BULLETIN *of the*
MALAYSIAN MATHEMATICAL
SCIENCES SOCIETY
http://math.usm.my/bulletin

Bull. Malays. Math. Sci. Soc. (2) **37**(2) *(2014), 411–423*

Evaluation and Computation of Diagnostic Tests: A Simple Alternative

[1]NAHID SULTANA SUMI, [2]M. ATAHARUL ISLAM AND [3]MD. AKHTAR HOSSAIN
[1,3] Department of Statistics, Biostatistics & Informatics, University of Dhaka,
Dhaka 1000, Bangladesh
[2]Department of Applied Statistics, East West University Jahurul Islam City, Aftabnagar, Dhaka-1212, Bangladesh
[1]sumi_sbi@du.ac.bd, [2]mataharul@yahoo.com, [3]akhtar_sbi@du.ac.bd

Abstract. Methods of evaluating and comparing the performance of diagnostic tests are of increasing importance in medical science. When a test is based on an observed variable that lies on a continuous scale, an assessment of the overall value of the test can be made through the use of a Receiver Operating Characteristic (ROC) curve. The ROC curve descibes the discrimination ability of a diagnosis test for the diseased subjects from the non-diseased subjects. The area under the ROC curve (AUC) represents the probability that a randomly chosen diseased subject will have higher probability of having disease than a randomly chosen non-diseased subject. For comparing two diagnostic systems, the difference between AUCs is often used. In this paper we have investigated various methods of the comparison of equality of two AUCs and proposed a McNemar test for the comparison of two diagnostic test procedures. The proposed test is based on an optimal cut-off point that discriminates the individuals in actually positive or actually negative cases for which we have a 2×2 contingency table where we can apply the McNemar test. The operating characteristics of the proposed test are evaluated using extensive simulation over a wide range of parameters.

2010 Mathematics Subject Classification: 62P10, 92-08, 62-07

Keywords and phrases: Diagnostic test, receiver operating characteristic (ROC) curve, area under the ROC curve (AUC), McNemar test.

1. Introduction

The Receiver Operating Characteristic (ROC) curve is an effective method of evaluating the quality or performance of diagnostic tests. The ROC curve is a plot of the sensitivity (or, true positive rate) of a test (Y-axis) vs. false positive rate (or, 1-specificity) of this test (X-axis) for all possible cut-off points. The ROC plot provides a comprehensive picture of the ability of a test to make the distinction between diseased and non-diseased individuals being examined over all decision thresholds. The accuracy of the diagnostic test is measured by the area under the ROC curve (AUC). The area under an empirical ROC curve can be computed by trapezoidal rule [1]. The area computed by trapezoidal rule under an empirical ROC curve is equal to the Mann-Whitney statistic for comparing distributions of values from the two samples [6].

Communicated by V. Ravichandran.
Received: May 18, 2012; *Revised:* December 5, 2012.

There are several ways to calculate the area under a ROC curve. First, the trapezoidal rule can be used but gives an underestimation of the area for continuous data [20]. Second, it is possible to get a better approximation of the curve by fitting the data to a binormal model with maximum likelihood estimates [4, 5]. After that it is possible to get a better estimate of the area. A third way to calculate the area is to use the Mann-Whitney U statistic which is also known as the non-parametric Wilcoxon statistic. That is, no assumptions on the distributions of the data are done since Wilcoxon is a distribution free statistic [1, 6].

There are several ways to express the reader's confidence in the presence of a disease such as a binary result which is either positive or negative for the disease, a discrete rating scale such as five or six-point scale, and a continuous scale such as a percent-confidence scale from 0 to 100 percent. In most of the ROC analyses of radiological tests, a discrete rating scale with five or six categories has been used. A variety of tests has been suggested by many authors to test the equality of the two ROC curves and also the equality of AUCs. We discuss some of them in details in this paper.

Since AUC is the global measure of accuracy, many permutation tests are developed for comparing AUCs. We here give a review of Bandos *et al.*'s permutation test for comparing ROCs on the basis of AUCs for continuous scale data [2]. DeLong *et al.* gave a conventional nonparametric test for comparing AUCs for continuous scale data [3]. Because using a continuous scale is desirable theoretically, Wagner *et al.* [18], here we propose an alternative based on McNemar test [12] for the comparison of two diagnostic tests based on continuous scale data as well as for discrete binary data. We considered a matched-pairs design where a single response variable for each subject is observed in a matched pair. The data structure is such that we have recoded each subject's rating as 'positive' or 'negative' on each of two diagnostic procedures and our interest is in testing whether the proportions of 'positive' responses are the same on the first and second procedure with account taken of the correlation of the bivariate ratings. And then we compare the proposed method with the conventional nonparametric test suggested by DeLong *et al.* [3] and permutation test by Bandos *et al.* [2].

2. Estimation of AUC

Suppose X and Y denote the patients without disease and with disease, respectively. Let z be a threshold. Then $P(X > z) = G(z)$ and $P(Y > z) = F(z)$ where $F(z)$ is nothing but sensitivity and $1 - G(z)$ represents specificity. So by these functions we can represent ROC curve, that is, ROC curve is a plot of $F(z)$ versus $G(z)$ for all possible thresholds, z. The area under the ROC curve (AUC) is defined as the probability that a randomly chosen diseased subject will have higher probability of having disease than a randomly chosen non-diseased subject. Probabilistically, $AUC = P(Y > X)$, where $AUC =$ area under the ROC curve, X be the test result of patients without disease and Y be the test result of patients with disease. For discrete case, $AUC = P(Y > X) + \frac{1}{2}P(X = Y)$. In continuous case, $P(X = Y) = 0$. The area under an empirical ROC curve can be computed by trapezoidal rule [1]. Hanley and McNeil [6] 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 [6] suggested for computing the area

under the ROC curve is given as,

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

where

$$\begin{aligned} A &= \text{Area under the ROC curve} \\ M &= \text{Number of 'abnormal' or 'diseased' subjects} \\ N &= \text{Number of 'normal' or 'non-diseased' subjects} \\ Y_j &= \text{The test score of } j\text{th patient with disease} \\ X_i &= \text{The test score of the } i\text{th patient without disease} \end{aligned}$$

g is a function comparing X_i with Y_j such that

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

So for the mth modality or diagnostic test the area under ROC curve can be computed as,

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

3. Different methods of comparing diagnostic tests

Comparative assessment of diagnostic tests is now an interesting field in Medical Science. A lot of researches have been developed to compare diagnostic tests or accuracy of diagnostic test procedures utilizing concepts of ROC curves and AUCs. In this study, we have discussed the very commonly used methods of comparing diagnostic test procedures and proposed an alternative procedure to compare two diagnostic test procedures based on McNemar test [12]. The suggested procedure is very easy to understand, to apply and to interpret, where applicable.

3.1. Conventional test by DeLong *et al.*

For comparing two diagnostic systems, the difference between *AUCs* is often used. The best known nonparametric procedure for correlated *AUCs* is given by DeLong *et al.* [3]. They have pointed out that when ROC curves are derived from tests performed on the same individuals, statistical analysis must take into account the correlated nature of the data. They developed a totally nonparametric approach to solve this problem by using the theory of generalized U statistics. They utilized the method of structural components provided by Sen [15] to generate consistent estimates of the elements of the variance-covariance matrix of a vector of U statistics, and the resulting test statistic has asymptotically a χ^2 distribution.

3.2. Permutation test by Bandos *et al.*

Bandos *et al.* [2] described exact and asymptotic permutation tests procedures to test the equality of two correlated ROC curves which is designed to have increased power to detect differences in the *AUCs*. For a paired design, the difference in two *AUCs* is estimated using the appropriately transformed ratings or the ranks of the ratings for actually negative and actually positive subjects and a permutation test is suggested to test statistical significance of observed difference.

The permutation tests are based on exchangeability. Exchangeability means that the joint distribution of the rank-ratings is symmetric with respect to its arguments. They considered null hypothesis of equality of ROC curve under the assumption of exchangeability. The distribution of the differences in the estimated *AUCs* over all permutations is obtained and the rejection region is identified based on percentile values for selected nominal level.

As a member of U statistics the non-parametric estimator of the *AUC* difference is known to be asymptotically normally distributed under quite general conditions [7]. Under the assumption of asymptotic normality of the U statistic and the additional assumption of exchangeability, they also constructed a simple asymptotic test procedure. In a simulation study, they showed that with small samples there is a good agreement between the exact and asymptotic test.

3.3. Proposed test based on McNemar test

Comparative diagnostic medicine studies commonly produce matched data since, frequently, all the examinations under comparison are performed on each subject. The McNemar test pertains to matched pairs of dichotomous test results. The results of each diagnostic test fall into two categories, positive and negative. The data are succinctly presented in a two-by-two array with the rows corresponding to the results of one diagnostic test and the columns to the results of the other; each element of the array is the number of observed cases with the particular combination of test results.

Diagnostic test procedure commonly demands 'Yes' or 'No' decisions and some diagnostic test procedure provide such dichotomous outcomes. To compare such two test procedures, we can have a 2×2 contingency table with ease and can use McNemar test. But all diagnostic procedures are not truly dichotomous and one, who wants to perform a comparison of two such test procedures, needs to convert continuous diagnostic ratings into dichotomous test results such as positive or negative. We convert each continuous diagnostic test into a dichotomous test based on an optimal cut-off point and can have a 2×2 contingency table and the proposed McNemar test can be applied. The ultimate choice of cut-off depends on the nature of the diagnostic test or study. For a practical dataset, we must know the cut-off to discriminate any two conditions of a diagnostic test procedure. For example, for a patient to be diabetic it is necessary to have sugar level above 11.5. So here cut-off value is 11.5 and a 2×2 contingency table can be obtained to apply McNemar test for the comparison of two diagnostic test procedures to see whether these diagnostic procedures differ or not.

For two diagnostic tests producing continuous ratings $\{X_i^m\}_{i=1}^N$, $\{Y_j^m\}_{j=1}^M$ in the mth modality for N actually negative and M actually positive subjects, we have to order the subjects so that $\{X_i^m\}_{i=1}^N$, $\{Y_j^m\}_{j=1}^M$ be the transformed ratings or results in the mth modality for N actually negative and M actually positive subjects. Suppose we have an optimal

cut-off point at z_m for mth modality. Then all the results greater than z_m are considered as positive and less than or equal to z_m are considered as negative. Classifying the subjects using optimal cut-off points, a 2×2 contingency table can be obtained for each diagnostic procedure.

Table 1. A 2×2 contingency table for mth $(m = 1, 2)$ diagnostic test procedure.

Test result for diagnostic procedure m	Observed (true) status		Row sums
	Positive	Negative	
Positive	a_m	b_m	$a_m + b_m$
Negative	c_m	d_m	$c_m + d_m$
Column sums	$a_m + c_m$	$b_m + d_m$	n_m

The number in the first cell of Table 1, a_m, is number of actually positive subjects with positive test results $(y_j^m > z_m)$. Similarly, b_m is number of actually negative subjects with positive test results $(x_i^m > z_m)$, c_m is number of actually positive subjects with negative test results $(y_j^m \leq z_m)$, d_m is number of actually negative subjects with negative test results $(x_i^m \leq z_m)$. Now for each diagnostic test as we obtain a 2×2 contingency table at the optimal cut-off point, we can see whether the diagnostic test procedure has any effect on the true disease (observed) status. If both diagnostic test procedures have significant effects, we can combine the two diagnostic test procedures. We will obtain a matched pair data from this combination of two diagnostic tests and subsequently a contingency Table 2.

Table 2. A 2×2 contingency table for two diagnostic test procedures.

Diagnostic test procedure 1	Diagnostic test procedure 2		Row sums
	Positive	Negative	
Positive	a	b	$a+b$
	(P_a)	(P_b)	(P_{a+b})
Negative	c	d	$c+d$
	(P_c)	(P_d)	(P_{c+d})
Column sums	$a+c$	$b+d$	n
	(P_{a+c})	(P_{b+d})	(1)

In Table 2 the frequency a represents positive test results on both test procedures; b represents positive test results on test procedure 1 but negative test results on the test procedure 2; c represents negative test results on test procedure 1 but positive test results on the test procedure 2; and finally d represents negative test results on both test procedures. Corresponding cell probabilities are shown in parentheses. The hypothesis of interest is whether the two diagnostic test procedures are different or not in their performances. An equivalent statement of the null hypothesis is that the marginal probabilities of positive result on the first and second procedures are equal and the alternative hypothesis is that the equality does not hold, that is,

$$H_0 : P_{a+b} = P_{a+c} \quad \text{versus} \quad H_1 : P_{a+b} \neq P_{a+c}$$

or equivalently,

$$H_0 : P_a + P_b = P_a + P_c \quad \text{versus} \quad H_1 : P_a + P_b \neq P_a + P_c$$

these competing hypotheses reduce to,

$$H_0 : P_b = P_c \quad \text{versus} \quad H_1 : P_b \neq P_c$$

Thus, the test of homogeneity of performances of two diagnostic procedures is also a test of symmetry investigating differences in types of discordance, {positive, negative} and {negative, positive} in Table 2.

The McNemar test statistic follows a chi-square distribution with one degree of freedom and has the form,

$$\chi^2 = \frac{(b-c)^2}{b+c} \qquad \text{(without continuity correction)}$$

$$\chi^2 = \frac{(|b-c|-1)^2}{b+c} \qquad \text{(with continuity correction)}$$

Since the McNemar test employs a continuous distribution to approximate a discrete probability distribution, some sources recommended that a correction for continuity be employed in computing the test statistic. When the sample size is small, in the interest of accuracy, the exact binomial probability for the data should be used.

Despite several advantages of being simple to calculate, easy to understand and readily applicable, the proposed test lacks the quality of providing evidence of inferiority or superiority of one diagnostic procedure over the other. The assessment of inferiority or superiority of a diagnostic procedure in reference to another can be a desirable issue in many medical researches. However, they are not immediately obtainable because the accuracy of diagnostic procedures is measured in both sensitivity and specificity simultaneously [16] which consequently requires the knowledge of true disease status of subjects [10]. The alternative hypothesis to be tested will be then one sided [19]. The problem arises when the true disease status or any other gold standard is not known. In many practical instances, the true status (gold standard) may be either unknown or difficult to use. For example, it takes a large cohort and a long follow-up period to determine the hip fracture status in osteoporosis studies.

The proposed test concludes about the equivalence of two diagnostic procedures from testing a two sided alternative hypothesis. The idea is very simple to compare only the discordance of two diagnostic procedures and does not require any knowledge of true disease status. Thus, it tests for equivalence, not for inferiority or superiority. This issue also explores a very advantageous robust feature of the proposed test that it is invariably applicable in all the instances irrespective of having knowledge of true status (gold standard) or not whereas other traditional test methods [2, 3] based on AUC estimates cannot be used without knowing true status (gold standard). The extension of the proposed approach for one sided comparison of inferiority or superiority needs further investigation.

4. Finding optimal cut-off point

For the use of McNemar test one crucial consideration is to define optimal cut-off point. It is a very important point to consider the optimal cut-off point. For a clinical application it is important to decide, which cut-off point should be used to discriminate diseased and non-diseased subjects. Though in most practical applications yielding continuous ratings

well known cut-off point exist, we are discussing a method to select optimal cut-off point (if unknown) based on Youden Index under normality assumption [14].

Assume that the continuous ratings of a specific diagnostic procedure are normally distributed, such that the non-diseased subjects (X), or true negatives, have mean μ_x and variance σ_x^2, and the diseased subjects (Y), or true positives, have mean μ_y and variance σ_y^2, and $\mu_y > \mu_x$. Under these assumptions, sensitivity $(q(z))$ and specificity $(p(z))$ can be written as,

$$q(z) = P(X \geq z) = \Phi(\frac{\mu_x - z}{\sigma_x})$$
$$p(z) = P(Y \leq z) = \Phi(\frac{z - \mu_y}{\sigma_y})$$

for a given cut-point z, where Φ denotes the standard normal distribution function. Accordingly, test measurements falling at or below z are negative results and those above z are positive. The Youden Index (J) is defined as $\max\{q(z) + p(z) - 1\}$ for all z.

The optimal cut-point occurs at an intersection of the probability density functions of non-diseased and diseased subjects. The number of intersections is a function of the variances of the non-diseased and diseased subjects. One simple case is that of equal variance in non-diseased and diseased subjects, $\sigma_x^2 = \sigma_y^2$, where only one intersection exists and z is simply the midpoint between means, $\frac{(\mu_y - \mu_x)}{2}$. In the case of unequal variance, the intersections can be found by the following quadratic equation

$$z_{1,2} = \frac{\mu_y(v^2-1) - u \pm v\sqrt{u^2 + (v^2-1)\sigma_y^2 ln(v^2)}}{v^2 - 1}$$

where $u = \mu_y - \mu_x$ and $v = \frac{\sigma_y}{\sigma_x}$. Let us first order the intersections, $z_1 < z_2$. If $v > 1$, then J occurs at z_2; alternatively, if $v < 1$, then J occurs at z_1. When data on both non-diseased and diseased subjects are available, appropriate estimates for μ_x, σ_x^2, μ_y, σ_y^2 and consequently the optimal cut-off point can be calculated. Methods to find optimal cut-off point in instances where the continuous diagnostic ratings comes from non-normal distributions are also available in literature [14].

5. Simulation study

We have performed extensive computer simulations to investigate and compare type I error and statistical power of the proposed McNemar test, conventional nonparametric test of DeLong *et al.* [3] and asymptotic test of Bandos *et al.* [2]. In our simulations we assumed equal correlation across the test procedures for the ratings of diseased and non-diseased subject rated on both continuous and binary scales.

For continuous scale, the rating values of the non-diseased subject were generated from a standard bivariate normal distribution and those of diseased subjects from bivariate normal distribution with mean and variance, μ_1, σ_1^2 and μ_2, σ_2^2 for the two test procedures 1 and 2, respectively. Data are generated for a set of correlation (ρ) between the ratings from two diagnostic tests. The areas under the ROC curve A_1 and A_2 are given by $\Phi\left(\frac{\mu_1}{(1+\sigma_1^2)^{\frac{1}{2}}}\right)$ and $\Phi\left(\frac{\mu_2}{(1+\sigma_2^2)^{\frac{1}{2}}}\right)$, where Φ is the standard normal cumulative distribution function.

If the binary variable X denotes the outcomes of a diagnostic test (1 for positive and 0 for negative), the distribution of X is fully determined by the single value $p_X = p(X = 1)$, which is also the expectation of X. The variance of X is $\text{var}(X) = p_X q_X$ where $q_X = 1 - p_X = p(X = 0)$. For two competing diagnostic tests employed on same set of subjects, we can assume two such binary variables (X, Y) which are not necessarily independent. Then the joint distribution of X and Y is determined by p_X, p_Y and either $p_{XY}, p_{X|Y}$ or $p_{Y|X}$ where $p_{XY} = p(X = 1, Y = 1)$, $p_{X|Y} = p(X = 1|Y = 1)$, $p_{Y|X} = p(Y = 1|X = 1)$ and the correlation coefficient of X and Y can be expressed as,

$$\rho = \frac{p_{XY} - p_X p_Y}{\sqrt{p_X q_X p_Y q_Y}} \quad \text{such that} \quad p_{XY} = \rho\sqrt{p_X q_X p_Y q_Y} + p_X p_Y$$

where $-1 \leq \rho \leq +1$ and $\rho = 0$ implies that X and Y are independent. For more details see [11]. For binary scale, correlated binary ratings were generated with required marginal probabilities (P_{a+b}, P_{a+c}) to obtain specific difference $(P_b - P_c)$ between the probabilities of discordant matches in Table 2. For non-diseased subjects, the binary ratings are generated with fixed marginal probabilities $(0.30, 0.35)$. The algorithm discussed in Leisch *et al.* [9] is used for simulating correlated binary data. The same algorithm is also used by Islam *et al.* [8] in their recent work of bivariate binary model for testing dependence in outcomes. The statistical computing software R version 2.11.1 [13] is used to perform the simulation study.

For continuous ratings simulated values of ρ ranged from 0 to 0.6 and the values of other parameters were selected to produce the difference between two *AUCs* in the range from 0 to 0.3. For binary ratings, ρ was considered in the range 0 to 0.5 and the marginal probabilities were selected to produce difference between probabilities of discordant matches in the range 0 to 0.2. Because of difficulties in generating correlated binary data for specific marginal probabilities, we had to choose some selected parameter combinations. For each considered scenario, 2000 replications were used in the computer simulations. Table 3 compares the type I error and the statistical power of the proposed McNemar test to that of conventional nonparametric *AUC* test developed by DeLong *et al.* [3] and to asymptotic permutation *AUC* test developed by Bandos *et al.* [2] for continuous type ratings. Table 4 compares the same for binary type ratings. The estimates of type I error (the gray shaded entries in Table 3 and Table 4) and the estimates of statistical power are obtained when the true *AUC*s for two tests are equal and different respectively. Rejection regions for the tests are determined using a nominal significance level of 0.05.

In case of continuous ratings, for lower *AUCs* the proposed McNemar test demonstrates a more conservative type I error and consequently an elevated power compared to the other two tests. But for moderate correlation among the modalities and increasing sample sizes the scenario tends to be more stable and all three tests comes to be very close in both type I error and statistical power. The proposed test exhibits more false positive rate for lower correlation which is because of the reason that McNemar test is appropriate to be used with correlated data. With increasing value of correlation the estimates of false positive rate tend to diminish. For higher values of *AUCs*, the type I error estimates are comparable for all sample sizes and correlation values. In cases with higher *AUCs* and lower correlation values the proposed test outperforms the tests by DeLong *et al.* [3] and Bandos *et al.* [2]. Though for moderate correlation values with higher *AUCs* other two tests provide better statistical

Table 3. Empirical type I error and statistical power of different tests for continuous ratings.

AUC	Mean	Variance	Sample size		$\rho = 0$			$\rho = 0.4$			$\rho = 0.6$		
A_1, A_2	μ_1, μ_2	$\sigma_1^2 = \sigma_2^2$	N	M	D^a	B^b	Mc^c	D^a	B^b	Mc^c	D^a	B^b	Mc^c
Type I error and statistical power													
0.6, 0.6	0.36, 0.36	1.0	20	20	0.050	0.041	0.066	0.049	0.045	0.060	0.052	0.053	0.051
			40	40	0.046	0.044	0.073	0.048	0.049	0.065	0.051	0.051	0.049
Type I error			60	60	0.059	0.058	0.096	0.041	0.041	0.062	0.046	0.046	0.057
			100	100	0.048	0.048	0.088	0.044	0.043	0.084	0.044	0.043	0.077
			150	150	0.044	0.043	0.098	0.043	0.043	0.073	0.047	0.047	0.072
0.6, 0.7	0.36, 0.74	1.0	20	20	0.122	0.100	0.184	0.172	0.163	0.205	0.226	0.215	0.200
			40	40	0.189	0.178	0.335	0.298	0.288	0.388	0.398	0.387	0.454
Power			60	60	0.230	0.286	0.459	0.450	0.440	0.554	0.588	0.576	0.633
			100	100	0.442	0.431	0.679	0.638	0.629	0.782	0.801	0.792	0.877
			150	150	0.609	0.605	0.842	0.809	0.802	0.915	0.937	0.933	0.963
0.6, 0.8	0.36, 1.19	1.0	20	20	0.405	0.365	0.469	0.571	0.524	0.559	0.724	0.627	0.604
			40	40	0.706	0.679	0.804	0.871	0.838	0.884	0.956	0.943	0.927
Power			60	60	0.683	0.850	0.940	0.976	0.968	0.979	0.998	0.992	0.991
			100	100	0.978	0.977	0.996	0.999	0.999	0.999	1.000	1.000	1.000
			150	150	0.997	0.997	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.6, 0.9	0.36, 1.81	1.0	20	20	0.817	0.767	0.763	0.939	0.904	0.836	0.986	0.969	0.884
			40	40	0.991	0.984	0.983	0.999	0.999	0.992	1.000	1.000	0.999
Power			60	60	0.999	0.998	0.999	1.000	1.000	1.000	1.000	1.000	1.000
			100	100	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			150	150	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.7, 0.7	0.74, 0.74	1.0	20	20	0.048	0.042	0.049	0.047	0.045	0.030	0.050	0.052	0.030
			40	40	0.041	0.039	0.045	0.050	0.049	0.048	0.051	0.050	0.042
Type I error			60	60	0.062	0.058	0.048	0.037	0.036	0.047	0.047	0.049	0.040
			100	100	0.034	0.034	0.066	0.052	0.051	0.052	0.050	0.050	0.051
			150	150	0.049	0.049	0.065	0.042	0.042	0.052	0.055	0.055	0.050
0.7, 0.8	0.74, 1.19	1.0	20	20	0.137	0.126	0.118	0.197	0.187	0.129	0.254	0.246	0.151
			40	40	0.232	0.221	0.229	0.351	0.340	0.272	0.471	0.461	0.325
Power			60	60	0.362	0.349	0.354	0.527	0.513	0.418	0.679	0.669	0.494
			100	100	0.562	0.552	0.577	0.745	0.734	0.680	0.871	0.859	0.770
			150	150	0.730	0.724	0.756	0.904	0.899	0.870	0.970	0.967	0.912
0.7, 0.9	0.74, 1.81	1.0	20	20	0.532	0.498	0.357	0.697	0.657	0.415	0.825	0.781	0.468
			40	40	0.858	0.833	0.694	0.960	0.947	0.779	0.991	0.981	0.842
Power			60	60	0.954	0.944	0.893	0.996	0.994	0.930	1.000	0.999	0.970
			100	100	0.999	0.998	0.985	1.000	1.000	0.997	1.000	1.000	1.000
			150	150	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

[a] *D*– Conventional *AUC* test (DeLong *et al.*)
[b] *B*– Approximation to permutation *AUC* test (Bandos *et al.*)
[c] *Mc*– McNemar test

power than the McNemar test, however, for increasing sample sizes McNemar test gives statistical power close to others.

When binary ratings data are considered, in all scenarios of parameter settings and for all sample sizes (small or large), the proposed McNemar test demonstrates superior statistical power and less conservative type I error compared to tests by DeLong *et al.* [3] and Bandos *et al.* [2]. In summary, for continuous type ratings, our simulations demonstrate that the proposed McNemar test provides close agreement of type I error to the nominal level. For smaller *AUC* values this agreement comes with moderate and higher correlation among the

Table 4. Empirical type I error and statistical power of different tests for discrete binary ratings.

Marginal probability			Sample size		$\rho = 0.00$			$\rho = 0.25$			$\rho = 0.50$		
P_{a+b}	P_{a+c}	$P_b - P_c$	N	M	D^a	B^b	Mc^c	D^a	B^b	Mc^c	D^a	B^b	Mc^c
Type I error and statistical power													
0.60	0.60	0.00	20	20	0.066	0.060	0.028	0.078	0.063	0.023	0.072	0.055	0.025
			40	40	0.060	0.055	0.039	0.070	0.069	0.040	0.070	0.066	0.049
Type I error			60	60	0.069	0.067	0.049	0.081	0.075	0.057	0.094	0.093	0.061
			100	100	0.085	0.082	0.069	0.098	0.096	0.081	0.108	0.105	0.092
			150	150	0.116	0.113	0.080	0.125	0.123	0.094	0.136	0.133	0.120
0.60	0.70	0.10	20	20	0.062	0.055	0.063	0.068	0.050	0.070	0.077	0.053	0.082
			40	40	0.072	0.065	0.132	0.074	0.070	0.147	0.076	0.069	0.221
Power			60	60	0.080	0.070	0.205	0.077	0.070	0.249	0.098	0.095	0.337
			100	100	0.090	0.088	0.299	0.093	0.088	0.381	0.118	0.110	0.560
			150	150	0.103	0.099	0.440	0.113	0.111	0.558	0.148	0.141	0.765
0.60	0.80	0.20	20	20	0.113	0.103	0.148	0.147	0.107	0.184	0.185	0.141	0.237
			40	40	0.183	0.166	0.344	0.232	0.214	0.409	0.304	0.269	0.585
Power			60	60	0.244	0.223	0.511	0.321	0.294	0.610	0.446	0.423	0.795
			100	100	0.377	0.364	0.720	0.490	0.460	0.847	0.644	0.610	0.960
			150	150	0.522	0.498	0.908	0.626	0.604	0.961	0.807	0.788	0.994
0.70	0.70	0.00	20	20	0.066	0.057	0.032	0.072	0.058	0.024	0.079	0.061	0.025
			40	40	0.058	0.055	0.043	0.060	0.059	0.042	0.067	0.061	0.049
Type I error			60	60	0.077	0.072	0.056	0.085	0.080	0.059	0.096	0.095	0.062
			100	100	0.086	0.085	0.061	0.095	0.091	0.076	0.137	0.131	0.087
			150	150	0.099	0.097	0.087	0.119	0.117	0.099	0.164	0.160	0.130
0.70	0.80	0.10	20	20	0.063	0.052	0.065	0.061	0.047	0.076	0.077	0.055	0.086
			40	40	0.070	0.066	0.138	0.071	0.069	0.161	0.079	0.069	0.245
Power			60	60	0.077	0.073	0.215	0.087	0.083	0.268	0.099	0.096	0.381
			100	100	0.091	0.085	0.353	0.098	0.093	0.433	0.131	0.123	0.583
			150	150	0.111	0.104	0.481	0.111	0.107	0.607	0.167	0.158	0.784
0.70	0.90	0.20	20	20	0.128	0.109	0.158	0.153	0.113	0.197	0.199	0.154	0.257
			40	40	0.199	0.185	0.373	0.252	0.239	0.446	0.337	0.302	0.628
Power			60	60	0.279	0.260	0.565	0.358	0.326	0.666	0.474	0.445	0.840
			100	100	0.423	0.407	0.786	0.502	0.474	0.884	0.705	0.672	0.974
			150	150	0.585	0.566	0.932	0.697	0.675	0.979	0.853	0.821	0.999

[a] $D-$ Conventional AUC test (DeLong *et al.*)
[b] $B-$ Approximation to permutation AUC test (Bandos *et al.*)
[c] $Mc-$ McNemar test

modalities. And for larger sample sizes the proposed test produces very comparable statistical power. Furthermore for discrete binary ratings, the proposed McNemar test possesses better operating characteristics than the other tests in all parameter settings considered. The performance of the proposed McNemar test in case of continuous type ratings might be suppressed in our simulation study due to difficulty in choosing optimal cut-off point to classify the subjects. To make it more evident in Section 6, we have considered a practical data set where a real cut-off point exists and conducted a bootstrapping power analysis to compare the statistical power of all the three tests.

6. Example

In this section we have presented an application and comparison of proposed and other considered tests using a practical data set adopted from Venkatraman *et al.* [17]. In patients with clinically localized primary testicular cancer it is important to determine the necessity

for an operation to remove any disease that may have spread to retroperitoneal lymph nodes. These nodes can be evaluated by computed tomography to determine the necessity for this operation. In the considered data set, the size of the largest node detected by computed tomography was used as diagnostic criterion and the goal of the study was to determine if the accuracy of this criterion is different for anterior versus posterior nodes. The 'gold standard' diagnosis is the presence of any nodal disease at surgery.

The test result recorded is the average size in millimeters of the largest node detected by three independent readers. Anything smaller than 4 millimeters is considered undetectable by the naked eye. The null hypothesis to test is that the sizes of anterior and posterior nodes possess equivalent diagnostic information.

The estimates of *AUCs* for diagnostic procedures 1 and 2 are 0.787 and 0.568 respectively and the estimated Pearson correlation coefficient between the ratings under two procedures is 0.165. To test equality of performances of these two diagnostic procedures, the conventional test by DeLong *et al.* [3], asymptotic permutation test by Bandos *et al.* [2] and the proposed McNemar test come to agreement of significant different performances yielding two sided p-values 0.0068, 0.019 and 0.0027 respectively. To search for more specific information regarding statistical power of tests, we have conducted a bootstrapping study where for each of considered sample sizes, 2000 random samples were taken from the data and rejection rates are computed.

Table 5. Bootstrapping statistical power for different tests.

Sample size		Rejection rate		
N	M	D[a]	B[b]	Mc[c]
20	20	0.557	0.448	0.660
30	30	0.728	0.651	0.859
40	40	0.844	0.790	0.944
60	60	0.959	0.943	0.994
100	100	0.994	0.991	1.000

[a] $D-$ Conventional AUC test (DeLong *et al.*)
[b] $B-$ Approximation to permutation AUC test (Bandos *et al.*)
[c] $Mc-$ McNemar test

Table 5 and Figure 1 demonstrates that for all sample sizes proposed McNemar test provides superior rejection rate and in large sample sizes tests by DeLong *et al.* [3] and Bandos *et al.* [2] shows rejection rates very closed to the McNemar test.

7. Conclusion

The area under the ROC curve (AUC) can be used for the comparison of two or more diagnostic tests. By the comparison of two *AUCs*, it is investigated which one of the two diagnostic tests is more suitable for discriminating non-diseased from diseased subjects. Since most of the diagnostic researches yield matched data, it is important to take into account the correlated nature of the diagnostic tests. The McNemar test is used for the correlated data.

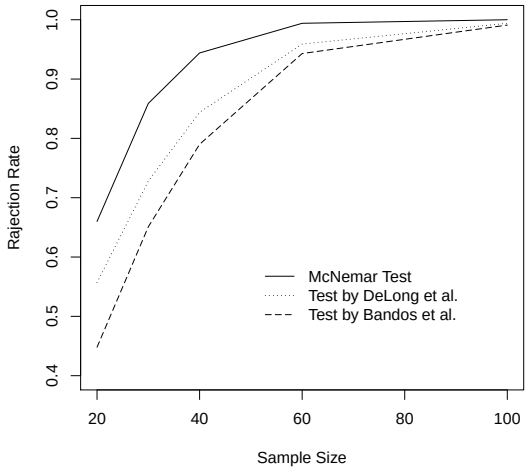

Figure 1. Bootstrapping power curves to compare different tests

Two *AUCs* are compared to test whether two diagnostic tests have the same ability to discriminate diseased and non-diseased subjects. If two diagnostic test procedures have same discriminating power then it can be said that they have come from same population. The McNemar test is used to test whether two populations differ significantly. If the two populations do not differ then they should have same discriminating ability to correctly identify diseased and non-diseased subjects. The McNemar test is based on an optimal cut-off point. The optimal cut-off point varies according to the clinical application of diagnostic tests. The simulation study, we have conducted, shows that the proposed McNemar test can be a very suitable alternative to the test by DeLong *et al.* [3] and test by Bandos *et al.* [2] that are very cumbersome to compute. An application to practical data set and bootstrapping study also supports our claim. Since the McNemar test is easy to compute as well as easy to communicate to the potential uses of the procedure, we can use this test conveniently. The strength of our proposed method is that it has easy implementation to discriminate diagnostic test procedures even by non-statisticians. Knowledge of true status of subjects or any other gold standard is not required to employ the proposed test. The idea of the McNemar test can also be used to construct confidence regions.

Acknowledgement. The authors are grateful to the anonymous reviewers for their insightful comments that contributed to improvement in the exposition of this paper to a great extent.

References

[1] D. Bamber, The area above the ordinal dominance graph and the area below the receiver operating characteristic graph, *J. Mathematical Psychology* **12** (1975), no. 4, 387–415.

[2] A. I. Bandos, H. E. Rockette and D. Gur, A permutation test sensitive to differences in areas for comparing ROC curves from a paired design, *Stat. Med.* **24** (2005), no. 18, 2873–2893.

[3] E. R. DeLong, D. M. DeLong and D. L. Clark-Pearson, Comparing the area under two or more correlated receiver operating characteristic curves: A nonparametric approach, *Biometrics* **44** (1988), 837–845.
[4] D. D. Dorfman and E. Alf, Maximum likelihood estimation of parameters of signal detection theory-a direct solution, *Psychometrika* **33** (1968), 117–124.
[5] D. D. Dorfman and E. Alf, Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals-rating method data, *J. Math. Psych.* **6** (1969), 487–496.
[6] J. A. Hanley and J. B. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, *Radiology* **143** (1982), 29–36.
[7] W. Hoeffding, A class of statistics with asymptotically normal distribution, *Ann. Math. Statistics* **19** (1948), 293–325.
[8] M. A. Islam, R. I. Chowdhury and L. Briollais, A bivariate binary model for testing dependence in outcomes, *Bull. Malays. Math. Sci. Soc. (2)* **35** (2012), no. 4, 845–858.
[9] F. Leisch, A. Weingessel and K. Hornik, *On the generation off correlated artificial binary data*, Working paper series. Working paper No. 13, Vienna University of Economics and Business Administration, August 2–6, 1090 Wien, Austria, 1998.
[10] Y. Lu, H. Jin and H.K. Genant, On the non-inferiority of a diagnostic test based on paired observations, *Statist. Med.* **22** (2003), 3029–3044.
[11] A. W. Marshall and I. Olkin, A family of bivariate distributions generated by the bivariate Bernoulli distribution, *J. Amer. Statist. Assoc.* **80** (1985), no. 390, 332–338.
[12] Q. McNemar, Note on the sampling error of the differences between correlated proportions or percentages, *Psychometrika* **12** (1947), 153–157.
[13] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2010.
[14] E. F. Schisterman, N. J. Perkins, A. Liu A and H. Bondell, Optimal cut-point and its corresponding Yuden index to discriminate individuals using pooled blood samples, *Epidemiology* **16** (2005), 73–81.
[15] P. K. Sen, On some convergence properties of U-statistics, *Calcutta Statist. Assoc. Bull.* **10** (1960), 1–18.
[16] S.A. Swets, Sensitivities and specificities of diagnostic tests, *J. Am. Med. Assoc.* **248** (1982), no. 5, 548–549.
[17] E. S. Venkatraman and C. B. Begg, A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment, *Biometrika* **83** (1996), no. 4, 835–848.
[18] R. F. Wagner, S. V. Beiden and C. E. Metz, Continuous versus categorical data for ROC analysis: Some quantitative considerations, *Acad. Radiol.* **8** (2001), 328–334.
[19] X.-H. Zhou, N. A. Obuchowski and D. K. McClish, *Statistical Methods in Diagnostic Medicine*, Wiley Series in Probability and Statistics, Wiley-Interscience, New York, 2002.
[20] M. H. Zweig and G. Campbell, Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine, *Clin. Chem.* **39** (1993), 561–577.