Comparing Two Correlated ROC Curves: A Generalized Pivotal Quantity Approach

Yonggang Zhao, PhD, MBA[1], Qianqiu Li, PhD[2]
[1]*Skyview Research, Philadelphia, PA*
[2]*Johnson & Johnson, Spring House, PA*

Abstract
Receiver operating characteristic (ROC) analysis is commonly used to evaluate the overall accuracy of continuous diagnostic tests. In this research, we propose a generalized pivotal quantity (GPQ) approach to compare the areas under two correlated ROC curves. The GPQ-based empirical powers of the equivalence tests for the area difference are simulated and compared with those by maximum likelihood and nonparametric methods.

Key Words: receiver operating characteristic (ROC), generalized pivotal quantity (GPQ), maximum likelihood, nonparametric, diagnostic tests, accuracy

1. Introduction

In medical practice, receiver operating characteristic (ROC) is widely used in evaluation of a continuous diagnostic test that distinguishes between diseased and normal cases. When two or more diagnostic tests are assessed, one primary interest is to compare the accuracy of the tests. If the tests are performed on the same study subjects, accounting for correlation between ROC curves is necessary. In this article, we compare the accuracy of two continuous diagnostic tests via equivalence testing for the difference in the areas of the correlated ROC curves.

In literature, several methods have been proposed in evaluation of correlated ROC curves. For example, DeLong et al (1998) compared the areas on the basis of correlated U statistics. Hanley and McNeil (1983) estimated the correlation of the two areas using Pearson correlation coefficients. Venkatraman and Begg (1996) applied permutation test for equality of paired ROC curves. Metz et al (1984) tested the equivalence of correlated ROC curves using a likelihood ratio test based on discretized continuous measurements. More recently, Li (2007) adopted generalized pivotal quantities for comparing the AUCs of ROC curves. Gallas and Pesce (2009) compared partially correlated ROC curves by taking into account the practical possibility that some subjects may be evaluated only by one but not both of the diagnostic tests. Wan and Zhang (2008), and Zhang and Zhang (2014) introduced semi-parametric methods for comparisons of correlated ROC curves and stated that their methods are more efficient than both of parametric and non-parametric counterparts. Bantis and Feng (2016) compared the AUCs under correlated ROCs given fixed specificity or sensitivity level.

In the past, most of the efforts were focused on estimation and evaluation of equality of correlated ROC curves. Published papers on equivalence testing for the areas under correlated ROC curves are sparse. Zhou et al (2002) used the Delong's non-parametric method for the equivalence test. Another attempt on this aspect was made by Liu et al (2006). The authors converted the area difference to the standardized difference for

assessing equivalence of paired areas under ROC curves, and compared the standardized difference method and its bootstrap version to Delong's method and bootstrap procedure of Delong's method in terms of the equivalence test powers. Considering lack of literature on comparing different methods for the equivalence testing on the basis of the difference in areas of correlated ROC curves, and the fact that the equivalence test results may rely heavily on the statistical methods and sample sizes, we tackle this problem by examining the performance of three methods applied to the equivalence testing: non-parametric method (Delong et al (1988)), maximum likelihood method, and generalized pivotal quantity approach. The equivalence test powers were empirically calculated based on simulation data from bivariate normal distributions under different scenarios. The bias in terms of estimation of the area difference between correlated ROC curves was also obtained for comparisons. The methods are detailed in next section.

2. Definitions and Methods

2.1 Notations and Assumptions

It is assumed that two diagnostic tests are used for each subject in both the diseased population and the non-diseased population. Define $X = (X_1, X_2, \cdots, X_n)$ and $Y = (Y_1, Y_2, \cdots, Y_m)$ as the vectors of bivariate random variables for the diseased and non-diseased subjects, respectively. Let $X_1, X_2, \cdots, X_n$ denote independent and identically distributed (iid) diagnostic test results from the diseased subjects, with mean vector μ and variance-covariance matrix Σ. For $i = 1, \cdots, n$,

$$X_i = \begin{pmatrix} X_{i1} \\ X_{i2} \end{pmatrix} \sim N(\mu, \Sigma) \quad where\ \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix},\ \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}, \qquad \rho_X = \frac{\sigma_{12}}{\sigma_1 \sigma_2}$$

The population correlation of the bivariate normal distribution is denoted as ρ_X.

For any i, the two measurements $X_{i1}\ and\ X_{i2}$ from the i-th subject are assumed to be correlated with a non-zero coefficient ρ_X,

$$\rho_X = \frac{\sigma_{12}}{\sigma_1 \sigma_2}$$

Similarly, let $Y_1, Y_2, \cdots, Y_m$ be independent and identically distributed test results from the non-diseased subjects, with mean η and variance Ψ. For $j = 1, \cdots, m$,

$$Y_j = \begin{pmatrix} Y_{j1} \\ Y_{j2} \end{pmatrix} \sim N(\eta, \Psi),\ \ where\ \eta = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix},\ \Psi = \begin{pmatrix} \psi_1^2 & \psi_{12} \\ \psi_{12} & \psi_2^2 \end{pmatrix},\ \ \rho_Y = \frac{\psi_{12}}{\psi_1 \psi_2}$$

Also, a non-zero and common correlation coefficient ρ_Y is assumed for two diagnostic test results from all non-diseased subjects.

For any i and j, X_{i1} and Y_{j1} represent the results from the first diagnostic test, and X_{i2} and Y_{j2} refer to the results from the second diagnostic test.

Under the above assumptions, for each of the two diagnostic tests, a receiver operating characteristic curve can be used to measure the overall accuracy. The ROC curve is a plot

of the sensitivity (or true positive rate) against its false positive rate (1-specificity), constructed by changing the decision thresholds that define positive and negative test results. Specifically, at each of the pre-specified threshold values, paired values of sensitivity and specificity can be computed and then used to generate one corresponding point on the ROC curve. Different diagnostic tests can be visually compared if the ROC curves do not intersect each other. Nevertheless, formal comparisons are commonly carried out using the total areas under the ROC curves (AUC). In next section, the applied methods for estimation and comparisons of the areas under correlated ROC curves are described in details.

2.2 Estimation and Comparisons of Areas under ROC Curves

Without loss of generality, we assume that a larger test result indicates greater likelihood of the disease. Then the AUC for k-th diagnostic test (k=1,2) can be expressed as

$$\theta_k = P_r\left(X_{ik} > Y_{jk}\right) \quad for\ any\ i\ and\ j$$

Under the above bivariate normal distributions, the AUC for k-th diagnostic test (k=1,2) can be rewritten as a function of μ_k, η_k, σ_k^2 and ψ_k^2 as follows,

$$\theta_k = \Phi\left(\frac{\mu_k - \eta_k}{\sqrt{\sigma_k^2 + \psi_k^2}}\right)$$

The second expression is used in application of the maximum likelihood and generalized pivotal quantity methods for estimation and comparisons of the correlated AUCs. On the other hand, using the first expression, the non-parametric method (Delong et al (1998)) estimates the AUCs and their correlation regardless of normality of the test data.

2.2.1 Non-Parametric Method

The nonparametric method by Delong et al (1998) is based on the correlated Mann-Whitney U statistics, under the assumption that the test results from the same subject are correlated. Specifically, the AUC for the k-th diagnostic test is given by

$$\theta_k = \frac{1}{nm}\sum_{j=1}^{m}\sum_{i=1}^{n} I\left(X_{ik}, Y_{jk}\right)$$

Where $I(a,b)$ is the indicator function defined as follows.

$$I(a,b) = \begin{cases} 1 & if\ a > b \\ 0.5 & if\ a = b \\ 0 & if\ a < b \end{cases}$$

The variance of the difference in areas must take into account the correlation. That is,

$$VAR\left(\hat{\theta}_1 - \hat{\theta}_2\right) = VAR\left(\hat{\theta}_1\right) + VAR\left(\hat{\theta}_2\right) - 2COV\left(\hat{\theta}_1, \hat{\theta}_2\right)$$

The covariance estimate is the weighted average of the covariance of the non-diseased and diseased estimates

$$COV(\hat{\theta}_1, \hat{\theta}_2) = \frac{1}{n} s_1^{12} + \frac{1}{m} s_0^{12}$$

and for k=1,2

$$VAR(\hat{\theta}_{\mathrm{k}}) = \frac{1}{n} s_1^{kk} + \frac{1}{m} s_0^{kk}$$

Where, with $g, h = 1, 2$

$$s_1^{g,h} = \frac{1}{n-1} \sum_{i=1}^{n} \left(V_1^g(X_i) - \hat{\theta}_{\mathrm{g}}\right)\left(V_1^h(X_i) - \widehat{\ }_{\mathrm{h}}\right)$$

$$s_0^{g,h} = \frac{1}{m-1} \sum_{j=1}^{m} \left(V_0^g(Y_j) - \hat{\theta}_{\mathrm{g}}\right)\left(V_0^h(Y_j) - \hat{\theta}_{\mathrm{h}}\right)$$

$$V_1^g(X_i) = \frac{1}{m} \sum_{j=1}^{m} I(X_{ig}, Y_{jg}) \text{ and } V_0^g(Y_j) = \frac{1}{n} \sum_{i=1}^{n} I(X_{ig}, Y_{jg}), \quad i = 1, \cdots, n;\ j = 1, \cdots, m$$

Where $I(a, b)$ is defined as before.

2.2.2 Maximum Likelihood Method

For $i = 1, \cdots, n$, $X_i = \begin{pmatrix} X_{i1} \\ X_{i2} \end{pmatrix}$ is a bivariate normal random variable with the following joint pdf,

$$\mathrm{p}(\mathrm{x_{i1}}, \mathrm{x_{i2}}) = \frac{1}{2\pi\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2} \left(\begin{pmatrix} \mathrm{x_{i1}} \\ \mathrm{x_{i2}} \end{pmatrix} - \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \right)' \Sigma^{-1} \left(\begin{pmatrix} \mathrm{x_{i1}} \\ \mathrm{x_{i2}} \end{pmatrix} - \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \right) \right)$$

The log-likelihood function is

$$\sum_{i=1}^{n} log\left(p(x_{i1}, x_{i2})\right)$$

Then the maximum likelihood (ML) estimates are obtained by maximizing the above log-likelihood function. Specifically, for k=1,2,

$$\hat{\mu}_{\mathrm{k}} = \frac{1}{n} \sum_{i=1}^{n} x_{ik} \ \ and \ \ \hat{\Sigma} = \begin{pmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12} \\ \hat{\sigma}_{12} & \hat{\sigma}_2^2 \end{pmatrix} = \frac{1}{n} \sum_{i=1}^{n} \left(\begin{pmatrix} x_{i1} \\ x_{i2} \end{pmatrix} - \begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{pmatrix} \right)\left(\begin{pmatrix} x_{i1} \\ x_{i2} \end{pmatrix} - \begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{pmatrix} \right)'$$

Similarly, based on the data from the non-diseased group, the ML estimates for $\eta = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix}$ and $\Psi = \begin{pmatrix} \psi_1^2 & \psi_{12} \\ \psi_{12} & \psi_2^2 \end{pmatrix}$ are

$$\hat{\eta}_{\mathrm{k}} = \frac{1}{n}\sum_{i=1}^{n} y_{ik} \;\; and \;\; \hat{\Psi} = \begin{pmatrix} \hat{\psi}_1^2 & \hat{\psi}_{12} \\ \hat{\psi}_{12} & \hat{\psi}_2^2 \end{pmatrix} = \frac{1}{m}\sum_{j=1}^{m} \left(\begin{pmatrix} y_{j1} \\ y_{j2} \end{pmatrix} - \begin{pmatrix} \hat{\eta}_1 \\ \hat{\eta}_2 \end{pmatrix} \right) \left(\begin{pmatrix} y_{j1} \\ y_{j2} \end{pmatrix} - \begin{pmatrix} \hat{\eta}_1 \\ \widehat{}_2 \end{pmatrix} \right)'$$

Using these ML estimates, under normality, the AUC estimate for the k-th diagnostic test can be computed as follows,

$$\hat{\theta}_k = \Phi\left(\frac{\hat{\mu}_{\mathrm{k}} - \hat{\eta}_{\mathrm{k}}}{\sqrt{\hat{\sigma}_k^2 + \hat{\psi}_k^2}} \right), \qquad k = 1,2$$

For estimating the variances $\widehat{Var}(\hat{\theta}_k)\; for\; k = 1,2$ and the covariance $\widehat{COV}(\hat{\theta}_1, \hat{\theta}_2)$, we introduce $\hat{a}_{\mathrm{k}}$ and $\hat{b}_{\mathrm{k}}$ for k=1,2,

$$\hat{a}_{\mathrm{k}} = \frac{\hat{\mu}_{\mathrm{k}} - \hat{\eta}_{\mathrm{k}}}{\sqrt{\hat{\sigma}_k^2}} \;\; \text{and} \;\; \hat{b}_{\mathrm{k}} = \frac{\hat{\psi}_k^2}{\hat{\sigma}_k^2}$$

and rewrite the equation of $\hat{\theta}_k$ below,

$$\hat{\theta}_k = \Phi\left(\frac{\hat{a}_{\mathrm{k}}}{\sqrt{1 + \hat{b}_k^2}} \right), \qquad k = 1,2$$

Then the variances and the covariance are calculated using the Delta-method formulas derived by Liu and Schisterman (2003).

2.2.3 Generalized Inference

The principles of generalized inference on correlated ROC curves was outlined by Li (2007). For computational convenience, we adopted the generalized pivotal quantities according to Johnson and Wichern (2008), Bebu and Mathew (2008). Let $S(x)$ be the matrix of sums of squares of the cross-products based on the data from the diseased group,

$$SSX = \sum_{i=1}^{n} \left(\begin{pmatrix} X_{i1} \\ X_{i2} \end{pmatrix} - \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \right) \left(\begin{pmatrix} X_{i1} \\ X_{i2} \end{pmatrix} - \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \right)' = \begin{pmatrix} SSX_{11} & SSX_{12} \\ SSX_{12} & SSX_{22} \end{pmatrix}$$

and define $\sigma_1^* = \sigma_1 - \frac{\sigma_{12}}{\sigma_2}$ and $SSX_{11}^* = SSX_{11} - \frac{SSX_{12}^2}{SSX_{22}}$. Using the fact that SSX has a Wishart distribution $W(\Sigma, n-1)$, the following three variables are independent and have either chi-square or standard normal distribution (Johnson and Wichern (2008)).

$$C_{22} = \frac{SSX_{22}}{\sigma_2} \sim \chi_{n-1}^2, \;\; C_{11}^* = \frac{SSX_{11}^*}{\sigma_1^*} \sim \chi_{n-2}^2 \;\; and \; Z = \frac{\left(SSX_{12} - \frac{\sqrt{\sigma_{12}}}{\sigma_2} SSX_{22} \right)}{\sqrt{\sigma_1^* SSX_{22}}} \sim N(0,1)$$

Then, given ssx as the observed SSX, the three quantities below are the generalized pivotal quantities with respect to the variance-covariance matrix Σ for the diseased group, as they do not depend on any parameters, and their observed values are $\sigma_2, \sqrt{\sigma_{12}}\; and\; \sigma_{11}$. Let $R_\Sigma = \begin{pmatrix} a_{11}^2 & a_{12} \\ a_{12} & a_{22}^2 \end{pmatrix}$, where

$$a_{22} = \frac{ssx_{22}}{C_{22}}, \qquad a_{12} = \frac{ssx_{12}}{C_{22}} - \frac{Z\sqrt{ssx_{11}^{*} ssx_{22}}}{\sqrt{C_{11}^{*} C_{22}}}, \quad a_{11} = \frac{ssx_{11}^{*}}{C_{11}^{*}} + \frac{a_{12}^2}{a_{22}}$$

Similarly, we can get the below generalized pivotal quantities ($b_{11}^2, b_{12}, b_{22}^2$) with respect to the variance-covariance matrix Ψ for the non-diseased group. Define $R_{\Psi} = \begin{pmatrix} b_{11}^2 & b_{12} \\ b_{12} & b_{22}^2 \end{pmatrix}$.

$$b_{22} = \frac{ssy_{22}}{C_{22}}, \qquad b_{12} = \frac{ssy_{12}}{C_{22}} - \frac{Z\sqrt{ssy_{11}^{*} ssy_{22}}}{\sqrt{C_{11}^{*} C_{22}}}, \quad b_{11} = \frac{ssy_{11}^{*}}{C_{11}^{*}} + \frac{a_{12}^2}{a_{22}}$$

The generalized pivotal quantity for $\theta_1 - \theta_2$ is given by

$$\mathrm{R}_{\theta_1-\theta_2} = \Phi\left(\sqrt{\mathrm{R}_1}\right) - \Phi\left(\sqrt{\mathrm{R}_2}\right) \quad where\ R_k = R'_{\mu_k-\eta_k}\left(R_{\Sigma_k} + R_{\Psi_k}\right)^{-1} R_{\mu_k-\eta_k}$$

R_k depends on the three terms below, as the functions of the sample averages and the observed matrices of sum of squares of the cross products.

$$R_{\mu_k-\eta_k} = \bar{x}_k - \bar{y}_k - Z\left(\frac{R_{\Sigma_k}}{n} + \frac{R_{\Psi_k}}{m}\right)^{\frac{1}{2}}, \quad Z \sim N(0,1)$$

Where R_{Σ_k} and R_{Ψ_k} are defined as before.

To obtain $\hat{\theta}_1 - \hat{\theta}_2$ and its variance estimate for each setting, $R_{\mu_k-\eta_k}$, R_{Σ_k} and R_{Ψ_k} were simulated 1,000 times from the standard normal distributions and the chi-square distributions as defined before, then $\mathrm{R}_{\theta_1-\theta_2}$ at each simulation was calculated and its median across 1,000 simulations was used as $\hat{\theta}_1 - \hat{\theta}_2$. The variance of $\mathrm{R}_{\theta_1-\theta_2}$ was calculated as the variance estimate of $\hat{\theta}_1 - \hat{\theta}_2$.

2.3 Equivalence Test

It is assumed that accuracy of two diagnostic tests will be compared using the equivalence test for the difference in areas of correlated ROC curves, with Δ as equivalence margin,

$$H_0: |\theta_1 - \theta_2| \geq \Delta \text{ vs. } H_A: |\theta_1 - \theta_2| < \Delta$$

Then, using normal approximation, the test power can be expressed below, given confidence level of $1 - \alpha$ and true area difference of d.

$$P\left(\frac{\Delta - \left|\hat{\theta}_1 - \hat{\theta}_2\right|}{\sqrt{\widehat{VAR}\left(\hat{\theta}_1 - \hat{\theta}_2\right)}} > Z_{1-\alpha} \mid |\theta_1 - \theta_2| = d < \Delta\right)$$

3. Simulations

Simulation was performed to compare the performance of the three methods applied in the equivalence testing. We focus on comparisons of the equivalence test powers. As the

simulation requires quite a few of parameters, it is only practical to consider some useful settings (total 1792 settings; 1792 = 2 x 4 x 4 x 8 x 7) as given below. For each setting, 5,000 replicates of X and Y were simulated. Then each replicate of X and Y per setting was used for calculation of $\hat{\theta}_1$, $\hat{\theta}_2$, $\widehat{VAR}(\hat{\theta}_1 - \hat{\theta}_2)$, and the test powers.

- Δ=0.1 or 0.2
- Sample sizes (n, m): (50,50), (50,500), (500,50), (500,500)
- Correlation coefficients (ρ_X, ρ_Y): (0.1,0.1), (0.1,0.6), (0.6,0.6)
- θ_k: $\theta_1 =$ 0.3, 0.8; θ_2 via $|\theta_1 - \theta_2| = p\Delta$ p =0.2, 0.8
- Standard deviations: seven cases were specified

Table 1: Cases of standard deviations used for simulation

SD	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7
ψ_1	1	1	1	1	1	10	10
ψ_2	1	1	1	1	10	1	10
σ_1	1	1	10	10	10	10	10
σ_2	1	10	1	10	10	10	10

- **Means:** μ_k calculated via the standard deviations, θ_k, and $\eta_k = 10\psi_k$

4. Results

In each setting, the test powers were summarized in terms of medians across 5,000 replicates. Then for each of the three methods, a setting was counted once if the corresponding median power is over 80%. It is observed that the Delong's method produced total 155 successful settings, and the MLE and the GPQ methods yielded more successful settings (i.e., 187 and 197 settings, respectively). Table 2 below summarizes the numbers of successful settings by one or more methods.

Table 2: Numbers of counted settings summarized by comparison and counting methods

Counting Method	Successful Method(s)					
	Only GPQ	Only Delong's	Only MLE	Delong's & GPQ	MLE & GPQ	All Three
Median Power > 80%	22	1	19	7	21	147

The influence of sample sizes and population parameters can also be reflected by the numbers of successful settings summarized in Table 3. Increasing the sample size in one or both groups would result in higher chance of success (i.e., with median power > 0.8). None of the settings with n=50 and m=50 achieved at least 80% median power. Elevation of standard deviations appears to have little impact on the number of successful settings for all the three methods. Increasing the correlation tends to yield higher success rate. In summary, regardless of the sample sizes and the correlations, the MLE method and the GPQ method yielded better results than the Delong's method.

Table 3: Numbers of counted settings summarized over sample sizes and population parameters

Method	Sample Sizes (n, m)			Standard Deviations			Correlations (ρ_X, ρ_Y)		
	(50,500)	(500,50)	(500,500)	Case 1	Each of Case 2 ~ 6	Case 7	(0.1,0.1)	(0.1,0.6)	(0.6,0.6)
Delong's	24	43	88	23	20~22	25	41	54	60
MLE	41	38	108	26	24~29	26	46	61	80
GPQ	25	60	112	27	25~35	29	52	63	82

Table 4 presents the numbers of successful settings under different hypothesis testing parameters. The evident difference between the Delong's method and the other two methods occurs at the two scenarios with $p = 0.8$ (i.e., $(\theta_1, \Delta, p) = (0.3,0.2,0.8)$ and $(0.8,0.2,0.8)$). As the two scenarios correspond to the largest $|\theta_1 - \theta_2|$, the results suggest to chose the MLE method or the GPQ method when the true AUC difference is large.

Table 4: Numbers of counted settings summarized by hypothesis testing parameters

Method	Hypothesis Testing Parameters: (θ_1, Δ, p)					
	(0.3,0.1,0.2)	(0.3,0.2,0.2)	(0.3,0.2,0.8)	(0.8,0.1,0.2)	(0.8,0.2,0.2)	(0.8,0.2,0.8)
Delong's	21	53	0	23	54	4
MLE	26	48	6	28	53	26
GPQ	24	53	8	34	54	24

The above tables only present the numbers of counted or successful settings. In order to understand the performance of all three methods across all settings also including those unsuccessful ones (i.e., with median power < 80%), the 5%, 50% and 95% percentiles of all calculated powers were also obtained. Moreover, considering that the test powers can be impacted by the estimation bias with respect to $\theta_1 - \theta_2$, we also summarized the bias using 5%, 50% and 95% percentiles and present them together with the percentiles of the test powers. The Delong's method tends to produce larger bias than those from the other two methods.

Table 5: Percentiles of equivalence test powers and bias of $\theta_1 - \theta_2$ across all settings

Method	Power: Median / (5%,95%)	Bias: Median / (5%,95%)
Delong's	41.2% / (7.2%,>99.9%)	0.0284 / (0.0115, 0.0530)
MLE	60.6% / (<0.1%, >99.9%)	0.0264 / (0.00960, 0.0469)
GPQ	64.2% / (11.1%, >99.9%)	0.0263 / (0.0105, 0.0452)

5. Discussion

As shown by the above results, in most of the scenarios (all under bivariate normal distributions), the GPQ method tends to outperform the other two methods and the Delong's method appears to be the least preferable. However, only about 10% of the settings are able to achieve 80% median of the powers, it would be desirable to summarize the results with lower median powers and to extend the simulation to consider other scenarios. For example, more attention may be focused on the distributions other than normal, or the comparisons among more than two ROC curves. In addition, if historical

data or prior knowledge is available, then choice of the methods and the equivalence margins may be examined with use of the prior information.

References

L. E. Bantis, Z. Feng (2016). Comparison of Two Correlated ROC Curves at a Given Specificity or Sensitivity Level, Statistics in Medicine, 35(24), pp. 4352-4367.

E. R. DeLong, D. M. DeLong, D. L. Clarke-Pearson (1988). Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves. Biometrics, 44(3), pp. 837-845.

B. D. Gallas, L. Pesce (2009). Comparison of ROC methods for partially paired data, Proceedings of SPIE – The International Society for Optical Engineering.

J. A. Hanley, B. J. McNeil (1983). A Method of Comparing the Areas under Receiver Operating Characteristic Curves Derived from the Sample Cases. Radiology, 148, pp. 839-843.

R. A. Johnson, D. W. Wichern (2008). Applied Multivariate Statistical Analysis, 6th Edition, New York: Wiley.

C. R. Li (2007). Exact Inferences on Paired ROC Curves, Dissertation.

J. P. Liu, M. C. Ma, C. Y. Wu, J. Y. Tai (2006). Tests of Equivalence and Non-inferiority for Diagnostic Accuracy Based on the Paired Areas under ROC Curves. Statistics in Medicine, 25, pp. 1219-1238.

A. Liu, E. F. Schisterman (2003). Comparison of Diagnostic Accuracy of Biomarkers with Pooled Assessments, Biometrical Journal, 45, pp 631-644.

C. E. Metz, P. L. Wang, H. B. Kroman (1984). A New Approach for Testing the Significance of Differences between ROC Curves for Correlated Data. In: Deconick F, editor. Information Processing in Medical Imaging.1st edition. Nijhoff, pp. 432-445.

M. B. Rao, C. R. Rao (2014). Handbook of Statistics: Computational statistics with R, Amsterdam: Elsevier.

X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J. C. Sanchez, M. Muller (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves, BMC Bioinformatics, 7, 77.

L. Tiana, A. Vexlera, L. Yan, E. F. Schisterman (2009). Confidence interval estimation of the difference between paired AUCs based on combined biomarkers, J. Stat Plan Inference, 139(10), pp. 3725-3732.

S. Wan, B. Zhang (2008). Comparing Correlated ROC Curves for Continuous Diagnostic Tests under Density Ratio Models. Computational Statistics & Data Analysis, 53(1), pp. 233-245.

D. Zhang, B. Zhang (2014). Semiparametric Empirical Likelihood Confidence Intervals for the Difference of Areas under Two Correlated ROC Curves under Density Ratio Model, 56(4), pp. 678-696.

X. H. Zhou, N. A. Obuchowski, D. K. McClish (2002). Statistical Methods in Diagnostic Medicine, John Wiley and Sons, New York.