COMPARISON OF TWO OR MORE CORRELATED AUCS IN PAIRED SAMPLE DESIGN

Okeh Uchechukwu Marius[1] and Julian Ibezimako Mbegbu[2]

[1]Department of Mathematics and Computer Science, Ebonyi State University Abakaliki, Nigeria.
[2]Department of Statistics, Nnamdi Azikiwe University Awka, Nigeria

ABSTRACT: *The performance of a diagnostic test when test results are measured on a binary or ordinal scale can be evaluated using the measures of sensitivity and specificity. In particular, when it is measured on a continuous scale, the assessment of the performance of a diagnostic test is always over the range of possible cut-off points for the predictor variable. This is achieved by the use of a receiver operating characteristic (ROC) curve which is a graph of sensitivity against 1-specificity across all possible decision cut-offs values from a diagnostic test result. This curve evaluates the diagnostic ability of tests to discriminate the true state of subjects and compare the performance of two alternative diagnostic tests performed on the same subject. These tasks of comparing diagnostic tests is always better achieved using a summary measure of accuracy across all possible ranges of cut-off values called the area under the receiver operating characteristic curve (AUC).So many parametric and nonparametric methods exist for comparing two or more correlated AUCs in diagnostic tests when the data is paired. In this paper, we proposed a simple and easy to understand chi-square method of comparing two or more AUCs in a paired sample design. The proposed method which does not require the knowledge of true status of subjects or gold standard in evaluating the accuracy of tests unlike the existing methods, it offers reliable statistical inferences even in small sample problems and circumvent the difficulties of deriving the statistical moments of complex summary statistics as seen in the Delong et al method. The proposed method provides for further analysis to determine the possible reason for rejecting the null hypothesis of equality of AUCs. The proposed method when applied on real data, was shown to be better than the Delong et al method as it avoids the lengthy and more difficult procedures of estimating the variances of two AUCs as a way of determining if two AUCs differ significantly. The method is validated using the Cochran Q test and was shown to compare favourably.*

KEYWORDS: Chi-Square Test, Delong et al, Cochran Q Test, Cut-Off Value, AUC, ROC, Predicted Probability, Dichotomous Data

INTRODUCTION

In medical sciences, the use of diagnostic procedures is based on clinical investigations or laboratory experiments or trials purposely to classify subject into diseased or non-diseased. These procedures makes for vital decision making aided with advanced machines/tools to detect any given condition. For decades now, receiver operating characteristic curve (ROC) analysis has been used as a popular technique of evaluating the performance or ability of a test to discriminate between alternative health status. The ROC curve represents a graph of sensitivity against 1-specificity across various cut-off values of diagnostic test. It assesses the effectiveness of continuous diagnostic test results to differentiate between groups of healthy and diseased individuals (Greiner et al., 2000; Zhou et al., 2002; Pepe, 2004). It is also a common tool for assessing the performance of various classification tools such as diagnostic tests, and to compare accuracy between tests or predictive models. The ROC curve was originated in the theory of signal detection in the years 1950-1960 (Green and Swets, 1966; Egan, 1975) to discriminate between signal and noise. It can provide a direct and visual comparison of two or more diagnostic tests on a single set of scales. It is possible to compare different tests at all decision cut-offs by constructing the ROC curves. For statistical analysis, a recommended numerical index of accuracy associated with an ROC curve are often better used to summarize the information provided for the ROC curve into a single global value or index(Swets and Picket,1982). This index is called area under the ROC curve. AUC takes values between 0.5 (which corresponds to the diagonal ROC curve that passes through the points (0,0) and (1,1)) and 1 (representing perfect test where all cases are correctly classified). AUC represents the diagnostic accuracy of the test Y, so that the larger the area the better the diagnostic accuracy of Y. This means that values closer to 1 indicate that Y optimally discriminates between healthy and diseased subjects, while values near 0.5 indicate that the test is not informative (Zhou et al, 2002). According to Mann-Whitney (1947), AUC is the probability that the observed test result of a randomly selected subject from the diseased population (Y_1) is larger than the observed test result of a randomly selected subject from the non-diseased population (Y_0).For comparing two diagnostic processes, the difference between AUCs is often used. In diagnostic imaging it is generally known that the changes due to subjects represent a major component of the overall changes of the AUC. To better control for the sources of changes when comparing diagnostic tests, a paired study design is often advised because it usually induces positive correlation between the tests results of the same subjects. This paper is devoted to reviewing some existing methods for comparing AUC. Finally, it is motivated to developing a test statistic for comparing the AUCs of diagnostic tests.

A number of methods exist for comparing two AUCs for the paired sample case whereby each subject given a condition of interest has test results coming from two or more diagnostic tests in which case they are matched or paired. When presented with two tests used in detecting a certain condition, it is not always feasible to simply directly compare two ROC curves. Therefore rather than comparing the two ROC curves visually, the AUC for the two ROC curves are compared. Another way to compare these procedures is to compare their respective AUCs to determine if the two AUCs are significantly different. This is achieved by taken into account the variances of both AUCs.

Existing Methods For Comparing Two AUCs In Paired Sample Data

Parametric (Binormal ROC Curve) Method

The parametric analysis assuming the binormal model was developed by Dorfman and Alf Jr.(1969), McClish (1989)and later implemented and further developed by Metz et al(1998). To compare the AUCs of two diagnostic test results for paired sample design and given the viability of the binormal assumption according to McClish(1989), the hypothesis for the equality of two AUCs denoted respectively as $AUC_1 \text{ and } AUC_2$ can be tested using the test statistic given as

$$z = \frac{\left(AUC_1 - AUC_2\right)}{\sqrt{V\left(AUC_1 - AUC_2\right)}} \qquad 1$$

$$where$$
$$V\left(AUC_1 - AUC_2\right) = V(AUC_1) + V(AUC_2) - 2C\,\text{ov}(AUC_1, AUC_2)$$
$$and\ Cov\left(AUC_1, AUC_2\right) = \rho SE\left(AUC_1\right) SE\left(AUC_2\right)$$

McClish (1989) derived $V(A\hat{U}C)$ as

$$V(A\hat{U}C) = f_1^2 V(\hat{a}_1) + f_2^2 V(\hat{a}_2) + 2 f_1 f_2 Cov(\hat{a}_1, \hat{a}_2)$$
$$where$$
$$f_1 = \frac{e^{-a_1^2/2(1+a_2^2)}}{\sqrt{2\pi(1+a_2^2)}},\ f_2 = -\frac{a_1 a_2 e^{-a_1^2/2(1+a_2^2)}}{\sqrt{2\pi(1+a_2^2)^3}},\ \hat{a}_1 = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\hat{\sigma}_1}\ and\ \hat{a}_2 = \frac{\hat{\sigma}_2}{\hat{\sigma}_1}$$

The variance of AUC can be estimated by substituting estimators for the parameters a_1 and a_2.

From equation 1, $Cov(A\hat{U}C_1, A\hat{U}C_2) = \rho SE(A\hat{U}C_1).SE(A\hat{U}C_2)$ according to Metz et al (1984) is an estimate of the covariance between the two correlated AUC's in parametric approach of comparative study of two diagnostic procedures. Where ρ and SE denote the correlation coefficient between the two estimated AUC's and the standard error (i.e. the square root of variance) of estimate of AUC's respectively. If the two diagnostic tests are not examined on the same subjects, obviously the two estimated AUC's are independent and the covariance term would be zero.

Non-parametric methods

DeLong et al(1988) developed a consistent empirical (nonparametric) estimator of the covariance matrix for several AUC estimators in a paired design.. The conventional nonparametric test for comparing correlated AUCs proposed by DeLong et al.(1988) uses a consistent variance estimator and relies on asymptotic normality of the AUC estimator. Comparing the AUC of paired sample design by DeLong et al (1988) using the empirical non-parametric method is based on the previous work by Zhou et al(2002) that a Z-test for this comparison of the AUCs of two diagnostic test for paired sample design is

$$Z = \frac{AUC_1 - AUC_2}{\sqrt{Var\left(AUC_1 - AUC_2\right)}} \qquad 2$$

$$where$$

$$Var\left(AUC_1 - AUC_2\right) = Var(AUC_1) + Var(AUC_2) - 2Cov\left(AUC_1 - AUC_2\right)$$

And each variance is defined as

$$Var(AUC_t) = \frac{S_{Y_{t1}}}{n_{t1}} + \frac{S_{Y_{t0}}}{n_{t0}} \qquad 3$$

$$where$$

$$S_{Y_{ti}} = \frac{\sum_{j=1}^{n_{ti}} \left[Var(Y_{tij}) - AUC_t\right]^2}{n_{ti} - 1}, t = 1,2; i = 0,1; j = 0,1.$$

The variance of the components Y_{t0j} *and* Y_{t1i} are respectively defined as

$$Var\left(Y_{t0j}\right) = \frac{\sum_{i=1}^{n_{t1}} F\left(Y_{t1i}, Y_{t0j}\right)}{n_{t1} - 1} and\ Var\left(Y_{t1i}\right) = \frac{\sum_{j=1}^{n_{t0}} F\left(Y_{t1i}, Y_{t0j}\right)}{n_{t0} - 1}, t = 1,2 \qquad 4$$

Since

$$F\left(Y_{1i}, Y_{0j}\right) = \begin{cases} 1\ if\ \ Y_1 > Y_0 \\ 0.5\ if\ \ Y_1 = Y_0 \\ 0\ if\ \ Y_1 < Y_0 \end{cases}$$

$$\therefore$$

$$AUC_t = \frac{1}{n_{t1}} \sum_{i=1}^{n_{t1}} Var(Y_{t1i}) = \frac{1}{n_{t0}} \sum_{j=1}^{n_{t0}} Var(Y_{t0j}), t = 1,2. \qquad 5$$

Note here that Y_{t0j} *and* Y_{t1i} are the observed diagnostic test results for the *jth and ith* subjects in group t that are not diseased and diseased respectively.

Also

$$Cov\left(AUC_1, AUC_2\right) = \frac{S_{Y_{11}Y_{21}}}{n_1} + \frac{S_{Y_{10}Y_{20}}}{n_0} \tag{6}$$

$$where$$

$$S_{Y_{11}Y_{21}} = \frac{1}{n_1 - 1}\sum_{j=1}^{n_1}\left[Var(Y_{11j}) - AUC_1\right]\left[Var(Y_{21j}) - AUC_2\right]$$

$$And$$

$$S_{Y_{10}Y_{20}} = \frac{1}{n_0 - 1}\sum_{j=1}^{n_0}\left[Var(Y_{10j}) - AUC_1\right]\left[Var(Y_{20j}) - AUC_2\right]$$

When the variances are estimated, one can calculate the AUC for the two diagnostic tests and then make comparison.

Proposed Chi-Square Test Statistic For Comparing The Equality Of Two Or More Correlated AUCs

Interest is to develop a simple and easy to understand method of testing the equality of AUCs arising from two or more diagnostic tests across different diagnostic tests. We here propose a chi-square test for the comparison of two or more diagnostic tests based on continuous, ordinal or binary scale data. Given measurement of test results on continuous scale, we dichotomize the results as positive or diseased (coded 1) and negative or non-diseased (coded 0) using a cut-off value c and present the information as coded in a contingency table.

Suppose n is a random sample of subjects drawn from a population of subjects for this study and x_{ij} is the sample test result for the ith subject at jth diagnostic test T, i = 1, 2, …, n and j = 1, 2,…,T,

Let

$$y_{ij} = \begin{cases} 1, if \;\; x_{ij} \geq c \\ 0, otherwise \end{cases} \tag{7}$$

Where y_{ij} is the continuous diagnostic test result drawn from population Y.

Based on the classification of y_{ij} in equation 7, the format of the data obtained is presented in Table 1.

Table 1. Format of the data for the test results of the ith subject at the jth diagnostic test.

	Diagnostic test dependent result			
Subjects	1	2	..	T

1	x_{11}	x_{12}	..	x_{1T}
2	x_{21}	x_{22}	..	x_{2T}
3	x_{31}	x_{32}	..	x_{3T}
…			..	
n	x_{n1}	x_{n2}	..	x_{nT}

This pattern of coding is appropriate if interest is to compare the AUCs obtained from diagnostic tests processes carried out on the same set of subjects. The coding is such that if a subject's test result is $x_{ij} \geq c$, that subject is considered diseased or response positive(coded 1) while a subject whose test result is $x_{ij} < c$ is declared non-diseased or response negative to the disease (coded 0).

To develop the test statistic for testing the equality of two or more AUCs across different diagnostic tests,

Let

$$\pi_j = p\left(y_{ij} = 1\right) \qquad 8$$

And

$$f_j = \sum_{i=1}^{n} y_{ij} \qquad 9$$

Where f_j indicates the number of subjects that are diseased or responding positive in the jth diagnostic test while the corresponding subjects who are not diseased or those responding negative is $n - f_j$.

Let

$$n_{1.} = f = \sum_{j=1}^{T} f_j \qquad 10$$

be the total number of subjects who are diseased for all the diagnostic tests T and let

$$n_{2.} = \sum_{j=1}^{T} n - f_j = n(T-1) - f \qquad 11$$

be the total number of subjects who are non-diseased or responding negative for all the diagnostic tests. Hence

$$E(y_{ij}) = \pi_j ; Var(y_{ij}) = \pi_j (1 - \pi_j) \qquad 12$$

and

$$E(f_j) = n\pi_j ; Var(f_j) = n\pi_j (1 - \pi_j) \qquad 13$$

Where π_j is the population proportion of subjects that are diseased or those responding positive for the jth diagnostic test.

Its sample estimate and sample variance are respectively

$$A_j = \frac{f_j}{n} \qquad 14$$

And

$$Var(A_j) = \frac{\hat{\pi}_j \left(1 - \hat{\pi}_j\right)}{n} = \frac{f_j (n - f_j)}{n^3} \qquad 15$$

Where A_j is actually the area under a portion of the AUC curve of jth diagnostic test. If the proportions of positive response or diseased subjects are equal for all the diagnostic test T, then the common proportion can be estimated as

$$A = \frac{f}{nT} \qquad 16$$

These results are presented in a 2 × T contingency table.

Table 2. 2 × T Contingency table for the Analysis of diagnostic Test Dependent Measurements.

	Diagnostic Test Measurements				
Observations	1	…..	… …	T	Total
Number of diseased subjects (f_j)	f_1	…..	….	f_T	f
Number of non-diseased subjects ($n - f_j$)	$n - f_1$	…..	…	$n - f_T$	$nT - f$
Total	n	…..	…	n	nT

Proportion(A_j)	A_1	…..	….	A_T	$A=\frac{f}{nT}$

Based on table 2, the observed numbers of diseased and non-diseased subjects for the *jth* diagnostic test are respectively

$$o_{1j} = f_j \ and \ o_{2j} = n - f_j \qquad 17$$

The corresponding expected numbers of diseased and non-diseased subjects are respectively

$$e_{1j} = \frac{nf}{nT} \ and \ e_{2j} = \frac{n(nT-f)}{nT} \qquad 18$$

The null hypothesis that the AUC of two or more diagnostic tests are equal is stated as

$$H_0 : AUC_1 = AUC_2 = ... = AUC_T \ versus \ H_1 : AUC_1 \neq AUC_2 \neq ... \neq AUC_T \qquad 19$$

Since AUC summarizes the accuracy or discriminating power of a diagnostic test, we shall subsequently be viewing the test of hypothesis for AUC in terms of the proportion of positive response rate, π_j of subjects to tests because the ability of a test to discriminate among alternative health status (positive and negative response) gives a summary of the diagnostic accuracy of a test. In other words, the probability of positive response of a test if obtained indicates a summary of the accuracy or discriminating power of a diagnostic test given that the proportion of negative response is just a relationship.

The corresponding test statistic for testing this hypothesis is

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{T}\frac{\left(o_{ij} - e_{ij}\right)^2}{e_{ij}} \qquad 20$$

Whose distribution is approximately of the chi-square type having T–1 degrees of freedom and it can be used to test the null hypothesis of equality of AUCs across diagnostic tests.

Writing equation 20 in terms of Equations 17 and 18, we have

$$\chi^2 = \sum_{j=1}^{T}\left(\frac{\frac{(f_j - nf)^2}{nT}}{\frac{nf}{nT}}\right) + \sum_{j=1}^{T}(n - f_j) - \frac{\frac{n(nT-f)^2}{nT}}{\frac{n(nT-f)}{nT}} \qquad 21$$

This simplifies to

$$\chi^2 = \frac{nT^2}{f(nT-1)} \sum_{j=1}^{T} \frac{(f_j - f)^2}{T} \qquad 22$$

Equation 22 has also a distribution of the chi-square type with T–1 degrees of freedom.

Rewriting equation 20 in terms of the proportions stated in equations 14 and 16, we have an equivalent expression given as

$$\chi^2 = n\sum_{j=1}^{T} \frac{\left(A_j - A\right)}{Aq}, iff \ nT > 30; q = 1 - A \qquad 23$$

At a given level of significance (α), the null hypothesis H_0 is rejected if

$$\chi^2 \geq \chi^2_{1-\alpha;T-1} \qquad 24$$

Otherwise it is accepted.

SUBSEQUENT ANALYSIS IF NULL HYPOTHESIS IS REJECTED

When the null hypothesis of equations 19 is rejected, it means that differences exist in the AUC across diagnostic tests or the proportion of positive response across diagnostic tests. Therefore, it is of interest to determine which of the AUC or equivalently the proportion of positive response among the diagnostic tests that has contributed to the rejection of H_0.In particular, interest may be to determine if the accuracy of diagnostic test result is improving successively over testing trials or procedures. Let π_j be the proportion of subjects that are diseased or those responding positive for the jth diagnostic test.

Now let π_j *and* π_k be the population proportions of subjects that are diseased or those responding positive at the *jth and kth* diagnostic tests respectively for $j, k = 1,..,T; j \neq k.$

Its corresponding sample estimates are respectively,

$$A_j = \frac{f_j}{n} \ and \ A_k = \frac{f_k}{n} (as\ in\ equ\ 14).$$

Where A_j *and* A_k are the areas under a portion of the AUC of *jth and kth* diagnostic tests respectively. Interest here may be in testing the null hypothesis that the proportion of diseased subjects in jth diagnostic test is at most equal to the proportion of diseased subjects in kth diagnostic test. The null hypotheses may be expressed as

$$H_0 : \pi_j \leq \pi_k \text{ versus } H_1 : \pi_j < \pi_k ; j \neq k \qquad 25$$

To test the null hypothesis of equation 25, where the sample estimates of $\pi_j - \pi_k$ are $A_j - A_k$.

Let

$$z = \frac{\left(A_j - A_k\right) - \pi_0}{s_e\left(A_j - A_k\right)} \qquad 26$$

Where

$$s_e\left(A_j - A_k\right) = \sqrt{Var(A_j) + Var(A_k) - 2Cov(A_j, A_k)}; se = s\tan dard\ deviation$$

Where

$$\begin{aligned} Cov(A_j, A_k) &= E\left[A_j A_k\right] - E\left[A_j\right]E\left[A_k\right] \\ &= \frac{E\left[f_j f_k\right]}{n^2} - \frac{E\left[f_j\right]E\left[f_k\right]}{n}, (\ equation\ 2.12) \\ &= \frac{\sum_{i=1}^{n}\sum_{s=1}^{n} E\left[y_{ij} y_{sk}\right]}{n^2} - \pi_j \pi_k, (equation\ 2.11) \end{aligned}$$

Now $y_{ij}.y_{sk}$ assumes the value 1 provided y_{ij} *and* y_{sk} both assume the value 1 with probability $\pi_j \pi_k$.

Therefore,

$$\frac{\sum_{i=1}^{n}\sum_{s=1}^{n} E\left[y_{ij} y_{sk}\right]}{n^2} = \frac{n^2\left(\pi_j \pi_k\right)}{n^2} = \pi_j \pi_k$$

Hence,

$$Cov(A_j, A_k) = \pi_j \pi_k - \pi_j \pi_k = 0$$

Therefore,

$$Var(A_j - A_k) = Var(A_j) + Var(A_k) \qquad 2.25$$

Based on the null hypothesis of equation 2.23, the statistic *z* of equation 2.24 becomes

$$z = \frac{(A_j - A_k)}{s_e(A_j - A_k)} - \pi_0 = \frac{(A_j - A_k) - \pi_0}{\sqrt{Var(A_j) + Var(A_k)}}$$

which is the unit normal distribution.

Hence,

$$\chi^2 = z^2 = \frac{\left((A_j - A_k) - \pi_0\right)^2}{Var(A_j) + Var(A_k)} \qquad 2.26$$

has approximately a distribution of the chi-square type with 1 degree of freedom where A_j is already given in equation 2.12

And

$$Var(A_j) = \frac{f_j\left(n - f_j\right)}{n^2}, (equation\ 2.13)$$

Under the null hypothesis of equation 2.23 and overall estimate of π_j such as $A_j \; is \; A$ given in equation (2.14), the estimate of $Var(A_j)$ is

$$Var(A_j) = \frac{A(1-A)}{n} = \frac{f\,(nT - f\,)}{n^3T^2} \qquad 2.27$$

The corresponding test statistic is given as:

$$\chi^2 = z^2 = \frac{\left(\left(\frac{f_j}{n} - \frac{f_k}{n}\right) - \pi_0\right)^2}{Var(A_j) + Var(A_k)} = \frac{\left(\left(\frac{f_j}{n} - \frac{f_k}{n}\right) - \pi_0\right)^2}{2Var(A)} \qquad 2.28$$

Where under the null hypothesis

$$Var(A_j) = Var(A_k) = Var(A) = \frac{f(nT-f)}{n^3T^2}, (equation\ 2.21)$$

The test statistic of equation 3.30 is now given as

$$\chi^2 = \frac{n^3T^2\left(\frac{f_j}{n} - \frac{f_k}{n}\right)}{f(nT-f)} - \pi_0 \qquad 2.29$$

In terms of A_j *and* A given in equations 2.12 and 2.14, equation 2.29 becomes

$$\chi^2 = \frac{n\left[(A_j - A_k) - \pi_0\right]^2}{2A(1-A)} \qquad 2.30$$

which has approximately a chi-square distribution with 1 degree of freedom.

The test statistic of equation 2.30 can be used to test the null hypothesis of equation 2.23 and can also be compared with a well chosen critical value of the chi-squared distribution with T–1 degrees of freedom at a specified α level. This is to make type 1 error become smaller and minimizes errors in conclusions.

Application To Real Data

The proposed methods can be applied to real data obtained from a retrospective study of pregnant women at risk for gestational diabetes mellitus (GDM) at certain hospitals in Ebonyi State Nigeria.The records of a total of 1113 pregnant women who had earlier tested positive after screening using 1 hour 50g Glucose Challenge Test (GCT) and who were also subjected to diagnosis using 2-hour 75g OGTT as well as 3-hours 100g OGTT according to WHO(1999) and National Diabetic Data Group(NDDG,1979) criteria were taken. This was to compare the efficacy of these two diagnostic procedures, These pregnant women were seen to have positive risk factors and aged between 15-45 years at less than 24 weeks and between 24-28 weeks of gestation.

Women who were known diabetics, or who were suffering from any chronic illness were excluded from the study. After obtaining permission from the hospitals' Research and Ethics Committee, assess was granted into the record units of the antenatal wards of these hospitals where the medical history of the patients were kept in a proforma containing general information on demographic characteristics such as body mass index, maternal age, previous fetal weight and vital clinical histories such as obstetric history of GDM and family history of diabetes were taken.

The GDM response variables (tests results) for the two tests, namely 75g OGTT and 100g OGTT represents the paired data for the pregnant women. These data type is suitable for comparing the accuracy of two tests in terms of their AUCs. Under this arrangement, the null hypothesis of interest which is testing of equality of the proportion of positive response is equivalent to testing the equality of AUCs for the tests. This comparison will be evaluated using the proposed method.

The research interest is to compare two correlated AUCs of diagnostic tests which also are equivalent to comparing the probability of positive response for paired sample design. To do this, we code the data for this work based on the specification of equation 7 to generate the corresponding data of 1's and 0's. In other to calculate the chi-square test statistic of equation 2.21 for testing the null hypothesis of no difference among the proportion of positive response in paired sample design, we evaluate the data for the work to have table 3.

Table 3: Computation of total number of diseased, non-diseased and proportion of diseased.

	Diagnostic test 1	Diagnostic test 2	Total
No of 1's (f_j)	146	149	295
No of 0's $(n-f_j)$	967	964	1931
Total n	1113	1113	2226
Proportion of 1's (p_j)	0.1311	0.1339	0.265

Now, to test the null hypothesis of equation 2.17 which is equivalent to testing the homogeneity of AUCs for paired sample tests, we use the proportion of 1's or diseased pregnant women of the data in equation 2.21, to calculate the chi-square test statistic as

$$\chi^2 = \frac{1113\left[(-0.1339)^2 + (-0.1311)^2\right]}{(0.265)(0.735)} = \frac{1113\left[(0.01793) + (0.01719)\right]}{0.194775} = \frac{1113[0.03512]}{0.194775} = \frac{39.08856}{0.194775} = 200.96$$

At 5% level of significance, where c=2, the chi-square is $\chi^2_{0.95,1} = 3.841$.

This means that the proportion of positive response for the two diagnostic tests differ significantly. In other words, the two AUC for the tests are different. From Table 3, the proportion of pregnant women who have GDM increased after the second diagnostic test.

Furthermore, our interest may be to determine which of the test is superior or other wise. This is also the same as carrying out further analysis to determine which of the test that may have contributed to our rejecting the null hypothesis of equality of AUCs in equation 2.17. This means that we need to test the null hypothesis of equation 2.23. From Table 3, put $p_2 = 0.1339$ *and* $p_1 = 0.1311$, in equation 2.30 to have

$$\chi^2 = \frac{1113(0.1311 - 0.1339)^2}{2(0.265)(0.735)} = \frac{1113 \times 0.00000784}{0.38955} = 0.0224.$$

If $\chi^2 = 0.0224$ is compared with its critical value at $c-1=2-1=1 and\ \alpha = 0.05$, we accept the null hypothesis of equation 2.23 and conclude that there exist no significant different between the two diagnostic tests. This simply means that $AUC_2 > AUC_1$.This means that that the second diagnostic test (100g OGTT), that is AUC_2 is preferred to first diagnostic test (75g OGTT), that is AUC_1 because the second test was able to discover few more pregnant women who actually have plasma glucose level of at least 7.8mmol/l(GDM positive patients).

Comparison Of The Proposed Method With Delong Et Al (1988) Method

Using the same coded data meant for comparing AUC, we obtained the estimates of AUCs for the diagnostic tests as 0.687, and 0.752 respectively for the first and second diagnostic test respectively. To test the null hypothesis of equation 2.17 for the homogeneity of AUCs, the non-parametric test by DeLong et al.(1988) and the proposed chi-square test yielded significant results with their p-values as 0.0068 and 0.0027 respectively.

Validation Of The Proposed Method Using Cochran Q Test

To make the proposed method valid in terms of efficiency, we illustrate using Cochran Q test for dichotomous data since the same null hypothesis of equality of AUCs (proportion of diseased pregnant women) across diagnostic tests can suitably be tested. Using the paired coded data which is also applicable, we let B_i be the sum of the number of 1's in row i, the pregnant women, where i=1,2,….,1113 and $Z_{j,k}$ be the sum of the number of 1's in column j and k, where j is test 1 and k is test 2. Then the statistic for Cochran's Q test is given by

$$Q = (T-1)\frac{\left[\sum_{j=1}^{T} Z_j^2 - \left(\sum_{j=1}^{T} Z_j\right)^2 \Big/ T\right]}{\sum_{i=1}^{n} B_i - \left(\sum_{j=1}^{T} B^2\right)\Big/ T} = (2-1)\frac{\left[(146)^2 + (149)^2 - (146+149)^2/2\right]}{295 - \left((2)^2 + (1)^2 + (2)^2 + ... + (1)^2\right)\Big/ 2} = \left[43517 - 43513\right] = 4$$

Which has T-1=2-1=1 degrees of freedom. Since $Q = 4$ is greater than $3.841 = \chi^2_{0.95;1,}$, we reject the null hypothesis of equation 2.17 which stated the equality of diagnostic tests in terms of their AUCs and conclude that the proportion of pregnant women responding positive (GDM positive patients) and indeed the AUCs differs significantly across tests. This conclusion is the same as that obtained when the proposed method is applied to the same data set. The advantage that the proposed method has over Cochran Q test is that it is capable of finding out why null hypothesis is rejected in the first instance. This implies that subsequent analysis when the null hypothesis is rejected which applies to the proposed method does not apply in the Cochran method.

SUMMARY

The proposed Chi-square test method is simple and easy to compute as well as easy to understand or communicate to the potential uses of the procedure. The strength of our method is that it has easy implementation to discriminate diagnostic test procedures even by non-statisticians. Knowledge of true status of subjects or any other gold standard is not required to employ the proposed method in analysis.

The proposed method of comparing two or more AUCs can widely be used whenever conventional solutions such as the Delong et al method are questionable, difficult to derive or unavailable. The proposed method offers reliable statistical inferences even in small sample problems and circumvent the difficulties of deriving the statistical moments of complex summary statistics.The proposed method is employed when there is no need for the knowledge of true status of subjects or no requirement of gold standard. Just like what was seen in Cochran Q test, Delong et al method cannot be used to carry out further analysis when the null hypothesis is rejected. The proposed method has the capacity of comparing even more than two AUCs, while other existing methods such as the Delong et al(1988) can only compare two AUCs. The proposed method avoids the lengthy and more difficult procedures of estimating the variances of two AUCs as a way of determining if two AUCs differ significantly, rather it developed a simple test statistic for testing the hypothesis of equality or otherwise.

CONCLUSION

In validating the proposed chi-square test statistic using Cochran Q test for dichotomous data, we conclude that significant difference exists in both tests. We also compared the proposed chi-square test method with the conventional nonparametric test suggested by DeLong et al in 1988 by comparing two AUCs to determine the statistical power of the two diagnostic tests in discriminating non-diseased from diseased subjects. Result also showed that the proposed chi-square test statistic is a very suitable alternative having higher statistical power when compared to the test by Delong et al. (1988) that are very cumbersome to compute. We conclude that the proportion of pregnant women having GDM differs significantly across diagnostic tests. The advantage that the proposed method has over Cochran Q test is that the chi-square test provides the opportunity of further analysis to know why the null hypothesis was rejected in the first instance. This implies that subsequent analysis is suggested in the proposed method, where by one tries to find out by further analysis why the null hypothesis is rejected.

DISCUSSION

The method of calculating AUC from predicted probability of positive response avoids the computational complex procedures of the maximum likelihood estimation (MLE) and numerical integration methods which not only involves lengthy calculations but also have restrictive assumptions about the distribution of diagnostic test results since there are parametric methods. It is note worthy that estimates from parametric methods such as the method of MLE are inconsistent thereby giving a misleading picture of the regression relationship (Pepe, 2003). Our method of calculating AUC is unique in so many ways: it incorporates predicted probability of positive response in the construction of ROC curve and indeed AUC, it uses the α prediction rule which enables the construction of very smooth ROC curve because several values of α normally will produce smoothness in the curve, the AUC calculated is also diagnostic test dependent since the test result depends on the test for the subjects and finally the method is not only simpler and straight forward but also it avoids the iteration procedure which is rigorous, time consuming and liable to errorrneous results.

DeLong et al (1988) proposed a nonparametric approach to compare the correlated AUCs using the theory of the U-statistics. The disadvantage of this method is that it has computational burden although that can be alleviated by the development of faster computers. Computational burden can still be substantial in binormal ROC curve as a method of calculating AUC because a number of iterative procedures that are involved in obtaining estimators, for instance MLE of AUC (Dorfmann & Alf 1969; Metz, Herman & Shen 1998).

The chi-square test employs a continuous distribution to approximate a discrete probability distribution. Apart from being simple to calculate, easy to understand and readily applicable, the chi-square test statistic provides the quality evidence of inferiority or superiority of one diagnostic process over the other. It does this by providing the opportunity for subsequent analysis to determine the inferiority or other wise of a test when the null hypothesis of interest is rejected. For instance, if the null hypothesis of equality of AUCs is rejected, there is need for further analysis to ascertain which of the AUC or diagnostic test that has contributed to the rejection of the null hypothesis. This is not possible when the Delong et al (1988) method is used for comparing between two AUCs. In using any existing methods, the idea of assessing inferiority or superiority of a diagnostic procedure are not immediately obtainable because sensitivity and specificity simultaneously measures the accuracy of diagnostic procedures (Swets et al, 1982) which always requires the knowledge of true disease status of subjects (Pepe,2003). The proposed method does not require the knowledge of the true disease status or the gold standard may not be known. This makes the proposed method to have robust feature since it is invariably applicable in all the instances especially where it does require having the knowledge of true status (gold standard).This is not the same with other traditional test methods such as Bandos et al(2005) and Delong et al(1988)which must require the knowledge of true status (gold standard) in estimating the AUC.

The chi-square test statistic is recommended for comparing the equality of correlated AUCs in paired sample design.

REFERENCES

Aoki, K., J. Misumi, T. Kimura, W. Zhao and T. Xie, 1997. Evaluation of cutoff levels for screening of gastric cancer using serum pepsinogens and distributions of levels of serum pepsinogens I, II and of PG I/PG II ratios in a gastric cancer case-control study. *Journal of Epidemiology*, **7**, 143 – 151.

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.

Buros, Amy and Tubbs, Jack D.(2013). Applying the Jonckheere-Terpstra Statistic to AUC Regression. Department of Statistical Science, Baylor University, Waco, TX 76706. www.louisville.edu/sphis/bb/srcos-2013/….

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), 837-845.

Dorfman DD, Alf JrE. Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals – rating-method data. *Journal of Mathematical Psychology* 1969; **6**: 487-496.

Egan, JP. (1975). Signal detection theory and ROC Analysis. New York: Academic Press.

Green, DM. and Swets, JA. (1966). Signal detection theory and psychophysics. John Wiley & Sons, Inc. New York.

Greiner, M., Pfeiffer, D., and Smith, RD. (2000). Principals and practical application of the Receiver Operating Characteristic analysis for diagnostic tests. Preventive Veterinary Medicine, 45:23–41.

Hsiao, J. K., J. J. Barko and W. Z. Potter, 1989. Diagnosing diagnoses: receiver operating characteristic methods and psychiatry. *Archives of General Psychiatry*, **46**, 664 – 667.

J. A. Hanley and B. J. McNeil, “The meaning and use of the area under a receiver operating characteristic (ROC) curve,” *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

Lasko, T. A., J. G. Bhagwat, K. H. Zou and L. Ohno-Machado, 2005. The Use of Receiver Operating Characteristic Curves in Biomedical Informatics. Journal of Biomedical Informatics, **38**, 404 – 415.

Mann, H.B & Whitney D.R (1947).On a test whether one of two random variables is stochastically larger than the other.Ann.Math.Statist;18,pp.50-60.

McClish, DK. (1989). Analyzing a portion of the ROC curve. Medical Decision Making, 9:190–195.

Metz CE, Wang PL, Kronman HB. A new approach for testing the significance of differences between ROC curves from correlated data. In: Deconick F, editor. Information processing in medical imaging. The Hague: Nijhoff; 1984. p.432–45.

Metz, C. E., 1989. Some practical issues of experimental design and data analysis in radiological ROC studies.*Investigation Radiology*, **24**, 234 – 245.

Metz, CE., Herman, BA., and Shen, JH. (1998). Maximum likelihood estimation of Receiver Operating Characteristic (ROC) curves from continuously-distributed data. Statistics in Medicine, 17:1033–1053.

Metzger BE, Buchanan TA, Coustan DR, de Leiva A, Dunger DB, Hadden DR, et al. Summary and recommendations of the fifth international workshop-conference on gestational diabetes mellitus. *Diabetes Care* 2007; 30: S 251.

National Diabetes Data Group. Classification and diagnosis of diabetes mellitus and other categories of glucose intolerance. Diabetes 1979; 28: 1039–1057

Pepe, M.S. (2003). *The statistical evaluation of medical test for classification and prediction*. Oxford: Oxford University Press.

Pepe, MS. (2004). The statistical evaluation of medical tests for classification and prediction. 1st ed. Oxford University Press, USA.

Swets, J. A. and R. M. Picket, 1982. Evaluation of diagnostic systems: methods from signal detection theory. Academic Press, New York.

WHO. Definition, Diagnosis and Classification of Diabetes Mellitus and its Complications. Report of a WHO consultation. Part 1: Diagnosis and Classification of Diabetes Mellitus. Geneva: Department of Non-communicable Disease Surveillance, World Health Organization, 1999.

Zhou, X. H., N. A. Obuchowski and D. K. McClish, 2002. Statistical methods in diagnostic medicine. Wiley, New York.