Dhaka Univ. J. Sci. 61(1): 131-138, 2013 (January)

Comparing Two Diagnostic Processes: Operating Characteristics of Nonparametric Methods

Md. Akhtar Hossain and Nahid Sultana Sumi

Department of Statistics, Biostatistics & Informatics, University of Dhaka, Dhaka - 1000, Bangladesh

Received on 19.07.2012. Accepted for Published on 18. 09. 2012

Abstract

The area under ROC curve (AUC) is frequently used as a measure for the effectiveness of diagnostic processes. The aim of this paper is to explore and evaluate several nonparametric test methods of comparing the effectiveness and performance of two competing diagnostic processes producing quantitative ratings. These nonparametric methods make use of ROC curves when comparing the diagnostics processes. An extensive simulation study is performed to investigate the operating characteristics of the test methods in a wide range of settings.

Keywords: Diagnostic process, Receiver operating characteristic (ROC) curve, Area under curve (AUC), Nonparametric tests.

I. Introduction

A great use of diagnostic processes is made in medical studies based on clinical observations or laboratory methods to specify which individuals are classified as nondiseased or as diseased. Diagnostic processes provide important medical decision making with improved technology to detect disease. During the decades, receiver operating characteristic curve (ROC) analysis has been used as a popular method of evaluating the performance or discriminatory power of diagnostic processes. The ROC curve is a plot of the diagnostic process's sensitivity versus 1-specificity at various observed value of the process. It has been used in many areas such as radiology[1], psychiatry[2], epidemiology[3], biomedical informatics[4], non-destructive testing[5] and manufacturing inspection systems[6].

For statistical analysis, a recommended index of accuracy associated with an ROC curve is the area under the curve[7]. The area under the ROC curve (AUC) is interpreted as the probability that the observed value of the diagnostic process will be greater for a randomly selected diseased individual than for a randomly selected nondiseased individual assuming that the higher values of a diagnostic process are associated with diseased individuals, while lower values are associated with nondiseased[8]. Thus, AUC lies between 0 and 1 and the greater the AUC, the better the discriminatory power of the diagnostic process[9].

For comparing two diagnostic processes, the difference between AUCs is often used. In the field of diagnostic imaging it is widely recognized that the variability due to subjects represents a substantial component of the overall variability of the AUC. To better control for the sources of variability when comparing diagnostic processes, a paired study design is often implemented. This type of design usually induces positive correlation between the ratings of the same subjects.

Various parametric and nonparametric methods have been suggested to compare the accuracy of two diagnostic processes within a paired design setting. DeLong *et. al*[8] developed a conventional fully nonparametric approach leading to an asymptotically normal test statistic. Venkatraman and Begg[10] prescribed a permutation test for testing equality of two ROC curves at every operating point. Bandos *et. al*[11] described an exact nonparametric method to test equality of two correlated ROC curves. Their method modifies the permutation test for comparing correlated ROC curves by Venkatraman and Begg[10] by using an AUC difference index rather than an index of equality of ROC curves at each operating point.

The scope of this article is limited to a brief review and comparing performances of widely used nonparametric methods suggested by DeLong *et. al*[8], Venkatraman and Begg[10] and Bandos *et. al*[11].

II. Estimation of AUC

For two diagnostic processes, suppose there are N individuals without disease and M individuals with disease. Suppose X^m and Y^m $(m = 1, 2)$ denote the corresponding patients without disease and with disease, respectively. Corresponding bivariate outcomes should be x_i^m $(i = 1, 2, \cdots, N)$ and y_j^m $(j = 1, 2, \cdots, M)$ respectively for two diagnostic processes on the same N nondiseased and M diseased individuals. Bivariate cumulative distribution functions are denoted by $F(x^1, x^2)$ and $G(y^1, y^2)$, and their corresponding marginal $F_m(x^m), G_m(y^m)$ $(m = 1, 2)$. Bamber[12] noted that the area under the ROC curve is equal to $P(Y > X)$. Let A_m $(m = 1, 2)$ be the areas under the respective ROC curves of diagnostic process 1 and 2. The methods of DeLong *et. al*[8] and Bandos *et. al*[11] test the null hypothesis $H_0 : A_1 = A_2$ versus the alternative $H_1 : A_1 \neq A_2$.

The area under an empirical ROC curve can be computed by trapezoidal rule[12]. Hanley and McNeil[13] showed that the area computed by trapezoidal rule under an empirical ROC curve is equal to the Mann - Whitney U statistic for comparing distributions of values from the two samples. The formula that Hanley and McNeil[13] suggested for computing the area under the ROC curve is given as

$$A = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} g\left(X_i, Y_j\right)$$

where, A = Area under the ROC curve, M = Number of diseased individuals, N = Number of nondiseased individuals, Y_j = The test score of j^{th} patient with disease, X_i = The test score of the i^{th} patient without disease and g is a function comparing X_i with Y_j such that

$$g\left(X_i, Y_j\right) = \begin{cases} 1 & \text{if } \; Y_j > X_i \\ 0.5 & \text{if } \; Y_j = X_i \\ 0 & \text{otherwise} \end{cases}$$

So for the m^{th} diagnostic process the area under ROC curve can be computed as

$$A_m = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} g\left(X_i^m, Y_j^m\right)$$

III. Nonparametric Methods of Comparing Diagnostic Processes

DeLong *et. al*'s Conventional Test

DeLong *et. al*[8] has developed a totally nonparametric approach for comparing areas from two samples on the same subjects by using the theory of generalized U statistics. The method of structural components is used to generate an estimated covariance matrix, and the resulting test statistic has asymptotically a χ^2 distribution.

For m different diagnostic measures with $\left\{Y_j^r\right\}$, $\left\{X_i^r\right\}$ $\left(i = 1, 2, \ldots, N;\; j = 1, 2, \ldots, M\right)$ and $\hat{A}_r, r = 1, 2, \ldots, m$,

$$V_{10}^r\left(Y_j\right) = \frac{1}{N}\sum_{i=1}^{N} g\left(Y_j, X_i\right),\; j = 1, 2, \ldots, M \quad \text{and}$$

$$V_{01}^r\left(X_i\right) = \frac{1}{M}\sum_{j=1}^{M} g\left(Y_j, X_i\right),\; i = 1, 2, \ldots, N$$

The $m \times m$ matrices S_{10} and S_{01} with $\left(r,s\right)^{th}$ elements,

$$S_{10}^{r,s} = \frac{1}{M-1}\sum_{j=1}^{M}\left[V_{10}^r\left(Y_j\right) - \hat{A}_r\right]\left[V_{10}^s\left(Y_j\right) - \hat{A}_s\right] \quad \text{and}$$

$$S_{01}^{r,s} = \frac{1}{N-1}\sum_{i=1}^{N}\left[V_{01}^r\left(X_i\right) - \hat{A}_r\right]\left[V_{01}^s\left(X_i\right) - \hat{A}_s\right]$$

Now the estimated covariance matrix for the vector $\left(\hat{A}_1, \hat{A}_2, \ldots, \hat{A}_m\right)$ of estimated areas under the curves of $S^{r,s}$ is obtained as, $S = \frac{1}{M}S_{10} + \frac{1}{N}S_{01}$.

DeLong *et. al*[8] thus, showed that for any contrast LA', where L is a row vector of coefficients, $\frac{L\hat{A}' - LA'}{\left[LSL'\right]^{\frac{1}{2}}}$ has a standard normal distribution. Squaring this, the test statistic then takes the form,

$$\left(\hat{A} - A\right)L'\left[LSL'\right]^{-1}L\left(\hat{A} - A\right)'$$

which has a chi-square distribution with degrees of freedom equal to the rank of LSL'.

Bandos *et. al*'s Area Test

Bandos *et. al*[11] derived exact and asymptotic permutation test methods to test the equality of two correlated ROC curves which are designed to have increased power to detect differences in the AUCs. If $\left\{X_i^m\right\}_{i=1}^{N}$, $\left\{Y_j^m\right\}_{j=1}^{M}$ be the ratings observed in the diagnostic process m for N actually nondiseased and M actually diseased individuals and $\left\{x_i^m\right\}_{i=1}^{N}$, $\left\{y_j^m\right\}_{j=1}^{M}$ be appropriately transformed ratings, an unbiased nonparametric estimator for the area under the ROC curve for diagnostic process m can be written as $\hat{A}_m$. For a paired design, the difference in two AUCs can be estimated as,

$$\hat{A}_1 - \hat{A}_2 = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\psi\left(x_i^1, y_j^1\right)}{NM} - \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\psi\left(x_i^2, y_j^2\right)}{NM}$$

where,

$$\psi\left(x_i^1, y_j^1\right) - \psi\left(x_i^2, y_j^2\right) = \begin{cases} 1 & x_i^1 < y_j^1, x_i^2 > y_j^2 \\ \frac{1}{2} & x_i^1 < y_j^1, x_i^2 = y_j^2 \;\text{ or }\; x_i^1 = y_j^1, x_i^2 > y_j^2 \\ 0 & x_i^1 < y_j^1, x_i^2 < y_j^2 \;\text{ or }\; x_i^1 > y_j^1, x_i^2 > y_j^2 \;\text{ or }\; x_i^1 = y_j^1, x_i^2 = y_j^2 \\ -\frac{1}{2} & x_i^1 > y_j^1, x_i^2 = y_j^2 \;\text{ or }\; x_i^1 = y_j^1, x_i^2 < y_j^2 \\ -1 & x_i^1 > y_j^1, x_i^2 < y_j^2 \end{cases}$$

As a member of U statistics the nonparametric estimator of the AUC difference is known to be asymptotically normally distributed under quite general conditions[14]. Based on this property and the additional assumption of exchangeability, they constructed a simple asymptotic test procedure with test statistic

$$\frac{\hat{A}_1-\hat{A}_2}{\sqrt{Var_{\Omega}\left(\hat{A}_1-\hat{A}_2\right)}} \xrightarrow{d} N(0,1)$$

where, Ω is the parameter space.

Venkatraman and Begg's Permutation Test

A general problem is that although the two diagnostic processes may have different ROC curves, they may have same area. But one diagnostic process may genuinely be superior to the other despite having the same area. In order to detect the differences between two ROC curves Venkatraman and Begg[10] developed a simple permutation test. The test proposed by Venkatraman and Begg[10] is thus for the equality of the underlying ROC curves and is executed by permuting the labels of the two diagnostic processes within each diseased and nondiseased subject. Such an approach implicitly assumes that both diagnostic processes are exchangeable within subject and requires an appropriate transformation, such as ranks, for diagnostic processes differing in scale. Let $t_k^m=\left(x_i^m, y_j^m\right)$ $\left(m=1,2;k=1,2,\cdots,n;i=1,2,\cdots,N;j=1,2,\cdots,M\right)$ be the observed results from a total of $n=N+M$ (N nondiseased and M diseased) individuals under the m^{th} diagnostic process. For empirical calculation of the test of Venkatraman and Begg[10], the entire data set can be denoted by $\left\{T_k^1, T_k^2, D_k; k=1,2,\cdots,N+M\right\}$ where $D_k=1$ if the case is diseased and $D_k=0$ if the case is nondiseased. The corresponding ranks of $\left\{T_k^1\right\}$ and $\left\{T_k^2\right\}$ be denoted by $\left\{R_k\right\}$ and $\left\{S_k\right\}$, respectively. Using the rank statistics of the observed diagnostic values and setting $l=1,2,\ldots,n-1$ an empirical error matrix is defined by

$$e_{kl}=\begin{cases} 1 & \text{if } \left(R_k\le l, S_k>l, D_k=0\right) \text{ or } \left(R_k>l, S_k\le l, D_k=1\right), \\ -1 & \text{if } \left(R_k>l, S_k\le l, D_k=0\right) \text{ or } \left(R_k\le l, S_k>l, D_k=1\right), \\ 0 & \text{otherwise} \end{cases}$$

The statistic $e_{.l}=e_{1l}+e_{2l}+\cdots+e_{nl}$ is a measure of the closeness of the two ROC curves at the l^{th} order statistic. The corresponding overall test statistic is $E=\sum_{l=1}^{n-1}\left|e_{.l}\right|$.

If the two diagnostic processes are evaluated on the same metric, and there is no systematic measurement bias, then the test values for any subject can be directly exchanged to generate the permutation distribution. If $\left(q_1,q_2,\ldots,q_n\right)$ represent a sequence of 0's and 1's, then a permuted data set $\left\{T_k^{1*}, T_k^{2*}\right\}$ indexed by that sequence is given by

$T_k^{1*}=q_kT_k^1+\left(1-q_k\right)T_k^2$,

$T_k^{2*}=q_kT_k^2+\left(1-q_k\right)T_k^1 \quad \left(k=1,2,\ldots,n\right)$.

A new set of ranks $\left\{R_k^*, S_k^*\right\}$ is evaluated based on $\left\{T_k^{1*}, T_k^{2*}\right\}$ and a corresponding statistic E^* is computed. The permutation distribution is the distribution which assigns a uniform mass to each value of E^* given by all the 2^n sequences of 0's and 1's. Since this may be a very large number, a sampling scheme is used where $\left(q_1,q_2,\ldots,q_n\right)$ is a random permutation generated by n fair coin tosses and the process is repeated a sufficiently large number of times to obtain a stable p - value.

If the direct exchangeability of T_k^1 and T_k^2 is not considered to be an appropriate assumption, then it is necessary to rely on the ranked samples to evaluate the p - value. In this case each permuted set of ranks is generated by randomly exchanging pairs of ranks and reranking them. That is, the set of ranks $\left\{R_k^*, S_k^*\right\}$ will be first generated by using

$R_k^*=q_kR_k+\left(1-q_k\right)S_k$,

$S_k^*=q_kS_k+\left(1-q_k\right)R_k \quad \left(k=1,2,\ldots,n\right)$

This process will invariably introduce numerous ties, so it is necessary to have a second randomization step to break the ties, that is, to generate $\left\{R_k^{**}, S_k^{**}\right\}$, where

$R_k^{**}=J\left(R_k^*\right) \quad S_k^{**}=J\left(S_k^*\right) \quad \left(k=1,2,\ldots,n\right)$

where $J(.)$ represents the process by which tied ranks are re-ranked by randomization.

IV. Simulation Study

An extensive simulation study has been performed to compare empirical test sizes (Type I errors) and power of the nonparametric test for different underlying AUC differences, correlation between diagnostic processes and different sample sizes. Four different practically possible scenarios as presented in Figure 1 are covered in this simulation study. For data generation purpose, we have assumed two continuous measurements for each nondiseased individual from a bivariate normal distribution centered at $\mu_x = 0$ with both measurements having a marginal variance of 1.0. That is $\mu_{x^m} = 0$ and

$\sigma^2_{x^m} = 1,\ m = 1,\ 2$. So we have

$$\Phi^{-1}(A_m) = \frac{\mu_{y^m}}{\sqrt{1+\sigma^2_{y^m}}},\ m = 1,\ 2 \tag{1}$$

where, Φ^{-1} is the percentile of standard normal distribution[15]. As two ROC curves drawn from measurements having equal variances never cross each other, two continuous measurements for each diseased individual from a bivariate normal distribution centered at μ_y with both measurements having a marginal variance of 1.0 are taken for diagnostic processes with non-crossing ROC curves. For diagnostic processes having crossing ROC curves unequal variances ($\sigma^2_{y^1} = 1.0$ & $\sigma^2_{y^2} = 4.0$) are assumed. In all the cases, three different correlation values ($\rho = 0.25, 0.50, 0.75$) are considered. The values of μ_y are directly determined from A_1 and A_2 from equation (1).

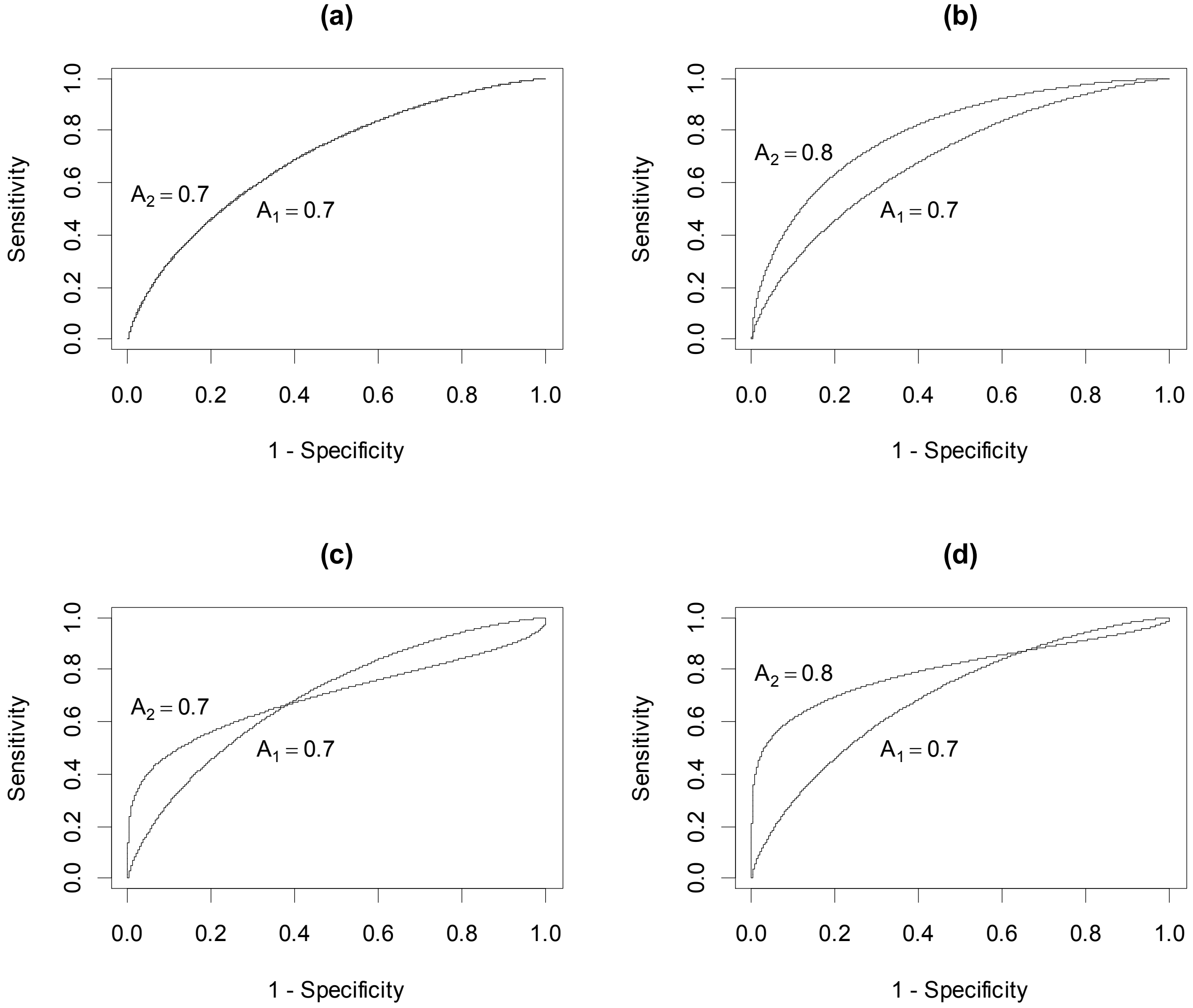

Fig. 1. ROC curves. (a). Non-crossing ROC curves with same areas; (b). Non-crossing ROC curves with different areas; (c) Crossing ROC curves with same areas and (d) Crossing ROC curves with different areas.

Table. 1. Empirical Test size when comparing two diagnostic tests with same areas and non-crossing ROC curves [A_1 - Area of diagnostic process 1; A_2 - Area of diagnostic process 2; D - DeLong *et. al* Test; V - Venkatraman & Begg Test; B - Bandos *et. al* Test].

Area	Sample Size	$\rho = 0.25$			$\rho = 0.50$			$\rho = 0.75$		
A_1, A_2	N, M	D	V	B	D	V	B	D	V	B
.60, .60	20, 20	**.074**	.056	**.065**	.060	.047	.050	.047	.052	.040
	40, 40	.059	.044	.050	.046	.043	.039	.053	.051	.049
	80, 80	.058	.058	.054	.053	.054	.048	.042	.044	.040
.70, .70	20, 20	.060	.050	.052	**.065**	.054	.049	**.066**	.055	.052
	40, 40	.054	.047	.044	.061	.055	.057	.062	.054	.060
	80, 80	.050	.047	.049	.045	.049	.044	.052	.053	.052
.80, .80	20, 20	.050	**.032**	**.030**	**.067**	.044	.049	.064	.043	.052
	40, 40	.046	**.035**	.041	.040	.037	**.035**	.042	.048	.042
	80, 80	.043	.036	.038	.056	.056	.053	.049	.063	.051
.90, .90	20, 20	.044	.040	.042	.056	.045	.042	.045	.040	.045
	40, 40	.043	.036	.037	.048	.045	.047	.046	.060	.049
	80, 80	.042	.037	.042	.051	.054	.053	**.065**	.053	.058

For each scenario 1000 replications are computed and both empirical test size and empirical power are obtained for sample sizes 20, 40 and 80. The rejection region for the tests are determined using a nominal significance level of $\alpha = 0.05$. The empirical nominal values are compared with the approximate 95% confidence interval (0.036, 0.064) around nominal size of 0.05 based on a binomial sample of 1000 repetitions.

For both non-crossing and crossing ROC curves, given the values of AUCs and variances, the mean values of diagnostic scores for diseased and nondiseased individuals can be obtained from relation (1) and the variance-covariance matrix can be constructed as

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix}$$

V. Result and Discussion

Table 1 presents the empirical test sizes obtained for the discussed nonparametric test methods when testing for equality in performance of two diagnostic process having non-crossing ROC curves and same areas. An example of such configuration is shown in Figure 1(a). The bold entries in Table 1 are the empirical test sizes obtained outside of the approximate 95% confidence interval (.036, .064). It is clear from results that, empirical size of the test suggested by Venkatraman and Begg[10] seems to be less conservative than tests by DeLong *et. al*[8] and Bandos *et. al*[11]. This is especially evident with smaller sample sizes.

Along with the empirical nominal sizes we also have considered the statistical power of the test methodologies to assess their performance. The power of a statistical hypothesis test procedure is defined as 1 – Type II error that is, the rate of rejecting the null hypothesis when it was false. Table 2 depicts the calculated empirical power of the nonparametric test methods when testing equality of performance of two diagnostic processes having non-crossing ROC curves and different area values. As portrayed in Figure 1(b), in this case, one diagnostic process is uniformly superior in performance than the other. The results calculated for a number of set up in Table 2 make it clear that the test suggested by DeLong *et. al*[8] exhibits better power than the other two. The power of the tests increases with increasing correlation and sample size. Though the test by Venkatraman and Begg[10] gives lowest power in almost all set up, it's power is very closed to test by Bandos *et. al*[11]. For large sample sizes and higher areas differences, the empirical power for each of all three nonparametric tests tends to others.

Table. 2. Empirical power when comparing two diagnostic tests with different areas and non-crossing ROC curves [A_1 - Area of diagnostic test 1; A_2 - Area of diagnostic test 2; D - DeLong *et. al* Test; V - Venkatraman & Begg Test; B - Bandos *et. al* Test].

Area	Sample Size	$\rho = 0.25$			$\rho = 0.50$			$\rho = 0.75$		
A_1, A_2	N, M	D	V	B	D	V	B	D	V	B
.60, .70	20, 20	.153	.129	.127	.176	.164	.170	.262	.242	.261
	40, 40	.256	.235	.239	.325	.304	.310	.504	.464	.491
	80, 80	.410	.404	.402	.574	.542	.568	.870	.817	.854
.60, .80	20, 20	.470	.430	.430	.629	.585	.591	.832	.795	.781
	40, 40	.764	.743	.733	.894	.874	.870	.994	.989	.988
	80, 80	.973	.970	.968	.993	.993	.992	1.00	1.00	1.00
.60, .90	20, 20	.906	.894	.864	.965	.958	.937	.998	.999	.991
	40, 40	.998	.996	.995	1.00	1.00	.999	1.00	1.00	1.00
	80, 80	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
.70, .80	20, 20	.181	.168	.164	.211	.194	.203	.317	.302	.331
	40, 40	.330	.314	.325	.407	.384	.397	.660	.598	.639
	80, 80	.544	.523	.534	.712	.680	.702	.929	.893	.918
.70, .90	20, 20	.655	.624	.601	.746	.722	.695	.932	.924	.894
	40, 40	.929	.915	.900	.977	.969	.965	.999	.999	.999
	80, 80	.998	.997	.996	1.00	1.00	1.00	1.00	1.00	1.00
.80, .90	20, 20	.252	.236	.246	.290	.278	.292	.429	.394	.434
	40, 40	.466	.448	.447	.605	.578	.589	.838	.789	.828
	80, 80	.773	.761	.769	.881	.863	.868	.989	.987	.987

Table. 3. Empirical power when comparing two diagnostic tests with same areas but different (crossing) ROC curves [A_1 - Area of diagnostic test 1; A_2 - Area of diagnostic test 2; D - DeLong *et. al* Test; V - Venkatraman & Begg Test; B - Bandos *et. al* Test].

Area	Sample Size	$\rho = 0.25$			$\rho = 0.50$			$\rho = 0.75$		
A_1, A_2	N, M	D	V	B	D	V	B	D	V	B
.60, .60	20, 20	.066	.096	.061	.059	.079	.043	.062	.131	.050
	40, 40	.057	.147	.054	.071	.186	.065	.046	.236	.049
	80, 80	.050	.348	.051	.047	.434	.047	.046	.574	.050
.70, .70	20, 20	.061	.093	.059	.044	.080	.042	.057	.112	.056
	40, 40	.060	.136	.057	.050	.196	.045	.055	.282	.057
	80, 80	.048	.383	.048	.035	.463	.041	.056	.612	.059
.80, .80	20, 20	.042	.078	.042	.045	.084	.044	.057	.114	.059
	40, 40	.041	.122	.044	.051	.163	.050	.040	.236	.041
	80, 80	.044	.323	.044	.047	.410	.053	.047	.532	.054
.90, .90	20, 20	.040	.057	.054	.038	.060	.053	.050	.084	.061
	40, 40	.044	.093	.046	.044	.112	.049	.051	.151	.055
	80, 80	.048	.207	.049	.041	.252	.046	.046	.337	.050

Table 3 and Table 4 demonstrate the empirical statistical power of the test methods for testing null hypothesis of equal performance of two diagnostic processes having crossing ROC curves. Particularly, in Table 3, we assumed that the two processes have same overall area under curve but they have different ROC curves and consequently different in their performance. As described in Figure 1(c), this is a case where each of diagnostic processes has partially better performance than the other but none performs uniformly better.

Table. 4. Empirical power when comparing two diagnostic tests with different areas as well as different (crossing) ROC curves [A_1 - Area of diagnostic test 1; A_2 - Area of diagnostic test 2; D - DeLong *et. al* Test; V - Venkatraman & Begg Test; B - Bandos *et. al* Test].

Area	Sample Size	$\rho = 0.25$			$\rho = 0.50$			$\rho = 0.75$		
A_1, A_2	N, M	D	V	B	D	V	B	D	V	B
.60, .70	20, 20	.165	.203	.152	.179	.221	.171	.223	.283	.212
	40, 40	.209	.310	.206	.273	.418	.277	.311	.519	.318
	80, 80	.414	.669	.412	.465	.775	.475	.547	.886	.571
.60, .80	20, 20	.487	.504	.456	.475	.508	.418	.544	.597	.497
	40, 40	.705	.755	.678	.791	.856	.768	.848	.903	.831
	80, 80	.952	.982	.948	.977	.991	.997	.992	.997	.992
.60, .90	20, 20	.842	.845	.815	.912	.920	.897	.945	.958	.948
	40, 40	.992	.994	.990	.999	.998	.998	.999	.999	.999
	80, 80	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
.70, .80	20, 20	.187	.211	.179	.192	.234	.191	.222	.282	.230
	40, 40	.304	.423	.302	.309	.446	.316	.389	.574	.417
	80, 80	.484	.739	.497	.597	.812	.608	.628	.888	.656
.70, .90	20, 20	.560	.596	.545	.623	.661	.612	.695	.735	.703
	40, 40	.853	.886	.844	.903	.938	.900	.952	.967	.956
	80, 80	.991	.998	.990	.998	1.00	.998	.999	1.00	.999
.80, .90	20, 20	.242	.269	.242	.243	.275	.255	.294	.353	.330
	40, 40	.402	.483	.412	.482	.576	.516	.565	.662	.594
	80, 80	.712	.839	.726	.774	.907	.790	.848	.950	.871

The configurations considered in Table 3 show very interesting results. From these results, it is apparent that both the test methods by DeLong *et. al*[8] and Bandos *et. al*[11] ignore the difference in the ROC curves (the crossing nature) when testing for equality in performance of two diagnostic processes. As a result, these two test methods show very little power in all the set up. A reverse result is observed for Vekatraman and Begg[10] test method. Despite of same area values, it takes the difference in ROC curves and exhibits far better power than other two test methods in all set up. Table 4 compares two diagnostic processes those are different in areas as well as in ROC curves. Figure 1(d) elucidates this configuration. As seen in Table 3, the test by Venkatraman and Begg[10] performs better again in this configuration. The same argument makes the difference here too. The virtue of tracking difference in the ROC curves along with the difference in areas keeps the Venkatraman and Begg[10] method ahead.

VI. Conclusion

Comparing classification and discriminatory performance of two diagnostic processes is of great interest in many practical research fields including medical science, signal processing, engineering, bioinformatics etc. There are a number of methodologies devised for this purpose. Little knowledge regarding operating conditions of these methods often confuses the researchers to select appropriate method in their respective applications. In this article three very commonly used and competing nonparametric test methodologies to compare performance of two competing diagnostic processes namely, DeLong *et. al*[8] test, Venkatraman and Begg[10] test and Bandos *et. al*[11] test, are discussed and the operating characteristics of these methods are explored and compared through extensive simulation. The simulation study depicts that when the two diagnostic processes only differ in areas and have non-crossing ROC curves, the DeLong's test[8] method exhibits better operating characteristics though it provides little conservative measures of Type I error. The method by Bandos *et. al*[8] can be considered as a very close alternative in such configurations. On the other hand, if the two diagnostic processes have crossing ROC curves with same or different areas, the Venkatraman and Begg's method[10] can be employed without a second thought. DeLong's[8] and Bandos's[11] test methods expose very weak operating characteristics in these scenarios. It is expected that, the findings of this study will be of help for researchers to avoid confusion and to select between the competing test methods more confidently.

………………..

1. Metz, C. E., 1989. Some practical issues of experimental design and data analysis in radiological ROC studies. *Investigation Radiology*, **24**, 234 – 245.

2. Hsiao, J. K., J. J. Barko and W. Z. Potter, 1989. Diagnosing diagnoses: receiver operating characteristic methods and psychiatry. *Archives of General Psychiatry*, **46**, 664 – 667.

3. Aoki, K., J. Misumi, T. Kimura, W. Zhao and T. Xie, 1997. Evaluation of cutoff levels for screening of gastric cancer using serum pepsinogens and distributions of levels of serum pepsinogens I, II and of PG I/PG II ratios in a gastric cancer case-control study. *Journal of Epidemiology*, **7**, 143 – 151.

4. Lasko, T. A., J. G. Bhagwat, K. H. Zou and L. Ohno-Machado, 2005. The Use of Receiver Operating Characteristic Curves in Biomedical Informatics. *Journal of Biomedical Informatics*, **38**, 404 – 415.

5. Nockemann, C., H. Heidt and N. Thomsen, 1991. Reliability in NDT: ROC study of radiographic weld inspections. *Nondestructive Testing and Evaluation International*, **24**, 235 – 245.

6. Somoza, E., D. Mossman and L. McFeeters, 1990. The info-ROC technique: a method for comparing and optimizing inspection systems. In *Review of Progress in quantitative Nondestructive Evaluation*, D. O. Thomson, D. E. Chimenti (eds.). Plenum Press, New York.

7. Swets, J. A. and R. M. Picket, 1982. Evaluation of diagnostic systems: methods from signal detection theory. Academic Press, New York.

8. DeLong, E. R., D. M. DeLong and D. L. Clarke-Pearson, 1988. Comparing the area under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, **44**, 837 – 844.

9. Zhou, X. H., N. A. Obuchowski and D. K. McClish, 2002. Statistical methods in diagnostic medicine. Wiley, New York.

10. Venkatraman, E. S. and C. B. Begg, 1996. A distribution-free procedure for comparing receiver operating charac-teristic curves from a paired experiment. Biometrika, 83, 835 – 848.

11. Bandos, A. I., H. E. Rockette and D. Gur, 2005. A permutation test sensitive to differences in areas for comparing ROC curves from a paired design. Statistics in Medicine, 24, 2873 – 2893.

12. Bamber, D., 1975. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, **12**, 387 – 415.

13. Hanley, J. A. and B. J. McNeil, 1982. The meaning and use of the Area under Receiver Operating Characteristic (ROC) Curve. *Radiology*, **143**, 29 – 36.

14. Hoeffding, W., 1948. A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, **19(3)**, 293 – 325.

15. Weiand, S., M. H. Gail, B. R. James and K. L. James, 1989. A Family of non-parametric statistics for comparing diagnostic markers with paired or unparied data. *Boimetrika*, **76**, 585 – 592.