

A STUDY ON PARAMETRIC APPROACHES TO COMPARE AREAS UNDER TWO CORRELATED ROC CURVES

Nahid Sultana Sumi* and Md. Akhtar Hossain

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

Abstract

Receiver operating characteristic (ROC) analysis is the commonly accepted method for comparing diagnostic markers. The overall accuracy of a diagnostic marker is summarized by the area under the ROC curve (AUC). One way to compare the accuracies of diagnostic markers is to compare their AUCs. This paper presents a study of the operating characteristics of the commonly used parametric methods of comparing AUCs of two correlated ROC curves. An extensive simulation study is conducted to explore and compare the operating conditions of the testing methodologies.

Key words: Diagnostic markers, receiver operating characteristic curve, area under curve, parametric tests

Introduction

A great use of diagnostic markers is made in medical studies based on clinical observations or laboratory methods to specify which individuals are classified as normal (nondiseased) or as abnormal (diseased). Throughout the medicine and public health studies such markers are important. Diagnostic medicine has progressed successfully in the last several decades and the trend promises to continue well into the next millennium. Technological improvements provide new methods for detecting disease or physical impairment. Diagnostic markers provide important medical decision making. Thus statistical methods of evaluating and comparing the performance of such markers are of great importance. The receiver operating characteristic (ROC) curve is a useful graphical and statistical tool for evaluating and comparing diagnostic markers. It has been used in many areas such as radiology (Metz 1989), psychiatry (Hsiao *et al.* 1989), epidemiology (Aoki *et al.* 1997) and biomedical informatics (Lasko *et al.* 2005). But the effectiveness of continuous diagnostic markers in distinguishing between diseased and nondiseased individuals using ROC curves has been increased recently (Greiner *et al.* 2000, Pepe 2003).

The ROC curve is a plot of the diagnostic marker's sensitivity versus 1-specificity at various observed values of the marker. The discriminatory power of a diagnostic marker

* Corresponding author. Email: sumi_sbi@du.ac.bd

measures the ability to correctly discriminate individuals or subjects into 'diseased/abnormal' and 'nondiseased/normal' subjects. To evaluate the discriminatory ability of a diagnostic marker, it is obvious to summarize the information of ROC curve into single global measure (Greiner *et al.* 2000). The most commonly used global measure for the overall accuracy of a diagnostic marker is summarized by the area under the ROC curve (AUC) because of its meaningful interpretation. AUC is interpreted as the probability that the observed value of the diagnostic marker will be greater for a randomly selected abnormal subject than for a randomly selected normal subject assuming that the higher values of a diagnostic marker are associated with 'abnormal' subjects, while lower values are associated with 'normal' (DeLong *et al.* 1988). Thus, AUC lies between 0 and 1 and the greater the AUC, the better the discriminatory power of the diagnostic marker (Zhou *et al.* 2002).

Since AUC is a measure of the overall performance of a diagnostic marker, the overall diagnostic performance of different markers can be compared by comparing their AUCs. By the comparison of areas under the two ROC curves we can estimate which one of two diagnostic markers is more suitable for discriminating nondiseased subjects from diseased subjects or any other two conditions of interest. When comparing the AUCs of two diagnostic markers, equal AUC values means that the two markers yield the same overall diagnostic performance, but does not necessarily mean that the two ROC curves of the two markers are identical. 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 markers, a paired study design is often implemented. This type of design usually induces positive correlation between the ratings of the same subjects. Several parametric and non-parametric methods have been suggested to compare the accuracy of two diagnostic markers within a paired design setting. The scope of this article is limited to a brief review and comparing performances of widely used parametric methods suggested by Hanley and McNeil (1983) and Wieand *et al.* (1989). Hanley and McNeil (1983) proposed a test statistic for comparing diagnostic markers from the same individual and Wieand *et al.* (1989) prescribed a parametric test to compare diagnostic markers assuming that the diagnostic marker values must follow the normal distribution. In the following sections these two widely used parametric procedures are discussed and their operating conditions are investigated through extensive simulation studies.

Comparison of AUCs

For two diagnostic markers, suppose there are N individuals without disease and M individuals with disease. Suppose (X_1, X_2) and (Y_1, Y_2) denote the diagnostic ratings of corresponding patients without disease and with disease for the two diagnostic markers

respectively. Corresponding bivariate outcomes should be (x_{1i}, x_{2i}) ($i = 1, 2, \dots, N$) and (y_{1j}, y_{2j}) ($j = 1, 2, \dots, M$), respectively for two diagnostic markers 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 $F_l(x_l)$, $G_l(y_l)$ ($l = 1, 2$). Bamber (1975) noted that the area under the ROC curve is equal to $\Pr(Y > X)$. Let AUC_l ($l = 1, 2$) be the areas under the respective ROC curves of diagnostic marker 1 and 2, respectively. Here we discuss the parametric tests of the hypothesis $H_0 : AUC_1 = AUC_2$ versus $H_1 : AUC_1 \neq AUC_2$.

Hanley and McNeil Approach

Hanley and McNeil (1983) developed a method of comparing the AUCs derived from the same samples. For the two observed areas under ROC curves AUC_1 and AUC_2 of diagnostic markers 1 and 2, their method of assessing the difference in the AUCs derived from the same set of patients is defined by the following standardized normal z - score

$$z = \frac{AUC_1 - AUC_2}{\sqrt{SE_1^2 + SE_2^2 - 2r SE_1 SE_2}}$$

where SE_1 and SE_2 refer to the estimated standard errors of the corresponding ROC areas. It is mentioned that using the correlation coefficient for the ratings of nondiseased subjects by the two diagnostic markers and the correlation coefficient for that from the diseased subjects, between area correlation r can be obtained assuming Gaussian distribution for the two diagnostic markers. That is, they first calculate the Pearson correlation coefficient between the two diagnostic markers, then combining this number and the size of the observed areas the between area correlation r is obtained through a complicated formula that we like to skip in this paper¹. Moreover, Hanley and McNeil (1983) provided a working table to find the between area correlation r for different combinations of average correlation between markers and average of observed areas. The quantity z is then referred to tables of the normal distribution and values of z above some cutoffs, for example, $z \geq 1.96$, are taken as evidence that the "true" ROC areas are different.

Wieand *et al.* Approach

Wieand *et al.* (1989) described a test to compare two diagnostic markers. The test that they suggested is based on the assumption that diagnostic markers 1 and 2 have bivariate normal distribution for both diseased and nondiseased individuals. That is,

¹ The mathematical derivation of this formula is available on request.

$X \sim N(\mu_x, \sigma_x^2)$ and $Y \sim N(\mu_y, \sigma_y^2)$. Let μ_{x_l} and $\sigma_{x_l}^2$ ($l=1,2$) represent the means and variances of the diagnostic marker for nondiseased individuals while μ_{y_l} and $\sigma_{y_l}^2$ ($l=1,2$) represent those for diseased individuals. Then AUC_l can be defined as $AUC_l = \Phi(\delta_l)$ with

$$\delta_l = \frac{\mu_{y_l} - \mu_{x_l}}{\sqrt{\sigma_{y_l}^2 + \sigma_{x_l}^2}}, \quad (l=1,2) \quad (1)$$

Note that Φ is the standard normal cumulative distribution function. The hypothesis $\Delta\delta = \delta_1 - \delta_2 = 0$ refers to the hypothesis $H_0 : AUC_1 = AUC_2$. The test statistic for testing this hypothesis is $T = \hat{\delta}_1 - \hat{\delta}_2 = \Delta\hat{\delta}$. Wieand *et al.* (1989) derived the asymptotic variance of $\Delta\hat{\delta}$ using delta method (Rao 1973) which takes into account the correlation between two diagnostic markers 1 & 2. The hypothesis is tested by considering the statistic $\frac{\Delta\hat{\delta}}{\sqrt{Var(\Delta\hat{\delta})}}$ as approximately a $N(0, 1)$ variate. The asymptotic variance of $\Delta\hat{\delta}$ is obtained using delta method (Rao 1973).

Simulation study

We have performed an extensive simulation to compute empirical test sizes and power of the parametric test approaches given by Hanley and McNeil (1983) and Wieand *et al.* (1989) for different underlying AUC differences, correlation between diagnostic markers and different sample sizes. We have considered the scenarios with non-crossing as well as crossing ROC curves. Two such scenarios of non-crossing and crossing ROC curves are shown in Figure 1. In our simulation study 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 and correlation ρ (0.25, 0.50, 0.75). That is $\mu_{x_l} = 0$ and $\sigma_{x_l}^2 = 1$, $l = 1, 2$.

So from (1) we have

$$\Phi^{-1}(AUC_l) = \delta_l = \frac{\mu_{y_l}}{\sqrt{1 + \sigma_{y_l}^2}}, \quad l = 1, 2 \quad (2)$$

where Φ^{-1} is the percentile of standard normal distribution. Similarly, 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 for non-crossing ROC curves; for crossing ROC curves $\sigma_{y_1}^2 = 1.0$ and $\sigma_{y_2}^2 = 4.0$ and correlation ρ (0.25,

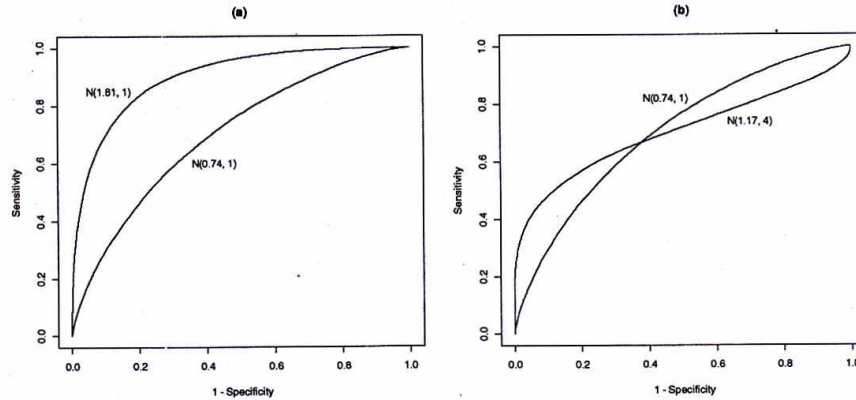


Figure 1. (a). Non-crossing ROC curves and (b). Crossing ROC curves.

0.50, 0.75) are assumed. Equal variances assumed to present ROC curves not crossing each other and unequal variances are assumed to attain the true situation where the ROC curves cross each other. The values of μ_y are directly determined from AUC_1 and AUC_2 from equation (2). We have performed simulations for different sample sizes such as 20, 40, 80 in all cases and the rejection region for the tests are determined using a nominal significance level of $\alpha=0.05$. For each scenario 1000 replications are computed and both empirical test size and empirical power are obtained. 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 repetitions². R: A language and environment for statistical computing version 2.15.0 (R Development Core Team 2012) is used to conduct the simulation work. Bivariate normal data are generated using function `rmvnorm` in R package `mvtnorm` (Alan Genz *et al.* 2012).

For both non-crossing and crossing ROC curves, given the value of AUCs and variances the mean values of diagnostic scores for diseased and nondiseased individuals can be obtained from relation (2) and the variance-covariance matrix can be constructed

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

² 100(1- α)% confidence interval of binomial proportion for large sample size is $\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, where

\hat{p} is the estimate of proportion of successes in a binomial trial, $z_{\frac{\alpha}{2}}$ is the $\frac{\alpha}{2}$ upper quantile of a standard normal distribution and n is the sample size.

Result and Discussion

Table 1 presents the empirical nominal sizes obtained for the two discussed methodologies when testing for equality of two non-crossing ROC curves for different combinations of correlation and sample sizes. Table 2 depicts the similar results for a null hypothesis of equality of two crossing ROC curves. For non-crossing ROC curves (Table 1) we see the method of Hanley and McNeil (1983) provides more stable estimate of empirical test size whereas for several setup of simulation the Wieand *et al.* (1989) approach gives test size estimate outside of the 95% confidence interval of considered nominal value 0.05 (the bold entries). Similarly in case of crossing ROC curves (Table 2) Wieand *et al.* (1989) approach is more conservative than Hanley and McNeil (1983) especially for high correlation among the diagnostic markers.

Table 1. Empirical test size for non-crossing ROC curves [HM - Hanley & McNeil test; W - Wieand *et al.* test].

Area		Mean		Variance		Sample Size		$\rho = 0.25$		$\rho = 0.50$		$\rho = 0.75$	
AUC_1	AUC_2	μ_{y_1}	μ_{y_2}	$\sigma_{y_1}^2$	$\sigma_{y_2}^2$	N	M	HM	W	HM	W	HM	W
0.70	0.70	0.74	0.74	1.00	1.00	20	20	0.044	0.048	0.042	0.049	0.054	0.058
						40	40	0.055	0.051	0.058	0.050	0.054	0.051
						80	80	0.055	0.049	0.047	0.053	0.048	0.046
0.75	0.75	0.95	0.95	1.00	1.00	20	20	0.048	0.055	0.051	0.054	0.051	0.056
						40	40	0.052	0.049	0.051	0.040	0.056	0.048
						80	80	0.050	0.048	0.060	0.064	0.059	0.061
0.80	0.80	1.19	1.19	1.00	1.00	20	20	0.053	0.055	0.049	0.056	0.050	0.037
						40	40	0.058	0.053	0.061	0.055	0.059	0.045
						80	80	0.052	0.049	0.058	0.057	0.055	0.058
0.85	0.85	1.47	1.47	1.00	1.00	20	20	0.049	0.047	0.045	0.040	0.045	0.041
						40	40	0.061	0.051	0.059	0.045	0.063	0.046
						80	80	0.053	0.054	0.055	0.053	0.053	0.051
0.90	0.90	1.81	1.81	1.00	1.00	20	20	0.046	0.034	0.045	0.030	0.044	0.028
						40	40	0.061	0.047	0.058	0.047	0.055	0.042
						80	80	0.051	0.054	0.052	0.046	0.049	0.046
0.95	0.95	2.33	2.33	1.00	1.00	20	20	0.051	0.017	0.051	0.015	0.051	0.012
						40	40	0.057	0.044	0.057	0.038	0.055	0.033
						80	80	0.049	0.047	0.054	0.045	0.053	0.038

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 is false. Table 3 and Table 4 demonstrate statistical power of the test procedures for testing equality of two non-crossing and two crossing ROC curves, respectively.

Table 2. Empirical test size for crossing ROC curves [HM - Hanley & McNeil test; W - Wieand *et al.* test].

Area		Mean		Variance		Sample Size		$\rho = 0.25$		$\rho = 0.50$		$\rho = 0.75$	
AUC_1	AUC_2	μ_{y_1}	μ_{y_2}	$\sigma_{y_1}^2$	$\sigma_{y_2}^2$	N	M	HM	W	HM	W	HM	W
0.70	0.70	0.74	1.17	1.00	4.00	20	20	0.052	0.057	0.054	0.062	0.055	0.081
						40	40	0.056	0.052	0.055	0.060	0.063	0.086
						80	80	0.060	0.067	0.046	0.062	0.047	0.076
0.75	0.75	0.95	1.51	1.00	4.00	20	20	0.058	0.068	0.062	0.077	0.057	0.078
						40	40	0.058	0.052	0.050	0.057	0.059	0.082
						80	80	0.052	0.060	0.041	0.066	0.047	0.073
0.80	0.80	1.19	1.88	1.00	4.00	20	20	0.065	0.063	0.058	0.069	0.061	0.079
						40	40	0.058	0.051	0.062	0.066	0.064	0.081
						80	80	0.053	0.057	0.042	0.057	0.043	0.078
0.85	0.85	1.47	2.32	1.00	4.00	20	20	0.057	0.057	0.057	0.059	0.058	0.071
						40	40	0.060	0.050	0.058	0.061	0.058	0.066
						80	80	0.054	0.058	0.046	0.061	0.047	0.071
0.90	0.90	1.81	2.87	1.00	4.00	20	20	0.059	0.043	0.058	0.040	0.059	0.046
						40	40	0.058	0.050	0.058	0.057	0.060	0.061
						80	80	0.056	0.060	0.048	0.059	0.052	0.076
0.95	0.95	2.33	3.68	1.00	4.00	20	20	0.058	0.027	0.064	0.027	0.070	0.027
						40	40	0.058	0.039	0.060	0.043	0.060	0.048
						80	80	0.056	0.048	0.052	0.060	0.051	0.072

For non-crossing ROC curves (Table 3) the test of Hanley and McNeil (1983) shows uniformly better performance in terms of statistical power than the Wieand *et al.* (1989) approach for all simulation setup. Both the approaches exhibits lower power for smaller differences in AUCs and for smaller sample sizes and with increasing sample sizes the power also increases. When two crossing ROC curves are considered to compare (Table 4) our simulation study shows that though for AUCs ≤ 0.85 and small differences between AUCs the Wieand *et al.* (1989) approach has slightly higher power than Hanley and McNeil (1983) approach, in all other situations the later dominates the former in terms of statistical power.

Table 3. Empirical test power for non-crossing ROC curves [HM - Hanley & McNeil test; W - Wieand *et al.* Test].

Area		Mean		Variance		Sample Size		$\rho = 0.25$		$\rho = 0.50$		$\rho = 0.75$	
AUC_1	AUC_2	μ_{y_1}	μ_{y_2}	$\sigma_{y_1}^2$	$\sigma_{y_2}^2$	N	M	HM	W	HM	W	HM	W
0.70	0.75	0.74	0.95	1.00	1.00	20	20	0.081	0.078	0.101	0.090	0.148	0.132
						40	40	0.122	0.127	0.142	0.141	0.231	0.222
						80	80	0.163	0.153	0.254	0.217	0.424	0.392
0.70	0.80	0.74	1.19	1.00	1.00	20	20	0.207	0.191	0.266	0.238	0.417	0.377
						40	40	0.326	0.317	0.449	0.427	0.735	0.681
						80	80	0.540	0.517	0.737	0.704	0.949	0.916
0.70	0.85	0.74	1.47	1.00	1.00	20	20	0.396	0.374	0.509	0.476	0.793	0.710
						40	40	0.675	0.652	0.842	0.805	0.974	0.957
						80	80	0.926	0.912	0.985	0.981	1.000	1.000
0.70	0.90	0.74	1.81	1.00	1.00	20	20	0.692	0.644	0.835	0.766	0.982	0.949
						40	40	0.939	0.920	0.982	0.976	1.000	1.000
						80	80	0.998	0.999	1.000	1.000	1.000	1.000
0.70	0.95	0.74	2.33	1.00	1.00	20	20	0.935	0.903	0.987	0.959	1.000	0.995
						40	40	0.999	0.998	1.000	1.000	1.000	1.000
						80	80	1.000	1.000	1.000	1.000	1.000	1.000
0.75	0.80	0.95	1.19	1.00	1.00	20	20	0.094	0.088	0.111	0.099	0.171	0.132
						40	40	0.124	0.123	0.157	0.147	0.259	0.240
						80	80	0.192	0.178	0.278	0.250	0.471	0.434
0.75	0.85	0.95	1.47	1.00	1.00	20	20	0.242	0.219	0.296	0.279	0.487	0.425
						40	40	0.386	0.364	0.521	0.489	0.801	0.758
						80	80	0.641	0.628	0.826	0.791	0.976	0.960
0.75	0.90	0.95	1.81	1.00	1.00	20	20	0.480	0.449	0.629	0.547	0.876	0.798
						40	40	0.793	0.768	0.910	0.884	0.993	0.986
						80	80	0.967	0.964	0.995	0.994	1.000	1.000
0.75	0.95	0.95	2.33	1.00	1.00	20	20	0.841	0.779	0.936	0.876	0.995	0.968
						40	40	0.986	0.977	1.000	0.996	1.000	1.000
						80	80	1.000	1.000	1.000	1.000	1.000	1.000
0.80	0.85	1.19	1.47	1.00	1.00	20	20	0.103	0.091	0.118	0.096	0.194	0.145
						40	40	0.155	0.132	0.190	0.168	0.307	0.257
						80	80	0.225	0.215	0.329	0.313	0.554	0.508
0.80	0.90	1.19	1.81	1.00	1.00	20	20	0.293	0.244	0.358	0.296	0.600	0.497
						40	40	0.494	0.455	0.635	0.595	0.892	0.842
						80	80	0.771	0.748	0.895	0.887	0.994	0.989
0.80	0.95	1.19	2.33	1.00	1.00	20	20	0.671	0.568	0.809	0.690	0.965	0.870
						40	40	0.929	0.897	0.977	0.959	1.000	0.998
						80	80	0.998	0.996	1.000	1.000	1.000	1.000
0.85	0.90	1.47	1.81	1.00	1.00	20	20	0.122	0.094	0.149	0.104	0.237	0.157
						40	40	0.188	0.160	0.231	0.198	0.406	0.327
						80	80	0.319	0.283	0.429	0.398	0.688	0.639
0.85	0.95	1.47	2.33	1.00	1.00	20	20	0.414	0.308	0.537	0.379	0.806	0.588
						40	40	0.711	0.651	0.847	0.782	0.978	0.941
						80	80	0.938	0.911	0.984	0.973	1.000	1.000
0.90	0.95	1.81	2.33	1.00	1.00	20	20	0.187	0.108	0.238	0.122	0.362	0.181
						40	40	0.298	0.229	0.397	0.299	0.629	0.497
						80	80	0.507	0.472	0.687	0.610	0.896	0.841

Table 4. Empirical test power for crossing ROC curves [*HM* - Hanley & McNeil test; *W* - Wieand *et al.* test].

Area		Mean		Variance		Sample Size		$\rho = 0.25$		$\rho = 0.50$		$\rho = 0.75$	
AUC_1	AUC_2	μ_{y_1}	μ_{y_2}	$\sigma_{y_1}^2$	$\sigma_{y_2}^2$	<i>N</i>	<i>M</i>	<i>HM</i>	<i>W</i>	<i>HM</i>	<i>W</i>	<i>HM</i>	<i>W</i>
0.70	0.75	0.74	1.51	1.00	4.00	20	20	0.095	0.094	0.100	0.119	0.119	0.140
						40	40	0.120	0.135	0.129	0.149	0.141	0.189
						80	80	0.165	0.167	0.210	0.211	0.241	0.281
0.70	0.80	0.74	1.88	1.00	4.00	20	20	0.185	0.191	0.215	0.232	0.252	0.293
						40	40	0.296	0.302	0.350	0.364	0.410	0.457
						80	80	0.502	0.495	0.594	0.604	0.659	0.723
0.70	0.85	0.74	2.32	1.00	4.00	20	20	0.380	0.383	0.420	0.429	0.500	0.518
						40	40	0.619	0.605	0.688	0.694	0.787	0.797
						80	80	0.891	0.872	0.935	0.930	0.974	0.979
0.70	0.90	0.74	2.87	1.00	4.00	20	20	0.637	0.611	0.688	0.682	0.782	0.784
						40	40	0.911	0.896	0.940	0.927	0.968	0.968
						80	80	0.998	0.994	0.998	0.997	1.000	1.000
0.70	0.95	0.74	3.68	1.00	4.00	20	20	0.895	0.878	0.926	0.909	0.961	0.955
						40	40	0.994	0.988	1.000	0.996	1.000	1.000
						80	80	1.000	1.000	1.000	1.000	1.000	1.000
0.75	0.80	0.95	1.88	1.00	4.00	20	20	0.099	0.107	0.106	0.117	0.122	0.145
						40	40	0.116	0.129	0.129	0.153	0.153	0.195
						80	80	0.188	0.183	0.212	0.217	0.253	0.294
0.75	0.85	0.95	2.32	1.00	4.00	20	20	0.219	0.217	0.243	0.264	0.289	0.316
						40	40	0.346	0.342	0.393	0.408	0.472	0.513
						80	80	0.589	0.578	0.683	0.682	0.756	0.799
0.75	0.90	0.95	2.87	1.00	4.00	20	20	0.460	0.450	0.510	0.493	0.584	0.591
						40	40	0.731	0.713	0.795	0.786	0.873	0.878
						80	80	0.946	0.939	0.977	0.971	0.996	0.989
0.75	0.95	0.95	3.68	1.00	4.00	20	20	0.779	0.725	0.841	0.797	0.896	0.880
						40	40	0.972	0.962	0.985	0.974	0.995	0.992
						80	80	1.000	1.000	1.000	1.000	1.000	1.000
0.80	0.85	1.19	2.32	1.00	4.00	20	20	0.106	0.108	0.109	0.129	0.125	0.152
						40	40	0.134	0.135	0.148	0.163	0.180	0.209
						80	80	0.224	0.209	0.247	0.258	0.303	0.343
0.80	0.90	1.19	2.87	1.00	4.00	20	20	0.257	0.245	0.295	0.292	0.347	0.369
						40	40	0.433	0.419	0.498	0.504	0.575	0.606
						80	80	0.705	0.701	0.790	0.786	0.871	0.876
0.80	0.95	1.19	3.68	1.00	4.00	20	20	0.599	0.553	0.666	0.604	0.749	0.706
						40	40	0.889	0.859	0.924	0.907	0.955	0.947
						80	80	0.993	0.990	0.997	0.995	0.998	0.998
0.85	0.90	1.47	2.87	1.00	4.00	20	20	0.122	0.109	0.134	0.117	0.147	0.153
						40	40	0.171	0.171	0.191	0.199	0.222	0.244
						80	80	0.284	0.274	0.327	0.337	0.393	0.439
0.85	0.95	1.47	3.68	1.00	4.00	20	20	0.390	0.336	0.427	0.368	0.500	0.443
						40	40	0.631	0.613	0.699	0.682	0.780	0.768
						80	80	0.907	0.874	0.936	0.921	0.969	0.968
0.90	0.95	1.81	3.68	1.00	4.00	20	20	0.167	0.123	0.189	0.136	0.218	0.166
						40	40	0.262	0.239	0.285	0.268	0.356	0.345
						80	80	0.469	0.429	0.532	0.504	0.608	0.615

Conclusion

The area under ROC curve (AUC) is recognized as a global measure of accuracy of a diagnostic marker and consequently comparing AUCs of two diagnostic markers is of practical importance to comment about discriminatory power of them. In this article two very commonly used and competing parametric methodologies to compare AUCs of two correlated diagnostic markers namely, Hanley and McNeil (1983) approach and Wieand *et al.* (1989) approach, are discussed. The operating characteristics of the methods are explored and compared through extensive simulation. The simulation study depicts that though the Hanley and McNeil (1983) approach is little complex and tedious to apply, it outperforms the Wieand *et al.* (1989) approach except some very restrictive situations in terms of empirical test size and statistical power. The findings of this study will be of help for researchers to select between the competing methods that this article has discussed.

References

- Alan Genz, Frank Bretz, Tetsuhisa Miwa, Xuefei Mi, Friedrich Leisch, Fabian Scheipl, Torsten Hothorn. 2012. mvtnorm: Multivariate Normal and t Distributions. R package version 0.9-9992.
- Aoki, K., Misumi, J., Kimura, T., Zhao, W. and Xie, T. 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. *J. Epidemiology* 7: 143 - 151.
- 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.
- DeLong, E. R., DeLong, D. M. and Clarke-Pearson, D. L. 1988. Comparing the area under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 44: 837 - 844.
- Hanley, J. A. and McNeil, B. J. 1983. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* 148: 839 - 843.
- Greiner, M. J., Pfeiffer, D. and Smith, R. M. 2000. Principles and Practical Application of the Receiver-operating Characteristic Analysis for Diagnostic Tests. *Preventive Veterinary Medicine* 45: 23 - 41.
- Hsiao, J. K., Barko, J. J. and Potter, W. Z. 1989. Diagnosing diagnoses: receiver operating characteristic methods and psychiatry. *Archives of General Psychiatry* 46: 664 - 667.
- Lasko, T. A., Bhagwat, J. G., Zou, K. H. and Ohno-Machado, L. 2005. The Use of Receiver Operating Characteristic Curves in Biomedical Informatics. *Journal of Biomedical Informatics* 13 38: 404 - 415.
- Metz, C. E. 1989. Some practical issues of experimental design and data analysis in radiological ROC studies. *Investigation Radiology* 24: 234 - 245.

- Pepe, M. S. 2003. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford Statistical Science Series.
- R Development Core Team. 2012. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- Rao, C. R. 1973. *Linear Statistical Inference and its Applications*, 2nd ed. Wiley: New York.
- Weiland, S., Gail M. H., James, B. R. and James, K. L. 1989. A Family of Non-parametric Statistics for Comparing Diagnostic Markers with Paired or Unpaired Data. *Biometrika* 76: 585 - 592.
- Zhou, X. H., Obuchowski, N. A. and McClish, D. K. 2002. *Statistical Methods in Diagnostic Medicine*. Wiley: New York.

(Manuscript received on 13 June, 2012; revised on 16 July 2012)