Journal of Mathematics and Statistics

Original Research Paper

A Permutation Test for Comparing Two Correlated Receiver Operating Characteristic Curves

[1]Okeh Uchechukwu Marius and [2]Onyeagu Sidney

[1]Department of Industrial Mathematics and Applied Statistics, Ebonyi State University Abakaliki Nigeria, Nigeria
[2]Department of Statistics, Nnamdi Azikiwe University Awka, Anambra State Nigeria

Article history
Received: 07-11-2019
Revised: 26-01-2020
Accepted: 23-05-2020

Corresponding Author:
Okeh Uchechukwu Marius
Department of Industrial Mathematics and Applied Statistics, Ebonyi State University Abakaliki Nigeria, Nigeria
Email: uzomaokey@ymail.com

Abstract: The area under the Receiver Operating Characteristic (ROC) curve (AUC) is a summary measure when comparing two ROC curves. However, this summary measure is less informative when two ROC curves cross and have the same AUCs. In order to detect differences between ROC curves and to be able to tackle the problem of exchangeability of the labels between two diagnostic tests within subject, an alternative permutation test based on between-subject permutations of the labels of the subjects within each diagnostic test is proposed for assessing a change in the AUCs in a continuous matched pair of data from two diagnostic test procedures having both non-diseased and diseased subject in each of the test. The Wilcoxon signed rank test statistic was modified as a permutation test under the null hypothesis of equality of AUCs. An algorithm for carrying out complete enumeration of all the distinct permutations of the paired test results was developed which provides exact p-values. Using simulated data, the proposed test compares in statistical power to the modified sign test proposed by Braun and Alonzo but the proposed test has better operating characteristics, that is greater statistical power to detect a crossing alternative and is less conservative in test size and in the range of parameters of at least 0.8 of AUCs on the average with a correlation of at least 0.4 and small to moderately large sample sizes. Similarly in applying real life data, the proposed test has the more likelihood of rejecting null hypothesis of equality of AUC_1 and AUC_2 at nominal level of 0.05 with the proposed test having a p-value of 0.0312 against the Braun and Alonzo's test with a p-value of 0.0387. This is because the proposed test is modified to adjust for the presence of zero differences in values and considers the signs of values as well as the absolute ranks of values. Also the estimates of AUC_1 and AUC_2 for the two diagnostic tests are 0.668 and 0.887 respectively showing that AUC_2, that is 2hour 100g Oral Glucose Tolerance Test (OGTT) is superior to AUC_1 (2hour 70g OGTT) at a time that the specificity is greater than 0.7.

Keywords: Permutation Test, Exchangeability, Asymptotic Approximation, Algorithm, Two Diagnostic Test Procedures, Area Under the ROC Curve (AUC), Modified Wilcoxon Signed Rank Test, Receiver Operating Characteristic (ROC) Curve

Introduction

Nonparametric inference for a difference in Areas Under the Curve (AUCs) for paired studies was first proposed by DeLong *et al*. (1988). They developed a conventional fully nonparametric approach to compare two correlated AUCs of two diagnostic tests for paired samples of subjects by using the asymptotic theory of generalized *U* statistics (Hoeffding, 1948) and used the jackknife to estimate the covariance of the 2 *U*-statistics all lead to an asymptotically normal test statistic. The test by DeLong *et al.* is limited by the fact that the AUC has an unbiased non-parametric estimator called the indicator variable that requires the comparison of all the number of subjects responding positive and negative. Other nonparametric inference procedures include those based upon an analysis of variance of jackknife pseudo-values (Dorfman *et al.*, 1992; Song, 1997) and bootstrap-based methods (Campbell, 1994; Moise, 1988). However, these methods are valid for large sample size, so that computational time could be long and their test of difference in AUC is not valid in small samples. A competing nonparametric approach that is valid for small sample size is permutation test. Permutation based procedures are specific to hypothesis testing. A permutation procedure constructs a permutation sample space, which consists of the equally likely permutation samples. The permutation samples are created by interchanging the units of the data that are assumed to be “exchangeable” under the null hypothesis. The permutation sample space is the exact probability space of the possible arrangements of the data under the null hypothesis given the original sample. This natural permutation test is characterized by exchanging the paired units when two diagnostic test procedures are to be compared with paired data. Three permutation tests for paired Receiver Operating Characteristic (ROC) studies currently exist: One proposed by Venkatraman and Begg (1996), one from Bandos *et al.* (2005) and the other from Braun and Alonzo (2008). However, Venkatraman and Begg (1996) and Bandos *et al.* (2005) proposed a permutation tests concerning correlated Receiver Operating Characteristic (ROC). The test of Bandos *et al.* (2005) directly tests for an equality of AUCs, the test of Venkatraman and Begg (1996) is more general and tests for equality of the underlying ROC curves, while the test by Braun and Alonzo (2008) compares the ROC curves but is designed to have increased power of detecting a difference in AUCs. As a result, the test of Venkatraman and Begg is less powerful for testing equality of AUCs but more general in testing for the equality of the overall ROC curves. While the test by Venkatraman and Begg is specifically designed to detect any differences between two ROC curves at every operating point, the test of Braun and Alonzo is designed to detect differences in AUCs. By comparison, the statistical power of the permutation test by Bandos *et al.* (2005) is more than the nonparametric approach employed by DeLong *et al.* (1988) in terms of when the AUCs are large, small sample sizes and moderate correlation between diagnostic test procedures. Meanwhile, the estimator proposed by DeLong *et al.* (1988) possesses an upward bias which on the one hand results in an improved (compared to the unbiased estimator) type I error of the statistical test for equality of the AUCs when AUCs are small, but on the other hand results in loss of statistical power when AUCs are large (Bandos, 2005; Bandos *et al.* 2005). Bandos *et al.* (2005) compared the performance of their test to that of DeLong *et al.* (1988) via simulation and found that the permutation test had greater power than the nonparametric test developed by DeLong *et al.* (1988) when there was moderate correlation between diagnostic tests, large AUCs and small sample sizes. The permutation tests by Bandos *et al.* as well as Venkatraman and Begg requires exchangeability of the two diagnostic test procedures within the non-diseased and diseased labels of subjects. These permutation tests require that both diagnostic tests are exchangeable within subject and require an appropriate transformation, such as ranks, because the measurements of test results are on different scales. This means that both of these tests assume the same condition of exchangeability of the diagnostic results under the null hypothesis, but differ with respect to their sensitivity to specific alternatives and the availability of an asymptotic version.

We propose an alternative permutation test that does not require data transformation due to the presence of zero differences or tied absolute values of differences which makes test results to be taken on at most the ordinal scale and exchangeability of two diagnostic test procedures rather requires between subjects permutation of the non-diseased and diseased labels of subjects within a given diagnostic test procedure. This permutation test is based on the works of Braun and Alonzo (2008) that in their work used sign test as their permutation test. While sign test considers the direction of units measured, our test considers both the direction and magnitudes of the units of interest. In an effort to assess a difference in AUCs of the two diagnostic tests, an algorithm for computing the exact permutations of the test statistic will be implemented. In the next section, we propose our permutation test and show that it is equal to modified Wilcoxon signed rank test statistic. In section 3, we shall also present an algorithm for computing the exact distribution of the permutation test. In section 4, we describe the simulation and real life data, apply the proposed test on the data and present the results. In section 5, we discuss the result of the simulation in terms of the operating characteristics of the proposed test, compare the test size and power of our test and a competing test as well as compare the power of the two tests using real life data. In section 6, we make our summary and conclusions.

Proposed Permutation Test

The proposed method discussed here is a permutation test designed to compare the AUCs of two diagnostic test procedures given as AUC_1 and AUC_2 having a total number of n subjects and where subject labels are exchangeable within each diagnostic test under null hypothesis. Since an issue in a permutation test is to choose a test statistic that discriminates between the null and alternative hypothesis and given the fact that a popular choice is a test statistic developed in asymptotic theory, we therefore modify for use, Wilcoxon signed rank test statistic.

The procedure is such that a total number of N non-diseased subjects and M diseased subjects each received both diagnostic tests. Let the test results of diagnostic tests 1 and 2 for the non-diseased subject be X_{i1} and X_{i2} where $i = 1,\ldots,N$. Also let the test results of diagnostic tests 1 and 2 for the diseased subject be Y_{j1} and Y_{j2} where $j = 1,\ldots,M$. Also let $X = \{(X_{11}, X_{12}), (X_{21}, X_{22}),\ldots, (X_{N1}, X_{N2})\}$ denotes pairs of vector of measurement on non-diseased subjects and let $Y = \{(Y_{11}, Y_{12}), (Y_{21}, Y_{22}),\ldots, (Y_{M1}, Y_{M2})\}$ be the pairs of vector of measurement on diseased subjects. Therefore the difference in AUCs given as $AUC_\Delta = AUC_2 - AUC_1$ is estimated non-parametrically as:

$$A\hat{U}C_\Delta = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} Q\left(X_{im}, X_{jm}\right)$$
$$= \left[\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} Q\left(X_{i2}, \mathrm{Y}_{j2}\right) - \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} Q\left(X_{i1}, \mathrm{Y}_{j1}\right)\right] \quad (1)$$

where, Q(X_{im}, Y_{jm}) = S_{ij2} - S_{ij1} = S_{ijm} and S_{ijm} = $A(Y_{jm} > X_{im}) + \frac{1}{2} A(X_{im} > Y_{jm})$; $m = 1,2$:

$$S_{ij2} - S_{ij1} = \left[\left[A\left(Y_{j2} > X_{i2}\right) + \frac{1}{2}A\left(X_{i2} = Y_{j2}\right)\right]\right.$$
$$\left.-\left[A\left(Y_{j1} > X_{i1}\right) + \frac{1}{2}A\left(X_{i1} = Y_{j1}\right)\right]\right].$$

Consider according to Hanley and McNeil (1982), that this indicator function is:

$$S_{ijm} = \begin{cases} 1 \ \ if \ Y_{jm} > X_{im} \\ 0.5 \ \ if \ X_{im} = Y_{jm} \\ 0 \ \ if \ Y_{jm} < X_{im} \end{cases} \quad (2)$$

In other to test the null hypothesis H_0: AUC_2- AUC_1 = 0, we combine N and M subjects to have a total of n subjects and let $S_1 = \{S_{11}, S_{12},\ldots, S_{1N}, S_{1,N+1}, S_{1,N+2}\ldots., S_{1n}\}$ be n measurements arising from diagnostic test 1 while the subscripts p = 1,2,..,N shows test results for the non-diseased subjects while q = N +1, N +2,….,n shows test results for the diseased subjects. Based on this arrangement within diagnostic test 1, we compare every subject's test result to every other subject's test result. Thus:

$$R_{pq1} = A\left(S_{q1} > S_{p1}\right) + \frac{1}{2}A\left(S_{p1} = S_{q1}\right); if \ \ p \neq q \quad (3)$$

This implies that every diseased subject is compared to all non-diseased subjects and all (M-1) other diseased subjects. Similarly, every non-diseased subject is compared to all diseased subjects and all (N-1) other non-diseased subjects. Also let $S_2 = \{S_{21}, S_{22},\ldots, S_{2N}, S_{2,N+1}, S_{2,N+2}\ldots., S_{2n}\}$ be n measurements arising from diagnostic test 2 while the subscripts p = 1,2,..,N shows test results for the non-diseased subjects while q = N +1, N +2,….,n shows test results for the diseased subjects. Similarly within diagnostic test, 2, we compare every subject's test result to every other subject's test result, that is:

$$R_{pq2} = A\left(S_{q2} > S_{p2}\right) + \frac{1}{2}A\left(S_{p2} = S_{q2}\right); if \ \ p \neq q. \quad (4)$$

Given the above definitions, therefore R_{pq} = 1- R_{pqm}; m = 1,2.

To test the null hypothesis that AUC_Δ = 0, which is similar to testing the null hypothesis that the difference between paired samples is a distribution that is symmetric around zero, we adopt the transformation in Equation 2 whose indicator function is [1,0.5,0] and adjust for the presence of ties (zero difference) from the diagnostic pairs and disease status[0,1] and map to [1,0,-1]. Given the specifications above, we generalize the estimate of AUC_Δ as:

$$A\hat{U}C_\Delta = \frac{1}{NM}\sum_{p=1}^{N}\sum_{q=1}^{M} iT_{pq} = \frac{1}{NM}\sum_{p=1}^{N}\sum_{q=1}^{M} T_{pq} r\left|Q_{pq}\right| \quad (5)$$

Where:

$$T_{pq} = \begin{cases} 1, if \ \ p \ and \ q \ test \ result \ of \ \ subject \ is \ nondiseased(-) and \ \ diseased(+) \ respectively \\ -1, if \ \ p \ and \ q \ test \ result \ of \ \ subject \ is \ diseased(+) and \ \ nondiseased(-) \ respectively \\ 0, \ \ if \ \ p \ and \ q \ test \ result \ of \ \ subject \ are \ both \ diseased(+) \ or \ both \ nondiseased(-) \end{cases}$$

$$and \ \ r\left(Q_{pq}\right) = \left(R_{pq2} - R_{pq1}\right). \ Note \ that \ i = rank \ of \left(\left|Q_{pq}\right|\right).$$

Note that Q_{pq} is the difference between the sample pairs of S_1 being measurements arising from diagnostic test 1 and S_2 being measurements arising from diagnostic test 2. This is based on the exchangeability of the diseased and non-diseased labels of the subjects within each diagnostic test. The indicator function T_{pq} takes value 1 at the calibrated cut-off point *c* of a given diagnostic test if subject test result *p* is non-diseased and subject test result *q* is diseased. It takes -1 if subject test result *p* is diseased and subject test result *q* is non-diseased. Values of 0 represents cut-offs at which both subject test results *p* and *q* are diseased or non-diseased. Recall that the AUC is equivalent to two-sample Wilcoxon test statistic (Pardo and Franco-Pereira, 2017) and can be used to carry out test of symmetry around zero for paired samples. Based on that finding, the Equation 5 above which is the modified Wilcoxon Signed rank test statistic is equivalent to difference in AUCs and can be used as a test statistic for the test of symmetry around zero. This proposed test statistic is more powerful than the modified sign test statistic (Oyeka, 2009) proposed by Braun and Alonzo (2008) for comparing correlated ROC curves as it utilizes both the signs, T_{pq} and the absolute ranks of Q_{pq}.

When both diagnostic tests results are measured continuously, testing the hypothesis that $AUC_\Delta = 0$ is equal to testing the null hypothesis that $r(Q_{pq})$ is a symmetric distribution around zero. We therefore test the null hypothesis that $AUC_\Delta = 0$ by computing $AUC_\Delta = 0$ for every permutation of T_{pq}, the signs of the rank of $|Q_{pq}|$. Given that our permutation of T_{pq} requires exchanging the labels of non-diseased subject's test results *p* and diseased subject's test result *q*, it is the same as permuting among the subjects, the vector of test results of diseased/non-diseased labels. Therefore, the link between the true diseased status of a given subject as well as its test results arising diagnostic tests 1 and 2 are dislodged under this type of permutation arrangement. This permutation test is therefore valid if either one of the AUC of the diagnostic tests is equal to *t*, where *t* is a number in between 0.5 and 1 inclusive.

An Algorithm for Computing the Exact Distribution of the Permutation Test, $\hat{W}(A\hat{U}C_\Delta)$

To ensure that the probability of a type I error is exactly α, thus obtaining exact p-values, an algorithm for obtaining exact permutation distribution of the test statistic, $A\hat{U}C_\Delta$, is presented by implementing it in Intel Visual FORTRAN. This software package is to be used because it can carry out sampling without replacement, which increases the power of the permutation test. For a complete enumeration of all the paired permutations of the two diagnostic test results, the required number of permutations is given by:

$$\sum_{s=1}^{n}\binom{n}{s} = 2^n \ where\ n = M + N.$$

Therefore a paired sample design with *n* pairs has 2^{N+M} possible permutations of the variates with each permutation occurring with probability 2^{-N+M}.

Since $S_1 = \{S_{11}, S_{12},\ldots, S_{1N}, S_{1,N+1}, S_{1,N+2}\ldots., S_{1n}\}$ and $S_2 = \{S_{21}, S_{22},\ldots, S_{2N}, S_{2,N+1}, S_{2,N+2}\ldots., S_{2n}\}$ are *n* measurements arising from diagnostic tests 1 and 2 respectively where the subscripts $p = 1,2,..,N$ represents test results for the non-diseased subjects and $q = N+1, N+2,\ldots.,n$ representing test results for the diseased subjects, we consider AUC_Δ given in (5) as the test statistic and test the null hypothesis H_0: $AUC_1 = AUC_2$ versus H_1: $AUC_1 \neq AUC_2$.

Suppose the test statistic AUC_Δ and it is required that difference in AUCs should be computed for all pairs arising from diagnostic test 1 and 2, we therefore for simplicity replace our test statistic AUC_Δ with W. Let $W = (W_1, W_2, W_2,\ldots, W_m)$ be m distinct values of the test statistic *W*. The probability distribution of the test statistic *W* under the null hypothesis is given by:

$$P\left(W_l = w_0 \middle| H_0\right) = \sum_{k=1}^{f_l}\left(2^{-N+M}\right) = f_l\left(2^{-N+M}\right), \tag{6}$$

where, f_l is the frequency of occurrences of W_l. Given a particular value of n and significant level α, *c* being the critical value is in correspondence to the closest of α. The distinct occurrences of *W* are therefore all ordered in an increasing order of size. If the point occupied by the observed value of *W* is h, then the left and right side of the probability distribution of *W* has level of significance given as:

$$\alpha = P\left(W_h \leq c \middle| H_0\right) = \sum_{l=1}^{h}\sum_{k=1}^{f_l}\left(2^{-N+M}\right) = 2^{-N+M}\sum_{l=1}^{h} f_l \tag{7}$$

And:

$$\alpha = P\left(W_h \geq c \middle| H_0\right) = \left(2^{-N+M}\right)\sum_{l=h}^{m} f_l. \tag{8}$$

Since the alternative hypothesis suggests a two sided test, the left and right side are added up. Therefore, for a symmetric distribution of *W* around zero:

$$\sum_{l=1}^{h} f_l = \sum_{l=m-h+1}^{m} f_l. \tag{9}$$

Since permuted subjects labels are represented by S_1 and S_2 from diagnostic test 1 and 2 respectively, let $\{\theta_1, \theta_1, \ldots, \theta_n\}$ be a set of all distinct permutations resulting from S_1 and S_2 pairs from diagnostic test 1 and 2 such that θ_s is the s^{th} permutation.

The steps involved in the permutation test are defined as follows:

1. Calculate the Test Statistic, W_1 for the original observations θ_1
2. Obtain a distinct permutation θ_s
3. Calculate the Test Statistic for the distinct permutation, θ_s that is $W(\theta_s)$
4. Go back to Steps 2 and 3 and repeat for $s = 2,3,\ldots,2^n$, $n = N + M = sample\ size$
5. Now build the empirical cumulative probability distribution as:

$$p_0 = \hat{p}\left(W \le W_s\right) = \frac{1}{2^n}\sum_{s=1}^{2^n} T\left(W_1 - W_s\right)$$
$$where\ T = \begin{cases} 1\, if\ W_1 > W_s \\ 0\ if\ W_1 = W_s \\ -1\, if\ W_1 < W_s \end{cases} \tag{10}$$

6. Given the empirical cumulative probability distribution $\hat{p}$, if $p_0 \le \alpha$, we reject H_0

These steps compute the empirical cumulative probability distribution of W under the null hypothesis.

An Algorithm for Calculating the Exact Distribution of $\hat{W}$

The test statistic $\hat{W}$ is computed for each permutation in the complete enumeration of the distinct permutations. The distribution of the test statistic is obtained by tabulating the distinct values of the statistic against their probabilities of occurrence in the complete enumeration, bearing in mind that all the permutations are equally likely. The paired permutation is constructed by letting S_{sm} represent the paired test results of subjects in the two diagnostic tests 1 and 2, where $s = 60$; $m = 1,2$. See appendix A1 for the algorithm.

Examples

a. Simulation Description and Implementation

Test results from two diagnostic test procedures were simulated for the purpose of comparing the test sizes and statistical power of the proposed permutation test for various underlying AUC differences, different sample sizes and correlations between two diagnostic test procedures as follows. In other to generate data, we assumed and drew two continuous measurements for each non-diseased subject from a bivariate normal distribution centered at $\mu_X = 0$, with both measurements having a marginal variance of 1.0. So that for mth diagnostic test, we have $\mu_{x^m} = 0\ and\ \sigma^2_{x^m} = 1.0; m = 1, 2.$

Since two ROC curves taken from measurements with same variances cannot cross each other, we drew two continuous measurements for each diseased subject from a bivariate normal distribution centered at μ_Y, with both measurements having a marginal variance of 1.0 for diagnostic tests procedures having non-crossing ROC curves. The values in μ_Y are directly determined from AUC_1 and AUC_2 particularly from the Hanley and McNeil (1982) equation of AUC. We assume unequal variances such as $\sigma^2_{y^1} = 1.0\ and\ \sigma^2_{y^2} = 3.0$ for diagnostic tests procedures with crossing ROC curves. We also assume the correlation for all the scenarios to be ρ = 0.25, 0.5 and 0.75.

A total of 10000 replications are computed for a given case while the sample sizes of 20,40,60 and 80 are considered and used in obtaining both the type I error (test size) and statistical power are obtained for sample sizes 20, 40,60 and 80. A nominal significance level of 5% was used in determining the region of rejection for the tests. The exact values are compared with the asymptotic 95% confidence interval (0.036, 0.064) around a significant level of 5% is on the basis of 10000 simulation in each case.

From AUCs and variances values for both non-crossing and crossing ROC curves, the values of mean of diagnostic test results denoted as μ_X and μ_Y for non-diseased and diseased subjects respectively are obtained from the Hanley and McNeil (1982) equation of AUC while the variance-covariance matrix is constructed.

The main essence of data simulation is to evaluate the ability to control Test size (Type I error) and to achieve higher statistical power for the proposed permutation test as compared to other tests. To know the Test size (type I error) and statistical power of the normal approximation (asymptotic pattern) and exact values of various AUCs that are involved, how correlated subjects' test results are across diagnostic tests at different sample sizes. Here equal correlation is assumed for non-diseased and diseased subjects across the two diagnostic test results that are continuous while non-crossing as well as crossing of ROC curves are similarly considered. We compared the size and power of the permutation test to another method in terms of their exact permutation and their normal approximation. Because of enormous time required to implement the exact permutation procedure, the comparisons done here are limited to sample sizes that are small. In comparing the test size and statistical power of the proposed test in relation to a competing method, six tables were obtained as well as four scenarios showing the ROC curves with varying AUCs. These are presented below.

Table 1: Comparison of test size for the proposed test and that of Braun and Alonzo in terms of exact and asymptotic methods with different area and non-crossing ROC curves.

		$\rho = 0.25$				$\rho = 0.50$				$\rho = 0.75$			
		MWSRT		B & A		MWSRT		B & A		MWSRT		B & A	
AUC$_1$	*AUC*$_2$	EXACT	ASY	EXACT	ASY	EXACT	ASY	EXACT	ASY	EXACT	ASY	EXACT	ASY
0.6	0.7	0.046	0.045	0.043	0.036	0.047	0.043	0.046	0.044	0.049	0.044	0.038	0.035
0.6	0.8	0.050	0.047	0.047	0.043	0.054	0.050	0.052	0.050	0.056	0.050	0.054	0.050
0.7	0.8	0.065	0.063	0.064	0.060	0.075	0.068	0.072	0.071	0.085	0.079	0.079	0.074
0.7	0.9	0.092	0.088	0.091	0.087	0.113	0.107	0.111	0.110	0.142	0.132	0.140	0.135
0.8	0.9	0.127	0.122	0.123	0.120	0.168	0.160	0.165	0.164	0.221	0.204	0.220	0.220
0.6	0.7	0.039	0.036	0.039	0.034	0.043	0.038	0.042	0.038	0.042	0.038	0.041	0.040
0.6	0.8	0.046	0.045	0.043	0.049	0.049	0.045	0.045	0.043	0.050	0.045	0.046	0.045
0.7	0.8	0.062	0.059	0.060	0.057	0.069	0.064	0.063	0.060	0.081	0.073	0.078	0.077
0.7	0.9	0.086	0.083	0.085	0.082	0.110	0.102	0.107	0.105	0.136	0.124	0.136	0.129
0.8	0.9	0.126	0.120	0.125	0.122	0.171	0.159	0.170	0.170	0.223	0.201	0.222	0.220
0.6	0.7	0.032	0.030	0.030	0.026	0.034	0.038	0.032	0.032	0.032	0.030	0.030	0.026
0.6	0.8	0.036	0.034	0.028	0.023	0.040	0.042	0.036	0.034	0.044	0.042	0.042	0.041
0.7	0.8	0.053	0.050	0.051	0.047	0.064	0.072	0.060	0.060	0.075	0.072	0.074	0.073
0.7	0.9	0.080	0.075	0.078	0.073	0.104	0.122	0.102	0.100	0.137	0.132	0.132	0.130
0.8	0.9	0.122	0.115	0.120	0.118	0.174	0.179	0.171	0.172	0.231	0.228	0.227	0.217
0.6	0.7	0.022	0.020	0.021	0.020	0.026	0.031	0.022	0.020	0.023	0.021	0.022	0.022
0.6	0.8	0.026	0.023	0.021	0.018	0.032	0.040	0.032	0.031	0.032	0.029	0.031	0.031
0.7	0.8	0.039	0.035	0.036	0.034	0.034	0.037	0.034	0.032	0.070	0.057	0.065	0.063
0.7	0.9	0.029	0.023	0.025	0.022	0.026	0.024	0.025	0.022	0.042	0.039	0.041	0.040
0.8	0.9	0.022	0.019	0.022	0.017	0.020	0.017	0.018	0.015	0.022	0.020	0.021	0.018

Sample sizes of 10 for both non-diseased and diseased subjects were simulated

Table 2: Comparison of Test size for the proposed test and that of Braun and Alonzo in terms of exact and asymptotic methods with different area and crossing ROC curves

		$\rho = 0.25$				$\rho = 0.50$				$\rho = 0.75$			
		MWSRT		B & A		MWSRT		B & A		MWSRT		B & A	
AUC$_1$	*AUC*$_2$	EXACT	ASY	EXACT	ASY	EXACT	ASY	EXACT	ASY	EXACT	ASY	EXACT	ASY
0.6	0.7	0.050	0.047	0.048	0.037	0.053	0.048	0.048	0.046	0.052	0.048	0.050	0.045
0.6	0.8	0.054	0.050	0.050	0.047	0.058	0.054	0.055	0.054	0.061	0.059	0.057	0.053
0.7	0.8	0.068	0.066	0.064	0.060	0.080	0.076	0.076	0.074	0.090	0.087	0.086	0.083
0.7	0.9	0.097	0.093	0.093	0.080	0.120	0.119	0.116	0.116	0.142	0.139	0.141	0.140
0.8	0.9	0.132	0.128	0.131	0.130	0.174	0.173	0.173	0.168	0.218	0.208	0.215	0.214
0.6	0.7	0.042	0.040	0.040	0.037	0.045	0.038	0.042	0.040	0.045	0.041	0.044	0.043
0.6	0.8	0.046	0.044	0.044	0.040	0.050	0.048	0.045	0.044	0.053	0.046	0.053	0.052
0.7	0.8	0.065	0.063	0.065	0.063	0.075	0.066	0.072	0.072	0.083	0.082	0.080	0.076
0.7	0.9	0.094	0.088	0.093	0.087	0.115	0.109	0.115	0.110	0.141	0.138	0.138	0.134
0.8	0.9	0.136	0.127	0.134	0.132	0.178	0.173	0.176	0.174	0.224	0.218	0.222	0.220
0.6	0.7	0.036	0.032	0.034	0.030	0.037	0.035	0.033	0.032	0.040	0.037	0.038	0.036
0.6	0.8	0.040	0.038	0.037	0.034	0.045	0.037	0.043	0.042	0.046	0.042	0.036	0.033
0.7	0.8	0.058	0.055	0.055	0.052	0.069	0.059	0.064	0.062	0.082	0.076	0.075	0.074
0.7	0.9	0.087	0.086	0.085	0.083	0.112	0.108	0.112	0.110	0.140	0.137	0.138	0.136
0.8	0.9	0.129	0.125	0.126	0.122	0.182	0.175	0.189	0.185	0.232	0.227	0.230	0.224
0.6	0.7	0.026	0.023	0.023	0.020	0.026	0.022	0.025	0.023	0.027	0.023	0.025	0.022
0.6	0.8	0.029	0.024	0.027	0.022	0.035	0.033	0.034	0.032	0.038	0.035	0.033	0.030
0.7	0.8	0.044	0.038	0.043	0.041	0.060	0.058	0.058	0.054	0.071	0.068	0.070	0.067
0.7	0.9	0.073	0.069	0.072	0.070	0.104	0.100	0.102	0.100	0.141	0.136	0.135	0.133
0.8	0.9	0.022	0.020	0.019	0.016	0.039	0.028	0.037	0.027	0.034	0.029	0.028	0.022

Sample sizes of 10 for both non-diseased and diseased subjects were simulated

Table 3: Comparison of Test size for the proposed test and that of Braun and Alonzo test with same area and non-crossing ROC curves in term of their asymptotic approximation test

			p = 20, q = 20		p = 40, q = 40		p = 60, q = 60		p = 80, q = 80	
ρ	AUC_1	AUC_2	B & A	MWSRT	B & A	MWSRT	B & A	MWSRT	B & A	MWSRT
0.0	0.6	0.6	0.056	0.049	0.052	0.049	0.051	0.050	0.049	0.047
	0.7	0.7	0.052	0.048	0.050	0.048	0.051	0.049	0.048	0.046
	0.8	0.8	0.050	0.046	0.050	0.048	0.050	0.049	0.049	0.048
	0.9	0.9	0.039	0.044	0.048	0.046	0.048	0.049	0.048	0.047
0.25	0.6	0.6	0.053	0.049	0.052	0.050	0.053	0.052	0.053	0.050
	0.7	0.7	0.052	0.049	0.051	0.050	0.050	0.048	0.051	0.050
	0.8	0.8	0.048	0.047	0.049	0.048	0.050	0.050	0.050	0.049
	0.9	0.9	0.044	0.045	0.047	0.048	0.050	0.050	0.051	0.049
0.5	0.6	0.6	0.051	0.050	0.050	0.050	0.051	0.050	0.050	0.048
	0.7	0.7	0.048	0.048	0.050	0.050	0.049	0.050	0.047	0.046
	0.8	0.8	0.045	0.046	0.049	0.050	0.050	0.051	0.048	0.046
	0.9	0.9	0.041	0.041	0.047	0.049	0.050	0.051	0.049	0.047
0.75	0.6	0.6	0.044	0.047	0.038	0.042	0.046	0.046	0.045	0.046
	0.7	0.7	0.043	0.045	0.037	0.041	0.043	0.044	0.042	0.043
	0.8	0.8	0.037	0.041	0.038	0.040	0.042	0.045	0.044	0.046
	0.9	0.9	0.025	0.036	0.035	0.039	0.037	0.039	0.035	0.038

Table 4: Comparison of Test size for the proposed test and that of Braun and Alonzo test with same area and crossing ROC curves in terms of their asymptotic approximation test

			p = 20, q = 20		p = 40, q = 40		p = 60, q = 60		p = 80, q = 80	
ρ	AUC_1	AUC_2	B & A	MWSRT	B & A	MWSRT	B & A	MWSRT	B & A	MWSRT
0.0	0.6	0.6	0.057	0.054	0.055	0.054	0.052	0.052	0.051	0.051
	0.7	0.7	0.055	0.052	0.054	0.053	0.052	0.051	0.048	0.049
	0.8	0.8	0.033	0.037	0.032	0.035	0.049	0.050	0.047	0.046
	0.9	0.9	0.020	0.028	0.021	0.025	0.045	0.046	0.044	0.045
0.25	0.6	0.6	0.054	0.052	0.053	0.055	0.051	0.050	0.052	0.054
	0.7	0.7	0.053	0.052	0.052	0.053	0.053	0.054	0.052	0.053
	0.8	0.8	0.040	0.045	0.050	0.051	0.050	0.052	0.049	0.048
	0.9	0.9	0.019	0.023	0.039	0.043	0.043	0.044	0.043	0.044
0.5	0.6	0.6	0.052	0.054	0.050	0.052	0.051	0.053	0.053	0.054
	0.7	0.7	0.050	0.051	0.049	0.051	0.050	0.052	0.052	0.055
	0.8	0.8	0.045	0.047	0.047	0.049	0.046	0.049	0.053	0.054
	0.9	0.9	0.020	0.023	0.034	0.036	0.037	0.040	0.039	0.040
0.75	0.6	0.6	0.047	0.050	0.050	0.055	0.050	0.054	0.051	0.053
	0.7	0.7	0.045	0.048	0.046	0.049	0.047	0.050	0.049	0.050
	0.8	0.8	0.037	0.040	0.037	0.042	0.038	0.041	0.040	0.044
	0.9	0.9	0.015	0.024	0.026	0.035	0.032	0.039	0.038	0.040

Table 1 (Fig. 1) and Table 2 (Fig. 2) examine the comparison of Test size of the proposed permutation test and Braun and Alonzo's permutation test in terms of their exact and asymptotic methods for assessing a difference in AUC for two continuous diagnostic test procedures when the areas are different for non-crossing and crossing ROC curves respectively. Since large computational time was needed for carrying out the computation of exact permutation, the comparisons shown in Table 1 (Fig. 1) and Table 2 (Fig. 2) are limited to sample sizes that are small where result indicates that good agreement exists between the exact and normal approximation test. Table 1 (Fig. 1) and Table 2 (Fig. 2) shows that even with small sample size of 10 for each of non-diseased and diseased subjects, the normal approximation test is adequate while the exact permutation test required a little computer time to conduct. Subsequent Tables 3 to 6 considered simulating the operating characteristics of the normal approximation test for large sample sizes since the exact permutation test results are essentially equivalent.

In Table 3 (Fig. 3), we compared and presented the estimates for continuous data of the test size of the proposed asymptotic normal approximation test and normal approximation test proposed by Braun and Alonzo (2008). In Table 4 (Fig. 4) where the areas are same with crossing ROC curves, the test size is the statistical power, since the proposed method is designed to detect a difference in AUCs but formally test the null hypothesis for the equality of AUCs subject to exchangeability. In Table 3 (Fig. 3) and Table 4 (Fig. 4) where the AUCs are same, for moderately large sample sizes such as 40 to 60 with non-crossing ROC curves having at least moderately high correlation between diagnostic tests, the proposed test showed a less conservative test size compared to Braun and Alonzo's test. This effect is especially evident with smaller sample sizes. In Table 4 (Fig. 4) when the AUCs are the same with crossing ROC curves, the test size of the proposed test is very close to that of the Braun and Alonzo' test since both tests is for detecting a difference in AUCs. Therefore the two methods are not advisable to be used to detect crossing ROC curves when the AUCs are the same. The closeness of the test size and the nominal level of significance suggests that two permutation tests (proposed test as well as Braun and Alonzo, 2008) which in comparison provide an asymptotic normal approximation of test of equality of AUCs are comparable in statistical power.

Table 5: Comparison of power for the proposed test and that of Braun and Alonzo's test in terms of their approximations with different area and crossing ROC curves.

		p = 20, q = 20		p = 40, q = 40		p = 60, q = 60		p = 80, q = 80	
		ρ = 0.0		ρ = 0.25		ρ = 0.5		ρ = 0.75	
AUC$_1$	*AUC*$_1$	B & A	MWSRT	B & A	MWSRT	B & A	MWSRT	B & A	MWSRT
0.6	0.7	0.076	0.071	0.082	0.086	0.090	0.102	0.180	0.200
0.6	0.8	0.142	0.135	0.179	0.183	0.213	0.236	0.544	0.575
0.7	0.8	0.251	0.240	0.332	0.339	0.422	0.450	0.880	0.883
0.7	0.9	0.403	0.387	0.535	0.541	0.655	0.680	0.937	0.954
0.8	0.9	0.476	0.459	0.566	0.572	0.656	0.673	0.996	0.998
0.6	0.7	0.079	0.076	0.087	0.090	0.092	0.106	0.197	0.215
0.6	0.8	0.154	0.145	0.194	0.201	0.232	0.257	0.593	0.624
0.7	0.8	0.277	0.267	0.366	0.375	0.459	0.489	0.914	0.926
0.7	0.9	0.452	0.437	0.587	0.595	0.705	0.735	0.983	0.987
0.8	0.9	0.537	0.532	0.612	0.621	0.822	0.820	0.995	0.998
0.6	0.7	0.084	0.081	0.093	0.102	0.101	0.118	0.275	0.289
0.6	0.8	0.174	0.167	0.221	0.227	0.265	0.293	0.777	0.801
0.7	0.8	0.323	0.313	0.423	0.435	0.524	0.552	0.979	0.988
0.7	0.9	0.531	0.520	0.623	0.631	0.874	0.831	0.993	1.00
0.8	0.9	0.542	0.535	0.724	0.753	0.924	0.953	0.993	0.994
0.6	0.7	0.091	0.088	0.115	0.135	0.125	0.162	0.375	0.406
0.6	0.8	0.205	0.202	0.350	0.386	0.410	0.480	0.914	0.923
0.7	0.8	0.410	0.401	0.534	0.542	0.896	0.892	1.00	1.00
0.7	0.9	0.671	0.663	0.695	0.724	0.811	0.856	0.998	1.00
0.8	0.9	0.118	0.137	0.226	0.286	0.526	0.586	0.623	0.685

Table 6: Comparison of power for the proposed test and that of Braun and Alonzo in terms of their approximations with different area and non-crossing ROC curve

		p = 20, q = 20		p = 40, q = 40		p = 60, q = 60		p = 80, q = 80	
		ρ = 0.0		ρ = 0.25		ρ = 0.5		ρ = 0.75	
AUC$_1$	*AUC*$_1$	B & A	MWSRT	B & A	MWSRT	B & A	MWSRT	B & A	MWSRT
0.6	0.7	0.076	0.068	0.081	0.081	0.088	0.093	0.119	0.201
0.6	0.8	0.142	0.129	0.184	0.180	0.239	0.244	0.612	0.613
0.7	0.8	0.261	0.245	0.368	0.352	0.469	0.475	0.920	0.921
0.7	0.9	0.414	0.391	0.568	0.562	0.711	0.715	0.985	0.985
0.8	0.9	0.429	0.421	0.589	0.589	0.702	0.725	0.994	0.994
0.6	0.7	0.076	0.071	0.081	0.082	0.090	0.096	0.219	0.222
0.6	0.8	0.153	0.139	0.198	0.198	0.256	0.263	0.665	0.668
0.7	0.8	0.288	0.270	0.393	0.389	0.510	0.520	0.952	0.952
0.7	0.9	0.466	0.446	0.619	0.616	0.767	0.771	0.987	0.987
0.8	0.9	0.479	0.466	0.634	0.635	0.787	0.786	0.996	0.998
0.6	0.7	0.077	0.070	0.084	0.090	0.096	0.107	0.252	0.258
0.6	0.8	0.169	0.159	0.226	0.230	0.284	0.300	0.745	0.748
0.7	0.8	0.330	0.315	0.450	0.450	0.572	0.589	0.978	0.980
0.7	0.9	0.546	0.538	0.702	0.702	0.828	0.838	0.989	0.989
0.8	0.9	0.526	0.523	0.332	0.419	0.857	0.847	0.999	0.999
0.6	0.7	0.082	0.078	0.092	0.097	0.102	0.127	0.309	0.316
0.6	0.8	0.145	0.135	0.263	0.271	0.336	0.364	0.845	0.846
0.7	0.8	0.220	0.208	0.347	0.374	0.465	0.487	0.976	0.976
0.7	0.9	0.117	0.104	0.204	0.451	0.516	0.846	0.997	0.997
0.8	0.9	0.103	0.116	0.123	0.263	0.330	0.417	0.636	0.638

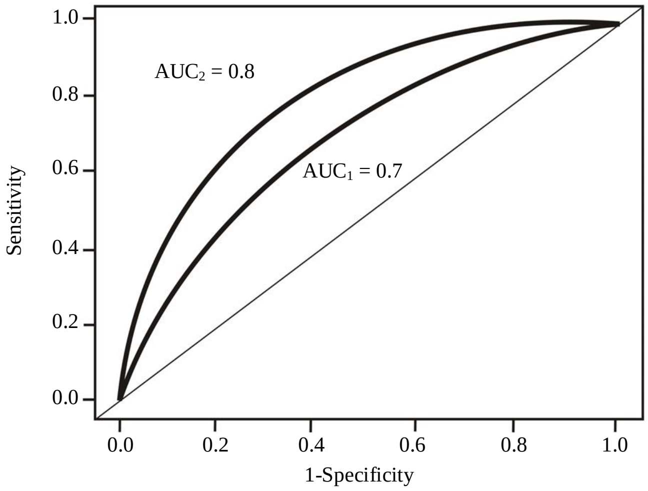

Fig. 1: Different AUCs with not-crossing ROC curves

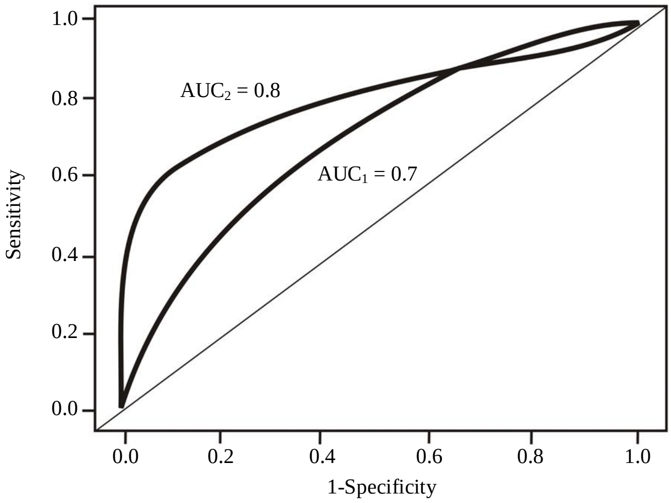

Fig. 2: Different AUCs with crossing ROC curves

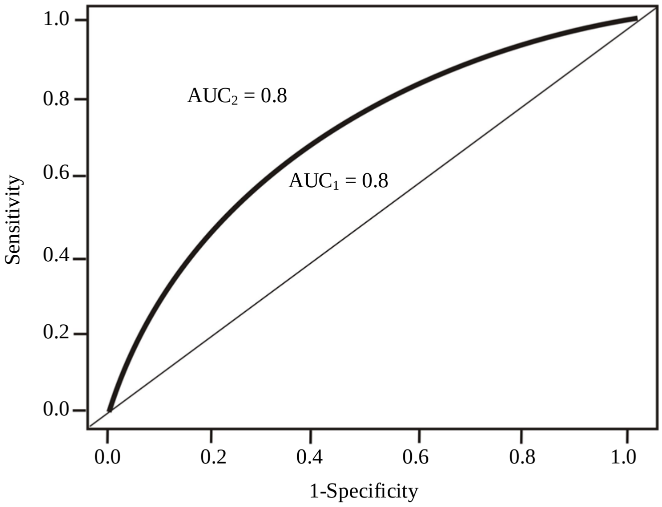

Fig. 3: The same AUCs with non-crossing ROC curves

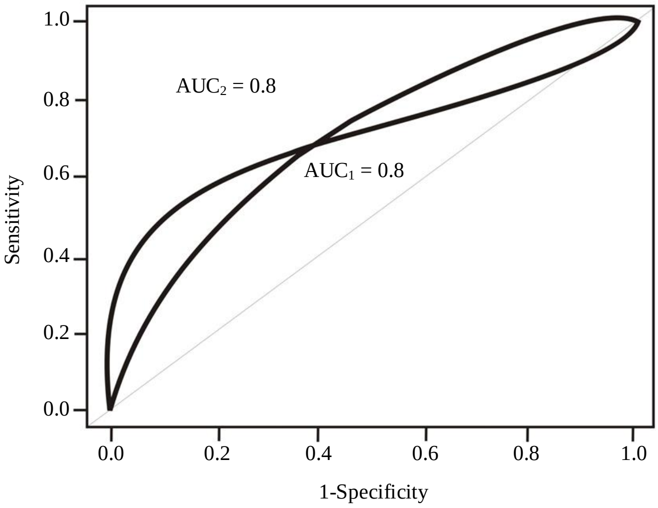

Fig. 4: The same AUCs with crossing ROC curves

In Table 5 and 6 when the different AUC is at least 0.8 with a correlation of $\rho\geq0.4$ having crossing and non-crossing ROC curves respectively, the proposed permutation test has greater statistical power compared to the test proposed by Braun and Alonzo (2008). This is because the proposed permutation test is less conservative in the stated range of parameters. When the correlation is less than 0.4 with different AUCs less than 0.8, Braun and Alonzo's test has slightly greater statistical power because at this region they test size is slightly high. As sample size increases, the operating characteristics of the two permutation tests near one another.

Therefore, in summary our simulations showed for the proposed permutation test the test size and nominal level of significance are in close agreement for sample sizes that are reasonably small. Again, for sample sizes that are small with large AUCs and moderate correlation between diagnostic tests the proposed test has operating characteristics that is better than the permutation test proposed by Braun and Alonzo (2008). Finally, the statistical power of the proposed permutation test to detect crossing ROC curves with same AUCs is near to the nominal level of significance. This means that for crossing of ROC curves to be detected, the AUCs of the two curves must be different under the range of parameters considered. The Test size and statistical power of each test were computed as the percentage of 10,000 simulations and the null hypothesis of $AUC_\Delta = 0$ was rejected at a nominal significant level of 0.05.We generated the permutation of the empirical probability distribution of $\hat{AUC}_\Delta$ in each simulation by generating 10,000 random permutations of the diseased and non-diseased labels.

b. Real Life Data Example

By simple random sampling method, a total of 60 pregnant women underwent two types of diagnostic tests for the in-depth confirmation of Gestational Diabetic Mellitus (GDM) such that their test results were paired or matched to each other. These diagnostic tests are a 75 g Oral Glucose Tolerance Test (OGTT) and a 100 g OGTT. The data is used to evaluate the feasibility of the proposed permutation test at a nominal level of 0.05. The characterization and criteria adopted for diagnosing antenatal mothers who underwent either 75 g OGTT/100 g OGTT were 2 h OGTT characterization while the criteria was ≥ 155 mg/dl for one to be considered diseased/positive (coded 1) for GDM while <155 mg/dl is considered non-diseased/negative (coded 0) for GDM. Exchangeability of the measured test results is a vital condition to achieve result given that these results are paired. If the null hypothesis is true, then we can infer that the subjects' test results in diagnostic 1 and 2 are exchangeable and so the permutation test is applied on raw scores and are not ranked. It showed that there exist a number of pairs with tied test results, even though the test results are continuous. The null hypothesis is that the 2 h 75 g OGTT contributes the same diagnostic information or accuracy as the 2 h 100 g OGTT. That is, AUC_1 and AUC_2 of the two diagnostic tests are equal. The real data if analyzed will evaluates the performance of the proposed estimates. It will compare the performance of the two diagnostic tests in terms of ROC curves between the two diagnostic tests and a crossing ROC curve will emerge (Fig. 5). The crossing ROC curves will have the areas for the two diagnostic test procedures (Fig. 5). In applying the data, the diagnostic test results need to have a bivariate bi-normal distribution. But according to Wang (2015), most powerful test does not exist for testing bivariate normal distribution. Therefore, for each test result, one resorted to checking only the univariate normality.

Checking for univariate normality of two diagnostic test results by Shapiro-Wilk test reveals that the p-values for the diagnostic tests 1 and 2 for the non-diseased subjects are respectively 0.6124 and 0.8975 while that of diseased subjects for the diagnostic tests 1 and 2 are respectively 0.6345 and 0.8765. The estimates of AUC_1 and AUC_2 for diagnostic tests are 0.668 and 0.887 respectively. Hence using the proposed permutation test, the p-value of 0.0312 is rejected at a nominal level of 0.05. Using the Braun and Alonzo's permutation test, the null hypothesis is also rejected since the P-value is 0.0387.

Discussion

The proposed permutation test can be used to compare the performances of diagnostic tests for paired sample design. It makes for the conduct of exact permutation test and makes for easy to implement approximation when the sample size is large. Our test which is used in testing the null hypotheses about paired ROC curves (in other words, the equality of AUCs) is designed to have increased power to detect a difference in the AUC. The need for an alternative permutation test based on between-subject permutations of the labels of the subjects within each diagnostic test for detecting differences between ROC curves was necessary so as to tackle the problem associated with few existing methods which is characterized by the exchangeability of the labels between two diagnostic tests within subject. In the real sense of it, the proposed test is for assessing a change in the AUCs in a continuous matched pair of data from two diagnostic test procedures having both diseased and non-diseased subject in each of the test. Here permutations are made between subjects particularly by shuffling the diseased and non-diseased labels of the subjects within each diagnostic test procedure. According to DeLong *et al*. (1988), the condition for having appropriate test size and increased statistical power stipulates the following: That the sample size for both the non-diseased and diseased subjects must not be more than 60, the average of two correlated AUCs must be at least 0.80 as well as the fact that the correlation within subjects test results is $\rho \geq 0.4$. At small average AUC, low correlation between diagnostic tests and at sample size higher than 60, the method by DeLong *et al*. (1988) has improved test size and greater or higher power than our test but these does not apply here where there is evaluation involving diagnostic tests more so when permutation test is required. For small sample sizes, the proposed permutation test and that of Braun and Alonzo have similar test size and statistical power. According to the simulation conducted by Venkatraman and Begg (1996), for non-crossing ROC curves, the statistical power of DeLong *et al*. has a higher power than that of Venkatraman and Begg. This is because the procedure of Venkatraman and Begg is designed to detect differences in ROC curves as against detecting differences only in AUCs. In other words, when ROC curves cross, the power of test is higher because it detects difference in ROC curves but if roc curves do not cross, DeLong *et al.*'s test that compare AUCs only have higher power. Therefore, Venkatraman and Begg (1996) test has lower power for non-crossing ROC curves as it detect differences in ROC curves while in such scenario, DeLong *et al.*'s test has higher power as it detects differences in AUCs. Our permutation test though tests the null hypothesis of equality of AUCs, it is designed to detect a difference in AUC as it compares the correlation in ROC curves when the ROC curves cross each other. While our permutation test formally tests a difference in ROC curves and detects a difference in AUC, it has higher power than DeLong *et al.*'s conventional test that only detects difference in AUCs. Result showed that our proposed test has comparable power to the test conducted by Bandos (2005) as well as Braun and Alonzo (2008) but has superior operating characteristics in some ranges of parameters as well as due to the fact that our test is designed to consider the value of signs as well as the absolute ranks of values as well while the test by Braun and Alonzo considered only the signs of values. However, the test by Venkatraman and Begg would have been a better option for use assuming our primary interest was to detect a difference in ROC curves at every operating point. In all our simulation result shows that our permutation test is slightly conservative but has an excellent power to detect a crossing alternative. The test size of the permutation test for sample sizes that are small was investigated using simulations. The algorithm for calculating the exact permutation distribution of $A\hat{U}C_{\Delta}$ enabled us to obtain a normal approximation to the exact procedure and this is suitable when the sample size is small. The presence of an asymptotic method provides a simple and exact approximation to the permutation test since exact permutation tests can be computationally burdensome if sample size increases.

Summary and Conclusion

The Test size and statistical power of each test were computed as the percentage of 10,000 simulations and the null hypothesis of $AUC_{\Delta} = 0$ were rejected at a nominal level of 0.05. Because the proposed permutation test is formally for testing the null hypothesis of equality of AUC, the rejection rate becomes the statistical power when the ROC curves cross each other. If the sample size is moderate and more especially for small sample sizes in a case of non-crossing ROC curves having equivalent and large AUC given the fact that the correlation between the diagnostic tests are moderate, the test size demonstrated by the proposed test is less conservative than the Braun and Alonzo test. In practical terms, it is not advisable to employ the proposed test in detecting crossing ROC curves when the AUCs from crossing ROC curves are equal because its rejection rate, talking the power is very close to that of Braun and Alonzo test (type I error). The proposed test makes provision for an approximate test of equality of AUCs due to the fact that the rejection rate is very close to the given level of significance. The power of the proposed test is greater than that of Bandos *et al*. as well as Braun and Alonzo's test if the correlation is at least 0.4 and the average of AUC is at least 0.80 for non-crossing ROC curves since the range of parameters of the proposed test is less conservative. The power of Braun and Alonzo's test is greater when the correlation is lower and the average AUCs is smaller than this, a situation seen at a region where the test size test of this competitive test is slightly elevated. As the sample size increases, the operating characteristics of these comparative tests get closer to each other. In particular, when the ROC curves cross, the rejection rate of the proposed test is higher when the correlations and average of AUCs are higher. Therefore, our simulations shows that the test size of the proposed test and the nominal value shows close agreement when the sample size is reasonably small. In addition, the proposed permutation test has better operating characteristics when the correlation between diagnostic tests is moderate at large average AUC and small sample sizes than Bandos *et al*. as well as Braun and Alonzo's tests. So the proposed test has power close to the significance level in detecting when ROC curves cross with equal AUCs within the range of parameters considered. This means that for the null hypothesis to be rejected, the AUCs of the two ROC curves must differ. We presented various Tables of comparisons of test size and statistical power of the proposed permutation test and that of the competing test in an effort to assess a difference in the AUCs of two diagnostic tests. In applying the proposed test on real data, we saw in the graph of ROC curves Fig. 5 that 2 h 100 g OGTT diagnostic test is superior at a time that the specificity is greater than 0.7. As soon as the specificity decreases, the disparity between the two diagnostic tests procedures reduces. In applying the proposed permutation test, the diagnostic test results need to have a bivariate bi-normal distribution. But according to Wang (2015), most powerful test does not exist for testing bivariate normal distribution. Therefore, for each test result, one resorted to checking only the univariate normality. Checking for normality of two diagnostic test results by Shapiro-Wilk test reveals that the P-values for the diagnostic tests 1 and 2 for the non-diseased subjects are respectively 0.6124 and 0.8975 while that of diseased subjects for the diagnostic tests 1 and 2 are respectively 0.6345 and 0.8765. Therefore, the null hypothesis for this univariate normal is rejected that the two diagnostic test procedures did not contribute similar information or that their accuracies are not the same. Hence using the proposed permutation test, the P-value of 0.0312 is rejected at a nominal level of 0.05. Using the Braun and Alonzo's permutation test, the null hypothesis of $AUC_{\Delta} = 0$ is rejected also since the P-value is 0.0387. Comparing the proposed test and that of Braun and Alonzo's permutation test in terms of their P-values, one will say that the proposed test is more powerful since it has the more likelihood of rejecting the null hypothesis. These results are consistent with the findings obtained by the proposed permutation test by Bandos *et al*. (2005).

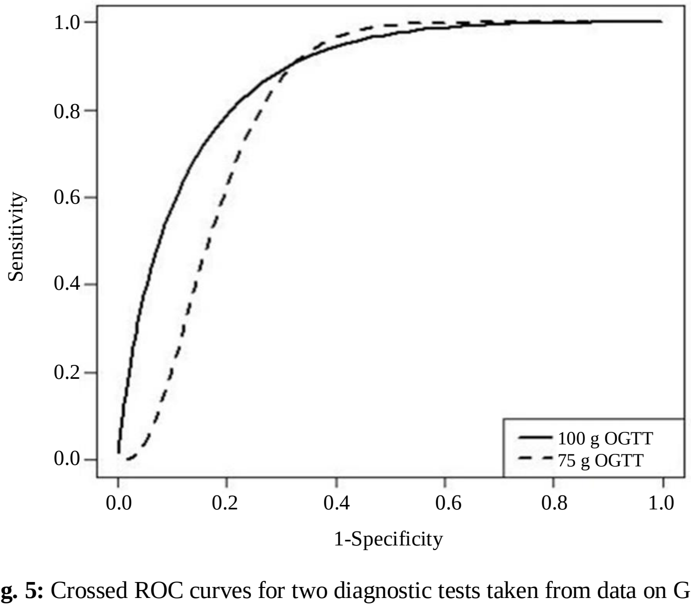

Fig. 5: Crossed ROC curves for two diagnostic tests taken from data on GDM

We therefore recommend the use of permutation tests for comparing two diagnostic tests that are correlated as it provides a more exact results with small sample sizes which is the demand of clinical practices. We suggest the use our proposed permutation test to generate a confidence interval for AUC_Δ as a complement to the hypothesis test as well as how permutation method can be applied if the test statistic is seen as McNemar test. It is vital to consider the use of a test statistic that will consider the use of absolute ranks as well as absolute magnitude of a test statistic that discriminates between the null hypothesis and alternative hypothesis. Under the present scenario, Wilcoxon signed-ranks test, which is our permutation test equivalent to AUC_Δ only use the absolute rank of Q_{pq} and not its absolute magnitude. Future study includes extending the proposed test to accommodate the “multiple-reader” setting – a commonly used design in which so many readers evaluate selected cases using different diagnostic tests.

Acknowledgement

I wish to acknowledge and appreciate the effort of Dr.C. H. Nwankwo for finding time to read through this work and make vital remarks where necessary.

Authors Contributions

Okeh Uchechukwu Marius: Proposed the test, searched for relevant literatures, collected the data and carried out the data analysis.

Onyeagu Sidney: He gave a suitable title to the work, organized the entire work and supervised the work to ensure that it is of a good standard and proof read the work for necessary corrections and finally approved it to be sent out for publication.

Ethics

The authors are liable to the materials, data and references used in this work.

Reference

Bandos, A., 2005. Nonparametric methods in comparing two ROC curves. Doctoral dissertation, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh.

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. Stat. Med., 24: 2873-2893. DOI: 10.1002/sim.2149

Braun, T.M. and T.A. Alonzo, 2008. A modified sign test for comparing paired ROC curves. Biostatistics, 9: 364-372. DOI: 10.1093/biostatistics/kxm036

Campbell, G., 1994. General methodology I: Advances in statistical methodology for the evaluation of diagnostic and laboratory tests. Stat. Med., 13: 499-508. DOI: 10.1002/sim.4780130513

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: 837-846. DOI: 10.2307/2531595

Dorfman, D.D., K.S. Berbaum and C.E. Metz, 1992. Receiver operating characteristic rating analysis: Generalization to the population of readers and patients with the jackknife method. Invest. Radiol., 27: 723-731. DOI: 10.1097/00004424-199209000-00015

Hanley, J.A. and B.J. McNeil, 1982. The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve. Radiology, 143: 29-36. DOI: 10.1148/radiology.143.1.7063747

Hoeffding, W., 1948. A class of statistics with asymptotically normal distribution. Annals Math. Stat., 19: 293-325. DOI: 10.1214/aoms/1177730196

Moise M., 1988. Interactive system for automatic generation of technologies specific to splinting processing, cybernetics and technical scientific revolution. Publishing House of the Romanian Academy, Bucharest.

Oyeka, C.A., 2009. An Introduction to Applied Statistical Methods. 8th Edn., Nobern Avocation Publishing Company, Enugu, Nigeria, ISBN-13: 978-2457-6-7.

Pardo, M.C. and A.C. Franco-Pereira, 2017. Non parametric ROC summary statistics. Stat. J., 15: 583-600.

Song, H.H., 1997. Analysis of correlated ROC areas in diagnostic testing. Biometrics, 53: 370-382. DOI: 10.2307/2533123

Venkatraman, E.S. and C.B. Begg, 1996. A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment. Biometrika, 83: 835-848. DOI: 10.1093/biomet/83.4.835

Wang, C.C., 2015. A MATLAB package for multivariate normality test. J. Stat. Comput. Simulat., 85: 166-188. DOI: 10.1080/00949655.2013.808638

Appendix A1.

An Algorithm for Calculating the Exact Distribution of $\hat{AUC}_\Delta$

```
1: for $s_1 \leftarrow 1,2$ do
```

```
  $A\hat{U}C_{\Delta 1} \leftarrow S_{1,s_1}$
  if $s_1 \leftarrow 1$ then
    $T_1 \leftarrow S_{1,2}$
  else
    $T_1 \leftarrow S_{1,1}$
  end if
  for $s_2 \leftarrow 1,2$ do
    $A\hat{U}C_{\Delta 2} \leftarrow S_{2,s_2}$
    if $s_2 \leftarrow 1$ then
      $T_2 \leftarrow S_{2,2}$
    else
      $T_2 \leftarrow S_{2,1}$
    end if
    for $s_3 \leftarrow 1,2$ do
      $A\hat{U}C_{\Delta 3} \leftarrow S_{3,s_3}$
      if $s_3 \leftarrow 1$ then
        $T_3 \leftarrow S_{3,2}$
      else
        $T_3 \leftarrow S_{3,1}$
      end if
      for $s_4 \leftarrow 1,2$ do
        $A\hat{U}C_{\Delta 4} \leftarrow S_{4,s_4}$
        if $s_4 \leftarrow 1$ then
          $T_4 \leftarrow S_{4,2}$
        else
          $T_4 \leftarrow S_{4,1}$
        end if
        for $s_5 \leftarrow 1,2$ do
          $A\hat{U}C_{\Delta 5} \leftarrow S_{5,s_5}$
          if $s_5 \leftarrow 1$ then
            $T_5 \leftarrow S_{5,2}$
          else
            $T_5 \leftarrow S_{5,1}$
          end if
          …………..
          for $s_{60} \leftarrow 1,2$ do
            $A\hat{U}C_{\Delta 60} \leftarrow S_{60,s_{60}}$
            if $s_{60} \leftarrow 1$ then
              $T_{60} \leftarrow S_{60,2}$
            else
              $T_{60} \leftarrow S_{60,1}$
            else if
              Compute $A\hat{U}C_{\Delta}$
          end for
        end for
      end for
    end for
  end for
  …………..
end for
```