Asian Journal of Mathematics & Statistics

Research Article
Comparison of Two Diagnostic Test Procedures Using Modified Wilcoxon Signed Rank Test

[1]U.M. Okeh and [2]I. Onyeagu Sidney

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

Abstract

Background and Objectives: 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. This study was aimed at assessing a difference in the AUCs of paired data where each non-diseased and diseased subject are both subjected to 2 diagnostic test procedures as well as to tackle the problem of exchangeability of the labels between two diagnostic tests within subject which characterizes previous studies. **Materials and Methods:** 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 was proposed. A real data was analyzed to compare the proposed test and that of standard test as well as the 2 diagnostic tests. **Results:** Test reveals that the p-values for 2 h 70 g OGTT and 2 h 100 g OGTT for the non-diseased subjects are, respectively 0.6124 and 0.8975 while that of diseased subjects 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. **Conclusion:** Result showed that the proposed test is more powerful than the standard test. Also 2 h 100 g OGTT diagnostic test is superior to 2 h 70 g OGTT diagnostic test at a time that the specificity is greater than 0.7.

INTRODUCTION

In nonparametric inference, they first derived the Area Under the Curve (AUC) test[1] and proposed a test where paired data accounted for correlations[2]. Similarly, they constructed a nonparametric area test to compare two empirical AUC estimates[2]. The AUC is defined as the probability that the observed value of the diagnostic test will be greater for a randomly selected diseased individual than for a randomly selected non-diseased individual if higher values of a diagnostic test are associated with diseased subjects, while lower values are associated with non-diseased[3]. They developed a totally nonparametric approach to compare two correlated AUCs of two diagnostic tests for paired samples of subjects by using the theory of generalized U statistics[3]. In other words, they developed a conventional fully nonparametric approach leading to an asymptotically normal test statistic[3]. The test by DeLong *et al.*[3] 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. When the comparison of AUCs of two Receiver Operating Characteristic (ROC) curves was done, it can be estimated that which 1 of 2 diagnostic tests is more suitable for discriminating non-diseased subjects from diseased subjects[4]. To overcome the challenges occasioned by the area test of DeLong *et al.*[3] which requires large sample sizes, a permutation test which thrives when the sample size is small was also proposed. In carrying out permutation tests involving diagnostic tests, two authors proposed a method for detecting any differences at every operating point between two ROC curves[5]. Similarly, other authors proposed a method that is sensitive to the difference in AUCs in diagnostic performance[6]. 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.*[6] has an easy-to-implement and precise approximation and better detects different ROC curves if they differ with respect to the AUC while it was aimed to increase the power to detect a crossing alternative[5]. Specifically, Bandos *et al.*[6] based their permutation test on the difference in areas and derived exact and asymptotic permutation test methods to test the equality of 2 correlated ROC curves which are designed to have increased power to detect difference in the AUC. The test of Bandos *et al.*[6] 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.*[6] compared the performance of their test to that of DeLong *et al.*[3] through simulation and found that the permutation test had greater power than the nonparametric test developed by DeLong *et al.*[3] when there was moderate correlation between diagnostic tests, large AUCs and small sample sizes. Bandos *et al.*[6] 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[7]. In order to obtain exact test in clinical trials which requires a given small sample size, Harris and Hardin[8] proposed Wilcoxon Signed Rank (WSR) test. This is because large-sample results are not acceptable in many clinical trials studies. The WSR 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. The WSR test statistic utilizes both the magnitudes and signs of differences unlike the sign test proposed by Braun and Alonzo[7] which utilizes only the signs of the differences between each observation and ignoring the magnitudes of these observations. Therefore, WSR test is expected to be more powerful test than the sign test. The essential assumptions for the WSR test are continuous and symmetric population distribution. Current study was aimed to assess a difference in the AUCs of paired data where non-diseased and diseased subject are both subjected to two diagnostic test procedures as well as to tackle the problem of exchangeability of the labels between two diagnostic tests within subject.

MATERIALS AND METHODS

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 mth diagnostic test, respectively. Then $x_i^m = i$ where, i = 1, 2,..., N and $y_j^m = j$ where, 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 $F_m(x^m)$, $G_m(y^m)$ (m = 1, 2) corresponds to the bivariate cumulative distribution functions given as F (x[1], x[2]) and G (y[1], y[2]). The AUC is equal P (Y>X), 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[9]. Let AUC_m (m = 1, 2) represents the AUCs of the ROC curves for the two diagnostic tests. The null hypothesis of the equality of two AUCs were tested[3,6]. Using the method of trapezoidal rule, the AUC for empirical ROC curve is computed[9], but Hanley and McNeil[1] 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[1], 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:

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

Where:
Q = Indicator function comparing X_i and Y_i
N = Number of non-diseased subjects
M = Number of diseased subjects
X_i = Test result of the ith 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 method: The comparison of 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 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:

$$X_{i1} \text{ and } X_{i2}$$

where, i = 1,...N.

Also let the test results of diagnostic tests 1 and 2 for the diseased subject be:

$$Y_{j1} \text{ and } Y_{j2}$$

where, j = 1,..., M.

Also let X = {(X_{11}, X_{12}), (X_{21}, X_{22}),..., (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}),..., (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 nonparametrically as:

$$\begin{aligned}A\hat{U}C_\Delta &= \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}Q\left(X_{im}, X_{jm}\right) \\ &= \left[\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}Q\left(X_{i2}, Y_{j2}\right) - \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}Q\left(X_{i1}, Y_{j1}\right)\right]\end{aligned} \qquad (4)$$

Where:

$$Q\left(X_{im}, Y_{jm}\right) = S_{ij2} - S_{ij1} = S_{ijm}$$

and:

$$S_{ijm} = A\left(X_{im} < Y_{jm}\right) + \frac{1}{2}A\left(X_{im} = Y_{jm}\right); m = 1, 2$$

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

Consider according to Hanley and McNeil[1], that this indicator function is:

$$S_{ijm} = \begin{cases} 1 \text{ if } X_{im} < Y_{jm} \\ 0.5 \text{ if } X_{im} = Y_{jm} \\ 0 \text{ if } X_{im} > Y_{jm} \\ m = 1, 2 \end{cases} \qquad (5)$$

In other to test the null hypothesis $H_0: AUC_2 - AIC_1 = 0$, the M and N subjects were combined 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 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, the comparison of every subject's test result to every other subject's test result. Thus:

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

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 $S_2 = \{S_{21}, S_{22},..., S_{2N}, S_{2,N+1}, S_{2,N+2},..., S_{2n}\}$ be n measurements arising from diagnostic test 2 while the subscripts p = 1, 2,..., N shows test results for the non-diseased subjects while q = N+1, N+2,..., n shows test results for the diseased subjects. Similarly within diagnostic test, comparison was done to every subjects test result to every other subjects test result, that is:

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

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, adoption of the transformation in Eq. 5 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, generalization of the estimate of AUC_Δ as:

$$\hat{AUC}_\Delta = \frac{1}{NM}\sum_{p=1}^{N}\sum_{q=1}^{M} iT_{pq} = \frac{1}{NM}\sum_{p=1}^{N}\sum_{q=1}^{M} T_{pq}r\left|Q_{pq}\right| \qquad (8)$$

Where:

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

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. The AUC is equivalent to two-sample Wilcoxon Test Statistic[10] and can be used to carry out test of symmetry around zero for paired samples. Based on that finding, the Eq. 5 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 modified Wilcoxon Signed Rank test is more powerful than the modified sign test statistic Oyeka[11] proposed by Braun and Alonzo[7] 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. The null hypothesis was tested that $AUC_\Delta = 0$ by computing AUC_Δ for every permutation of tested 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.

Data collection and study area: The study area used for this study was from Alex Ekwueme University Teaching Hospital, Abakaliki, Ebonyi state, which is a Tertiary Care Teaching Hospital with referrals from 13 General hospitals from 13 local government areas, 40 private hospitals and dispensaries. The data collection was approved by the Research and Ethics Committee. Records revealed that antenatal mothers with 24-28 weeks gestation period were the category of pregnant women who were involved in the study and whose data were recorded. As a procedure all pregnant women passed through a fasting and after lunch plasma glucose evaluation during the first antenatal visit in first trimester. As a protocol for screening of Gestational Diabetic Mellitus (GDM) subjects, all antenatal mothers that falls within this category of gestation period were given a 50 g Oral Glucose Challenge Test (OGCT) irrespective of the presence or absence of risk factors.

Available record showed that 50 g glucose was given as a solution in 200 mL water which was consumed within 5 min. This was done not minding whether the antenatal mothers fasted or not. They were not allowed to eat and drink for a period of 1 h. Blood sample was taken from them after 1 h. Those whose plasma glucose measurements after 1 h were $\geq$140 mg dL^{-1} was considered as diseased while those whose measurements after 1 h were recorded as <140 mg dL^{-1} was considered as non-diseased subjects. Out of a total of 2850 subjects screened in 2 years (January, 2016-December, 2017) period, 166 subjects tested positive for GDM. A total of 166 subjects received screening within this period and who tested positive for GDM. Using a 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 $\geq$155 mg dL^{-1} for one to be considered diseased/positive (coded 1) for GDM while <155 mg dL^{-1} 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 can be inferred 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 2 diagnostic tests are equal. The real data if analyzed will evaluate the performance of the proposed estimates. It will compare the performance of the 2 diagnostic tests in terms of ROC curves between the 2 diagnostic tests and a crossing ROC curve will emerge. The crossing ROC curves will have the areas for the 2 diagnostic test procedures. In applying the data, the diagnostic test results need to have a bivariate binormal distribution. Most powerful test does not exist for testing bivariate normal distribution[12]. Therefore, for each test result, one resorted to checking only the univariate normality.

RESULTS

Checking for univariate normality of two diagnostic test results by Shapiro-Wilk test reveals that the p-values for the
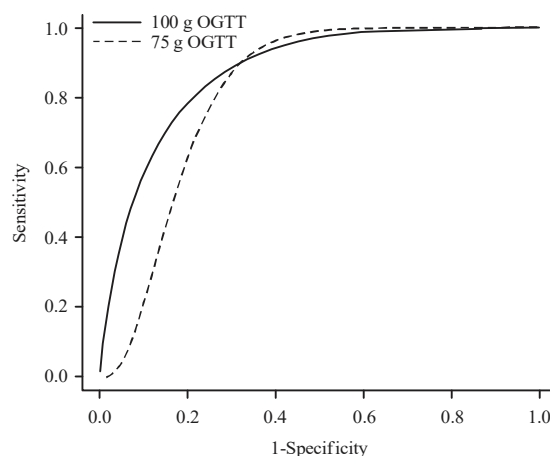

Fig. 1: Crossed ROC curves for two diagnostic tests taken from data on GDM

ROC: Receiver operating characteristic, GDM: Gestational diabetic mellitus, OGGTT: Oral glucose tolerance test

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 as present in Fig. 1. 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 standard permutation test, the null hypothesis is also rejected since the p-value is 0.0387.

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 based on proposed modified Wilcoxon Signed Rank test statistic. In comparing paired ROC curves, current 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 2 diagnostic tests having both diseased and non-diseased subject in each of the test where permutations are made between subjects particularly by shuffling the diseased and non-diseased labels of the subjects within each diagnostic test. It will be recalled 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 0.4 at least[3]. Therefore, at small average AUC, low correlation between diagnostic tests and at sample size higher than 60, the method by DeLong *et al.*[3] has improved test size and greater power than our permutation test otherwise permutation has improved test size and greater power. Venkatraman and Begg[5] found that for noncrossing ROC curves, the statistical power of DeLong *et al.*[3] is higher than that of Venkatraman and Begg[5] because the procedure of Venkatraman and Begg[5] 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 e.g., DeLong *et al.*[3] test. Therefore, Venkatraman and Begg[5] test has lower power for noncrossing ROC curves as it detect differences in ROC curves while in the same scenario, DeLong *et al.*[3] test has higher power as it detects differences in AUCs. The 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.*[3] conventional test that only detects difference in AUCs. Result showed that this proposed test has comparable power to the test conducted by Bandos *et al.*[6] as well as Braun and Alonzo[7], 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 this proposed test is designed to consider the signs of values as well as the absolute ranks of values. Braun and Alonzo[7] considered only the signs of values. This permutation test is slightly conservative but has an excellent power to detect a crossing alternative based on simulation results. 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 previous authors[5,6].

CONCLUSION

In applying the real data, the proposed permutation test is more powerful than the comparison test since it has the more likelihood of rejecting the null hypothesis. Graph of ROC curves showed 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. Also since the null hypothesis for the univariate normal is rejected given the disparity in the p-values of the diagnostic tests for non-diseased and diseased as well as the values of AUCs, the 2 diagnostic test procedures did not contribute equivalent diagnostic information.

The proposed test can be a very suitable alternative to the comparison test that only consider the direction of values (signs of differences). The strength of proposed test is that it has easy implementation to discriminate diagnostic test procedures even by non-statisticians. Since the MWSR test is easy to compute as well as easy to communicate to the potential uses of the procedure, this test can be used conveniently.

SIGNIFICANCE STATEMENT

The use of Modified Wilcoxon Signed Rank test as permutation test will circumvent the difficulties or reduce the computational burden associated with estimating the difference between two AUCs in a paired sample design. The method of comparing ROC curves using this test statistic is designed to assess a difference in AUCs for paired samples and it provides meaningful information than when ROC curves cross and has the same AUCs. Adjustment for the presence of tied absolute value of difference or zero value in the test statistic helps to increase the power and accuracy of tests since no data is lost due to absence of zero difference. Also the proposed test offers reliable statistical inferences for small sample sizes. Since permutations are made within each diagnostic test between subjects, the validity of the permutation test holds even when both diagnostic tests are measured on different scales. This permutation procedure in diagnostic tests research will resolve the problem of exchangeability of the labels between two diagnostic tests within subject which characterized previous existing tests. This study will help the researcher to uncover the critical areas of diagnostic tests comparisons that many other researchers were not able to explore. Thus a new theory based on between subject's label permutation patterns within each diagnostic test may be arrived at.

ACKNOWLEDGMENT

REFERENCES

1. 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.
2. Hanley, J.A. and B.J. McNeil, 1983. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology, 148: 839-843.
3. DeLong, E.R., D.M. DeLong and D.L. Clarke-Pearson, 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics, 44: 837-845.
4. Sumi, N.S. and M.A. Hossain, 2012. A study on parametric approaches to compare areas under two correlated ROC curves. Bangladesh J. Scient. Res., 25: 61-71.
5. 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.
6. 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. Statist. Med., 24: 2873-2893.
7. Braun, T.M. and T.A. Alonzo, 2008. A modified sign test for comparing paired ROC curves. Biostatistics, 9: 364-372.
8. Harris, T. and J.W. Hardin, 2013. Exact Wilcoxon signed-rank and Wilcoxon Mann-Whitney ranksum tests. Stata J., 13: 337-343.
9. Bamber, D., 1975. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. J. Math. Psychol., 12: 387-415.
10. Pardo, M.C. and A.M. Franco-Pereira, 2017. Non parametric ROC summary statistics. REVSTAT-Statist. J., 15: 583-600.
11. Oyeka, C.A., 2009. An Introduction to Applied Statistical Methods. 8th Edn., Nobern Avocation Publishing Company, Enugu, Nigeria.
12. Wang, C.C., 2015. A MATLAB package for multivariate normality test. J. Statist. Comput. Simulat., 85: 166-188.