International Journal of Physical Sciences Research

Vol.6, No.1, pp.15-28, 2022

Print ISSN 2515-0391 (Print)

Online ISSN 2515-0405 (Online)

A Modified Wilcoxon Signed Rank Test for Comparing Roc Curves in A Matched Pair Design

[1]Okeh U.M. [2]Onyeagu S. I.

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’s email address:uzomaokey@ymail.com

Citation: Okeh U.M. and Onyeagu S. I. (2023) A Modified Wilcoxon Signed Rank Test for Comparing Roc Curves in A Matched Pair Design, *International Journal of Physical Sciences Research*, Vol.6, No.1, pp.15-29

ABSTRACT: *Receiver Operating Characteristic (ROC) analysis has been used as a popular technique of comparing the performance of two paired diagnostic tests data while the area under the curve (AUC) summarizes the overall activities between two ROC curves. In other to assess a difference in the AUCs of paired data as well as to tackle the problem of exchangeability of the labels between two diagnostic tests within subject which characterizes previous studies, we propose a modified Wilcoxon signed-rank test to accommodate the presence of tied absolute values of differences for assessing a difference in the AUCs in a continuous matched pair of data. This assessment is based on between-subjects permutations particularly by exchanging the non-diseased and diseased labels of the subjects within each diagnostic test procedure, thus validating the permutation test since test results for each diagnostic test can be taken on a different scale. We as well derive the asymptotic normal approximation to the permutation test. 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. 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 2 hours 100g Oral Glucose Tolerance Test (OGTT) is superior to* AUC_1 *(2 hours 70g OGTT) at a time that the specificity is greater than 0.7.*

KEY WORDS: permutation test, exchangeability, asymptotic normal approximation, two diagnostic test procedures, area under the ROC curve (AUC), Modified Wilcoxon signed rank test, receiver operating characteristic (ROC) curve

INTRODUCTION

For decades now, ROC analysis has been used as a popular technique of evaluating the performance or ability of a test to discriminate between alternative health status or the true state of subjects (Kummar and Indrayan, 2011). 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. DeLong, et al. (1988) developed a totally nonparametric approach to compare two correlated areas under the receiver operating characteristic curves (AUCs) of two diagnostic tests for paired samples of subjects by using the theory of generalized U statistics. Their test 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, thus working with very large number of observations, so that computational time could be long. However, Venkatraman & Begg (1996) proposed a permutation test for detecting any differences at every operating point between two receiver operating characteristic (ROC) curves. Similarly, Bandos, et al. (2005) also proposed a permutation test that is sensitive to the difference in AUCs in diagnostic performance.

These permutation 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. It also requires diagnostic tests that are measured on identical scales. In the same way, Braun and Alonzo (2008) proposed a modified rank test that does not require a data transformation as seen in Venkatraman and Begg (1996) but showed that the modified test has the same power as Bandos, et al. (2005) though the result obtained by Bandos, et al. (2005) may prove to be 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).The sign test proposed by Braun and Alonzo(2008) for comparing paired ROC curves only gives knowledge about the direction(signs of differences) but not the magnitude between sampled pairs.

Meanwhile, sign test uses relatively smaller information from the data to be tested when there exists reasonable number of zeros and tied observations and so the greater is the number of zeros difference, the greater is the information that will be lost as a result of reduction in the number of smaller size that will be examined. Applying the permutation procedure employed previously in the paper by Braun and Alonzo (2008), we shall propose a permutation test for determining a difference in AUCs in a paired sample design where both diseased and non-diseased subjects are each meant to be tested using two different diagnostic procedures. Here permutations are made between subjects within a particular diagnostic procedure by exchanging the labels of diseased and non-diseased subjects, thus validating the test since test results for each diagnostic test can even come from different scale of measurement. This permutation procedure provides an exact and powerful test for assessing a difference in AUC and as well allows developing a precise and easy-to-apply approximation given that the sample size is small. Our permutation test will be based on Wilcoxon signed rank (WSR) test. WSR test statistic utilizes both the magnitudes and signs of differences. WSR test is expected to be more powerful test than the sign test (Oyeka, 1990). The real advantage of the WSR test is robustness of efficiency (Kotz, Johnson and Read (1988). The essential assumptions for the WSR test are continuous and symmetric population distribution. There is need to modify WSR test statistic to accommodate tied absolute value of differences.

ESTIMATION OF AUC

Given two diagnostic tests having N non-diseased subjects and M diseased subjects, let $X^m \; and \; Y^m (m=1,2)$ represents the subjects that are non-diseased and diseased in the m^{th} diagnostic test respectively. Then $x_i^m = (i=1,2,...,N)$ and $y_j^m = (j=1,2,...,M)$ are respectively the corresponding bivariate test results for the two diagnostic tests with *N* non-diseased and *M* diseased subjects. Therefore the marginal $F_m\left(x^m\right), G_m\left(y^m\right)\left(m=1,2\right)$ corresponds to the bivariate cumulative distribution functions given as $F\left(x^1, x^2\right) \; and \; G\left(y^1, y^2\right)$. According to Bamber (1975), the AUC is equal $P\left(Y > X\ \right)$,which is the probability that the diseased subjects whose test results are positive is greater than the non-diseased subjects whose test results are negative. Let $AUC_m\left(m=1,2\right)$ represents the AUCs of the ROC curves for the two diagnostic tests. The null hypotheses of the equality of two AUCs were tested by DeLong et al.(1988) and Bandos, et al.(2005).Using the method of trapezoidal rule, the AUC for empirical ROC curve is computed (Bamber,1975). But Hanley and McNeil (1982) demonstrated that AUC obtained using the trapezoidal rule under an empirical ROC curve is equivalent to the Mann - Whitney *U* statistic for comparing test results from two samples.

According to Hanley and McNeil (1982), the AUC for a given diagnostic test is given by

$$A\hat{U}C = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} Q\left(X_i, Y_j\right) \qquad 1$$

Where $Q\left(X_i, Y_j\right) = \begin{cases} 1 \; if \; Y_j > X_i \\ 0 \, if \; Y_j < X_i \\ 0.5 \, if \; Y_j = X_i \end{cases}$

And Q is the indicator function comparing $X_i \; and \; Y_j$, N = Number of non-diseased subjects, M = Number of diseased subjects, X_i = Test result of the i^{th} non-diseased subject, Y_j =Test result of *jth* diseased subject. For m^{th} diagnostic test the AUC is given by

$$A\hat{U}C_m = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} Q\left(X_i^m, Y_j^m\right) \qquad 2$$

When the sampled test results are paired, $A\hat{U}C_{\Delta}$ represented as $A\hat{U}C_2 - A\hat{U}C_1$ is given by

$$A\hat{U}C_2 - A\hat{U}C_1 = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\left[Q\left(X_i^1, Y_j^1\right) - Q\left(X_i^2, Y_j^2\right)\right] \qquad 3$$

This shows the difference in the AUCs between two diagnostic tests.

PROPOSED 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 $i = 1, \ldots, \mathrm{N}$. Also let the test results of diagnostic tests 1 and 2 for the diseased subject be $Y_{j1} \; and \; Y_{j2}$ where $j = 1, \ldots, \mathrm{M}$. Also let $X = \left\{\left(X_{11}, \mathrm{X}_{12}\right), \left(X_{21}, \mathrm{X}_{22}\right), \ldots, \left(X_{N1}, \mathrm{X}_{N2}\right)\right\}$ denotes pairs of vector of measurement on non-diseased subjects and let $Y = \left\{\left(Y_{11}, \mathrm{Y}_{12}\right), \left(Y_{21}, \mathrm{Y}_{22}\right), \ldots, \left(Y_{M1}, \mathrm{Y}_{M2}\right)\right\}$ 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

$$A\widehat{U}C_{\Delta} = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} Q\left(X_{im}, X_{jm}\right) = \left[\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} Q\left(X_{i2}, \mathrm{Y}_{j2}\right) - \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} Q\left(X_{i1}, \mathrm{Y}_{j1}\right)\right] \qquad 4$$

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

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

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

$$S_{ijm} = \begin{cases} 1 \ \ if\ \ Y_{jm} > X_{im} \\ 0.5 \ \ if\ \ X_{im} = Y_{jm} \\ 0 \ \ \ if\ \ Y_{jm} < X_{im} \end{cases} \qquad 5$$

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 $S_1 = \left\{S_{11}, \mathrm{S}_{12}, ..., \mathrm{S}_{1N}, \mathrm{S}_{1,\mathrm{N}+1}, \mathrm{S}_{1,\mathrm{N}+2} ...., \mathrm{S}_{1n}\right\}$ be n measurements arising from diagnostic test 1 while the subscripts $p = 1, 2, .., \mathrm{N}$ shows test results for the non-diseased subjects while $q = N + 1, \mathrm{N}+2, ..., \mathrm{n}$ 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,

$$R_{pq1} = A\left(S_{q1} > S_{p1}\right) + \frac{1}{2}A\left(S_{p1} = S_{q1}\right); iff\ \ p \neq q \qquad 6$$

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

$$R_{pq2} = A\left(S_{q2} > S_{p2}\right) + \frac{1}{2}A\left(S_{p2} = S_{q2}\right); iff\ \ p \neq q. \qquad 7$$

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

To test the null hypothesis that $\mathrm{AUC}_\Delta = 0$, 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) 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

$$\mathrm{A\hat{U}C}_\Delta = \frac{1}{NM}\sum_{p=1}^{N}\sum_{q=1}^{M} iT_{pq} = \frac{1}{NM}\sum_{p=1}^{N}\sum_{q=1}^{M} T_{pq} r\left|Q_{pq}\right| \qquad 8$$

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 $r\left(Q_{pq}\right)$ 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 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.

Proposed Asymptotic Test for the Approximation to Permutation Test

In other to test for the hypothesis $A\hat{U}C_{\Delta} = 0$, we assume that AUC_{Δ} is symmetric around zero. We also assume that the populations from where samples are drawn must be quantitative data measured on at most the ordinal scale. This means that the paired test results could continuous or discrete. We continue to use without loss of generality that Q_{pq} is the difference between S_1 *and* S_2 measurements of test results from diagnostic test 1 and 2 respectively. Also $r\left(\left|Q_{pq}\right|\right)$ is the rank of the absolute value of Q_{pq}. Recall that in equation 8, $i = r\left(\left|Q_{pq}\right|\right)$. Note that testing the null hypothesis $AUC_{\Delta} = 0$ is equivalent to testing the null hypothesis that $r\left(Q_{pq}\right)$ has a symmetric distribution around zero. Based on the specifications of T_{pq} in equation 8, let n be the total number of subjects with p non-diseased and q diseased labels of subject in each of the diagnostic test so that T_{pq} represents the indicator function (outcome values) which were previously defined.

$$Let\ \pi_{pq} = P(T_{pq}) \qquad 9$$
$$where\ pq = either +1, 0, or -1.$$

That is,

$$\pi_{+} = P\left(T_{pq} = 1\right), \pi_{0} = P\left(T_{pq} = 0\right), \pi_{-} = P\left(T_{pq} = -1\right) \qquad 10$$

$$where\ \pi_{+} + \pi_{0} + \pi_{-} = 1$$

To validate or justify equation (8), let

$$W_{+}\ and\ W_{-}\ be\ respectively\ defined\ as$$

$$W_+ = \sum_{p=1}^{N}\sum_{q=1}^{M} iT_{pq} \qquad 11$$

$$where\ T_{pq} = \begin{cases} 1, if \ \ p < q \\ 0, if \ \ p = q \end{cases} ;i = r(|Q_{pq}|)$$

$$and$$

$$W_- = \sum_{p=1}^{N}\sum_{q=1}^{M} iT_{pq} \qquad 12$$

$$where\ T_{pq} = \begin{cases} -1, if \ \ p > q \\ \ 0, if \ \ p = q \end{cases} ;i = r(|Q_{pq}|)$$

Therefore

$$W = W_+ + \ W_- \qquad 13$$

Where W is here defined as the modified Wilcoxon signed rank test statistic which is the difference between the sums of signed ranks to absolute values of subjects test results with positive difference denoted as W_+ and negative difference denoted as W_- defined in *n* pairs of observations. The sum of these modified Wilcoxon signed ranks $\left(W_+ \ and \ W_-\right)$ gave the test statistic equivalent to that stated in equation (8) since 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. Instinctively, $\pi_+ \ and \ \pi_-$ are negatively related since their sum is the numbers 1 to n. Consider in a sample of n pairs of test results of subjects, the frequency of occurrences of +1,0 and -1 in the frequency distribution of T_{pq}. The mean and variance of $\left(T_{pq}\right)$ are respectively given as

$$E\left(T_{pq}\right) = n\left(\pi_+ - \pi_-\right) \qquad 14$$

$$and$$

$$Var\left(T_{pq}\right) = n\pi_+\left(1-\pi_+\right) + n\pi_-\left(1-\pi_-\right) + 2n\pi_+\pi_- \qquad 15$$

Again

$$E\left(W\right)=\sum_{p=1}^{N}\sum_{q=1}^{M}i\left(T_{pq}\right)=\sum_{p=1}^{N}\sum_{q=1}^{M}\left(i.E\left(T_{pq}\right)\right);$$

$$E\left(W\right)=\frac{n(n+1)}{2}.n\left(\pi_{+}-\pi_{-}\right) \qquad 16$$

The estimated value of the probability values $\pi_{+}-\pi_{-}$ is gotten as $\hat{\pi}_{+}-\hat{\pi}_{-}$ from $W=\frac{n+1}{2\left(\hat{\pi}_{+}-\hat{\pi}_{-}\right)}$.

Similarly,

$$Var\left(W\right)=Var\left(\sum_{p=1}^{N}\sum_{q=1}^{M}i\left(T_{pq}\right)\right)$$

$$Var\left(W\right)=\sum_{p=1}^{N}\sum_{q=1}^{M}\left(i^{2}.Var\left(T_{pq}\right)\right) \qquad 17$$

But $Cov(T_{pq},T_{qp})=0\sin ce\ p\neq q.$

$$Var(W)=\frac{n(n+1)(2n+1)}{6}\left(n\pi_{+}\left(1-\pi_{+}\right)+n\pi_{-}\left(1-\pi_{-}\right)+2n\pi_{+}\pi_{-}\right) \qquad 18$$

Note that $\pi_{+},\pi_{0}\ and\ \pi_{-}$ are respectively the probabilities that $p<q$ that is, p non-diseased test results of subject is less than q diseased test results of subject, $p=q$ that is, where p and q test results of subjects are both non-diseased or diseased, and $p>q$ that is, p non-diseased test result of subject is greater than q diseased test result of subject. They are estimated as the rate of occurrences of 1's, 0's and -1's in the distribution of p non-diseased and q diseased test results of subjects in T_{pq}.

Thus, $\hat{\pi}_+ = \frac{r^+}{n}, \hat{\pi}_- = \frac{r^0}{n} \; and \; \hat{\pi}_- = \frac{r^-}{n}$ 19

where r^+, r^0 *and* r^- are the rate of occurrences of 1,0 and -1 in the distribution of T_{pq}.

Given the null hypothesis, $AUC_\Delta = 0$,the test statistics is given as

$$Z = \frac{W - \frac{(n+1)}{2(\pi_+ - \pi_-)}}{\sqrt{\frac{n(n+1)(2n+1)}{6}\left(n\pi_+(1-\pi_+) + n\pi_-(1-\pi_-) + 2n\pi_+\pi_-\right)}} \qquad 20$$

Which has standard normal distribution under H_0 for fairly large sample size n. Where the probabilities, namely π_+ *and* π_- are replaced by their sample estimates. Since W_+ is defined as the sum of signed ranks to absolute values of subjects test results with positive differences and given the probability of occurrence of positive difference only, Wilcoxon's statistic by its specification does not provide explicitly for the possible occurrence of negative differences, so that π_- is automatically set equal to zero in equation 20. Under H_0, the mean of W is set to zero also and because permutation based procedures for comparing ROC curves permits that the $H_0 : AUC_\Delta = \frac{1}{2}$, according to Venkatraman and Begg (1996),Venkatraman (2000), Bandos, et al. (2005, 2006), as well as Braun and Alonzo (2008),we rewrite the test statistic as

$$Z = \frac{2W}{\sqrt{\frac{2n(n+1)(2n+1)}{3}\left(n\pi_+(1-\pi_+) + n\pi_-(1-\pi_-) + 2n\pi_+\pi_-\right)}} \qquad 21$$

This is however the asymptotic test for the null hypothesis $AUC_\Delta = 0$ when the sample size is fairly large. Practically, π_+ *and* π_- are replaced by their sample estimates in equation 19.

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 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 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 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 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.

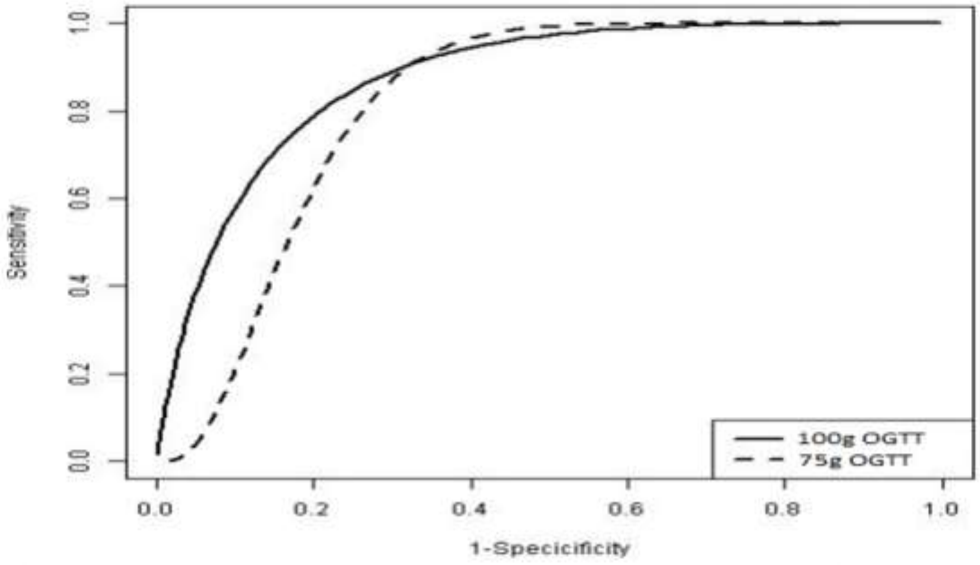

Figure 1: Crossed ROC curves for two diagnostic tests taken from data on GDM.

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 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 CONCLUSIONS

In applying the proposed test on real data, we saw in the graph of ROC curves figure 1 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. 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). 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.

REFERENCE

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

Bandos, A. (2005). Nonparametric methods in comparing two ROC curves. Doctoral dissertation, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh. (http://etd.library.pitt.edu /ETD/available/etd-07292005-012632/)

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.

Bandos, A. I., Rockette, H. E., and Gur, D. (2006). A permutation test for comparing ROC curves in multireader studies. A multi-reader ROC, permutation test. Academic radiology,13(4):414.

Braun, T. M. and Alonzo, T. A. (2008). A modified sign test for comparing paired ROC curves. Biostatistics 9 (2): 364-372.

DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L. (1988). Comparing the area under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44(3), 837-846.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143: 29-36.

Kotz, S., Johnson, L.N. and Read, B.C. (1988) Encyclopedia of statistical sciences volume 8. John Wiley & Sons, Inc., New York.

Kummar R, Indrayan A. Receiver operating characteristic (ROC) curve for medical researchers. Indian Pediatr 2011; 48: 277-89.

Oyeka, Cyprian A. (2009); An Introduction to Applied Statistical Methods 8th edition, Nobern Avocation Publishing Company, Enugu, Nigeria (ISSN 978-2457-6-7).

Pardo,M.C and Franco-Pereira,A.C(2017). Non Parametric ROC summary statistics. Statistical journal volume 15, number 4, october 2017, 583–600.

Venkatraman ES, Begg CB.(1996). A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment. Biometrika **83**(4): 835- 848.

Venkatraman, E. S. (2000). A permutation test to compare receiver operating characteristic curves. Biometrics 56: 1134-1138.

Wang, C. C. (2015). A MATLAB Package for multivariate normality test. Journal of Statistical Computation and Simulation 85 (1): 166-188.