. This EL-based interval has good coverage accuracy.

2.4 Bootstrap intervals for the AUC

In this section, we introduce five bootstrap confidence intervals for the AUC. In order to calculate a bootstrap estimate for δ, we need to draw a bootstrap resample of size n, Y_i^*'s, with replacement from the diseased sample Y_i's, and a separate bootstrap resample of size m, X_j^*'s, with replacement from the non-diseased sample X_j's. This bootstrap resampling strategy has been successfully used to construct confidence intervals for sensitivity of a continuous-scale test in Zhou and Qin.[24] Based on bootstrap samples, we can calculate a bootstrap estimate for the AUC which is

$$\widehat{\delta}^* = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} I(Y_j^* \geq X_i^*).$$

After repeating the process B times, B bootstrap estimates for the AUC are obtained:

$$\{\widehat{\delta}_b^* : b = 1, 2, \ldots, B\}$$

Let $\widehat{\delta}^*_{(1)}, \widehat{\delta}^*_{(2)}, \ldots, \widehat{\delta}^*_{(B)}$ denote the ordered values of $\widehat{\delta}_b^*$'s. Based on these $\widehat{\delta}_b^*$'s, we can construct different bootstrap-based confidence intervals for δ.

2.4.1 *Bootstrap percentile confidence interval for the AUC*

Let $\widehat{G}$ be the empirical distribution function of $\widehat{\delta}_b^*$'s,. We can apply the bootstrap percentile method described by Efron and Tibshirani[22] to get the $(1-\alpha)$-th bootstrap percentile (BP) confidence interval for δ:

$$(l_5, u_5) = \left(\widehat{G}^{-1}(\alpha/2), \widehat{G}^{-1}(1-\alpha/2)\right) = \left(\widehat{\delta}^*_{([B\alpha/2])}, \widehat{\delta}^*_{([B(1-\alpha)/2])}\right),$$

where $[x]$ represents the interger part of x, $\widehat{G}^{-1}(p)$ is the $100p$-th percentile of $\widehat{G}$.

Shao and Tu[23] (p. 132) provided a justification of the bootstrap percentile method. The assumption required for a good performance of the method is that the estimator for the parameter of interest has a known distribution or asymptotic distribution. Since the Mann–Whitney two-sample statistic $\widehat{\delta}$ is asymptotically normal, the BP interval for δ is asymptotically valid and its performance depends on how good the normal approximation is. The BP interval is simple but may not be very accurate unless sample sizes are very large.

2.4.2 *Bootstrap percentile-**t** confidence interval for the AUC*

The following method for constructing confidence interval of the AUC is derived from the standard bootstrap percentile-t method. Let S^* be the S calculated from bootstrap sample. We define the bootstrap distribution of $\widehat{\delta}^*$ as

$$K(x) = P^* \left\{ \left(\frac{mn}{m+n} \right)^{1/2} \frac{\widehat{\delta}^* - \widehat{\delta}}{S^*} \leq x \right\},$$

where P^* is the conditional probability distribution given the original samples. As Efron and Tibshirani[22] shown in their book, percentile-t method estimates the distribution function K directly from the data. By using B bootstrap re-samples, we can calculate $(mn/(m+n))^{1/2}\,(\widehat{\delta}^* - \widehat{\delta})/S^*$ and get

$$\left\{ k_b^* = \left(\frac{mn}{m+n} \right)^{1/2} \frac{\widehat{\delta}_b^* - \widehat{\delta}}{S_b^*} : b = 1, 2, \ldots, B \right\},$$

where S_b^* is the b-th bootstrap replicate of S. Then, the distribution function K can be estimated by the empirical distribution $\widehat{K}$ of k_b^*'s. The $(1-\alpha)$ bootstrap percentile-t (BPT) confidence interval for δ is given by

$$\begin{aligned}(l_6, u_6) &= \left(\widehat{\delta} - \left(\frac{(m+n)S^2}{mn} \right)^{1/2} \widehat{K}^{-1}(\alpha/2), \widehat{\delta} + \left(\frac{(m+n)S^2}{mn} \right)^{1/2} \widehat{K}^{-1}(1-\alpha/2) \right) \\ &= \left(\widehat{\delta} - \left(\frac{(m+n)S^2}{mn} \right)^{1/2} k^*_{([B\alpha/2])}, \widehat{\delta} + \left(\frac{(m+n)S^2}{mn} \right)^{1/2} k^*_{([B(1-\alpha)/2])} \right),\end{aligned}$$

where $k^*_{(b)}$'s represent the ordered values of k_b^*'s, $\widehat{K}^{-1}(p)$ is the $100p$-th percentile of $\widehat{K}$.

2.4.3 *Confidence intervals for the AUC calculated with bootstrap variance estimate*

Zhou and Qin[24] proposed a bootstrap method to construct confidence intervals for sensitivity at a fixed level of specificity of a continuous-scale diagnostic test. Using similar procedure, we can construct two additional confidence intervals for the AUC.

After drawing B bootstrap re-samples, we can obtain a bootstrap variance estimate for $\widehat{\delta}$:

$$V^* = \frac{1}{B-1}\sum_{b=1}^{B}(\widehat{\delta}_b^* - \bar{\delta}^*)^2$$

where $\bar{\delta}^* = \frac{1}{B}\sum_{b=1}^{B}\widehat{\delta}_b^*$.

Based on bootstrap variance estimate V^*, the first $(1-\alpha)$ confidence interval for the AUC, called BV1 interval, can be constructed as follows:

$$(l_7, u_7) = \bar{\delta}^* \mp Z_{1-\alpha/2}V^{*1/2}.$$

The second one, called BV2 interval, is given by

$$(l_8, u_8) = \widehat{\delta} \mp Z_{1-\alpha/2}V^{*1/2}.$$

2.4.4 *Bootstrap bias correction and acceleration confidence interval*

The bootstrap Bias Correction and Acceleration (BCa) method for construction of confidence interval is an improved version of the percentile method.[22] The endpoints of the BCa interval are given by percentiles of the bootstrap distribution, but they are not same as the ones described earlier. Under the setting for the interval estimation of the AUC, the percentiles used for this method depend on bias-correction W and acceleration a, which are defined by

$$W = \Phi^{-1}\left(\frac{1}{B}\sum_{b=1}^{B} I(\widehat{\delta}_b^* \le \widehat{\delta})\right),$$

where $\Phi^{-1}(\cdot)$ is the inverse function of the standard normal distribution function, and

$$a = \frac{1}{6}\sum_{j=1}^{n}\frac{d_j^3}{\left(\sum_{j=1}^{n} d_j^2\right)^{3/2}},$$

where $d_j = \widehat{\delta} - \widehat{U}_j$, $j = 1, 2, \ldots, n$. Using W and a, we can calculate the adjusted nominal level as

$$\widetilde{\alpha} = \Phi\left(W + \frac{W + z_\alpha}{1 - a(W + z_\alpha)}\right),$$

where z_α is the α-th quantile of standard normal distribution. Therefore, the $(1-\alpha)$-th BCa interval for the AUC is given by

$$(l_9, u_9) = \left(\widehat{\delta}^*_{([B\widetilde{\alpha}/2])}, \widehat{\delta}^*_{([B(1-\widetilde{\alpha})/2])}\right).$$

3 Simulation study

To compare the finite sample performances of the confidence intervals presented in Section 2, we conduct a simulation study to evaluate the coverage accuracy of these intervals when AUC is at level of 0.70, 0.80 (moderate accuracy), 0.90 and 0.95 (high accuracy). The study is conducted for eight combinations of sample sizes of (m, n) such as (25, 25), (50, 50), (80, 80), (100, 100), (50, 80), (80, 50), (70, 100) and (100, 70). For each combination of sample sizes, we generate 5000 random samples of size m from the non-diseased population and of size n from the diseased population, respectively. The 95% confidence intervals for the AUC are computed by using the nine different methods explained earlier. Since five of the confidence intervals for the AUC are based on bootstrap method, based on our extensive simulation studies, we recommend drawing $B \geq 150$ (here we take $B = 400$) bootstrap re-samples from the original samples.

In the simulation study, binormal and exponential models are evaluated for non-diseased and diseased populations. In binormal model, the distribution function F of non-diseased population is chosen to be a standard normal distribution function with mean $\mu_0 = 0$ and standard deviation $\sigma_0 = 1$. The distribution function G of diseased population is chosen to be a normal distribution function but with the mean $\mu = \sqrt{5}\Phi^{-1}(\delta)$ and the standard deviation $\sigma = 2$ where Φ is the standard normal cumulative distribution function. For binormally distributed test results, the AUC is

$$\delta = \Phi\left(\frac{\mu - \mu_0}{\sqrt{\sigma^2 + \sigma_0^2}}\right).$$

In exponential model, F is chosen to be a standard exponential distribution with rate $\upsilon = 1$, while G is chosen to be an exponential distribution with rate $\theta = 1/\delta - 1$. The corresponding AUC is

$$\delta = \frac{\upsilon}{\upsilon + \theta}.$$

Tables 1–4 show results obtained from the simulation study.

When AUC is at level 0.7, we observe that the coverage probabilities of all the intervals are close to each other, but EL and BP intervals have slightly shorter interval length. As AUC increases, the coverage probabilities of MW, DL, BP, BV1, BV2 and BCa intervals decrease, and fall well below the nominal level for high values of AUC. Tables 1–4 also show that EL intervals and LT intervals have similar coverage accuracy for all the AUC values considered here. The EL coverage is slightly less than the nominal level, whereas the LT coverage is at or slightly above the nominal level in most cases. It may explain why the length of EL interval is slightly shorter than that of LT interval. One problem with the LT interval is that it may break down when AUC $\widehat{\delta}$ is close to one (see Tables 1–4 where 'NA' means 'not available' for AUC = 0.90, 0.95).

Among the bootstrap-based intervals for the AUC, the BPT interval has the best coverage accuracy although it has the longest interval length. The other four bootstrap intervals often show much lower coverage probabilities than the EL and LT

Table 1 Normal distribution: coverage probability of 95% confidence intervals for the AUC. EL: EL-based interval. MW: Mann–Whitney interval. DL: DeLong's interval. LT: Logit-Transformation-based interval. BP: Bootstrap Percentile interval. BPT: Bootstrap Percentile-*t* interval. BV1 and BV2: Two intervals for the AUC calculated with bootstrap variance estimate. BCa: Bootstrap Bias Correction and Acceleration interval

AUC	(m, n)	EL	MW	DL	LT	BP	BPT	BV1	BV2	BCa
0.70	(25,25)	0.9310	0.9309	0.9325	0.9549	0.9333	0.9733	0.9273	0.9276	0.9276
	(50,50)	0.9429	0.9417	0.9433	0.9541	0.9380	0.9623	0.9366	0.9376	0.9390
	(80,80)	0.9461	0.9437	0.9441	0.9552	0.9453	0.9530	0.9466	0.9460	0.9446
	(100,100)	0.9479	0.9464	0.9445	0.9547	0.9480	0.9546	0.9480	0.9490	0.9473
	(50,80)	0.9465	0.9465	0.9455	0.9554	0.9433	0.9586	0.9456	0.9460	0.9413
	(80,50)	0.9427	0.9428	0.9421	0.9532	0.9456	0.9666	0.9456	0.9470	0.9470
	(70,100)	0.9430	0.9433	0.9497	0.9503	0.9430	0.9536	0.9433	0.9426	0.9380
	(100,70)	0.9451	0.9437	0.9408	0.9527	0.9410	0.9586	0.9400	0.9400	0.9393
0.80	(25,25)	0.9275	0.9245	0.9213	0.9551	0.9330	0.9790	0.9246	0.9236	0.9303
	(50,50)	0.9407	0.9379	0.9357	0.9538	0.9300	0.9690	0.9316	0.9343	0.9290
	(80,80)	0.9431	0.9418	0.9367	0.9530	0.9450	0.9690	0.9426	0.9406	0.9366
	(100,100)	0.9489	0.9454	0.9400	0.9514	0.9380	0.9590	0.9433	0.9430	0.9410
	(50,80)	0.9468	0.9412	0.9434	0.9513	0.9310	0.9610	0.9253	0.9270	0.9256
	(80,50)	0.9420	0.9347	0.9347	0.9512	0.9313	0.9656	0.9303	0.9290	0.9306
	(70,100)	0.9472	0.9440	0.9445	0.9513	0.9290	0.9570	0.9430	0.9420	0.9393
	(100,70)	0.9443	0.9368	0.9399	0.9477	0.9383	0.9613	0.9366	0.9363	0.9366
0.90	(25,25)	0.8892	0.8845	0.8846	NA	0.9010	0.9406	0.8863	0.8860	0.8900
	(50,50)	0.9352	0.9204	0.9161	0.9468	0.9150	0.9700	0.9240	0.9246	0.9300
	(80,80)	0.9411	0.9264	0.9311	0.9471	0.9310	0.9670	0.9376	0.9373	0.9346
	(100,100)	0.9468	0.9356	0.9332	0.9486	0.9220	0.9540	0.9330	0.9330	0.9293
	(50,80)	0.9458	0.9281	0.9317	0.9547	0.9340	0.9630	0.9363	0.9356	0.9396
	(80,50)	0.9330	0.9174	0.9194	0.9438	0.9246	0.9626	0.9160	0.9153	0.9160
	(70,100)	0.9434	0.9297	0.9349	0.9498	0.9290	0.9580	0.9403	0.9406	0.9440
	(100,70)	0.9401	0.9265	0.9257	0.9444	0.9290	0.9653	0.9180	0.9173	0.9253
0.95	(25,25)	0.8400	0.8300	0.8160	NA	0.8453	0.8513	0.8173	0.8186	0.9000
	(50,50)	0.8964	0.8818	0.8793	0.9289	0.8840	0.9490	0.9020	0.9040	0.9160
	(80,80)	0.9252	0.9064	0.9072	0.9414	0.9180	0.9610	0.9056	0.9063	0.9120
	(100,100)	0.9340	0.9142	0.9188	0.9366	0.9180	0.9660	0.9196	0.9196	0.9246
	(50,80)	0.9269	0.9060	0.9065	0.9462	0.9150	0.9720	0.9043	0.9046	0.9190
	(80,50)	0.9205	0.8810	0.8827	0.9263	0.8896	0.9433	0.8780	0.8783	0.8883
	(70,100)	0.9351	0.9138	0.9164	0.9478	0.9230	0.9660	0.9120	0.9110	0.9166
	(100,70)	0.9273	0.9000	0.9054	0.9361	0.9146	0.9636	0.9046	0.9036	0.9096

intervals when the AUC is higher than 0.80. Furthermore, the bootstrap intervals are computationally the most extensive intervals among all the intervals considered here.

In summary, we recommend the use of EL intervals or the LT intervals for the AUC when the underlying distributions for diseased and non-diseased populations are unknown. When AUC ≥ 0.95, the BPT interval is also a good alternative confidence interval for the AUC even though it is slightly conservative.

Table 2 Normal distribution: average length of 95% confidence intervals for the AUC

AUC	(m, n)	EL	MW	DL	LT	BP	BPT	BV1	BV2	BCa
0.70	(25,25)	0.2905	0.3033	0.3035	0.2971	0.2943	0.3330	0.2986	0.2986	0.2924
	(50,50)	0.2088	0.2134	0.2133	0.2110	0.2087	0.2231	0.2115	0.2115	0.2081
	(80,80)	0.1658	0.1681	0.1682	0.1670	0.1654	0.1736	0.1675	0.1675	0.1652
	(100,100)	0.1485	0.1501	0.1502	0.1494	0.1480	0.1545	0.1499	0.1499	0.1478
	(50,80)	0.1744	0.1770	0.1771	0.1757	0.1739	0.1826	0.1763	0.1763	0.1736
	(80,50)	0.2013	0.2056	0.2058	0.1757	0.2016	0.2161	0.2041	0.2041	0.2011
	(70,100)	0.1540	0.1559	0.1560	0.1551	0.1533	0.1599	0.1554	0.1554	0.1530
	(100,70)	0.1723	0.1749	0.1749	0.1550	0.1718	0.1813	0.1740	0.1740	0.1712
0.80	(25,25)	0.2490	0.2567	0.2567	0.2583	0.2507	0.3104	0.2551	0.2551	0.2489
	(50,50)	0.1499	0.1519	0.1809	0.1519	0.1502	0.1626	0.1798	0.1798	0.1762
	(80,80)	0.1188	0.1197	0.1428	0.1200	0.1185	0.1259	0.1424	0.1424	0.1400
	(100,100)	0.1064	0.1071	0.1277	0.1073	0.1071	0.1126	0.1276	0.1276	0.1255
	(50,80)	0.1246	0.1257	0.1501	0.1262	0.1252	0.1327	0.1489	0.1489	0.1467
	(80,50)	0.1726	0.1749	0.1749	0.1753	0.1715	0.1894	0.1741	0.1741	0.1707
	(70,100)	0.1102	0.1109	0.1322	0.1111	0.1107	0.1165	0.1322	0.1322	0.1304
	(100,70)	0.1473	0.1487	0.1488	0.1490	0.1463	0.1577	0.1482	0.1482	0.1455
0.90	(25,25)	0.1765	0.1779	0.1775	NA	0.1719	0.2815	0.1764	0.1764	0.1716
	(50,50)	0.1070	0.1069	0.1270	0.1101	0.1058	0.1252	0.1274	0.1274	0.1240
	(80,80)	0.0849	0.0846	0.1010	0.0860	0.0841	0.0930	0.1009	0.1009	0.0987
	(100,100)	0.0759	0.0757	0.0903	0.0767	0.0752	0.0817	0.0900	0.0900	0.0878
	(50,80)	0.0886	0.0882	0.1055	0.0902	0.0883	0.0973	0.1051	0.1051	0.1030
	(80,50)	0.1247	0.1238	0.1233	0.1281	0.1202	0.1513	0.1223	0.1223	0.1187
	(70,100)	0.0783	0.0781	0.0933	0.0794	0.0782	0.0847	0.0929	0.0929	0.0910
	(100,70)	0.1061	0.1054	0.1051	0.1082	0.1031	0.1199	0.1048	0.1048	0.1021
0.95	(25,25)	0.1146	0.1128	0.1154	NA	0.1117	0.2126	0.1166	0.1166	0.1210
	(50,50)	0.0723	0.0711	0.0851	0.0765	0.0706	0.1013	0.0857	0.0857	0.0830
	(80,80)	0.0582	0.0571	0.0680	0.0600	0.0569	0.0704	0.0683	0.0683	0.0662
	(100,100)	0.0522	0.0513	0.0610	0.0531	0.0512	0.0603	0.0609	0.0609	0.0593
	(50,80)	0.0604	0.0593	0.0704	0.0626	0.0595	0.0727	0.0705	0.0705	0.0685
	(80,50)	0.0850	0.0826	0.0827	0.0907	0.0810	0.1368	0.0829	0.0829	0.0800
	(70,100)	0.0537	0.0527	0.0631	0.0546	0.0522	0.0612	0.0626	0.0626	0.0608
	(100,70)	0.0733	0.0711	0.0713	0.0760	0.0699	0.0993	0.0713	0.0713	0.0690

4 An illustration example: pancreatic cancer biomarker

Pancreatic cancer is a disease that is difficult to be diagnosed at its early stage. CA-19-9 (a carbohydrate antigen) is a biomarker for pancreatic cancer measured on a continuous positive scale. Wieand *et al.*[9] reported a study on the diagnostic accuracy of CA-19-9 in detecting pancreatic cancer. Concentrations of CA-19-9 in sera (ML) from 51 'control' patients with pancreatitis and 90 'cases' pancreatic cancer were collected. This dataset has been used by numerous statisticians to illustrate statistical techniques for

Table 3 Exponential distribution: coverage probability of 95% confidence intervals for the AUC

AUC	(m, n)	EL	MW	DL	LT	BP	BPT	BV1	BV2	BCa
0.70	(25,25)	0.9348	0.9398	0.9400	0.9610	0.9350	0.9740	0.9316	0.9320	0.9350
	(50,50)	0.9447	0.9445	0.9436	0.9532	0.9383	0.9620	0.9383	0.9390	0.9420
	(80,80)	0.9418	0.9418	0.9428	0.9532	0.9376	0.9520	0.9353	0.9356	0.9346
	(100,100)	0.9448	0.9464	0.9451	0.9508	0.9513	0.9600	0.9520	0.9523	0.9496
	(50,80)	0.9445	0.9455	0.9491	0.9518	0.9436	0.9596	0.9486	0.9483	0.9430
	(80,50)	0.9419	0.9434	0.9428	0.9521	0.9496	0.9646	0.9466	0.9453	0.9476
	(70,100)	0.9489	0.9483	0.9487	0.9525	0.9463	0.9550	0.9436	0.9450	0.9450
	(100,70)	0.9445	0.9429	0.9478	0.9510	0.9466	0.9610	0.9486	0.9486	0.9430
0.80	(25,25)	0.9253	0.9251	0.9241	0.9572	0.9390	0.9800	0.9273	0.9273	0.9400
	(50,50)	0.9446	0.9394	0.9390	0.9551	0.9270	0.9570	0.9430	0.9453	0.9473
	(80,80)	0.9459	0.9432	0.9438	0.9547	0.9280	0.9540	0.9366	0.9373	0.9356
	(100,100)	0.9507	0.9456	0.9426	0.9478	0.9380	0.9540	0.9433	0.9450	0.9473
	(50,80)	0.9469	0.9444	0.9468	0.9550	0.9370	0.9620	0.9480	0.9463	0.9480
	(80,50)	0.9439	0.9400	0.9390	0.9522	0.9460	0.9696	0.9420	0.9423	0.9420
	(70,100)	0.9477	0.9410	0.9468	0.9541	0.9450	0.9660	0.9420	0.9420	0.9403
	(100,70)	0.9449	0.9437	0.9442	0.9531	0.9416	0.9583	0.9406	0.9410	0.9386
0.90	(25,25)	0.9025	0.8961	0.8898	NA	0.9060	0.9476	0.8853	0.8833	0.9150
	(50,50)	0.9321	0.9200	0.9200	0.9482	0.9240	0.9740	0.9176	0.9170	0.9296
	(80,80)	0.9423	0.9291	0.9345	0.9514	0.9240	0.9600	0.9296	0.9286	0.9310
	(100,100)	0.9499	0.9380	0.9362	0.9485	0.9340	0.9610	0.9326	0.9330	0.9373
	(50,80)	0.9467	0.9312	0.9316	0.9535	0.9270	0.9610	0.9290	0.9293	0.9346
	(80,50)	0.9331	0.9140	0.9182	0.9546	0.9300	0.9713	0.9216	0.9226	0.9276
	(70,100)	0.9467	0.9364	0.9313	0.9520	0.9310	0.9670	0.9413	0.9403	0.9406
	(100,70)	0.9432	0.9271	0.9293	0.9525	0.9310	0.9650	0.9236	0.9250	0.9260
0.95	(25,25)	0.8800	0.8000	0.8148	NA	0.8476	0.8520	0.8153	0.8156	0.8600
	(50,50)	0.8977	0.8817	0.8875	NA	0.9000	0.9460	0.8800	0.8796	0.8983
	(80,80)	0.9296	0.9090	0.9060	0.9398	0.9160	0.9660	0.9093	0.9103	0.9190
	(100,100)	0.9412	0.9174	0.9129	0.9473	0.9110	0.9660	0.9160	0.9180	0.9263
	(50,80)	0.9329	0.9080	0.9010	0.9446	0.9220	0.9750	0.9103	0.9123	0.9193
	(80,50)	0.9230	0.8817	0.8772	0.9421	0.8860	0.9533	0.8726	0.8733	0.8850
	(70,100)	0.9356	0.9117	0.9166	0.9435	0.9340	0.9710	0.9130	0.9130	0.9216
	(100,70)	0.9360	0.9058	0.9010	0.9431	0.9106	0.9646	0.9010	0.9013	0.9100

diagnostic tests.[2] Here, we are interested in estimating the AUC of CA-19-9 and finding a range of the AUC.

An estimate for the AUC of CA19-9 is 0.862 based on MW estimator. Since the distributions of measurements for CA-19-9 in 'control' and 'case' groups are unknown, based on our simulation studies, we should use the EL interval or LT interval for the AUC as the range of global diagnostic accuracy of CA-19-9. The 95% EL and LT intervals for the AUC are [0.793, 0.913], [0.791, 0.912], respectively. Note that the EL and LT intervals have almost no difference in this example. Both intervals suggest

Table 4 Exponential distribution: average length of 95% confidence intervals for the AUC

AUC	(m, n)	EL	MW	DL	LT	BP	BPT	BV1	BV2	BCa
0.70	(25,25)	0.2816	0.2932	0.2932	0.2878	0.2855	0.3176	0.2900	0.2900	0.2855
	(50,50)	0.2013	0.2054	0.2055	0.2033	0.2016	0.2141	0.2043	0.2043	0.2016
	(80,80)	0.1598	0.1619	0.1618	0.1609	0.1596	0.1669	0.1614	0.1614	0.1599
	(100,100)	0.1431	0.1446	0.1447	0.1438	0.1423	0.1481	0.1440	0.1440	0.1422
	(50,80)	0.1762	0.1788	0.1790	0.1771	0.1761	0.1850	0.1784	0.1784	0.1764
	(80,50)	0.1872	0.1905	0.1903	0.1773	0.1870	0.1982	0.1895	0.1895	0.1865
	(70,100)	0.1537	0.1554	0.1555	0.1547	0.1533	0.1599	0.1553	0.1553	0.1537
	(100,70)	0.1612	0.1633	0.1632	0.1547	0.1610	0.1689	0.1630	0.1630	0.1610
0.80	(25,25)	0.2408	0.2481	0.2479	0.2492	0.2410	0.2872	0.2453	0.2453	0.2408
	(50,50)	0.1725	0.1746	0.1745	0.1748	0.1713	0.1865	0.1739	0.1739	0.1712
	(80,80)	0.1366	0.1376	0.1377	0.1378	0.1354	0.1436	0.1371	0.1371	0.1352
	(100,100)	0.1223	0.1230	0.1229	0.1232	0.1210	0.1274	0.1226	0.1226	0.1209
	(50,80)	0.1478	0.1489	0.1484	0.1489	0.1464	0.1559	0.1483	0.1483	0.1470
	(80,50)	0.1628	0.1645	0.1648	0.1490	0.1622	0.1769	0.1646	0.1646	0.1616
	(70,100)	0.1292	0.1301	0.1301	0.1303	0.1283	0.1353	0.1300	0.1300	0.1284
	(100,70)	0.1396	0.1407	0.1406	0.1302	0.1385	0.1479	0.1403	0.1403	0.1381
0.90	(25,25)	0.1737	0.1752	0.1745	NA	0.1698	0.2643	0.1739	0.1739	0.1672
	(50,50)	0.1254	0.1247	0.1244	0.1290	0.1223	0.1489	0.1244	0.1244	0.1217
	(80,80)	0.0990	0.0984	0.0986	0.1010	0.0972	0.1093	0.0986	0.0986	0.0968
	(100,100)	0.0887	0.0882	0.0883	0.0898	0.0870	0.0957	0.0883	0.0883	0.0866
	(50,80)	0.1044	0.1036	0.1036	0.1064	0.1020	0.1143	0.1035	0.1035	0.1019
	(80,50)	0.1206	0.1199	0.1198	0.1064	0.1174	0.1439	0.1194	0.1194	0.1165
	(70,100)	0.0919	0.0914	0.0913	0.0933	0.0898	0.0987	0.0912	0.0912	0.0898
	(100,70)	0.1031	0.1025	0.1023	0.0933	0.1005	0.1152	0.1022	0.1022	0.0996
0.95	(25,25)	0.1123	0.1115	0.1162	NA	0.1103	0.2120	0.1152	0.1152	0.1180
	(50,50)	0.0881	0.0859	0.0858	NA	0.0828	0.1401	0.0848	0.0848	0.0822
	(80,80)	0.0705	0.0688	0.0685	0.0731	0.0668	0.0890	0.0680	0.0680	0.0662
	(100,100)	0.0630	0.0617	0.0614	0.0650	0.0606	0.0751	0.0616	0.0616	0.0600
	(50,80)	0.0726	0.0707	0.0702	0.0759	0.0697	0.0920	0.0711	0.0711	0.0693
	(80,50)	0.0856	0.0835	0.0830	0.0759	0.0812	0.1412	0.0832	0.0832	0.0803
	(70,100)	0.0643	0.0629	0.0629	0.0668	0.0619	0.0766	0.0630	0.0630	0.0614
	(100,70)	0.0738	0.0720	0.0719	0.0664	0.0704	0.1010	0.0718	0.0718	0.0696

that CA-19-9 has moderate to high level of diagnostic accuracy in detecting pancreatic cancer.

5 Discussion

The main purpose of a diagnostic test is to determine if a patient has or does not have the disease. Because of its significant role, the accuracy of this test is very important. The area under the ROC curve measures the discrimination ability of a diagnostic test. In order to report AUC properly, it is necessary to construct a confidence interval for its value. In this article, we have discussed and compared nine non-parametric methods for constructing confidence intervals of the AUC. Among the nine intervals, the MW interval is the simplest but usually not the best in terms of coverage probability, particularly when AUC is high and the sample sizes for diseased and non-diseased subjects are small and unequal. Both EL and LT intervals for the AUC have good coverage accuracy. The EL interval has nice asymptotic property and is also simple to implement. The LT interval depends on the asymptotic normality assumption of logit transformation of MW estimator. It may break down when the observed AUC $\widehat{\delta}$ is close to one. It can be used when the normal approximation is true and $\widehat{\delta}$ is not close to one. The BPT interval is slightly conservative, but it has good coverage accuracy when AUC ≥ 0.95. It may be a good alternative interval for the AUC when the AUC is extremely high. A S-plus code implementing the recommended methods is available from the authors.

6 Acknowledgments

This work is supported in part by NSF DMS 0603913. We would like to thank the editor and referees for their valuable comments which led to substantial improvements in this article.

References

1 Zhou XH, Obuchowski NA McClish DM. *Statistical Methods in Diagnostic Medicine*. Wiley & Sons, 2002.

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

3 Hanley JA. Receiver operating characteristic (ROC) methodology: the state of the art. *Clinical Reviews in Diagnostic Imaging* 1989; **29**: 307–35.

4 Swets JA. The relative operating characteristic in psychology. *Science* 1973; **182**: 990–1000.

5 Swets JA. Measuring the accuracy of diagnostic systems. *Science* 1988; **240**: 1285–93.

6 Shapiro DE. The interpretation of diagnostic tests. *Statistical Methods in Medical Research* 1999; **8**: 113–34.

7 Bamber DC. The area above the ordinal dominance graph and the area below the receiver operating characteristic curve graph. *J Math Psychology* 1975; **12**: 387–415.

8 Wolfe DA, Hogg RV. On constructing statistics and reporting data. *American Statistician* 1971; **25**: 27–30.

9 Wieand S, Gail MH, James BR, James KR. A family of non-parametric statistics for comparing diagnostic markers with paired and unpaired data. *Biometrika* 1989; **76**: 585–92.

10 Metz CE, Herman BA, Shen J-H. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. *Statistics in Medicine* 1998; **17**: 1033–53.

11 Tong H. On the estimation of $P(Y < X)$ for exponential families. *IEEE Transactions on Reliability* 1977; **26**: 54–56.

12 Pham T, Almhana J. The generalized gamma distribution: its hazard rate and stress-strength model. *IEEE Transactions on Reliability* 1995; **44**: 392–97.

13 Gupta CG, Brown N. Reliability studies of the skew-normal distribution and its application to a stress-strength model. *Communications in Statistics: Theory and Methods* 2001; **30**: 2427–45.

14 Obuchowski NA, Lieber ML. Confidence bounds when estimated ROC area is 1.0. *Academic Radiology* 2002; **9**: 526–30.

15 Obuchowski NA, Lieber ML. Confidence intervals for the receiver operating characteristic area in studies with small samples. *Academic Radiology* 1998; **5**: 561–71.

16 Hanley JA, Hajian-Tilaki KO. Sampling variability of non-parametric estimates of the area under receiver operating characteristic curves: an update. *Academic Radiology* 1997; **4**: 49–58.

17 DeLong EA, DeLong DM, Clarke-Pearson DL. Comparing the area under receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988; **44**: 837–45.

18 Sen PK. A note on asymptotically distribution-free confidence bounds for $P(X < Y)$ based on two independent samples. *Sankhyã, Series A* 1967; **29**: 95–102.

19 Qin GS, Zhou XH. Empirical likelihood inference for the area under the ROC curve. *Biometrics* 2006; **62**: 613–22.

20 Rao ACR, Collinson PO, Canepa-Anson R, Josepha SP. Troponin T measurement after myocardial infarction can identify left ventricular ejection of less than 40%. *Heart* 1998; **80**: 223–25.

21 Pepe MS, Cai T. The analysis of placement values for evaluating discriminatory measures. *Biometrics* 2004; **60**: 528–35.

22 Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Chapman and Hall, 1993.

23 Shao J, Tu D. *The Jackknife and Bootstrap*. Springer-Verlag, 1995.

24 Zhou XH, Qin GS. Improved confidence intervals for the sensitivity of a receiver operating characteristic-scale test at a fixed level of specificity. *Statistics in Medicine* 2005; **24**: 465–77. DOI:10.1002/sim.1563