Open Science Journal of Statistics and Application
2018; 5(2): 13-21
http://www.openscienceonline.com/journal/osjsa
ISSN: 2381-4950 (Print); ISSN: 2381-4969 (Online)

An Asymptotic Multivariate Test for Testing the Equality of the Average Areas Under Independent Receiver Operating Characteristic Curves - Simulation Study and an Application

Marina Roshini Sooriyarachchi*, Nuzhi Ahmed Meyen

Department of Statistics, University of Colombo, Colombo, Sri Lanka

Email address

roshinis@hotmail.com (M. R. Sooriyarachchi), NuzhiM@hotmail.com (N. A. Meyen)

*Corresponding author

To cite this article

Marina Roshini Sooriyarachchi, Nuzhi Ahmed Meyen. An Asymptotic Multivariate Test for Testing the Equality of the Average Areas Under Independent Receiver Operating Characteristic Curves - Simulation Study and an Application. *Open Science Journal of Statistics and Application.* Vol. 5, No. 2, 2018, pp. 13-21.

Received: April 18, 2018; **Accepted**: May 15, 2018; **Published**: July 23, 2018

Abstract

The Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) Curve has become a popular summary measure of the curve. In a previous paper, the authors proposed an asymptotic bivariate test for comparing AUCs for paired data. In this case, the test statistic derived was found to follow a distribution proportional to the Beta distribution. This test can also be applied to the multivariate case for independent data as shown in this paper. The properties of the developed test are examined by using simulation studies for the scenario of multivariate independent ROC curves. The general method is illustrated for this case by applying it to a published data set in the Rockit manual. The simulation studies found that the developed test has good properties for large samples.

Keywords

Multivariate Test, Receiving Operating Characteristic (ROC) Curve, Area Under the Curve (AUC), Beta Distribution, Maximum Likelihood Estimates, Simulation

1. Introduction

1.1. Background

Receiver Operating Characteristic (ROC) curves were developed in the 1950s for studying radio signals. In modern times this procedure has been adapted for decision making in medicine, agriculture, biology and so on. The goodness of a diagnostic test can be measured using sensitivity and specificity of the test. The sensitivity is how good the test is at detecting true positives and pertains to the true positive fraction, specificity is the ability of the test to detect true negatives and pertains to the true negative fraction. A ROC curve is a plot of the sensitivity versus 1 - specificity as the test threshold is varied ([1, 2]). The most popular summary measure of a ROC curve is the area under the curve (AUC) and alternative diagnostic tests have been compared by comparing their AUCs ([3, 4]). However, since of late other measures different from AUC have also been looked at though these have still to become popular [5]).

The classic paper of Hanley and McNeil [6] first popularized the theory for comparing two AUCs pertaining to two independent ROC curves. In Hanley and McNeil [6] they estimate the AUC of a ROC curve based on the well-known nonparametric Wilcoxon statistic. They also derived closed-form expressions for the approximate standard error of the AUC for ROC curves. Hanley and McNeil [6] derive a test statistic for detecting the difference between the areas under two ROC curves for an unpaired design (independent data). Another popular method of comparing AUCs is the method of DeLong, DeLong and Clarke-Pearson [7]. It is preferred over some other non-parametric methods, for small

sample sizes [8]. The Mann Whitney method is used in DeLong, DeLong and Clarke-Pearson [7] for estimating the AUC and its standard errors. Hanley and McNeil [9] criticized the trapezoidal rule (Mann Whitney test) used in the non-parametric estimation for underestimating the AUC. They indicated preference for the Dorfman and Alf [10] method in this regard. Subsequently, Park, Goo and Jo [11] showed that the estimate of the AUC based on the Wilcoxon statistic also underestimates the true value of the AUC and they also recommended the maximum likelihood approach of Dorfman and Alf [10] for the estimation of the AUC. The Dorfman and Alf [10] approach is implemented in two well-known software packages for ROC analysis, namely ROCKIT [11] and StAR [12]. Very recently, Martinez-Camblo [13] used a new non-parametric approach based on the Youden index to combine AUC results of ROC curves for Meta-analysis. He found that his method was better than some older reference nonparametric methods. To implement this new method the package NSROC [14] can be used.

1.2. Objectives

In section 1.1 the nonparametric methods of Hanley and McNeil [6] and DeLong, DeLong and Clarke-Pearson [7] were clearly criticized due to these methods underestimating the AUC. The Martinez-Camblo's [13] new non-parametric approach can only be applied to Meta-analysis. The recommended method of estimation given as a consensus by all authors is the Dorfman and Alf [10] approach ([11, 12]). Park, Goo & Jo, [11] and Veragra, Normbuena, Ferrada, Slater & Melo, [12] went on to develop software, namely, ROCKIT and StAR respectively using Dorfman and Alf [10] method of estimation, however they did not study the properties of their test and only analyzed a few examples. Dorfman and Alf [10] initially developed code for comparing AUC's of two paired ROC curves. Thus the primary objective of this paper is to modify the asymptotic bivariate statistical test for the paired case and develop a multivariate statistical test to compare several AUCs of independent ROC curves that is based on the Dorfman and Alf approach of estimation of independent data and examine the properties of this test using large scale simulations. The second objective of this paper is to illustrate the developed methods on an example.

1.3. Brief Explanation of Methodology

The general theory is based on the independent case where an asymptotic multivariate test was used for comparing several AUCs at once. For large samples, the test statistic modified follows a distribution which is proportional to the Beta distribution with parameters depending on the number of AUC curves compared (p) and the number of independent quantities making up the AUC (n). The values of the estimates of the AUCs and their standard errors were based on Dorfman and Alf [10] maximum likelihood approach.

1.4. Data for the Example

The method developed is applied to independent data consisting of a dataset taken from the Rockit manual. The data consists of information on a 60 observer study of 3 different mammographic techniques applied to 20 CAD patients on each test. The estimated (data based) values of the AUCs and their variance-covariance matrix were obtained using the package ROCKIT [11]. Other example data sets are given in Schutts [2].

Section 2 gives a review of the literature pertaining to the problem. In section 3 the theorems, definitions, results and proofs related to modifying test statistic for the bivariate case and developing the multivariate test are presented. Section 4 consists of a simulation study to examine the properties of the test. Section 5 gives an illustration of the methodology developed in section 3 on an example. Conclusions and Discussion are given in Section 6.

2. Literature Review

Our paper is based on the comparison of the performance of binary classifiers by using Receiver Operating Characteristic (ROC) curves. The Area under the curve (AUC) is the most popular summary measure of ROC curves ([15, 3, 16]).

To test for significant differences between AUCs of independent ROC curves, the main factor that needs to be considered is the outcome distribution [12]. This will determine the approach to be used in estimating the AUCs and its variance-covariance matrix. Possible approaches are parametric, ([17, 1, 18]) semi-parametric ([19]) and non-parametric ([6, 7, 20]). For each approach, different methods of estimating the AUC have been used. For the parametric approach, Dorfman and Alf [10] method of fitting smooth curves based on the binormal assumption is usually used where the ROC curve can be completely described by two parameters estimated using Maximum Likelihood Estimation (MLE). The semi-parametric approach of Metz, Herman and Roe [19] is also based on a parametric binormal model yet the MLE used does not depend on an explicit expression of the likelihood function. In the non-parametric approach of Bamber [21] the trapezoidal rule equivalent to the Mann-Whitney U statistic is used.

3. Method

3.1. Estimating an AUC of an Independent ROC Curve Using the Dorfman and Alf Method

The signal detection paradigm on which ROC curves are based is important to understand the underlying principle behind ROC curve analysis. According to Grey and Morgan [22], the signal-detection paradigm consists simply of a subject successively choosing between a signal present population (with background noise), SN, or signal absent population (just noise), N. The model then assumes that the response of the subject can be represented by a random variable X with cumulative distribution function, $F_{SN}(x)$ if the signal was present, $F_N(x)$ if no signal was present.

For the purposes of this study $F_N(x) = \Phi(x), F_{SN}(x) = \Phi(bx - a)$ where b and a are the two principal parameters

of the ROC curve which can be seen to depend on the means and standard deviation of $F_N(x)$ and $F_{SN}(x)$ and $\Phi(.)$ denotes the cumulative density function of the standard normal distribution. The values of a and b along with other parameters of the ROC curve were estimated using the method of scoring proposed in Grey and Morgan [22].

Simulation: The method of scoring used is an iterative process which uses initial parameter estimates. The start for the initial iteration was used as the parameter estimates of the simple linear regression as given in Grey and Morgan (1972). Iteration continues until either, two successive iterates differ by less than 10^{-3} in all of their components and the final iterate is a possible solution. A degenerate solution for the parameter estimates of the ROC curve can occur from empty cells in the data matrix. Therefore, in order to overcome the problem of degeneracy similar to [23] the method of scoring developed adds a small positive constant in order to avoid degeneracy in the case of empty cells. Other more recent methods of simulation are given in Dobson et al. [24]

Calculation of the AUC and variance of the AUC:

It is possible to obtain the AUC of a ROC curve using the following formula $AUC = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right)$ where $\Phi(.)$ denotes the cumulative standard normal distribution.

In order to calculate the variance of the AUC, the delta method [25] is made use of, giving the formula as follows for the variance.

$$Var(\widehat{AUC}) = \left(\frac{\partial AUC}{\partial a}\right)^2 var(\hat{a}) + \left(\frac{\partial AUC}{\partial b}\right)^2 var(\hat{b}) + 2\left(\frac{\partial AUC}{\partial a}\right)\left(\frac{\partial AUC}{\partial b}\right) cov(\hat{a},\hat{b})$$

3.2. Modification of the Bivariate Test for the Paired Case (Seneratna, Sooriyarachchi, Meyen, 2015) to the Multivariate Test for the General Case of p AUC Curves for the Independent Case)

The theory given inr section 3.2 is applicable to independent data.

3.2.1. Relevant Theorems, Definitions and Results

Theorem 1 ([26]):

If $\underline{\mathbf{X}} \sim \boldsymbol{N}_{\boldsymbol{p}}\left(\underline{\boldsymbol{\mu}},\underline{\boldsymbol{\Sigma}}\right)$ is a random variable from a p-dimensional multivariate normal (Gaussian) distribution and $\underline{\mathbf{W}} \sim \boldsymbol{W}_{\boldsymbol{p}}\left(\underline{\boldsymbol{\Sigma}},\boldsymbol{n}\right)$ has a Wishart distribution where n is the number of independent quantities associated with $\underline{\boldsymbol{X}}$ then the distribution of $T^2 = n\left(\underline{\mathbf{X}}-\underline{\boldsymbol{\mu}}\right)' \underline{\mathbf{W}}^{-1}\left(\underline{\mathbf{X}}-\underline{\boldsymbol{\mu}}\right)$ is $\boldsymbol{T}^2(\boldsymbol{p},\boldsymbol{n})$ that is it follows a Hotelling's T-square distribution with parameters p and n. Here $\frac{1}{n}\underline{W}$ is the "sample variance" matrix of $\underline{\mathbf{X}}$. That is $\frac{1}{n}\underline{\mathbf{W}} = \underline{\hat{\boldsymbol{\Sigma}}}$ It can be shown that $\frac{n-p+1}{np}T^2 \sim F(p, n-p+1)$ where F is the F distribution. Here $\boldsymbol{\Sigma}$ is a diagonal matrix as the ROC curves are independent. That is the off-diagonal terms (covariance terms) of the matrix are zero.

Theorem 2 (, [27]):

The general form of the Hotelling's T^2 statistic is $T_G^2 = \left(\underline{\mathbf{X}}-\underline{\hat{\boldsymbol{\mu}}}\right)' \underline{\hat{\boldsymbol{\Sigma}}}^{-1}\left(\underline{\mathbf{X}}-\underline{\hat{\boldsymbol{\mu}}}\right)$ where $\underline{\mathbf{X}} \sim N_p\left(\underline{\boldsymbol{\mu}},\underline{\boldsymbol{\Sigma}}\right)$ and $\underline{\hat{\boldsymbol{\Sigma}}}$ is some estimator of $\underline{\boldsymbol{\Sigma}}$ and $\underline{\hat{\boldsymbol{\mu}}}$ is some estimator of $\underline{\boldsymbol{\mu}}$. Again $\boldsymbol{\Sigma}$ is a diagonal matrix as the ROC curves are independent. Thus the diagonal elements of $\underline{\hat{\boldsymbol{\Sigma}}}^{-1}$ are the reciprocal of the diagonal elements of $\underline{\hat{\boldsymbol{\Sigma}}}$ and the off-diagonal elements are zero. Wilks ([28]) and Gnanadesikan and Kettering [29]) showed that under the conditions described in theorem 1, the exact distribution of T_G^2 is proportional to the Beta distribution. That is $T_G^2 \frac{n}{(n-1)^2} \sim Beta\left(\frac{p}{2}, \frac{n-p-1}{2}\right)$.

Theorem 3 [30]):

If $\underline{\mathbf{X}} \sim N_p\left(\underline{\boldsymbol{\mu}},\underline{\boldsymbol{\Sigma}}\right)$ is a random variable from a p-dimensional multivariate normal distribution and $\frac{S}{n} = \frac{1}{n}\left(\underline{\mathbf{X}}-\underline{\boldsymbol{\mu}}\right)\left(\underline{\mathbf{X}}-\underline{\boldsymbol{\mu}}\right)'$ is the maximum likelihood estimator of the "Population Covariance matrix" $\underline{\boldsymbol{\Sigma}}$. Then the random matrix $\underline{\mathbf{S}}$ has a p-dimensional Wishart distribution with parameters n and $\underline{\boldsymbol{\Sigma}}$. Here too $\boldsymbol{\Sigma}$ is a diagonal matrix as the ROC curves are independent.

Result 1

If $\underline{\mathbf{X}}_i$ is a n by p matrix of p variables each having n elements and it has distribution $N_p\left(\underline{\boldsymbol{\mu}},\underline{\boldsymbol{\Sigma}}\right)$ then it follows that $\overline{\underline{\mathbf{X}}}$ (the sample mean of the $\underline{\mathbf{X}}_i$' s) has a distribution $N_p\left(\underline{\boldsymbol{\mu}},\frac{\underline{\boldsymbol{\Sigma}}}{n}\right)$. It follows from Theorem 3 that $S_{\mu}^2 = \left(\overline{\underline{\mathbf{X}}}-\underline{\boldsymbol{\mu}}\right)\left(\overline{\underline{\mathbf{X}}}-\underline{\boldsymbol{\mu}}\right)'$ has a p-dimensional Wishart distribution with parameters n and $\underline{\boldsymbol{\Sigma}}'$ where the variance-

covariance matrix of $\underline{\overline{\mathbf{X}}}$ is diagonal and $\underline{\overline{\boldsymbol{\Sigma}}}' = \frac{\underline{\boldsymbol{\Sigma}}}{n}$

3.2.2. Showing That the Asymptotic Distribution of the Test Statistic Developed for Testing the Equality of Several AUCs Is Proportional to the Beta Distribution

This was proved in Seneratna, Sooriyarachchi and Meyen [31] for the bivariate case for comparing AUC's of two paired ROC curves. Here it is illustrated how this case can be modified to the multivariate case for comparing AUC's of several independent ROC curves.

Let $\underline{\mathbf{AUC}} = \begin{pmatrix} AUC_1 \\ AUC_2 \\ . \\ . \\ AUC_p \end{pmatrix}_{p\times 1}$

Where AUC_i is the AUC of the i[th] ROC curve.

Let $\underline{\mathbf{A\hat{U}C}}$ be an estimate of $\underline{\mathbf{AUC}}$, let $\underline{\boldsymbol{\mu}}$ be the expected value of $\underline{\mathbf{A\hat{U}C}}$ and let $\underline{\boldsymbol{\Sigma}}$ be the associated diagonal variance-covariance matrix of $\underline{\mathbf{A\hat{U}C}}$. Then as $\underline{\mathbf{A\hat{U}C}}$ is the Dorfman and Alf (1969) maximum likelihood estimate (MLE) of $\underline{\mathbf{AUC}}$ and as MLE's are asymptotically normal (for large samples). That is $\underline{\mathbf{A\hat{U}C}} \sim N_p\left(\underline{\boldsymbol{\mu}}, \underline{\boldsymbol{\Sigma}}\right)$.

Suppose the estimate $\underline{\mathbf{A\hat{U}C}}$ of $\underline{\mathbf{AUC}}$ of a ROC curve is made up of the sum of n independent quantities where, n is a function of n_1 (the number of positive responses) and n_2 (the number of negative responses) [12]. The $\underline{\mathbf{A\hat{U}C}}$ is made up of $n_1 n_2$ quantities (pairs) of which n = min (n_1, n_2) are independent. Thus n is the number associated with $\underline{\mathbf{A\hat{U}C}}$. Let $n_t = n_1 + n_2$.

$\underline{\hat{\boldsymbol{\Sigma}}}$ is the Dorfman and Alf [10] MLE of the covariance matrix $\underline{\boldsymbol{\Sigma}}$ of the $\underline{\mathbf{A\hat{U}C}}$. According to Theorem 3, [30] the sampling distribution of the MLE of the $\left(\underline{\mathbf{A\hat{U}C}}-\underline{\boldsymbol{\mu}}\right)\left(\underline{\mathbf{A\hat{U}C}}-\underline{\boldsymbol{\mu}}\right)'$ matrix is asymptotically $W_p\left(\underline{\boldsymbol{\Sigma}}, n\right)$ as $\underline{\mathbf{A\hat{U}C}}$ has an asymptotic multivariate normal distribution. Thus asymptotically according to theorem 3, [30] and Result 1, $n\underline{\hat{\boldsymbol{\Sigma}}} \sim W_p\left(\underline{\boldsymbol{\Sigma}}, n\right)$

We want to test the null hypothesis (H_0) that all $\underline{\mathbf{AUC}}$s are the same on average versus the alternative hypothesis (H_1) that all $\underline{\mathbf{AUC}}$s are not the same on average.

That is $H_0 : \underline{\boldsymbol{\mu}} = \underline{\mathbf{K}}$ where $\underline{\mathbf{K}}$ is a constant vector, versus $H_1 : \underline{\boldsymbol{\mu}} \neq \underline{\mathbf{K}}$

As we do not know $\underline{\mathbf{K}}$ it has to be estimated. $\underline{\mathbf{K}}$ can be estimated as $\overline{\mathbf{K}}$ the simple average of the $\underline{\mathbf{A\hat{U}C}}$ (that is individual $\underline{\mathbf{A\hat{U}C}}_i$'s). That is $\overline{\mathbf{K}} = \frac{\sum_{i=1}^{p} A\hat{U}C_i}{p}$

From Theorem 2, the general form of the Hotelling's T^2 statistic [26] is $T_G^2 = \left(\underline{\mathbf{A\hat{U}C}}-\underline{\overline{\mathbf{K}}}\right)' \underline{\hat{\boldsymbol{\Sigma}}}^{-1} \left(\underline{\mathbf{A\hat{U}C}}-\underline{\overline{\mathbf{K}}}\right)$.

The dimensionality (p) needs to be reduced by 1 for estimating $\underline{\mathbf{K}}$. Therefore take q=p-1 instead of p. Then for large samples, $T_G^2 \frac{n}{(n-1)^2} \sim Beta\left(\frac{q}{2}, \frac{n-q-1}{2}\right)$.

Here p is the number of AUCs and n is the number of independent quantities used to calculate the AUCs. For the case of large samples (large n_1 and n_2) n will be large. Under this condition T_G^2 has an approximate chi-square distribution (under the null hypothesis) with q degrees of freedom ([32], [33]). The test statistic T_G^2 can be used to test H_o. The percentage points for the test statistic's distribution can be obtained by

$$\frac{n}{(n-1)^2} Beta\left(\frac{q}{2}, \frac{n-q-1}{2}\right)$$

3.3. The Use of ROCKIT

The software ROCKIT was developed in 2004 by Park, Goo and Jo [11] for analysis of ROC curves particularly with respect to the comparison of two AUCs. It uses the Dorfman and Alf [10] method of estimation of AUCs for comparing two AUCs. By analyzing the data for each ROC curve separately ROCKIT can be used to obtain the Dorfman and Alf [10] maximum likelihood estimates of the AUC's and their standard errors

4. Simulation Study

Simulation studies of the proposed test were carried out for the cases: 2 independent ROC curves and 3 independent ROC curves. Both the type I error and the power of the test were studied under each case. The study used a significance level of 5% for testing.

(i) Case 1: Comparison of 2 independent ROC curves

For the case when the number of ROC curves being

compared were 2 and were independent, data was simulated for 3 category rating scale data [10] for sample sizes of 20, 50, 100, 250 and 500 observations in total (i.e. Sample sizes of 10, 25, 50, 125 and 250 with respect to the positive and negative groups respectively). Following Cleeves [34] the degree of overlap of the two populations was controlled by generating observations from Gaussian distributions whose means differed by 0.5, 0.75 and 1 standard deviations. Additionally, data were simulated assuming equal variances in the two subpopulations, and assuming distributions with standard deviation ratios of 1:1.5. The a and b values (where a and b are parameters of the ROC curve, which are estimated using the method of scoring proposed by Dorfman and Alf for these combinations of values when simulated under the null hypothesis are as given in Table 1. Under the null hypothesis of equality in the two ROC curves, each of the 42 combinations of sample size, degree of overlap, and the standard deviation ratio was replicated 1000 times and is also included in table 1. This was used to study the type I error of the test for case I. Similarly simulations were carried out under the alternative hypothesis for two independent ROC curves by varying the a and b values as given in Table 2. This was used to study the power of the test for the 42 combinations. It can be seen from Table 2 that the Type I error decreases as the sample size is increased and approaches the stipulated 5% value. The 95% probability interval for α=5% and a 1000 trials is [0.036, 0.064]. A sample size of 250 seems to be the cut-off for an appropriate type I error. From Table 2 it is seen that the power of the test increases as the sample size is increased. When the overlap between the Gaussian distributions were less the test statistic performed better with respect to the power of the test.

Table 1. *Proportion of rejections of H_0 Under H_0: (comparing 2 independent ROC curves simultaneously).*

Sample size	a_1	a_2	b_1	b_2	Proportion of rejections
20	0.5	0.5	1.0	1.0	0.078
	0.5	0.5	0.67	0.67	0.082
	0.75	0.75	1.0	1.0	0.092
	1.0	1.0	1.0	1.0	0.075
50	0.5	0.5	1.0	1.0	0.059
	0.5	0.5	0.67	0.67	0.068
	0.75	0.75	1.0	1.0	0.069
	1.0	1.0	1.0	1.0	0.069
100	0.5	0.5	1.0	1.0	0.058
	0.5	0.5	0.67	0.67	0.05
	0.75	0.75	1.0	1.0	0.074
	1.0	1.0	1.0	1.0	0.07
250	0.5	0.5	1.0	1.0	0.051
	0.5	0.5	0.67	0.67	0.054
	0.75	0.75	1.0	1.0	0.057
	1.0	1.0	1.0	1.0	0.058
500	0.5	0.5	1.0	1.0	0.039
	0.5	0.5	0.67	0.67	0.052
	0.75	0.75	1.0	1.0	0.046
	1.0	1.0	1.0	1.0	0.069

Table 2. *Proportion of rejections of H_0 Under H_1: (comparing 2 independent ROC curves simultaneously).*

Sample size	a_1	a_2	b_1	b_2	Proportion of rejections
20	0.67	0.5	0.67	1.0	0.095
	0.67	0.33	0.67	0.67	0.115
	1.0	0.5	1.0	1.0	0.117
	1.0	0.33	1.0	0.67	0.152
50	0.67	0.5	0.67	1.0	0.09
	0.67	0.33	0.67	0.67	0.125
	1.0	0.5	1.0	1.0	0.177
	1.0	0.33	1.0	0.67	0.23
100	0.67	0.5	0.67	1.0	0.112
	0.67	0.33	0.67	0.67	0.174
	1.0	0.5	1.0	1.0	0.243
	1.0	0.33	1.0	0.67	0.34
250	0.67	0.5	0.67	1.0	0.198
	0.67	0.33	0.67	0.67	0.342
	1.0	0.5	1.0	1.0	0.501
	1.0	0.33	1.0	0.67	0.683
500	0.67	0.5	0.67	1.0	0.33
	0.67	0.33	0.67	0.67	0.619
	1.0	0.5	1.0	1.0	0.791
	1.0	0.33	1.0	0.67	0.938

(i) Case 2: Comparison of 3 independent ROC curves

Data was also simulated for the case when the number of ROC curves being compared was 3 and independent for 3 category rating scale data. The sample sizes considered were 20, 50, 100, 250 and 500 observations in total (i.e. Sample sizes of 10, 25, 50, 60, 70, 125 and 250 with respect to the positive and negative groups respectively). The results for simulation under the null hypothesis are given under Table 3 and under the alternative under Table 4. Once again it can be seen that the Type I error decreases as the sample size is increased and is within stipulated limit at a cut-off limit of 250 sample size. It is seen that the power of the test increases as the sample size is increased. The power for 3 independent ROC curves is larger than the corresponding case for 2 ROC curves. When the overlap between the Gaussian distributions were less the test statistic performed better with respect to the power of the test.

Table 3. *Proportion of rejections of H_0 Under H_0: (comparing 3 independent ROC curves simultaneously).*

Sample size	a_1	a_2	a_3	b_1	b_2	b_3	Proportion of rejections
20	0.5	0.5	0.5	1.0	1.0	1.0	0.1270
	0.5	0.5	0.5	0.67	0.67	0.67	0.1330
	0.75	0.75	0.75	1.0	1.0	1.0	0.1340
	1.0	1.0	1.0	1.0	1.0	1.0	0.1060
50	0.5	0.5	0.5	1.0	1.0	1.0	0.0700
	0.5	0.5	0.5	0.67	0.67	0.67	0.0720
	0.75	0.75	0.75	1.0	1.0	1.0	0.0640
	1.0	1.0	1.0	1.0	1.0	1.0	0.0790
100	0.5	0.5	0.5	1.0	1.0	1.0	0.0610
	0.5	0.5	0.5	0.67	0.67	0.67	0.0680
	0.75	0.75	0.75	1.0	1.0	1.0	0.0650
	1.0	1.0	1.0	1.0	1.0	1.0	0.0850
250	0.5	0.5	0.5	1.0	1.0	1.0	0.0490
	0.5	0.5	0.5	0.67	0.67	0.67	0.0580
	0.75	0.75	0.75	1.0	1.0	1.0	0.0420
	1.0	1.0	1.0	1.0	1.0	1.0	0.0580
500	0.5	0.5	0.5	1.0	1.0	1.0	0.0370
	0.5	0.5	0.5	0.67	0.67	0.67	0.0480
	0.75	0.75	0.75	1.0	1.0	1.0	0.0530
	1.0	1.0	1.0	1.0	1.0	1.0	0.0480

Table 4. *Under H_1: (comparing 3 independent ROC curves simultaneously).*

Sample size	a_1	a_2	a_3	b_1	b_2	b_3	Proportion of rejections
20	0.67	0.5	6.67	0.67	1.0	0.67	0.1370
	0.67	0.33	6.67	0.67	0.67	0.67	0.1580
	1.0	0.5	10	1.0	1.0	1.0	0.1730
	1.0	0.33	10	1.0	0.67	1.0	0.2030
50	0.67	0.5	6.67	0.67	1.0	0.67	0.0920
	0.67	0.33	6.67	0.67	0.67	0.67	0.1330
	1.0	0.5	10	1.0	1.0	1.0	0.2050
	1.0	0.33	10	1.0	0.67	1.0	0.2700
100	0.67	0.5	6.67	0.67	1.0	0.67	0.1280
	0.67	0.33	6.67	0.67	0.67	0.67	0.1950
	1.0	0.5	10	1.0	1.0	1.0	0.2470
	1.0	0.33	10	1.0	0.67	1.0	0.3680
250	0.67	0.5	6.67	0.67	1.0	0.67	0.1820
	0.67	0.33	6.67	0.67	0.67	0.67	0.3500
	1.0	0.5	10	1.0	1.0	1.0	0.5270
	1.0	0.33	10	1.0	0.67	1.0	0.7240
500	0.67	0.5	6.67	0.67	1.0	0.67	0.3320
	0.67	0.33	6.67	0.67	0.67	0.67	0.6610
	1.0	0.5	10	1.0	1.0	1.0	0.8430
	1.0	0.33	10	1.0	0.67	1.0	0.9640

5. Application

5.1. The Data

The data is from an example in the ROCKIT manual. It can be visualized as information on a 30 observer study of 3 different mammographic CAD techniques applied to 10 patients each, for the understanding of the reader. This method is an asymptotic one as explained in the methods section, however, it is difficult to find large published data sets to illustrate our method. Therefore, just to highlight the application of the method to the readers a small data set is used. The objective of the exercise is to identify the sensitivity of the 3 mamographic techniques as a means of predicting the CAD results (Coronary Artery Disease (CAD)). As different patients are used for the 3 different tests this results in independent data. The sample size of the data set consists of 60 complete cases, 20 complete cases for each test.

5.2. ROC Curves

Vergara, Normbuena, Ferrada, Slater and Melo [12] stated that in many real world applications, when a classification process is required for the prediction of discrete states, identifying the most optimum classifier has become a practically imperative exercise.

Hosmer and Lemeshow [35] explain that by plotting sensitivity values against (1-specificity), is obtained, what is known as the ROC curve, and the area under this curve (AUC) provides they explain, a measure of discrimination. As a rule of thumb Hosmer and Lemeshow [35] point out that: If AUC = 0.50, suggests no discrimination. That is, might as well flip a coin; If $0.7 \leq$ AUC < 0.8, acceptable discrimination; If $0.8 \leq$ AUC < 0.9, excellent discrimination; If AUC > 0.9, outstanding discrimination.

Agresti [36] states that in the use of most diagnostic tests, when test data do not fall into two obviously defined categories, ROC curves can be used. One of the primary reasons for the utilization of the ROC curve is, due to the fact that, it is simple and graphical, and does not depend on the prevalence. Further, Hosmer and Lemeshow [35] interestingly points out that though the model may not have a good fit, it may still have a good discrimination, calculated through the AUC.

Due to the popularity of the ROC curve, statistical packages such as ROCKIT, SAS and STAR are readily available in order to construct, estimate and compare ROC curves [12].

5.3. Use of ROCKIT for Obtaining Required Parameters for the Multivariate Test

Using ROCKIT [11] the areas under each independent ROC curves and their respective standard errors of the AUCs were obtained. The data was used for each test at a time so as to incorporate independence between the ROC curves. The method of estimation of these parameters in ROCKIT is the Dorfman and Alf method of maximum likelihood [10]

Table 5 gives for each ROC curve the estimated AUCs their standard errors.

Table 5. *Estimated AUCs, Standard errors and Sample sizes.*

Statistic (Estimate)	Test		
	1	2	3
AUC	0.7467	0.8473	0.7744
SE (AUC)	0.1107	0.0870	0.1022
Sample Size (n_1, n_2)	10,10	10,10	10,10

5.4. Application of the Multivariate Test to Independent Data

Using the Developed Test for Comparison of 3 Independent Curves (p=3)

(i) We want to test the null hypothesis (H_0) that the AUC's are all the same versus the alternative hypothesis that at least one AUC is different from the other two. Under H_0, the test statistic developed in section 3.2.2 is

$$T_G^2 = \left(\underline{\mathbf{A}}\underline{\hat{\mathbf{U}}}\underline{\mathbf{C}} - \overline{\mathbf{K}}\right)' \hat{\underline{\boldsymbol{\Sigma}}}^{-1} \left(\underline{\mathbf{A}}\underline{\hat{\mathbf{U}}}\underline{\mathbf{C}} - \overline{\mathbf{K}}\right)$$

Data in table 5 gives $\underline{\mathbf{A}}\underline{\hat{\mathbf{U}}}\underline{\mathbf{C}} = \begin{bmatrix} 0.7467 \\ 0.8473 \\ 0.7744 \end{bmatrix}$

$$\hat{\underline{\boldsymbol{\Sigma}}} = \begin{bmatrix} 0.0123 & 0 & 0 \\ 0 & 0.0076 & 0 \\ 0 & 0 & 0.0104 \end{bmatrix}$$

Using the values in table 2 gives $\overline{K} = \dfrac{[0.7467 + 0.8473 + 0.7744]}{3} = 0.7895$

Using the above expressed scaler, vector and matrix the value of the test statistic T_G^2 was determined to be 0.611. Here n_1 = number of patients with CAD (positive) = 10 and n_2 = number of patients without CAD (negative) = 10. Thus, n= minimum (n_1, n_2) = 10.

Under H0, $T_G^2 \dfrac{n}{(n-1)^2} \sim Beta\left(\dfrac{q}{2}, \dfrac{n-q-1}{2}\right)$

p = number of groups = 3 and q = p-1 = 2. As this is a two sided test α=0.025.

From Minitab, $Beta_{(1\cdot0,\ 3\cdot5),\ 2\cdot5\%} = 0.0072$ and $Beta_{(1,\ 3\cdot5),\ 97\cdot5\%} = 0.651$

Thus the 2.5% and 97.5% points of the test statistic are

$\dfrac{(9)^2}{10} Beta_{[1\cdot0\cdot3\cdot5,\ 2\cdot5\%]} = 0.0583$

$\dfrac{(9)^2}{10} Beta_{[1\cdot0,\ 3\cdot5,97\cdot5\%]} = 5.273$

As $0.0583 < 0.611 < 5.273$ we do not reject H_0 and conclude that the AUCs are the same. This leads to the conclusion that there is no difference in the diagnostic ability of the three mammographic tests in detecting CAD at the 5% significance level.

5.5. Conclusions from Results

The AUCs are not significantly different at the $\alpha = 5\%$ level. Therefore, in order to recommend one test this leads us to the conclusion that as the diagnostic power of all the tests are similar to select the most economical and administratively convenient test.

6. Discussion

In this section results obtained are discussed with respect to both statistical and medical findings. Further, some drawbacks of the research are discussed and further work suggested.

6.1. Statistical and Medical Findings

Several authors in the past ([6], [7] and [3] have dealt with the problem of comparing independent AUCs. However, all these authors have used non-parametric methods which have several drawbacks as mentioned in our introduction. Though the Maximum Likelihood method of Dorfman and Alf [10] has been proposed by several researchers as an alternative and better approach little work especially in the form of development of statistical tests and examining the properties of these tests using simulation studies have been carried out. This manuscript looks into the explained problem. The simulation studies show that the test has stipulated type one error for moderately large samples. The properties of the test such as type I error and power improve with increasing sample size and increasing number of AUC's (p's). Another advantage of our test is that it can be used to compare multiple alternative tests for independent samples, such as in the example. Up to date, there is no developed method except Delong, Delong, Clarke-Pearson method [7] for comparing several independent AUCs at once, however, as discussed this existing method has several drawbacks. This paper addresses this important need by developing a multivariate test for comparing several independent AUCs. For large samples (asymptotically) this test statistic has a distribution proportional to the Beta distribution, under the null hypothesis, provided that the estimated AUCs can be assumed to be normally distributed. This assumption of normality is one which all previous authors related to this subject have used.

Medically the most important conclusion reached was that all the tests have similar diagnostic power and our recommendation is to select the most economical and administratively convenient test.

6.2. Limitations and Further Work

One major limitation of the study was that the example used was rather a small one and it would have been better if a practically large sample size could have been used. Also, it would have been more useful if this multivariate method could have been applied to correlated AUC's as well. However, as the Dorfman and Alf (1969) algorithm has been developed only for pairwise comparison this was beyond the scope of this paper. A sequel paper is envisaged which applied this test to bivariate correlated data.

7. Conclusions

The developed test is based on multivariate theory.

The simulation study indicates that the developed test has good properties for large samples. This is because the Dorfman and Alf method [10] used to estimate the AUC's is based on the Maximum Likelihood (ML) approach. It is well known that ML estimates are asymptotically normal. Therefore the test statistic has a Beta distribution.

The example used to illustrate the method showed non-significance of results. In such an instance the more practical method should be selected.

This test can be used to compare AUCs of any number of independent ROC curves.

Authors' Contributions

The first author developed the multivariate test for comparing several independent AUCs, wrote the programs to do the multivariate test, conducted the multivariate test and wrote up the paper. The second author conducted the entire simulation study using C++ and FORTRAN 77 as well as helped to write the simulation section. He also modified the corroc2. f program.

Acknowledgements

The authors are grateful to Dr. Sameera Viswakula for helping with the use of package ROCKIT.

References

[1] Metz CE. Basic principles of ROC analysis. Seminars in Nuclear Medicine. 1978; 8 (4), 283-298.

[2] Schutts, Joshua William, "The Use of Receiver Operating Characteristic Curve Analysis for Academic Progress and Degree Completion" (2016). Dissertations. 357.

[3] Pepe M S. *The statistical evaluation of medical Tests for classification and prediction*. New York: Oxford University Press; 2003.

[4] Krzanowski WJ, Hand DJ. *ROC curves for continuous data*. CRC/Chapman and Hall; 2009.

[5] Grégoire Thomas, Louise C. Kenny, Philip N. Baker and Robin Tuytten. A novel method for interrogating receiver operating characteristic curves for assessing prognostic tests. *Diagnostic and Prognostic Research* (2017) 1: 17.

[6] Hanley JA, McNeil BJ. The meaning and Use of the Area under a Receiver Operating Characteristic (ROC) curve. *Radiology* 1982; 143 (1): 29-36.

[7] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometric. 1988, 44 (3),* 837-45.

[8] Cleves MA. Comparing Areas Under Receiver Operating Characteristics Curves from Two or More Probit or Logit Models. The STATA Journal, 2002; 2: 301-31.

[9] Hanley JA, McNeil BJ. A method of comparing the Areas under Receiver Operating Characteristic Curves Derived from the same cases. *Radiology* 1983; 148 (3): 839-843.

[10] Dorfman DD, Alf E. Maximum likelihood estimation of parameters of signal detection theory and determination of Confidence intervals-rating method data. *Journal of Mathematical Psychology* 1969; 6 (3): 487-496.

[11] . Park SH, Goo JM, Jo C. Receiver Operating Characteristic (ROC) curve: Practical review for radiologists. *Korean Journal of Radiology.* 2004; 5 (1): 11-18.

[12] Vergara IA, Normbuena T, Ferrada E, Slater AW, Melo F. StAR: a simple tool for the statistical comparison of ROC curves. *BMC Bioinformatics.* 2008; 9: 265 Open Access.

[13] Pablo Martı´nez-Camblor. Fully non-parametric receiver operating characteristic curve estimation for random-effects meta-analysis. *Statistical Methods in Medical Research* 2017, Vol. 26 (1) 5–20.

[14] Sonia Perez Fernandez (2017). Non-Standard ROC Curve Analysis. Package 'nsROC'. Cran package in R.

[15] Honghu-Liu, Li G, William G. Cumberland and Tongtong Wu. Testing Statistical Significance of the Area under a Receiving Operating Characteristics Curve for Repeated Measures Design with Bootstrapping Journal of Data Science 3 (2005), 257-278.

[16] Krzysztof Gajowniczek, Tomasz Ząbkowski, Ryszard Szupiluk (2014). ESTIMATING THE ROC CURVE AND ITS SIGNIFICANCE FOR CLASSIFICATION MODELS' ASSESSMENT. QUANTITATIVE METHODS IN ECONOMICS Vol. XV, No. 2, 2014, pp. 382-391.

[17] Hanley JA. The robustness of the "binormal" assumptions used in fitting ROC curves. *Medical Decision Making. 1988;* 8 (3), 197-203.

[18] Pepe, M., Longton, G., & Janes, H. (2009). Estimation and Comparison of Receiver Operating Characteristic Curves. *The Stata Journal*, *9* (1), 1.

[19] . Metz CE, Herman BA, Roe CA. Statistical Comparison of Two ROC-curve estimates obtained from partially-paired datasets. *Medical Decision Making* 1998a; 18 (1): 110-121.

[20] . Hall P, Zhou X. Nonparametric Estimation of Component Distributions in a Multivariate Mixture: *The Annals of Statistics. 2003:* 31 (1) 201-224.

[21] Bamber D. The Area above the Ordinal Dominance Graph and the Area below the Receiver Operating Characteristic Graph. J. Math Psychol, 1975; 12 (4): 387-415.

[22] . Grey, D. M. & Morgan, B. T. " Some aspects of ROC curve fitting: Normal and Logistic models."Journal of Mathematical Psychology, Volume 9, pp. 128-139, 1972.

[23] Dorfman, D. D. & Berbaum, K. S. "Degeneracy and discrete reciever operating characteristic rating data." Academic Radiology*,* 2 (10), pp. 907-915, 1995.

[24] Liu C, Dobson J, Cawley P. 2017 Efficient generation of receiver operating characteristics for the evaluation of damage detection in practical structural health monitoring applications. Proc. R. Soc. A 473: 20160736.

[25] Casella, G & Berger, R. L. "Statistical Inference." Duxbury Press. Second Edition, 2002.

[26] Hotelling H. *Multivariate Quality Control*. In C. Eisenhart, M. W. Hastay and W. A. Wallis, eds. Techniques of Statistical Analysis. New York: McGraw-Hill, 1947.

[27] Williams JD, Woodall WH, Birch JB, Sullivan JH. Distribution of Hotelling's T-squared Statistic Based on the Successive Differences Estimator. *Journal of Quality Technology.*

[28] Wilks SS. Multivariate Statistical outliers. *Sankhy*a *(Indian Statistical Journal)* 1963.

[29] Gnanadesikan R, Kettering JR. Robust Estimates, Residuals and Outlier detection with Multi-response data. *Biometrics* 1972; 28: 81-124.

[30] Mardia KV, Kent JT, Bibby JM. *Multivariate Analysis*, Academic Press, 1979.

[31] Seneratna, D., Sooriyarachchi, M. R. and Meyan, N. Bivariate Test for Testing the EQUALITY of the Average Areas under Correlated Receiver Operating Characteristic Curves (Test for Comparing of AUC's of Correlated ROC Curves*). American Journal of Applied Mathematics and Statistics*, 2015, Vol. 3, No. 5, 190-198.

[32] Hotelling H, Frankel LR. The transformation of Statistics to simplify their Distribution. *Annals of Mathematical Statistics*, 1938; 9 (2): 87-96.

[33] Wallace DL. Asymptotic Approximations to Distributions, *Annals of Mathematical Statistics*, 1958; 29 (3): 635-654.

[34] Cleeves, A. M. (2002). Comparative assessment of three common algorithms for estimating the variance of the area under the nonparametric receiver operating characteristic curve. *The Stata Journal, 2* (3), 280-289.

[35] Hosmer DW, Lemeshow S. *Applied logistic regression*: Wiley Series in probability and statistics. 2000.

[36] Agresti A. *An introduction to categorical data analysis*: Wiley-Interscience. 2007.