East African Scholars Journal of Engineering and Computer Sciences
Abbreviated Key Title: East African Scholars J Eng Comput Sci
ISSN 2617-4480 (Print) | ISSN 2663-0346 (Online) |
Published By East African Scholars Publisher, Kenya
OPEN ACCESS
Volume-2 | Issue-12 | Dec-2019 | DOI:10.36349/EASJECS.2019.v02i12.013

Research Article

An Algorithm for Computing the Exact Distribution of Modified Wilcoxon Signed Rank Test Statistic

Okeh UM[1] and Onyeagu Sidney I[2]

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

*Corresponding Author
Okeh UM

Abstract: This paper proposes an algorithm for the computation of exact probability distribution of the modified Wilcoxon signed-rank test. The implementation of the exact permutation algorithm will help in carrying out complete enumeration of all possible distinct rearrangements that is required without just sampling without replacement from the permutation sample space. The method is however suitable when samples are paired and can be adopted for all such situations requiring complete enumeration of all distinct permutations thus producing exact p-values which ensures that the probability of a type I error is exactly α. An extensive simulation study was carried out to compare the exact permutation and the asymptotic normal approximation of two competing permutation tests in terms of their type I error and statistical power. The algorithm is capable of breaking down completely all the problem associated with the permutation to ensure easy implementation and the required analysis. The algorithm was implemented in Intel Visual Fortran to compare the performances of two diagnostic test procedures on gestational diabetic mellitus.

Keywords: Wilcoxon signed-rank test, diagnostic test.

1. INTRODUCTION

Permutation tests (randomization test), are nonparametric statistics often used for statistical inferences about AUC. They are often associated with the early works of Fisher (1935) and are specific to hypothesis testing. A permutation test constructs a permutation sample space, which consists of equally likely permutation sample points created by interchanging the test results of the sample which are assumed to be "exchangeable" under the null hypothesis. Therefore, the permutation sample space is the exact probability space of the possible arrangements of the sample data under the null hypothesis given the original sample. For instance, when comparing two diagnostic tests having paired data, permutation tests here consists of exchanging the paired test results. With permutation tests, we randomly redistribute the overall test results into two groups of N nondiseased and M diseased subjects' labels, and calculate a test statistic of interest. The type of sampling involved in this reshuffling of the labels of subjects is called sampling without replacement. When this reshuffling or exchange happens say 1000 or 10,000 times, we generate a distribution of test results for the test statistic of interest under the null hypothesis of equality of no difference between the two sampled results from two populations of interest. Permutation tests provide exact distribution results when complete enumeration is possible. Permutation tests are generally confronted with the problems of high demand for space and time complexity during computation.

Good (2000) Summarized Five Steps for A Permutation Test As Follows:

1. Analyze the given problem.
2. Make choose of a test statistic and establish a rejection rule for distinguishing the null hypothesis from the alternative hypothesis.
3. Compute the test statistic for the original observations.
4. Rearrange the observations, compute the test statistic for every new arrangement and repeat this process until all permutations are obtained.
5. Construct the exact distribution for the test statistic based on Step 4. Step 4 is where the difficulty in permutation test lies because a complete

Quick Response Code

Journal homepage:
http://www.easpublisher.com/easjecs/

Article History
Received: 28.11.2019
Accepted: 08.11.2019
Published: 24.12.2019

enumeration of all distinct permutations of the experiment is required.

Good (2000) identified the sufficient condition for a permutation test to be exact and unbiased against shifts in the direction of higher values as the exchangeability of the paired sample test results or observed units.

In carrying out permutation tests involving diagnostic tests, Venkatraman & Begg (1996) proposed a method for detecting any differences at every operating point between two ROC curves. Similarly, Bandos *et al*., (2005) proposed a method that is sensitive to the difference in AUCs in diagnostic performance. These tests assume the same condition of exchangeability of the diagnostic test results under the null hypothesis, but differ in the sense that the permutation test by Bandos *et al*., has an easy-to-implement and precise approximation and better detects different ROC curves if they differ with respect to the AUC while Venkatraman and Begg (1996) aimed to increase the power to detect a crossing alternative. Specifically, Bandos *et al*., (2005) based their permutation test on the difference in areas and derived exact and asymptotic permutation test methods to test the equality of two correlated ROC curves which are designed to have increased power to detect difference in the AUC. The test of Bandos *et al*., (2005) directly tests for an equality of AUCs. This approach implicitly assumes that both diagnostic test procedures are exchangeable within subject and requires an appropriate transformation, such as ranks, for diagnostic test procedures differing in scale. Bandos *et al*., (2005) compared the performance of their test to that of DeLong *et al*., (1988) through 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. Bandos *et al*., (2005) test is limited by the fact that it requires the exchangeability of the diagnostic test procedures and do requires also the transformations of the original data if test results are measured on different scales. Therefore it requires diagnostic tests that are measured on identical scales. Therefore it is less powerful in settings in which the diagnostic test results are skewed since it requires diagnostic tests that are measured on identical scales (Braun and Alonzo, 2008).

Given the fact that clinical trials requires exact results and the purpose of matching in clinical trials is to increase the precision of the comparisons among the samples thus reducing variability among them, a suitable statistic for matched sample is inevitable. According to Harris and Hardin (2013), Wilcoxon signed rank test (WSRT) is a nonparametric test often used in clinical trials, in which it is common to have small samples and inferences about exact statistics are made. This is because large-sample results are not acceptable in many clinical trials studies. WSRT is the nonparametric counterpart to the two sample paired t test for paired samples. The test is based on the signed ranks of a random sample from a population which is continuous and symmetric around the median. This statistic uses the ranks of the absolute differences between the paired samples along with the sign of the difference. It uses the relative magnitudes of the data. This statistic can also be used to test for symmetry and to test for equality of location for paired samples.

2. PROPOSED METHOD

We wish to compare namely AUC_1 *and* AUC_2 which are respectively the AUCs of two diagnostic test procedures having a total number of n subjects. The procedure is such that a total number of N nondiseased subjects and M diseased subjects each received both diagnostic tests. Let the test results of diagnostic tests 1 and 2 for the nondiseased subject be X_{i1} *and* X_{i2} where $i = 1,...,\mathrm{N}.$ Also let the test results of diagnostic tests 1 and 2 for the diseased subject be Y_{j1} *and* Y_{j2} where $j = 1,...,\mathrm{M}.$ Also let $X = \{(X_{11},\mathrm{X}_{12}),(X_{21},\mathrm{X}_{22}),...,(X_{N1},\mathrm{X}_{N2})\}$ denotes pairs of vector of measurement on nondiseased subjects and let $Y = \{(Y_{11},\mathrm{Y}_{12}),(Y_{21},\mathrm{Y}_{22}),...,(Y_{M1},\mathrm{Y}_{M2})\}$ be the pairs of vector of measurement on diseased subjects. Note that $\hat{AUC}_\Delta$ *and* $\hat{W}$ are used interchangeably. Therefore the difference in AUCs given as $AUC_\Delta = AUC_2 - AUC_1$ is estimated nonparametrically as:

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

$$where\ Q(X_{im}, \mathrm{Y}_{jm}) = S_{ij2} - S_{ij1} = S_{ijm}\ and\ S_{ijm} = A(X_{im} < Y_{jm}) + \frac{1}{2}A(X_{im} = Y_{jm});\ m = 1,2.$$

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

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

$$S_{ijm} = \begin{cases} 1 \; if \;\; X_{im} < Y_{jm} \\ 0.5 \; if \;\; X_{im} = Y_{jm} \\ 0 \; if \;\; X_{im} > Y_{jm} \end{cases} \qquad 2$$
$$m = 1, 2.$$

In other to test the null hypothesis $H_0 : AUC_2 - AUC_1 = 0$, we combine $M \; and \; N$ subjects to have n subjects and let $S_1 = \{S_{11}, S_{12}, ..., S_{1N}, S_{1,N+1}, S_{1,N+2} ...., S_{1n}\}$ be n measurements arising from diagnostic test 1 while the subscripts $p = 1, 2, .., N$ shows test results for the nondiseased 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,

$$R_{pq1} = A(S_{p1} < S_{q1}) + \frac{1}{2} A(S_{p1} = S_{q1}); iff \;\; p \neq q \qquad 3$$

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

$$R_{pq2} = A(S_{p2} < S_{q2}) + \frac{1}{2} A(S_{p2} = S_{q2}); iff \;\; p \neq q. \qquad 4$$

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

To test the null hypothesis that $AUC_{\Delta} = 0$, 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) by mapping from the diagnostic pairs and disease status[0,1] to [1,0,-1].Given the specifications above, we generalize the estimate of AUC_{Δ} as

$$\hat{W} = \frac{1}{NM} \sum_{p=1}^{N} \sum_{q=1}^{M} iT_{pq} = \frac{1}{NM} \sum_{p=1}^{N} \sum_{q=1}^{M} T_{pq} r|Q_{pq}| \qquad 5$$

Where

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

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

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 nondiseased 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 nondiseased and subject test result q is diseased. It takes -1 if subject test result p is diseased and subject test result q is nondiseased. Values of 0 represents cut-offs at which both subject test results p and q are diseased or nondiseased. 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 $r\left(Q_{pq}\right)$ is a symmetric distribution around zero. We therefore test the null hypothesis that $AUC_{\Delta}=0$ by computing AUC_{Δ} for every permutation of T_{pq}, the signs of the rank of $\left|Q_{pq}\right|$. Given that our permutation of T_{pq} requires exchanging the labels of nondiseased 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/nondiseased 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.

3. EXACT PERMUTATION TEST

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}.

Let $S_1=\{S_{11},S_{12},...,S_{1N},S_{1,N+1},S_{1,N+2}....,S_{1n}\}\ and\ S_2=\{S_{21},S_{22},...,S_{2N},S_{2,N+1},S_{2,N+2},...,S_{2n}\}$ be n measurements arising from two diagnostic tests 1 and 2 respectively where the subscripts $p=1,2,...,N$ represents test results for the nondiseased subjects and $q=N+1,N+2,...,n$ representing test results for the diseased subjects, we consider AUC_{Δ} given in (equation 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 $W=\left(W_1,W_2,W_3,....,W_m\right)$ be m distinct values of the test statistic W. The probability distribution of the test statistic W under the null hypothesis is given by

$$P(W_l=w_o|H_o)=\sum_{k=1}^{f_l}(2^{-N+M})=f_l(2^{-N+M}), \qquad 6$$

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

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

And

$$\alpha=P(W_h\geq c|H_o)=(2^{-N+M})\sum_{l=h}^{m}f_l. \qquad 8$$

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

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

Since permuted subjects labels are represented by $S_1 \text{ and } S_2$ from diagnostic test 1 and 2 respectively, let $\{\theta_1, \theta_2, ..., \theta_n\}$ be a set of all distinct permutations resulting from $S_1 \text{ and } S_2$ 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 $s = 2, 3, ..., 2^n$, $n = N + M = sample\ size$
5. Now build the empirical cumulative probability distribution as

$$p_0 = \hat{p}(W \le W_s) = \frac{1}{2^n} \sum_{s=1}^{2^n} T(W_1 - W_s) \qquad 10$$

$$where\ T = \begin{cases} 1\ if\ W_1 > W_s \\ 0\ if\ W_1 = W_s \\ -1\ if\ W_1 < W_s \end{cases}$$

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.

4. Operating Characteristics

We have performed data simulations to investigate and compare test size (type I error) and statistical power of the proposed test and test of Braun and Alonzo (2008). Nondiseased and diseased test results of subjects from two diagnostic tests are modeled using binormal distribution as it provides flexibility and simulated. To achieve this, we apply binormal ROC model for simplicity and robustness (Hanley,1988).So, within the m^{th} diagnostic test, subjects test results are obtained from binormal distribution of nondiseased subjects as $X_i^m \xrightarrow{i.i.d} N\left(\mu_X^m, \sigma_X^m\right)$, and diseased subjects as $Y_j^m \xrightarrow{i.i.d} N\left(\mu_Y^m, \sigma_Y^m\right)$. Since sample of subjects for two diagnostic tests are paired, a correlation is introduced in the measurement of test results as $\left(Cov(X^1, X^2) = Cov(Y^1, Y^2) = \rho\right)$. Note that under the binormal model, the accuracy between two diagnostic tests can be evaluated by comparing two ROC curves with the capacity to detecting a difference in AUCs. The hypotheses of interest is $H_0 : AUC_1 = AUC_2 \text{ versus } H_1 : AUC_1 \ne AUC_2$.

The binormal ROC curve for the distribution of subjects test results within the m^{th} diagnostic test can be parameterized without actually transforming the data as

$$AUC_m = P\left(X^m < Y^m\right) \text{ and } b_m = \frac{\sigma_X^m}{\sigma_Y^m}$$

In other to model different patterns of correlation (ρ) between the paired test results of subjects and difference of AUCs as well as shapes of the ROC curve (***b***), we interchange the parameters of the distributions of the test results. For instance, to model non-crossing of ROC curves, we set ***b***=1 for the two diagnostic tests. In the same way, we set ***b***<1 in other to simulate for crossing ROC curves. This indicates greater changes among the test results of diseased subjects for one of the diagnostic tests. Several values of nondiseased (M) and diseased (N) subjects as well as the probability of diseased subjects are considered also. 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 nondiseased 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 $\mu_{x^m} = 0 \text{ and } \sigma^2_{x^m} = 1.0; m = 1, 2.$

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

Where Φ^{-1} is the percentile of standard normal distribution (Weiand *et al*., 1989). 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 noncrossing 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 assumed equal correlation across the test procedures for the test results of nondiseased and diseased subject measured on continuous scale thus, assuming the correlation for all the scenarios to be $\rho = 0.25, 0.5 \text{ and } 0.75.$ A total of 10,000 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 rejection region for the tests. The exact values are compared with the approximate 95% confidence interval around a nominal size of 0.05 is (0.036, 0.064) on the basis of 10,000 simulation in each case.

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

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

5. Simulation Results

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. We wish to know the test size (type I error) and statistical power of the asymptotic normal approximation and exact values of various AUCs that are involved as well as how correlated subjects' test results are across diagnostic tests at different sample sizes. Here equal correlations are assumed for nondiseased and diseased subjects across the two diagnostic test results for non-crossing and crossing of ROC curves. We compared the test size and power of the permutation test to the test by Braun and Alonzo(2008) in terms of their exact permutation and normal approximation. In comparing the test size and statistical power of the proposed test in relation to Braun and Alonzo(2008) method, a number of 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.

AUC_1	AUC_2	$\rho = 0.25$				$\rho = 0.50$				$\rho = 0.75$			
		MWSRT		**B & A**		**MWSRT**		**B & A**		**MWSRT**		**B & A**	
		EXACT	ASY	EXACT	ASY	EXACT	ASY	EXACT	ASY	EXACT	ASY	EXACT	ASY
0.6	0.7	.046	.045	.043	.036	.047	.043	.046	.044	.049	.044	.038	.035
0.6	0.8	.050	.047	.047	.043	.054	.050	.052	.050	.056	.050	.054	.050
0.7	0.8	.065	.063	.064	.060	.075	.068	.072	.071	.085	.079	.079	.074
0.7	0.9	.092	.088	.091	.087	.113	.107	.111	.110	.142	.132	.140	.135
0.8	0.9	.127	.122	.123	.120	.168	.160	.165	.164	.221	.204	.220	.220
0.6	0.7	.039	.036	.039	.034	.043	.038	.042	.038	.042	.038	.041	.040
0.6	0.8	.046	.045	.043	.049	.049	.045	.045	.043	.050	.045	.046	.045
0.7	0.8	.062	.059	.060	.057	.069	.064	.063	.060	.081	.073	.078	.077
0.7	0.9	.086	.083	.085	.082	.110	.102	.107	.105	.136	.124	.136	.129
0.8	0.9	.126	.120	.125	.122	.171	.159	.170	.170	.223	.201	.222	.220
0.6	0.7	.032	.030	.030	.026	.034	.038	.032	.032	.032	.030	.030	.026
0.6	0.8	.036	.034	.028	.023	.040	.042	.036	.034	.044	.042	.042	.041
0.7	0.8	.053	.050	.051	.047	.064	.072	.060	.060	.075	.072	.074	.073
0.7	0.9	.080	.075	.078	.073	.104	.122	.102	.100	.137	.132	.132	.130
0.8	0.9	.122	.115	.120	.118	.174	.179	.171	.172	.231	.228	.227	.217
0.6	0.7	.022	.020	.021	.020	.026	.031	.022	.020	.023	.021	.022	.022
0.6	0.8	.026	.023	.021	.018	.032	.040	.032	.031	.032	.029	.031	.031
0.7	0.8	.039	.035	.036	.034	.034	.037	.034	.032	.070	.057	.065	.063
0.7	0.9	.029	.023	.025	.022	.026	.024	.025	.022	.042	.039	.041	.040
0.8	0.9	.022	.019	.022	.017	.020	.017	.018	.015	.022	.020	.021	.018

Sample sizes of 10 for both nondiseased and diseased subjects were simulated.

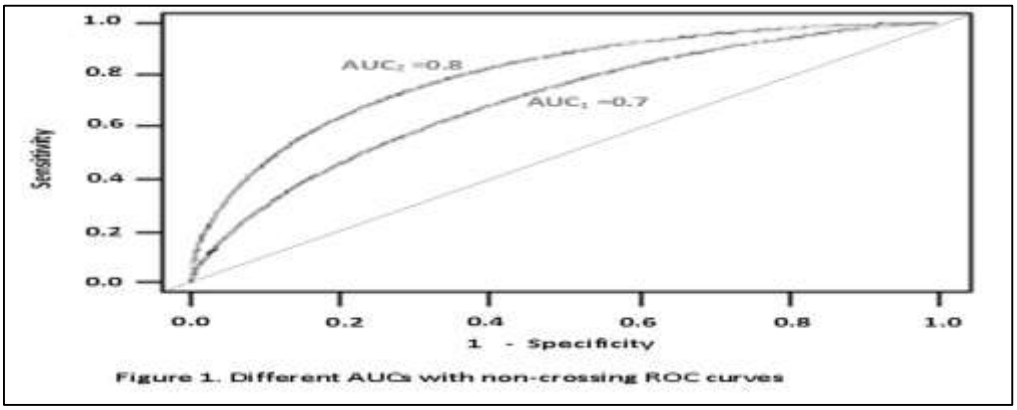

Figure 1. Different AUCs with non-crossing ROC curves

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.

AUC_1	AUC_2	$\rho = 0.25$				$\rho = 0.50$				$\rho = 0.75$			
		MWSRT		**B & A**		**MWSRT**		**B & A**		**MWSRT**		**B & A**	
		EXACT	ASY	EXACT	ASY	EXACT	ASY	EXACT	ASY	EXACT	ASY	EXACT	ASY
0.6	0.7	.050	.047	.048	.037	.053	.048	.048	.046	.052	.048	.050	.045
0.6	0.8	.054	.050	.050	.047	.058	.054	.055	.054	.061	.059	.057	.053
0.7	0.8	.068	.066	.064	.060	.080	.076	.076	.074	.090	.087	.086	.083
0.7	0.9	.097	.093	.093	.080	.120	.119	.116	.116	.142	.139	.141	.140
0.8	0.9	.132	.128	.131	.130	.174	.173	.173	.168	.218	.208	.215	.214
0.6	0.7	.042	.040	.040	.037	.045	.038	.042	.040	.045	.041	.044	.043
0.6	0.8	.046	.044	.044	.040	.050	.048	.045	.044	.053	.046	.053	.052
0.7	0.8	.065	.063	.065	.063	.075	.066	.072	.072	.083	.082	.080	.076
0.7	0.9	.094	.088	.093	.087	.115	.109	.115	.110	.141	.138	.138	.134
0.8	0.9	.136	.127	.134	.132	.178	.173	.176	.174	.224	.218	.222	.220
0.6	0.7	.036	.032	.034	.030	.037	.035	.033	.032	.040	.037	.038	.036
0.6	0.8	.040	.038	.037	.034	.045	.037	.043	.042	.046	.042	.036	.033
0.7	0.8	.058	.055	.055	.052	.069	.059	.064	.062	.082	.076	.075	.074
0.7	0.9	.087	.086	.085	.083	.112	.108	.112	.110	.140	.137	.138	.136

0.8	0.9	.129	.125	.126	.122	.182	.175	.189	.185	.232	.227	.230	.224
0.6	0.7	.026	.023	.023	.020	.026	.022	.025	.023	.027	.023	.025	.022
0.6	0.8	.029	.024	.027	.022	.035	.033	.034	.032	.038	.035	.033	.030
0.7	0.8	.044	.038	.043	.041	.060	.058	.058	.054	.071	.068	.070	.067
0.7	0.9	.073	.069	.072	.070	.104	.100	.102	.100	.141	.136	.135	.133
0.8	0.9	.022	.020	.019	.016	.039	.028	.037	.027	.034	.029	.028	.022

Sample sizes of 10 for both nondiseased and diseased subjects were simulated.

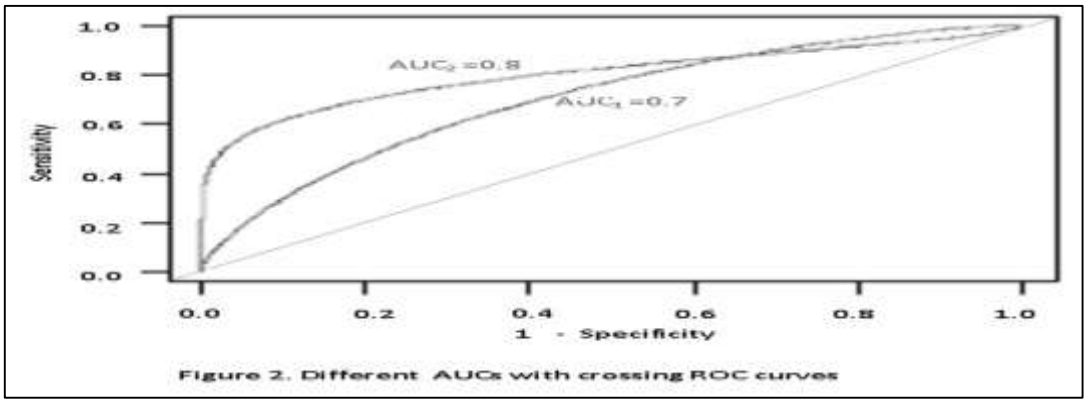

Figure 2. Different AUCs with crossing ROC curves

Tables 1 and 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 noncrossing and crossing ROC curves respectively. Since large computational time was needed for carrying out the computation of exact permutation, the comparisons shown in tables 1 and 2 are limited to sample sizes that are small where result indicates that good agreement exists between the exact and normal approximation test. Tables 1 and 2 shows that even with small sample size of 10 for each of nondiseased 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.

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.

ρ	AUC_1	AUC_2	$p=20, q=20$		$p=40, q=40$		$p=60, q=60$		$p=80, q=80$	
			B & A	**MWSRT**	**B & A**	**MWSRT**	**B & A**	**MWSRT**	**B & A**	**MWSRT**
0.0	.6	.6	.056	.049	.052	.049	.051	.050	.049	.047
	.7	.7	.052	.048	.050	.048	.051	.049	.048	.046
	.8	.8	.050	.046	.050	.048	.050	.049	.049	.048
	.9	.9	.039	..044	.048	.046	.048	.049	.048	.047
0.25	.6	.6	.053	.049	.052	.050	.053	.052	.053	.050
	.7	.7	.052	.049	.051	.050	.050	.048	.051	.050
	.8	.8	.048	.047	.049	.048	.050	.050	.050	.049
	.9	.9	.044	.045	.047	.048	.050	.050	.051	.049
0.5	.6	.6	.051	..050	.050	.050	.051	.050	.050	.048
	.7	.7	.048	.048	.050	.050	.049	.050	.047	.046
	.8	.8	.045	.046	.049	.050	.050	.051	.048	.046
	.9	.9	.041	.041	.047	.049	.050	.051	.049	.047
.75	.6	.6	.044	.047	.038	.042	.046	.046	.045	.046
	.7	.7	.043	.045	.037	.041	.043	.044	.042	.043
	.8	.8	.037	.041	.038	.040	.042	.045	.044	.046
	.9	.9	.025	.036	.035	.039	.037	.039	.035	.038

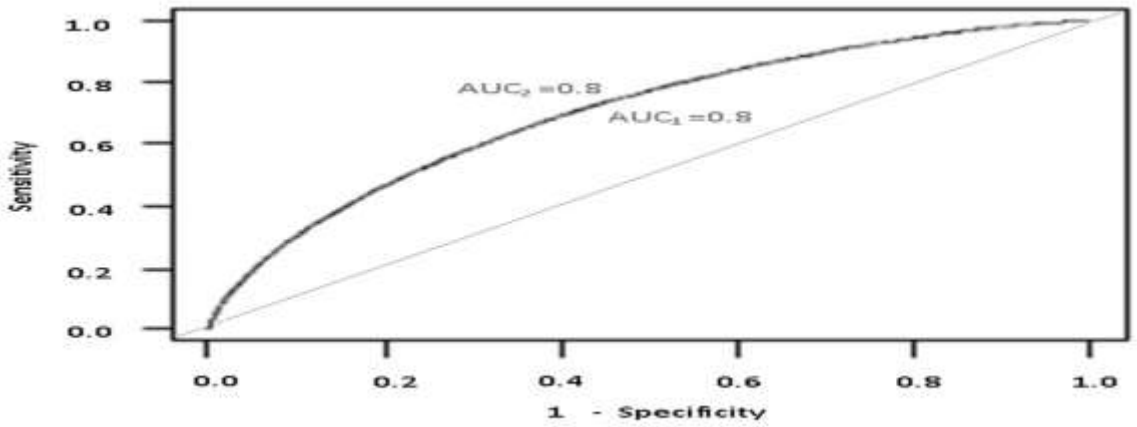

Figure 3. The same AUCs with non-crossing ROC curves

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.

ρ	AUC_1	AUC_2	$p=20, q=20$		$p=40, q=40$		$p=60, q=60$		$p=80, q=80$	
			B & A	**MWSRT**	**B & A**	**MWSRT**	**B & A**	**MWSRT**	**B & A**	**MWSRT**
0.0	.6	.6	.057	.054	.055	.054	.052	.052	.051	.051
	.7	.7	.055	.052	.054	.053	.052	.051	.048	.049
	.8	.8	.033	.037	.032	.035	.049	.050	.047	.046
	.9	.9	.020	.028	.021	.025	.045	.046	.044	.045
0.25	.6	.6	.054	.052	.053	.055	.051	.050	.052	.054
	.7	.7	.053	.052	.052	.053	.053	.054	.052	.053
	.8	.8	.040	.045	.050	.051	.050	.052	.049	.048
	.9	.9	.019	.023	.039	.043	.043	.044	.043	.044
0.5	.6	.6	.052	.054	.050	.052	.051	.053	.053	.054
	.7	.7	.050	.051	.049	.051	.050	.052	.052	.055
	.8	.8	.045	.047	.047	.049	.046	.049	.053	.054
	.9	.9	.020	.023	.034	.036	.037	0.040	.039	.040
.75	.6	.6	.047	.050	.050	.055	.050	.054	.051	.053
	.7	.7	.045	.048	.046	.049	.047	.050	.049	.050
	.8	.8	.037	.040	.037	.042	.038	.041	.040	.044
	.9	.9	.015	.024	.026	.035	.032	.039	.038	.040

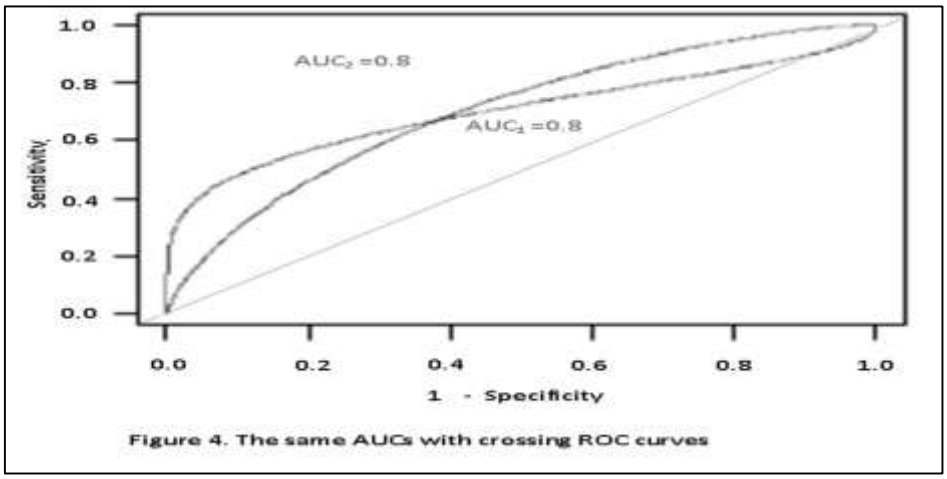

Figure 4. The same AUCs with crossing ROC curves

In Table 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 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 Tables 3 and 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 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 asymptotic approximations with different area and crossing ROC curves.

AUC_1	AUC_2	$p=20, q=20$		$p=40, q=40$		$p=60, q=60$		$p=80, q=80$	
		$\rho=0.0$		$\rho=0.25$		$\rho=0.5$		$\rho=0.75$	
		B & A	**MWSRT**	**B & A**	**MWSRT**	**B & A**	**MWSRT**	**B & A**	**MWSRT**
.6	.7	.076	.071	.082	.086	.090	.102	.180	.200
.6	.8	.142	.135	.179	.183	.213	.236	.544	.575
.7	.8	.251	.240	.332	.339	.422	.450	.880	.883
.7	.9	.403	.387	.535	.541	.655	.680	.937	.954
.8	.9	.476	.459	.566	.572	.656	.673	.996	.998
.6	.7	.079	.076	.087	.090	.092	.106	.197	.215
.6	.8	.154	.145	.194	.201	.232	.257	.593	.624
.7	.8	.277	.267	.366	.375	.459	.489	.914	.926
.7	.9	.452	.437	.587	.595	.705	.735	.983	.987
.8	.9	.537	.532	.612	.621	.822	.820	.995	.998
.6	.7	.084	.081	.093	.102	.101	.118	.275	.289
.6	.8	.174	.167	.221	.227	.265	.293	.777	.801
.7	.8	.323	.313	.423	.435	.524	.552	.979	.988
.7	.9	.531	.520	.623	.631	.874	.831	.993	1.00
.8	.9	.542	.535	.724	.753	.924	.953	.993	.994
.6	.7	.091	.088	.115	.135	.125	.162	.375	.406
.6	.8	.205	.202	.350	.386	.410	.480	.914	.923
.7	.8	.410	.401	.534	.542	.896	.892	1.00	1.00
.7	.9	.671	.663	.695	.724	.811	.856	.998	1.00
.8	.9	.118	.137	.226	.286	.526	.586	.623	.685

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

AUC_1	AUC_2	$p=20, q=20$		$p=40, q=40$		$p=60, q=60$		$p=80, q=80$	
		$\rho=0.0$		$\rho=0.25$		$\rho=0.5$		$\rho=0.75$	
		B & A	**MWSRT**	**B & A**	**MWSRT**	**B & A**	**MWSRT**	**B & A**	**MWSRT**
.6	.7	.076	.068	.081	.081	.088	.093	.119	.201
.6	.8	.142	.129	.184	.180	.239	.244	.612	.613
.7	.8	.261	.245	.368	.352	.469	.475	.920	.921
.7	.9	.414	.391	.568	.562	.711	.715	.985	.985
.8	.9	.429	.421	.589	.589	.702	.725	.994	.994
.6	.7	.076	.071	.081	.082	.090	.096	.219	.222
.6	.8	.153	.139	.198	.198	.256	.263	.665	.668
.7	.8	.288	.270	.393	.389	.510	.520	.952	.952
.7	.9	.466	.446	.619	.616	.767	.771	.987	.987
.8	.9	.479	.466	.634	.635	.787	.786	.996	.998
.6	.7	.077	.070	.084	.090	.096	.107	.252	.258
.6	.8	.169	.159	.226	.230	.284	.300	.745	.748
.7	.8	.330	.315	.450	.450	.572	.589	.978	.980

.7	.9	.546	.538	.702	.702	.828	.838	.989	.989
.8	.9	.526	.523	.332	.419	.857	.847	.999	.999
.6	.7	.082	.078	.092	.097	.102	.127	.309	.316
.6	.8	.145	.135	.263	.271	.336	.364	.845	.846
.7	.8	.220	.208	.347	.374	.465	.487	.976	.976
.7	.9	.117	.104	.204	.451	.516	.846	.997	.997
.8	.9	.103	.116	.123	.263	.330	.417	.636	.638

In Table 5 and 6 when the different AUC is at least 0.8 with a correlation of $\rho \geq 0.4$ 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 nondiseased labels.

6. Real Data Example and Results

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 75g Oral Glucose Tolerance Test (OGTT) and a 100g 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 75g OGTT /100g OGTT were 2hr OGTT characterization while the criteria was ≥ 155mg/dl for one to be considered diseased/positive (coded 1) for GDM while <155mg/dl is considered nondiseased/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 2hours 75g OGTT contributes the same diagnostic information or accuracy as the 2hours 100g 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. The crossing ROC curves will have the areas for the two diagnostic test procedures. In applying the data, the diagnostic test results need to have a bivariate binormal 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.

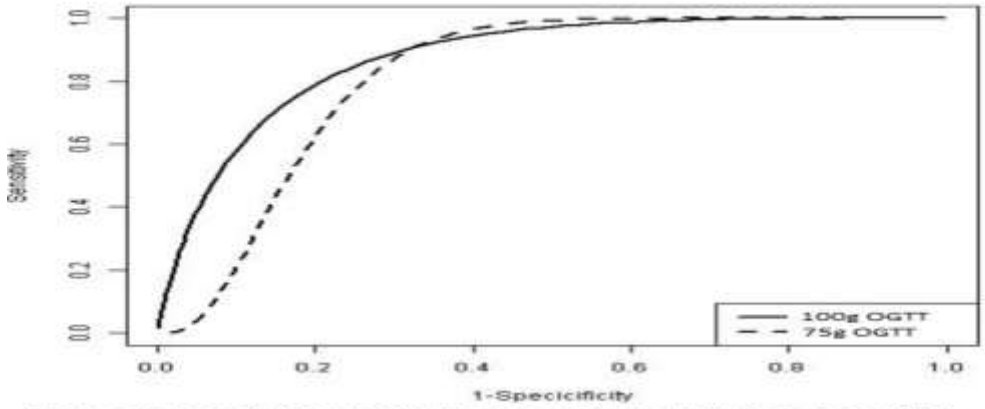

Figure 5. Crossed ROC curves for two diagnostic tests taken from data on GDM.

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 nondiseased 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.

7. DISCUSSION

The proposed permutation test compared the performances of two diagnostic tests for paired sample design. It conducted exact permutation test by implementing an algorithm derived for the purpose and derived an asymptotic normal approximation for large sample size based on proposed modified Wilcoxon signed rank test statistic. In comparing paired ROC curves, our design is to have increased power to detect a difference in the AUC. The proposed permutation test which is based on between-subject permutations of the labels of the subjects within each diagnostic test for detecting differences between ROC curves was necessary to tackle the problem of exchangeability of the labels between two diagnostic tests within subject. The proposed test is designed to assess a change in the AUCs in a continuous matched pair of data from two diagnostic tests having both diseased and nondiseased subject in each of the test where permutations are made between subjects particularly by shuffling the diseased and nondiseased labels of the subjects within each diagnostic test.

1. It will be recalled that DeLong *et al*., (1988) found that to have appropriate test size and increased statistical power, the necessary conditions are that the sample size for subject labels must be at most 60, the average of two AUCs must be at least 0.80 and the correlation within subjects test results should be at least 0.4. Therefore, 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 power than our permutation test otherwise permutation has improved test size and greater power.
2. Venkatraman and Begg(1996) found that for noncrossing ROC curves, the statistical power of DeLong *et al*., is higher than that of Venkatraman and Begg 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 a given test is higher because it detects difference in ROC curves but if ROC curves do not cross, the test that compares only the equality of AUCs has higher power eg. DeLong *et al*., test. Therefore, Venkatraman and Begg (1996)test has lower power for noncrossing ROC curves as it detect differences in ROC curves while in the same scenario, DeLong *et al*., test has higher power as it detects differences in AUCs.
3. 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.
4. Result showed that our proposed test has comparable power to the test conducted by Bandos (2005) as well as Braun and Alonzo (2008) who also proposed permutation tests but has superior operating characteristics in some ranges of parameters owing to the pattern of between subjects permutations as well as the fact that our proposed test is designed to consider the signs of values as well as the absolute ranks of values. Braun and Alonzo (2008) considered only the signs of values. Our permutation test is slightly

conservative but has an excellent power to detect a crossing alternative based on simulation results.

5. The algorithm for calculating the exact permutation distribution of $\hat{AUC}_{\Delta}$ enabled us to obtain a normal approximation to the exact procedure for small sample size. The presence of an asymptotic normal approximation method provides a simple and exact approximation to the permutation test for large sample size.
6. Using the real data to illustrate the feasibility of the proposed permutation test showed that the null hypothesis of equality of diagnostic information is rejected on account of one diagnostic test showing superiority over another and the proposed test showing higher power over existing tests. These results are consistent with the findings obtained by the proposed permutation test by Bandos *et al*., (2005) as well as Venkatraman and Begg(1996).
7. The problem with permutation tests has been high computational demands, viz; space and time complexities. Available permutation procedures can sample from the permutation sample space rather than carrying out complete enumeration of all possible distinct permutations. These available procedures cannot avoid the possibility of drawing the same sample more than once, thereby reducing the power of the permutation test, see Opdyke(2003). This study formulates and implements a sure way of obtaining exact permutation distribution of paired observations by ensuring that a complete enumeration of all the distinct permutations is achieved. This produces exact p-values and ensures that the probability of a type I error is exactly α. The algorithm can be extended to any sample size, depending on the processor speed and memory space of the computer being used to implement the algorithm.

1. SUMMARY AND CONCLUSIONS

1. With two diagnostic tests having different AUCs with non-crossing ROC curves, DeLong *et al*., area test which tests for the equality of AUC would have been a good option to have better operating characteristics with little conservative measures of test size (Type I error) than any other test such as Venkatraman and Begg (1996) that considers testing to detect difference in ROC curves.
2. Since most of the diagnostic researches yield matched data, it is important to take into account the correlated nature of the diagnostic tests to reduce variability among values.
3. When two diagnostic tests have crossing ROC curves with same or different AUCs, our proposed test in considering accessing a difference in AUCs have better operating characteristics with little conservative measures of test size than the Braun and Alonzo's test of (2008). Based on simulation result, our permutation test is slightly conservative and has an excellent power to detect a crossing alternative.
4. The p-value obtained through the exact permutation approach is exact. This process can be difficult because of its computational intensive. For small sample sizes, the exact permutation distribution of a test statistic and its asymptotic equivalent can be quite discrepant.
5. A proposed algorithm was implemented in Intel Visual FORTRAN for computing the exact distribution of the paired test results of two diagnostic tests by carefully enumerating all the distinct permutations of the test results. The permutation algorithm presented in this study beats the limitations and difficulties inherent in the exact permutation approach that probably led to the introduction of other approximate methods, which do not truly provide the exact distribution of a test statistic.
6. Since the proposed permutation test formally test the null hypothesis of the equality of AUC, the rejection rate otherwise called test size becomes the statistical power when the ROC curves cross each other.
7. For small and moderate sample sizes with same and large AUC as well as for moderate correlation between the diagnostic tests with non-crossing ROC curves, the test size shown by the proposed test is less conservative than the Braun and Alonzo test. This means that it is not advisable to employ the proposed permutation test in detecting crossing ROC curves when its AUCs are the same because its test size, talking about the power is very close to that of Braun and Alonzo test(type I error).
8. The proposed permutation test makes provision for an approximate test of equality of AUCs due to the fact that the test size is very close to the given level of significance.
9. As the sample size increases, the operating characteristics of these comparative tests (proposed test as well as Braun and Alonzo,2008) get closer to each other. In particular, when the ROC curves cross, the test size or 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.
10. 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.
11. 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.

12. In applying the real data, when we compared the proposed test to Braun and Alonzo's permutation test in terms of their p-values, the proposed permutation test is more powerful since it has the more likelihood of rejecting the null hypothesis. Graph of ROC curves in figure 5 showed that 2 hours 100g 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. Also since the null hypothesis for the univariate normal is rejected given the disparity in the p-values of the diagnostic tests for nondiseased and diseased as well as the values of AUCs, the two diagnostic test procedures did not contribute equivalent diagnostic information. This shows that 2 hours 100g OGTT diagnostic test is more suitable for discriminating non-diseased from diseased subjects than 2 hours 70g OGTT diagnostic test procedure meaning that the two procedures come from different population.
13. Simulation study showed that the proposed test can be a very suitable alternative to the test by Braun and Alonzo (2008) that only consider the direction of values. An application to real data s*et al*so supports our claim.
14. Since the MWSR 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 test is that it has easy implementation to discriminate diagnostic test procedures even by non-statisticians.

We 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.

APPENDIX A1

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

$1: for\ s_1 \leftarrow 1,2\ do$
$2: W_1 \leftarrow S_{1,s_1}$
$3: if\ s_1 \leftarrow 1\ then$
$4: TR_1 \leftarrow S_{1,2}$
$5: else$
$6: TR_1 \leftarrow S_{1,1}$
$7: end\ if$
$8: for\ s_2 \leftarrow 1,2\ do$
$9: W_2 \leftarrow S_{2,s_2}$
$10: if\ s_2 \leftarrow 1\ then$
$11: TR_2 \leftarrow S_{2,2}$
$12: else$
$13: TR_2 \leftarrow S_{2,1}$
$14: end\ if$
$15: for\ s_3 \leftarrow 1,2\ do$
$16: W_3 \leftarrow S_{3,s_3}$
$17: if\ s_3 \leftarrow 1\ then$
$18: TR_3 \leftarrow S_{3,2}$
$19: else$
$20: TR_3 \leftarrow S_{3,1}$
$21: end\ if$
$22: for\ s_4 \leftarrow 1,2\ do$
$23: W_4 \leftarrow S_{4,s_4}$
$24: if\ s_4 \leftarrow 1\ then$
$25: TR_4 \leftarrow S_{4,2}$
$26: else$
$27: TR_4 \leftarrow S_{4,1}$
$28: end\ if$

$29: \ for\ s_5 \leftarrow 1,2\ do$
$30: W_5 \leftarrow S_{5,s_5}$
$31: if\ \ s_5 \leftarrow 1\ then$
$32: TR_5 \leftarrow S_{5,2}$
$33: else$
$34: TR_5 \leftarrow S_{5,1}$
$35: end\ if$
$36: .....$
$37: \ for\ s_{60} \leftarrow 1,2\ do$
$38: W_{60} \leftarrow S_{60,s_{60}}$
$39: if\ \ s_{60} \leftarrow 1\ then$
$40: TR_{60} \leftarrow S_{60,2}$
$41: else$
$42: TR_{60} \leftarrow S_{60,1}$
$43: else\ if$
$44: Compute\ the\ test\ \ statistic\ W$
$45: end\ \ for$
$46: end\ \ for$
$47: end\ \ for$
$48: end\ \ for$
$49: end\ \ for$
$50: .....$
$51: end\ \ for$

REFERENCE

1. Bandos, A. (2005). *Nonparametric methods in comparing two correlated ROC curves* (Doctoral dissertation, University of Pittsburgh).
2. Bandos, A. I., Rockette, H. E., & Gur, D. (2005). A permutation test sensitive to differences in areas for comparing ROC curves from a paired design. *Statistics in medicine*, *24*(18), 2873-2893.
3. Braun, T. M., & Alonzo, T. A. (2007). A modified sign test for comparing paired ROC curves. *Biostatistics*, *9*(2), 364-372.
4. DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 837-845.
5. Fisher, R.A. (1935). Design of experiments. Oliver and Boyd, Edinburgh.
6. Good, P. (2000). Permutation tests: a practical guide to resampling methods for testing.
7. Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29-36.
8. Hanley, J. A. (1988). The robustness of the" binormal" assumptions used in fitting ROC curves. *Medical decision making*, *8*(3), 197-203.
9. Harris, T., & Hardin, J. W. (2013). Exact Wilcoxon signed-rank and Wilcoxon Mann–Whitney ranksum tests. *The Stata Journal*, *13*(2), 337-343.
10. Opdyke, J. D. (2003). Fast permutation tests that maximize power under conventional Monte Carlo sampling for pairwise and multiple comparisons. *Journal of Modern Applied Statistical Methods*, *2*(1), 5.
11. Oyeka, C.A. (2009). An Introduction to Applied Statistical Methods 8th edition, Nobern Avocation Publishing Company, Enugu, Nigeria (ISSN 978-2457-6-7).
12. Pardo, M. C., & Franco-Pereira, A. M. (2017). Non parametric ROC summary statistics. *REVSTAT*, *15*(4), 583-600.
13. Venkatraman, E. S., & Begg, C. B. (1996). A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment. *Biometrika*, *83*(4), 835-848.
14. Wang, C. C. (2015). A MATLAB package for multivariate normality test. *Journal of statistical Computation and simulation*, *85*(1), 166-188.
15. Wieand, S., Gail, M. H., James, B. R., & James, K. L. (1989). A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika*, *76*(3), 585-592.