DETECTING OUTLIERS AND INFLUENTIAL DATA POINTS IN RECEIVER OPERATING CHARACTERISTIC (ROC) ANALYSIS

by

Amy H. Klym

B.S., University of Pittsburgh, 1997

Submitted to the Graduate Faculty of

the Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2007

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This thesis was presented

by

Amy H. Klym

It was defended on

February 28, 2007

and approved by:

Andriy Bandos, PhD, Research Assistant Professor, Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

David Gur, ScD, Professor, Department of Radiology
School of Medicine
University of Pittsburgh

Thesis Advisor: Howard E. Rockette, PhD, Professor, Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

DETECTING OUTLIERS AND INFLUENTIAL DATA POINTS IN RECEIVER OPERATING CHARACTERISTIC (ROC) ANALYSIS

Amy H. Klym, M.S.

University of Pittsburgh, 2007

Receiver operating characteristic (ROC) studies and analyses are often used to evaluate medical tests and are very useful in the field of radiology to evaluate a single diagnostic imaging system, to compare the accuracy of two or more diagnostic imaging systems, or to assess observer performance. There have been many refinements in the development of different ROC type study designs and the corresponding statistical analysis. These methods have become increasingly important and ROC methods are the principal approach for evaluating imaging technologies and/or observer performances. The systems that are often evaluated using ROC methodology include digital and radiographic images of the chest and breast. An improved method of evaluating diagnostic imaging systems contributes to the development of better diagnostic methods; hence, improving imaging systems for diagnoses of breast and lung cancer would have major public health significance. In our work with observer performance studies, in which receiver operating characteristic (ROC) analysis is used, we have noted that some contributions of readers and cases can substantially alter the conclusions of the analysis. To the best of our knowledge, to date there is no statistical test cited in the statistical literature that addresses the detection and influence of outliers on the estimate of the area under the ROC curve. Evaluating outliers may be especially important for the ROC model since subtle (difficult) cases have the potential for being missed by a reader (e.g. a difficult positive case is rated as an unquestionably negative case), and can have a considerable influence on the estimated area under the ROC curve, especially if the study has a small set of cases. Therefore, we believe it is important to develop a method for detecting and measuring the influence of outliers for ROC models. The development of this method will involve deriving a test statistic for outliers based on the jackknife influence values and conducting a preliminary validation of the test.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENT

I would like to take this opportunity to thank those who helped make this thesis possible. First, I would like to extend my deepest gratitude to my committee members. In particular, I thank my thesis advisor Dr. Howard Rockette for his guidance, comments, and suggestions through out the preparation of this thesis. I would also like to thank Dr. Andriy Bandos for his time, advice, and help with SAS programming and Dr. David Gur for his continued encouragement and support.

I would also like to extend my appreciation to my fellow colleagues from Radiology Imaging Research, especially Glenn Maitz for his programming expertise and Jill King for listening and offering words of encouragement.

I would also like to offer special thanks to my husband, Michael, for his unending encouragement and support, and to my two children, Laurie and Mitchell, for their patience.

Lastly, this is work is supported in part by Grants EB001694, EB002106, and EB003503 (to the University of Pittsburgh) from the National Institute for Biomedical Imaging and Bioengineering (NIBIB), National Institute of Health.

1.0 INTRODUCTION

Outliers are those outermost, or apparently peculiar, data points that lie beyond the data range in either direction. In linear regression analysis outliers are defined as being those data points that have much larger residuals, in absolute value, than all other residuals in a set of data [1]. Usually data points with residuals lying three or more standard deviations away from the mean of the residuals is considered an outlier. An outlier may be the result of a data entry error, sampling error, an indication of some other problem, or a "true" outlier (i.e. not an error or any other problem with the data set). Whatever the cause, outliers can significantly affect the outcome of the analysis being performed, particularly if the sample size is small. Outliers that significantly affect the analysis are considered influential data points, and some may have more impact than others. It should be emphasized that not all outliers are influential data points; therefore, prior to final analysis, outliers need to be identified and closely inspected to determine the impact they may have on the outcome of any analyses of the data set in question. Depending on the analysis being performed, such as linear regression, logistic regression, or proportional hazards model, there are diagnostics that can be performed to detect these extreme data points. For example, in the case of linear regression analysis, jackknife residuals and leverages are used to detect outliers and Cook's distance is used to determine their influence. It is important to identify those observations that cause disproportionately large influence on model performance.

In our work with observer performance studies, in which receiver operating characteristic (ROC) analysis is used, we note that some contributions of readers and cases can substantially alter the conclusions of the analysis [2,3]. To the best of my knowledge, to date there is no method cited in the statistical literature that addresses the detection and influence of outliers on the estimate of the area under the ROC curve. Evaluating outliers may be especially important for the ROC model since subtle (difficult) cases have the potential for being miss-diagnosed by a reader (e.g. a difficult positive case is rated as an unquestionably negative case), and this can have a considerable influence on the estimated area under the ROC curve, especially if the study has a small set of cases. Therefore, we believe it is relevant to develop a method for detecting and measuring the influence of outliers for ROC models, and we propose to develop this method.

1.1 BACKGROUND

1.1.1 ROC Analysis

Receiver operating characteristic (ROC) studies and analyses are often used to evaluate medical tests and are very useful in the field of radiology to evaluate a single diagnostic imaging system, to compare the accuracy of two or more diagnostic imaging systems, or to assess observer performance [4-6]. ROC analysis was first introduced into the medical field to assess medical decision making by Lusted in the 1960s [7, 8]. Since then ROC analysis has undergone many refinements, including development of more flexible study designs and improved statistical techniques [9-21]. ROC studies have become increasingly important and are presently the principal methodology for evaluating imaging technologies and/or observer performances. This can be seen in the literature where there is a considerable amount of studies in which ROC type

methodology is used for this purpose [22-35]. Typically ROC studies require many experienced readers evaluating a large set of typical (easy) and subtle (difficult) cases where truth is known in order to ascertain statistically reliable results; however, many ROC studies do not include large sets of cases and readers because it is to costly and time consuming.

Sensitivity and specificity are measures of how well a medical test performs, and these two measurements are needed to perform ROC analysis [36]. Sensitivity is the fraction of patients who are diagnosed with disease who truly have the disease, and specificity is the fraction of patients who are not diagnosed with disease who truly do not have disease. However, sensitivity and specificity depend on an upper limit (threshold value or cut point) determined by the individual evaluating the patient (e.g. a radiologist) which classifies the patient as diseased (abnormal) or non-diseased (normal). As the upper limit changes sensitivity and specificity will change (e.g. if the upper limit is increased, sensitivity will decrease while specificity will increase).

The ROC curve is a plot of the trade off between sensitivity (the true positive fraction (TPF) of diseased cases) and 1-specificity (the false positive fraction (FPF) of diseased cases) [37]. Specifically, the empirical ROC curve plots the TPF (y axis) vs. the FPF (x axis) for all possible threshold values and line segments are used to connect these values. A smooth ROC curve can be obtained by fitting a statistical model which assumes an underlying binormal distribution (i.e., assuming a normal distribution for both populations of diseased and non-diseased patients). Figure 1.1.1 gives an example of a hypothetical empirical ROC curve and a smooth ROC curve.

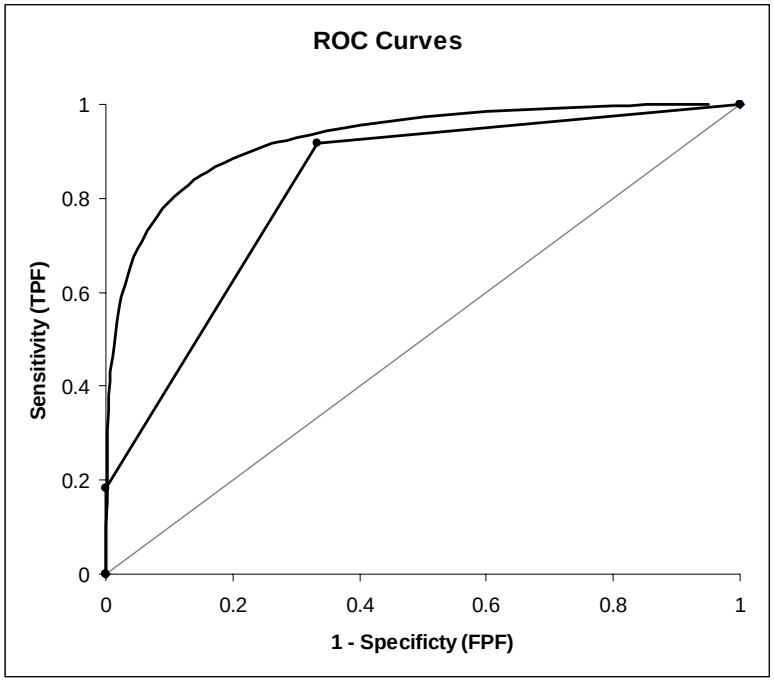

Figure 1.1.1: Example of a Smooth ROC curve and an Empirical ROC curve

A number of summary indices can be used to measure the diagnostic accuracy of a test [6, 38, and 39]. One of the most frequently used measures of accuracy is the area under the ROC curve (AUC). The AUC can be interpreted as the probability of making a correct decision, namely, correctly identifying a randomly chosen abnormal patient with greater suspicion for disease than that of a randomly chosen normal patient [6, 40]. For example, an AUC equal to 0.50 corresponds to the situation where both the randomly chosen abnormal and normal patients have the same suspicion for disease; essentially the decision could be made with a flip of coin. Therefore, a diagnostic or screening test yielding an AUC of 0.50 would not be useful or beneficial. An AUC greater than 0.50 would indicate the test has some discriminating ability (i.e., the test can differentiate between diseased and non-diseased patients). It is perhaps thought that an AUC in the range of 0.50 to 0.60 would not be considered a good indicator; however,

what constitutes as a "good" AUC all depends on the particular medical condition being studied and medical test being used [29, 41-46].

To perform an ROC study a basic design may involve a reader and a set of test cases; however, a more typical ROC study involves multiple readers and two or more imaging systems to be evaluated [6]. The test cases are comprised of abnormal (subject with disease) and normal (subject without disease) findings where truth is known. The reader then classifies the set of cases either on a scale of 1 – 5 (considered rating-scale data) or 0 – 100 (often considered to be continuous data). Usually those subjects evaluated as normal are given a rating at the lower end of the scale and those subjects evaluated as abnormal are given a rating at the higher end of the scale. Depending on the distribution of the data and the type of data (continuous or ordinal), there are several methods for estimating the AUC.

For continuously distributed data and assuming a binormal model, a smooth ROC curve can then be estimated from the data and the AUC is estimated using the parameters (a,b) which are functions of the means and standard deviations of the two distributions (diseased and non-diseased populations). If the data is not binormally distributed but can be transformed to binormality then this direct or simple method will also work on the transformed data. This direct or simple method creates an unbiased estimate if the binormality assumption is true, or if the original data can be transformed. Another method for estimating AUC if the underlying binormality assumptions are not met but there exists a monotonic transformation is based on two transformation models described by Zou and Hall [47]. One model employs a Box-Cox transformation to the data to obtain binormality and the other is semiparametric where the transformation of the data is unspecified.

Most often the data is not binormal and sometimes can not be transformed. If the binormality assumption is not valid and the data is not transformable to a binormal distribution then there are alternative methods to estimating the AUC. One alternative is to obtain parameter estimates $(\hat{a},\hat{b})$ through computer software programs developed by Metz et al. [48, 49] based on the Dorfman and Alf method [50]. When using this method, for continuous distributions, the data is binned so that it is similar to rating scale data and then algorithms developed for rating scale data can be applied. The program ROCKIT [48, 49] is used for obtaining maximum likelihood estimates (MLEs) from rating scale data and/or continuous data. This program is also used for comparing two datasets that can be paired, partially paired, or unpaired with regard to differences between ROC index estimates and parameters. Another alternative would be to use the nonparametric method developed by Delong et al [11] for continuous rating data. This latter method does not assume binormality or even that the data is transformable to a binomial distribution. Lastly, Alonzo and Pepe developed an ROC regression model which can be also used to fit the binormal ROC curve without making assumptions about distribution of the ratings [51].

1.1.2 Cook's Distance Measure

In linear regression analysis Cook's distance (d_i) measures the influence of a single data point (observation) on the model estimates. This is done by deleting the i^{th} observation from the data set and running the model again to assess any significant changes in the estimated parameters. Specifically, given n observations $(Y_i, X_{i1}, X_{i2}, ..., X_{ik})$ where k is the number of independent parameters and i = 1, 2, …,n, Kleinbaum et al. express the least squares regression model of the

observed dependent variable Yi as $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} + E_i$, where Ei is the error term for the ith response, and $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + ... + \hat{\beta}_k X_k$ is the fitted least squares regression model, and $\hat{Y}i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + ... + \hat{\beta}_k Xi_k$ is the predicted response model at the i^{th} data point. They further define the i^{th} residual e_i as the observed values of Y_i minus the predicted values of $\hat{Y}_i$ ($e_i = Yi - \hat{Y}_i, i = 1,2,...,n$), and the estimate of population variance calculated from the sample of the n residuals as $S^2 = \frac{1}{n-k-1}\sum_{i=1}^{n} e_i^2$. Cook's distance is given by

$d_i = \frac{e_i^2 h_i}{(k+1)S^2(1-h_i)^2}$ where $h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{(n-1)S_x^2}$ (leverage value for the i^{th} observation) and $S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. A value greater than 1 for an observation would suggest influence on the model estimates based on an F random variable with k and n-k-1 degrees of freedom.

1.2 STATEMENT OF PROBLEM

Outliers can have substantial influence on the estimated AUC. For example, in a study where there are 30 positive cases and 50 negative cases, one missed positive case (i.e. rated as absolutely normal) would shift the estimate of the area by 2%. Therefore, motivated by the general principle of Cook's distance to measure influential data points by deleting the i^{th} observation and running the model again to assess any significant changes in the model estimates, we would like to be able to identify cases that may have a considerable influence on the estimated AUC; specifically, by developing a statistical test which will be used to establish if

such a case exists and if it is statistically influential. The test will be used to measure the individual effect of an observation (case) on the estimated AUC and will be investigated for both an estimator based on the binormal model and the standard nonparametric estimator. Specifically, when assuming the binormal model estimates of the parameters $(\hat{a},\hat{b})$ will be obtained directly from the means and standard deviations of the two distributions, and these will be used to obtain an estimate of the AUC. When no assumption is made regarding the underlying distribution estimates will be obtained using the procedure developed by Delong et al [11].

In both scenarios (parametric and nonparametric estimation of the AUC) simulations will be used to validate the test by estimating type I error for various sample sizes, distributional assumptions, and underlying AUC values. Statistical power will also be determined for both scenarios characterizing the occurrence of an outlier when an actually negative (normal) subject is rated as abnormal.

2.0 CONVENTIONS AND DEFINITIONS

We consider a population of subjects where true disease status is independently verified and known for each subject; therefore, we have a population of normal (N) and abnormal (M) subjects based on the presence or absence of disease giving us a total of T subjects (N+M=T). In addition, we consider a study in which a reader classifies the population of subjects on a continuous rating scale from 0 – 100 as to the presence or absence of disease/abnormality. Ratings at the higher end of the scale are more indicative of the presence of disease; therefore, lower ratings are less indicative of the presence of disease. Let $X_1, X_2, ..., X_n$ be the independent rating data for the normal subjects and $Y_1, Y_2, ..., Y_m$ be the independent rating data for the abnormal subjects.

2.1 METHODS

As stated previously, there are several different methods for fitting an ROC curve. For the purpose of our simulations we will focus on the parametric method which assumes an underlying binormal distribution and the method proposed by Delong et al for the nonparametric method. For the parametric method the parameters (a, b) are related directly to the means and standard deviations of the two distributions by the formulas $a = \frac{\mu_Y - \mu_X}{\sigma_Y}$ and $b = \frac{\sigma_X}{\sigma_Y}$, where μ_X and

μ_Y are the means and σ_X and σ_Y are the standard deviations of the binormal distribution of test cases [52]. The estimate of the AUC is then given by

$$\hat{A}_z = \Phi(\frac{\hat{a}}{\sqrt{1+\hat{b}^2}})$$

where

$$\hat{a} = \frac{\overline{Y} - \overline{X}}{S_y} \text{ and } \hat{b} = \frac{S_x}{S_y}$$

and an estimate of the variance of the A_z is given by

$$Var\ (\hat{A}) = f^2 Var\ (\hat{a}) + g^2 Var\ (\hat{b}) + 2\ fgCov\ (\hat{a}, \hat{b})$$

where

$$f = \frac{e^{-\hat{a}^2/2(1+\hat{b}^2)}}{\sqrt{2\pi(1+\hat{b}^2)}}, \quad g = -\frac{\hat{a}\hat{b}e^{-\hat{a}^2/2(1+\hat{b}^2)}}{\sqrt{2\pi(1+\hat{b}^2)^3}}$$

and

$$\hat{Var}(\hat{a}) = \frac{M(\hat{a}^2+2) + 2N\hat{b}^2}{2NM}$$

$$\hat{Var}(\hat{b}) = \frac{(M+N)\hat{b}^2}{2NM}$$

$$\hat{Cov}(\hat{a}, \hat{b}) = \frac{\hat{a}\hat{b}}{2N} \qquad [13, 52].$$

To obtain the nonparametric estimate of the AUC we used the procedure proposed by Delong et al [11] where the estimate of the AUC is given by

$$\hat{A}_D = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\Psi(X_i, Y_j), \text{ where } \Psi(X_i, Y_j) = \begin{cases} 1 & X < Y \\ 1/2 & X = Y \\ 0 & X > Y \end{cases}$$

and variance of the AUC is estimated by

$$Var\left(\hat{A}_D\right) = \frac{\frac{1}{M-1}\sum_{j=1}^{M}[V(Y_j) - \hat{A}_D]^2}{M} + \frac{\frac{1}{N-1}\sum_{i=1}^{N}[V(X_i) - \hat{A}_D]^2}{N},$$

and the X and Y components are computed as

$$V(X_i) = \frac{1}{M}\sum_{j=1}^{M}\Psi\left(X_i, Y_j\right),\ V(Y_j) = \frac{1}{N}\sum_{i=1}^{N}\Psi\left(X_i, Y_j\right) \qquad [11, 52].$$

2.2 DERIVATION OF TEST

As stated previously, we would like to be able to identify cases that may have a considerable influence on the estimated area under the ROC curve. We will measure influential data points by deleting the ith observation and re-estimating area to assess any significant changes in the estimated area. The null hypothesis we wish to test is that the area under the ROC curve (AUC) denoted by A_T is equal to the area with the ith observation (case) deleted denoted by $A_{(-i)}$; that is

H_O: $A_T = A_{(-i)}$ against H_A: $A_T \neq A_{(-i)}$. To test this hypothesis we propose the test statistic:

$$Z = \frac{\hat{A}_{(-i)} - \hat{A}_T}{\sqrt{Var(\hat{A}_T - \hat{A}_{(-i)})}}$$

Under the null hypothesis we will investigate whether this test statistic is asymptotically normally distributed. We derive the variance of the differences by utilizing the jackknife technique.

Jackknife is a technique used to reduce bias and was first developed by Quenouille (1956) [53, 54] and named by Tukey (1958) [55]. Later, the jackknife method was applied to ROC analysis [16]. The basic idea of the jackknife technique involves estimating a parameter θ called the jackknife estimator which we denote by $\tilde{\theta}$. The jackknife estimator is obtained using pseudovalues, which are estimates where the ith observation has been deleted. The ith pseudovalue ($\tilde{\theta}_i$) is defined as

$$\tilde{\theta}_i = n\hat{\theta}_n - (n-1)\hat{\theta}_{(-i)}$$

where $\hat{\theta}_n$ is the parameter estimate with all the data, $\hat{\theta}_{(-i)}$ is the parameter estimate with the ith observation deleted, and n is the total sample size. The pseudovalues are known to be asymptotically identically and independently distributed (iid) normal random variables. The jackknife estimate is defined to be the mean of the pseudovalues

$$\tilde{\theta} = \frac{1}{n}\sum_{i=1}^{n}\tilde{\theta}_i$$

and the variance of the pseudovalues is the estimate of the variance of $\tilde{\theta}$. If we apply this jackknife method to the MLE estimate of area under the ROC curve, this procedure would estimate the area (denoted by A) T times since there are T total cases. Let $\hat{A}_{(-i)}$ denote the estimate of A with the ith case deleted, and let $\tilde{A}_i$ denote the ith pseudovalue where

$$\tilde{A}_i = T\hat{A}_T - (T-1)\hat{A}_{(-i)}$$

and can be rewritten as

$$\tilde{A}_i = \hat{A}_T + (T-1)(\hat{A}_T - \hat{A}_{(-i)}).$$

Then the jackknife estimate denoted by $\tilde{A}$ is the mean of the pseudovalues so that

$$\tilde{A} = \frac{1}{T}\sum_{i-1}^{T} \tilde{A}_i$$

and since $\tilde{A}_i$ are asymptotically iid

$$Var\left[\tilde{A}\right] \approx \frac{1}{T} Var[\tilde{A}_i].$$

In addition,

$$Var[\tilde{A}] \approx Var[\hat{A}_T] = \sigma^2;$$

therefore, this implies

$$Var[\tilde{A}_i] = T\sigma^2.$$

To derive $Var(\hat{A}_T - \hat{A}_{(-i)})$ we know in some cases of unbiased statistic the estimated area under the ROC curve is equal to the jackknife estimate, that is $\hat{A}_T = \tilde{A}$ and

$$\hat{A}_T - \hat{A}_{(-i)} = \frac{\tilde{A}_i - \hat{A}_T}{T-1} \approx \frac{\tilde{A}_i - \tilde{A}}{T-1}.$$

Therefore,

$$Var[\hat{A}_T - \hat{A}_{(-i)}] = \frac{1}{(T-1)^2} Var[\tilde{A}_i - \hat{A}_T] \approx \frac{1}{(T-1)^2} Var[\tilde{A}_i - \tilde{A}]$$

$$= \frac{1}{(T-1)^2}\left[Var(\tilde{A}_i) + Var(\tilde{A}) - 2Cov(\tilde{A}_i, \tilde{A})\right] \approx \frac{1}{(T-1)^2}\left[T\sigma^2 + \sigma^2 - \frac{2}{T}T\sigma^2\right] = \frac{\sigma^2}{(T-1)}.$$

Based on the above derivation, the test statistic for the hypothesis H_O: $A_T = A_{(-i)}$ is

$$Z = \frac{\hat{A}_{(-i)} - \hat{A}_T}{\sqrt{Var(\hat{A}_T)/(T-1)}} \qquad 2.2.a$$

where T is the total number of normal (N) and abnormal (M) cases (i.e., T = N+M). The distributional properties of the test statistic will be investigated by simulating the type I error in testing H_o: $A_T = A_{(-i)}$ and estimating statistical power for a specified number of cases.

2.3 SIMULATION DESCRIPTION

To determine the adequacy of our assumption of a normal approximation of the test statistic we modeled rating data for samples of normal and abnormal subjects generated from a binormal distribution. Let X_i, i = 1, …n, be independently normally distributed random variables of ratings for normal subjects, i.e., $X_i \overset{i.i.d.}{\sim} N(\mu_X, \sigma_X^2)$, and Y_j, j = 1, …m, be independently normally distributed random variables of ratings for abnormal subjects, i.e., $Y_j \overset{i.i.d.}{\sim} N(\mu_Y, \sigma_Y^2)$. Parametric and nonparametric estimates of the area under the ROC curve (AUC) and the variance of the AUC were then estimated on the complete dataset. To obtain parametric estimates we used parameters (a,b) so that $\hat{a} = \frac{\overline{Y} - \overline{X}}{S_y}$ and $\hat{b} = \frac{S_x}{S_y}$, and to obtain nonparametric estimates we used the procedure developed by Delong et al. [11]. Then using the same methods we obtained parametric and nonparametric estimates on the data with the i^{th} subject removed and computed the test statistic given in 2.2.a. We investigated the distributional properties of the test statistic obtained by both the parametric and nonparametric methods.

The variance estimator of the numerator of the test statistic, also called the jackknife influence value (Efron and Tibshirani 1993), derived in section 2.2 is based on the one-sample jackknife technique. Such a variance estimator estimates the variance when jackknife

pseudovalues can be based on either normal or abnormal subjects with the probability proportional to the sample prevalence. Namely, the variance estimator provides an estimate of the variability of the jackknife influence value based on a subject randomly selected from the sample with a structure similar to the one in the observed sample.

This procedure allows for a simplification in the simulation model when ratings for normal and abnormal subjects are normally distributed with the same variance and an unbalanced sample is used. Namely, we can simulate test statistics based only on a deleted normal subject. The results of the observed test will be the same due to the equality of the distributions of the pseudovalues based on normal or abnormal subjects.

We adopt this simplified approach. A more general simulation model applicable to when sample sizes or variances are unequal can be constructed by randomly choosing between deleting a normal or abnormal subject with probability proportional to the sample prevalence.

2.4 SIMULATION STUDY TO ESTIMATE TYPE I ERROR

Simulations were performed using the software package SAS (version 9.1; SAS Institute, Cary, NC). In SAS, the simulation model is defined by a set of parameters specified by the user. By changing the parameters we are able to simulate data for different sample sizes (total number of normal and abnormal subjects (T=N+M)), AUCs, slope parameter b, and the mean and standard deviation of the normally distributed random variables of ratings for normal subjects which would then determine the distribution of ratings for abnormal subjects. Specifically, simulations were performed for 1) different sample sizes which included N=20 and M=20, N=40 and M=40, N=60 and M=60, and N=100 and M=100, 2) different values of AUC which included 0.50, 0.65,

0.75, 0.85, and 0.95, 3) $b = 1$, and 4) ratings for normal subjects distributed as $X_i \overset{i.i.d.}{\sim} N(16,64)$. Note the derived test statistic with the considered estimation approaches is invariant with respect to location-scale transformations. For considered scenarios we simulated from 10,000 datasets. Software code is provided in Appendix A.

2.4.1 Results

Observed type I error rates of the test statistic at an alpha = 0.05 are provided in Table 2.4.1.1 and illustrated in Figure 2.4.1.1 for both the parametric and nonparametric estimate methods for different values of AUC, different sample sizes, $b = 1$, and ratings for normal subjects distributed $X_i \overset{i.i.d.}{\sim} N(16,64)$. Figures 2.4.1.2 – 2.4.1.5 illustrates the distributions of the test statistic for the parametric and nonparametric estimation methods for the smallest and largest values of AUC and sample sizes.

When AUC is estimated parametrically the type I error seems to be appropriate for all considered parameters. A slight variability of the estimates can be attributed to the sampling error since all the estimates are within the bounds of reasonable values observable with 10,000 replications.

For the nonparametric estimation of AUC the type I error increases with AUC for all sample sizes until it reaches an AUC = 0.95 then it decreases. In addition, type I error of the test statistic is very low for AUC = 0.50 and decreases with increasing sample size. Figures 2.4.1.2 – 2.4.1.4 demonstrate that the distribution of the statistic with the nonparametric estimation of AUC is not normally distributed. For AUC = 0.50 and sample sizes T = 40 and T = 200, the distribution appears to be almost uniform and then skews to the right at AUC = 0.95.

Table 2.4.1.1: Type I Error Rate

Method	AUC	Total Sample Size (T) 40 subjects	80 subjects	120 subjects	200 subjects
parametric	**0.50**	0.056	0.053	0.053	0.051
	0.65	0.052	0.048	0.047	0.045
	0.75	0.052	0.049	0.049	0.050
	0.85	0.053	0.053	0.053	0.053
	0.95	0.052	0.053	0.053	0.054
nonparametric	**0.50**	0.019	0.009	0.005	0.002
	0.65	0.041	0.039	0.038	0.041
	0.75	0.059	0.060	0.061	0.061
	0.85	0.065	0.066	0.064	0.065
	0.95	0.059	0.053	0.054	0.053

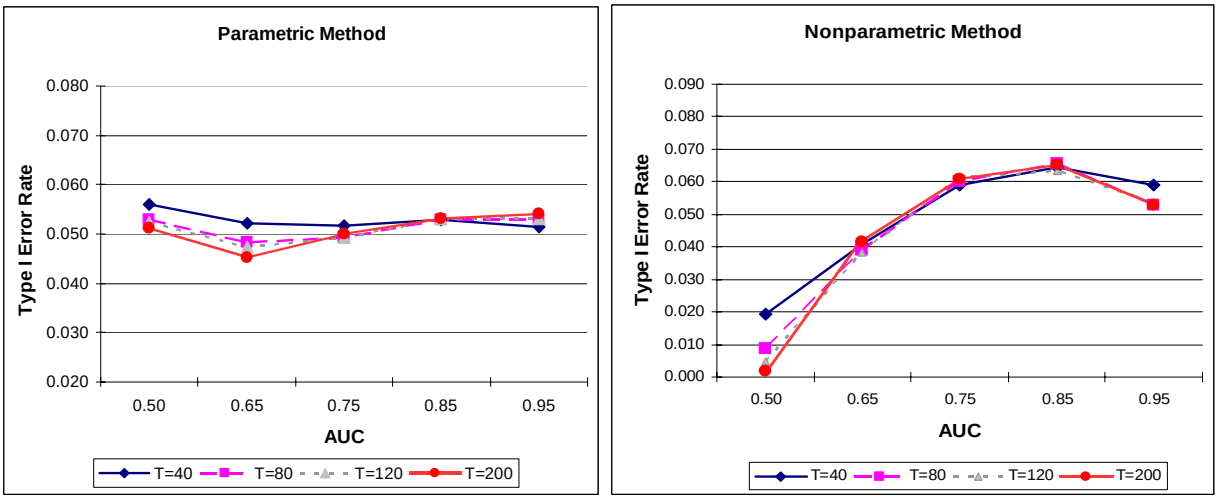

Figure 2.4.1.1: **Type I Error of the test statistic for both parametric and non-parametric estimation of AUC.**

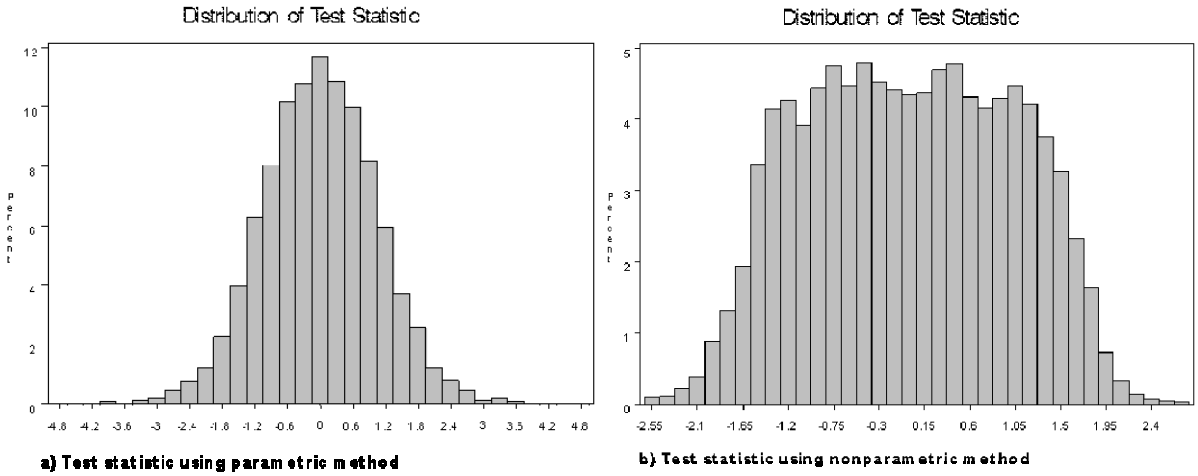

Figure 2.4.1.2: **Distributions of test statistic for small sample size and low AUC**

a) Distribution of test statistic using the parametric method for estimating AUC = 0.50, sample size T = 40, and b=1; b) Distribution of test statistic using the nonparametric method for estimating AUC = 0.50, sample size T = 40, and b=1.

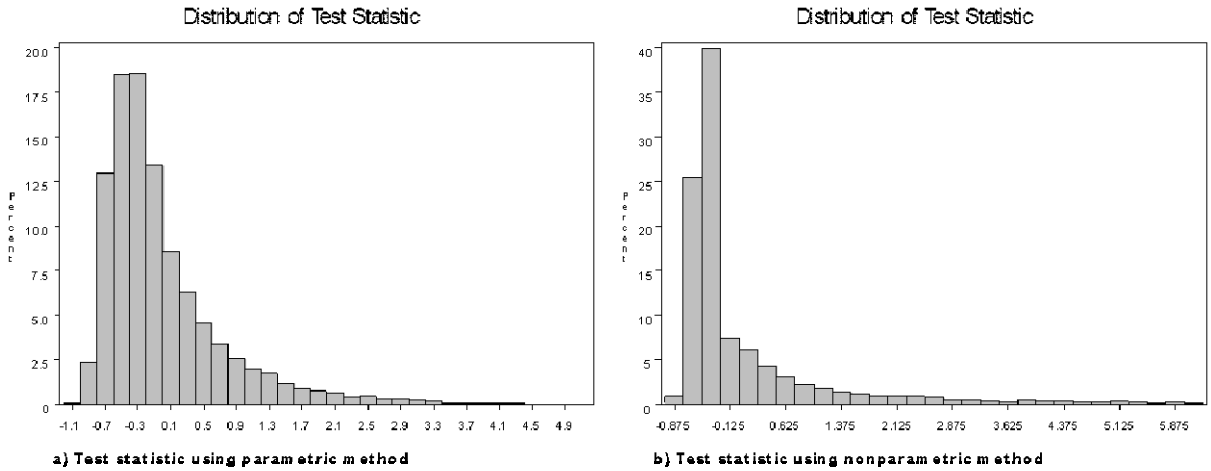

Figure 2.4.1.3: Distributions of test statistic for small sample size and high AUC

a) Distribution of test statistic using the parametric method for estimating AUC = 0.95, sample size T = 40, and b=1; b) Distribution of test statistic using the nonparametric method for estimating AUC = 0.95, sample size T = 40, and b=1

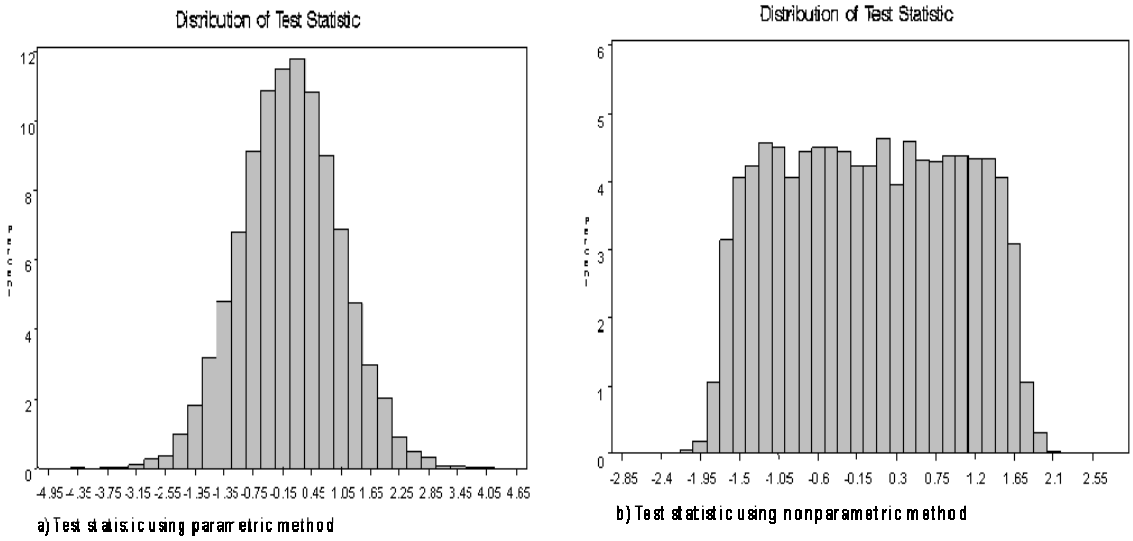

Figure 2.4.1.4: **Distributions of test statistic for large sample size and low AUC**

a) Distribution of test statistic using the parametric method for estimating AUC = 0.50, sample size T = 200, and b=1; b) Distribution of test statistic using the nonparametric method for estimating AUC = 0.50, sample size T = 200, and b=1.

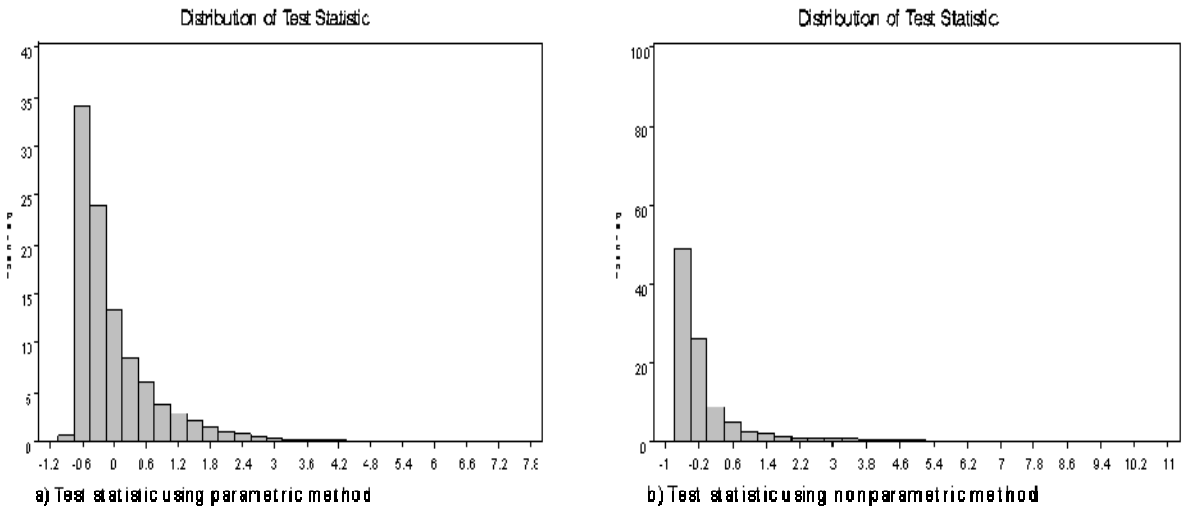

Figure 2.4.1.5: **Distributions of test statistic for large sample size and high AUC**

a) Distribution of test statistic using the parametric method for estimating AUC = 0.95, sample size T = 200, and b=1; b) Distribution of test statistic using the nonparametric method for estimating AUC = 0.95, sample size T = 200, and b=1.

2.4.2 Summary

For most scenarios the type I error of the test statistic is more dependent on how the AUC is estimated (parametric or nonparametric estimation) and the value of the true AUC, rather than the sample size. Namely, we found for the nonparametric estimation of AUC the test statistic failed to achieve the nominal type I error for many underlying AUCs and did not appear to approach normality in the range we investigated; therefore, based on our limited simulation study for type I error, the considered statistic based on nonparametric AUC estimator does not appear to be a useful method for detecting or assessing influence of outliers on the estimated AUC. However, the parametric test statistic provided reasonable type I error rate for both small and large sample sizes and for AUCs in the range we investigated.

2.5 SIMULATION STUDY TO ESTIMATE POWER

Simulations were performed using the same program as described for estimating type I error, but with some modifications to the program. To estimate the power (i.e., probability of rejecting the null hypothesis when it is known to be false) of the test statistic we assume a specific observation (rating) came from a population with a different true AUC. If this is true we can then generate an observation from a different population of normally distributed normal or abnormal subjects and include it in the generated sample of normal or abnormal rating data. We then estimate AUC for the entire sample and then estimate AUC again with the observation from the different population deleted to calculate the test statistic and estimate power. Similar to the type I error investigation, for the considered parameters of the simulation study (i.e. equal sample sizes and variances of ratings) we can simplify the numerical investigation by considering only normal outliers in the normal set of cases. Thus, we estimated statistical power by generating an observation (outlier) from a different population of normal subjects. Since absolute values greater than three standard deviations are unlikely we generate our different population of normal subjects with mean three standard deviations away from the generated sample mean of our original distribution of normal rating data to obtain our outlier. Note we will have greater statistical power to detect outliers greater than three standard deviations away from the mean of our original distribution of normal rating data. We again simulated from 10,000 datasets.

2.5.1 Results

The observed power estimates of the test statistic are provided in Table 2.5.1.1 and illustrated in Figure 2.5.1.1 for both the parametric and nonparametric estimate methods for the scenario of a

normal subject being rated as abnormal. We estimated power for different values of AUC which included 0.50, 0.65, 0.75, 0.85, and 0.95, different sample sizes which included N=20 and M=20, N=40 and M=40, N=60 and M=60, and N=100 and M=100, $b = 1$, rating data for a sample of normal and abnormal subjects generated from a binormal distribution with normal subjects distributed $X_i \overset{i.i.d.}{\sim} N(35,64)$, and to generate a higher rating for a normal subject the normal single rating data (i.e., the outlier) was generated from a population of normal subjects distributed $X_i \overset{i.i.d.}{\sim} N(59,64)$.

We observe increasing statistical power with increasing sample size with the exception of estimating AUC by the nonparametric method for AUC = 0.50 where the statistical power decreases with increasing sample size. We would expect increasing power with increasing sample size since as AUC increases the variance of AUC decreases much faster than sample increases; the unexpected result for AUC of 0.50 seems to be caused by rapid decrease of the type I error rate. Another unusual observation is the decrease in power at AUC = 0.95 for the nonparametric method, this phenomenon is also likely to be caused by a corresponding decrease in the type I error rate. We also observe increasing power as AUC increases until AUC reaches 0.95 where it levels off.

Table 2.5.1.1: **Power when a single normal rating is generated from a different population**

Method	Average AUC with outlier	AUC with out outlier	Total Sample Size (T) 40 subjects	80 subjects	120 subjects	200 subjects
parametric	**0.48**	**0.50**	0.723	0.795	0.818	0.834
	0.62	**0.65**	0.853	0.878	0.884	0.892
	0.72	**0.75**	0.875	0.895	0.900	0.906
	0.82	**0.85**	0.884	0.902	0.909	0.913
	0.93	**0.95**	0.884	0.902	0.909	0.913
nonparametric	**0.49**	**0.50**	0.139	0.102	0.072	0.042
	0.63	**0.65**	0.633	0.753	0.802	0.852
	0.73	**0.75**	0.842	0.898	0.910	0.919
	0.83	**0.85**	0.892	0.916	0.922	0.926
	0.93	**0.95**	0.889	0.900	0.905	0.909

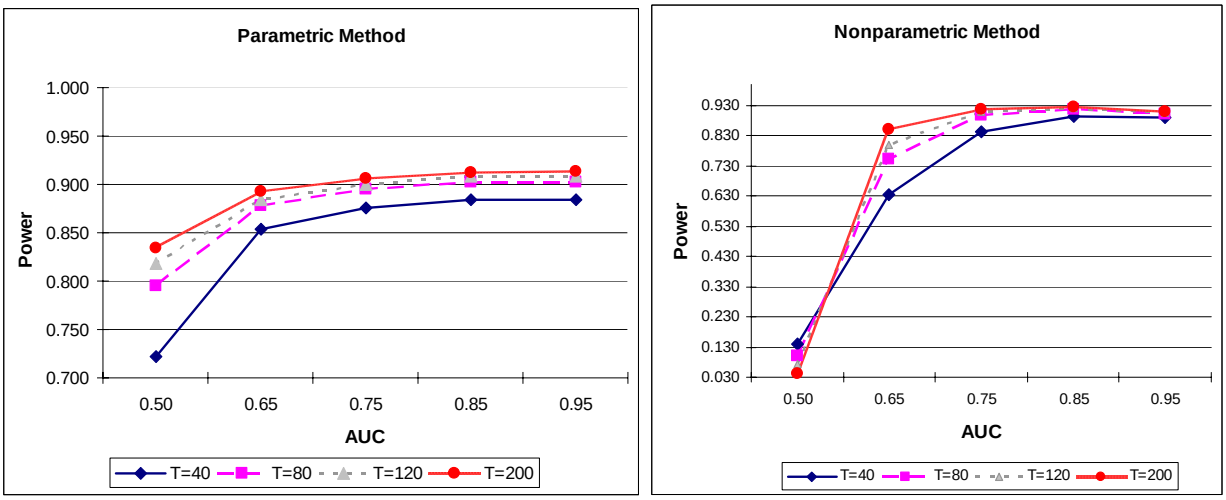

Figure 2.5.1.1: **Power when a single normal rating is generated from a different population**

2.5.2 Summary

In our limited simulation study reasonable power was achieved for the parametric test statistic. Specifically, reasonable power was achieved for AUC values in the range of 0.65 to 0.95 for $T \geq 40$ and for AUC = 0.50 reasonable power was achieved for $T \geq 120$. However, for the nonparametric estimation method we observed similar patterns in the power as we did for the type I error rates. Therefore, based on the type I error and power results, the proposed test would be useful when AUC is estimated by parametric methods and not when AUC is estimated by nonparametric methods.

3.0 CONCLUSION AND DISCUSSION

To the best of our knowledge, outliers and influential data points in ROC analysis have not been investigated to date. In this work we attempted to derive a test statistic that is approximately normally distributed asymptotically for the detection of outliers and evaluation of potential influence on one of the more frequently used measures of accuracy (i.e., the area under the ROC curve – AUC) in ROC analysis. Research regarding the detection and the evaluation of outliers is important in determining the accuracy of conclusions. Evaluating outliers may be especially important for the ROC model since subtle (difficult) cases have the potential for being miss-diagnosed by a reader (e.g. a difficult positive case is rated as an unquestionably negative case), and this can have a considerable influence on the estimated area under the ROC curve. Once an outlier(s) is flagged and evaluated as to whether it causes considerable influence on the estimated parameter of interest then the outlier can be investigated further. Namely, is the outlier a result of a data entry error, sampling error, an indication of some other problem, or a "true" outlier (i.e. not an error or any other problem with the data set)? Whatever the cause, after further investigation, more accurate or suitable conclusions can then be made regarding the outcome of the analysis of the data set in question.

In our study we investigated the distributional properties of the derived test statistic when AUC is estimated assuming the binormal distribution and by the nonparametric method proposed by Delong et al [11] to assess if the statistic is a reasonable test. We estimated type I error and

power for various parameters and scenarios. We found for the nonparametric estimation of AUC the test statistic failed to achieve the nominal type I error for many underlying AUCs and did not appear to approach normality in the range we investigated. The developed test statistic when based on the nonparametric estimator of the AUC appeared to have a highly non-normal distribution. Therefore, in its current form the nonparametric test does not appear to be a useful method for detecting or assessing influence of outliers on the estimated AUC. The reasons for such unusual behavior of the nonparametric test statistic warrant further investigation. An alternative approach could be to explore a modified permutation test for the detection and influence of outliers.

The parametric test statistic provided reasonable type I error for both small and large sample sizes and for all considered AUCs. In addition, in our limited simulation study reasonable power was achieved for the parametric test statistic. Specifically, reasonable power was achieved for AUC values in the range of 0.65 to 0.95 for $T \geq 40$ and for AUC = 0.50 reasonable power was achieved for $T \geq 120$. Therefore, the proposed test would be useful when AUC is estimated by parametric methods for $T \geq 40$ for AUC values in the range of 0.65 to 0.95 and for AUC values in the range of 0.50 to 0.95 for larger sample sizes (i.e., $T \geq 120$). In these particular situations the test statistic would be a useful tool for the assessing the influence, if any, of a single observation on the estimated AUC provided the data has an underlying binormal distribution. In addition, the test would be easy to implement and normal tables can be used to obtain the significance level.

One of the limitations of the considered parametric test includes the underlying binormality assumption of the data. In reality, most often, ROC type data is not binormally distributed. However, the simple parametric method described in this work can be used on

transformed data. Namely, if the data has one-to-one transformational properties then the data can be transformed to binormality and AUC can then be estimated by the parametric method on the transformed data.

Several issues or limitations apply to both parametric and nonparametric estimation methods. First, we observed the derived distribution of the test statistic skew to the right as the AUC increased. This phenomenon is due to the fact that as the AUC increases the AUC approaches its absolute bound of 1.0. Some skew-alleviating transformations, such as logit or probit, can be considered to address this problem.

Second, the formula for the variance estimator derived in section 2.2 reflects the variability of the numerator (i.e., the jackknife influence values) when these are based on the subject randomly selected from a sample with fixed prevalence of abnormal subjects. While such property permits making inferences blindly to the true status of the subjects (hence are likely to be useful for a more complex problem of detecting an outlier), in some instances it may also be useful to implement another approach. Namely, for investigating a given observation with known truth status (i.e., conditioning on the true status) one might want to know the variability of the jackknife influence values based only on normal or abnormal subjects. The formula for the corresponding variance estimator can also be derived for the nonparametric AUC estimator and might be derivable for a general AUC estimator. This problem, however, is outside the scope of this work.

Third, in the regression setting a modified residual is often considered where the modification is obtained by excluding the "tested" observation from the variance estimation. A similar approach can be implemented in our problem; however, our preliminary studies do not indicate any substantial improvement in using such modification.

Lastly, we investigated only the case of equal sample sizes. Often ROC data does not include equal sample sizes and in fact these studies usually include a larger sample of normal subjects. We might expect an outlier to have greater influence on the AUC if the outlier is located in the smaller abnormal sample. Possible future work in this area could include a more in depth investigation of the possible effects of different sample sizes and where the outlier is located (i.e., normal or abnormal subjects) on the type I error and power of the parametric test developed and investigated in this work. However, for this investigation a two-sample jackknife approach should be considered.

APPENDIX A

[SAS PROGRAM]

```
%macro simulations(n_sim=, n_x=, n_y=, t_auc=, b=, mu_x=, s_x=);

      %do sim=1 %to &n_sim;
      /*************************************************
            Simulated Dataset
      *************************************************/

      Data rdata;
            www=mod(&sim,10);
            sss = &sim;
            if www = 0 then do;
                  put 'sim=' sss;
                  end;

            seed = 56832+&sim*100;

            n_x=&n_x;                              /*Number of cases*/
            n_y=&n_y;
            n_t=n_x+n_y;


            t_auc=&t_auc;
            b=&b;

            /*Normal cases*/
            mu_x=&mu_x;
            s_x=&s_x;

            /*Abnormal cases*/
            s_y=s_x/b;
            mu_y=mu_x+s_y*sqrt(1+b**2)*probit(t_auc);


            id_x=0;                                /*Id's for subjects*/
            id_y=0;
            do id = 1 to n_t;
                  call rannor(seed, z);
```

```
                        if id <= n_x then do;
                              id_x=id;
                              rating = mu_x + (s_x*z);
                              abnormal = 0;
                     /* Abnormality status, 0-normal (X)*/
                              end;
                        else  do;
                              id_y=id-n_x;
                              rating = mu_y + (s_y*z);
                              abnormal = 1;
                     /*Abnormality status, 1-abnormal (Y)*/
                              end;
                  output;
                  id_x=0;
                  id_y=0;
            end;
      run;
      title1 "Rating Dataset";



      /* This section of the program is for the power simulation where we
         get a single normal data point from a different distribution*/
      Data single;
            seed2 = 634285+&sim*100;


            mu_x=59;                        /*Normal cases*/
            s_x=8;

            call rannor (seed2,z);
            rating = mu_x + (s_x*z);
            abnormal = 0;
            id=10000;
            output;
    run;



      data rboth;
            set rdata single;
      run;

      proc sort data=rboth;
            by abnormal;
      run;



/*Run the programs on Complete data*/
      %nmle_auc(datain=rboth,rating=rating,abnormal=abnormal,
                  dataout=nmle_out, create=1, key=0);

      %delong_auc(datain=rboth,rating=rating,abnormal=abnormal,
                  dataout=delong_out, create=1, key=0);
```

```
/*Jackknife all subjects*/
title1      "jackknifed data";
%jackknife_data(data=rboth,id=id, id_min=10000, id_max=10000);

/*Merging the results*/
data all_out;
      merge nmle_out delong_out ;
      by key;
run;



/* Compute statistic */
proc sort data=all_out;
      by key;
run;

data one_sim;
      set all_out;
      retain auc_nmle0 s2_nmle0 auc_delong0 s2_delong0 nt0;
            if key = 0 then do;
                  auc_nmle0 = auc_nmle;
                  s2_nmle0 = s2_nmle;
                  auc_delong0 = auc_delong;
                  s2_delong0 = s2_delong;
                  nt0 = nt;
            end;

      /*calculate statistic*/
      L_nmle = (auc_nmle - auc_nmle0)/(sqrt(s2_nmle0/(nt0-1)));
      L_delong = (auc_delong - auc_delong0)/(sqrt(s2_delong0/(nt0-1)));

      /*calculate probability*/
      normp1 = 1 - probnorm(ABS(L_nmle));
      normp2 = 1 - probnorm(ABS(L_delong));

            if normp1 <= .025 then do;
                  sum1 = 1;
                  end;
            else do;
                  sum1 = 0;
                  end;


            if normp2 <= .025 then do;
                  sum2 = 1;
                  end;
            else do;
                  sum2 = 0;
                  end;


      sim=&sim;

      if key = 0 then delete;
run;
```

```
            %if &sim=1 %then  %do;
                                    data all_sim;
                                                set one_sim;
                                    run;
                              %end;
                              %else %do;
                                    data all_sim;
                                          set all_sim one_sim;
                                    run;
                              %end;

      %end;


%mend;
option nonotes;
option pagesize=100;
%simulations(n_sim=10000, n_x=99, n_y=100, t_auc=.65, b=1, mu_x=35, s_x=8)

proc means data=all_sim mean var;
      var auc_nmle0 auc_delong0 s2_nmle0 s2_delong0
            auc_nmle auc_delong
            L_nmle L_delong sum1 sum2;
run;


Proc univariate data=all_sim;
      var L_nmle L_delong;
      title1      "Distribution of Test Statistic";
      histogram L_nmle L_delong/ normal(noprint color=red)
        kernel(noprint color=blue) cfill=ligr;

run;

/*********************************************
      Naive MLE Estimates
*********************************************/
%macro nmle_auc(datain=,rating=,abnormal=,
                        dataout=,create=1,key=);

Proc iml;

      use &datain;
      read all var {&rating} into x where (&abnormal = 0);
      read all var {&rating} into y where (&abnormal = 1);

      mux=x[:];
      muy=y[:];

      ssx=(x-mux);
      ssx2=ssx##2;
      ssqx=sum(ssx2);

      ssy=(y-muy);
      ssy2=ssy##2;
      ssqy=sum(ssy2);
```

```
      nx=nrow(x);
      ny=nrow(y);
      nt=nx+ny;

      sigma_x=sqrt(ssqx/(nx-1));
      sigma_y=sqrt(ssqy/(ny-1));

      /* ROC Parameters */
      a=(muy-mux)/(sigma_y);
      b=sigma_x/sigma_y;

      s2_a=(ny*(a**2+2)+(2*nx*b**2))/(2*nx*ny);
      s2_b=((ny+nx)*b**2)/(2*ny*nx);
      cov_ab=(a*b)/(2*nx);

      /* Naive MLE of the AUC */
      auc_nmle=probnorm(a/(sqrt(1+(b**2))));

      /* Variance of Naive MLE of the AUC*/
      pi=constant('PI');
      f=exp(-(a**2/(2*(1+b**2))))/
            sqrt(2*pi*(1+b**2));
      g=-a*b*exp(-(a**2/(2*(1+b**2))))/
            sqrt(2*pi*(1+b**2)**3);
      s2_nmle=(f**2*s2_a)+(g**2*s2_b)+(2*f*g*cov_ab);

      /*Output*/
      if &create  then  create &dataout var {key auc_nmle s2_nmle nt};
                        else  edit &dataout;

      key=&key;
      append var {key auc_nmle s2_nmle nt};
quit;

%mend;


/*********************************************
      Nonparametric Estimates
*********************************************/
%macro delong_auc(      datain=,rating=,abnormal=,
                              dataout=, create=, key=);

proc iml;

      use &datain;
      read all var {&rating} into x where (&abnormal = 0);
      read all var {&rating} into y where (&abnormal = 1);

      nx=nrow(x);
      ny=nrow(y);


      /*Matrix of order indicators*/
      temp_x=repeat(x,1,ny);
      temp_y=repeat(y`,nx,1);
```

```
      psi=(temp_x<temp_y)+0.5*(temp_x=temp_y);

      /*Nonparametric AUC estimator*/
      auc_delong=psi[:];

      /*Variance of the nonparametric AUC (DeLong)*/
      Vx=psi[,:]-auc_delong;
      Vy=psi[:,]-auc_delong;

      s2_delong=Vx[##]/(nx*(nx-1))+Vy[##]/(ny*(ny-1));

      /*Output*/
      if &create  then  create &dataout var {key auc_delong s2_delong};
                        else  edit &dataout;

      key=&key;
      append var {key auc_delong s2_delong};
quit;
%mend;

/*********************************************
      Jackknifing the data
*********************************************/
%macro jackknife_data(data=, id=, id_min=, id_max=,id_step=);
      %do i=&id_min %to &id_max;
            data i_minus;
                  set &data;
                  if &id^=&i;
            run;

            %nmle_auc(datain=i_minus,rating=rating,abnormal=abnormal,
               dataout=nmle_out, create=0, key=&i);

            %delong_auc(datain=i_minus,rating=rating,abnormal=abnormal,
               dataout=delong_out, create=0, key=&i);

      %end;

%mend;
```

BIBLIOGRAPHY

1. Kleinbaum DG, Kupper LL, Muller KE, Nizam A. Applied regression analysis and other multivariable methods. Pacific Grove: Duxbury Press, 1998.

2. Rockette HE, King JL, Medina JL, Eisen HB, Brown ML, Gur D. Imaging systems evaluation: effect of subtle cases on the design and analysis of receiver operating characteristic studies. AJR Am J Roentgenol. 1995; 165(3):679-683.

3. Gur D, Rockette HE, Armfield DR, et al. Prevalence effect in a laboratory environment. Radiology. 2003 Jul;228(1):10-4.

4. Swets JA. ROC analysis is applied to the evaluation of medical imaging techniques. Invest Radiol. 1979; 14:109-121.

5. Metz CE. ROC methodology in radiologic imaging. Invest Radiol. 1986; 21:720-733.

6. Hanley JA. Receiver operating characteristic (ROC) methodology: The state of the art. Critical Rev Diagn Imaging 1989; 29:307-335.

7. Lusted, LB. Logical analysis in Roentgen diagnosis. *Radiology*, 1960; 74:178-193.

8. Lusted, LB. Introduction to Medical Decision Making, Thomas Springfield, IL, 1968

9. Hanley JA and McNeil BJ. A method of comparing the areas under the receiver operating characteristic curves derived from the same cases. Radiology, 1983; 148:839-843.

10. McClish DK. Comparing the areas under more than two independent ROC curves. Med Decis Making, 1987; 7:149-155.

11. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics, 1988; 44:837-845.

12. Berbaum KS, Dorfman DD, Franken EA. Measuring observer performance by ROC analysis: Indications and complications. Investigative Radiology, 1989; 24: 228-233.

13. McClish DK. Analyzing a portion of the ROC curve. Med Decis Making, 1989; 9:190-195.

14. Thompson ML and Zucchini W. On the statistical analysis of ROC curves. Statistics in Medicine, 1989; 8:1277-1290.

15. Good WF, Gur D, Straub WH, Feist JH. Comparing imaging systems by ROC studies: Detection versus interpretation. Investigative Radiology, 1989; 24:932-933.

16. Dorfman DD, Berbaum KS, Metz CE. Receiver operating characteristic rating analyses generalization to the population of readers and patients with jackknife method. Invest Radiol, 1992; 27:723:731.

17. Obuchowski NA and Rockette HE. Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests: An ANOVA approach with dependent observations. Comm Statist Simulat, 1995; 24:285-308.

18. Obuchowski NA. Multireader receiver operating characteristic studies: A comparision of study designs. Acad Radiol, 1995; 2:709-716.

19. Zhou XH and Gatsonis CA. A simple method for comparing correlated ROC curves using incomplete data. Statistics in Medicine, 1996; 15:1687-1693.

20. Beiden SV, Wagner RF, Campbell G. Components of variance models and multiple bootstrap experiments: An alternative method for random effects, receiver operating characteristics analysis. Acad Radiol, 2000; 7:341-349.

21. Franken EA Jr, Berbaum KS, Marley SM, et al. Evaluation of a digital workstation for interpreting neonatal examinations. A receiver operating characteristic study. Invest Radiol. 1992; 27(9):732-7.

22. Mushlin AI, Detsky AS, Phelps CE, et al. The accuracy of magnetic resonance imaging in patients with suspected multiple sclerosis. The Rochester-Toronto Magnetic Resonance Imaging Study Group. JAMA. 1993; 269(24):3146-51.

23. MacMahon H, Engelmann R, Behlen FM, et al. Computer-aided diagnosis of pulmonary nodules: results of a large-scale observer test. Radiology. 1999; 213(3):723-6.

24. Good WF, Sumkin JH, Ganott M, et al. Detection of masses and clustered microcalcifications on data compressed mammograms: An observer performance study. AJR Am J Roentgenol. 2000; 175(6):1573-6.

25. Iinuma G, Ushio K, Ishikawa T, Nawano S, Sekiguchi R, Satake M. Diagnosis of gastric cancers: comparison of conventional radiography and digital radiography with a 4 million-pixel charge-coupled device. Radiology. 2000; 214(2):497-502.

26. Uozumi T, Nakamura K, Watanabe H, Nakata H, Katsuragawa S, Doi K. ROC analysis of detection of metastatic pulmonary nodules on digital chest radiographs with temporal subtraction. Acad Radiol. 2001; 8(9):871-8.

27. Jiang Y, Nishikawa RM, Schmidt RA, Toledano AY, Doi K. Potential of computer-aided diagnosis to reduce variability in radiologists' interpretations of mammograms depicting microcalcifications. Radiology 2001; 220:787-794.

28. Lowe LH, Draud KS, Hernanz-Schulman M, et al. Nonenhanced limited CT in children suspected of having appendicitis: Prospective comparison of attending and resident interpretations. Radiology 2001; 221:755-759.

29. Fuhrman CR, Britton CA, Bender T, et al. Observer performance studies: Detection of single versus multiple abnormalities of the chest. AJR Am J Roent. 2002; 179:1551-1553.

30. Pisano ED, Cole EB, Kistner EO, et al. Interpretation of digital mammograms: Comparison of speed and accuracy of soft-copy versus printed-film display. Radiology 2002; 223:483-488.

31. Kakeda S, Nakamura K, Kamada K, et al. Improved detection of lung nodules by using a temporal subtraction technique. Radiology 2002; 224:145-151.

32. Huo Z, Giger ML, Vyborny CJ, Metz CE. Breast cancer: Effectiveness of computer-aided diagnosis – observer study with independent database of mammograms. Radiology 2002; 224:560-568.

33. Shiraishi J, Abe H, Engelmann R, Aoyama M, MacMahon H, Doi K. Computer-aided diagnosis to distinguish benign from malignant solitary pulmonary nodules on radiographs: ROC analysis of radiologists' performance – initial experience. Radiology 2003; 227:469-474.

34. Obuchowski NA. Receiver operating characteristic curves and their use in radiology. Radiology 2003; 229:3-8.

35. Kakeda S, Moriya J, Sato H, et al. Improved detection of lung nodules on chest radiographs using commercial computer-aided diagnosis system. AJR Am J Roent. 2004; 182:505-510.

36. Metz CE. Basic principles of ROC analysis. Seminars in Nuclear Med. 1978; 8(4):283-298.

37. van Erkel AR, Pattynama PM. Receiver operating characteristic (ROC) analysis: Basic principles and applications in radiology. European Journal of Radiology 1998; 27:88-94.

38. Swets JA, Pickette RM. Evaluation of diagnostic systems: Methods from signal detection theory. New York: Academic Press, 1982.

39. Swets JA. Signal detection theory and ROC analysis in psychology and diagnostics: collected papers. New Jersey: Lawrence Erlbaum Associates, Inc., 1996.

40. Hanley JA and McNeil BJ. The meaning and use of the area under the receiver operating characteristic curve. Radiology 1982; 143:29-36.

41. Li Q, Li F, Shiraishi J, Katsuragawa S, Sone S, Doi K. Investigation of new psychophysical measures for evaluation of similar images on thoracic computed tomography for distinction between benign and malignant nodules. Med Phys. 2003; 30(10):2584-2593.

42. Shah SK, McNitt-Gray MF, De Zoysa KR, et al. Solitary pulmonary nodule diagnosis on CT: results of an observer study. Acad Radiol. 2005; 12(4):496-501.

43. Title RS, Harper K, Nelson E, Evans T, Tello R. Observer performance in assessing anemia on thoracic CT. AJR Am J Roentgenol. 2005; 185(5):1240-12444.

44. Nardone G, Rocco A, Staibano S, et al. Diagnostic accuracy of the serum profile of gastric mucosa in relation to histological and morphometric diagnosis of atrophy. Aliment Pharmacol Ther. 2005; 22(11-12):1139-1146.

45. Ouwendijk R, Kock MC, van Dijk LC, van Sambeek MR, Stijnen T, Hunink MG. Vessel wall calcifications at multi-detector row CT angiography in patients with peripheral arterial disease: effect on clinical utility and clinical predictors. Radiology. 2006; 241(2):603-608.

46. Gurung J, Maataoui A, Khan M, et al. Automated detection of lung nodules in multidetector CT: influence of different reconstruction protocols on performance of a software prototype. Rofo. 2006; 178(1):71-77.

47. Zou KH and Hall WJ. Two transformation models for estimating an ROC curve derived from continuous data. Journal of Applied Statistics 2000; 27:621-631.

48. Metz CE, Herman BA, Roe CA. Statistical comparison of two ROC curve estimates obtained from partially-paired datasets. Med Decis Making 1998; 18: 110.

49. Metz CE, Herman BA, Shen J-H. Maximum-likelihood estimation of ROC curves from continuously-distributed data. Stat Med 1998; 17: 1033.

50. Dorfman DD and Alf E Jr. Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals – rating method data. Math Psych. 1969; 6:487-496.

51. Alonzo TA and Pepe MS. Distribution-free ROC analysis using binary regression techniques. Biostatistics 2002; 3:421-432.

52. Zhou XH, Obuchowski NA, McClish DK. Statistical Methods in Diagnostic Medicine. New York: Wiley & Sons Inc., 2002.

53. Quenouille MH. Approximate test of correlation in time series. J Roy Stat Soc Ser B 1949; 11:68-84.

54. Quenouille MH. Notes on bias in estimation. Biometrika 1956; 43:353-360.

55. Tukey JW. Bias and confidence in not quite large samples. Ann Math Statist 1958; 29:614. Abstract.