UW Biostatistics Working Paper Series

5-21-2003

Improved Confidence Intervals for the Sensitivity at a Fixed Level of Specificity of a Continuous-Scale Diagnostic Test

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

Gengsheng Qin
Georgia State University, gqin@mathstat.gsu.edu

1. Introduction

The accuracy of a test can be measured by its ability to correctly classify patients as diseased or non-diseased. When the response of a test is binary, the accuracy of the test is usually represented by its sensitivity and specificity. The sensitivity is defined as the proportion of positive test results among diseased patients, and the specificity is defined as the proportion of negative test results among non-diseased patients. When dealing with a continuous-scale test, a cut-off point is usually chosen so that a fixed value of specificity is achieved (typically the 80%, 90%, or 95%)[1]. Since in practice the distribution of test results for non-diseased patients is unknown, the cut-off point corresponding to a fixed value of specificity has to be estimated using the data from non-diseased patients. Therefore, we have to adjust for the uncertainty associated with the estimated cut-off point when we construct a confidence interval for the sensitivity of the test at the cut-off point that provides this fixed value of specificity.

Linnet (1987) [2] proposed both parametric and non-parametric methods for constructing confidence intervals for the sensitivity of a test at a fixed value of specificity, accounting for the random variation associated with the estimated cut-off point. Platt et al. (2000) [1] pointed out several shortcomings in Linnet's methods and then proposed to use Efron's bias-corrected acceleration (BCa) bootstrap interval; they have shown through simulation studies that the BCa bootstrap interval has better coverage accuracy than Linnet's interval. However, as shown in this paper, this BCa bootstrap interval can still have poor coverage accuracy in many circumstances.

In this paper, we proposed two new intervals for the sensitivity of a test at a fixed value of specificity. The proposed intervals may be regarded as extensions of Agresti and Coull's interval [3] for a binomial proportion, but the extensions are non-trivial because the sensitivity at a fixed value of specificity is not a simple binomial proportion. Simulation studies indicated that the new intervals have better coverage accuracy and shorter length than the BCa bootstrap interval.

This paper is organized as follows. In section 2 we state our problem and introduce necessary notation. In Section 3 we describe existing methods for interval estimation of the sensitivity of a continuous-scale test at a fixed level of its specificity. In section 3 we propose two new intervals that are based on extensions of Agresti and Coull's idea for confidence intervals of binomial proportions. In Section 4 we conduct simulation studies to assess the finite-sample performance of the new intervals with the best existing interval. In Section 5 we illustrate the application of the proposed method in two real examples.

2. Problem and notation

Let Y and X be results of a continuous-scale test for a diseased and a non-diseased patient, respectively. For a given cut-off point c, we can define sensitivity and specificity of the test as

$$Se = P(Y \geq c),\, Sp = P(X \leq c),$$

respectively. Let F_1 and F_0 be the distribution functions of Y and X, respectively. We can then write $Se = 1 - F_1(c)$ and $Sp = F_0(c)$. Therefore, for a fixed value of specificity at p, the corresponding sensitivity of the test is $R(p) = 1 - F_1(F_0^{-1}(p))$, where $F_0^{-1}(p)$ is the inverse function of $F_0(p)$. Our goal in this paper is to construct confidence intervals for $R(p)$ at a fixed level of specificity p.

Let $Y_1, \ldots, Y_m$ be test results of a random sample of diseased patients and $X_1, \ldots, X_n$ be test results of a random sample of non-diseased patients. Based on these observations, we wish to construct $(1 - \alpha)100\%$ confidence intervals for the sensitivity $R(p)$ of the test at a fixed value of specificity p.

3. Existing Confidence Intervals

Note that $R(p) = P(Y_1 \geq F_0^{-1}(p))$. So an obvious estimator for $R(p)$ is the observed sensitivity at p-th sample quantile of test results in the sample of non-diseased patients. More specifically, let $\widehat{F}_0$ be the empirical distribution function based on $X_1, \ldots, X_n$. Then, we can define this obvious estimator by

$$\widehat{R}(p) = \frac{\sum_{i=1}^{m} I_{[Y_i \geq \widehat{F}_0^{-1}(p)]}}{m}, \qquad (1)$$

where I_A is an indicator variable, and $\widehat{F}_0^{-1}(p) = \sup\{t : \widehat{F}_0(t) \leq p\}$.

Treating $\widehat{F}_0^{-1}(p)$ as fixed, we would obtain the naive variance of $\widehat{R}(p)$ as

$$\widehat{Var}_N(\widehat{R}(p)) \quad = \quad \frac{\widehat{R}(p)(1 - \widehat{R}(p))}{m}$$

and the corresponding $(1-\alpha)100\%$ confidence interval (hereafter NV interval) for $R(p)$ as

$$\left(\widehat{R}(p) - z_{1-\alpha/2}\sqrt{\widehat{Var}_N(\widehat{R}(p))}, \widehat{R}(p) + z_{1-\alpha/2}\sqrt{\widehat{Var}_N(\widehat{R}(p))} \right),$$

where z_α is the α-th quantile of the standard normal distribution [1].

Linnet (1987) [2] showed that the coverage probability of the above NV interval could fall far below the nominal confidence level, as expected, because it only considered the variability in the test values from diseased patients and ignored the random variation due to the estimated quantile of the test values from non-diseased patients. Recogning this problem, Linnet (1987) [2] proposed an idea to account for this extra random variation in the variance formula for $\widehat{R}(p)$, but he didn't give the general formula explicitly. However, it is an easy task to derive a general variance formula for $\widehat{R}(p)$ using Linnet's idea. The key assumption in Linnet's variance formula for $\widehat{R}(p)$ is that the variance of $\widehat{R}(p)$ can be approximated by the sum of the two terms:

$$Var(\widehat{R}(p)) = \frac{R(p)(1 - R(p))}{m} + f_1^2(F_0^{-1}(p))Var(\widehat{F}_0^{-1}(p)),$$

where f_1 is the density of the distribution function F_1, and $Var(\widehat{F}_0^{-1}(p))$ represents the variance of the sample quantile $\widehat{F}_0^{-1}(p)$. Note that the sample quantile $\widehat{F}_0^{-1}(p)$ has the following asymptotic normal distribution:

$$\widehat{F}_0^{-1}(p) - F_0^{-1}(p) \sim N\left(0, \frac{(1-p)p}{nf_0^2(F_0^{-1}(p))}\right),$$

where f_0 is the density of the distribution function F_0. Therefore, Linnet's variance formula for $\widehat{R}(p)$ can be written as

$$Var(\widehat{R}(p)) = \frac{R(p)(1-R(p))}{m} + \frac{(1-p)p}{nf_0^2(F_0^{-1}(p))} f_1^2(F_0^{-1}(p)). \tag{2}$$

It turns out that the variance $Var(\widehat{R}(p))$ for $\widehat{R}(p)$ has some nice asymptotic properties. In fact several authors have shown that when both m and n are large, $\widehat{R}(p)$ has an approximately normal distribution with mean $R(p)$ and variance $Var(\widehat{R}(p))$, given by (2) [4].

By substituting unknown quantities in (2) by their corresponding sample estimates, we obtain the following estimated variance for $\widehat{R}(p)$:

$$\widehat{Var}_{LN}(\widehat{R}(p)) = \frac{\widehat{R}(p)(1-\widehat{R}(p))}{m} + \frac{(1-p)p}{n\widehat{f}_0^2(\widehat{F}_0^{-1}(p))} \widehat{f}_1^2(\widehat{F}_0^{-1}(p)), \tag{3}$$

where $\widehat{f}_1$ and $\widehat{f}_0$ are the empirical density estimates from samples of diseased and non-diseased patients, respectively, and $\widehat{F}_0^{-1}(p)$ is the p-th sample quantile in the sample of non-diseased patients. It is worth noting that Linnet's variance estimate for $\widehat{R}(p)$, given by Platt et al. (2000) [2], contains some minor typos (see their formula for $\widehat{V_{LNP}(\hat{S}e)}$ on page 315 of their paper). However, the typos in their paper should not affect their simulation conclusions. Using $\widehat{Var}_{LN}(\hat{R}(p))$, we obtain the $(1-\alpha)100\%$ Linnet confidence interval (hereafter LN interval) for $R(p)$ as follows:

$$\left(\widehat{R}(p) - z_{1-\alpha/2}\sqrt{\widehat{V}ar_{LN}(\widehat{R}(p))}, \widehat{R}(p) + z_{1-\alpha/2}\sqrt{\widehat{V}ar_{LN}(\widehat{R}(p))}\right).$$

From the variance formula for $\widehat{Var}_{LN}(\widehat{R}(p))$ in (3), we see that the performance of Linnet' interval may be greatly affected by poor empirical density estimation. Platt et al. (2000)[1] studied this issue and found via a simulation study that Linnet' interval could perform poorly under certain circumstances, particularly when test responses of a diseased and a non-diseased patient were not normally distributed. To avoid estimation of the variance of $\widehat{R}(p)$, Platt et al. (2000) [1] proposed to use Efron's bias correction and acceleration (BCa) bootstrap method to construct a confidence interval for $R(p)$.

4. New Confidence Intervals

We first assume that the distribution function F_0 for test results of non-diseased patients is known and let $W_i = I_{[Y_i \geq F_0^{-1}(p)]}$, $i = 1, \cdots, m$. Then, we can see that W_i's are Bernoulli random variables with the proportion $R(p) = P\left(Y_i \geq F_0^{-1}(p)\right)$. Letting $R_0(p) = \sum_{i=1}^m W_i/m$, we obtain the standard $(1-\alpha)100\%$ Wald interval for $R(p)$ as follows:

$$\left(R_0(p) - z_{1-\alpha/2}\sqrt{R_0(p)(1-R_0(p))/m}, R_0(p) + z_{1-\alpha/2}\sqrt{R_0(p)(1-R_0(p))/m}\right).$$

However, it has been shown that the Wald interval has poor coverage accuracy, particularly for small sample sizes [5]. In order to improve the coverage accuracy of the Wald interval for binomial proportions, Agresti and Coull (1998) [3] proposed an easy-to-use interval, called the AC interval. Applying the AC interval to our setting, we obtain the following $(1-\alpha)100\%$ confidence interval for $R(p)$:

$$\left(\widetilde{R}_0(p) - z_{1-\alpha/2}\sqrt{\widehat{Var}_{AC}(\widetilde{R}_0(p))}, \widetilde{R}_0(p) + z_{1-\alpha/2}\sqrt{\widehat{Var}_{AC}(\widetilde{R}_0(p))}\right) \tag{4}$$

where

$$\widetilde{R}_0(p) = \frac{\sum_{i=1}^m I_{[Y_i \geq F_0^{-1}(p)]} + z_{1-\alpha/2}^2/2}{m + z_{1-\alpha/2}^2}, \tag{5}$$

and

$$\widehat{Var}_{AC}(\widetilde{R}_0(p)) = \frac{\widetilde{R}_0(p)(1-\widetilde{R}_0(p))}{m + z_{1-\alpha/2}^2}. \tag{6}$$

Since $z_{1-\alpha/2}$ is approximately equal to 2, when $\alpha = 0.05$, the AC interval is regarded as the adjusted Wald interval by adding two successes and two failures to Bernoulli observations. Agresti and Coull

(1998) [3] strongly recommended this interval because their simulation study has shown that it has good coverage accuracy even for small sample sizes.

Since $F_0^{-1}(p)$ is unknown, we can not directly use the AC interval for $R(p)$, defined by (4). One way of overcoming this problem is to replace the unknown $F_0^{-1}(p)$ in the formula (4) by its sample quantile $\widehat{F}_0^{-1}(p)$, resulting in a plug-in type interval,

$$\left(\widetilde{R}(p) - z_{1-\alpha/2}\sqrt{\widehat{Var}_{AC}(\widetilde{R}(p))}, \widetilde{R}(p) + z_{1-\alpha/2}\sqrt{\widehat{Var}_{AC}(\widetilde{R}(p))}\right)$$

where

$$\widetilde{R}(p) = \frac{\sum_{i=1}^{m} I_{[Y_i \geq \widehat{F}_0^{-1}(p)]} + z_{1-\alpha/2}^2/2}{m + z_{1-\alpha/2}^2}, \tag{7}$$

and

$$\widehat{Var}_{AC}(\widetilde{R}(p)) = \frac{\widetilde{R}(p)(1-\widetilde{R}(p))}{m + z_{1-\alpha/2}^2}. \tag{8}$$

However, our simulation study, not reported here, has shown that the plug-in type AC interval has poor coverage accuracy.

The problem with the plug-in AC interval is still underestimation of the variance for the estimator $\widetilde{R}(p)$ in (7), which is derived by replacing $F_0^{-1}(p)$ in (5) with $\widehat{F}_0^{-1}(p)$. When we replace $F_0^{-1}(p)$ by its sample quantile $\widehat{F}_0^{-1}(p)$ in $W_i = I_{[Y_i \geq F_0^{-1}(p)]}$, the resulting random variables $I_{[Y_i \geq \widehat{F}_0^{-1}(p)]}$ are no longer independent. Hence, the variance estimate, given by (8), also underestimates the true variance of $\widetilde{R}(p)$.

In this paper we propose a bootstrap method to estimate the variance of $\widetilde{R}(p)$. After we obtain such an appropriate variance estimate, we then apply Agresti and Coull's idea to obtain confidence intervals for $R(p)$. We summarize the procedure for computing the bootstrap variance in the following steps:

1. Draw a resample of size m, Y_i^*'s, with replacement from the diseased sample Y_i's and a

separate resample of size n, X_j^*'s, with replacement from the non-diseased sample X_j's.

2. Calculate the bootstrap version of $\widetilde{R}(p)$,

$$\widetilde{R}^*(p) = \frac{\sum_{i=1}^m I_{[Y_i^* \geq \widehat{F}_0^{-1*}(p)]} + z_{1-\alpha/2}^2/2}{m + z_{1-\alpha/2}^2},$$

   where $\widehat{F}_0^{-1*}(p)$ is the p-th sample quantile based on the bootstrap resample X_j^*'s.

3. Repeat the first two steps B times to obtain the set of bootstrap replications $\{\widetilde{R}^{*b}(p) : b = 1, 2, \cdots, B\}$ (it is recommended that $B \geq 200$; in this paper, we take $B = 500$). Then, the proposed bootstrap variance estimator $V^*(p)$ is defined by

$$V^*(p) = \frac{1}{B-1}\sum_{b=1}^{B}(\widetilde{R}^{*b}(p) - \bar{R}^*(p))^2,$$

   where $\bar{R}^*(p) = (1/B)\sum_{b=1}^B \widetilde{R}^{*b}(p)$.

Now we propose the following two new intervals for $R(p)$. The first $(1-\alpha)100\%$ level interval, called BTI interval, for $R(p)$ is defined by

$$\left(\widetilde{R}(p) - z_{1-\alpha/2}\sqrt{V^*(p)}, \widetilde{R}(p) + z_{1-\alpha/2}\sqrt{V^*(p)}\right)$$

where $\widetilde{R}(p)$ is defined by (7). The second $(1-\alpha)100\%$ level interval, called BTII interval, for $R(p)$ is defined by

$$\left(\bar{R}^*(p) - z_{1-\alpha/2}\sqrt{V^*(p)}, \bar{R}^*(p) + z_{1-\alpha/2}\sqrt{V^*(p)}\right).$$

5. Simulation Studies for the Confidence Intervals

We conducted two simulation studies to evaluate coverage accuracy and interval length of the newly proposed intervals for $R(p)$ when p =80% or 90% in finite-sample sizes. Since Platt et al. (2000) [1] have already shown that their BCa bootstrap interval has better coverage accuracy than currently used methods, including Linnet's interval, for non-normally distributed data and has at least as good as currently used methods when test responses follow normal distributions, in our simulation studies we only included the BCa bootstrap interval for comparison purposes. In both simulation studies, we generated 5,000 random samples of size m from the distribution function F_1 for test responses of diseased patients and another independent random sample of the sample size n from the distribution function F_0 for test responses of non-diseased patients, and we took $(m, n) = (20, 20)$, $(50, 50)$, and $(40, 20)$ in the both simulation studies to represent a small sample size, moderate sample size, and unequal sample size setting. In the simulation studies, the distributions F_0 and F_1 were chosen to represent a setting of normally distributed data and a setting of non-normally distributed data, as used in Platt et al (2000) [1].

In the first simulation study, we chose the distributions F_0 and F_1 to be beta distributions with parameters (a_0, b_0) and (a_1, b_1), respectively. The parameter settings for a_0, a_1, b_0, b_1 were chosen so that various values of test' sensitivity were achieved when its specificity was fixed at either 80% or 90%. In Table 1 we report the corresponding sensitivity when test's specificity is fixed at either 80% or 90%. We display coverage probabilities and interval lengths for the BCa intervals and the two newly proposed intervals (BTI and BTII) with the nominal level of 95% in Table 2 when $n = m = 20$, in Table 3 when $n = m = 50$, and in Table 4 when $m = 40$ and $n = 20$, respectively. From the results in Tables 2-4, we see that generally the both proposed BTI and BTII intervals have better coverage accuracy and shorter interval length than the BCa interval, regardless of sample sizes and true values of sensitivity and specificity considered here. The BTII interval performs the best among the three intervals compared under simulated scenarios here. The improvement of the BTII interval over the BCa interval in coverage accuracy can be huge when sample sizes are samll or unqual and when the test's sensitivity is greater than 0.93. For example, when $n = m = 20$, at the fixed specificity of 80% with the true sensitivity of 93% (Run 2), the coverage probability of the BTII interval with the nominal level of 95% is 92.55% whileas that of the BCa interval with the nominal level of 95% is only 71.25%.

In the second simulation study, we chose the distributions F_0 and F_1 to be the standard normal distribution and a normal distribution with mean μ and variance 1, where μ was chosen to give different levels of the desired sensitivity of the test when test's specificity was fixed at 80% or 90%. Table 5 displays the parameter settings for μ and corresponding sensitivities when the specificity was fixed at either 80% or 90%. We present coverage probabilities and interval lengths for the BCa intervals and the two newly proposed intervals (BTI and BTII) with the nominal level of 95% in Table 6 when $n = m = 20$, in Table 7 when $n = m = 50$, and in Table 8 when $m = 40$ and $n = 20$, respectively. From results in Tables 6-8, we see a very similar pattern on coverage accuracy and interval length of the various intervals as in the first simulation study.

In summary, our simulation studies suggest that the two newly proposed BTI and BTII intervals tend to have better coverage accuracy and shorter interval lengths than the existing BCa interval regardless of the sample sizes and for both normal and non-normal data we considered here. Among the three intervals compared here, the BTII interval tends to have the best coverage accuracy and shortest length. In addition, the new intervals are computanionally simpler than the BCa interval.

6. Two Real Applications

We illustrated the application of the proposed methods in two real studies. The first study investigated the accuracy of dematoscopy in distingushing patients with malignant melanoma (MM) from those without MM; the second study assessed the the accuracy of cerebrospinal fluid CK-BB isoenzyme in predicting the future outcome of severe head trauma.

6.1 Dermoscope Example

The most deadly kind of skin disease is malignant melanoma (MM), and early detection of MM combined with excision of MM is the only way to cure patients with MM. Stolz et al. (1994) [6] studied the accuracy of clinical evaluation with the aid of dermatoscopy in detecting malignant melanoma by using the ABCD rule (Asymmetry, irregular Border, different Colors, and Diameter larger than 6mm). The dermatoscopy is a hand-held instrument for skin surface microscopy at 10 times magnification [7]. The study sample consists of 21 patients with MM and 51 patients with benign melanocytic lesions, and the gold standard used in the study is biopsy. To be sure that dermatoscopy has a high change of ruling out patients without MM, dermatologists want the specificity of dermatoscopy to be at least 90% for detecting patients without MM and want to know what the corresponding sensitivity of dermatoscop is in detecting patients with MM. Therefore, It is an interest to construct a confidence interval for the sensitivity of dermatoscopy when its specificity is fixed at 90%.

The estimated sensivitiy of the dermatoscopy at the fixed 90% level of specificity is 0.71, and the corresponding 95% confidence intervals are: [1.00,1.00] using the BCa bootstrap method, [0.357,1.00] using the BTI bootstrap method, and [0.285,0.934] using the BTII bootstrap method. It is worth noting the BCa bootstrap method gave a degenerated confidence interval. According to our simulation result, we would use [0.285,0.934] from the BTII bootstrap method as the 95% confidence interval for the sensitivity of the dermatoscopy at the fixed 90% level of specificity.

6.2 Severe Head Trauma Example

Hans et al. (1985) [8] conducted a study on assessing the accuracy of cerebrospinal fluid CK-BB isoenzyme measured within 24 hours of injury as a means of predicting the outcome of severe head trauma. Investigators are interested in determining whether patients will have a poor outcome (death, vegetative state or severe disability) after suffering a severe head trauma. A sample of 60 subjects admitted to a hospital with severe head trauma are considered with 19 eventually having moderate to full recovery, with the remaining 41 poor or no recovery (Hans et al., 1985).

Investigators want the specificity of CK-BB to be at least 90% for predicting poor recovery or death from severe head injury. They want to know what the corresponding sensitivity of CK-BB is. Therefore, one main research question is how to construct a confidence interval for the corresponding sensitivity of CK-BB when one chooses a decision threshold to achieve a minimum value of required specificity of 90%.

The estimated sensitivity at the fixed 90% level of specificity is 0.634, and the corresponding 95% confidence intervals are: [0.415,0.829] using the BCa bootstrap method, [0.419,0.825] usig the BTI bootstrap method, and [0.429,0.836] using the BTII bootstrap method. Based on our simulation result, we would use [0.429,0.836] as the 95% confidence interval for the sensitivity of the fluid CK-BB isoenzyme at the fixed 90% level of its specificity.

7. Discussion

When the response of a test is continuous, we need to choose a cut-off point to compute its sensitivity and specificity. A patient is classified as diseased if the test's response is above the chosen cut-off point and as non-diseased if the test's response is below the chosen cut-off point. For a continuous-scale test, we are often interested in estimating the sensitivity of the test at a cut-off point that yields a pre-determined level of specificity [2]. The current best existing interval is Platt's BCa bootstrap interval. Since this interval is based on the traditional Wald-type interval for the binomial proportion, as shown in this paper it can still have poor coverage accuracy. In this paper, we have proposed two new intervals (BTI and BTII) that are based on an improved version of the Wald-type interval. We have shown in the simulations that the proposed BTII has the best performance in term of coverage accuracy and interval length. However, the performance of the proposed BTII interval is still not satisfactory when the true sensitivity is very high (≥ 0.95)

and sample sizes are small. It is a future research topic to improve the coverage accuracy of the proposed intervals when the sensitivity is very high.

If we do not know which level of specificity a test should be fixed at when we are estimating the sensitivity of the test, we need to construct a receiver operating characteristic curve (ROC) to represent the accuracy of the test. Our proposed methods can also be used to construct pointwise confidence bands for an ROC curve. Whereas our method may be considered as a non-parametric method for the interval estimation of the ROC curve of a continuous-scale test, Metz et al. (2000) [9] has proposed a latent semi-parametric method for estimation of the ROC curve. It is a future research topic to compare these two approaches. The use of the jackkinife or bootstrap method has also been considered by Dorfman et al. (1992) [10] for the analysis of ROC curve areas of ordinal-scale tests.

A S-plus code implementing the proposed BTI and BTII intervals are available from the authors.

ACKNOWLEDGMENTS

We would like to thank Dr. Robert W. Platt for sending us his S code for the implementation of the BCa bootstrap interval.

References

[1] Platt RW, Hanley JA and Yang H (2000). Bootstrap confidence intervals for the sensitivity of a quantitative diagnostic test *Statistics in Medicine* 2000 **19**, 313-322.

[2] Linnet K. Comparison of quantitative diagnostic tests: type I error, power, and sample size. *Statistics in Medicine* 1987 **6**, 147-158.

[3] Agresti A and Coull BA. Approximate is better than "exact" for interval estimation of Binomial proportions. *The American Statistician* 1998 **52**, 119-126.

[4] Li G, Tiwari RC, and Well M. Quantile comparison functions in two-sample problems: with applications to comparisons of diagnostic markers. *Journal of the American Statistical Association* 1996 **91**, 689-698.

[5] Brown LD, Cai TT, and DasGupta A. Interval estimation for a binomial proportion. *Statistical Science* 2001 **16**, 101-117.

[6] Stolz W, Bilek P, Landthaler M, Merkle T, and Braun-Falco O. Skin surface microscopy. *Lancet* 1989 **2**, 864-865.

[7] Stolz W, Riemann A, Cognetta AB, Pillet L, Abmayr W, Holzel D, Bilek P, Nachbar F, Landthaler M, and Braun-Falco O. ABCD rule of dermatoscopy: a new practical method for ealy recognition of malignant melanoma. *Eurpean Journal of Dermatology* 1994 **4**, 521-527.

[8] Hans P, Albert A, Born JD, and Chapelle JP. Derivation of a bioclinical index in severe head trauma. *Intensive Care Medicine* 1985 **11**, 186-191.

[9] Metz CE, Herman BA, and Shen JH. Maximum-likelihood estimation of ROC curves from continuously-distributed data. *Stat Med 1998*, 17, 1033-1053.

[10] Dorfman DD, Berbaum, KS, and Metz, CE. Receiver operating characteristic rating analysis: generalization to the population of readers and patients with the jackknife method. *Invest. Radiol* 1992, 27, 723-731.

Table 1. Parameter settings for beta distributions in the first simulation, where a_1 and b_1 denote two parameters in a beta distribution for a diseased patient, and a_0 and b_0 denote two parameters in a beta distribution for a non-diseased patient. The specificity of the test is fixed at either 80% or 90%, and the corresponding sensitivity is calculated based on a given beta distribution.

Run	(a_1, b_1)	(a_0, b_0)	Specificity (p)	Sensitivity ($R(p)$)
1	(4, 1)	(1, 3.5)	0.90	0.95
2	(3, 1)	(1, 3)	0.80	0.93
3	(3, 1)	(1, 3)	0.90	0.85
4	(4, 2)	(2, 4)	0.80	0.82
5	(3, 2)	(2, 3)	0.80	0.55

Table 2. Coverage errors and probabilities and average lengths of the various intervals with the nominal level of 95% when data are generated from beta distributions with $m = n = 20$.

Run	Method	Coverage errors		Coverage	Average
		Lower	Upper	probability	length
1	BCa	0.3575	0.0175	0.6250	0.2640
	BTI	0.0000	0.1905	0.8095	0.2278
	BTII	0.0000	0.1605	0.8395	0.2278
2	BCa	0.2410	0.0465	0.7125	0.3192
	BTI	0.0000	0.0975	0.9025	0.2457
	BTII	0.0000	0.0745	0.9255	0.2457
3	BCa	0.0860	0.0215	0.8925	0.4663
	BTI	0.0520	0.0235	0.9245	0.3741
	BTII	0.0430	0.0110	0.9460	0.3741
4	BCa	0.0485	0.0455	0.9060	0.4883
	BTI	0.0320	0.0300	0.9380	0.3875
	BTII	0.0270	0.0185	0.9545	0.3875
5	BCa	0.0085	0.0565	0.9350	0.6005
	BTI	0.0495	0.0250	0.9255	0.5223
	BTII	0.0405	0.0155	0.9440	0.5223

Note:

Coverage errors refer to the proportions of runs in which the lower (or upper) limit of the confidence interval excluded the true $R(p)$ (each is expected to be 0.025); Coverage probability is the proportions of runs in which the confidence interval contained the true $R(p)$; the average length is the mean of lengths of confidence intervals for $R(p)$.

Table 3. Coverage errors and probabilities and average lengths of the various intervals with the nominal level of 95% when data are generated from beta distributions with $m = n = 50$.

Run	Method	Coverage errors		Coverage	Average
		Lower	Upper	probability	length
1	BCa	0.0875	0.0510	0.8615	0.2059
	BTI	0.0280	0.0230	0.9490	0.1678
	BTII	0.0275	0.0135	0.9590	0.1678
2	BCa	0.0355	0.0495	0.9150	0.2131
	BTI	0.0155	0.0265	0.9580	0.1783
	BTII	0.0220	0.0250	0.9530	0.1783
3	BCa	0.0085	0.0355	0.9560	0.3234
	BTI	0.0280	0.0155	0.9565	0.2829
	BTII	0.0295	0.0140	0.9565	0.2829
4	BCa	0.0105	0.0420	0.9475	0.3112
	BTI	0.0255	0.0215	0.9530	0.2812
	BTII	0.0305	0.0210	0.9485	0.2812
5	BCa	0.0250	0.0375	0.9375	0.4030
	BTI	0.0430	0.0215	0.9355	0.3858
	BTII	0.0320	0.0140	0.9540	0.3858

Table 4. Parameter settings for normal distributions in the second simulation study. The response of a non-diseased patient has the standard normal distribution $N(0, 1)$, and the response of a diseased patient has a normal distribution with mean μ and variance 1 $N(\mu, 1)$. The specificity of the test is fixed at either 80% or 90%, and the corresponding sensitivity is calculated based on $N(\mu, 1)$.

Run	μ	Specificity	Sensitivity
		(p)	$(R(p))$
1	2.9264	0.90	0.95
2	2.5631	0.90	0.90
3	2.1231	0.90	0.80
4	2.4865	0.80	0.95
5	1.6832	0.80	0.80

Table 5. Coverage errors and probabilities and average lengths of the various intervals with the nominal level of 95% when data are generated from normal distributions with $m = n = 20$.

Run	Method	Coverage errors		Coverage	Average
		Lower	Upper	probability	length
1	BCa	0.3695	0.0145	0.6160	0.2511
	BTI	0.0000	0.2605	0.7395	0.2196
	BTII	0.0000	0.2165	0.7835	0.2196
2	BCa	0.1900	0.0620	0.7480	0.3690
	BTI	0.0470	0.0290	0.9240	0.3061
	BTII	0.0465	0.0130	0.9405	0.3061
3	BCa	0.0555	0.0475	0.8970	0.5191
	BTI	0.0455	0.0170	0.9375	0.4278
	BTII	0.0475	0.0080	0.9445	0.4277
4	BCa	0.3535	0.0230	0.6235	0.2519
	BTI	0.0000	0.2335	0.7665	0.1990
	BTII	0.0000	0.2020	0.7980	0.1990
5	BCa	0.0250	0.0655	0.9095	0.5216
	BTI	0.0220	0.0245	0.9535	0.4159
	BTII	0.0190	0.0170	0.9640	0.4159

Table 6. Coverage errors and probabilities and average lengths of the various intervals with the nominal level of 95% when data are generated from normal distributions with $m = n = 50$

Run	Method	Coverage errors		Coverage	Average
		Lower	Upper	probability	length
1	BCa	0.0925	0.0430	0.8645	0.1955
	BTI	0.0230	0.0315	0.9455	0.1572
	BTII	0.0225	0.0175	0.9600	0.1572
2	BCa	0.0145	0.0450	0.9405	0.2745
	BTI	0.0270	0.0165	0.9565	0.2293
	BTII	0.0220	0.0105	0.9675	0.2293
3	BCa	0.0150	0.0430	0.9420	0.3577
	BTI	0.0430	0.0210	0.9360	0.3201
	BTII	0.0365	0.0150	0.9485	0.3201
4	BCa	0.0735	0.0435	0.8830	0.1775
	BTI	0.0205	0.0380	0.9415	0.1443
	BTII	0.0200	0.0225	0.9575	0.1443
5	BCa	0.0105	0.0535	0.9360	0.3291
	BTI	0.0320	0.0255	0.9425	0.2977
	BTII	0.0255	0.0160	0.9585	0.2977

Table 7. 95% confidence intervals for $R(p)$ in the illustrative example.

Fixed level of specificity(p)	method	Estimated sensitivity ($\widehat{R}(p)$)	Confidence intervals	Lengths
0.95	BCa	0.689	(0.524, 0.778)	0.254
	BTI	0.689	(0.561, 0.801)	0.240
	BTII	0.689	(0.560, 0.800)	0.240
0.90	BCa	0.756	(0.678, 0.892)	0.214
	BTI	0.756	(0.643, 0.847)	0.204
	BTII	0.756	(0.627, 0.831)	0.204
0.80	BCa	0.778	(0.678, 0.856)	0.178
	BTI	0.778	(0.682, 0.851)	0.169
	BTII	0.778	(0.680, 0.849)	0.169
0.70	BCa	0.811	(0.656, 0.889)	0.233
	BTI	0.811	(0.706, 0.891)	0.185
	BTII	0.811	(0.711, 0.896)	0.185
0.50	BCa	0.889	(0.778, 0.944)	0.166
	BTI	0.889	(0.801, 0.945)	0.144
	BTII	0.889	(0.798, 0.942)	0.144