STATISTICS IN MEDICINE
Statist. Med. (2008)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3366

Bayesian bootstrap estimation of ROC curve

Jiezhun Gu[1,*,†], Subhashis Ghosal[2] and Anindya Roy[3]

[1]*Duke Clinical Research Institute, Duke University Medical Center, Durham, NC, U.S.A.*
[2]*Department of Statistics, North Carolina State University, Raleigh, NC, U.S.A.*
[3]*Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, MD, U.S.A.*

SUMMARY

Receiver operating characteristic (ROC) curve is widely applied in measuring discriminatory ability of diagnostic or prognostic tests. This makes the ROC analysis one of the most active research areas in medical statistics. Many parametric and semiparametric estimation methods have been proposed for estimating the ROC curve and its functionals. In this paper, we propose the Bayesian bootstrap (BB), a fully nonparametric estimation method, for the ROC curve and its functionals, such as the area under the curve (AUC). The BB method offers a bandwidth-free smoothing approach to the empirical estimate, and gives credible bounds. The accuracy of the estimate of the ROC curve in the simulation studies is examined by the integrated absolute error. In comparison with other existing curve estimation methods, the BB method performs well in terms of accuracy, robustness and simplicity. We also propose a procedure based on the BB approach to test the binormality assumption. Copyright © 2008 John Wiley & Sons, Ltd.

KEY WORDS: area under the curve (AUC); Bayesian bootstrap; integrated absolute error; ROC curve; testing binormality

1. INTRODUCTION

Since its introduction in the context of electronic signal detection [1], receiver operating characteristic (ROC) curve has become the method of choice for quantification of accuracy of medical diagnostic tests. The diagnostic variables $X \sim F$ for the group without disease and $Y \sim G$ for those with disease are well defined. The ROC curve is a plot of the true positive fraction

*Correspondence to: Jiezhun Gu, Duke Clinical Research Institute, Duke University Medical Center, P.O. Box 17969, Durham, NC 27715, U.S.A.
†E-mail: jiezhun.gu@duke.edu

Contract/grant sponsor: NSF; contract/grant number: DMS-0349111

Received 1 May 2007
Accepted 16 May 2008

(TPF) as a function of the false positive fraction (FPF), or sensitivity versus one minus specificity, and is obtained by varying the threshold criterion distinguishing between a positive and negative diagnosis. Some features such as the invariance property and interpretation of the area under the curve (AUC) as $\Pr(Y>X)$ make the ROC analysis extremely popular in diagnostics research.

There is extensive literature for ROC analysis for continuous diagnostic variables based on independent observations. Semiparametric (SP) methods for ROC analysis have been particularly popular because in addition to specific parametric features, the presence of nonparametric components make these models considerably flexible. Under the SP framework, there are several different approaches. The simplest one is the binormal model [1], which assumes normality of each diagnostic test variable after a common monotonically increasing transformation. The intercept and slope in the binormal model can be estimated by several methods, such as by Hsieh and Turnbull [2], Metz *et al.* [3], Zou and Hall [4], Pepe [5, 6], Cai and Moskowitz [7], among others. However, Goddard and Hinberg [8] provided examples where binormality failed. Similarly, bi-gamma [9] and bi-beta [10] models have also been considered. Li *et al.* [11] proposed a model where F and G are specified nonparametrically and parametrically, respectively. Qin and Zhang [12] modeled the functional form of the likelihood ratio to estimate the parameters. A normal mixture model was studied by Hall and Zhou [13]. Among completely nonparametric methods, the kernel estimates of F and G were discussed by Zou *et al.* [14], Lloyd [15] and others. Because AUC is an important index for the ROC curve, the estimation method has been discussed by many including Bamber [16], Brownie *et al.* [17], DeLong *et al.* [18] and Qin and Zhou [19].

Within the nonparametric framework, the empirical estimator of ROC was studied by Hsieh and Turnbull [2], along with its asymptotic property. Li *et al.* [20] obtained the weak convergence theory for the ROC estimator under censoring by plugging-in the product-limit estimators.

The empirical counterpart of the ROC curve is inherently discrete due to the finite choice of the weights. The discreteness creates conceptual problems in inversion of the estimated ROC curve. Moreover, since the true ROC curve is generally perceived to be a smooth continuous function, the discontinuous estimator is unappealing. One option is to smooth the empirical estimator, but the final estimator depends on the choice of the smoothing parameter. The choice of the smoothing parameter is a non-trivial issue. In addition, construction of confidence bands based on asymptotic distribution is complicated. Our motivation is to develop a procedure that can bypass the smoothing step and easily produce credible bands. Our approach is to generate an ensemble of replica of the ROC curve and compute relevant quantities based on the ensemble. The average of the ensemble is used as a point estimator and the variation in the ensemble is used to construct credible regions. The Bayesian bootstrap (BB) [21] is a resampling procedure, which is similar to the bootstrap but gives smoother choices of weights. We propose a smoother estimator of the ROC curve and related credible bands by using the BB method. The BB method closely resembles a non-parametric Bayesian analysis using the Dirichlet process prior with precision parameters converging to zero, and makes inference based on the 'BB posterior distribution'. In other words, because the BB method corresponds to the posterior of a 'non-informative' prior for which the prior base measure in a Dirichlet process has been chosen to be the null measure, generating samples from the BB posterior distribution reduces to finite-dimensional random variate generations, i.e. assigning the Dirichlet distribution to the weights at observations. A more detailed explanation is given in the following section. We also avoid the problem of inverting the randomly generated survival function, $\bar{F}(x)$. Another appealing feature of our methodology is that it readily produces standard errors,

DOI: 10.1002/sim

credible intervals and credible bands for associated summary measures such as the AUC and the partial AUC (pAUC). The BB method is not based on large sample techniques, and in principle applies to any sample size.

Our simulations show that these bands and intervals have approximate frequentist validity even for very small sample sizes. This phenomenon has been theoretically explained by strong approximation theory [22]. In comparison with other existing curve estimation methods, the BB method performs well in terms of accuracy, robustness, simplicity and smoothness. We also propose a procedure to test the binormality assumption as an application.

Our methodology is explained in Section 2. A procedure for testing the binormality assumption is presented in Section 3. Results from simulation studies are displayed in Section 4 and real data analyses are given in Section 5.

2. METHODOLOGY

The purpose of using the BB method is to produce valid curve estimates as well as credible bands for any ROC curve.

Let $X \sim F$ and $Y \sim G$ be two independent continuous variables, for instance, two diagnostic variables coming from two populations, one without disease and one with disease, respectively. By varying the decision threshold value c_t (if $X > c_t$ or $Y > c_t$, false or true positive event occurs) and plotting the TPF (sensitivity) versus the FPF (one minus specificity), the ROC curve is obtained: $\{(P(X>c_t), P(Y>c_t))\} = \{(t, R(t))\}$, where $c_t \in \mathbb{R}$, $t = P(X>c_t)$ is called the FPF. Hence, when t is given, $c_t = \bar{F}^{-1}(t) = F^{-1}(1-t)$, where $F^{-1}(\xi) = \inf\{x : F(x) \geqslant \xi\}$. Mathematically, we can express the functional form of ROC curve [6, p. 106] as follows:

$$R(t) = \bar{G}(\bar{F}^{-1}(t)) = P(Y>c_t) = P(Y>\bar{F}^{-1}(t)) = P(\bar{F}(Y) \leqslant t) \tag{1}$$

where $\bar{F}(u) = P(X>u)$ and $\bar{G}(u) = P(Y>u)$ are survival functions of X and Y, respectively. A commonly used index to compare the accuracy of the modalities is given by AUC A and its estimate $\hat{A}$ defined as

$$A = \int_0^1 R(t)\,\mathrm{d}t \quad \text{and} \quad \hat{A} = \int_0^1 \hat{R}(t)\,\mathrm{d}t \tag{2}$$

where $\hat{R}(t)$ is some estimate of $R(t)$.

The accuracy of estimation for the entire ROC curve can be measured by the integrated absolute error (IAE) [23]: $\mathrm{IAE} = \int_0^1 |\hat{R}(t) - R(t)|\,\mathrm{d}t$.

Clearly $|\hat{A} - A| \leqslant \mathrm{IAE}$. To construct a uniform credible band for ROC, it is advantageous to map the domain to the real line via a transformation ψ, such as the logistic transformation (LT) $\psi(x) = \log(x/(1-x)), x \in (0, 1)$. The maximum possible estimation error in the ψ-scale is $\varepsilon(\psi, R, \hat{R}) = \sup\{|\psi(\hat{R}(t)) - \psi(R(t))| : t \in (0, 1)\}$. The width of a uniform $100(1-\alpha)$ per cent credible band for the transformed ROC, denoted by $d_\alpha(\psi, R)$, is given by

$$d_\alpha = d_\alpha(\psi, R) = 100(1-\alpha) \text{ per cent percentile of the distribution of } \varepsilon(\psi, R, \hat{R}) \tag{3}$$

Thus, the uniform $100(1-\alpha)$ per cent credible band for the ROC curve can be constructed by

$$\psi^{-1}(\psi(\hat{R}(t)) - d_\alpha) \leqslant R(t) \leqslant \psi^{-1}(\psi(\hat{R}(t)) + d_\alpha) \tag{4}$$

DOI: 10.1002/sim

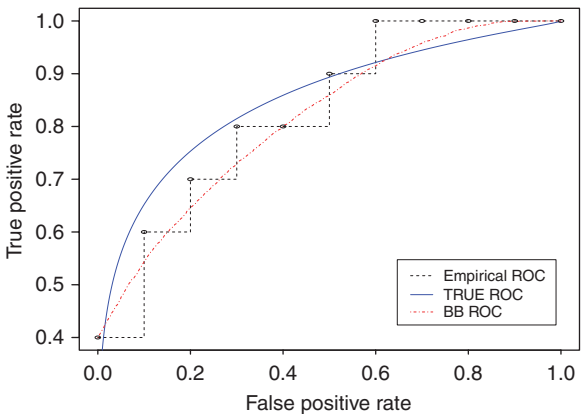

Figure 1. Comparison of empirical and the BB estimate of ROC with the true ROC (simulation data set: $X_1, \ldots, X_m \sim$ i.i.d. N$(0, 1)$, $Y_1, \ldots, Y_n \sim$ i.i.d. N$(1.868, 1.5^2)$, $m = n = 10$, 5000 BB resamples, grid points on $[0, 1]$ are chosen at equal intervals of length 0.001).

The transformation–retransformation technique automatically ensures that the credible band lies within the unit square. In practice, d_α has to be estimated, usually by some resampling technique. If the uniform credible band for ROC on $t \in (0, 1)$ is too wide, other alternatives such as pointwise $100(1-\alpha)$ per cent credible band or uniform $100(1-\alpha)$ per cent credible band restricted on a small subinterval of interest should be considered instead.

As shown below, the motivation of our BB estimator is twofold. On the one hand, we can view the BB estimator as a bandwidth-free smoothing of the empirical estimate. On the other hand, we argue that it is a non-informative limit of a Bayesian estimate based on the Dirichlet process prior. The motivations of our BB estimator are as follows:

1. Empirical ROC estimators [2] are easily obtained by plugging the empirical counterparts in the expression for the ROC. In order to have a continuous estimator of the ROC curve, the jumps in the empirical cumulative distribution function (CDF) can be interpolated linearly. Bootstrap method [24] can be used to get the error of the curve estimate. However, inherent discreteness of the estimate is partially due to finite choice of the weights. The BB method proposed below assigns the Dirichlet distribution to the weights. By forming an ensemble of estimators and averaging, it provides a smoother version of the bootstrap. Figure 1 gives an illustration of these differences even when the sample size is small.
2. To implement a Bayesian analysis, a natural choice for priors on F and G is Dirichlet process, denoted as DP, with certain pairs of precision M and center measure ξ, say $F \sim \text{DP}(M_1, \xi_1)$, $G \sim \text{DP}(M_2, \xi_2)$. Conditional on the data $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$, the posterior of $(F|\text{data})$ is $\text{DP}(M_1+m, (M_1\xi_1 + m\mathbb{F}_m)/(M_1+m))$, and that of $(R|F, \text{data})$ is $\text{DP}(M_2+n, (M_2\xi_2 \circ \bar{F}^{-1} + n\mathbb{G}_n \circ \bar{F}^{-1})/(M_2+n))$. Unfortunately, the above informative posterior can be obtained only by the Sethuraman [25] representation of the Dirichlet process and function inversion. The Sethuraman representation involves generating an infinite collection of random variables, which is computationally very intensive. Instead, we consider the non-informative limit of the Dirichlet processes by letting $M_1 \to 0$ and $M_2 \to 0$. This simplifies the procedure to an easier computational problem. In fact, we do not even have to specify

DOI: 10.1002/sim

the center measures ξ_1 and ξ_2. We need only to generate from the uniform distribution over the simplex, which can be done quite easily; see Remark (1). Our BB estimator is based on this simplification, i.e. the posterior of $(F|\text{data})$ is DP$(m, \mathbb{F}_m)$ and $(R|F, \text{data})$ is DP $(n, \mathbb{G}_n \circ \bar{F}^{-1})$.

The BB estimator of the ROC curve and its associated summary measures can be computed as follows: Recall $R(t) = \Pr(\bar{F}(Y) \leqslant t),\ t \in \mathscr{D} \subset [0, 1]$, where $\mathscr{D}$ denotes a prespecified set of FPF of interest. If we can impute the variable $Z = \bar{F}(Y)$ by plugging-in the survival function $\bar{F}$ of X, generated from the BB resampling distribution given $X_1, \ldots X_n$, then the CDF of Z based on the BB resampling distribution constitutes one realization of the ROC curve from the corresponding posterior distribution. This argument leads to the following computational steps:

Step 1 (*Imputing the placement variables based on the BB resampling distribution*): Let $Z_j = \bar{F}^{\#}(Y_j) = 1 - F^{\#}(Y_j),\ F^{\#}(u) = \sum_{i=1}^{m} p_i 1(X_i \leqslant u),\ (p_1, \ldots, p_m) \sim$ Dirichlet $(m; 1, \ldots, 1)$ independent of others. This is equivalent to generating $\bar{F}^{\#} \sim$ DP $(m, \mathbb{F}_m)$ and evaluating Z_j at Y_j (Z_j is also called non-disease placement value [6, p. 105] evaluated at Y_j). The difference between our method and Pepe's method lies in the fact that we choose the survival function using the BB resampling distribution instead of the empirical one.

Step 2 (*Generating one random realization of the ROC curve*): Generate the realization of $R^{\#}_{m,n}(t)$, the CDF of $Z_1, \ldots, Z_n$, where $R^{\#}_{m,n}(t) = \sum_{j=1}^{n} q_j 1(Z_j \leqslant t),\ (q_1, \ldots, q_n) \sim$ Dirichlet $(n; 1, \ldots, 1)$ independent of others. This is equivalent to generating $R^{\#}_{m,n}(t) \sim$ DP $(n, \mathbb{G}_n \circ \bar{F}^{-1})$ evaluated at $t \in \mathscr{D}$, where $\bar{F}$ is generated by Step 1. Let the random realization of AUC corresponding to $R^{\#}_{m,n}$ be denoted by $A^{\#}$, plus some subscript to indicate the index of the BB realization.

Step 3 (*Averaging the ensemble of random ROC curves*): The BB estimate, denoted as $\hat{R}^{\text{BB}}_{m,n}(t)$, is obtained by averaging the random realizations of the ROC curves, i.e. $\hat{R}^{\text{BB}}_{m,n}(t) = \text{mean}(R^{\#}_{m,n}(t))$, $t \in \mathscr{D}$. Similarly, we obtain the BB estimate of AUC denoted as $\hat{A}^{\text{BB}}$ by substituting $\hat{R}^{\text{BB}}_{m,n}(t)$ into (2).

Because of two levels of random variations and averaging over them, the BB estimate is much smoother than the empirical one. Note that we do not need a kernel to smooth it out.

Remark 1

A convenient method for generating $(p_1, \ldots, p_m) \sim$ Dirichlet $(m; 1, \ldots, 1)$ is to generate $w_1, \ldots, w_m \sim$ i.i.d. exponential distribution with rate 1 and substitute $p_i = w_i / \sum_{j=1}^{m} w_j,\ i = 1, \ldots, m$.

In order to compute error estimates for the BB estimators of the ROC curve and associated indices, the above steps need to be repeated K times (where K is a reasonably large number). For example, the BB standard error of $\hat{A}^{\text{BB}}$ is given by

$$s = \sqrt{\frac{1}{K-1} \sum_{l=1}^{K} (A^{\#}_l - \hat{A}^{\text{BB}})^2} \tag{5}$$

In addition, $100(1-\alpha)$ per cent BB credible interval for A can be obtained from

$$\text{the percentiles of } \{A^{\#}_l,\ l = 1, \ldots, N\} \text{ at level } \alpha \tag{6}$$

DOI: 10.1002/sim

To obtain a uniform credible band for R based on BB samples, we may estimate d_α by the $100(1-\alpha)$ per cent percentile of the sample $\sup\{|\psi(R^{(l)}(t))-\psi(\hat{R}(t))|: t\in(0,1), l=1,\ldots,N\}$, where $R^{(l)}(t)$ and $\hat{R}(t)$ are lth random realization and BB estimate of $R(t)$, respectively, and substitute $\hat{d}_\alpha$ in (4).

Remark 2

Multivariate measurements in the ROC analysis appear fairly common when subjects undergo two diagnostic tests [26]. Our method immediately extends to the multivariate situation where the ROC function is estimated componentwise for measurement vectors of both disease and non-disease groups. One of the easiest solutions is to apply a common set of BB weights to all components of measurements corresponding to the disease group and similar to the non-disease group. The rationale behind this assignment is that in this situation, the joint distribution would have been given a Dirichlet process prior, whose non-informative limit gives the BB resampling scheme using the same set of weights for each component.

3. APPLICATION TO TESTING BINORMALITY

Because the binormal model is popularly used in practice, it is important to validate the model assumption before using it. Several methods are available, such as the one based on the linearity property of TPF and FPF on the 'normal-deviate axes', the graphic method mentioned by Swets [27], a residual plot using bootstrap sampling method proposed by Cai and Moskowitz [7] and goodness-of-fit tests proposed by Dorfman and Alf [28], Lin and Mudholkar [29], Bozdogan and Ramirez [30].

Our procedure for testing binormality is motivated as follows. In the binormal model, $H(X)\sim$ Normal$(0,1)$ and $H(Y)\sim$Normal (μ,σ^2), where $H(x)=\Phi^{-1}(F(x))$ and $\mu>0$ by convention; hereafter, Φ^{-1} and Φ denote the quantile function and CDF of the standard normal distribution, respectively. By plugging-in a kernel smoothed empirical estimate $\tilde{F}_m$ of F, we can estimate $H(x)$ by $\hat{H}(x)=\Phi^{-1}(\tilde{F}_m(x))$, where $\tilde{F}_m=\Phi_{\sigma_m}*F_m$, '$*$' stands for the convolution operation and σ_m is the bandwidth; see Zou *et al.* [14], Lloyd [15], Zhou and Harezlak [31], among others, for a discussion about the choice of the bandwidth. Now μ and σ can be estimated by the sample mean and sample standard deviation of $\hat{H}(Y_1),\ldots,\hat{H}(Y_n)$, which are denoted by $\hat{\mu}$ and $\hat{\sigma}$, respectively. Under the null hypothesis that the binormal model is true, the ROC function is given by $R(t)=\Phi(a+b\Phi^{-1}(t))$, where $a=\mu/\sigma$, $b=1/\sigma$.

Thus, we consider the test statistic $T=\sup_t|\hat{R}(t)-\Phi(\hat{a}+\hat{b}\Phi^{-1}(t))|$, where $\hat{R}(t)$ is the empirical (or the BB) estimate of $R(t)$, $\hat{a}=\hat{\mu}/\hat{\sigma}$ and $\hat{b}=1/\hat{\sigma}$. We reject the null hypothesis of binormality for large value of T, that is, at level α we reject if $T\geqslant c_\alpha$. However, it is hard to analytically compute or approximate the resulting cut-off point c_α. We employ a resampling technique, more specifically the BB, to estimate c_α. In view of the strong approximation result of Gu and Ghosal [22], the sampling distribution of any test statistic $\gamma_{m,n}(F_m,G_n)$ is approximately equal to the BB resampling distribution of $\gamma_{m,n}(F_m^\#,G_n^\#)$ conditional on the data; here $G^\#(u)=\sum_{j=1}^n q_i 1(Y_i\leqslant u)$ and the rest of the notations are identical to the ones introduced in Section 2. In this case, we may define $T^\#=\gamma_{m,n}(F_m^\#,G_n^\#)=\sup_t|R_{m,n}^\#(t)-\Phi(a^\#+b^\#\Phi^{-1}(t))|$, where $R_{m,n}^\#(t)$ is defined in Section 2, $a^\#$ and $b^\#$ are obtained through $\mu^\#$ and $\sigma^\#$ which in turn are estimated by the sample mean and sample standard deviation of $H^\#(Y_1),\ldots,H^\#(Y_n)$, $H^\#(x)=\Phi^{-1}(\tilde{F}_m^\#(x))$ and $\tilde{F}_m^\#=\Phi_{\sigma_m}*F_m^\#$. Then c_α can

DOI: 10.1002/sim

be estimated by $\hat{c}_\alpha = 100(1-\alpha)$ per cent percentile of the BB distribution of $T^\#$ conditional on the given samples. Thus, we reject the null hypothesis of binormality if $T \geqslant \hat{c}_\alpha$.

4. SIMULATION STUDIES

In order to compare the performance of the BB estimator with some of the existing alternative procedures, we conduct simulations under various scenarios:

1. *Comparison of curve fitting with some SP estimation methods*: By examining the IAEs, we shall compare the accuracy of the estimates of the ROC curve obtained by the BB method with the following methods: the kernel density estimator of Zou *et al.* [14] abbreviated as KZ method; Lloyd's kernel CDF estimator [15] abbreviated as LD; parametric method abbreviated as PA, in which case we assume knowing the diagnostic variables' parametric forms and the method of moments is used to estimate the parameters in F and G; BN-G (ROC GLM method by Pepe [5]); BN-T Box-Cox [4] and SP (SP location-scale models by Pepe [6, p. 112]). Note that BN-G and BN-T assume a binormal model. The purpose of including the parametric method is to check whether the loss of efficiency of the BB method is significant when the parametric assumption holds true. We also compare coverage probabilities and corresponding average lengths of 90 per cent credible intervals for AUC obtained by the BB method with BN-G, BN-T and SP methods. The AUC estimates proposed by Zou *et al.* [14] and Lloyd [15] will not be presented in this simulation. For BN-G, BN-T and SP methods, the bootstrap is used to estimate standard errors or construct confidence intervals. The distributions of (F, G) used to generate the data are chosen as lognormal, location-scale exponential, gamma and beta (abbreviated as A, B, C, D, respectively) for different combinations of the parameters (see Table I). We replicate each simulation 1000 times and resample 1000 times in each replication.

From the simulation results shown in Table I and Figure 2 (the boxplots of IAE are shown using the first data set of A, B, C, D, respectively in Table I), we observe that the proposed BB method performs well with regard to accuracy and robustness. The IAEs obtained by the PA method can be regarded as the true ones and the IAEs obtained by the BB method and the BN-T method are comparable with those obtained by the PA method. However, the BN-T method gives lower coverage probabilities of AUC in some cases (see Table I). The BB method gives the better ROC curve estimate than the kernel estimates proposed by Zou *et al.* [14] and Lloyd [15].

2. *Comparison with some nonparametric estimation methods*: There are several nonparametric estimation methods available to estimate the AUC. Qin and Zhou [19] conducted extensive simulations to compare the accuracy and efficiency of the estimates using various methods, which are EL [19], MW (Mann–Whitney two-sample rank statistics), LT (by Pepe [6, p. 107]), standard percentile bootstrap (PB) and percentile-t bootstrap (PTB). Two out of the three simulation models used are the same as those used in Qin and Zhou [19]. They are normal with mean and standard deviation $(0, 1)$ and $(5^{1/2}\Phi^{-1}(\mathrm{AUC}), 2)$ for F and G, respectively; exponential with rate 1 for F and rate $(1/\mathrm{AUC}-1)$ for G, where the probability density function of exponential distribution with rate λ is defined by $f_\lambda(x) = \lambda \mathrm{e}^{-\lambda x}$. We shall compare the performance of the BB estimator with these estimators. From simulation results (see Table II), the BB estimator performs well, especially, the BB intervals tend to be shorter.

Testing binormality assumption: Limited simulation results shown in Table III for testing binormality indicate that the test is consistent, and somewhat too conservative. More investigation will

DOI: 10.1002/sim

Table I. Coverage probabilities of AUC and the corresponding average lengths of the 90 per cent CI shown in parentheses.

Data		$m=n=15$				$m=n=50$			
u_x, σ_x	u_y, σ_y	BB	BN-G	BN-T	SP	BB	BN-G	BN-T	SP
A									
0, 1	1, 1	0.899	0.886	0.866	0.923	0.893	0.871	0.882	0.900
		(0.262)	(0.254)	(0.250)	(0.276)	(0.149)	(0.141)	(0.143)	(0.176)
0, 1	3, 3	0.861	0.859	0.801	0.820	0.886	0.890	0.873	0.150
		(0.230)	(0.233)	(0.205)	(0.275)	(0.136)	(0.134)	(0.122)	(0.162)
B									
0, 1	1, 1	0.875	0.862	0.860	0.880	0.886	0.888	0.856	0.885
		(0.305)	(0.288)	(0.285)	(0.304)	(0.172)	(0.159)	(0.158)	(0.167)
0, 1	3, 3	0.854	0.857	0.910	0.930	0.902	0.772	0.925	0.919
		(0.098)	(0.097)	(0.076)	(0.110)	(0.057)	(0.062)	(0.045)	(0.065)
C									
1, 1	2, 1	0.886	0.876	0.840	0.873	0.886	0.851	0.860	0.867
		(0.244)	(0.234)	(0.223)	(0.244)	(0.140)	(0.132)	(0.133)	(0.146)
1, 1	5, 3	0.852	0.861	0.858	0.882	0.906	0.837	0.886	0.847
		(0.101)	(0.105)	(0.084)	(0.108)	(0.059)	(0.061)	(0.051)	(0.057)
D									
0.15, 0.15	0.2, 0.3	0.891	0.878	0.867	0.865	0.898	0.892	0.774	0.821
		(0.321)	(0.323)	(0.336)	(0.348)	(0.182)	(0.177)	(0.212)	(0.184)
0.15, 0.15	0.5, 0.45	0.896	0.876	0.874	0.834	0.886	0.912	0.796	0.662
		(0.333)	(0.332)	(0.440)	(0.341)	(0.189)	(0.187)	(0.266)	(0.211)

Our simulation results are based on 1000 simulated data sets and corresponding 1000 BB resamples. Data are generated by lognormal, location-scale exponential, gamma and beta distributions (abbreviated as A, B, C, D, respectively) with different combinations of the parameters (A: X and Y data sets are generated from the lognormal with corresponding normal parameters (u_x, σ_x) and (u_y, σ_y), respectively; B: X's and Y's are generated from the exponential distribution with rate 0.5 and the location and scale parameters (u_x, σ_x) and (u_y, σ_y), respectively; C: X's and Y's are generated from gamma distribution with mean and standard error (u_x, σ_x) and (u_y, σ_y), respectively; D: X's and Y's are generated from beta distribution with mean and standard error (u_x, σ_x) and (u_y, σ_y), respectively). The grid points on $[0, 1]$ are chosen at equal intervals of length 0.05.

be needed to study its properties, design improvements and compare our test procedure with other alternatives.

5. REAL DATA ANALYSES

Two examples will be utilized to illustrate how the BB method can be used to construct a credible band for the ROC curve and a credible interval for the AUC estimate.

1. We will use the data set published by Wieand *et al.* [32]. This study was based on 51 patients as control group diagnosed as pancreatitis and 90 patients as case group diagnosed as pancreatic cancer by two biomarkers, which were a cancer antigen (CA 125) and a carbohydrate antigen (CA 19-9). For the purpose of illustration, we only choose biomarker CA 19-9. The BB estimates are based on 5000 resamples and grid points at even intervals of length 0.01 on $[0, 1]$. We only consider a pointwise 90 per cent credible band in this case (see Figure 3).

DOI: 10.1002/sim

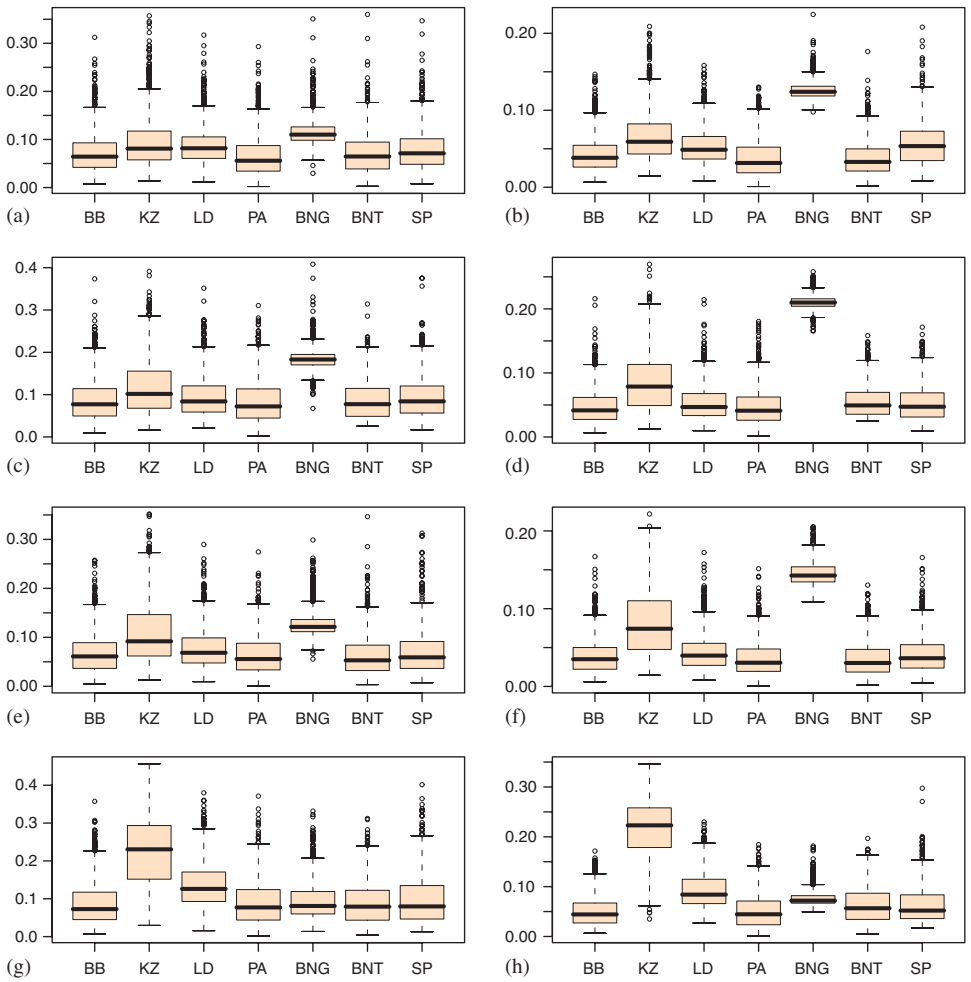

Figure 2. (a) and (b) The boxplots of IAE by the lognormal data sets; (c) and (d) location-scale exponential data sets; (e) and (f) gamma data sets; and (g) and (h) beta data sets. Left and right sides of the graphs are shown for $m=n=15$ and $m=n=50$, respectively. In each graph unit, boxplots of IAE are simulated by BB, Zou *et al.* [14], Llody [15], parametric method, BN-G, BN-T and SP sequentially using the first data set of A, B, C, D, respectively in Table I.

Our BB estimate (corresponding 95 per cent credible interval) of AUC is 0.8542 which is similar to those of Qin and Zhang [12] and Wan and Zhang [33], but the corresponding confidence interval (0.7834, 0.8995) is slightly narrower.

Using the procedure given in Section 3 for testing binormality, we fail to reject the binormality assumption of biomarker CA 125 and CA 19-9 at an alpha level of 0.05, based on $T=0.1427$ and $c_{0.05}=0.2943$ for CA 125, and $T=0.3978$ and $c_{0.05}=0.5349$ for CA 19-9. These results are based on 1000 BB resamples and grid points at even intervals of length 0.05 on $[0,1]$.

DOI: 10.1002/sim

Table II. Coverage probabilities and corresponding average lengths of 95 per cent CI for AUC (in parentheses) obtained by BB and other nonparametric methods based on our simulation and the information contained in Qin and Zhou [19].

Data/AUC	BB	EL	MW	LT	PB	PTB
Normal						
0.8	0.9438	0.9407	0.9379	0.9538	0.9300	0.9690
	(0.1765)	(0.1783)	(0.1808)	(0.1808)	(0.1746)	(0.1971)
0.9	0.9315	0.9352	0.9204	0.9468	0.9150	0.9700
	(0.1234)	(0.1281)	(0.1271)	(0.1326)	(0.1228)	(0.1591)
0.95	0.9066	0.8964	0.8818	0.9289	0.8840	0.9490
	(0.0823)	(0.0874)	(0.0850)	(0.0930)	(0.0814)	(0.1360)
Exponential						
0.8	0.9465	0.9446	0.9394	0.9551	0.9270	0.9570
	(0.1708)	(0.1725)	(0.1746)	(0.1748)	(0.1692)	(0.1887)
0.9	0.9396	0.9321	0.9200	0.9482	0.9240	0.9740
	(0.1212)	(0.1254)	(0.1247)	(0.1290)	(0.1198)	(0.1501)
0.95	0.9049	0.8977	0.8817	NA	0.9000	0.9460
	(0.0823)	(0.0881)	(0.0859)	NA	(0.0838)	(0.1427)

Our simulation results are based on 10 000 simulated data sets and corresponding 1000 BB resamples. The grid points on $[0, 1]$ are chosen at equal intervals of length 0.005, $(m, n) = (50, 50)$.

Table III. Rejection rates for binormality assumption based on BB testing procedure at $\alpha = 0.05$.

Data		Rejection rate		
(u_x, σ_x)	(u_y, σ_y)	$m = n = 15$	$m = n = 50$	$m = n = 100$
A $(0, 1)$	$(1, 1)$	0	0	0
D $(0.15, 0.15)$	$(0.2, 0.3)$	0.191	0.688	0.916

Our simulation results are based on 1000 simulated data sets and corresponding 1000 BB resamples. Data are generated by lognormal, and beta distributions (abbreviated as A, D, respectively) with parameters (A: X and Y data sets are generated from the lognormal with corresponding normal parameters (u_x, σ_x) and (u_y, σ_y), respectively; D: X's and Y's are generated from beta distribution with mean and standard error (u_x, σ_x) and (u_y, σ_y), respectively). The grid points on $[0, 1]$ are chosen at equal intervals of length 0.05.

2. We also analyze the data sets published by Titomir *et al.* [34] who proposed a new approach to detect the left and right ventricular hypertrophies (LVH and RVH). This study was based on 147 subjects with LVH, 60 subjects with RVH and 143 healthy subjects without hypertrophy. The gold standard test relies on clinical and instrumental data, such as roentgenography and echocardiography. Some noninvasive electrocardiographic measurements can better distinguish the subject from the LVH and RVH by dipole electrocardiotopography. These measurements could be spatiotemporally related to the heart activation process. Titomir *et al.* [34] defined the new measurements denoted as ILVH and LRVH for LVH and RVH, respectively, where ILVH = integral indices of activation duration for the left ventricles (IDLV)$*$(maximum depolarization vector and the QRS interval duration), IRVH = integral indices of activation duration for the right ventricles (IDRV)$*$(module of the vector with components equal to the waves R_z and S_x of the scalar vectorcardiographic curves). The new measurements ILVH and IRVH were compared with the sums of wave amplitudes, $R_x + S_z$ and $R_z + S_x$ for LVH and RVH, respectively.

DOI: 10.1002/sim

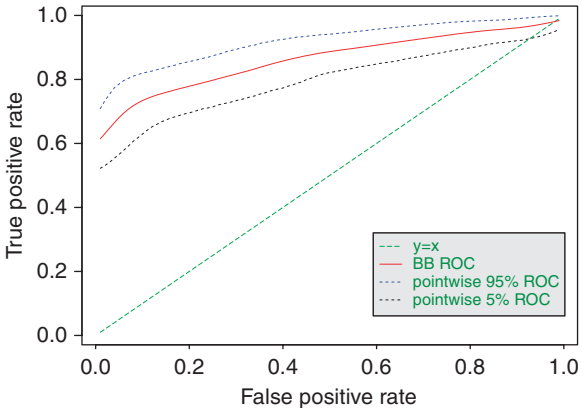

Figure 3. Pointwise 90 per cent credible band of the ROC curve using biomarker CA 19-9.

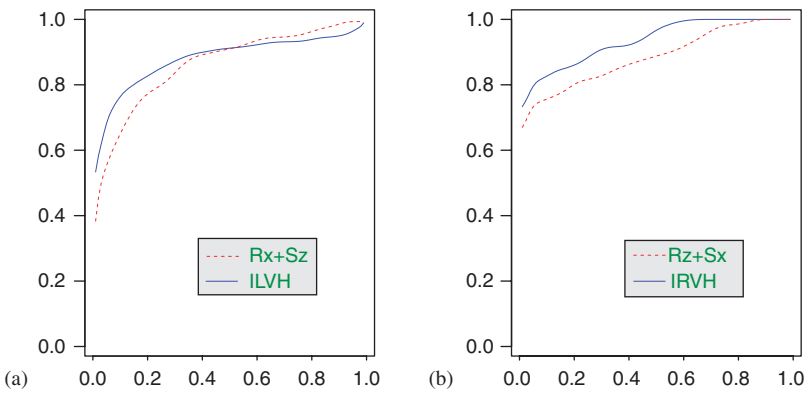

Figure 4. (a) and (b) BB estimates of ROC curves for diagnostic tests for left and right ventricular hypertrophies are based on 5000 resamples and grid points at equal intervals of length 0.01, respectively. Diagnostic tests ILVH and IRVH are plotted in solid lines, tests $R_x + S_z$ and $R_z + S_x$ are plotted in dotted lines.

Because each patient had more than one diagnostic test, we will apply the BB estimation method under multivariate setting. The resulting BB estimates of the ROC curves for diagnostic measurements ILVH and $R_x + S_z$ for LVH, IRVH and $R_z + S_x$ for RVH are shown in Figure 4(a) and (b), respectively. We used 5000 BB resamples. The grid points are chosen at even intervals of length 0.01 on $[0, 1]$. The corresponding BB estimates of pAUC ($t \in [0, 0.3]$) and AUC for LVH and RVH, along with their differences and 99 per cent credible intervals are provided in Tables IV and V, respectively. pAUC or AUC is chosen for LVH and RVH based on clinical consideration to compare the accuracy of the tests. The ILVH measurement comes out to be significantly better than $R_x + S_z$, when $t \in [0, 0.3]$. Similarly, the IRVH measurement turns out to be significantly better than $R_x + S_z$, when $t \in [0, 1]$.

Our results are consistent with those of Titomir *et al.* [34]. The difference lies in that they use normal approximation, whereas we use the Bayesian bootstrap technique.

DOI: 10.1002/sim

Table IV. Comparison of diagnostic tests ILVH and $R_x + S_z$ for left ventricular hypertrophy based on partial AUC ($t \in [0, 0.3]$).

	BB's pAUC		Difference of pAUC		
Diagnostic test	Estimate	Std. dev.	Estimate	Std. dev.	99 per cent CI
ILVH	0.2233	0.0102	0.0242	0.0096	(0.0019, 0.0512)
$R_x + S_z$	0.1991	0.0123			

The BB estimate of pAUC is based on 5000 BB resamples. The grid points on $[0, 1]$ are chosen at equal intervals of length 0.01.

Table V. Comparison of diagnostic tests IRVH and $R_z + S_x$ for right ventricular hypertrophy based on AUC.

	BB's AUC		Difference of AUC		
Diagnostic test	Estimate	Std. dev.	Estimate	Std. dev.	99 per cent CI
ILVH	0.9175	0.0196	0.0503	0.0132	(0.0239, 0.0926)
$R_x + S_z$	0.8672	0.0297			

The BB estimate of AUC is based on 5000 BB resamples. The grid points on $[0, 1]$ are chosen at equal intervals of length 0.01.

APPENDIX: MATLAB CODE

The Matlab code to implement our BB estimate of the ROC curve based on one test is given as follows:

```
%Given data: x,      m observations from nondisease group
%            y,      n observations from disease group
%            grid,  the length of equal intervals of FPF
%            rep,    resample size

%Define FPF and helper vectors, based on the information given before;
t= [grid:grid:1-grid] % FPF vector
ot=ones(length(t),1)  % vector of 1 with the same length as vector t
onx=ones(m,1);ony=ones(n,1)
% vectors of 1 with the same length as x and y, respectively

%AUC function (using Simpson's method);
function [auc] =auc(roctrue,grid) %input ROC curve vector as roctrue.
auc=1/3*grid*(roctrue(1)+roctrue(length(roctrue))
  +2*sum(roctrue(2:(length(roctrue)-1)))
  +2*sum(roctrue(2:2:(length(roctrue)-1))))

%%%%%%%%%%%%%%%%%%%%BB estimate of ROC, AUC;
for r=1:rep
```


DOI: 10.1002/sim

```
% note: to generate Dirichlet weight vectors p and q
p=exprnd(1,1,m);p=p/sum(p);q=exprnd(1,1,n);q=q/sum(q);
z=p*(x*ony'> onx*y');
roc(r,:) = q*(z'*ot'<ony*t);
aucbb(r) = auc(roc(r,:),grid);
end;
rocbb=mean(roc)% rocbb--BB estimate of ROC
aucbb=auc(rocbb, grid)% BB estimate of AUC
```

For the multivariate setting, the Matlab code can be obtained easily following Remark (2). Other sampling error information can be obtained easily.

ACKNOWLEDGEMENTS

We want to thank an associate editor and two reviewers for their dedicated comments. Also, we sincerely thank Eduard Aidu and Vladimir Trunov for providing us their data sets. Research of the second author is partially supported by NSF grant number DMS-0349111.

REFERENCES

1. Green DM, Swets JA. *Signal Detection Theory and Psychophysics*. Wiley: New York, 1966.
2. Hsieh F, Turnbull BW. Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *The Annals of Statistics* 1996; **24**:25–40.
3. Metz CE, Herman BA, Shen J. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. *Statistics in Medicine* 1998; **17**:1033–1053.
4. Zou KH, Hall WJ. Two transformation models for estimating an ROC curve derived from continuous data. *Journal of Applied Statistics* 2000; **27**:621–631.
5. Pepe MS. An interpretation for the ROC curve and inference using GLM procedures. *Biometrics* 2000; **56**: 352–359.
6. Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford Statistical Science Series. Oxford University Press: Oxford, 2003.
7. Cai T, Moskowitz C. Semi-parametric estimation of the binormal ROC curve for a continuous diagnostic test. *Biostatistics* 2004; **5**:573–586.
8. Goddard MJ, Hinberg I. Receiver operator characteristic (ROC) curves and non-normal data: an empirical study. *Statistics in Medicine* 1990; **9**:325–337.
9. Dorfman DD, Berbaum KS, Metz CE, Lenth RV, Hanley JA, Dagga HA. Proper receiver operating characteristic analysis: the bigamma model. *Academic Radiology* 1997; **4**:138–149.
10. Zou KH, Warfield SK, Bharatha A, Tempany CM, Kaus MR, Haker SJ, Wells III WM, Jolesz FA, Kikinis R. Statistical validation of image segmentation quality based on a spatial overlap index. *Academic Radiology* 2004; **11**:178–189.
11. Li G, Tiwari RC, Wells MT. Semiparametric inference for a quantile comparison function with applications to receiver operating characteristic curves. *Biometrika* 1999; **86**:487–502.
12. Qin J, Zhang B. Using logistic regression procedures for estimating receiver operating characteristic curves. *Biometrika* 2003; **90**:585–596.
13. Hall P, Zhou XH. Nonparametric estimation of component distributions in a multivariate mixture. *The Annals of Statistics* 2003; **31**:201–224.
14. Zou KH, Hall WJ, Shapiro DE. Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. *Statistics in Medicine* 1997; **16**:2143–2156.
15. Lloyd CJ. Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems. *Journal of the American Statistical Association* 1998; **93**:1356–1364.
16. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology* 1975; **12**:387–415.

DOI: 10.1002/sim

17. Brownie C, Simonoff JS, Hochberg Y, Reiser B. Estimating $\Pr(X<Y)$ in categorized data using 'ROC' analysis. *Biometrics* 1988; **44**:615–621.
18. 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.
19. Qin G, Zhou XH. Empirical likelihood inference for the area under the ROC curve. *Biometrics* 2006; **62**:613–622.
20. Li G, Tiwari RC, Wells MT. Quantile comparison functions in two-sample problems, with application to comparisons of diagnostic markers. *Journal of the American Statistical Association* 1996; **91**:689–698.
21. Rubin DB. The Bayesian bootstrap. *The Annals of Statistics* 1981; **9**:130–134.
22. Gu J, Ghosal S. Strong approximations for resample quantile processes and application to ROC methodology. *Journal of Nonparametric Statistics* 2008; **20**:229–240.
23. Moise A, Clement B, Raissis M, Nanopoulos P. A test for crossing receiver operating characteristic (ROC) curves. *Communications in Statistics—Theory and Methods* 1988; **17**:1985–2003.
24. Efron B. Bootstrap methods: another look at the jackknife. *Annals of Statistics* 1979; **7**:1–26.
25. Sethuraman J. A constructive definition of Dirichlet priors. *Statistica Sinica* 1994; **4**:639–650.
26. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* 1983; **148**:839–843.
27. Swets JA. Form of empirical ROCs in discrimination and diagnostic tasks: implications for theory and measurement of performance. *Psychological Bulletin* 1986; **99**:181–198.
28. Dorfman DD, Alf E. Maximum likelihood estimation of parameters of signal detection theory—a direct solution. *Psychometrika* 1968; **33**:117–124.
29. Lin C, Mudholkar G. A simple test for normality against asymmetric alternatives. *Biometrika* 1980; **67**:455–461.
30. Bozdogan H, Ramirez DE. Testing for model fit: assessing the Box–Cox transformations of multivariate data to 'near' normality. *Computational Statistics Quarterly* 1986; **3**:127–150.
31. Zhou XH, Harezlak J. Comparison of bandwidth selection methods for kernel smoothing of ROC curves. *Statistics in Medicine* 2002; **21**:2045–2055.
32. 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.
33. Wan S, Zhang B. Smooth semiparametric receiver operating characteristic curves for continuous diagnostic tests. *Statistics in Medicine* 2007; **26**:2565–2586.
34. Titomir LI, Trunov VG, Aidu EAI, Sakhnova TA, Blinova EV. New approaches to the diagnosis of left and right ventricular hypertrophy by means of dipolar electrocardiotopography. *The Anatolian Journal of Cardiology* 2007; **7**(Suppl. 1):29–31.

DOI: 10.1002/sim