UW Biostatistics Working Paper Series

3-14-2005

New Confidence Intervals for the Difference between Two Sensitivities at a Fixed Level of Specificity

Gengsheng Qin
Georgia State University, gqin@gsu.edu

Yu-Sheng Hsu
Georgia State University, matysh@langate.gsu.edu

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

Suggested Citation

Qin, Gengsheng; Hsu, Yu-Sheng; and Zhou, Xiao-Hua, "New Confidence Intervals for the Difference between Two Sensitivities at a Fixed Level of Specificity" (March 2005). *UW Biostatistics Working Paper Series.* Working Paper 244.
http://biostats.bepress.com/uwbiostat/paper244

1. Introduction

The accuracy of a diagnostic test can be measured by its sensitivity and specificity, which are defined as the probabilities of correctly identifying the diseased and the non-diseased individual respectively. In many medical applications, we have two (or more) continuous-scale diagnostic tests to the same set of individuals, some of whom are non-diseases, some diseased. In this situation, it is of interest to know which test is better for them. When we have a minimally acceptable value for the specificity of both tests, our focus of analysis is on comparison of sensitivities of two tests at this minimal specificity.

Greenhouse and Mantel (1950) and Linnet (1987) proposed nonparametric procedures for the comparison of two sensitivities at a fixed level of specificity. Wieand et al (1989) studied asymptotic behaviors of these nonparametric procedures and generalized them to a comparison of two weighted average of sensitivities.

Consider two diagnostic tests T_1 and T_2 that yield continuous measurements. Assume that both tests are performed on the same m controls (non-diseased) and n cases (diseased). Let (X_{1j}, X_{2j}), $j = 1, 2, \cdots, m$, be i.i.d. bivariate outcomes from the population with a joint distribution $F(x_1, x_2)$ that represents the non-diseased group, (Y_{1k}, Y_{2k}), $k = 1, 2, \cdots, n$, be i.i.d. bivariate outcomes from population with a joint distribution $G(y_1, y_2)$ that represents the diseased group. Denote the marginal distribution functions of X_i and Y_i by $F_i(x_i)$ and $G_i(y_i)$, respectively, $i = 1, 2$. For a given cut-off point c, the sensitivity and specificity of the test T_i, $i = 1, 2$, are defined by

$$S_i(c) = P(Y_i \geq c) = 1 - G_i(c), \quad Sp_i(c) = P(X_i \leq c) = F_i(c),$$

respectively. Therefore, for a fixed value of specificity at p, the sensitivity of test T_i is $S_i(p) = 1 - G_i(F_i^{-1}(p))$, where $F_i^{-1}(p) = inf\{t : F_i(t) \geq p\}$, $i = 1, 2$. The parameter of interest is the difference between two sensitivities at the same fixed value of specificity p_0,

$$\Delta = S_1(p_0) - S_2(p_0).$$

Let $\widehat{G}_i$ be the empirical distribution of G_i, based on the sample $X_{i1}, \ldots, X_{im}$, and let $\widehat{F}_i^{-1}(p)$ be the empirical estimate for the p-th quantile of F_i, $i = 1, 2$, based on the sample $Y_{i1}, \ldots, Y_{in}$. The non-parametric estimator for Δ proposed by Linnet (1987) and Wieand et al (1989) is given as follows:

$$\hat{\Delta} = \widehat{S}_1(p_0) - \widehat{S}_2(p_0),$$

where $\widehat{S}_i(p_0) = 1 - \widehat{G}_i(\widehat{F}_i^{-1}(p_0))$. Let $N = m + n$. Wieand et al (1989) have shown that

$$N^{1/2}\left(\hat{\Delta} - \Delta\right) \sim N\left(0, \sigma^2\right) \qquad (1)$$

where

$$\begin{aligned}
\sigma^2 &= \sigma_1^2 + \sigma_2^2 - 2\sigma_{12}, \\
\sigma_i^2 &= (1-\lambda)^{-1} S_i(p_0)(1 - S_i(p_0)) + \lambda^{-1}(1-p_0)p_0 \frac{g_i^2(F_i^{-1}(p_0))}{f_i^2(F_i^{-1}(p_0))} \qquad (i = 1, 2), \\
\sigma_{12} &= (1-\lambda)^{-1}\left[G\left(F_1^{-1}(p_0), F_2^{-1}(p_0)\right) - G_1(F_1^{-1}(p_0))G_2(F_2^{-1}(p_0))\right], \\
&\quad + \lambda^{-1}\left[F\left(F_1^{-1}(p_0), F_2^{-1}(p_0)\right) - p_0^2\right] \frac{g_1(F_1^{-1}(p_0))}{f_1(F_1^{-1}(p_0))} \frac{g_2(F_2^{-1}(p_0))}{f_2(F_2^{-1}(p_0))}, \\
\lambda &= m/(m+n).
\end{aligned}$$

where f_i and g_i are the density functions of F_i and G_i respectively.

We can use the normal approximation (1) to construct a confidence interval for the difference between two sensitivities at the same fixed level of specificity if a good estimate for σ^2 is available. However, the estimation of σ^2 requires the estimation of density functions f_i and g_i, the estimation of bivariate distribution functions $F(x_1, x_2)$ and $G(y_1, y_2)$, and the estimation of quantiles $F_i^{-1}(p)$. Therefore, the performance of the normal approximation based confidence interval (hereafter called WGJ interval) is very sensitive to the choice of the smoothing parameters in density and distribution estimations. Selection of a satisfactory smoothing parameters in this context is problematic.

In this paper, we propose three new intervals for the difference between sensitivities of two diagnostic tests at a fixed value of specificity. The major advantage of the new intervals over the normal approximation based interval is that we don't need to use density and distribution estimation. In addition, the new intervals are computationally simple and easy to implement in practice. Our simulation studies also indicate that the new intervals perform better than the existing normal approximation based interval in terms of coverage accuracy and interval length.

The rest of this paper is organized as follows. In Section 2 we propose three new confidence intervals. In Section 3 we conduct simulation studies to assess the finite-sample performance of the new intervals. In Section 4 we illustrate the application of the proposed methods in a real example.

2. New confidence intervals

In this section, we construct $(1-\alpha)100\%$ confidence intervals for the difference Δ of two sensitivities at the same fixed value of specificity p_0. Note that

$$\Delta \;=\; S_1(p_0) - S_2(p_0) = P\left(Y_{1k} \geq F_1^{-1}(p_0)\right) - P\left(Y_{2k} \geq F_2^{-1}(p_0)\right).$$

If F_i were known, an obvious estimator of Δ would be the difference between the observed sensitivities at p_0-th quantiles $F_1^{-1}(p_0)$ and $F_2^{-1}(p_0)$, which would be defined as

$$\Delta_0 = \frac{1}{n}\sum_{k=1}^{n} I_{[Y_{1k} \geq F_1^{-1}(p_0)]} - \frac{1}{n}\sum_{k=1}^{n} I_{[Y_{2k} \geq F_2^{-1}(p_0)]}, \tag{2}$$

where I_A is the indicator function of A. We can also regard Δ_0 as the difference between two sample proportions of binomial distributions with proportions $S_i(p_0)$, $i = 1, 2$. Because F_i's are in fact unknown, replacing the $F_i^{-1}(p_0)$ by $\hat{F}_i^{-1}(p_0)$ in (2), we obtain an estimator for Δ. That is,

$$\hat{\Delta}_0 = \frac{1}{n}\sum_{k=1}^{n} I_{[Y_{1k} \geq \hat{F}_1^{-1}(p_0)]} - \frac{1}{n}\sum_{k=1}^{n} I_{[Y_{2k} \geq \hat{F}_2^{-1}(p_0)]}. \tag{3}$$

Since the indicator variables $I_{[Y_{i1} \geq \hat{F}_i^{-1}(p_0)]}$, $I_{[Y_{i2} \geq \hat{F}_i^{-1}(p_0)]}$, $\cdots$, $I_{[Y_{in} \geq \hat{F}_i^{-1}(p_0)]}$ are not independent, $\hat{\Delta}_0$ is no longer the difference between two simple binomial proportions. Therefore, the usual methods for construction of confidence interval for the difference between two binomial proportions, such as one proposed by Agresti and Caffo (2000), can not be directly applicable here. However, noticing the relationship between $\hat{\Delta}_0$ and a two-sample binomial problem, we can construct intervals for Δ based on a variation of $\hat{\Delta}_0$ by combining bootstrap method with the technique by Agresti and Caffo (2000). Depending on whether there is a correlation between the test results from two diagnostic tests, we propose the following different procedures for the confidence intervals of Δ.

2.1 A paired uncorrelated samples

When the test results from two diagnostic tests are conditionally uncorrelated within the diseased groups, $\hat{\Delta}_0$ can be considered as the difference between two independent sample proportions. Using the technique by Agresti and Caffo (2000), we proposed the following potentially better estimator for Δ_0 instead of $\hat{\Delta}_0$:

$$\hat{\Delta} = \hat{S}_1(p_0) - \hat{S}_2(p_0), \tag{4}$$

where

$$\hat{S}_i(p_0) = \frac{\sum_{k=1}^{n} I_{[Y_{ik} \geq \widehat{F}_i^{-1}(p_0)]} + z_{1-\alpha/2}^2/2}{n + z_{1-\alpha/2}^2}, \quad i = 1, 2, \tag{5}$$

and $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of standard normal distribution when the confidence level is $1-\alpha$. Since $z_{1-\alpha/2}^2 = 1.96^2$ is approximately equal to 4 when $\alpha = 0.05$, $\hat{S}_i(p_0)$ may be regarded as an adjusted estimate for binomial proportion $S_i(p_0)$ by adding two successes and two failures to Bernoulli observations. We use $\hat{\Delta}$ here rather than the standard $\hat{\Delta}_0$ as the estimate for Δ_0 because the simulation study by Agresti and Coull (1998) showed that the adjusted Wald intervals for $S_i(p_0)$ based on $\hat{S}_i(p_0)$ have good coverage accuracy even for small sample sizes. Although $\hat{\Delta}$ is the difference of two conditionally uncorrelated proportions, it is still difficult to find a good variance estimate for $\hat{\Delta}$ because of the dependence among the indicator variables

$I_{[Y_{ik} \geq \widehat{F}_i^{-1}(p)]}, k = 1, 2, \cdots, n$. Therefore, the most often used Wald interval cannot be directly applicable here. In this paper we propose to use a bootstrap method to estimate the variance of $\hat{\Delta}$. We summarize the procedure for computing the bootstrap variance in the following steps:

1. For each $i = 1, 2$, draw a resample of size n, Y_{ik}^* $(k = 1, ..., n)$ with replacement from the diseased patient sample Y_{ik} $(k = 1, ..., n)$, and a separate resample of size m, X_{ij}^* $(j = 1, ..., m)$ with replacement from the non-diseased patient sample X_{ij} $(j = 1, ..., m)$.

2. Calculate the bootstrap version of $\hat{S}_i(p_0)$ (i=1,2), and $\hat{\Delta}$,
$$\begin{aligned} \hat{S}_i^*(p_0) &= \frac{\sum_{k=1}^n I_{[Y_{ik}^* \geq \widetilde{F}_i^{-1}(p_0)]} + z_{1-\alpha/2}^2/2}{n + z_{1-\alpha/2}^2}, \\ \hat{\Delta}^* &= \hat{S}_1^*(p_0) - \hat{S}_2^*(p_0), \end{aligned}$$
where $\widetilde{F}_i^{-1}(p)$ is the p-th sample quantile based on the bootstrap resample X_{ij}^*'s.

3. Repeat the first two steps B times to obtain the set of bootstrap replications $\{\hat{S}_{ib}^*(p_0), \hat{\Delta}_b^* : b = 1, 2, \cdots, B\}$, $i = 1, 2$.

Then, the bootstrap variance estimator V^* for $\hat{\Delta}$ is defined as follows:

$$V^* = V_1^* + V_2^*,$$

where

$$V_i^* = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{S}_{ib}^*(p_0) - \bar{S}_i^*(p_0) \right)^2, \quad i = 1, 2,$$

and $\bar{S}_i^*(p_0) = (1/B) \sum_{b=1}^B \hat{S}_{ib}^*(p_0)$, $i = 1, 2$.

Here we want to point out that the above procedure can easily be extended to the case of two independent samples with different sample sizes. For simplicity, we only consider the paired conditionally uncorrelated samples in this paper.

2.2 A paired dependent samples

When two diagnostic tests are applied to the same patients, the test results from two diagnostic tests are most likely correlated. Because of the dependence of the paired samples, we propose to use the bootstrap procedure defined as before except that the constant $z_{1-\alpha/2}$ in (5) and in the second step be taken to be $(z_{1-\alpha/2})^{1/2}$. When $\alpha = 0.05$, $(z_{1-\alpha/2})^{1/2} \approx \sqrt{2}$, and

$$\hat{S}_i(p_0) = \frac{\sum_{k=1}^{n} I_{[Y_{ik} \geq \widehat{F}_i^{-1}(p_0)]} + 1}{n+2}, \quad i = 1, 2.$$

Therefore, $\hat{\Delta}$ may be regarded as an adjusted estimate for the difference between two binomial proportions by adding one success and one failure to Bernoulli observations. Our extensive simulation study indicated that the confidence intervals for Δ resulting from this modification have better coverage accuracy than that of adding two successes and two failures method proposed for two independent (or paired uncorrelated) samples. The bootstrap variance estimator V^* for $\hat{\Delta}$ is then defined as follows:

$$V^* = V_1^* + V_2^* - 2V_{12}^*$$

where V_i^* (i=1, 2) are defined as before, and

$$V_{12}^* = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{S}_{1b}^*(p_0) - \bar{S}_1^*(p_0) \right) \left(\hat{S}_{2b}^*(p_0) - \bar{S}_2^*(p_0) \right).$$

2.3 New bootstrap intervals for Δ

Now we can propose new intervals for Δ. The first two $(1-\alpha)100\%$ confidence intervals for Δ are bootstrap intervals based on the bootstrap variance estimator V^*. They are defined as follows:

(i) The first one, called BTI interval, is

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

where $\hat{\Delta}$ is defined by (4).

(ii) The second one, called BTII interval, is

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

where $\bar{\Delta}^* = (1/B)\sum_{b=1}^{B}\hat{\Delta}_b^*$.

The above two intervals require variance estimation of $\hat{\Delta}$. The third interval for Δ is a BCa-type bootstrap interval, which does not require the direct variance estimation. Efron and Tibshirani suggest the use of the BCa intervals as they provide more stable results and bettter coverage probabilities with fewer bootstrap resamples that do the percentile intervals. The following is a modified BCa interval for Δ in the setting of comparing two sensitivities at the same fixed level of specificity :

$$\left(\hat{\Delta}^*_{(B\hat{\alpha}/2)}, \hat{\Delta}^*_{(B(1-\hat{\alpha}/2))}\right),$$

where

$$\begin{aligned}
\hat{\alpha} &= \Phi\left(w + \frac{w + z_\alpha}{1 - a(w + z_\alpha)}\right), \\
w &= \Phi^{-1}\left(\frac{1}{B}\sum_{b=1}^{B} I_{[\hat{\Delta}_b^* \le \hat{\Delta}]}\right), \\
a &= \frac{1}{6}\frac{\sum_{k=1}^{n} l_k^3}{\left(\sum_{k=1}^{n} l_k^2\right)^{3/2}}, \\
l_k &= \left(I_{[Y_{1k} \ge \hat{F}_1^{-1}(p_0)]} - I_{[Y_{2k} \ge \hat{F}_2^{-1}(p_0)]}\right) - \left(\hat{S}_1(p_0) - \hat{S}_2(p_0)\right),
\end{aligned}$$

Φ is the standard normal distribution, and $\hat{\Delta}^*_{(b)}$ is the b-th ordered value among $\{\hat{\Delta}_b^*, b = 1, 2, \cdots, B\}$.

3. Simulation Studies for the Confidence Intervals

In this section, we conduct two simulation studies to evaluate coverage accuracy and interval length of the newly proposed intervals for Δ when the specificity p is taken to be 80% or 90% in finite-sample sizes. In both studies, We generated 2,000 random samples of size n from $G(y_1, y_2)$ for test responses of diseased patients, and another independent random samples of size m from $F(x_1, x_2)$ for test responses of non-diseased patients. The normal approximation based interval (WGJ interval), proposed by Wieand et al (1989), is also included in these studies for comparison.

In the first study, $G(y_1, y_2)$ is chosen to be a bivariate normal distribution having means $E(Y_1) = \mu_1$, $E(Y_2) = \mu_2$, and with a common standard deviation 2 and correlation ρ; $F(x_1, x_2)$ is chosen to be a bivariate normal distribution with means $E(X_1) = 0$, $E(X_2) = 0$, and with a common standard deviation 1 and correlation ρ. Thus $S_i(p) = 1 - \Phi\{\frac{\Phi^{-1}(p)-\mu_i}{2}\}$, for $i = 1, 2$. For $\Delta = 0$, we choose $\mu_1 = \mu_2$ such that the sensitivity $S_i(p)$ of the test T_i $(i = 1, 2)$ varies over the points 0.95, 0.90, 0.80, 0.70, 0.60, 0.50, 0.40, 0.30, 0.20, 0.10, respectively.

In the second study, the distributions $G(y_1, y_2)$, $F(x_1, x_2)$ are chosen to be different bivariate exponential distributions that have exponential distributions as their marginal distributions. Depending on the possible correlation between the test results from two diagnostic tests, we use two different procedures to generate the random samples of test responses. First we choose the correlation as zero ($\rho = 0$), and then we generate two independent samples, X_{11}, X_{12}, $\cdots$, X_{1m} and X_{21}, X_{22}, $\cdots$, X_{2m}, from standard exponential distribution; and two independent samples, Y_{11}, Y_{12}, $\cdots$, Y_{1n}, and Y_{21}, Y_{22}, $\cdots$, Y_{2n}, from exponential distributions with rates λ_1, λ_2, respectively. Therefore, $S_i(p) = \exp[\lambda_i \log(1 - p)]$, for $i = 1, 2$. Second, we choose a positive correlation ($\rho > 0$), we first generate random samples, $U_{i1}, U_{i2}, \cdots, U_{im}$, from a exponential distribution with rate 0.5, for $i = 1, 2, 3$; and random samples, $V_{i1}, V_{i2}, \cdots, V_{in}$, from a exponential distribution with rate l_i, for $i = 1, 2$; and a random sample, $V_{31}, V_{32}, \cdots, V_{3n}$, from a exponential distributions with rate 0.01. Then, the simulated test responses for non-diseased patients are $X_{ij} = \min(U_{ij}, U_{3j})$, $i = 1, 2$, $j = 1, 2, \cdots, m$, which are random samples from two standard exponential distributions with correlation ρ; and those for diseased patients are $Y_{ik} = \min(V_{ik}, V_{3k})$, $i = 1, 2$, $k = 1, 2, \cdots, n$, which are random samples from two exponential distributions with correlation ρ and rates $l_1 + 0.01$, $l_2 + 0.01$, respectively. Under this setting, $S_i(p)$ is $\exp[(l_i + 0.01) \log(1 - p)]$, for $i = 1, 2$. Similar to the first simulation study, we choose

λ_i, l_i ($i = 1, 2$) such that $\Delta = 0$ as the sensitivity $S_1(p)$ varies over the points 0.50, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95, respectively.

The computation of WGJ interval is complicated by estimating the unknown underlying density functions f_i and g_i, and bivariate distribution functions $F(x_1, x_2)$ and $G(y_1, y_2)$. In the simulation studies, we use the same method as that by Wieand et al (1989) to estimate the asymptotic variance σ^2. That is, the consistent estimate of σ^2 is obtained by substituting kernel density estimators for f_i and g_i, the empirical distribution functions and sample quantiles for corresponding population distribution functions and quantiles.

Tables 1-4 goes here

We summary the average coverage probabilities and average interval lengths over selected values of sensitivities $S_1(p)$'s for the WGJ interval and the three newly proposed intervals (BCa, BTI and BTII) in Tables 1-2 when the underlying distributions are bivariate normal distributions and in Tables 3-4 when the underlying distributions are the bivariate exponential distributions. Since the averaging coverage probabilities do not provide information on effects of particular values of $S_1(p)$ and $S_2(p)$ on coverage probability, we also plot the coverage probabilities of Δ when $S_1(p)$ varies over the points chosen above. Figures 1-4 display the coverage coverage probabilities of Δ for the four intervals as functions of sensitivity $S_1(p)$ when $(m, n) = (20, 20), (50, 50)$, and $(30, 50)$, respectively.

Figures 1-4 goes here

From Tables 1-4 and Figures 1-4, we make the following observations.

(1) When the correlation $\rho = 0$, the newly proposed BTI and BTII intervals have uniformly better coverage accuracy than the WGJ interval across all the sensitivity levels at the specificity levels considered here. The coverage probabilities of WGL interval are below the nominal level

for most of the sensitivity levels. The BTI, BTII intervals outperform the WGJ interval, particularly for small to moderate sample size ($n, m \leq 20$).

(2) When the correlation ρ is positive, BCa performs better than the other methods in terms of coverage probability for most of the sensitivities levels.

(3) Although WGJ interval occasionally has good coverage probability, it generally has longer average interval length than the newly proposed intervals, and sometimes its average interval length is twice long as the newly proposed intervals. Moreover, the computation of WGJ interval is the most complicated.

In summary, our simulation studies suggest that the newly proposed BTI and BTII intervals perform better than the existing WGJ interval for independent samples, and BCa performs better than the WGJ interval for paired dependent samples. In addition, the new intervals are computationally much simpler than the WGJ interval. Among the three new intervals, we recommend the BCa interval for paired dependent samples and the BTI and BTII intervals for independent samples.

4 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) studied the accuracy of clinical evaluations with or without 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. 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 (Venkatraman, 1996). Hence, we have two tests for detecting MM; the first test is the clinical assessment without the aid of dermatoscopy, and the second test is the clinical assessment with the aid of dermatoscopy.

To be sure that the two tests have a high change of ruling out patients without MM, dermatologists want the specificity of the tests to be at least 90% for detecting patients without MM and want to know what the relative corresponding sensitivities of the two tests are in detecting patients with MM. Therefore, It is an interest to construct a confidence interval for the difference of sensitivities of the two tests when their specificities are fixed at 90% (or 95%).

Ninety-five percent confidence intervals for the difference in sensitivities between the two clinical assessments without and with the aid of dermatoscopy at the two fixed levels of specificities (90% and 95%) are shown in Table 5. All confidence intervals are containing zero. Therefore, we conclude that there is no significant advantage to adopt the clinical assessment with the aid of dermatoscopy in detecting MM.

Table 5 goes here

5. Discussion

There are different ways for comparing the accuracy of two continuous-scale tests, depending on whether we can specify a commonly minimal acceptable value for the specificity of both tests. If we can, we would be interested in comparing sensitivities of the two tests at the same fixed level of their specificities. If we cannot, specify a minimally acceptable value for the specificity of test, our interest would be comparison of the whole or partial ROC curve.

In this paper, we have focused our attention on the situation where we can specify a commonly minimal acceptable value for the specificity of both tests. We have proposed BTI, BTII and BCa confidence intervals for sensitivity at a fixed level of specificity, and have shown via simulation that the newly proposed methods outperform the existing method in terms of the coverage accuracy and interval length. Among the three new intervals, BTI and BTII are based on the techniques for the confidence intervals between *two independent binomial proportions* proposed by Agresti and Caffo (2000), it is expected that BTI and BTII perform better than the other methods for independent samples. Agresti and Caffo (2000) didn't discuss the confidence intervals for *paired dependent binomial proportions*. We applied the similar technique to paired dependent samples with adjustment for variance estimate (see section 2.2), and proposed a BCa interval for the difference between two sensitivities for paired dependent samples. Our simulation study indicated that this method works and BCa method performs better than BTI and BTII. One possible reason is that BCa method better captures the dependence between the paired samples and produced a better adjusted confidence level. The theoretical comparison of these methods are difficult. Edgeworth expansion or saddlepoint approximation for the coverage probabilities may shed some light on this problem (see Zhou, Tsao and Qin, 2004; Zhou and Qin, 2005).

References

Agresti A, and Caffo BA. Simple and effective confidence intervals for proportions and difference of proportions result from adding two successes and two failures. *The American Statistician* 2000; **54**:280-288.

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

Greenhouse SW, and Mantel N. The evaluation of diagnostic tests. *Biometrics* 1950; **6**:399-412.

Friedman JH. Multivariate Adaptive Regression Splines. *Annals of Statistics* 1991; **19**:1067.

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

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

Venkatraman ES and Begg CB. A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment. *Biometrika* 1996; **83**:835-848.

Wieand S, Gail MH, James BR, and James KR. A family of non-parametric statistics for comparing diagnostic markers with paired and unpaired data. *Biometrika* 1989; **76**:585-592.

Zhou XH, Tsao M and Qin GS. New intervals for the difference between two independent binomial proportions. *J Statist Plan Infer* 2004; **123**:97-115;

Zhou XH, and Qin GS. A new confidence interval for the difference between two binomial proportions of paired data. *J Statist Plan Infer* 2005; **128**:527-542.

Table 1. Level 95% confidence interval for Δ. Bivariate normal distribution with $\rho = 0$

Specificity	sample size	Method	Ave. coverage probability	Average length
0.90	m=10, n=10	WGJ	0.9072	1.0629
		BCa	0.9181	0.4634
		BTI	0.9528	0.5664
		BTII	0.9662	0.5664
	m=20, n=20	WGJ	0.9225	0.7664
		BCa	0.9331	0.4300
		BTI	0.9584	0.4966
		BTII	0.9701	0.4966
	m=50, n=50	WGJ	0.9377	0.4752
		BCa	0.9418	0.3222
		BTI	0.9588	0.3535
		BTII	0.9682	0.3535
	m=30, n=50	WGJ	0.9212	0.5348
		BCa	0.9418	0.3511
		BTI	0.9603	0.3842
		BTII	0.9687	0.3842
0.80	m=10, n=10	WGJ	0.9082	1.0373
		BCa	0.9230	0.4692
		BTI	0.9582	0.5716
		BTII	0.9675	0.5716
	m=20, n=20	WGJ	0.9217	0.7454
		BCa	0.9312	0.4175
		BTI	0.9557	0.4797
		BTII	0.9645	0.4797
	m=50, n=50	WGJ	0.9336	0.4651
		BCa	0.9388	0.3101
		BTI	0.9555	0.3382
		BTII	0.9645	0.3382
	m=30, n=50	WGJ	0.9234	0.5153
		BCa	0.9360	0.3308
		BTI	0.9561	0.3614
		BTII	0.9667	0.3614

Table 2. Level 95% confidence interval for Δ. Bivariate normal distribution with $\rho = 0.5$

Specificity	sample size	Method	Ave. coverage probability	Average length
0.90	m=10, n=10	WGJ	0.9015	0.9369
		BCa	0.9570	0.5286
		BTI	0.9279	0.4960
		BTII	0.9520	0.4960
	m=20, n=20	WGJ	0.9315	0.7051
		BCa	0.9646	0.4649
		BTI	0.9212	0.3917
		BTII	0.9470	0.3917
	m=50, n=50	WGJ	0.9333	0.4303
		BCa	0.9688	0.3347
		BTI	0.9241	0.2647
		BTII	0.9332	0.2647
	m=30, n=50	WGJ	0.9189	0.4862
		BCa	0.9708	0.3598
		BTI	0.9182	0.2864
		BTII	0.9414	0.2864
0.80	m=10, n=10	WGJ	0.9075	0.8988
		BCa	0.9595	0.5280
		BTI	0.9279	0.4889
		BTII	0.9562	0.4889
	m=20, n=20	WGJ	0.9352	0.6720
		BCa	0.9623	0.4493
		BTI	0.9193	0.3760
		BTII	0.9448	0.3760
	m=50, n=50	WGJ	0.9318	0.4100
		BCa	0.9687	0.3225
		BTI	0.9155	0.2525
		BTII	0.9336	0.2525
	m=30, n=50	WGJ	0.9232	0.4693
		BCa	0.9696	0.3435
		BTI	0.9142	0.2700
		BTII	0.9378	0.2700

Table 3. Level 95% confidence interval for Δ. Bivariate exponential distribution with $\rho = 0$

Specificity	sample size	Method	Ave. coverage probability	Average length
0.90	m=10, n=10	WGJ	0.9040	1.1603
		BCa	0.9185	0.4479
		BTI	0.9652	0.6083
		BTII	0.9769	0.6083
	m=20, n=20	WGJ	0.9132	0.8423
		BCa	0.9277	0.4430
		BTI	0.9591	0.5303
		BTII	0.9712	0.5303
	m=50, n=50	WGJ	0.9329	0.5069
		BCa	0.9358	0.3431
		BTI	0.9580	0.3810
		BTII	0.9688	0.3810
	m=30, n=50	WGJ	0.9151	0.5721
		BCa	0.9379	0.3721
		BTI	0.9581	0.4174
		BTII	0.9685	0.4174
0.80	m=10, n=10	WGJ	0.8906	1.1659
		BCa	0.9249	0.4769
		BTI	0.9652	0.6291
		BTII	0.9753	0.6291
	m=20, n=20	WGJ	0.9124	0.8516
		BCa	0.9311	0.4484
		BTI	0.9644	0.5311
		BTII	0.9730	0.5311
	m=50, n=50	WGJ	0.9286	0.5191
		BCa	0.9370	0.3393
		BTI	0.9576	0.3766
		BTII	0.9668	0.3766
	m=30, n=50	WGJ	0.9095	0.5954
		BCa	0.9325	0.3695
		BTI	0.9555	0.4121
		BTII	0.9671	0.4121

Table 4. Level 95% confidence interval for Δ. Bivariate exponential distribution with $\rho > 0$

Specificity	sample size	Method	Ave. coverage probability	Average length
0.90	m=10, n=10	WGJ	0.9198	0.6707
		BCa	0.9317	0.5253
		BTI	0.8915	0.5020
		BTII	0.9277	0.5020
	m=20, n=20	WGJ	0.9285	0.5006
		BCa	0.9494	0.4871
		BTI	0.8924	0.4076
		BTII	0.9174	0.4076
	m=50, n=50	WGJ	0.9387	0.4882
		BCa	0.9592	0.3570
		BTI	0.8895	0.2794
		BTII	0.9075	0.2794
	m=30, n=50	WGJ	0.9301	0.5663
		BCa	0.9577	0.3848
		BTI	0.8933	0.3047
		BTII	0.9161	0.3047
0.80	m=10, n=10	WGJ	0.9014	1.1545
		BCa	0.9363	0.5596
		BTI	0.8873	0.5173
		BTII	0.9223	0.5173
	m=20, n=20	WGJ	0.9236	0.8271
		BCa	0.9497	0.4928
		BTI	0.8834	0.4081
		BTII	0.9103	0.4081
	m=50, n=50	WGJ	0.9323	0.5054
		BCa	0.9551	0.3532
		BTI	0.8858	0.2767
		BTII	0.9043	0.2767
	m=30, n=50	WGJ	0.9209	0.5717
		BCa	0.9572	0.3841
		BTI	0.8884	0.3011
		BTII	0.9105	0.3011

Table 5. 95% Confidence interval for the difference of sensitivities between the two clinical assessments without and with the aid of dermatoscopy

Specificity	WGJ	BTI	BTII	BCa
0.90	(-0.538, 0.729)	(-0.220, 0.394)	(-0.302, 0.312)	(-0.261, 0.261)
0.95	(-1.000, 1.000)	(-0.346, 0.346)	(-0.336, 0.357)	(-0.609, 0.479)

Figure 1: Coverage probability of 95% confidence interval for Δ. Bivariate normal distribution with $\rho = 0$

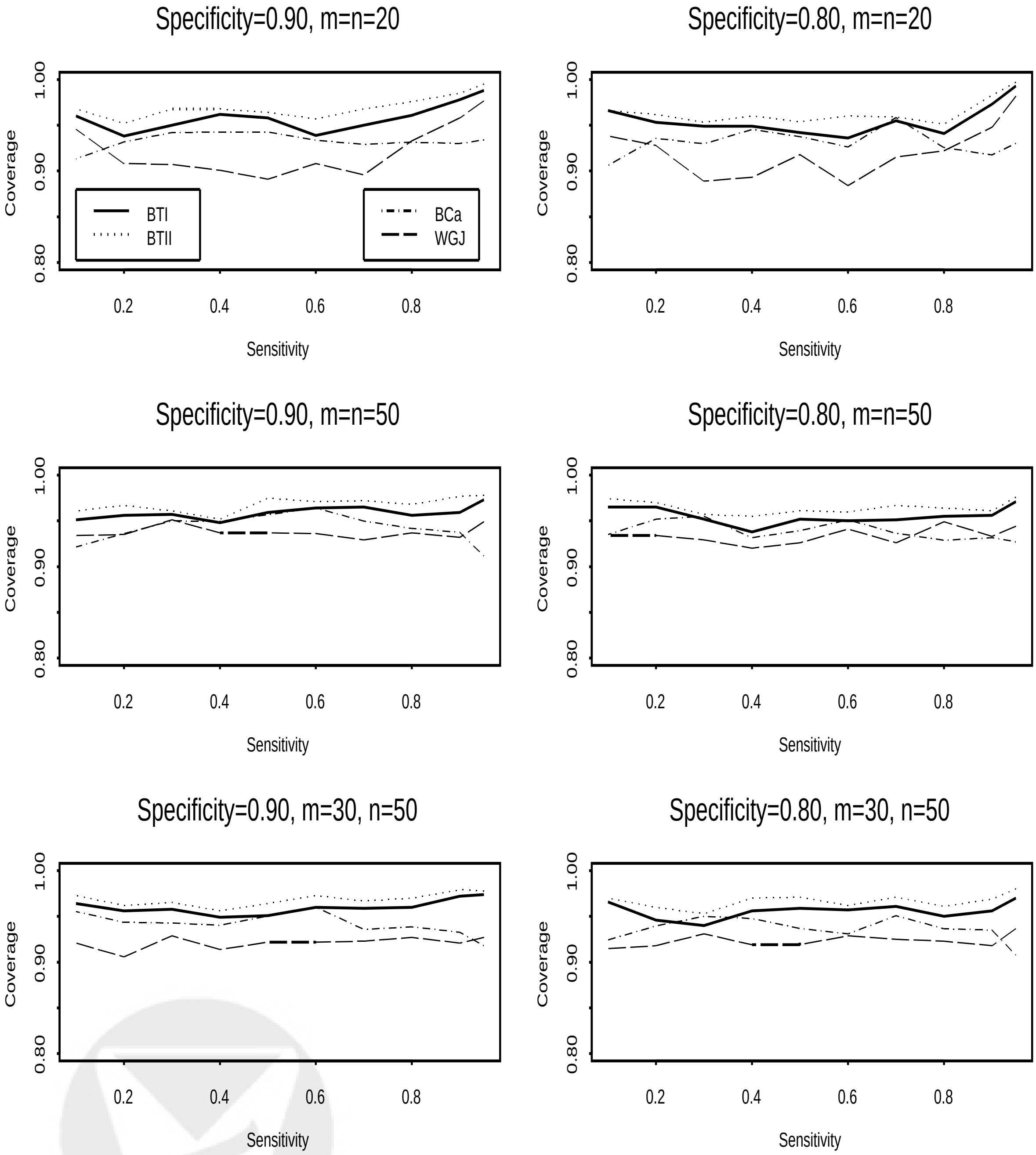


Figure 2: Coverage probability of 95% confidence interval for Δ. Bivariate normal distribution with $\rho = 0.5$

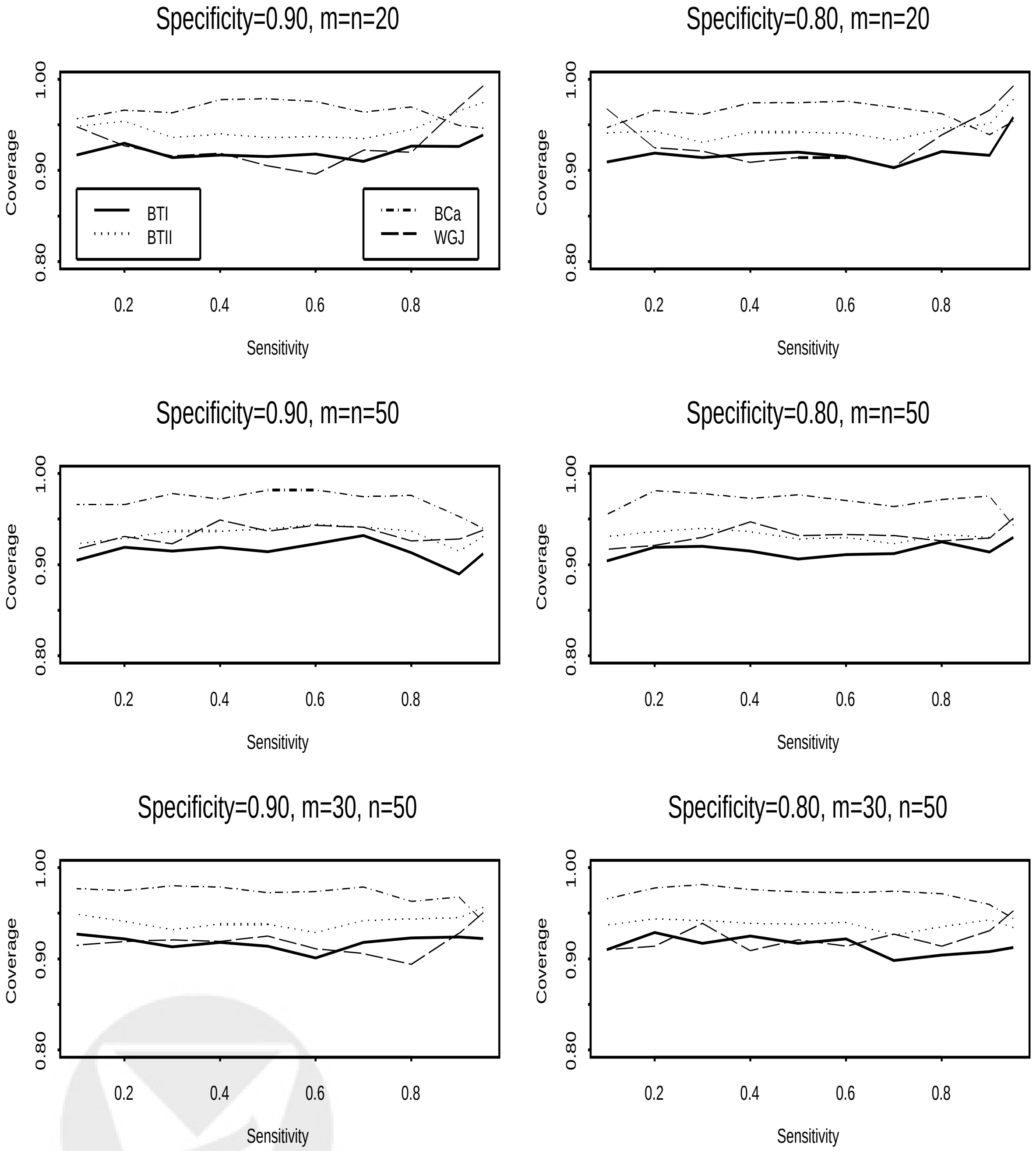


Figure 3: Coverage probability of 95% confidence interval for Δ. Bivariate exponential distribution with $\rho = 0$

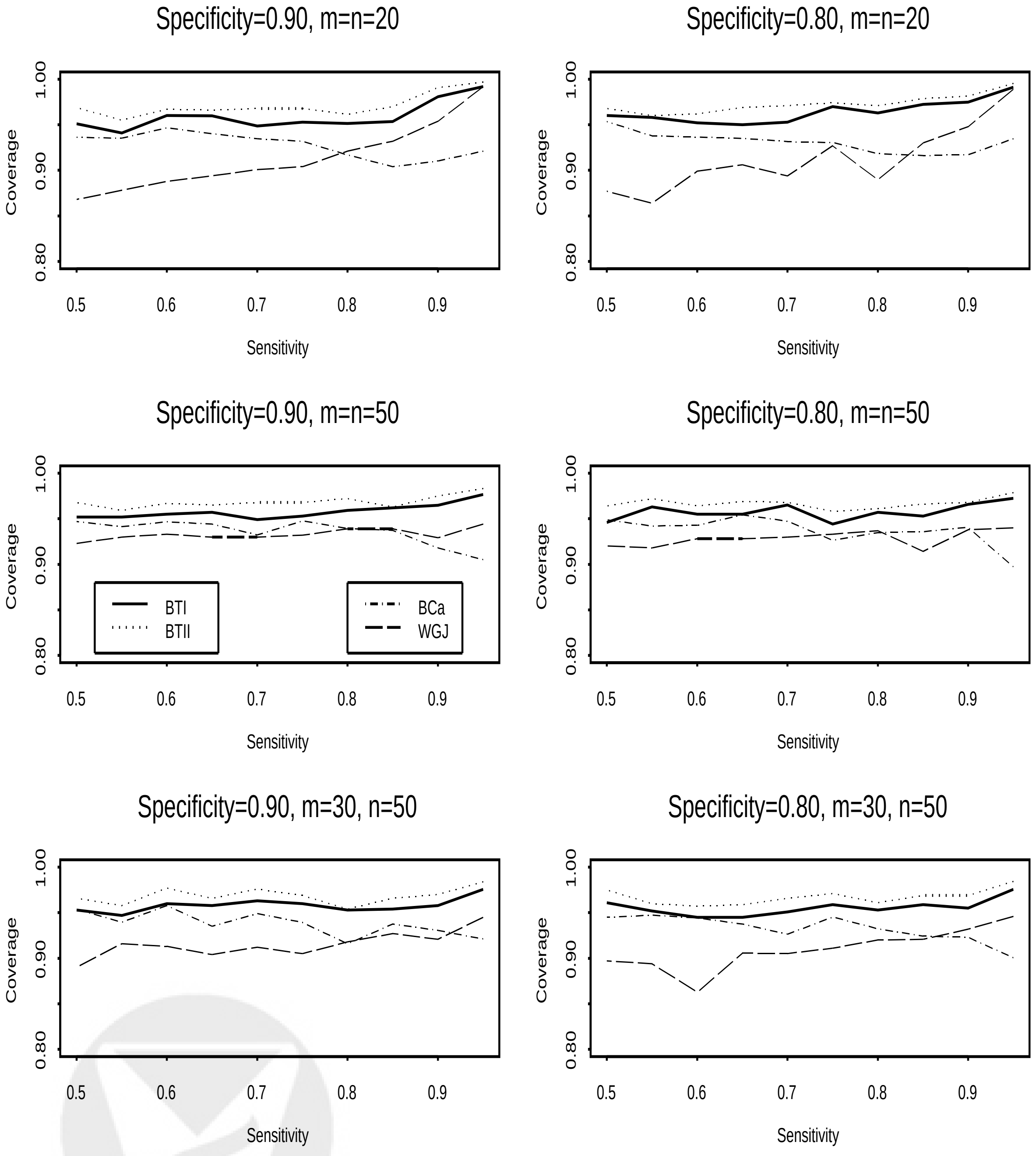
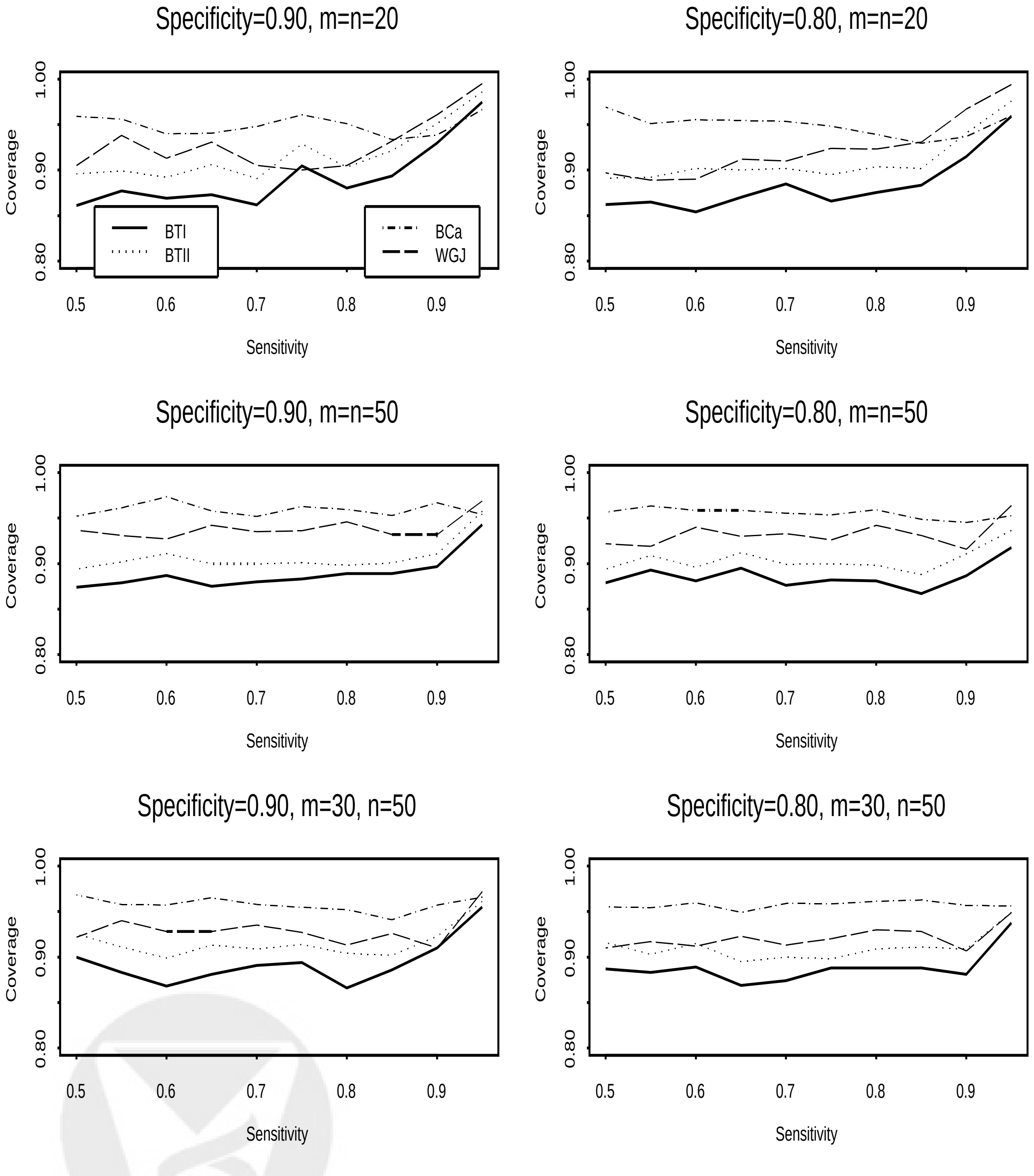


Figure 4: Coverage probability of 95% confidence interval for Δ. Bivariate exponential distribution with $\rho > 0$


Specificity=0.90, m=n=20
Specificity=0.80, m=n=20
Specificity=0.90, m=n=50
Specificity=0.80, m=n=50
Specificity=0.90, m=30, n=50
Specificity=0.80, m=30, n=50
Coverage
Sensitivity
BTI
BTII
BCa
WGJ