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)

D. M. Senaratna*, M.R. Sooriyarachchim, N. Meyen
Department of Statistics, University of Colombo, Colombo 3, Sri Lanka
*Corresponding author: roshinis@hotmail.com

Abstract  Methodology developed for comparing correlated ROC curves are mainly based on nonparametric methods. These nonparametric methods have several disadvantages. In this paper the authors propose an asymptotic bivariate test for comparing pairs of AUCs for independent data based on the Dorfman and Alf maximum likelihood approach. The properties of the test are examined by using simulation studies. The method is illustrated on an example of angiogram results from Sri Lanka. The test applied to the example found that there was a significant difference in the predictive power of three different cut-offs examined.

Keywords: bivariate test, Receiving Operating Characteristic (ROC) curve, Area under the Curve (AUC), Beta Distribution, Angiogram, Cardiac Stress Test (CST)


1. Introduction

1.1. Background

In modern times ROC curves have been widely used for medical decision making among other applications. The goodness of a diagnostic test can be measured using sensitivity and specificity of the test. An ROC curve is a plot of the sensitivity versus 1 - specificity as the test threshold is varied [25]. One of the main uses of ROC curve analysis is to detect the comparative benefits of alternative diagnostic tests in the field of medicine. The most popular summary measure of an ROC curve is the area under the curve (AUC) and alternative diagnostic tests have been compared by comparing their AUCs [17,29].


1.2. Objectives

The nonparametric methods of used by authors using nonparametric methods have been criticized in the literature due to these methods underestimating the AUC when dealing with both independent and correlated data. A method which has received much acclaim in this area is the method of Dorfman and Alf [8] extended by Metz, Wang, and Kronman [26] to compare two correlated ROC curves. For correlated ROC curves, Goo & Jo, 2004 and Veragra, Normbuena, Ferrada, Slater & Melo, 2008 went on to develop software, namely, ROCKIT and StAR for comparing 2 correlated ROC curves using Metz, Wang, and Kronman [26] method. However they did not study
the properties of their test and only analyzed a few examples. Thus the first objective of this paper is to develop an asymptotic bivariate statistical test that is based on Metz, Wang, and Kronman [26] approach of estimation for correlated data and examine the properties of these tests using large scale simulations.

One issue that should be borne in mind is that Metz, Wang, and Kronman [26] developed the approach for correlated data only for comparing 2 ROC curves at a time. The secondary objective of this paper is to determine the most appropriate cut-off of a cardiac stress test to determine angiogram results.

1.3. Brief Explanation of Methodology

The theory is developed for the correlated case where an asymptotic bivariate test was derived for comparing two AUCs at once. For large samples the test statistic derived follows a distribution which is proportional to the Beta distribution with parameters depending on the number of AUC curves compared (2) and the number of independent quantities making up the AUC (n). The values of the estimates of the AUCs, and their standard errors and correlations between pairs of AUCs were based on Metz, Wang, and Kronman [26] maximum likelihood approach. The paired case is a special case of the general case as here \( p = 2 \) and this test thus becomes an asymptotic bivariate test.

1.4. Data for the Example

The method developed is applied to correlated data consisting of all patients who had a Cardiac stress test (CST) at the Sri Jayawardenepura General Hospital in Sri Lanka during 2008 and 2009. Data collected includes the CST result, angiogram result, demographic information on patient, history of cardiovascular disease of patient and information on related diseases. One problem with the data was the fact that due to the high cost of an angiogram many patients that had passed the CST were not subjected to an angiogram resulting in missing data. The problem was overcome by using multiple imputation [18,19,23]. The procedure of Royston [31] was used to impute the missing values. The software used for this purpose was Stata10’s ICE module. The estimated (data based) values

<table>
<thead>
<tr>
<th>Variable</th>
<th>Type</th>
<th>Levels</th>
<th>Total Number used in the study</th>
<th>Number imputed that were used in the study</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0 Angiogram Results</td>
<td>Nominal</td>
<td>0 – No Disease 1 – Disease Present</td>
<td>144 96</td>
<td>34 78</td>
<td>Disease status identified through the angiogram</td>
</tr>
<tr>
<td>2.0 Cardiac Stress Test Result</td>
<td>Nominal/Ordinal</td>
<td>0 - Not carried out 1 - Stage 1 Difficulty 2 - Stage 2 Difficulty 3 - Stage 3 Difficulty or other signs for concern 4 – Completed Bruce protocol test or completed up to stage 3 or beyond</td>
<td>49 81 55 42 13</td>
<td>0 0 0 0 0</td>
<td>Performance with respect to the CST</td>
</tr>
<tr>
<td>3.0 Strong Indication of the Disease expressed by medical consultants</td>
<td>Binary</td>
<td>0 – No 1 – Yes</td>
<td>Not used For study</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4.0 Age</td>
<td>Continuous</td>
<td></td>
<td>240 0</td>
<td></td>
<td>Patients Age</td>
</tr>
<tr>
<td>5.0 Gender</td>
<td>Binary</td>
<td>0 – Female 1- Male</td>
<td>91 149</td>
<td>0 0</td>
<td>-</td>
</tr>
<tr>
<td>6.0 Hypertension</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.1 Hypertension - History</td>
<td>Binary</td>
<td>0 – No 1 – Yes</td>
<td>109 131</td>
<td>0 0</td>
<td>-</td>
</tr>
<tr>
<td>6.2 Hypertension - Pressure</td>
<td>Continuous</td>
<td></td>
<td>240 0</td>
<td>0 0</td>
<td>Pressure on admission</td>
</tr>
<tr>
<td>6.3 Hypertension – Pulse</td>
<td>Continuous</td>
<td></td>
<td>240 0</td>
<td>0 0</td>
<td>Pulse on admission</td>
</tr>
<tr>
<td>7.0 Cholesterol</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.1 Cholesterol – LDL</td>
<td>Continuous</td>
<td></td>
<td>Not used</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>7.2 Cholesterol – Triglyceride</td>
<td>Continuous</td>
<td></td>
<td>Not used</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>7.3 Cholesterol - HDL</td>
<td>Continuous</td>
<td></td>
<td>Not used</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>7.4 Cholesterol - Total</td>
<td>Continuous</td>
<td></td>
<td>Not used</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>8.0 Diabetes Mellitus</td>
<td>Binary</td>
<td>0 – No 1 – Yes</td>
<td>157 83</td>
<td>0 0</td>
<td>-</td>
</tr>
<tr>
<td>9.0 Family History of disease or known related factors</td>
<td>Binary</td>
<td>0 – No 1 – Yes</td>
<td>135 105</td>
<td>0 0</td>
<td>-</td>
</tr>
<tr>
<td>10.0 Cigarette Consumption</td>
<td>Binary</td>
<td>0 – No 1 – Yes</td>
<td>158 82</td>
<td>0 0</td>
<td>-</td>
</tr>
<tr>
<td>11.0 Alcohol Consumption</td>
<td>Binary</td>
<td>0 – No 1 – Yes</td>
<td>167 73</td>
<td>0 0</td>
<td>-</td>
</tr>
<tr>
<td>12.0 Marital Status</td>
<td>Binary</td>
<td>0 – No 1 – Yes</td>
<td>5 240</td>
<td>0 0</td>
<td>-</td>
</tr>
<tr>
<td>13.0 Date</td>
<td>Date</td>
<td></td>
<td>240 0</td>
<td></td>
<td>-</td>
</tr>
</tbody>
</table>
of the AUCs and their variance-covariance matrix were obtained using the package ROCKIT [28]. Table 1 gives the variables collected for the purpose of the study together with their levels, total number of observations selected from each level for the analysis and the number of imputed values selected for the analysis. This last statistic is related to section 5.2.

Section 2 gives a review of the literature pertaining to the problem. In section 3 the theorems, definitions, results and proofs related to deriving the new test statistic and developing the bivariate 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 two correlated Receiver Operating Characteristic (ROC) curves. The Area under the curve (AUC) is the most popular summary measure of ROC curves [14,29].

To test for significant differences between two correlated AUCs of ROC curves, the main factor that needs to be is the outcome distribution. This will determine the approach to be used in estimating the AUCs and its variance-covariance matrix. Possible approaches are parametric, semi-parametric and non-parametric. In the case of comparing correlated curves the two best known methods, namely, Hanley & McNeil, 1983 and Delong, Delong & Clarke-Pearson both use nonparametric methods where the AUC and its variance covariance matrix are estimated using Wilcoxons method and Mann Whitney method respectively. For each approach, different methods of estimating the AUC have been used. For the parametric approach, that is suggested in the paper Dorfman and Alf [8] method of fitting smooth curves based on the binormal assumption is used where the ROC curve can be completely described by two parameters estimated using Maximum Likelihood Estimation (MLE).

The literature on two sample tests for the comparison of AUCs of ROC curves have two important papers regarding dealing with missing data. Spritzler, DeGruttola and Pei [34] discuss that the usual tests used in this case have poor properties when the data is missing and suggest two alternative tests to be used when data is missing completely at random (MCAR) and missing at random (MAR) respectively. They show that their tests have improved properties. Harel and Zhou explain that when all subjects are screened using a common test and only a subset of these subjects are tested using a gold standard test, then there is a bias in the estimated parameter estimates. Harel and Zhou explain that for the case of correlated data the procedure works efficiently only for pairwise comparison. The algorithm given for estimating correlation for paired data (Corroc2.F available in website http://www.bio.ri.ccf.org/doc/corroc2.f) has therefore been given only for pairwise comparisons. Thus, section 3.2 discusses the special case of the asymptotic bivariate test for p=2 (correlated case).

3. Method

3.1. Estimating the AUC’s of Two Correlated ROC Curve Using the Dorfman and Alf Method

Grey and Morgan [10], explain the signal-detection paradigm on which the estimation is based. Dorfman and Alf [8] show how the AUC curve can be estimated using Maximum Likelihood methods where the curve can be determined by two parameters, namely, a and b.

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 [10].

Simulation: The method of scoring used is an iterative process which uses initial parameter estimates. Grey and Morgan [10] clearly explains this. Degenerate solution for the parameter estimates of the ROC curve can occur from empty cells in the data matrix. Dorfman and Berbaum, 1995 explain how to overcome this.

Calculation of the AUC and variance of the AUC:

It is possible to obtain the AUC of an ROC curve using the parameters a and b [8] In order to calculate the variance of the AUC, the delta method (Casella and Berger [5]) is made use of.

3.2. Development of a Bivariate Test for Comparing the AUC’s of 2 Correlated ROC Curves

The theory developed under section 3.2 is applicable to correlated data. Metz, Wang, Kronman [26] have extended the Dorfman and Alf [8] procedure to be applied only for the comparison of two correlated ROC curves at a time. They argue that for the case of correlated data the procedure works efficiently only for pairwise comparison. The algorithm given for estimating correlation for paired data (Corroc2.F available in website http://www.bio.ri.ccf.org/doc/corroc2.f) has therefore been given only for pairwise comparisons. Thus, section 3.2 discusses the special case of the asymptotic bivariate test for p=2 (correlated case).

3.2.1. Relevant Theorems, Definitions and Results

Using the theorems proposed by Hotelling [16], Williams, Woodall, Birch & Sullivan and Mardia, Kent & Bibbey [22] and the following result the proof in section 3.2.2. was derived.

Result 1

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

3.2.2. Proof (proving that the asymptotic distribution of
the test statistic developed for testing the equality of
two correlated AUCs is proportional to the Beta
distribution)

Let

$$AUC = \begin{bmatrix} AUC_1 \\ AUC_2 \end{bmatrix}$$

Where $AUC_i$ is the AUC of the $i^{th}$ ROC curve where
$i=1,2$

Let $\hat{AUC}$ be an estimate of $AUC$, let $\mu$ be the
expected value of $\hat{AUC}$ and let $\Sigma$ be the associated
variance-covariance matrix of $\hat{AUC}$. Then as $\hat{AUC}$ is
the Dorfman and Alf [8] maximum likelihood estimate
(MLE) of $AUC$ and as MLE’s are asymptotically normal
(for large samples). That is $\hat{AUC} \sim N_2(\mu, \Sigma)$

Suppose the estimate $\hat{AUC}$ of $AUC$ of an 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) [35]. The
$\hat{AUC}$ is made up of $n_1n_2$ quantities (pairs) of which $n = \min(n_1, n_2)$ are independent. Thus $n$ is the number
associated with $\hat{AUC}$. Let $n_1 = n + n_2$.

$\hat{\Sigma}$ is the Dorfman and Alf [8] MLE of the covariance
matrix $\Sigma$ of the $\hat{AUC}$. According to Theorem 3 [22] the
sampling distribution of the MLE of the
$$(\hat{AUC} - \mu)(\hat{AUC} - \mu)'$$
matrix is asymptotically $W_p(\Sigma, n)$ as $\hat{AUC}$ has an asymptotic multivariate normal
distribution. Thus asymptotically according to theorem 3
[22] and Result 1, $n\hat{\Sigma} \sim W_p(\Sigma, n)$.

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

That is $H_0: \mu = K$ where $K$ is a constant vector,
versus $H_1: \mu \neq K$.

As we do not know $K$ it has to be estimated. $K$ can be
estimated as $K$ the simple average of the $\hat{AUC}$ (that is
individual $\hat{AUC}_i$ ’s). That is $K = \frac{1}{2} \sum_{i=1}^{2} AUC_i$.

From Theorem 2, the general form of the Hotelling’s
$T^2$ statistic [16] is

$$T^2_G = (\hat{AUC} - K)' \hat{\Sigma}^{-1} (\hat{AUC} - K).$$

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

Here $p=2$ 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.
The test statistic $T^2_G$ can be used to test $H_0$. The
percentage points for the test statistic’s distribution can be
obtained by

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

3.2.3. Duncan’s Multiple Range Test

Theorem 4 [1]:

When there are more than two correlated curves to be
compared then multiple comparisons have to be used of
each pair. In this case as repeated tests are used, the
overall type I error rate increases with the number of
pairwise comparisons. One method of keeping the overall
type I error rate to $\alpha$ would be to use a much lower
pairwise type I error rate ($\alpha'$). If $N$ is the number of all
possible pairwise comparisons then ($\alpha'$) can be shown to be

$$\alpha = 1 - \left(1 - \alpha'\right)^N$$

where $N$ is the number of all possible pairwise
comparisons ($N = p(p-2)$). One such method of
determining the value of $\alpha'$ is attributed to Bonferroni [3].

3.3.2. The test for Pairwise Comparisons

The test developed for comparing two correlated AUC
curves can be applied here for all-pairwise-comparisons
taking $p=2$. The stringent significance level of $\alpha'$ is used
for all tests.

3.3.3. The Use of ROCKIT

The software ROCKIT was developed in 2004 by Park,
Goo and Jo [28] for analysis of ROC curves particularly
with respect to the comparison of two AUCs. It uses the
Dorfman and Alf [8] method of estimation of AUCs for
comparing two AUCs. However it does not use stringent
significance levels to adjust for multiple comparisons. Thus
ROCKIT was used solely for the purpose of
obtaining the parameter estimates of the AUCs and its
variance-covariance matrix, but the pairwise tests for
comparisons of AUCs was done manually.

4. Simulation Study

Simulation studies of the proposed test were carried out
for the case of 2 correlated 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: Comparison of 2 correlated ROC curves.
Furthermore correlated ROC curves were also simulated taking \(a, b, r_n\) and \(r_s\) values similar to those in Metz et al. [26] which are typical parameter values for clinical data. Here \(r_n\) and \(r_s\) denote the correlation coefficients of the bivariate normal "noise only" and bivariate normal "signal present" densities respectively. Data were simulated for 5 category rating scale data for sample sizes of 100 and 500 (i.e. sample sizes of 50 and 250 with respect to the positive and negative groups respectively). From Table 2 it can be seen that as the sample size increases the Type I error generally tends to be less ‘conservative’ and more in line with the stipulated size of the test. To compare the power of the test data was simulated for 5 category rating scale data for sample sizes of 100 and 500 (i.e. sample sizes of 50 and 250 with respect to the positive and negative groups respectively). From the results given in Table 3 it can be seen that the power of the test increases as the sample size is increased. For the paired case a sample size of 100 seems to be satisfactory. When the overlap between the Gaussian distributions were less the test statistic performed better with respect to the power of the test.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>(a_1)</th>
<th>(a_2)</th>
<th>(b_1)</th>
<th>(b_2)</th>
<th>(r_n)</th>
<th>(r_s)</th>
<th>Proportion of rejections</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>1.3</td>
<td>1.3</td>
<td>0.85</td>
<td>0.85</td>
<td>0.5</td>
<td>0.85</td>
<td>0.0380</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>1.0</td>
<td>0.85</td>
<td>0.85</td>
<td>0.5</td>
<td>0.85</td>
<td>0.0360</td>
</tr>
<tr>
<td></td>
<td>1.7</td>
<td>1.7</td>
<td>0.85</td>
<td>0.85</td>
<td>0.5</td>
<td>0.85</td>
<td>0.0310</td>
</tr>
<tr>
<td></td>
<td>1.9</td>
<td>1.9</td>
<td>0.85</td>
<td>0.85</td>
<td>0.5</td>
<td>0.85</td>
<td>0.0470</td>
</tr>
<tr>
<td>500</td>
<td>1.3</td>
<td>1.3</td>
<td>0.85</td>
<td>0.85</td>
<td>0.5</td>
<td>0.85</td>
<td>0.0370</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>1.0</td>
<td>0.85</td>
<td>0.85</td>
<td>0.5</td>
<td>0.85</td>
<td>0.0370</td>
</tr>
<tr>
<td></td>
<td>1.7</td>
<td>1.7</td>
<td>0.85</td>
<td>0.85</td>
<td>0.5</td>
<td>0.85</td>
<td>0.0470</td>
</tr>
<tr>
<td></td>
<td>1.9</td>
<td>1.9</td>
<td>0.85</td>
<td>0.85</td>
<td>0.5</td>
<td>0.85</td>
<td>0.0450</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample size</th>
<th>(a_1)</th>
<th>(a_2)</th>
<th>(b_1)</th>
<th>(b_2)</th>
<th>(r_n)</th>
<th>(r_s)</th>
<th>Proportion of rejections</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>1.0</td>
<td>1.3</td>
<td>0.5</td>
<td>1.0</td>
<td>0.5</td>
<td>0.85</td>
<td>0.6000</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>1.7</td>
<td>0.5</td>
<td>1.0</td>
<td>0.5</td>
<td>0.85</td>
<td>0.3110</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>2.0</td>
<td>0.5</td>
<td>1.0</td>
<td>0.5</td>
<td>0.85</td>
<td>0.1340</td>
</tr>
<tr>
<td></td>
<td>1.3</td>
<td>1.7</td>
<td>0.5</td>
<td>1.0</td>
<td>0.5</td>
<td>0.85</td>
<td>0.7830</td>
</tr>
<tr>
<td></td>
<td>1.3</td>
<td>2.0</td>
<td>0.5</td>
<td>1.0</td>
<td>0.5</td>
<td>0.85</td>
<td>0.5770</td>
</tr>
<tr>
<td></td>
<td>1.7</td>
<td>2.0</td>
<td>0.5</td>
<td>1.0</td>
<td>0.5</td>
<td>0.85</td>
<td>0.9160</td>
</tr>
<tr>
<td>500</td>
<td>1.0</td>
<td>1.3</td>
<td>0.5</td>
<td>1.0</td>
<td>0.5</td>
<td>0.85</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>1.7</td>
<td>0.5</td>
<td>1.0</td>
<td>0.5</td>
<td>0.85</td>
<td>0.9660</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>2.0</td>
<td>0.5</td>
<td>1.0</td>
<td>0.5</td>
<td>0.85</td>
<td>0.7010</td>
</tr>
<tr>
<td></td>
<td>1.3</td>
<td>1.7</td>
<td>0.5</td>
<td>1.0</td>
<td>0.5</td>
<td>0.85</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>1.3</td>
<td>2.0</td>
<td>0.5</td>
<td>1.0</td>
<td>0.5</td>
<td>0.85</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>1.7</td>
<td>2.0</td>
<td>0.5</td>
<td>1.0</td>
<td>0.5</td>
<td>0.85</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

5. Application

5.1. The Data

The data are from the Sri Jayawardanapura General Hospital, in Sri Lanka and was collected with the objective of identifying the sensitivity of the Cardiac Stress Test (CST) as a means of predicting the angiogram results (Coronary Artery Disease (CAD)). This involved the determining of a suitable “cut-off” in the CST for predicting the CAD results efficiently. As the same data set was used for all “cut-offs” this results in paired data.

The Sri Jayawardanapura General Hospital was chosen due to the administrative infeasibility of obtaining a simple random sample from the entire record of past patients a convenience sample had been selected using the bed head tickets (BHT) of those who had undergone a CST in 2008 or 2009. The data collected included: the CAD result which was finally coded as passed or failed, the CST result which was determined to have four levels (1 - patient had difficulty at stage 1, 2 - difficulty at stage 2, 3 - difficulty at stage 3 or had other signs for concern, 4 – Completed Bruce protocol test or completed test up to stage 3 or beyond and hence declared as out of danger by medical experts), information on possible prognostic factors such as age, sex, hypertension, Diabetes Mellitus, alcohol and Cigarette consumption and family history. A seemingly important prognostic factor, cholesterol level was not included as most patients were taking cholesterol reducing drugs. How the status of the CAD result is obtained from the gold standard angiogram is explained in Cardiac Catheterization and Angiogram [37]. Clearly, it is obtained by methods independent of CST. The sample size of the data set was determined using the design review procedure of Bolland, Sooriyarachchi and Whitehead [2]. The sample size was calculated using Stata 10 and this was found to be 419 but there were only 202 complete records at the end of the data collection.

5.2. Missing Values and Their Imputation

A problem regarding this data was the amount of missingness. A majority of patients passing the CST had not done the angiogram due to the high cost of angiograms. To determine sensitivity and specificity, we sought to collect data on CAD status and CST level; however, typically those with higher levels of CST are less willing to do an angiogram resulting in missing data
for CAD status. When CST = 4 there are 13 observations for which all CAD values are missing. Seneratna and Sooriyarachchi [33] discuss verification bias with respect to this data set and explain that there is little cause for alarm as explained previously in this paper. As the missingness in the CAD result does not depend on the result of the CAD status itself, but on the CST, according to the definitions of Little and Rubin [19] and Carpenter and Kenward [4] this data can be considered as missing at random (MAR). Apart from the missing values in the response variable (CAD) there were also several missing values in the explanatory variables. The opinion of the medical doctors involved in the study regarding the missingness of the explanatory variables was that these values could be biased as those missing values were not conditional on another variable and hence according to the definition of Little and Rubin [19] are missing not at random (MNAR). Current research [6], indicates that while using imputed missing values that are missing at random (MAR) or missing completely at random (MCAR) do not bias the results, the same may not be the case for missing values which are missing not at random (MNAR). In this case one has to be very careful in using these values due to possible bias. In order to increase the sample size and thus the power of hypothesis tests used and to study the sensitivity and specificity of the CST on a reasonably large sample it was decided to impute only the CAD values.

For the purpose of imputation, Stata10’s ICE module [31] was used. Since the response variable in this study was a dichotomous one, and hence, the final model to be used was decided to be a logistic model, as recommended by Schafer [32] this same logistic model was used for the imputation procedure. Once this model was fitted the predicted values along with its standard error was calculated by Stata. Then from this posterior distribution, Stata randomly selects a value which is none other than the imputed observation. In this study following Van-Leeuwen, Zweers, Opmeer, Ballegooie, Brugge and Valk [18], 100 such imputations were averaged out to determine the missing values for CAD. The averaged out observations were grouped as ‘diseased’ if the estimated probability of disease from the model was greater than 0.5 or else grouped as ‘not diseased’. If a different threshold (different from 0.5) can be reasoned out to be more appropriate then that threshold can be used. However, the missing values in the sample’s explanatory variables were not imputed and this resulted in a substantial decrease in the effective sample size to 240 patients.

Further, the work carried out by Spritzler, DeGruttola and Pei [34] depicted the properties of the AUC in the presence of missingness and recommended against the use of the trapezoidal rule. Therefore, the Dorfman and Alf [8] method was adopted. In this study the statistical modules ICE developed for the statistical package Stata 10, were used. ICE was first developed by Royston [31] in 2004 and the newest version MI is available for Stata 11.

5.3. ROC Curves

Hosmer and Lemeshow [15] 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 [15] point out that: If AUC = 0.50, suggests no discrimination. That is, might as well flip a coin; If 0.7 ≤ AUC < 0.8, acceptable discrimination; If 0.8 ≤ AUC < 0.9, excellent discrimination; If AUC > 0.9, outstanding discrimination.

In order to determine whether the performance of CST may be over-stated because first it was used to impute the status of CAD and then it was used again to generate ROC curves, a before and after analysis (not reported here) was done for determining the association between CST and CAD variables. Here the Pearson’s Chi-Square test was used. The p-value of the test before and after was 0.019 and less than 0.001 respectively. Thus the significance before imputation is maintained and increased after imputation. This could be because of the increase in power due to the increase in sample size.

5.4. Modeling the Different “Cut-offs”

Receiver Operating Characteristic (ROC) curves were used to identify the best cut-off for the CST. Three different cut-off’s were examined, the first being rating 1 of CST versus the other ratings (grouping 1), the second being rating 1 and 2 of CST versus the other ratings (grouping 2) and the third being rating 1, 2, 3 of CST versus rating 4 of CST (grouping 3). In order to construct the ROC curves 3 different cut-off models were created.

For each of these cut-off’s logistic models were fitted between CST and CAD after adjusting for important confounding variables. As it was of interest to see the effects of including all the candidate regressors just so that nothing obvious is missed, the backward elimination procedure was used for selecting important confounding variables.

On using backward elimination the cut-off 1 model results in selecting the variables Age, Sex, Hypertension (HT), Diabetes Mellitus (DM), family history (FH) and Alcohol (Alc). The cut-off 2 model includes the variables CST, Age, Sex, HT, DM, FH and Alc. The cut-off 3 model includes the main effects CST, Age, Sex, HT, DM, FH and Alc and the two factor interactions CST*age, CST*Sex, CST*DM. Here different combination of variables are used for the different cut-offs. One might argue that common covariates are more suitable. The justification for our choice is that we want to find the best cut-off after adjusting for important covariates. Our third objective was this.

<table>
<thead>
<tr>
<th>Statistic (Estimate)</th>
<th>Grouping 1 versus 2</th>
<th>Grouping 1 versus 3</th>
<th>Grouping 2 versus 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUC</td>
<td>0.9011</td>
<td>0.9040</td>
<td>0.9079</td>
</tr>
<tr>
<td>SE (AUC)</td>
<td>0.0195</td>
<td>0.0191</td>
<td>0.0187</td>
</tr>
<tr>
<td>Correlation between the two AUCs of the ROC curves</td>
<td>0.9488</td>
<td>0.8046</td>
<td>0.8201</td>
</tr>
</tbody>
</table>
5.5. Use of ROCKIT for Obtaining Required Parameters for the Multivariate Test

Using the predicted values from the models related to each cut-off, in ROCKIT [28] the areas under each ROC curve, their respective standard errors and the correlations between each pair of AUCs were obtained. The method of estimation of these parameters in ROCKIT is the Dorfman and Alf method of maximum likelihood [8].

Table 4 gives for each pair of ROC curves (for each cut-off or grouping) the estimated AUCs their standard errors and correlation.

5.6. Application of the Bivariate Test to Paired Data

5.6.1. Using the Developed Test for Comparison of Pairwise AUCs (p=2)

(i) We want to test the null hypothesis (H₀) that AUC₁ is the same as AUC₂ versus the alternative hypothesis that the two AUCs are not the same. Under H₀, the test statistic developed in section 3.2.2 is 

\[ T_G^2 = \left( \frac{AUC - K}{\sum^{1}} \right) \Sigma^{-1} \left( AUC - K \right). \]

Data in Table 2 gives AUC = \( \begin{pmatrix} 0.9011 \\ 0.9039 \end{pmatrix} \) 
\( \Sigma = \begin{pmatrix} 0.00038 & 0.00036 \\ 0.00036 & 0.00037 \end{pmatrix} \) 
\( \Sigma^{-1} = \begin{pmatrix} 0.00038 & -0.00036 \\ -0.00036 & 0.00037 \end{pmatrix} \)

Using the values in Table 2 gives 
\( T_G^2 = \left( \frac{0.9011 + 0.9039}{2} \right) = 0.9025. \)

Using MATLAB the value of the test statistic \( T_G^2 \) was determined to be 0.196. Here \( n_1 = \) number of patients with CAD (positive) = 96 and \( n_2 = \) number of patients without CAD (negative) = 144. Thus, \( n = \) minimum \((n_1, n_2) = 96. \)

The simulations in table 6 indicate that the total sample size (96+144=240) can be considered to be large enough for asymptotic properties to hold for the paired case. Thus asymptotically,

Under H₀, 

\[ T_G^2 \sim \frac{\sum^{1}}{(n-1)^2} \times \frac{\sum^{2}}{q} \times \frac{n-q-1}{2}. \]

\( p = \) number of groups = 2 and \( q = p-1 = 1. \) As this is a two sided test \( \alpha = 0.025 \) giving an \( \alpha' = 0.0085 \) for the Bonferroni correction.

From Matlab, \( Beta(0.5,47), 0.88\% = 1.2138e-006 \) and \( Beta(0.5, 47), 99.15\% = 0.00714. \)

Thus the 0.85% and 99.15% points of the test statistic are 
\[
\left( \frac{95}{96} \right)^2 Beta(0.5,47,0.88\%) = 0.0001141
\]

As 0.0001141 < 0.196 < 6.712 we do not reject \( H_0 \) and conclude that the AUCs are the same.

5.6.2. Pairwise Comparison of AUC Curves

As the results were significant indicating differences between the three AUCs the three groups were compared pairwise using Duncan’s multiple range test [1]. Here, 

\[ \alpha = 1 - (1 - \alpha')^N \] where \( N = 3 \) (=\(^C_2\) ) and \( \alpha = 5\% \). Thus 
\[ \alpha' = 1 - (1 - 0.05)^{1/2} = 0.17 \quad \alpha' = 0.0085. \]

Comparing AUCs of groups 1 and 2

\[ H_0 : AUC_3 = AUC_1 \]
\[ H_1 : AUC_3 \neq AUC_1 \]

Using Hanley and McNeil [12] the test statistic is 

\[
\left[ \frac{0.006361 + 0.000372}{-2 \times 0.9503 \times 0.0190 \times 0.0193} \right]^{1/2} = 0.5298.
\]

Based on the normality assumption, the test statistic p-value = 0.7019 and thus the results are not significant at \( \alpha' \) % level.

Similarly AUCs of groups 3 and 1 and groups 3 and 2 can be compared and the respective test statistics are 2.4118 (resulting in a p-value of 0.007937) and 2.1688 (resulting in a p-value of 0.01505). While these tests indicate a significant difference between the AUCs of groups 3 and 1 there is no significant difference between the AUCs of groups 3 and 2.

However if one sided test are carried out with the alternative hypotheses being 

\[ H_1 : AUC_3 > AUC_1 \] 
\[ H_1 : AUC_3 > AUC_2 \]

while the first null hypothesis is not rejected the other two null hypotheses are rejected at \( \alpha' \) % level indicating that the AUC of group 3 is larger than the AUCs of groups 1 and 2.

The reason for using the one sided tests is to satisfy our practical objective, objective 3. Here we were interested in determining whether in the local scenario too the best cutoff for discrimination is the global “Bruce-Protocol” cutoff. For that we had to recommend one of the 3 cutoffs as the best.

5.7. Conclusions from Results

The AUCs are not all the same and \( H_0 \) is rejected at the \( \alpha = 5\% \) level. Duncan’s multiple range test was initially used to do two sided tests for all-pairwise-comparison methods. This indicated that the only significant difference at \( \alpha \% \) significance level was between the AUCs of groups 3 and 1. In order to recommend one cut-off one-sided significance tests were carried out for all-pairwise-comparison methods and this indicated that the AUC of group 3 was significantly larger than the AUCs of groups 1 and 2.
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 [11,12,20,21,25,27] have dealt with the problem of comparing two AUCs. Often it may be required to compare multiple alternative tests, for example, one might want to compare the cardiac stress test, the electrocardiogram and the echocardiogram with respect to their abilities to diagnose coronary artery disease using the angiogram as the gold standard. Up to date there is no developed method except Delong, Delong, Clarke-Pearson method [7] for comparing several AUCs at once, however, as discussed this existing method has several drawbacks. This paper addresses this important need by developing a bivariate test and using multiple comparisons. The significance level is adjusted for these comparisons by using the Bonferroni correction [3] for comparing several 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. As there were several missing values, making it almost impossible to estimate the specificity and sensitivity of the test the missing value imputation method for Van- Leeuwen, Zweers, Opmeer, Ballegooie, Brugge and Valk [18] was used for the response variable. However, this same technique could not be used for the missing values in the explanatory variables as these values were found to be missing not at random (MNAR) as defined by Little and Rubin [19]. On rejecting the null hypothesis pertaining to the test it was found that ROCKIT can be used to compare all pairwise AUCs. However ROCKIT does not adjust for multiple testing and uses the required significance level for all tests, making the overall significance level inflated. In this paper, we suggest to use a stringent significant level based on the Bonferroni adjustment [3] for all pairwise-comparison methods so as to keep the overall significance level within required limits.

Medically the most important conclusion reached was that the Bruce-protocol cut-off for the CST as a diagnostic of CAD was seen to be the most favorable cut-off for this Sri Lankan data which is a finding consistent with world standards. The other two cut-offs examined are significantly different to this one. Some other medically important findings were that age, sex, Diabetes Mellitus, hypertension, alcohol and family history are all prognostic factors for CAD. In addition, CST was important as a diagnostic of CAD not only on its own, but also in combination with age, sex and Diabetes Mellitus status in order to predict CAD results. Also prognostic variables hypertension, alcohol and family history should be taken into account before coming to a conclusion. It is noteworthy to mention here that some studies in the past have found that moderate amounts of alcohol have been beneficial for CAD. However, in this study the variable alcohol intake is binary and the levels are ‘yes’ and ‘no’, thus we could not test this hypothesis.

6.2. Limitations and Further Work

One major limitation of the study was that the observations having missing values for the explanatory variables had to be removed from the analysis as these could not be imputed. Statistically, there was the issue of these values being NMAR and the doctors were aware that there was bias in this missingness. Imputation of NMAR missing values are a current area of research. This requires further study though no software is available for the purpose. Due to this limitation the sample of size 300 was reduced to 240 resulting in a reduction in power. In addition the results would have been more generalizable if data was obtained from hospitals all over Sri Lanka without restricting it to a single general hospital. This was done because the angiogram is a very expensive and relatively new technique in Sri Lanka and very few hospitals have the facility for doing it. Asymptotically the test statistic follows a distribution proportional to the Beta distribution (under the null hypothesis) as asymptotically the AUCs have an approximately normal distribution. The other major drawback was the fact that Metz, Wang and Kronman [26] have developed their method of determining the AUC’s and their standard errors and correlations for correlated data only for paired data and thus only a bivariate procedure could be used.

7. Contributions

The first author was involved in the collection and cleaning of the data, missing value imputation and analyzing the example especially with respect to model fitting and the use of ROCKIT to obtain the necessary parameters. The second author developed the bivariate test for comparing two AUCs at a time, included the Bonferroni adjustment for multiple comparison of pairs of AUCs, wrote the programs to do the bivariate test, conducted the bivariate test and wrote up the paper. The third 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. Dilhari Attygalle of the Department of Statistics who suggested the imputation of missing values. We are also indebted to Dr. N.L.Amarasena, Consultant Cardiologist of the Sri Jayawardanapura General Hospital for providing the data. Finally we would like to thank Dr. K.H. Abeywickrama for sending us some key publications required for this research.

References


