LSSP #752840, VOL 00, ISS 00

A Simulation Based Study for Comparing Tests Associated With Receiver Operating Characteristic (ROC) Curves

D. N. JAYASEKARA AND M. R. SOORIYARACHCHI

QUERY SHEET

This page lists questions we have about your paper. The numbers displayed at left can be found in the text of the paper for reference. In addition, please review your paper as a whole for correctness.

Q1. Au: Primary and Secondary classifications are required. Please provide the appropriate primary and secondary classifications from http://www.ams.org/mathscinet/msc/msc.html.
Q2. Au: Please suggest whether this is year 1982a or 1982b.
Q3. Au: Please suggest whether this is year 1982a or 1982b.
Q4. Au: Please suggest whether this is year 1982a or 1982b.
Q5. Au: Please suggest whether this is year 1982a or 1982b.
Q6. Au: Please suggest whether this is year 1982a or 1982b.
Q7. Au: Please suggest whether this is year 1982a or 1982b.
Q8. Au: Please suggest whether this is year 1982a or 1982b.
Q9. Au: Please suggest whether this is year 1982a or 1982b.
Q10. Au: Please suggest whether this is year 1982a or 1982b.
Q11. Au: Please provide Figures 1–15 citation in text.
Q12. Au: Please suggest whether this is year 1982a or 1982b.
Q13. Au: Please suggest whether this is year 1982a or 1982b.
Q14. Au: Please suggest whether this is year 1982a or 1982b.
Q15. Au: As there are two Hanley & McNeil 1982 references in the list as well as text citations, they have been named as Hanley & McNeil 1982a and Hanley & McNeil 1982b, but it is not clear as to which of the citations should bear the year label 1982a and which one 1982b. Also please check and suggest if Hanley & McNeil 1983 is identical to Hanley & McNeil 1982b.
Q16. Au: Ref. Hanley & McNeil 1984 is not cited in the text. Please cite it at its appropriate place in the text.
Q17. Au: Metz et al. 1998 is not cited in the text. Please cite it at an appropriate place in the text.

TABLE OF CONTENTS LISTING

The table of contents for the journal will list your paper exactly as it appears below:

Communications in Statistics—Simulation and Computation®, 00: 1–24, 2014

ISSN: 0361-0918 print / 1532-4141 online
DOI: 10.1080/03610918.2012.752840

A Simulation Based Study for Comparing Tests Associated With Receiver Operating Characteristic (ROC) Curves

D. N. JAYASEKARA AND M. R. SOORIYARACHCHI

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

Receiver Operating Characteristic curves and the Area Under Curve (AUC) are widely used to evaluate the predictive accuracy of diagnostic tests. The parametric methods of estimating AUCs are well established while nonparametric methods, such as Wilcoxon's method, lack proper research. This study considered three standard error techniques, namely, Hanley and McNeil, Hanley and Tilaki, and DeLong methods. Several parameters were considered, while measuring the predictor on a binary scale. The normality and type I error rate was violated for Hanley and McNeil's method while asymptotically DeLong's method performed better. Hanley and Tilaki's Jackknife method and DeLong's method performed equally well.

Keywords Area under curve (AUC); DeLong's method; Hanley and McNeil's method; Hanley and Tilaki's method; Receiver operating characteristic (ROC) curve; Simulation study.

Mathematical Subject Classification . Q1

1. Introduction

Receiver operating characteristic (ROC) curves were first adopted to predict the presence of Japanese aircrafts from radar signals, following the attack on Pearl Harbour in 1941 (Green and Swets, 1966). Since then it has been widely used to evaluate the predictive accuracy of models, algorithms or technologies that produce the predictions. It may often involve classification of a certain outcome into two or more categories. The ROC curve is a probability scale, two-dimensional plot of Sensitivity versus 1-Specificity for a given classifier with continuous or ordinal output score and is calculated using all possible cutoffs (Agresti, 2007). Sensitivity or the "True Positive Rate" (TPR) is the probability of a positive test in a person known to have a positive outcome, while the Specificity also known as "True Negative Rate" (TNR) describes the probability of a negative test in a person known to have a negative outcome (Nettleman, 1988). The AUC measures the strength of association between the underlying test and the outcome status and is widely used to measure the classification power of diagnostic tests.

Hanley and McNeil (1982) first developed the theory for comparing the AUCs pertaining to two ROC curves for unpaired data. The AUCs were estimated nonparametrically by Q2

Received July 18, 2012; Accepted November 20, 2012
Address correspondence to Prof. M. R. Sooriyarachchi, Department of Statistics, University of Colombo, Colombo 03, Sri Lanka; E-mail: roshini@mail.cmb.ac.lk

the Wilcoxon's method. However, the Wilcoxon's statistic is an estimate of the true area under ROC curve for infinitely large samples and sufficiently continuous rating scale and often underestimates the true AUC (Hanley and McNeil, 1982). Their work was extended in 1983 to determine a method of comparing areas under two ROC curves for paired designs. This approach involves calculating the correlations induced by the paired nature for both the diseased and nondiseased groups, separately. A table containing the average of the two correlations along with the average of the areas under the two curves is used to arrive at an estimated correlation between the two areas. However, for measures that are not on a continuous rating scale, Hanley and McNeil (1982, 1983) method heavily relies on Gaussian modeling assumptions for estimating the variances of the two areas. Hence, Hanley and McNeil (1982, 1983) method is not a completely nonparametric approach. However, Hanley and Tilaki (1997) proposed a method to account for the paired nature of data through Jackknife method that could be effectively used in simulations. An alternative methodology for comparing two or more ROC curves using a more completely nonparametric approach was introduced by Delong et al. (1988), which exploits the properties of Wilcoxon statistic. A covariance matrix is estimated using the method of structural components. Hanley and Tilaki (1997) observed the "twin-like" nature in results obtained by Jackknife and Delong et al. (1988) methods.

Cleves (2002) performed a simulation study to compare the two algorithms proposed by Hanley and McNeil (1982) and, DeLong et al. (1988) for estimating the standard error of the estimated AUCs under one sample design. He found that when the outcome of the diagnostic test was measured on a continuous scale, both Hanley and McNeil's (1982) and DeLong et al. (1988) methods performed similarly well. It was found that when the outcome of the diagnostic test was measured on a discrete ordinal scale, the methods developed by DeLong et al. (1988) outperformed Hanley and McNeil's (1982) method. This was true regardless of sample size and distance between population means. However, Cleves (2002) study was restricted to one sample analysis. Also, Cleves' (2002) study was limited to comparing variances. He did not examine the normality, type I error rate, and power. Thus, a complete study is imperative. Binary classifiers pose more of a challenge as several assumptions have to be made regarding the estimated AUC and its standard error for both methods. This problem has led to fewer complete studies being done on binary classifiers. For these reasons, this study is based purely on binary classifiers. Thus, the main objective of the study was to analyze the behavior and the sensitivity of the Wilcoxon test statistic under different study designs. The study facilitated in identifying the effect of sample size on the normality of the AUCs and distribution of the test statistic while determining the power of the test and type I error rates for various parameter combinations. This further enabled the comparison of Hanley and McNeil's (1982, 1983), Hanley and Tilaki (1997) and, DeLong et al. (1988) standard error calculation techniques in terms of the performance.

2. Simulation

Data were simulated by assuming the diagnostic test to produce results on a binary scale. Simulations were carried out for one sample, two independent and two correlated sample designs with varying sample sizes such as 20, 50, 75, 100, 250, and 500. The predictability of a classifier was varied by the degree of correlation between the observed and predicted outcomes. The simulations were performed under both null and alternative hypotheses of the respective study design. Each combination of parameters was replicated 5,000 times.

To determine the effect of sample size on the normality of estimated AUCs, the Chi-square Goodness of Fit test was applied to the empirical null distribution. In the case of the alternative hypothesis, the Chi-square test was performed assuming the empirical mean of AUCs to be equal to average of estimated AUCs for large simulations such as 5,000.

3. Methodology

3.1. *Algorithm for Estimating the AUC and its Variances*

Suppose a sample of N individuals undergoes a test for predicting the presence or absence of a condition. Assume the diagnostic variable to be binary. In the case of a dichotomous diagnostic variable, the value 1 represents the positive or the "abnormal" outcome while 0 represents the negative or the "normal" outcome. Let the positive group contain m number of individuals while the negative group contain n (N–m) number of individuals. Let X_i, $i = 1, 2, \ldots, m$ and $Y_j, j = 1, 2, \ldots, n$ be the outcome for the diagnostic test for both positive and negative groups, respectively. The Wilcoxon statistic estimates the probability θ, that a randomly selected observation from the population represented by the positive group will be less than or equal to a randomly selected observation from the population represented by the negative group. It can be computed as,

$$\hat{\theta} = \frac{1}{mn} \sum_{j=1}^{n} \sum_{i=1}^{m} \varphi(X_i, Y_j) \quad \text{where} \quad \varphi(X, Y) = \begin{cases} 1 & X > Y \\ 0.5 & X = Y \\ 0 & X < Y \end{cases}$$

The $\hat{\theta}$ represents the estimated AUC derived using the Wilcoxon method.

The Hanley and McNeil's (1982) variance is formulated as follows. Let Q_1 be the Q10 probability that two randomly selected positive ("abnormal") subjects both having a higher score than a randomly selected negative ("normal") subject, and let Q_2 be the probability that one randomly selected positive ("abnormal") subject will have a higher score than any two randomly selected negative ("normal") subjects. The standard error of $\hat{\theta}$ is given by the following equation.

$$\mathrm{SE}(\hat{\theta}) = \sqrt{\frac{\theta(1-\theta) + (\mathrm{m}-1)(Q_1 - \theta^2) + (n-1)(Q_2 - \theta^2)}{mn}} \ldots\ldots \tag{1}$$

When the underlying distributions of the negative group (X_n) and the positive group (X_m) are Gaussian, gamma or negative exponentials, Q_1 and Q_2 can be expressed as simple functions,

$$Q_1 = \theta/(2-\theta) \quad \text{and} \quad Q_2 = 2\theta^2/(1+\theta).$$

The standard error formula (1) could be used both under one sample and two independent sample situations.

By adhering to the notations given in 3.1, DeLong et al. (1988) variance for each AUC could be computed as follows. For each of the positive subjects i,

$$V_{10}(X_i) = \frac{1}{n} \sum_{j=1}^{n} \varphi(X_i, Y_j) \quad \text{and} \quad S_{10} = \frac{1}{m-1} \sum_{i=1}^{m} (V_{10}(X_i) - \hat{\theta})^2$$

and similarly, the following is defined for each negative subject j,

$$V_{01}(Y_j) = \frac{1}{m}\sum_{i=1}^{m}\varphi(X_i, Y_j) \quad \text{and} \quad S_{01} = \frac{1}{n-1}\sum_{j=1}^{n}(V_{01}(Y_j) - \hat{\theta})^2.$$

Then, DeLong's variance of the estimated AUC is given by,

$$\mathrm{Var}(\hat{\theta}) = \frac{S_{10}}{m} + \frac{S_{01}}{n}.$$

In machine learning and statistics, classification is the problem of identifying to which of a set of response categories an observation belongs. An algorithm that implements classification is known as a classifier.

In the presence of two classifiers r_1 and r_2, the components of the covariance term is,

$$[S_{10}]_{r_1,r_2} = \frac{1}{m-1}\sum_{i=1}^{m}\left(\frac{1}{n}\sum_{j=1}^{n}\varphi\left(X_i^{r_1}, Y_j^{r_1}\right) - \hat{\theta}^{r_1}\right)\left(\frac{1}{n}\sum_{j=1}^{n}\varphi\left(X_i^{r_2}, Y_j^{r_2}\right) - \hat{\theta}^{r_2}\right)$$

$$[S_{01}]_{r_1,r_2} = \frac{1}{n-1}\sum_{j=1}^{n}\left(\frac{1}{m}\sum_{i=1}^{m}\varphi\left(X_i^{r_1}, Y_j^{r_1}\right) - \hat{\theta}^{r_1}\right)\left(\frac{1}{m}\sum_{i=1}^{m}\varphi\left(X_i^{r_2}, Y_j^{r_2}\right) - \hat{\theta}^{r_2}\right).$$

The covariance is,

$$\mathrm{Cov}(\hat{\theta}^{r_1}, \hat{\theta}^{r_2}) = \frac{[S_{10}]_{r_1,r_2}}{m} + \frac{[S_{01}]_{r_1,r_2}}{n}\angle.$$

Therefore, the $\mathrm{Var}(\hat{\theta}^{r_1} - \hat{\theta}^{r_2})_{\mathrm{DeLong}} = \mathrm{Var}(\hat{\theta}^{r_1}) + \mathrm{Var}(\hat{\theta}^{r_2}) - 2\mathrm{Cov}(\hat{\theta}^{r_1}, \hat{\theta}^{r_2})$.

The following algorithm is the Jackknife technique proposed by Hanley and Tilaki (1997) to estimate the standard error for paired study designs. The method is developed based on pseudo values that are constructed for each observation. This can be determined by calculating the summary statistic with and without the observation in question. For example, if the summary ROC index is the AUC, then the AUC pseudo value (pAUC) corresponding to observation i is,

$$\mathrm{pAUC}_{(i)} = (m+n)\mathrm{AUC} - (m+n-1)\mathrm{AUC}_{(-i)} \ldots\ldots\ldots \tag{2}$$

The variance of individual AUCs is defined as,

$$\mathrm{Var[AUC]} = \text{Variance of mean of all } (m+n) \text{ pAUCs} = \frac{\text{Variance of all pAUCs}}{m+n}.$$

This variance depends on the pseudo values obtained by the Eq. (2). The covariance term is calculated as follows,

$$\mathrm{Covar}[\mathrm{AUC}_1, \mathrm{AUC}_2] = \frac{\mathrm{Covar}(\mathrm{pAUC}_1, \mathrm{pAUC}_2)}{m+n},$$

where, $\mathrm{Covar}[\mathrm{pAUC}_1, \mathrm{pAUC}_2] = \mathrm{Correlation} * \mathrm{SD}(\mathrm{pAUC}_1) * \mathrm{SD}(\mathrm{pAUC}_2))$.

3.2. *Simulation Study Design*

For a binary predictor, there are two distributions that should be considered for the positive or negative outcomes. Overlap exists between these two distributions as no classifier is perfect at predicting the positive or negative status. Thus, the degree of overlap between the two outcomes was considered as a parameter. Therefore, correlated binary variables were simulated using Park et al. (1996) method to represent the observed and predicted outcomes. The concept lies on the property that any Poisson random variable could be expressed as a convolution of several other independent Poisson random variables.

The study of one sample analysis considered two binary variables Y, X as response and explanatory variables, respectively. The null hypothesis tested was that there is no classification of variable Y by X. This is equivalent to the expected area under ROC curve between Y and X being 0.5 (Hosmer and Lemeshow, 2000). The alternative hypothesis represents the case where the classifier is suitable for predicting a certain outcome. The definition of suitability could be given by the overlap of the two distributions of the binary outcomes corresponding to the case where Y could be classified by X. Four scenarios were simulated under the alternative hypothesis where the correlation between Y and X was set to 0.2, 0.5, 0.75, and 0.9, which also depicts the gradual increase of predictability.

The analysis of two independent samples depicts the scenario in which the classifiers are tested on two completely independent samples. Consider a binary response variable Y_1 and a binary explanatory variable X_1 for classifier 1 and, another two binary variables Y_2 and X_2 to represent the response and explanatory variables for classifier 2. Let $Y_{1,}$ X_1 and Y_2, X_2 be correlated. However, $Y_{1,}$ X_1 is completely independent of Y_2, X_2. Thus, Y_1 is classified by X_1 and similarly, Y_2 is classified by X_2. The null hypothesis of interest is that there is no difference between the predictability for classifier 1 and classifier 2. This is equivalent to the expected area under ROC curves 1 and 2 being alike. The null hypothesis was simulated under four correlations 0.0, 0.3, 0.6, and 0.7, where equal predictabilities were given to both classifiers. The alternative hypothesis represents the case where both classifiers have different discrimination abilities between cases and controls. This could be simulated such that Y_1, X_1 and Y_2, X_2 be correlated by amounts ρ_1 and ρ_2, respectively. Four scenarios were simulated under the alternative hypothesis with correlations (ρ_1, ρ_2) being (0.6, 0.5), (0.6, 0.3), (0.7, 0.3), and (0.8, 0.2). The correlations were selected such that the differences between the correlations are increased by 0.1, 0.3, 0.4, and 0.6.

Paired samples give rise to the analysis of two correlated samples. The null hypothesis depicts the scenario in which the predictability of two classifiers is equal and is tested on two completely correlated samples. Consider two binary response variables Y_1 and Y_2, and a binary explanatory variable X to illustrate classifier 1 and 2, respectively. The variables Y_1 and X are correlated while variables Y_2 and X are also correlated. Since the explanatory variable X is common to both, it illustrates the scenario of correlated samples. Four scenarios were considered under the null hypothesis with correlations 0.0, 0.3, 0.6, and 0.7 for both classifiers. In order to depict the alternative hypothesis, where the predictabilities are different four scenarios with correlations (0.4, 0.3), (0.5, 0.3), (0.5, 0.2) and (0.6, 0.2) were considered with 0.1 increment in the difference of predictabilities. At each simulation, the estimated AUCs, the standard error of the estimated AUCs and the test statistic (Z_0) were calculated. Percentage points of Z_0 falling under different quartiles of the standard normal curve were obtained to perform the Goodness of Fit Test. The test statistics under H_0 for each case are given by the following formulae.

$$\text{One sample: } Z_0 = \frac{\widehat{\text{AUC}} - 0.5}{\text{SE}(\widehat{\text{AUC}})}. \tag{3}$$

$$\text{Two independent samples: } Z_0 = \frac{\widehat{\text{AUC}}_1 - \widehat{\text{AUC}}_2}{\text{SE}(\widehat{\text{AUC}}_1 - \widehat{\text{AUC}}_2)}. \tag{4}$$

Two correlated samples (DeLong):

$$Z_0 = \frac{\widehat{\text{AUC}}_1 - \widehat{\text{AUC}}_2}{\sqrt{\text{var}(\widehat{\text{AUC}}_1) + \text{var}(\widehat{\text{AUC}}_2) - 2\text{cov}(\widehat{\text{AUC}}_1, \widehat{\text{AUC}}_2)}}. \tag{5}$$

Two correlated samples (HT):

$$Z = \frac{\widehat{\text{AUC}}_1 - \widehat{\text{AUC}}_2}{\sqrt{\underbrace{\text{var}(\widehat{\text{AUC}}_1)}_{X} + \underbrace{\text{var}(\widehat{\text{AUC}}_2)}_{Y} - k * \rho * \underbrace{\text{SD}(\text{p}\widehat{\text{AUC}}_1) * \text{SD}(\text{p}\widehat{\text{AUC}}_2)}_{S})}}. \tag{6}$$

4. Results

4.1. *One Sample Case*

Table 1 gives the results for one sample analysis.

4.1.1. *Normality of the test statistic under H_0.* Table 1a illustrates the Chi-square values obtained for the Normal Goodness of Fit test under the null hypothesis of the methods HM and DeLong for one sample case. For all Chi-square goodness of fit tests in this research, 14 groups have been used. According to Table 1a, it is clear that the normality does not hold for all sample size combinations under the binary predictor for HM. However, in contrast to HM method, the asymptotic normality of the estimated AUCs could be clearly observed for DeLong's method as the Chi-square goodness of fit statistic reduces with increasing sample size and becomes nonsignificant for samples of size 250 and above for DeLong's method.

4.1.2. *Normality of the test statistic under H_1.* The true mean of the AUCs is unknown as there is no method to relate the true AUC to a given correlation under H_1. However, it is reasonable to assume that the true AUC is approximately equal to the $E(\widehat{\text{AUC}}) = \frac{\sum_{i=1}^{5,000} \widehat{\text{AUC}}_i}{5,000}$ for very large simulation such as 5,000. Using the above, the test statistic was recalculated and the percentage of standardized normal values falling to each quartile was checked. Table 1b illustrates the Chi-sqaure values obtained for the Normal GOF test under the alternative hypothesis. Table 1b clearly shows the complete violation of normality under the alternative hypothesis when HM standard errors are used. Interestingly, the normality is lost for classifiers with high predictability even under DeLong standard error. However, unlike HM method, the DeLong's method achieves the normality for classifiers

Table 1
Summarized results under one sample analysis

(a) Goodness of fit test results under H_0 for one sample analysis

Sample size	HM	DeLong
20	367.483*	137.393*
50	239.983*	65.9106*
75	250.604*	39.3410*
100	235.464*	33.1397*
250	171.023*	17.7067
500	168.004*	7.9976

(b) Goodness of fit test results under H_1 for one sample analysis

Sample	HM Correlations under H_1				DeLong Correlations under H_1			
size	0.2	0.5	0.2	0.5	0.2	0.5	0.2	0.5
20	287.9*	773.9*	1156.6*	3305.2*	315.7*	908.7*	805.6*	3027.7*
50	158.7*	230.8*	784.5*	1930.8*	70.3*	163.5*	809.1*	1141.7*
75	195.7*	189.2*	439.3*	1867.2*	68.1*	118.2*	629.2*	2194.4*
100	127.1*	177.5*	353.8*	1670.9*	49.2*	77.3*	330.0*	1797.8*
250	133.1*	125.8*	162.1*	436.5*	41.8*	31.7*	133.8*	398.8*
500	132.7*	115.0*	100.3*	165.0*	12.3	16.5	52.5*	223.0*

Note: Table value $= x^2_{(\alpha,k-c)} = 23.6848$ where $\alpha = 0.05$, $k = 14$ and $C = 0$ as no models were fitted. The asterisk (*) represents significant values.

(c) Significance level of HM/DeLong methods for one sample analysis

Method	Tail	Sample size 20	50	75	100	250	500
HM	Lower	0.0096*	0.0118*	0.0108*	0.0116*	0.0114*	0.0110*
	Upper	0.0072*	0.0098*	0.0080*	0.0104*	0.0122*	0.0114*
DeLong	Lower	0.032*	0.033*	0.024	0.028	0.024	0.025
	Upper	0.027	0.029	0.024	0.023	0.025	0.025

(d) Power of the tests under one sample analysis

Sample	HM Correlations under H_1				DeLong Correlations under H_1			
	0.2	0.5	0.2	0.5	0.2	0.5	0.2	0.5
20	0.127	0.594	0.952	0.998	0.151	0.634	0.961	0.998
50	0.232	0.940	1.000	1.000	0.317	0.966	1.000	1.000
75	0.320	0.993	1.000	1.000	0.420	0.997	1.000	1.000
100	0.412	0.999	1.000	1.000	0.539	1.000	1.000	1.000
250	0.819	1.000	1.000	1.000	0.886	1.000	1.000	1.000
500	0.985	1.000	1.000	1.000	0.995	1.000	1.000	1.000

with low predictabilities given the sample sizes are large. Under both methods, the normality improves with increasing sample size.

4.1.3. *Type I error rates.* In order to determine whether the type I error rate was maintained by the test, it was checked whether the proportions fall within the 95% probability interval (0.0207, 0.0293) for $\alpha/2 = 0.025$. Here $\alpha/2$ is used since two-sided tests are considered. According to Table 1c, the type I error rate is not maintained by the HM method for all sample sizes. The type I error rate is not maintained for smaller sample sizes such as 20 and 50 by DeLong's method while achieved for larger sample sizes.

4.1.4. *Power of the tests.* Table 1d illustrates the power of the test for varying sample sizes and correlations between the observed and the predicted outcomes for both HM and DeLong methods for one sample analysis. It is clear that the power of the tests increases with respect to both increasing sample size and predictability.

Comparing both methods, it is evident that DeLong's method outperforms Hanley and McNeil's method when the normality, type I error rates, and power of the tests are considered for one sample case.

4.2. *Two Independent Samples Case*

Table 2 gives the results for two independent samples' analysis.

4.2.1. *Normality of the test statistic under H_0.* Table 2a illustrates the Normal GOF test results under the simulation of the null hypothesis of two independent samples. According to Table 2a, the normality is lost for all combinations under HM's method. Even though normality is not achieved for all combinations of sample size and correlations, the normality improves when the predictability of the two classifiers improves for a given sample size as the Chi-square value decreases. Also, there seems to be no improvement in the normality with respect to increasing sample size. In contrast to HM's method, the normality holds for large samples such as size 50 and above under DeLong's method. The normality is achieved with respect to both increasing predictability and sample size.

4.2.2. *Normality of the test statistic under H_1.* Table 2b illustrates the Normal GOF test results obtained under the simulation of the alternative hypothesis under HM and DeLong's methods for two independent samples. Similar to null hypothesis, the normality does not hold under HM method while DeLong's method outperforms HM method as the normality is conserved for sample with size above 50.

4.2.3. *Type I error rates.* Table 2c presents type I error rates under two independent samples. The 95% probability interval for $\alpha/2 = 0.025$ is (0.0207, 0.0293). According to Table 2c, it is evident that type I error rates are not maintained by the HM method for all combinations of parameters under two independent samples. However, type I error is maintained by the DeLong's method on average for large samples such as size 250 and 500 when the predictability of both classifiers increases.

4.2.4. *Power of the tests.* Table 2d illustrates the power of the tests under two independent samples. Analyzing Table 2d, it is clear that the power of the test increases with respect to both increasing sample size and difference between the predictabilities of the classifiers

Table 2
Summarized results under two independent samples analysis

(a) Goodness of fit test results under H_0 for two independent samples

Sample size	HM Correlations under H_0 0.0/0.0	0.3/0.3	0.6/0.6	0.7/0.7	DeLong Correlations under H_0 0.0/0.0	0.3/0.3	0.6/0.6	0.7/0.7
20	133.50*	117.18*	69.25*	48.66*	39.09*	25.37*	61.30*	58.20*
50	160.80*	120.56*	60.62*	65.04*	15.04	8.98	22.71	10.08
75	145.07*	134.43*	62.98*	61.36*	11.41	9.68	22.24	11.92
100	150.76*	170.71*	81.26*	66.43*	18.96	10.52	18.65	18.64
250	161.47*	142.04*	76.42*	74.53*	15.14	15.86	9.50	10.30
500	188.27*	146.76*	79.41*	62.45*	4.50	9.62	18.38	4.64

(b) Goodness of fit test results under H_1 for two independent samples

Sample size	HM Correlations under H_1 0.6/0.5	0.6/0.3	0.7/0.3	0.8/0.2	DeLong Correlations under H_1 0.6/0.5	0.6/0.3	0.7/0.3	0.8/0.2
20	52.58*	61.50*	78.17*	91.19*	35.73*	28.30*	31.52*	31.30*
50	102.02*	88.00*	118.30*	135.37*	14.81	9.74	13.80	15.24
75	90.55*	134.61*	139.65*	119.97*	19.92	11.59	10.72	20.67
100	98.42*	126.75*	149.86*	123.58*	23.73	9.34	19.86	23.10
250	101.17*	115.53*	91.98*	155.15*	14.64	11.34	16.91	19.68
500	103.09*	146.77*	100.41*	106.46*	14.75	7.92	22.54	8.57

Note: Table values of 2.a and 2. b are, $x^2_{(\alpha,k-c)} = 23.6848$ where $\alpha = 0.05$, $k = 14$ and $C = 0$ as no models were fitted. The asterisk (*) represents significant values.

(c) Significance level of HM and DeLong methods for two independent samples

Method	Correlation	Tail	Sample size 20	50	75	100	250	500
HM	0.0/0.0	Lower	0.0162*	0.0136*	0.0138*	0.0130*	0.0138*	0.0110*
		Upper	0.0190*	0.0128*	0.0126*	0.0126*	0.0100*	0.0102*
	0.3/0.3	Lower	0.0162*	0.0104*	0.0142*	0.0142*	0.0108*	0.0114*
		Upper	0.0172*	0.0156*	0.0134*	0.0110*	0.0118*	0.0124*
	0.6/0.6	Lower	0.0196*	0.0150*	0.0154*	0.0128*	0.0170*	0.0180*
		Upper	0.0178*	0.0182*	0.0200*	0.0190*	0.0196*	0.0160*
	0.7/0.7	Lower	0.0198*	0.0170*	0.0156*	0.0168*	0.0132*	0.0170*
		Upper	0.0158*	0.0180*	0.0212	0.0214	0.0172*	0.0194*
DeLong	0.0/0.0	Lower	0.0304*	0.0266	0.0270	0.0282	0.0304*	0.0258
		Upper	0.0304*	0.0260	0.0242	0.0256	0.0246	0.0224
	0.3/0.3	Lower	0.0268	0.0242	0.0252	0.0234	0.0212	0.0224
		Upper	0.0338*	0.0280	0.0272	0.0256	0.0264	0.0262
	0.6/0.6	Lower	0.0316*	0.0294*	0.0274	0.0242	0.0264	0.0264
		Upper	0.0310*	0.0306*	0.0330*	0.0312*	0.0274	0.0258
	0.7/0.7	Lower	0.0338*	0.0258	0.0250	0.0258	0.0222	0.0240
		Upper	0.0304*	0.0270	0.0300*	0.0294*	0.0250	0.0242

(d) Power of the tests under two independent samples

Sample size	HM Correlations under H_1 0.6/0.5	0.6/0.3	0.7/0.3	0.8/0.2	DeLong Correlations under H_1 0.6/0.5	0.6/0.3	0.7/0.3	0.8/0.2
20	0.043	0.146	0.244	0.536	0.205	0.069	0.321	0.625
50	0.059	0.306	0.563	0.926	0.390	0.090	0.661	0.954
75	0.079	0.432	0.760	0.987	0.530	0.114	0.827	0.994
100	0.092	0.561	0.872	0.998	0.657	0.136	0.914	0.999
250	0.205	0.945	1.000	1.000	0.966	0.265	1.000	1.000
500	0.389	0.999	1.000	1.000	0.999	0.474	1.000	1.000

under both methods. However, the powers of tests based on DeLong's standard error are higher than the respective HM tests.

4.3. *Two Correlated Samples Case*

Table 3 gives the results for two correlated samples analysis.

4.3.1. *Normality of the test statistic under H_0.* Table 3a presents the Normal GOF test results under two correlated samples when HT (Jackknife method) and DeLong methods are used. Unlike in the two independent sample analysis, the Chi-square values under both methods increase as the predictability of the two classifiers increases under a given sample size. This results in the violation of normality for classifiers with high predictabilities when tested on same set of data. However, for a given classifier, the normality improves when the sample size is gradually increased. Interestingly, the HT and DeLong methods seem to behave similarly.

4.3.2. *Normality of the test statistic under H_1.* Table 3b depicts the Normal GOF test values under the alternative hypothesis of two correlated samples. According to Table 3b, the normality seems to hold for combinations of classifiers with different predictabilities when tested on paired samples with size above 100 on average. The similarity between HT and DeLong methods could be further observed in this. However, the normality holds true much better under the scenarios of alternative than the null.

4.3.3. *Type I error rates.* In order to determine whether the type I error rate was maintained by the test, it was checked whether the proportions fall within the 95% probability interval (0.0207, 0.0293) for $\alpha/2 = 0.025$. The parameter combinations given in bold lettering in Table 3c resulted in the test statistic being infinity due to perfect classification. Thus, Table 3e gives the corresponding intervals for those parameter combinations. Through Table 3c, it is clear that for both methods, type I error rates are not maintained for tests with smaller sample sizes.

4.3.4. *Power of the tests.* Table 3d illustrates the power of the test under two correlated samples. Analyzing Table 3d, it is clear that the powers of the HT and DeLong tests increase with respect to both increasing sample size and difference between the predictabilities of the classifiers. Interesting point to note is that the powers of the tests associated with both HT and DeLong's methods are approximately alike. Also, under both methods, the increase in power is less for classifiers with small and slightly different predictabilities. However, an increase in the power is seen as the difference in classification increases.

4.4. *Comparison with Respect to Empirical Variance*

In order to compare the standard errors under HM, HT, and DeLong methods, each standard error technique was compared with the empirical standard error generated by the estimated AUCs. The ratio between empirical to HM or DeLong standard deviations were found for comparison.

$$\text{Ratio} = \frac{\text{Emperical Standard Deviation}}{\text{HM or DeLong Standard Deviation}}. \tag{7}$$

Table 3
Summarized results under the analysis of two correlated samples

(a) Goodness of fit test results under H_0 for two correlated samples

Sample size	HT Correlations under H_0 0.0/0.0	0.3/0.3	0.6/0.6	0.7/0.7	DeLong Correlations under H_0 0.0/0.0	0.3/0.3	0.6/0.6	0.7/0.7
20	133.81*	178.14*	4983.62*	5472.97*	135.25*	177.71*	5057.06*	5526.48*
50	35.45*	27.29*	114.41*	247.85*	33.69*	26.45*	116.08*	254.18*
75	16.43	29.36*	29.63*	156.41*	14.14	22.96	22.00	144.90*
100	20.45	16.18	51.01*	63.97*	20.36	14.46	43.23*	55.45*
250	22.09	16.63	37.65*	44.45*	15.31	13.34	13.63	15.80
500	10.29	31.11*	14.80	46.58*	4.07	25.91*	5.83	23.11

(b) Goodness of fit test results under H_1 for two correlated samples

Sample size	HT Correlations under H_1 0.4/0.3	0.5/0.3	0.5/0.2	0.6/0.2	DeLong Correlations under H_1 0.4/0.3	0.5/0.3	0.5/0.2	0.6/0.2
20	55.03*	80.74*	81.60*	97.98*	61.24*	81.44*	84.36*	93.84*
50	41.01*	20.58	13.16	18.42	43.84*	24.04*	14.60	16.73
75	10.66	19.63	34.08*	19.46	9.88	23.91*	34.12*	19.03
100	14.24	12.19	18.26	19.63	16.17	12.92	21.86	18.98
250	7.85	15.98	5.64	11.56	6.18	13.49	5.70	10.96
500	6.72	11.29	26.75*	19.22	7.56	10.60	16.75	12.73

Note: Table values of 2.a and 2. b are, $x^2_{(\alpha,k-c)} = 23.6848$ where $\alpha = 0.05$, $k = 14$ and $C = 0$ as no models were fitted. The asterisk (*) represents significant values.

(c) Significance level of HT and DeLong methods for two correlated samples

Method	Correlation	Tail	Sample size 20	50	75	100	250	500
HT	0.0/0.0	Lower	0.0326*	0.0286	0.0276	0.0254	0.0212	0.0270
		Upper	0.0370*	0.0276	0.0260	0.0276	0.0244	0.0252
	0.3/0.3	Lower	0.0322*	0.0288	0.0280	0.0250	0.0268	0.0264
		Upper	0.0336*	0.0286	0.0252	0.0268	0.0248	0.0248
	0.6/0.6	Lower	0.0287	0.0304*	0.0280	0.0270	0.0262	0.0226
		Upper	0.0346*	0.0272	0.0274	0.0258	0.0250	0.0246
	0.7/0.7	Lower	0.0209	0.0266	0.0282	0.0260	0.0260	0.0238
		Upper	0.0265	0.0318*	0.0256	0.0248	0.0232	0.0262
DeLong	0.0/0.0	Lower	0.0366*	0.0276	0.0256	0.0282	0.0250	0.0252
		Upper	0.0326*	0.0286	0.0276	0.0252	0.0210	0.0268
	0.3/0.3	Lower	0.0334*	0.0288	0.0244	0.0264	0.0248	0.0232
		Upper	0.0322*	0.0288	0.0280	0.0246	0.0270	0.0258
	0.6/0.6	Lower	0.0346*	0.0256	0.0280	0.0266	0.0234	0.0254
		Upper	0.0287	0.0302*	0.0286	0.0278	0.0260	0.0238
	0.7/0.7	Lower	0.0265	0.0314*	0.0260	0.0250	0.0234	0.0262
		Upper	0.0209	0.0266	0.0276	0.0256	0.0262	0.0234

(d) Power of the tests under two correlated samples

Sample size	HT Correlations under H_1 0.4/0.3	0.5/0.3	0.5/0.2	0.6/0.2	DeLong Correlations under H_1 0.4/0.3	0.5/0.3	0.5/0.2	0.6/0.2
20	0.060	0.112	0.180	0.273	0.060	0.111	0.180	0.272
50	0.077	0.177	0.327	0.536	0.078	0.177	0.328	0.537
75	0.091	0.238	0.454	0.709	0.090	0.239	0.454	0.709
100	0.114	0.296	0.573	0.825	0.115	0.296	0.572	0.826
250	0.194	0.619	0.929	0.995	0.196	0.621	0.930	0.995
500	0.349	0.887	0.999	1.000	0.349	0.886	0.999	1.000

(e) Missing value summary under H_0 of two correlated sample analysis

Parameter combination	No. of missing values	Lower tail	Upper tail
$N = 20$, pho = 0.6/0.6	59	0.02065	0.02935
$N = 20$, pho = 0.7/0.7	206	0.02058	0.02942
$N = 50$, pho = 0.7/0.7	2	0.02067	0.02933

Thus, a ratio greater than unity indicates an underestimation of the empirical value, while a lesser value to unity indicates an overestimation. Comparisons were made by plotting the ratio against varying sample sizes (plots are not presented here due to space limitations). The empirical standard error remains as the reference line for comparison. The scale of the following plots was decided based on the Cleves (2002) paper. Summarizing the results observed under one sample analysis, the HM standard error overestimated while the DeLong standard error was close to the empirical standard error for all sample sizes. It was further observed that the distinction between HM and empirical standard errors were reduced when the predictabilities were improved. Similar to one sample analysis, the HM standard error overestimated the true empirical value while DeLong maintained it for varying sample sizes and correlations of null and alternative hypothesis of the two independent samples. Unlike one sample and two independent samples, the Jackknife method of HT and DeLong standard errors preformed equally well and was approximately close to the empirical value for all sample sizes and correlations of both null and alternative hypotheses of the two correlated samples.

4.5. *Analytical Explanation of HM and HT Results*

The HM standard error was derived in order to comprehend the behavior of HM standard error with the results obtained (see Annex). There are two problems in the standard error calculation.

(a) $\hat{\theta}$ is known to underestimate θ for Wilcoxon's method (Hanley and McNeil, 1982). Therefore, $\widehat{\mathrm{var}(\hat{\theta})}$ will be an overestimate ($\hat{\theta}$ squared terms mostly affect the variance negatively).

(b) When the correlation (ρ) increases between the observed and the predicted outcomes, the ties between them also increase. That is, the effect from $P(x_A = x_N)$ also increases. Hanley and McNeil (1982) have derived the above equation assuming the outcome of interest to be continuous. However, the simulation is conducted for a binary case. Since ties correspond to a positive component in the estimated variance of $\hat{\theta}$, not accounting for ties deflate the $\widehat{\mathrm{var}(\hat{\theta})}$. Therefore, in the presence of increasing correlation (i.e., increasing ties), the underestimation is more.

Thus, the above mentioned reasons (a) and (b) are two forces acting against each other as (a) gives an overestimation while (b) gives an underestimation.

4.5.1. *Explanation of the results in the light of the findings*. This section provides a brief explanation to the results observed with one sample, two independent and correlated samples, both under null and alternative hypotheses.

(a) *One sample:* The test statistic for one sample is given by Eq. (3). The top part of the test statistic is always an underestimation. Bradley's (1996) view point is that this underestimation decreases with respect to increasing sample size. There is no overestimation in variance due to underestimation of AUC as we take the AUC to be 0.5 under H_0, in the variance calculation. However, the binary classifier will have ties, even though the correlation is zero (i.e., $\mu_{AUC} = 0.5$) as there are only two values 0 and 1. This will result in an underestimation of the variance resulting inflated values of the test statistic (Z_0), which will violate both the normality assumption and the significance level.

(b) *Two independent samples:* The test statistic under null hypothesis is given by, Eq. (4). The underestimation of the AUCs in the top part of the equation is cancelled out as it is the difference between estimated AUCs. The variance is overestimated due to the estimated AUCs while underestimated due to the ties generated by the correlations. Slight underestimation is seen when the correlations are low such as 0.0 or 0.3 as the effect from ties is small. When the correlations increase up to 0.6 or 0.7, there are both overestimation due to estimated AUC and underestimation due to increasing ties. This results in the two forces cancelling out while giving an improved result with the increasing correlation (ρ).

(c) *Two correlated samples:* The test statistic under null is given by Eq. (6). Since the pseudo values are generated by a function of AUCs (which is a difference in AUCs), the underestimation of the true AUC by the Wilcoxon statistic is cancelled out. Hence, there is no underestimation from the top part of the test statistic and in the individual variance terms. When the correlation between the observed and the predicted is low such as 0.0 or 0.3, the correlation (ρ) between AUC_1 and AUC_2 also becomes negligible. Thus, the effect from the term S in Eq. (6) is zero. In contrast, when the correlation is increased up to 0.6 or 0.7, the correlation (ρ) between AUC_1 and AUC_2 also becomes significant and the effect from the term S is significant. Therefore, the bottom part of the test statistic keeps deflating as the correlation increases. This results in the test statistics inflation with respect to increasing correlation, which results in the rejection of the null hypothesis.

4.6. *Analytical Explanation of DeLong Results*

It is important to note that, the only assumption DeLong et al. (1988) have made is the large sample approximation (refer Annex). The problems associated with DeLong standard error are the underestimation of θ due to Wilcoxon's method and the large sample approximation.

4.6.1. *Explanation of the results in the light of the findings.* This section provides a brief explanation to the results observed with one sample, two independent and two correlated samples, with respect to DeLong standard error.

(a) *One sample:* Unlike in Hanley and McNeil's method, there is no effect from the ties. The variance formula should work well with large samples. Interestingly, this was clearly seen with large samples. The normality was held true for samples with size above 250. Thus, there is no effect from the standard error to the test statistic Z_0. The only effect is the underestimation due to Wilcoxon's method.

(b) *Two independent samples:* The underestimations by the estimated AUCs are now cancelled out since a difference in means is considered. Since the DeLong standard error is a large sample approximation, the results obtained from the simulation process showed a violation in the normality for small samples such as size 20. Thus, the results obtained through two independent samples are explainable.

(c) *Two correlated samples:* Similar to HT method, the effect from the covariance term is significant when the correlation is increased from 0 to 0.7. Therefore, the bottom part of the test statistic (5) decreases and results in the overall test statistic to be inflated. Thus, for increasing correlations the normality is violated with great degree. When the correlation is small, the only negative effect is from the large sample approximation.

5. An Example Based on Real Data

The methodology was illustrated using data from the Cardiology Unit of the Sri Jayawardenapura Hospital in Sri Lanka. The "Gold Standard" test for detecting coronary artery disease in patients is the angiogram. However, due to its high expense a substitute test in the form of cardiac stress test (CST) is first carried out to determine the necessity for doing an angiogram. The angiogram results have two levels failed or passed, while the CST results have four levels, namely,

1. Stage 1 difficulty
2. Stage 2 difficulty
3. Stage 3 or higher difficulty or minor difficulties
4. Completed the CST or patient was diagnosed as adequately stressed.

In order to use the CST results as a substitute for the angiogram, the CST is dichotomized using two cut-points in order to determine which cut-point results in better classification of the angiogram result. In the first categorization levels 1, 2 or 3 are considered as failure while stage 4 is considered as a success and this is considered as classifier one (r_1). In the second categorization levels 1 or 2 is considered a failure while stages 3 or 4 is considered a success and this is considered as classifier two (r_2). As the tests

Table 4
Summarized conclusions under all methods

Summarized properties of the tests under three sample designs			
	One sample	Two independent samples	Two correlated samples
Normality of the estimated AUCs	– Normality under DeLong's method is assured for samples with size above 20. – Normality is violated for all sample sizes under HM method.	– Normality under DeLong's method holds for samples with size 50 and above. – Normality is violated for all sample sizes under HM method.	– Both HT and DeLong methods require large samples (above 500) on average to achieve normality.
Type I error	– Type I error is maintained for large samples (above 50) by DeLong's method. – Type I error is violated under HM	– Type I error under DeLong's method is maintained for large samples (above 250). – Type I error is lost under HM.	– Type I error is maintained by both HT and DeLong's methods for samples with size above 75.
Power	– DeLong's method outperforms HM	– DeLong's method outperforms HM	– Powers of HT and DeLong methods are approximately equal

Q11

4C/Art

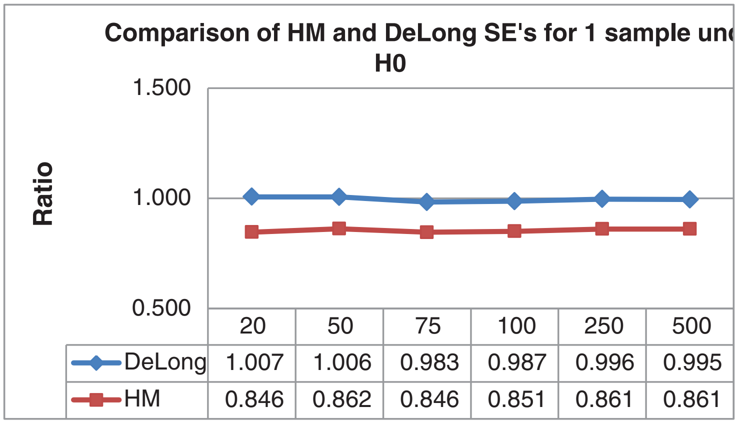


Figure 1. Ratios of empirical standard error to HM/DeLong standard errors under null hypothesis of the one sample. (color figure available online)

4C/Art

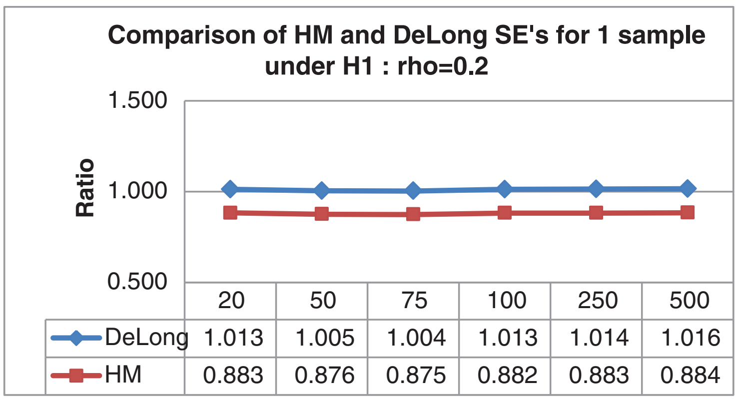


Figure 2. Ratios of empirical standard error to HM/DeLong standard errors when the correlation between observed and the predicted outcomes is set to 0.2 under alternative hypothesis. (color figure available online)

4C/Art

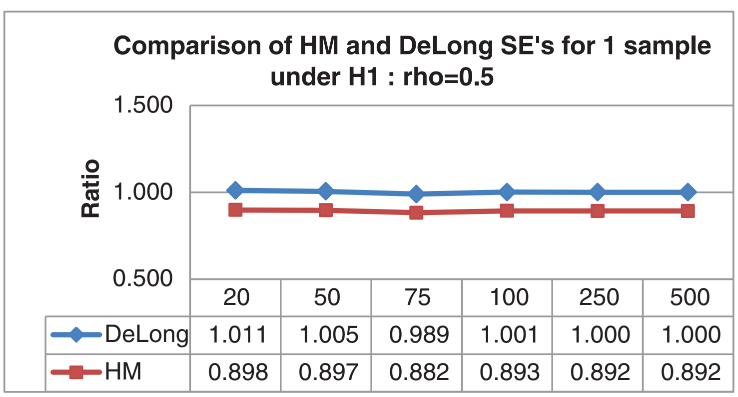


Figure 3. Ratios of empirical standard error to HM/DeLong standard errors when the correlation between observed and the predicted outcomes is set to 0.5 under alternative hypothesis. (color figure available online)

4C/Art

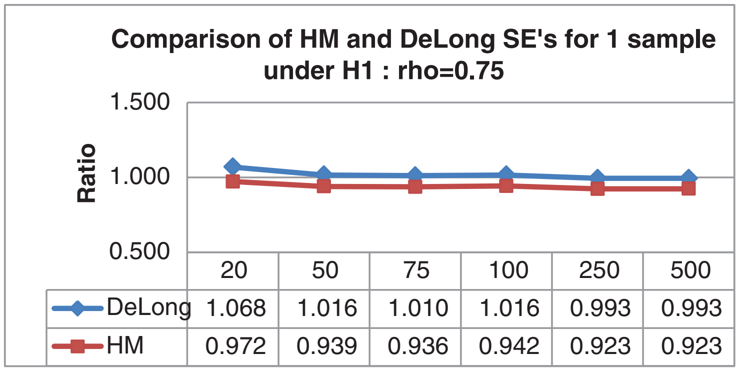


Figure 4. Ratios of empirical standard error to HM/DeLong standard errors when the correlation between observed and the predicted outcomes is set to 0.75 under alternative hypothesis. (color figure available online)

were carried out on the same patients, this was an example of the two samples correlated case. The null hypothesis was that AUC_1 corresponding to classifier one as a predictor of angiogram results was the same as AUC_2 that corresponds to classifier two as a predictor of angiogram results. The alternative hypothesis was that AUC_1 and AUC_2 were different in terms of predictive ability of angiogram results.

In order to test the null hypothesis both Hanley-Tilaki (HT) and DeLong methods for two correlated samples were used. The values of the estimated AUC_1, AUC_2 their standard errors and their covariance for both HT and DeLong's methods were: 0.345, 0.33, 0.031, 0.031, and 0.000653, respectively, giving a Z value of 0.604 and a p-value of 0.5456. The data corresponded to a correlation of 0.68 between AUC_1 and AUC_2. As the p-value is greater than 0.05 the null hypothesis is not rejected at the 5% level and it is concluded that there is no significant difference between AUC_1 and AUC_2 and thus, there is no significant difference between the two classifiers r_1 and r_2 in their classifying ability of the angiogram results.

4C/Art

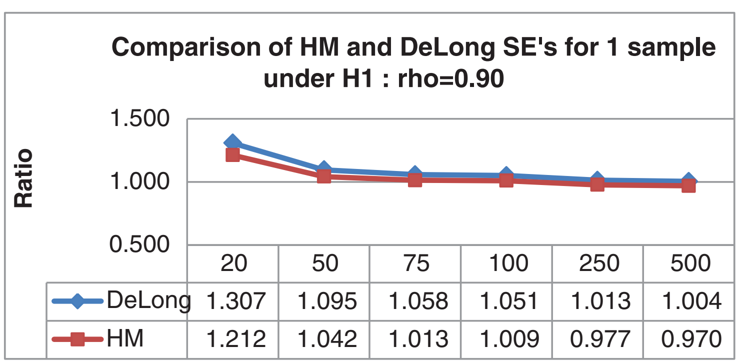


Figure 5. Ratios of empirical standard error to HM/DeLong standard errors when the correlation between observed and the predicted outcomes is set to 0.90 under alternative hypothesis. (color figure available online)

4C/Art

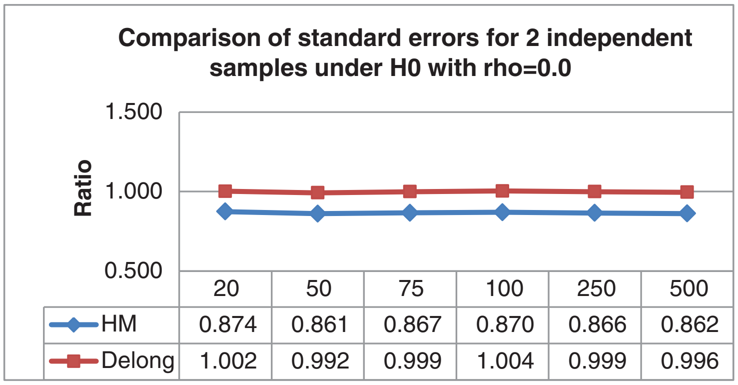


Figure 6. Ratios of empirical standard error to HM/DeLong standard errors when the correlation between observed and the predicted outcomes of both independent samples is set to 0.0 under null hypothesis. (color figure available online)

4C/Art

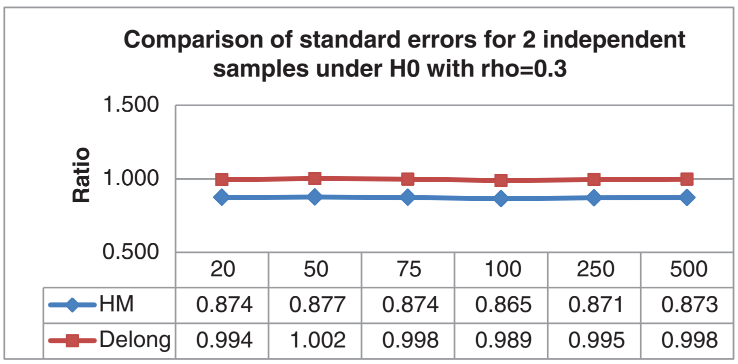


Figure 7. Ratios of empirical standard error to HM/DeLong standard errors when the correlation between observed and the predicted outcomes of both independent samples is set to 0.3 under null hypothesis. (color figure available online)

4C/Art

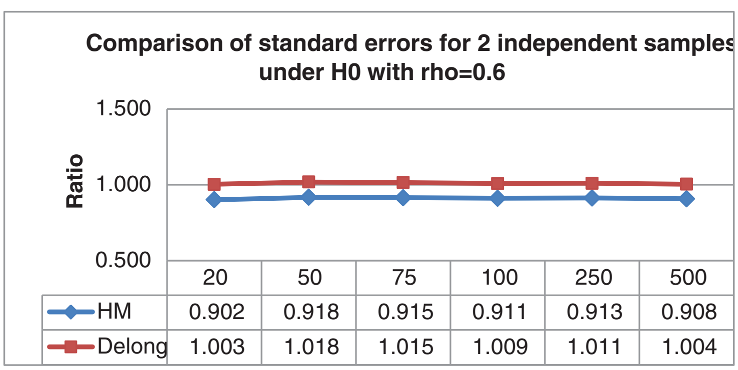


Figure 8. Ratios of empirical standard error to HM/DeLong standard errors when the correlation between observed and the predicted outcomes of both independent samples is set to 0.6 under null hypothesis. (color figure available online)

4C/Art

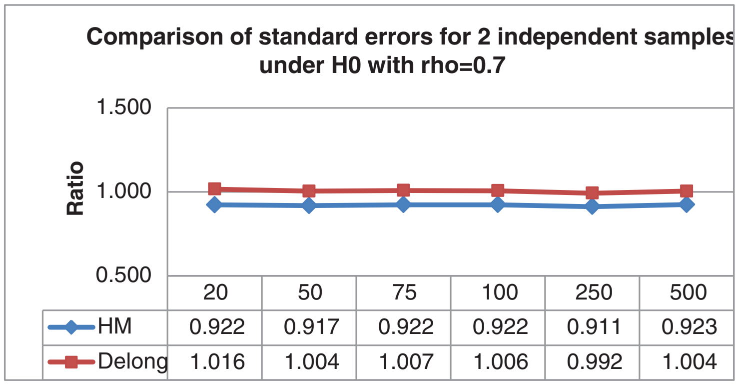


Figure 9. Ratios of empirical standard error to HM/DeLong standard errors when the correlation between observed and the predicted outcomes of both independent samples is set to 0.7 under null hypothesis. (color figure available online)

4C/Art

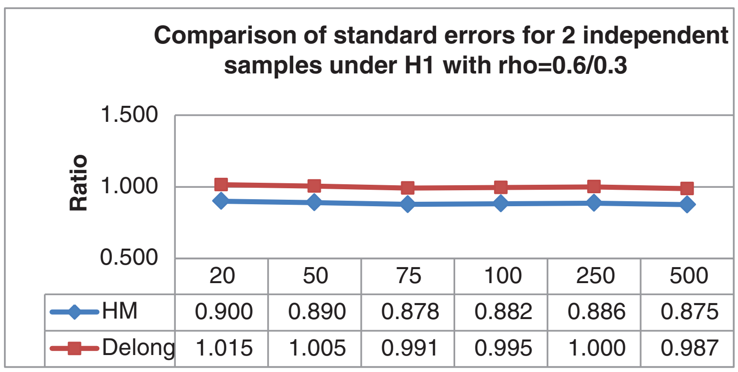


Figure 10. Ratios of empirical standard error to HM/DeLong standard errors when the correlation between observed and the predicted outcomes of the two independent samples is set to 0.6 and 0.3 under alternative hypothesis. (color figure available online)

4C/Art

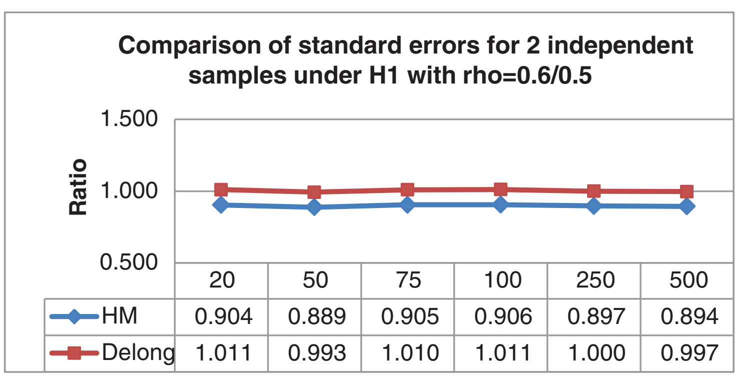
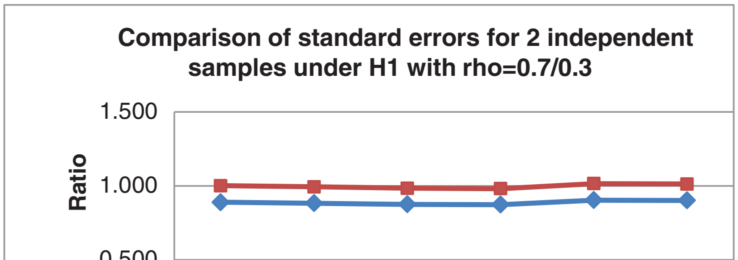
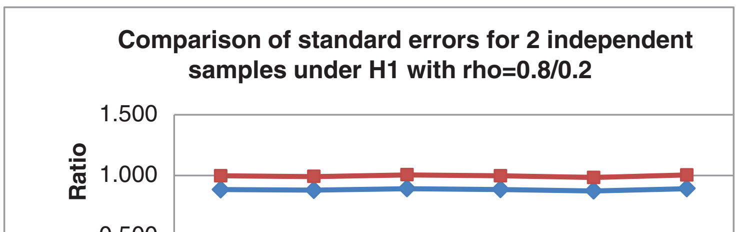


Figure 11. Ratios of empirical standard error to HM/DeLong standard errors when the correlation between observed and the predicted outcomes of the two independent samples is set to 0.6 and 0.5 under alternative hypothesis. (color figure available online)



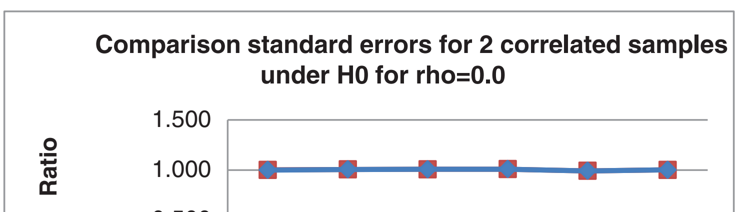
	20	50	75	100	250	500
HM	0.889	0.882	0.875	0.873	0.903	0.901
Delong	1.001	0.993	0.984	0.981	1.015	1.012

Figure 12. Ratios of empirical standard error to HM/DeLong standard errors when the correlation between observed and the predicted outcomes of the two independent samples is set to 0.7 and 0.3 under alternative hypothesis. (color figure available online)



	20	50	75	100	250	500
HM	0.882	0.878	0.890	0.884	0.872	0.890
Delong	0.996	0.990	1.004	0.996	0.982	1.003

Figure 13. Ratios of empirical standard error to HM/DeLong standard errors when the correlation between observed and the predicted outcomes of the two independent samples is set to 0.8 and 0.2 under alternative hypothesis. (color figure available online)



	20	50	75	100	250	500
HHT ratio	1.001	1.005	1.007	1.008	0.991	1.001
Delong ratio	1.000	1.005	1.007	1.008	0.991	1.001

Figure 14. Ratios of empirical standard error to HT/DeLong standard errors when the correlation between observed and the predicted outcomes of both correlated samples is set to 0.0 under null hypothesis. (color figure available online)

4C/Art

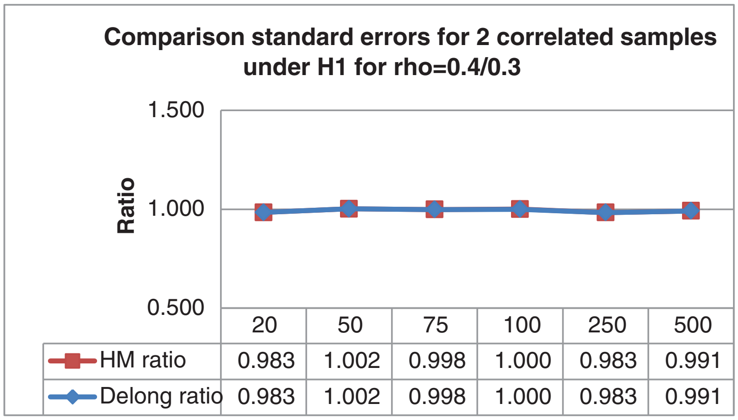


Figure 15. Ratios of empirical standard error to HT/DeLong standard errors when the correlation between observed and the predicted outcomes of the two correlated samples is set to 0.4 and 0.3 under alternative hypothesis. (color figure available online)

The low values of AUC_1 and AUC_2 indicate that for both classifiers, r_1 and r_2 used, the predictive ability of CST as a predictor for angiogram results is poor. Therefore, determining the predictive ability of r_1 and r_2 after adjusting for other prognostic factors such as gender, age, smoking, alcohol, family history, hypertension, diabetes, etc., using logistic models is recommended.

As both the Hanley and Tilaki method and the DeLong method give identical results, this illustrates their twin-like nature observed in the simulations.

6. Overall Discussion

Summarizing the results for one sample and two independent samples, it is advisable to use DeLong et al. (1988) algorithm if one is interested in proceeding with nonparametric techniques as it is very consistent and robust for samples with size above 250, even though the calculations are more difficult than compared to Hanley and McNeil (1982, 1983) methods. Both, DeLong et al. (1988), and Hanley and Tilaki (1997) methods perform similarly well for two correlated samples. The normality was achieved for various parameter combinations as explained in the previous sections. Thus, the assumption of normality of the area under curve often made in the literature of many previous researches could now be validated under certain conditions. Furthermore, the type I error rate was also controlled asymptotically for DeLong et al. (1988) and Hanley and Tilaki (1997) while lost for Hanley and McNeil (1982, 1983) on average. It was found that the Hanley and McNeil (1982, 1983) variance overestimated the true empirical variance, while the DeLong et al. (1988) variance maintained close to the empirical variance for all sample sizes and correlations under one sample and two independent samples. Interestingly, the "twin-like" nature between DeLong's and the Jackknife method as described by Hanley and Tilaki (1997) was clearly seen as both variances maintained close to the empirical variance for all combinations of sample sizes and correlations under two correlated samples. This was further highlighted in the example on real data.

7. Conclusions

The findings of this study are summarized in Table 4:

DeLong et al.'s (1988) method was consistent under all sample designs while Hanley and McNeil's (1982, 1983) method was not. Q12

Annex

Reference: Hanley and McNeil (1982) *and Appendix to Hanley and McNeil Radiology paper "A Method of Comparing the Areas under ROC curves derived from same cases."* Q13

Assume, without loss of generality, that higher values of a diagnostic test are associated with "abnormal" subjects, while lower values are associated with "normal" subjects. Further, assume that the diagnostic test is applied to n_N normal and n_A abnormal subjects. Let x_A $i = 1, 2, \ldots, n_A$ and $x_N, j = 1, 2, \ldots, n_N$ be the observed outcomes of the diagnostic test for the abnormal and normal subjects, respectively. Let the true area under curve be denoted as θ. The Wilcoxon statistic $(\hat{\theta})$ is given by,

$$\hat{\theta} = \frac{1}{n_A n_N} \sum_{i}^{n_A} \sum_{j}^{n_N} S(x_A, x_N) \quad \text{where} \quad S(x_A, x_N) \begin{cases} 1 & \text{if } x_A > x_N \\ 0.5 & \text{if } x_A = x_N \\ 0 & \text{if } x_A < x_N \end{cases} \quad \text{(Discrete only)}$$

The variance of the estimator is given as follows.

$$\operatorname{var}(\hat{\theta}) = \frac{1}{n_A^2 n_N^2} \left\{ \sum_{i}^{n_A} \sum_{i}^{n_N} \operatorname{var}[S(x_A, x_N)] + \sum_{ii \neq j}^{n_A} \sum_{j}^{n_N} \operatorname{cov}[S(x_A, x_N), S(x'_A, x'_N)] \right\} \quad \text{(A)}$$

Now consider $\operatorname{var}[S(x_A, x_N)]$, assuming $\alpha = S(x_A, x_N)$. It can be found as,

$$\operatorname{var}[\alpha] = E[\alpha^2] - (E[\alpha])^2$$

Hanley and McNeil (1982) assumes the data to be on a continuous scale. Thus, $P(x_A = x_N) = 0$ and $\theta = P(x_A > x_N)$. The expectations could be now written as follows, Q14

$$\begin{aligned} E[\alpha^2] &= 1^2 P(x_A > x_N) + 0.5^2 P(x_A = x_N) + 0^2 P(x_A < x_N) = \theta \\ E[\alpha] &= 1P(x_A > x_N) + 0.5P(x_A = x_N) + 0P(x_A < x_N) = \theta \\ \text{Thus,}\ & \operatorname{var}[S(x_A, x_N)] = \theta - \theta^2 = \theta(1 - \theta). \end{aligned} \quad \text{(B)}$$

Estimate it by $\hat{\theta}(1 - \hat{\theta})$. Now consider the covariance term in (A) and let $\beta = S(x'_A, x'_N)$. The covariance term could be written as, $\operatorname{cov}(\alpha, \beta) = E(\alpha\beta) - [E(\alpha)E(\beta)]$.

Most of the terms in the above equation are zero. As proved before $E(\alpha)E(\beta) = \theta^2$. The only non zero terms given by $E(\alpha\beta)$ is when, $x_A > x_N$ and $x'_A > x'_N$. This could be,

$$x_A > x_N x'_N \text{ or } x_A,\ x'_A > x_N.$$

Taking normal and abnormal cases separately,

<u>Abnormal</u> There are n_A such pairs

$$\sum_{ii \neq j}^{n_A} \sum_{j}^{n_N} \operatorname{cov}[S(x_A, x_N), S(x'_A, x'_N)] = n_A(n_A n_N - n_N)(Q_1 - \theta^2) \quad \text{(C)}$$

Normal There are n_N such pairs

$$\sum_{ii\neq j}^{n_A}\sum_{j}^{n_N}\mathrm{cov}[S(x_A,x_N),S(x'_A,x'_N)] = n_N(n_An_N-n_A)(Q_2-\theta^2) \tag{D}$$

where Q_1 is the probability that two randomly selected abnormal subjects will both have a higher score than a randomly selected normal subject, and Q_2 is the probability that one randomly selected abnormal subject will have a higher score than any two randomly selected normal subjects. Substituting B, C, and D to A,

$$\widehat{\mathrm{var}(\hat\theta)} = \frac{\{n_An_N\hat\theta(1-\hat\theta)+n_A(n_An_N-n_N)(Q_1-\hat\theta^2)+n_N(n_An_N-n_A)(Q_2-\hat\theta^2)\}}{n_A^2n_N^2}$$

$$\widehat{\mathrm{var}(\hat\theta)} = \frac{\hat\theta(1-\hat\theta)+(n_A-1)(Q_1-\hat\theta^2)+(n_N-1)(Q_2-\hat\theta^2)}{n_An_N}.$$

Reference: DeLong et al. (1988) *paper "Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach."*

Suppose a sample of N individuals undergo a test for predicting an event of interest or determining presence or absence of a medical condition. Adhere to the convention that higher values of the test variable are assumed to be associated with the event of interest, e.g., positive disease status. Let this group be denoted by C_1 and let the group of n $(= N - m)$ individuals who do not have the condition be denoted by C_2. Let $X_i, i = 1, 2, \ldots, m$ and $Y_j, j = 1, 2, \ldots, n$ be the values of the variable on which the diagnostic test is based for members of C_1 and C_2, respectively. Let the true area under curve be denoted as θ. The Wilcoxon statistic $(\hat\theta)$ is given by,

$$\hat\theta = \frac{1}{mn}\sum_{i}^{m}\sum_{j}^{n}S(X_i,Y_j) \text{ where } S(X_i,Y_j)\begin{cases}1 & \text{if } X_i > Y_j\\ 0.5 & \text{if } X_i = Y_j\\ 0 & \text{if } X_i < Y_j\end{cases} \quad \text{(discrete only).}$$

The variance of an estimated AUC is given by,

$$\widehat{\mathrm{var}(\hat\theta)} = \frac{\hat\varepsilon_{11}+(n-1)\hat\varepsilon_{10}+(m-1)\hat\varepsilon_{01}}{mn},$$

where $\varepsilon_{10} = E[S(X_i,Y_j)S(X_i,Y_k)]-\theta^2; j\neq k$, $\varepsilon_{01} = E[S(X_i,Y_j)S(X_k,Y_j)]-\theta^2; i\neq k$ and $\varepsilon_{11} = E[S(X_i,Y_j)S(X_i,Y_j)]-\theta^2$. It is important to note that the form of this equation is similar to the form of Hanley and McNeil's standard error formula. For large m, n, DeLong et al. (1988) made the following assumption: $\lim_{m,n\to\infty}\frac{\varepsilon_{11}}{mn} = 0$ (Reference: Hajian-Tilaki (1997)).

Assuming $\frac{\varepsilon_{11}}{mn} = 0$, the variance formula reduces to the following formula.

$$\widehat{\mathrm{var}(\hat\theta)} = \frac{(n-1)\hat\varepsilon_{10}+(m-1)\hat\varepsilon_{01}}{mn}. \tag{E}$$

Since $\varepsilon_{10} = E[S(X_i,Y_j)S(X_i,Y_k)]-\theta^2; j\neq k$ is a covariance term, it could be written as follows, $\varepsilon_{10} = E[S(X_i,Y_j)-E(S(X_i,Y_j))][S(X_i,Y_k)-E(S(X_i,Y_k))]$.

According to the defined values of $S(X_i,Y_j)$, $E(S(X_i,Y_j)) = \theta = E(\hat\theta)$.

Now $\hat{\varepsilon}_{10}$ is a sample covariance term. Thus,

$$\hat{\varepsilon}_{10} = \frac{\sum_{\substack{i=1 \\ i \neq k}}^{m} \sum_{j,k=1}^{n} [S(X_i, Y_j) - \hat{\theta}][S(X_i, Y_k) - \hat{\theta}]}{mn(n-1)}$$

$$= \frac{\sum_{i=1}^{m} \left[\sum_{j=1}^{n} (S(X_i, Y_j) - \hat{\theta}) \sum_{k=1}^{n} (S(X_i, Y_k) - \hat{\theta}) \right]}{mn(n-1)}$$

For large n, $\hat{\varepsilon}_{10} = \frac{\sum_{i=1}^{m} [nV_{10}(X_i) - n\hat{\theta}][(n-1)V_{10}(X_i) - (n-1)\hat{\theta}]}{mn(n-1)}$

$$\hat{\varepsilon}_{10} = \frac{\sum_{i=1}^{m} (V_{10}(X_i) - \hat{\theta})^2}{m} = \frac{(m-1)S_{10}}{m} \quad \text{where} \quad S_{10} = \frac{\sum_{i=1}^{m} (V_{10}(X_i) - \hat{\theta})^2}{m-1}.$$

Similarly, $\hat{\varepsilon}_{01} = \frac{\sum_{j=1}^{n} (V_{01}(Y_j) - \hat{\theta})^2}{n} = \frac{(n-1)S_{01}}{n}$ where $S_{01} = \frac{\sum_{j=1}^{n} (V_{01}(Y_j) - \hat{\theta})^2}{n-1}$.

The definitions of $V_{10}(X_i)$ and $V_{01}(Y_j)$ are given in the Methodology.

Substituting, to Eq. (E), $\widehat{\text{var}(\hat{\theta})} = \frac{(n-1)(m-1)S_{10}}{m^2 n} + \frac{(m-1)(n-1)S_{01}}{n^2 m}$.

For large m, n, $\lim_{m,n\to\infty} \frac{(n-1)(m-1)}{mn} = 1$.

Hence, for large sample sizes, $\widehat{\text{var}(\hat{\theta})} = \frac{S_{10}}{m} + \frac{S_{01}}{n}$.

This is the DeLong et al. (1988) standard error formula for estimating area under ROC curve for single sample. Thus, it is clear that it is a large sample formula.

Acknowledgments

The authors are grateful to Dr. N. L. Amarasena, Consultant Cardiologist and to the staff of the record room of the Sri Jayawardanapura General Hospital Sri Lanka, for providing the data for the example. The authors also thank Ms. Devini Senaratna for cleaning the data and putting it in to a presentable format.

References

Agresti, A. (2007). *An Introduction to Categorical Data Analysis*. Hoboken, NJ: Wiley Interscience.

Bradley, A. P. (1996). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* 30(7):1145–1159.

Cleves, M. A. (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.

DeLong, E. R., DeLong, D. M., Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* 44(3):837–845.

Green, D. M., Swets, J. M. (1966). *Signal Detection Theory and Psychophysics*. New York: Wiley.

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

Hanley, J. A., McNeil, B. J. (1982b). A method of comparing the areas under ROC curves derived from same cases. (An unpublished appendix.) Q15

Hanley, J. A., McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* 148:839–843.

Hanley, J. A., McNeil, B. J. (1984). Statistical approaches to the analysis of receiver operating characteristic (ROC) curves. *Medical Decision Making* 4(2):137–150.

Hanley, J. A., Tilaki, H. K. O. (1997). Sampling variability of nonparametric estimates of the areas under receiver operating characteristic curves: An update. *Statistics in Radiology* 4:49–58.

Hosmer, D. W., Lemeshow, S. (2000). *Applied Logistic Regression*. *Wiley Series in Probability and Statistics*. Hoboken, NJ: John Wiley & Sons, Inc.

Metz, C. E., Herman, B. A., Roe, C. A. (1998). Statistical comparison of two ROC-curve estimates obtained from partially-paired datasets. *Medical Decision Making* 18(1):110–121.

Nettleman, M. D. (1988). Receiver operator characteristic (ROC) curves. *Infection Control and Hospital Epidemiology* 9(8):374–377.

Park, C. G., Park, T., Shin, D. W. (1996). A simple method for generating correlated binary variates. *The American Statistician* 50(4):306–310.