2009

Confidence Interval Estimation of the Area under the Receiver Operating Characteristic Curve in the Presence of Measurement Error

Yanhong Li

Confidence Interval Estimation of the Area under the Receiver Operating Characteristic Curve in the Presence of Measurement Error

(Spine Title: C.I. Estimation of the Area under the ROC curve with M.E.)

(Thesis Format: Monograph)

by

Yanhong Li

Graduate Program in Epidemiology & Biostatistics

Submitted in partial fulfillment

of the requirements for the degree of Master of Science

School of Graduate and Postdoctoral Studies

The University of Western Ontario

London, Ontario

August 2009

ABSTRACT

Diagnosis of diseases is often based on biomarkers with continuous measurements. The discriminative ability of a biomarker can be depicted by a receiver operating characteristic (ROC) curve, which shows simultaneously the proportions of both abnormal and normal subjects correctly diagnosed at various cutoff points in the marker values. The area (A) under the ROC curve is commonly used to measure the ability of the marker to distinguish between two populations. Many biomarkers are subject to measurement error, which must be taken into account in statistical inference for A to avoid misleading results. Assuming a normal distribution for biomarker values, this thesis developed a confidence interval procedure for A adjusted for random measurement error that can be quantified by an external reliability study. The basis of the new procedure is the method of variance estimates recovery. Simulation results show that this procedure outperformed the one based on the Delta method. The methodology is illustrated by a data set from a study using thiobarbituric acid reaction substance to diagnose cardiovascular disease.

Keywords: diagnosis, coverage, Delta method, simulation

ACKNOWLEDGEMENTS

I would like to express my gratitude to all people who helped and inspired me during my study in The University of Western Ontario.

I especially would like to thank my supervisor, Dr. GuangYong Zou, who gave me guidance, encouragement and financial support to complete this thesis. His enthusiasm and hard working on research motivated me throughout my research and writing of this thesis. His great efforts to supervise made this thesis possible.

My sincere thank also goes to Dr. John Koval, whose patience greatly impressed me. His accessibility for help and insightful comments on my thesis enabled me to improve this thesis gradually.

I also appreciate the Department of Epidemiology and Biostatistics for offering me Schulich Graduate Scholarship and excellent academic environment to complete my studying in the University of Western Ontario.

I am indebted to Julia Taleban, Danuta Kowalik, Bin Zhang and Lihua Yue, with whom I always discussed. They offered me clear concepts, good suggestions and encouragement at any time when I had problems in writing this thesis.

Lastly, I give my special thank to my parents, Ruilun Li and Guifang Li, who loved me, taught me and supported me all the time in my life.

TABLE OF CONTENTS

LIST OF TABLES

Chapter 1

INTRODUCTION

1.1 What is a receiver operating characteristic curve?

Many diseases are diagnosed on the basis of biomarkers, which are indicators used to measure particular states of diseases or the effects of treatments. The discriminative ability of a biomarker can be depicted by a receiver operating characteristic (ROC) curve.

The ROC curve is a plot of test sensitivity (i.e. true positive rate) on the y-axis versus 1-specificity (i.e. false positive rate) on the x-axis for all possible cutoff points (see Fig.1.1 for three possible cutoff points). Cutoff points are decision thresholds used to classify people into "abnormal" and "normal" groups. Thus the ROC curve is a graphic means for assessing the ability of a diagnostic test to discriminate between "abnormal" and "normal" subjects. The shape of a ROC curve is determined by the amount of overlap between the distributions of test results of "abnormal" and "normal" subjects, reflecting the discriminating ability of the test.

1.2 Why is the ROC curve useful in epidemiology studies?

Diagnostic tests, such as laboratory tests (e.g. blood, urine tests), diagnostic imaging (e.g. X-rays, ultrasound) etc., play an important role in the practice of medical care. The performance of a diagnostic test or biomarker is a problem of diagnostic accuracy, which is the capability of a test to tell the difference between individuals with and without the disease of interest.

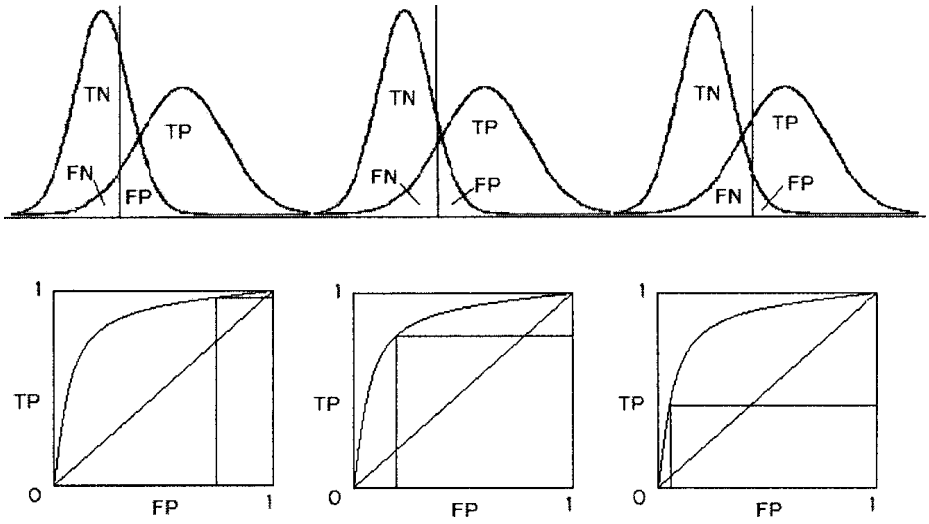

Figure 1.1: Each point on the ROC curve corresponds to a cutoff point. Note: "TN" means "true negative rate", "FN" means "false negative rate", "TP" means "true positive rate" and "FP" means "false positive rate".

There are a variety of performance measures to quantify the accuracy of a diagnostic test, such as sensitivity, specificity, positive and negative predictive values, accuracy and likelihood ratios. Of them, sensitivity and specificity are two commonly used measures to evaluate the diagnostic accuracy (Van Erkel and Pattynama, 1998). Sensitivity measures how well a test detects the presence of disease in the individuals who are actually in disease status. In other words, sensitivity is the true positive rate of correctly diagnosing an "abnormal" individual. Specificity quantifies how well a test identifies the absence of disease in the population who are in fact in healthy status. That is, specificity is the true negative rate of accurately recognizing a "normal" individual. If diagnostic tests give dichotomous results, such as "yes" or "no", sensitivity and specificity can be easily calculated by a 2×2 contingency table. In practice, however, many diagnostic tests provide ordinal results (e.g. the presence of disease - definitely, probably, possibly, probably not, definitely not) and continuous

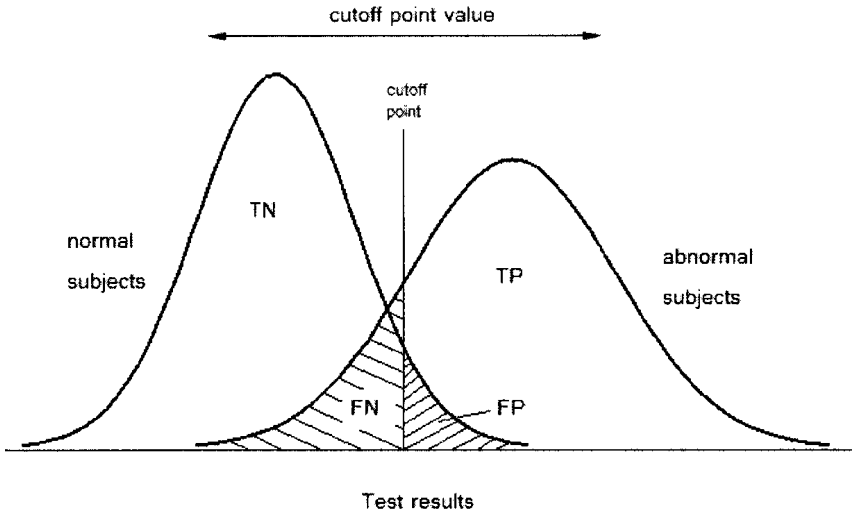

Figure 1.2: Binormal distributions. Note: "TN" means "true negative rate", "FN" means "false negative rate", "TP" means "true positive rate" and "FP" means "false positive rate".

results (e.g. blood pressure). Consequently, a cutoff point must be chosen to classify the ordinal and continuous results into dichotomous results. Since the distributions of test results that indicate the presence or absence of a disease usually overlap to some extent, any single cutoff point or decision criterion may misclassify some "abnormal" individuals as "normal", or some "normal" as "abnormal" (see Fig.1.2). Therefore, both sensitivity (i.e. true positive rate) and specificity (i.e. 1 - false positive rate) may vary as the cutoff point shifts. A lower cutoff point value will give a higher sensitivity and a lower specificity, or vice versa. Using sensitivity and specificity to describe the diagnostic accuracy of a test without presenting a cutoff point may be ambiguous and misleading. Also, it is difficult to compare two or more tests only based on a pair of sensitivity and specificity. Consequently, both sensitivity and specificity must be presented along with the corresponding cutoff point.

The ROC curve, on the other hand, can overcome the limitation that sensitivity

and specificity change as the cutoff point shifts. In fact, it is the tradeoff between the sensitivity and specificity of a diagnostic test, taking the variability of all possible cutoff points into account. Since both sensitivity and specificity are determined by the discriminative ability of a test, they are independent of disease prevalence. The ROC curve describes the test diagnostic accuracy apart from both cutoff point effect and disease prevalence (Metz, 1978). Moreover, the ROC curve offers a direct visual comparison for two or more tests on a common set of scales at all possible cutoff points, making it a more convenient tool for comparing two or more tests.

1.3 What is the area under the ROC curve?

Usually, when applying ROC curve to evaluate the discriminative capability of a diagnostic test, it is convenient to summarize the information of the ROC curve into a single index. Of several such summary indices, such as area-related, slope-related and intercept-related index (McNeil and Hanley, 1984; Shapiro, 1999; Greiner *et al.*, 2000), the area (A) under the ROC curve is a commonly used index in practice. The major reason is that A is independent of decision criteria, thus reducing the cutoff point effect on sensitivity and specificity (Van Erkel and Pattynama, 1998). It can be shown that $A = P(Y > X)$ (Bamber, 1975), where Y is the measured value of a biomarker on a randomly chosen "abnormal" subject, and X is the value of the same marker measured on a randomly chosen "normal" subject. Thus A is the probability that a randomly selected subject in "abnormal" group will have a higher test value than that of a randomly selected subject in "normal" group.

Assuming normality of biomarker measured values, A is given by:

$$\begin{aligned} P(Y > X) &= P(Y - X > 0) \\ &= P\left(\frac{(Y - X) - (\mu_Y - \mu_X)}{\sqrt{\sigma_Y^2 + \sigma_X^2}} > \frac{-(\mu_Y - \mu_X)}{\sqrt{\sigma_Y^2 + \sigma_X^2}}\right) \\ &\overset{\substack{under\ normal\ assumption}}{=} P\left(Z > \frac{-(\mu_Y - \mu_X)}{\sqrt{\sigma_Y^2 + \sigma_X^2}}\right) \\ &= P\left(Z < \frac{\mu_Y - \mu_X}{\sqrt{\sigma_Y^2 + \sigma_X^2}}\right) \\ &= \Phi\left(\frac{\mu_Y - \mu_X}{\sqrt{\sigma_Y^2 + \sigma_X^2}}\right). \end{aligned}$$

So A can be written as

$$A = \Phi(\delta), \qquad \delta = \frac{\mu_Y - \mu_X}{\sqrt{\sigma_Y^2 + \sigma_X^2}}, \tag{1.1}$$

where Φ is the standard normal cumulative distribution function, μ_Y and μ_X are the mean of test values on "abnormal" and "normal" subjects, while σ_Y^2 and σ_X^2 are the variance of test values on "abnormal" and "normal" subjects. In words, A is a function for the ratio of mean difference of test values on "abnormal" and "normal" subjects to the square root of the sum of variance for "abnormal" and "normal" subjects.

For non-normally distributed test results, A can be estimated based on the Mann-Whitney-Wilcoxon statistic (Bamber, 1975; Hanley and McNeil, 1982).

As a probability, the value of A can range from 0.0 to 1.0. A test with an area of 1.0 is a perfect test since the sensitivity (true positive rate) is 1.0 while the 1 - specificity (false positive rate) is 0.0. Namely, there is no overlap between the distributions of "abnormal" and "normal" subjects. The test can separate "abnormal" and "normal" subjects perfectly. On the other hand, A with the value of 0.5 gives no information for the accuracy of a test. A test with an area of 0.5 distinguishes the "abnormal" and "normal" subjects by pure chance. The ROC curve in this case is actually the diagonal of the unit square. Any improvement in false positive rate

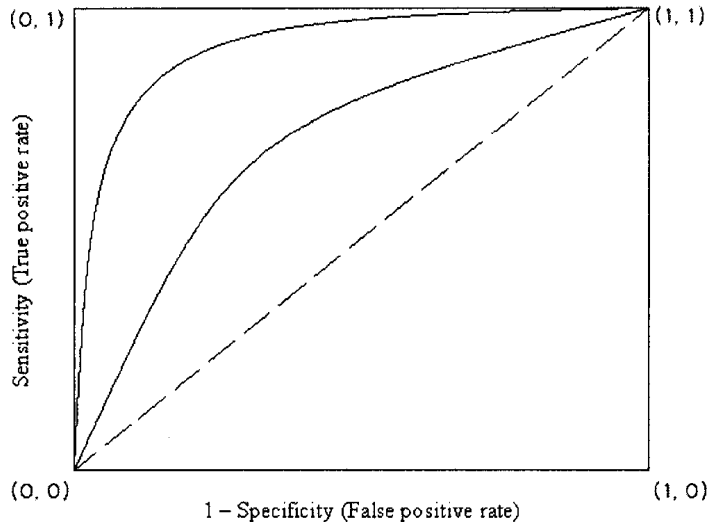

Figure 1.3: The comparison of the area under the ROC curve

comes along with a corresponding decline in the false negative rate. There is no difference between the distributions of "abnormal" and "normal" subjects, because the distributions of the "abnormal" and "normal" subjects totally overlap. A test with an area of 0.0 means that the test results are totally inaccurate. That is, this test can separate the "abnormal" and "normal" subjects as perfect as that with an area of 1.0, but mistakenly marks all the "abnormal" subjects as "normal" and all the "normal" individuals as "abnormal". In this case, it is convenient to compare the performance of the tests by changing the decision rule such that the test results can be converted, and the ROC curve thus can be flipped above the diagonal and have an area value greater than 0.5. Therefore, the practical range for A is from 0.5 to 1.0. A test with an area greater than 0.5 can provide evidence that it has the ability to distinguish between the "abnormal" and "normal" individuals. The closer to 1.0 the value of A is, the higher diagnostic accuracy the test has.

1.4 Confidence interval for the area under the ROC curve

Since A can summarize the information of the ROC curve, statistical inference in ROC curve study can be made by constructing a confidence interval for A. As shown in equation (1.1), A is a monotonic function of δ, a ratio of mean and standard deviation. Hence, the estimation of a confidence interval for A is equivalent to that for δ. An intuitive approach is to apply the Delta method, resulting in a symmetric confidence interval. However, as pointed out by Efron and Tibshirani (1993, p.180), such interval may perform poorly in practice due to its enforced symmetry. An alternative is to extend the method of variance estimates recovery as discussed by Zou and Donner (2008). This method was termed "MOVER" by Zou (2008) and can construct a confidence interval that reflects the asymmetric variance. This will be the focus in the rest of this thesis.

1.5 Why is measurement error a concern?

Measurement error is any discrepancy between true value and measured value (Koepsell and Weiss, 2003). It includes systematic error and random error. Systematic error can be caused by any factor that systematically affects measurements of the variables across samples, such as variation in laboratory equipment, inappropriate scale setting, etc. It is predictable error of measurement. Random error may be caused by any factor that randomly affects measurements of sample variables, such as temporal change in operator, data entry mistake before analysis. It is unpredictable.

Measurement error is a major source of bias in epidemiological studies. Usually, systematic error affects the mean of observed variable and is called bias. Random measurement error affects the variability around the mean and thus can introduce over-dispersion in the outcome. Both systematic and random errors can be either "differential" or "non-differential." Differential measurement error can cause the observed association between exposures and the outcome to have stronger or weaker

relationship than the truth, or can result in an opposite association, thus giving a totally fallacious conclusion for the study. Non-differential measurement error can attenuate the null value of no association when evaluating the association between exposures and the outcome. As a result, it gives a less power for a study to detect an association for a fixed sample size, or requires a larger study sample size to have the same power to detect the true relationship between exposures and the outcome, making a statistically significant difference more difficult to achieve (Armstrong *et al.*, 1992).

Normally, quality control measures in laboratories are used to calibrate measurement methods and prevent systematic error. Thus, the effect of non-differential systematic error on exposure-disease relationships is easy to predict. However, differential measurement error can introduce unpredictable bias into a study, even though it can be prevented by a good study design to some extent (Thomas *et al.*, 1993), and non-differential random measurement error is difficult to prevent. Therefore, the highest priority in study design and the conduct of study is to eliminate differential and non-differential measurement errors.

1.6 Why is a reliability study needed?

Measurement error can cause serious adverse effects on the results of a study. Thus, accurately assessing the amount of measurement error and then correcting error are important in an epidemiologic study.

Different from systematic error, which is a primary issue of validity that concerns how accurately a measurement represents the truth, measurement error is the basic concern of reliability, which focuses on the extent to which a test yields the same results across repeated measurements (Carmines and Zeller, 1979, p.11). In other words, a measure can be trusted if it is stable over time and consistent between observers. Consequently, a reliability study, which estimates the consistency of measurements,

is needed to obtain the measurement error information that can be used to adjust the measure for the effect of measurement error.

Reliability studies include external reliability study, which means the replication is conducted on an independent set of subjects at the same time with the main study, and internal reliability study, which means the replication is carried out on a subset of subjects of the main study. In this thesis, measurement error is estimated by an external reliability study. The details of reliability studies will be discussed in section (2.4) and section (3.1.2).

1.7 The objective of the thesis

Epidemiologists have recognized that the statistical analysis of epidemiological data must take the measurement error into account (Liu *et al.*, 1978; Ferrari *et al.*, 2007; Thiebaut *et al.*, 2008). The awareness that measurement error has been among the major weakness of epidemiologic studies stimulates methodological researches to correct the error. A number of statistical methods have been proposed for correcting measurement error (Fuller, 1987; Spiegelman *et al.*, 2005; Cole *et al.*, 2006).

In the studies of the area under the ROC curve, researchers have also developed methods to adjust the estimate and confidence interval due to measurement error.

Schisterman *et al.* (2001) presented a corrected estimate and confidence interval for A based on the Delta method, assuming the measurement error based on an external reliability study is normally distributed.

Since a confidence interval for A can be obtained through that for a ratio of mean difference to its standard deviation, we can first obtain the corresponding confidence limits, and then apply the MOVER to obtain the confidence interval for A. This is the foundation for the method proposed in this thesis.

This thesis will compare the Delta method and MOVER approach to construct a confidence interval for A, taking the measurement error into account. The methods

will be compared in terms of coverage probability, interval width and the symmetry of tail errors.

1.8 Organization of the thesis

The rest of the thesis is organized as follows. Chapter 2 presents the literature review for the development of the ROC curve and the area under the ROC curve. Chapter 3 describes the theoretical background of the Delta method and the MOVER approach and shows how to apply these two methods to construct an approximate confidence interval for the area under the ROC curve in the presence of random measurement error. Chapter 4 shows the results and conclusions of simulation studies for comparing the performances between the Delta method and the MOVER approach. An example of Thiobarbituric acid reaction substances (TBARS) biomarker is presented in Chapter 5, and the thesis is concluded with a discussion.

Chapter 2

LITERATURE REVIEW

2.1 A brief history of the ROC curve

The development of the ROC curve was based on signal decision theory (Wald, 1950). It can be traced back to the work done in electronic communications in the early 1940s and first appeared in the literature in the early 1950's (Van Meter and Middleton, 1954; Peterson *et al.*, 1954). It was Peterson and Birdsall who showed how to plot the data to get an ROC curve when applying statistical decision theory to radar detection problems, in which it was necessary for an observer to distinguish a signal plus noise from noise alone.

By the mid 1960's, the ROC curve had been widely used in visual and auditory experiments in psychology and psychophysics. Tanner and Swets (1954) studied the human observer's behavior in detecting light signals and briefly presented a new theory in the visual detection, assuming the false-alarm rate and correct detection varied together, as well as that the neural activity was a monotonically increasing function of light intensity, not necessarily linear. Based on Tanner and Swets' work, Swets *et al.* (1961) described the theory of ROC analysis adequately by four more vision experiments in addition to the first experiment described in Tanner and Swets (1954). The ROC curve was also used in the later work of Swets (1964, 1973) and Green and Swets (1966) on the detection and recognition of auditory and visual signals.

Lusted (1968), a radiologist, considered that the ROC curve for medical diagnosis had the same meaning as it did in signal detection studies, and then introduced ROC analysis into medical decision making. Lusted (1971*a*,*b*) applied signal detection

theory, of which the essential feature was the ROC curve, to evaluate the performance of criteria for radiologists' assistants and radiologic systems. From then on, ROC analysis began to be widely used in medical imaging and other medical diagnosis. Goodenough *et al.* (1974) used the ROC curve to described the detectability of the image of some tiny beads in a noisy background of a radiographic mottle. Metz (1986*a*) showed that ROC analysis was the most meaningful approach after comparing the advantages and limitations of various traditional techniques and ROC analysis to assess the diagnostic performance. Also, using the ROC curve, Han and Kim (1998) evaluated the diagnostic ability of several cephalometric measurements in determining the presence of two different class of skeletal patterns by ROC curve, and Boquete *et al.* (2003) assessed the diagnostic accuracy of an insulin-like growth factor and a binding protein in growth hormone-deficient children and adults. Today, ROC analysis plays a crucial role in the field of medical diagnosis.

2.2 The indices of accuracy based on the ROC curve

Several indices related to the ROC curve have been used to evaluate the accuracy of a test.

2.2.1 A two-parameter index

A two-parameter index termed $D(\Delta m, s)$ is used in the situation that the "abnormal" and "normal" individuals follow normal distributions with unequal variances. Δm denotes the difference between the means of two normal distributions with unequal variance. Its value is equal to the absolute normal-deviate value of false positive rate, $z(D|n)$, at the intercept point on the ROC curve where the normal-deviate value of true positive rate, $z(D|d)$, is zero. Usually, Δm is presented along with s, the slope of the ROC curve at the intercept point, as the two-parameter index termed $D(\Delta m, s)$ to give the information of the entire ROC curve.

2.2.2 Youden index

Youden index, J, which was first presented by Youden (1950), is a global measure of overall diagnostic effectiveness. It is a function of sensitivity and specificity. Faraggi (2000) and Reiser (2000) pointed out that it was the optimal cutoff point that maximizes the discriminating ability of a diagnostic marker when equal weight was given to sensitivity and specificity. Graphically, J is the maximum vertical distance between the ROC curve and the diagonal chance line. Its value ranges from 0.0, indicating an ineffective test, to 1.0, meaning a perfect effective test.

2.2.3 Likelihood ratio

Different from the "likelihood" used in statistical inference, a likelihood ratio here is defined as the ratio of the probability of a particular test result among individuals with the disease of interest to the probability of the same test result among individuals without the disease of interest (Fletcher *et al.*, 1988). In other words, likelihood ratio is the ratio of sensitivity (i.e. true-positive fraction) to (1 - specificity) (i.e. false-positive fraction). Actually, it is the slope of a ROC curve at a given cutoff point. It describes how many times more (or less) likely a particular test result can be found in "abnormal" individuals, compared to "normal" individuals, and summarizes the information of sensitivity and specificity. The disadvantage of using likelihood ratio is that it is an odds rather than a probability. Also, the report of the likelihood ratio must be accompanied by a cutoff point.

2.2.4 The area under the ROC curve

The area under the ROC curve, once was called A_z, is defined as the proportion of the total area of the ROC graph that lies under an ROC curve. This index is recommended for describing the ROC curve (Wolfe and Hogg, 1971; Swets and Pickett, 1982) because it is less affected by the location of the point on the curve and thus independent of

cutoff point effect.

Green and Swets (1966) explained the meaning of A in terms of the result of a signal detection experiment, in which A corresponds to the probability of correctly identifying "signal plus noise" from "noise" among two stimuli. Also, suppose Y is the measured value of a biomarker on a randomly chosen "abnormal" subject, and X is the value of the same marker measured on a randomly chosen "normal" subject, Bamber (1975) pointed out that $A = P(Y > X)$. Based on Bamber's result, Hanley and McNeil (1982) interpreted A as the probability that a randomly selected subject in "abnormal" group will have a higher test value than that of a randomly selected subject in "normal" group.

2.3 Confidence intervals for the area under the ROC curve

Usually, diagnostic test results can be measured in dichotomous (e.g. disease positive or negative), ordinal (e.g. the presence of disease - definitely, probably, possibly, probably not, definitely not) and continuous (e.g. blood pressure) scales. According to the types of diagnostic test outcomes, many approaches proposed in literatures for estimating A can be classified as parametric, non-parametric and semi-parametric approaches, respectively.

2.3.1 Parametric approach

A simple parametric approach for estimating A is to assume that X and Y, the diagnostic test measurements on the "normal" and "abnormal" subjects, respectively, are independent and distributed as classic models, such as normal, log-normal, exponential (Green and Swets, 1966) and gamma (Pham and Almhana, 1996) etc. The most popular model is the binormal model, which was developed by Dorfman and Alf (1969). It assumes that the test results of "abnormal" and "normal" subjects follow normal distributions with different means and variances. Dorfman and Alf (1969) and

Grey and Morgan (1972) applied maximum likelihood method to obtain the variance-covariance matrix and confidence interval for the binormal model. In this model, two parameters can be used to describe the ROC curve. Let a represents the "y-intercept" and b represents the "slope" of a particular ROC when the ROC curve is plotted as straight lines on "normal-deviate" axes. In terms of a pair of normal distributions, a stands for the difference between the conditional means, relative to the standard deviation of the distribution of the actually abnormal subjects, and b stands for the ratio of the standard deviations of the distributions for the actually normal subjects to the actually abnormal subjects (Metz, 1986*b*).

Then A can be expressed as

$$A = \Phi(\delta), \qquad \delta = \frac{a}{\sqrt{1+b^2}} = \frac{\mu_Y - \mu_X}{\sqrt{\sigma_Y^2 + \sigma_X^2}},$$

where Φ is the standard normal cumulative distribution function.

Using the Delta method (Rao, 1973), Wieand *et al.* (1989) derived the variance of δ, given by

$$\mathrm{var}(\widehat{\delta}) = \frac{1}{\widehat{\sigma}_X^2 + \widehat{\sigma}_Y^2}\left(\frac{\widehat{\sigma}_X^2}{n_X} + \frac{\widehat{\sigma}_Y^2}{n_Y}\right) + \frac{\widehat{\delta}^2}{2(\widehat{\sigma}_X^2 + \widehat{\sigma}_Y^2)^2}\left(\frac{\widehat{\sigma}_X^4}{n_X - 1} + \frac{\widehat{\sigma}_Y^4}{n_Y - 1}\right).$$

The advantage of the parametric approach is that it can create a smooth ROC curve that is defined by a small number of parameters, from which the statistical inferences are derived. However, this approach for the estimation of A is sensitive to the departure from assumptions. In practice, the assumptions in the parametric approach may not be completely satisfied (Goddard and Hinberg, 1990). In such a case, non-parametric methods that estimate the confidence interval of A without making distributional assumption are preferred.

2.3.2 Non-parametric approach

Non-parametric procedures are the appropriate methods to estimate A based on continuous test results in that they need no assumption about the distributions of the

"abnormal" and "normal" subjects, and no parameter is needed to model the ROC curve. Non-parametric methods use the empirical ROC curve, which asymptotic properties have been derived by Hsieh and Turnbull (1996), to evaluate A.

The Mann-Whitney-Wilcoxon statistic is one of the popular non-parametric interval estimators for A. When estimated by the trapezoidal rule, Bamber (1975) showed that A is linked to the Mann-Whitney U statistic, which is used for comparing distributions of test results from two samples. Based on the variance equation of Noether (1967), Bamber also derived a formula for the variance of the Mann-Whitney U for A. Hanley and McNeil (1982) indicated that A measured the same quantity that was estimated by the Wilcoxon statistic.

Kernel density estimation is another non-parametric method for the estimation of A. It calculates A based on the ranks of the observations in the combined sample and, therefore, is free of distributional assumptions (Campbell, 1994). Zou *et al.* (1997) showed that a smoothed ROC curve can be obtained using kernel density estimation. Lloyd and Yong (1999) recommended the kernel-based ROC estimator rather than the fully empirical estimator in terms of the asymptotic accuracy. This is consistent with the results of Hsieh and Turnbull (1996) who studied kernel-based and empirical-based estimators of the Youden index.

Recently, Zhou *et al.* (2005) estimated A based on ordinal-scale tests measurements without a gold standard by a non-parametric maximum likelihood method. Qin and Zhou (2006) proposed an alternative empirical likelihood based method for the confidence interval of A. They indicated that the empirical likelihood method for inference on A has better coverage accuracy than other nonparametric methods when A is close to 1.0. When tests are carried out on the same individuals, DeLong *et al.* (1988) proposed a non-parametric method to the calculate confidence interval of A based on the theory on generalized U-statistics. Also, the Bootstrap method can be applied to obtain a confidence interval for A (Efron and Tibshirani, 1993; Qin and Hotilovac, 2008).

2.3.3 Semi-parametric approach

The semi-parametric approach is an intermediate strategy between the parametric and non-parametric approaches. Metz *et al.* (1998) presented a semi-parametric algorithm, $LABROC4$, to estimate A by assuming that the underlying distributions of test results could be transformed to normal distributions by an unspecified monotone transformation, and then applying the maximum likelihood estimation to estimate A in the same way that a parametric approach does. In fact, the characteristic of a semi-parametric approach is that the transformation is unspecified and non-parametric, but the model is parametric after transformation (Zou *et al.*, 1997).

Recently, some new semi-parametric methods were proposed. Erkanli *et al.* (2006) describes a semi-parametric Bayesian approach for estimating ROC curves. The paper showed that this Bayesian estimation was similar to the kernel density estimation approach. Wan and Zhang (2007) proposed a new smooth semi-parametric ROC curve estimator based on a semi-parametric kernel distribution function estimator. They pointed out that the proposed estimators were more efficient than the traditional non-parametric kernel distribution estimators, as well as the non-parametric estimators proposed by Zou *et al.* (1997) and Lloyd (1998) in terms of asymptotic bias and variance of the proposed estimators.

2.4 The assessment of reliability

Reliability is the extent to which measurements remain the same over repeated tests on the same subjects. It can be assessed by the test-retest method, alternative-form method, split-halves method, internal consistency method (Carmines and Zeller, 1979).

2.4.1 Test-retest method

The test-retest method is the easiest way to assess the reliability of measurements. It gives the same test in the same way to the same subject at two points in time. The reliability coefficient is given by the correlation between two tests, $\rho_{x_1 x_2}$, where x_1 and x_2 are the two tests.

The straightforward and intuitively appealing procedure by which to assess reliability is the advantage of this method. The disadvantage is that effects of memory, learning and reactivity can confound reliability assessment.

2.4.2 Alternative-form method

The alternative-form method with refined test-retest technique involves similar, but not the same tests on two testing situations with the same subject. Of the two tests, the second test is an alternative form of the the first test without any systematic difference (i.e. with equal observed means and variance). The alternative-form method is widely used in assessing the reliability of all types of educational tests.

The alternative-form method is superior to the test-retest method because it minimizes the effects of memory. However, the disadvantage is that the alternative-form method has difficulty of developing an alternative form that is parallel to the first test, apart from sharing the other limitations with the test-retest method.

2.4.3 Split-halves method

The Split-halves method can simultaneously conduct two alternative forms of a test, which is typically divided into halves. Correlations between each half-test are determined. The reliability coefficient of the total test can be calculated by the Spearman-Brown prophecy formula (Spearman, 1910; Brown, 1910),

$$\rho_t = \frac{2\rho_{hh}}{1 + \rho_{hh}},$$

where ρ_t is the reliability coefficient for the total test, ρ_{hh} is the two half-tests correlation.

The advantage of this method is that it can be conducted on a single test administration, overcoming the limitations of the test-retest method and the alternative-form method. The disadvantage is that different reliability estimates will be obtained due to different possible splits.

2.4.4 Internal consistency method

Rather than calculating the reliability between arbitrary half-tests, the internal consistency method involves a single test administration and gives a unique reliability coefficient that is equivalent to the average of the correlations between any possible pair of items. The most popular reliability estimate to give the measure of internal consistency is Cronbach's α (Cronbach, 1951), which is defined as:

$$\alpha = \frac{N}{N-1}\left[1 - \frac{\sum \sigma^2(Y_i)}{\sigma_x^2}\right],$$

where N is the number of test items, $\sum \sigma^2(Y_i)$ is the sum of item variances, and σ_x^2 is the total variance.

A special case of Cronbach's α, Kuder-Richardson formula number 20 (KR20) (Kuder and Richardson, 1937), can be used to calculate the reliability of scales composed of dichotomously variables as follow:

$$KR20 = \frac{N}{N-1}\left[1 - \frac{\sum p_i q_i}{\sigma_x^2}\right],$$

where N is the number of dichotomous items, p_i is the proportion responding correctly to the i^{th} item, q_i = 1 - p_i, and σ_x^2 is the total variance.

The advantage of this method is that it provides a unique reliability coefficient on a single test administration without splitting. The disadvantage is that it involves more complex computations. Also, it requires equal expected means, observed variance and correlations between test items.

2.5 *Confidence intervals for the area under the ROC curve in the presence of measurement error*

In the studies of ROC curve, Coffin and Sukhatme (1996, 1997) studied the effects of measurement error on the parametric and nonparametric estimates of A and showed that ignoring measurement error could result in biased estimation of A. Faraggi (2000) considered the effect of neglecting measurement error on the confidence interval for A, based on a normal model, and pointed out that not taking measurement error into account could give seriously spurious results that understated the diagnostic efficacy of a biomarker. Reiser (2000) developed an adjusted confidence interval on A, taking measurement error into account, based on the information of measurement error offered by an internal study. Schisterman *et al.* (2001) also suggested a random measurement error correction method, based on an external study, for the estimator and confidence interval of A based on the Delta method. The authors showed that the result was affected by the correction for random measurement error. When measurement error is ignored, the effectiveness of the biomarker could be seriously understated. Tosteson *et al.* (2005) studied the effects of heterogenous measurement error on binormal ROC curves and corrected the estimators and confidence intervals for specific points on the curve by assuming that the measurement error is non-normally distributed.

Chapter 3

CONFIDENCE INTERVALS FOR THE AREA UNDER THE ROC CURVE WITH MEASUREMENT ERROR

3.1 The Delta Method

3.1.1 The estimation of the area under the ROC curve without measurement error

Suppose X_i and Y_i represent the values of a biomarker in the "normal" and "abnormal" subjects, respectively, and follow normal distributions. That is, $X_i \sim N(\mu_X, \sigma_X^2)$, $(i = 1,...,n_X)$, $Y_i \sim N(\mu_Y, \sigma_Y^2)$, $(i = 1,...,n_Y)$. When there is no random measurement error, the area under the ROC curve can be written as

$$A = \Phi(\delta), \qquad \delta = \frac{\mu_Y - \mu_X}{\sqrt{\sigma_Y^2 + \sigma_X^2}},$$

where Φ is the standard normal cumulative distribution function.

The above function shows that A is a monotonic function of δ. Therefore, to evaluate A is equivalent to calculate δ first and then obtain the standard normal cumulative distribution function of δ, which actually is the ratio of mean difference to the square root of the sum variance of Y and X. A can be estimated by substituting sample mean difference and variances into the above formula to get

$$\widehat{A} = \Phi\left(\frac{\overline{y} - \overline{x}}{\sqrt{S_y^2 + S_x^2}}\right),$$

where $\overline{x}$, $\overline{y}$, S_x^2, and S_y^2 denote the sample means and variances for the "normal" and "abnormal" population, respectively.

3.1.2 External reliability study

Schisterman *et al.* (2001) considered the case where an external reliability study was conducted to estimate measurement error. This external reliability study is independent of the main study, in which the variations of the observed values of the biomarker, $\widehat{\sigma}_x$ and $\widehat{\sigma}_y$, can be estimated.

Suppose, in the main study, the observed values of a biomarker on the "normal" and "abnormal" subjects are

$$x_i = X_i + \varepsilon_i^x \qquad i = 1, ..., n_x \tag{3.1}$$

$$y_i = Y_i + \varepsilon_i^y \qquad i = 1, ..., n_y, \tag{3.2}$$

where x_i and y_i are the observed values of the biomarker, X_i and Y_i are the true values of the biomarker, ε_i^x and ε_i^y are the measurement errors of X_i and Y_i, respectively, $\varepsilon_i^x \sim N(0, \sigma_\varepsilon^2)$ and $\varepsilon_i^y \sim N(0, \sigma_\varepsilon^2)$.

Assume that X_i, Y_i, ε_i^x, and ε_i^y are independent, and the variations that result in measurement error do not rely on the person's risk status or the true value of the biomarker. Therefore, the variances of ε_i^x and ε_i^y can be assumed to be equal. They can be estimated by an external reliability study.

Let w_{ij} denotes the jth observed value of the biomarker on the ith subject in an external reliability study.

$$w_{ij} = W_i + \varepsilon_{ij} \qquad i = 1, ..., n_0, \quad j = 1, ...p_i,$$

where W_i is the "true" value of the biomarker for the ith subject, and $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$.

Further let

$$n_f = \sum_{i=1}^{n_0} p_i - 1 \qquad \overline{w}_{i.} = \sum_{j=1}^{p_i} \frac{w_{ij}}{p_i},$$

where n_f reflects the degree of freedom for σ_ε^2.

The unbiased estimator of σ_ε^2 is

$$\widehat{\sigma}_\varepsilon^2 = \frac{\sum_{i=1}^{n_0}\sum_{j=1}^{p_i}(w_{ij} - \overline{w}_{i.})^2}{n_f}.$$

The sample variances for the observed x_i, y_i are

$$S_x^2 = \frac{\sum_{i=1}^{n_x}(x_i - \overline{x})^2}{n_x - 1}$$

$$S_y^2 = \frac{\sum_{i=1}^{n_y}(y_i - \overline{y})^2}{n_y - 1},$$

respectively, where $\overline{x}, \overline{y}$ are the sample means for x_i, y_i.

From equations (3.1), (3.2), we know

$$E(S_x^2) = \sigma_X^2 + \sigma_\varepsilon^2$$

$$E(S_y^2) = \sigma_Y^2 + \sigma_\varepsilon^2.$$

Thus, the unbiased estimators of σ_X^2 and σ_Y^2 are

$$\widehat{\sigma}_X^2 = S_x^2 - \widehat{\sigma}_\varepsilon^2$$

$$\widehat{\sigma}_Y^2 = S_y^2 - \widehat{\sigma}_\varepsilon^2.$$

The ratio of the true to observed variance is called the reliability of the observed value as a measure of the true value (Carmines and Zeller, 1979, p.31). The reliability can be measured by R, which is used as a "reliability index" in the social sciences. R reflects the amount of measurement error related to the inherent biomarker variability and can be written as

$$R = \frac{\sigma_X^2 + \sigma_Y^2}{\sigma_X^2 + \sigma_Y^2 + 2\sigma_\varepsilon^2}.$$

R ranges between 0 and 1. When R is close to 1, the measurement error is regarded as small and the reliability is high. If R close to 0 implies that the measurement error is relatively large and the reliability is small.

3.1.3 Area under the ROC curve in the presence of measurement error

To consider the random measurement error, A can be corrected as

$$A_c = \Phi(\delta), \qquad \delta = \frac{\mu_Y - \mu_X}{\sqrt{\sigma_X^2 + \sigma_Y^2 - 2\sigma_\varepsilon^2}}, \tag{3.3}$$

where Φ is the standard normal cumulative distribution function. Then the estimate of the corrected A is given by

$$\widehat{A}_c = \Phi(\widehat{\delta}), \qquad \widehat{\delta} = \frac{\overline{y} - \overline{x}}{\sqrt{S_x^2 + S_y^2 - 2\widehat{\sigma}_\varepsilon^2}},$$

where Φ is the standard normal cumulative distribution function. It is possible that $S_x^2 < \widehat{\sigma}_\varepsilon^2$ or $S_y^2 < \widehat{\sigma}_\varepsilon^2$, which gives $\widehat{\sigma}_X^2 (= S_x^2 - \widehat{\sigma}_\varepsilon^2) < 0$ or $\widehat{\sigma}_Y^2 (= S_y^2 - \widehat{\sigma}_\varepsilon^2) < 0$, resulting in an undefined $\sqrt{S_x^2 + S_y^2 - 2\widehat{\sigma}_\varepsilon^2}$. In such case, the negative variance estimate is replaced by a very small number (Rao, 1997).

The denominator of A_c is smaller than the denominator of A so that the area is larger if random measurement error is taken into account. In other words, if ignoring measurement error, A will be underestimated, and the ability of the biomarker to distinguish between "abnormal" subjects and "normal" subjects will be falsely regarded as less than what it really is.

3.1.4 Variance estimate by the Delta method

If random measurement error exists, the variance of $\widehat{\delta}$ can be estimated by the Delta method (see Appendix A) and written as the following:

$$\begin{aligned} \widehat{\mathrm{var}}(\widehat{\delta}) \;=\; & \left(\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}\right)(S_x^2 + S_y^2 - 2\widehat{\sigma}_\varepsilon^2)^{-1} + \\ & \frac{(\overline{y} - \overline{x})^2}{4(S_x^2 + S_y^2 - 2\widehat{\sigma}_\varepsilon^2)^3}\left(\frac{2S_x^4}{n_x - 1} + \frac{2S_y^4}{n_y - 1} + \frac{8\widehat{\sigma}_\varepsilon^4}{n_f}\right). \end{aligned}$$

3.1.5 *Confidence intervals for the area under the ROC curve in the presence of measurement error*

The approximate (1 - α)$\times$ 100 % confidence interval for δ is

$$\widehat{\delta} \pm z_{\alpha/2}\sqrt{\widehat{\text{var}}(\widehat{\delta})},$$

where $z_{\alpha/2}$ is the $\alpha/2$th upper quantile of the standard normal distribution. The corresponding confidence interval for A_c is

$$\left\{\Phi\left(\widehat{\delta} - z_{\alpha/2}\sqrt{\widehat{\text{var}}(\widehat{\delta})}\right), \Phi\left(\widehat{\delta} + z_{\alpha/2}\sqrt{\widehat{\text{var}}(\widehat{\delta})}\right)\right\}.$$

3.2 The MOVER Approach

3.2.1 *Traditional confidence interval approach*

The confidence interval constructed by the Delta method is a traditional two-sided confidence interval like the Wald interval, which is obtained by assuming the symmetric distribution of an estimate. For example, suppose that θ_i, $i = 1, 2$, are parameters of interest, and $\widehat{\theta}_i$, $i = 1, 2$, are independently distributed estimates. The endpoints for a $(1 - \alpha)100\%$ two-sided traditional confidence interval (l, u) for a sum, $\theta_1 + \theta_2$, are given by

$$(l, u) = \widehat{\theta}_1 + \widehat{\theta}_2 \mp z_{\alpha/2}\sqrt{\widehat{\text{var}}(\widehat{\theta}_1) + \widehat{\text{var}}(\widehat{\theta}_2)},$$

where $\widehat{\theta}_i$ and $\widehat{\text{var}}(\widehat{\theta}_i)$, $i = 1, 2$, are the estimates of θ_i and $\text{var}(\theta_i)$, $i = 1, 2$, respectively, and $z_{\alpha/2}$ is the $\alpha/2$th upper quantile of the standard normal distribution. Note that variances on the lower and upper limits of $\theta_1 + \theta_2$ are equal.

However, this confidence interval may perform poorly unless sample sizes are large, or the sampling data of $\widehat{\theta}_i$ follow normal distributions. This occurs because the

sampling distributions of θ_i may not be normally distributed when the sample sizes are small or moderate. In other words, the variances at the endpoints of a confidence interval for a parameter may not be equal. However, they are forced to be equal when using the traditional method to construct a confidence interval, leading to a poor coverage confidence interval. Hence, it is more reasonable to get the variances near the confidence limit boundaries rather than estimating the variance at the maximum likelihood estimate of the parameter.

3.2.2 The MOVER approach

The MOVER approach recovers the variance estimates in the neighborhood of the lower and upper confidence limits of a parameter. It only requires the reliable confidence limits, which have coverage probabilities close to the nominal, of the parameters involving in the estimate.

Suppose θ_1 and θ_2 are two parameters of interest with confidence limits (l_1, u_1) and (l_2, u_2), respectively. λ is the ratio of θ_1 to θ_2 (i.e. $\lambda = \theta_1/\theta_2$), which can be rewritten as $\theta_1 + (-\lambda\theta_2) = 0$. This suggests that the confidence interval of λ can be converted from the confidence interval for a sum, $\theta_1 + (-\lambda\theta_2)$, since the plausible value for λ can be obtained if θ_1 and θ_2 also have their plausible values within their confidence limits and satisfy the equations above. Hence, to obtain the confidence interval for λ, first construct the confidence interval for a simple sum, $\theta_1 + \theta_2$, then extend it to the one for $\theta_1 + (-\lambda\theta_2)$ and finally convert it back to the confidence interval for λ.

Let l and u be the minimum and maximum values of $\theta_1 + \theta_2$. By the central limit theorem, l and u are given by

$$l = \widehat{\theta}_1 + \widehat{\theta}_2 - z_{\alpha/2}\sqrt{\text{var}(\widehat{\theta}_1) + \text{var}(\widehat{\theta}_2)} \tag{3.4}$$

$$u = \widehat{\theta}_1 + \widehat{\theta}_2 + z_{\alpha/2}\sqrt{\text{var}(\widehat{\theta}_1) + \text{var}(\widehat{\theta}_2)}. \tag{3.5}$$

To reflect the asymmetry of the distribution for $\widehat{\theta}_1 + \widehat{\theta}_2$, the estimate of $\text{var}(\widehat{\theta}_i)$ can

be recovered by substituting the possible true values at the neighborhood of the confidence limits l and u instead of using the point estimates, because the possible true value $l_1 + l_2$ is closer to l, and the possible true value $u_1 + u_2$ is closer to u than the point estimate $\widehat{\theta}_1 + \widehat{\theta}_2$ is, respectively.

By the central limit theorem, $l_i, i = 1, 2$, is given by

$$l_i = \widehat{\theta}_i - z_{\alpha/2}\sqrt{\widehat{\mathrm{var}}(\widehat{\theta}_i)};$$

then $\widehat{\mathrm{var}}(\widehat{\theta}_i)$ can be recovered at $\theta_i = l_i$ as

$$\widehat{\mathrm{var}}(\widehat{\theta}_i)_{l_i} = (\widehat{\theta}_i - l_i)^2/z^2_{\alpha/2}. \tag{3.6}$$

Replaced the $\widehat{\mathrm{var}}(\widehat{\theta}_i)$ in (3.4) with the recovered $\widehat{\mathrm{var}}(\widehat{\theta}_i)_{l_i}$ in (3.6), l is obtained as

$$\begin{aligned} l &= \widehat{\theta}_1 + \widehat{\theta}_2 - z_{\alpha/2}\sqrt{\widehat{\mathrm{var}}(\widehat{\theta}_1)_{l_1} + \widehat{\mathrm{var}}(\widehat{\theta}_2)_{l_2}} \\ &= \widehat{\theta}_1 + \widehat{\theta}_2 - z_{\alpha/2}\sqrt{\frac{(\widehat{\theta}_1 - l_1)^2}{z^2_{\alpha/2}} + \frac{(\widehat{\theta}_2 - l_2)^2}{z^2_{\alpha/2}}} \\ &= \widehat{\theta}_1 + \widehat{\theta}_2 - \sqrt{(\widehat{\theta}_1 - l_1)^2 + (\widehat{\theta}_2 - l_2)^2}. \end{aligned} \tag{3.7}$$

Similarly, $\widehat{\mathrm{var}}(\widehat{\theta}_i)$ in (3.5) can be recovered at $\theta_i = u_i$ as

$$\widehat{\mathrm{var}}(\widehat{\theta}_i)_{u_i} = (u_i - \widehat{\theta}_i)^2/z^2_{\alpha/2}.$$

Then the upper limit in (3.5) is obtained as

$$u = \widehat{\theta}_1 + \widehat{\theta}_2 + \sqrt{(u_1 - \widehat{\theta}_1)^2 + (u_2 - \widehat{\theta}_2)^2}. \tag{3.8}$$

For the confidence interval of $\theta_1 + (-\lambda)\theta_2$, just replace the confidence limits (l_2, u_2) in (3.7) and (3.8) with the confidence limits of $-\lambda\theta_2$, which are $(-\lambda u_2, -\lambda l_2)$, and get

$$\begin{aligned} l &= \widehat{\theta}_1 - \lambda\widehat{\theta}_2 - \sqrt{(\widehat{\theta}_1 - l_1)^2 + \lambda^2(u_2 - \widehat{\theta}_2)^2} \\ u &= \widehat{\theta}_1 - \lambda\widehat{\theta}_2 + \sqrt{(u_1 - \widehat{\theta}_1)^2 + \lambda^2(\widehat{\theta}_2 - l_2)^2}. \end{aligned}$$

For the lower limit of λ (i.e. L_λ), since

$$Pr\left(\frac{\theta_1}{\theta_2} < L_\lambda\right) = \frac{\alpha}{2} \qquad \Longrightarrow \qquad Pr\left(\theta_1 - L_\lambda\theta_2 < 0\right) = \frac{\alpha}{2},$$

a possible minimum value of $\theta_1 + (-L_\lambda)\theta_2$ is

$$l = \widehat{\theta}_1 - L_\lambda\widehat{\theta}_2 - \sqrt{(\widehat{\theta}_1 - l_1)^2 + L_\lambda^2(u_2 - \widehat{\theta}_2)^2}.$$

Let $l = 0$ to solve for L_λ and get

$$L_\lambda = \frac{\hat{\theta}_1\hat{\theta}_2 - \sqrt{\hat{\theta}_1^2\hat{\theta}_2^2 - (2u_2\hat{\theta}_2 - u_2^2)(2l_1\hat{\theta}_1 - l_1^2)}}{2u_2\hat{\theta}_2 - u_2^2}. \tag{3.9}$$

Similarly, for the upper limit of λ (i.e. U_λ), since

$$Pr\left(\frac{\theta_1}{\theta_2} > U_\lambda\right) = \frac{\alpha}{2} \qquad \Longrightarrow \qquad Pr\left(\theta_1 - U_\lambda\theta_2 > 0\right) = \frac{\alpha}{2},$$

a possible maximum value of $\theta_1 + (-U_\lambda)\theta_2$ is

$$u = \widehat{\theta}_1 - U_\lambda\widehat{\theta}_2 + \sqrt{(u_1 - \widehat{\theta}_1)^2 + U_\lambda^2(\widehat{\theta}_2 - l_2)^2}.$$

Let $u = 0$ to solve for U_λ and get

$$U_\lambda = \frac{\hat{\theta}_1\hat{\theta}_2 + \sqrt{\hat{\theta}_1^2\hat{\theta}_2^2 - (2l_2\hat{\theta}_2 - l_2^2)(2u_1\hat{\theta}_1 - u_1^2)}}{2l_2\hat{\theta}_2 - l_2^2}. \tag{3.10}$$

In the process of deriving the confidence limits of λ, the MOVER approach uses the central limit theorem to recover the variance estimates of $\theta_i, i = 1, 2$, rather than imposing symmetry on the confidence interval as the Delta method does.

3.2.3 Confidence intervals for the area under the ROC curve in the presence of measurement error

Suppose the confidence limits of each parameter in equation (3.3) are available. The numerator and the denominator of δ can be denoted as θ_1 (i.e. $\mu_Y - \mu_X$) and θ_2

(i.e. $\sqrt{\sigma_X^2 + \sigma_Y^2 - 2\sigma_\varepsilon^2}$), respectively. Then the problem of constructing a confidence interval for δ is reduced to that of θ_1/θ_2.

To use the MOVER approach to estimate the confidence interval of δ, taking random measurement error into account, first apply equations (3.7) and (3.8) to get the confidence limits of θ_1 and θ_2, respectively, then use equations (3.9) and (3.10) to obtain confidence limits for the ratio of θ_1 to θ_2.

For the confidence limits of θ_1, substitute the confidence limits of μ_Y and μ_X, which are (l_Y, u_Y) and (l_X, u_X), respectively, for (l_1, u_1) and (l_2, u_2) in equations (3.7) and (3.8), then the confidence limits of θ_1 are given by

$$l_1 = \overline{y} - \overline{x} - \sqrt{(\overline{y} - l_Y)^2 + (u_X - \overline{x})^2}$$

$$u_1 = \overline{y} - \overline{x} + \sqrt{(u_Y - \overline{y})^2 + (\overline{x} - l_X)^2}.$$

For the confidence limits of θ_2, first, substitute the confidence limits of σ_X^2 and σ_Y^2, which are $(l_{\sigma_X^2}, u_{\sigma_X^2})$ and $(l_{\sigma_Y^2}, u_{\sigma_Y^2})$, respectively, for (l_1, u_1) and (l_2, u_2) in equations (3.7) and (3.8) to obtain the confidence limits of $\sigma_X^2 + \sigma_Y^2$ as

$$l_{\sigma_X^2+\sigma_Y^2} = \widehat{\sigma}_x^2 + \widehat{\sigma}_y^2 - \sqrt{(\widehat{\sigma}_x^2 - l_{\sigma_X^2})^2 + (\widehat{\sigma}_y^2 - l_{\sigma_Y^2})^2}$$

$$u_{\sigma_X^2+\sigma_Y^2} = \widehat{\sigma}_x^2 + \widehat{\sigma}_y^2 + \sqrt{(u_{\sigma_X^2} - \widehat{\sigma}_x^2)^2 + (u_{\sigma_Y^2} - \widehat{\sigma}_y^2)^2}.$$

Second, substitute the confidence limits of $\sigma_X^2 + \sigma_Y^2$ and σ_ε^2, which are $(l_{\sigma_X^2+\sigma_Y^2}, u_{\sigma_X^2+\sigma_Y^2})$ and $(l_{\sigma_\varepsilon^2}, u_{\sigma_\varepsilon^2})$, respectively, for (l_1, u_1) and (l_2, u_2) in equations (3.7) and (3.8) again, then the confidence limits of $[(\sigma_X^2 + \sigma_Y^2) - 2\sigma_\varepsilon^2]$ are given by

$$l_{(\sigma_X^2+\sigma_Y^2)-2\sigma_\varepsilon^2} = \widehat{\sigma}_x^2 + \widehat{\sigma}_y^2 - 2\hat{\sigma}_\varepsilon^2 - \sqrt{(\widehat{\sigma}_x^2 + \widehat{\sigma}_y^2 - l_{\sigma_X^2+\sigma_Y^2})^2 + (2u_{\sigma_\varepsilon^2} - 2\hat{\sigma}_\varepsilon^2)^2}$$

$$u_{(\sigma_X^2+\sigma_Y^2)-2\sigma_\varepsilon^2} = \widehat{\sigma}_x^2 + \widehat{\sigma}_y^2 - 2\hat{\sigma}_\varepsilon^2 + \sqrt{(u_{\sigma_X^2+\sigma_Y^2} - (\widehat{\sigma}_x^2 + \widehat{\sigma}_y^2))^2 + (2\hat{\sigma}_\varepsilon^2 - 2l_{\sigma_\varepsilon^2})^2}.$$

Since θ_2 is the square root of $(\sigma_X^2 + \sigma_Y^2 - 2\sigma_\varepsilon^2)$, its confidence limits are given by

$$\begin{aligned} l_2 &= \sqrt{l_{(\sigma_X^2+\sigma_Y^2)-2\sigma_\varepsilon^2}} \\ &= \sqrt{\widehat{\sigma}_x^2 + \widehat{\sigma}_y^2 - 2\hat{\sigma}_\varepsilon^2 - \sqrt{(\widehat{\sigma}_x^2 + \widehat{\sigma}_y^2 - l_{\sigma_X^2+\sigma_Y^2})^2 + (2u_{\sigma_\varepsilon^2} - 2\hat{\sigma}_\varepsilon^2)^2}} \end{aligned}$$

$$
\begin{aligned}
u_2 &= \sqrt{u_{(\sigma_X^2+\sigma_Y^2)-2\sigma_\varepsilon^2}} \\
&= \sqrt{\widehat{\sigma}_x^2 + \widehat{\sigma}_y^2 - 2\hat{\sigma}_\varepsilon^2 + \sqrt{(u_{\sigma_X^2+\sigma_Y^2} - (\widehat{\sigma}_x^2 + \widehat{\sigma}_y^2))^2 + (2\hat{\sigma}_\varepsilon^2 - 2l_{\sigma_\varepsilon^2})^2}}.
\end{aligned}
$$

For the confidence limits of δ, substitute the confidence limits of θ_1 and θ_2 derived above for the corresponding (l_1, u_1) and (l_2, u_2) in equations (3.9) and (3.10), then obtain the confidence limits of δ, (l_δ, u_δ).

To obtain the confidence interval of A_c, use the transformation defined as $(\Phi(l_\delta), \Phi(u_\delta))$.

Chapter 4

SIMULATION STUDY

4.1 Study design and data generation

To investigate the performance of the Delta method and the MOVER approach, a series of simulation studies, in which the data was generated with measurement error, were carried out.

4.1.1 Study design

4.1.1.1 Parameter selection

In the simulation studies, the parameters examined were sample size (n), the area under the ROC curve (A), the reliability index (R) and the variance of test values for the "abnormal" subjects (σ_Y^2). The sample sizes were selected based on the study of Schisterman *et al.* (2001). The sample size combination (n_X, n_Y, n_f) corresponds to sample sizes of "normal", "abnormal" subjects and the degree of freedom for the variance of measurement error, respectively. Values of the sample size combinations were small balanced (50, 50, 19), medium balanced (100, 100, 49), large balanced (1,000, 1,000, 199) and unbalanced (900, 50, 49) samples with measurement error. A was chosen so that the two procedures can be evaluated under different discriminative abilities of tests. Since A ranges from 0.5 to 1, the values of A were selected as 0.6, 0.7, 0.8, 0.9 and 0.95 to reflect a wide range of test accuracy in practice from low to high. The values of R were 0.2, 0.4, 0.6 and 0.8 to provide the determined amount of measurement error relative to the biomarker variability from small to large, and on the other hand, to represent the measurement error due to measurement process

from large to small. The variance of measurement error, σ_ε^2, was selected based on $\sigma_\varepsilon^2 = (\sigma_X^2 + \sigma_Y^2) \times \frac{(1-R)}{2R}$. The mean and variance of test values for the "normal" subjects were selected as $\mu_X = 0$ and $\sigma_X^2 = 1$. While the mean of test values for the "abnormal" subjects, μ_Y, was given by $\mu_Y = \Phi^{-1}(A) \times \sqrt{\sigma_X^2 + \sigma_Y^2} + \mu_X$, where the corresponding variance for the "abnormal" subjects, σ_Y^2, was chosen as 0.5, 1.0, 3.0 and 5.0, so that it was fairly close, equal, moderately close and the least close to σ_X^2. The nominal levels $\alpha = 0.05, 0.10$ were considered.

4.1.1.2 Method comparison

The two confidence interval construction procedures for A discussed in this thesis are regarded as parametric approaches, since the samples for "abnormal" and "normal" subjects used in the two approaches follow normal distributions. The performances of the two procedures were evaluated in terms of coverage probability, interval width and the symmetry of tail errors (i.e. non-coverage probabilities).

Here, the coverage probability is an estimate of the percentage that the true value of the parameter is contained in the range of the interval. It is expected that the calculated coverage probability is close to the nominal coverage 1 - α. For the 95% confidence interval with 10,000 simulations, it is preferable to have calculated coverage probability that falls in the range of 94.6-95.4 percent, which is $0.95 \pm 1.96\sqrt{\frac{0.05 \times 0.95}{10000}}$.

Also, the interval width is the range between the upper limit and the lower limit of the parameter. It is desirable to obtain the required coverage probability with the least width, which represents a more precise estimate of the parameter.

As for the symmetry of tail errors, the difference of the left and right tail errors is used as a proxy for the symmetry of tail errors. The left tail error is obtained by calculating the proportion that the true value is less than the lower limit, and the right tail error is the proportion that the true value is greater than the upper limit. If the confidence interval is constructed appropriately, then the overall two sided error should approximately equal to the nominal level α, and the miss coverage for each

side of the interval is required to be $\alpha/2$ (Efron and Tibshirani, 1993, p.156). Thus, for a 95% confidence interval, it is preferable to have left and right miss coverage equal to 2.5%.

In summary, for a good performance, a procedure is expected to have its coverage probability close to the nominal percent level with the least width and have symmetric tail errors. In this thesis, a 95% confidence interval constructed with approximately 95% coverage probability, approximately equal missing left and missing right coverage and a narrow width is preferred.

4.1.2 Data generation

The values of "normal" subjects, X, were generated from a normal distribution with mean $\mu_X = 0$ and variance $\sigma_X^2 + \sigma_\varepsilon^2$. Similarly, the values of "abnormal" subjects, Y, were obtained as normally distributed with mean μ_Y and variance $\sigma_Y^2 + \sigma_\varepsilon^2$. There were two ways to obtain the estimate of σ_ε^2. The first way was to calculate it by using w_{ij} and $\overline{w}_{i.}$, which were mentioned in the previous "External reliability study" section (section 3.1.2). Knowing the sampling variance of measurement error followed chi-square distribution, the second way was to obtain it directly by $\widehat{\sigma}_\varepsilon^2 \sim \frac{\sigma_\varepsilon^2}{n_f}\chi_{n_f}^2$, where $\chi_{n_f}^2$ was a chi-square variate with n_f degree of freedom. Since the calculation of $\frac{\sigma_\varepsilon^2}{n_f}\chi_{n_f}^2$ was simpler than the calculation of w_{ij} and $\overline{w}_{i.}$, the second way therefore was chosen to calculate the estimate of σ_ε^2 in the simulations for simplicity.

For each parameter combination, a total of 10,000 replicates were conducted. Confidence intervals were constructed at $\alpha = 0.05$ and 0.1 using both the Delta method and the MOVER approach as described in chapter 3. The coverage probability, interval width and the percentage of the true value of the parameter laid completely beyond the left (missing left) and right (missing right) sides of the interval were estimated to evaluate the performances of the two procedures.

4.2 Results

The simulation results are showed from Table 4.1 to Table 4.4, in which only results for $\alpha = 0.05$ and $\sigma_Y^2 = 0.5$ and 5.0 were presented, since the results for $\alpha = 0.1$ was similar to those for $\alpha = 0.05$, and the results for $\sigma_Y^2 = 1.0$ and 3.0 are similar to those for $\sigma_Y^2 = 0.5$ and 5.0.

4.2.1 For $n_X = n_Y = 50, n_f = 19$

Table 4.1 shows the simulation results for $n_X = n_Y = 50, n_f = 19$.

There is no obvious trend in the coverage probability for either approach as σ_Y^2 changes. For the Delta method, when $R = 0.2$, all of the estimated coverage probabilities for different values of A do not fall within the target range of coverage probability. Most of them are far away from the target range. Some coverage probabilities are even less than 90% when $A \geq 0.8$. This indicates that the coverage probability of a higher test accuracy is greatly affected when the measurement error is relatively large. For $R \geq 0.4$, few coverage probabilities fall inside the target range, while the majority still fail to fall inside the range. This implies that the coverage probability is not close to the nominal level when the sample size is relatively small for the Delta method. For the MOVER approach, some of the coverage probabilities fall inside the target range, and those fell outside the range are close to the target range.

The confidence interval widths for the two approaches are stable as the σ_Y^2 changes, but vary as A and R change. The interval widths in both approaches become narrower when R increases, which implies that the estimation of A is more precise when measurement error is relatively small. The interval width obtained by the MOVER approach is much narrower than its corresponding width in the Delta method when $R = 0.2$, while it is slightly wider when $R = 0.4$ and 0.6 if a test with lower accuracy is applied, but gets much narrower if a test with higher accuracy is applied. This implies that if a test with higher accuracy is applied, the MOVER

approach can give a more precise estimate than the Delta method when measurement error is relatively large. The interval widths obtained by both approaches are very close when $R = 0.8$.

As for the symmetry of tail errors, the difference between the left and right tail errors changes as A and R change. For the Delta method, it is seriously unbalanced for the left and right tail errors as A gets larger, especially when $R \leq 0.6$. For $R = 0.8$, the tail errors are approximately symmetric. This implies that the symmetry of tail errors of a confidence interval for a test with high accuracy is more sensitive to the measurement error in the Delta method. For a given A, the difference of tail errors decreases as R increases, which means the left and right tail errors approach symmetry if measurement error is relatively small. The difference of tail errors for the MOVER approach is different from those for the Delta method. The difference for the tail errors is getting smaller if a test with higher accuracy is applied. In summary, a more symmetric distribution of tail errors is obtained by the MOVER approach than by the Delta method when the test accuracy is high. Also, the tail errors are more symmetric when measurement error is small for both approaches.

4.2.2 For $n_X = n_Y = 100, n_f = 49$

Table 4.2 presents the simulation results for the combination of $n_X = n_Y = 100, n_f = 49$.

As with $n_X = n_Y = 50, n_f = 49$, there is no trend for the estimated coverage probabilities as σ_Y^2 changes for both approaches. For the Delta method, when $R = 0.2$, all of the coverage probabilities are still far away from the target range, and some of them are still less than 90%. Also, for $R = 0.4$ and 0.6, only one coverage probability falls inside the target range. Most of the others still fall beyond the range. When $R = 0.8$, some of the coverage probabilities are covered by the target range. This suggests that slightly increasing sample size has no obvious improvement on coverage probability when measurement error is large, but the coverage probability is improved

to some extent when measurement error is small. For the MOVER approach, most of the coverage probabilities are in the target range. Even though some are not included in the target range, but they are close to the range.

The interval width is similar to that for the combination of $n_X = n_Y = 50, n_f = 19$ for the Delta method and the MOVER approach. But the width is slightly narrower than the corresponding interval in the small sample combination, which means slightly increasing sample size can somewhat improve the precision of the estimate.

The difference of tail errors for both approaches are varied as A and R change. For the Delta method, the asymmetry of left and right tail errors is smaller than that when $n_X = n_Y = 50, n_f = 19$. However, similar to the case when $n_X = n_Y = 50, n_f = 19$, the asymmetry is more serious when A increases, especially for $R \leq 0.6$, while the tail errors are close to symmetry for $R = 0.8$. For a given A, the difference decreases as R increases in all cases. For MOVER approach, the difference of tail errors slightly decreases compared to the combination $n_X = n_Y = 50, n_f = 19$, except for $R = 0.2$. The left and right tail errors approach symmetry when $R \geq 0.6$. As A increases, the difference of tail errors gradually become smaller. On the whole, almost all of the differences for the MOVER approach are smaller than the correspondingly differences in the Delta method. Also, as a test with higher test accuracy is applied, the difference of the tail errors decreases in the MOVER approach, but increases in the Delta approach. As measurement error decreases, the difference decreases in both approaches.

4.2.3 For $n_X = 900, n_Y = 50, n_f = 49$

Table 4.3 presents the results for a unbalanced combination of $n_X = 900, n_Y = 50, n_f = 49$.

Unlike the balanced case, the coverage probability gets larger as σ_Y^2 increases in the Delta method, but there is no such trend in the MOVER approach. As with the Delta method in the balanced case of $n_X = n_Y = 100, n_f = 49$, most coverage

probabilities fall far beyond the target range when $R = 0.2$. Only one coverage probabilities are included in the target range when R is between 0.4 and 0.6. When $R = 0.8$, even though there is still only one coverage probability in the range, but most of the others fall close to the target range. For the MOVER approach, most of the coverage probabilities are in the target range. Those not included are very close to the target range.

As with the coverage probability, the width of the confidence interval gradually increases as σ_Y^2 increases for both approaches. Also, the width increases more when R is relatively large than those when R is small for a given A as σ_Y^2 increases. This suggests that as the σ_Y^2 increases, the confidence intervals become wider for both approaches when data is unbalanced, and the interval width is more affected by the σ_Y^2 if measurement error is small for a given test. Similar to the balanced data, if a test with higher accuracy is applied, the interval obtained by the MOVER approach is gradually getting much narrower than its corresponding width in the Delta method when $R = 0.2$, but is slightly wider when $R = 0.4$ and 0.6 and $A \leq 0.7$. When $R = 0.8$, the interval produced by both approaches are very close in values. Also, the values of the width in the unbalanced combination are close to that in the combination of $n_X = n_Y = 100, n_f = 49$.

As for the symmetry of tail errors, most of the difference of tail errors for the Delta method becomes larger as A increases. For a fixed A, the difference gets smaller as R increases. However, for the MOVER approach, the imbalance of the left and right tail errors slightly increases as the σ_Y^2 increases in most cases. The symmetry of tail errors in $n_X = 900, n_Y = 50, n_f = 49$ is somewhat less than that in $n_X = n_Y = 100, n_f = 49$ for the MOVER approach. This may be due to the imbalance of sample size. When $R = 0.2$, the symmetry is gradually improved in MOVER approach if a test with higher accuracy is applied. When A is fixed, the difference of tail errors decreases as measurement error decreases. In summary, the difference of tail errors slightly increases in the MOVER approach as σ_y^2 increases. If a test with higher accuracy

is applied, the difference of tail errors becomes larger in the Delta method, while becomes smaller in the MOVER approach. It also gets smaller as measurement error decreases in both approaches.

4.2.4 For $n_X = 1000, n_Y = 1000, n_f = 199$

Table 4.4 displays the simulation results of combination $n_X = n_Y = 1000, n_f = 199$.

Like the other balanced data, the coverage probability does not vary when σ_Y^2 changes. For the Delta method, when $R = 0.2$, most of the coverage probabilities still fall far beyond the target range. This suggests that if the measurement error is relatively large, increasing sample size has little effect on the improvement of coverage probability. When $R = 0.4$, most of the coverage probabilities are still not included in the range, but very close to it. When $R \geq 0.6$, most of the coverage probabilities are included in the range. For the MOVER approach, all of the coverage probabilities are in the target range and close to their nominal levels as expected. This suggests that greatly increasing sample size can greatly improve the performance of the procedures.

The widths for both approaches in this combination are much narrower than those that occur in the other combinations. As with the other balanced data, the interval width is stable when the σ_Y^2 changes, but decreases in both approaches as R increases. When $R = 0.2$, the interval width is wider in the MOVER approach than in the Delta method if the test accuracy is low, but quickly gets narrower than that in the Delta method as a test with higher accuracy is applied. The widths in both approaches are quite similar when $R \geq 0.4$.

The difference of the tail errors does not change obviously as σ_Y^2 changes. For the Delta method, as with the other sample size combinations, the asymmetry for the left and right tail errors is less serious when the value of A is small, but become serious unbalanced when A is large, especially when R is small. For the MOVER approach, the tail errors approach symmetry when a test with higher accuracy is applied and finally "converges" to the range of less than 1.0. Overall, if a test with

higher accuracy is applied, the differences of the left and right tail errors increases in the Delta method, but decreases in the MOVER approach. Also, the difference gets smaller as the measurement error decreases for the both approaches.

4.3 Conclusion

We compared the Delta method and the MOVER approach in terms of confidence interval coverage, interval width and symmetry of tail errors.

For coverage probability, as the variance of y changes, the coverage probability does not change for the balanced data, while it tends to increase for the unbalanced data under the Delta method. The proportion of coverage probability that is beyond the target range of coverage probability, (94.6%, 95.4%), is much higher for the Delta method than for the MOVER approach based on the sample size, reliability index and the test accuracy (see Fig.4.1, 4.2 and 4.3). For the Delta method, some coverage probabilities fall far away from the target range, especially in the cases of small sample size, relatively large measurement error and high test accuracy. In the case of large measurement error, it seems that increasing sample size has a slight effect on reducing the non-coverage probability in the Delta method. On the contrary, in the MOVER approach, most of the coverage probabilities fall in the target interval, and those not included in the target range are close to the range.

For the width of the confidence interval, as the variance of y changes, the interval width is stable for balanced data. However, the width becomes wider for both approaches if the data is unbalanced. Such increase in width is also greater if measurement error is smaller for a given test accuracy. This suggests that if measurement error is relatively small, the confidence interval is mainly affected by the variances of the biomarker values if the data is unbalanced. However, if the measurement error is relatively large, the confidence interval is greatly affected by the measurement error instead of the variances of the biomarker values. The interval widths in

both approaches are getting narrower and nearly equal when sample size is large and measurement error is relatively small (see Fig.4.4, 4.5). If sample size is small or measurement error is relatively large, the interval in the Delta method is wider than that in the MOVER approach. Also, it is slightly wider in the MOVER approach than that in the Delta method when a low accuracy test is applied, while it gets narrower than that in the Delta method when a high accuracy test is applied (see Fig. 4.6).

For the symmetry of tail errors (see Fig. 4.7, 4.8 and 4.9), the difference of tail errors in the Delta method is larger than that in the MOVER approach in terms of sample size, measurement error and test accuracy. Increasing sample size seems to have no effect on reducing the difference in the Delta method, but slightly reduces the difference in the MOVER approach. When measurement error is relatively large, the difference of the tail errors in the Delta method is much larger that in the MOVER approach. As the measurement error decreases, the difference tend to be more symmetric for both approaches. When a low accuracy test is applied, the difference of tail errors in the Delta method is slightly larger than that in the MOVER approach. However, as higher accuracy tests are applied, the difference in the Delta method is getting larger, while getting smaller in the MOVER approach. Hence, the MOVER approach gives more symmetric tail errors than the Delta method, especially when measurement error is large and test accuracy is high.

On the whole, in terms of coverage, interval width and symmetry of tail errors, the performance of the MOVER approach is much better than the Delta method in the presence of measurement error, especially when sample size is small, the measurement error is large and the accuracy of a test is high.

Table 4.1: Observed Coverage Probabilities for the Delta Method and the MOVER Approach to Construct a Two-Sided 95% Confidence Interval for the Area under the ROC Curve ($n_X = 50, n_Y = 50, n_f = 19$)

A	R	Delta method		MOVER approach	
		$\sigma_Y^2 = 0.5$	$\sigma_Y^2 = 5.0$	$\sigma_Y^2 = 0.5$	$\sigma_Y^2 = 5.0$
		CV(ML, MR)% width	CV(ML, MR)% width	CV(ML, MR)% width	CV(ML, MR)% width
0.6	0.2	97.90 (0.05, 2.05) 0.62	97.94 (0.05, 2.01) 0.62	93.79 (4.35, 1.86) 0.54	93.49 (4.64, 1.87) 0.54
	0.4	97.62 (0.27, 2.11) 0.44	97.61 (0.26, 2.13) 0.44	94.29 (4.02, 1.69) 0.50	94.14 (4.12, 1.74) 0.50
	0.6	96.48 (1.03, 2.49) 0.30	96.46 (1.06, 2.48) 0.30	94.87 (3.24, 1.89) 0.39	94.88 (3.32, 1.80) 0.39
	0.8	94.97 (2.39, 2.64) 0.24	95.04 (2.35, 2.61) 0.24	94.51 (2.97, 2.52) 0.25	94.41 (3.18, 2.41) 0.25
0.7	0.2	93.88 (0.00, 6.12) 0.64	94.01 (0.00, 5.99) 0.64	95.34 (3.30, 1.36) 0.49	95.33 (3.28, 1.39) 0.49
	0.4	96.11 (0.05, 3.84) 0.46	96.14 (0.03, 3.83) 0.46	94.89 (4.00, 1.11) 0.43	94.80 (4.01, 1.19) 0.43
	0.6	96.51 (0.51, 2.98) 0.29	96.59 (0.41, 3.00) 0.30	94.92 (3.63, 1.45) 0.34	94.95 (3.70, 1.35) 0.34
	0.8	95.16 (2.12, 2.72) 0.22	95.22 (2.08, 2.70) 0.23	94.65 (3.15, 2.20) 0.23	94.66 (3.20, 2.14) 0.24
0.8	0.2	89.78 (0.00, 10.22) 0.66	89.75 (0.00, 10.25) 0.66	95.86 (2.84, 1.30) 0.41	95.88 (2.77, 1.35) 0.41
	0.4	94.64 (0.00, 5.36) 0.47	94.59 (0.00, 5.41) 0.47	94.99 (3.79, 1.22) 0.33	94.94 (3.79, 1.27) 0.33
	0.6	96.21 (0.19, 3.60) 0.28	96.47 (0.14, 3.39) 0.28	95.05 (3.57, 1.38) 0.27	94.99 (3.66, 1.35) 0.28
	0.8	95.42 (1.77, 2.81) 0.20	95.30 (1.75, 2.95) 0.20	94.97 (3.14, 1.89) 0.20	94.81 (3.21, 1.98) 0.21
0.9	0.2	86.40 (0.00, 13.60) 0.67	86.49 (0.00, 13.51) 0.67	95.65 (2.60, 1.75) 0.29	95.58 (2.62, 1.80) 0.29
	0.4	92.62 (0.00, 7.38) 0.45	92.57 (0.00, 7.43) 0.45	95.20 (3.25, 1.55) 0.23	95.23 (3.17, 1.60) 0.23
	0.6	95.69 (0.03, 4.28) 0.23	95.75 (0.05, 4.20) 0.23	95.21 (3.34, 1.45) 0.18	94.98 (3.47, 1.55) 0.19
	0.8	95.65 (1.34, 3.01) 0.14	95.62 (1.33, 3.05) 0.15	95.08 (2.99, 1.93) 0.14	95.05 (3.11, 1.84) 0.15
0.95	0.2	84.66 (0.00, 15.34) 0.67	84.73 (0.00, 15.27) 0.67	95.70 (2.39, 1.91) 0.22	95.45 (2.59, 1.96) 0.22
	0.4	91.34 (0.00, 8.66) 0.43	91.39 (0.00, 8.61) 0.43	95.31 (2.97, 1.72) 0.16	95.22 (2.91, 1.87) 0.16
	0.6	95.24 (0.01, 4.75) 0.18	95.32 (0.01, 4.67) 0.19	95.35 (3.02, 1.63) 0.12	94.97 (3.32, 1.71) 0.12
	0.8	95.79 (1.14, 3.07) 0.10	95.83 (1.07, 3.10) 0.10	95.19 (2.80, 2.01) 0.10	95.14 (2.99, 1.87) 0.10

Note: CV means coverage probability. ML means missing left coverage probability. MR means missing right coverage probability.

Table 4.2: Observed Coverage Probabilities for the Delta Method and the MOVER Approach to Construct a Two-Sided 95% Confidence Interval for the Area under the ROC Curve ($n_X = 100, n_Y = 100, n_f = 49$)

A	R	Delta method		MOVER approach	
		$\sigma_Y^2 = 0.5$	$\sigma_Y^2 = 5.0$	$\sigma_Y^2 = 0.5$	$\sigma_Y^2 = 5.0$
		CV(ML, MR)% width	CV(ML, MR)% width	CV(ML, MR)% width	CV(ML, MR)% width
0.6	0.2	97.33 (0.02, 2.65) 0.50	97.32 (0.02, 2.66) 0.50	93.71 (4.38, 1.91) 0.50	93.81 (4.33, 1.80) 0.50
	0.4	96.70 (0.63, 2.67) 0.28	96.70 (0.56, 2.74) 0.29	94.90 (3.31, 1.79) 0.38	94.78 (3.41, 1.81) 0.38
	0.6	95.74 (1.74, 2.52) 0.20	95.78 (1.65, 2.57) 0.20	94.98 (2.90, 2.12) 0.22	95.06 (2.83, 2.11) 0.22
	0.8	95.32 (2.26, 2.42) 0.17	95.47 (2.10, 2.43) 0.17	95.16 (2.53, 2.31) 0.17	95.17 (2.58, 2.25) 0.17
0.7	0.2	94.11 (0.00, 5.89) 0.54	94.12 (0.01, 5.87) 0.54	94.78 (3.93, 1.29) 0.44	94.73 (3.94, 1.33) 0.44
	0.4	96.16 (0.12, 3.72) 0.30	96.26 (0.09, 3.65) 0.30	94.89 (3.55, 1.56) 0.33	94.89 (3.62, 1.49) 0.33
	0.6	96.21 (1.07, 2.72) 0.20	96.21 (1.03, 2.76) 0.20	95.12 (3.11, 1.77) 0.21	95.14 (3.02, 1.84) 0.22
	0.8	95.57 (1.99, 2.44) 0.16	95.71 (1.88, 2.41) 0.16	95.27 (2.62, 2.11) 0.16	95.36 (2.54, 2.10) 0.16
0.8	0.2	91.22 (0.00, 8.78) 0.57	91.21 (0.00, 8.79) 0.57	94.99 (3.40, 1.61) 0.34	94.92 (3.47, 1.61) 0.34
	0.4	95.30 (0.02, 4.68) 0.30	95.17 (0.03, 4.80) 0.30	95.00 (3.27, 1.73) 0.27	95.09 (3.31, 1.60) 0.27
	0.6	96.47 (0.51, 3.02) 0.18	96.50 (0.50, 3.00) 0.18	95.16 (3.01, 1.83) 0.19	95.21 (3.01, 1.78) 0.19
	0.8	95.69 (1.86, 2.45) 0.14	95.85 (1.73, 2.42) 0.14	95.44 (2.60, 1.96) 0.14	95.46 (2.64, 1.90) 0.14
0.9	0.2	88.84 (0.00, 11.16) 0.58	88.90 (0.00, 11.10) 0.58	95.10 (2.91, 1.99) 0.24	95.18 (2.90, 1.92) 0.24
	0.4	93.97 (0.00, 6.03) 0.27	94.15 (0.00, 5.85) 0.27	95.13 (2.85, 2.02) 0.18	95.18 (3.02, 1.80) 0.18
	0.6	96.37 (0.14, 3.49) 0.14	96.38 (0.12, 3.50) 0.14	95.31 (2.77, 1.92) 0.13	95.48 (2.82, 1.70) 0.13
	0.8	95.78 (1.65, 2.57) 0.10	95.82 (1.62, 2.56) 0.10	95.60 (2.42, 1.98) 0.10	95.65 (2.43, 1.92) 0.10
0.95	0.2	87.83 (0.00, 12.17) 0.58	87.90 (0.00, 12.10) 0.58	94.98 (2.83, 2.19) 0.18	95.01 (2.80, 2.19) 0.18
	0.4	93.24 (0.00, 6.76) 0.23	93.33 (0.00, 6.67) 0.24	95.03 (2.76, 2.21) 0.12	95.08 (2.91, 2.01) 0.12
	0.6	96.06 (0.03, 3.91) 0.10	96.31 (0.04, 3.65) 0.10	95.60 (2.55, 1.85) 0.09	95.43 (2.73, 1.84) 0.09
	0.8	95.83 (1.46, 2.71) 0.07	95.72 (1.56, 2.72) 0.07	95.64 (2.40, 1.96) 0.07	95.63 (2.45, 1.92) 0.07

Note: CV means coverage probability. ML means missing left coverage probability. MR means missing right coverage probability.

Table 4.3: Observed Coverage Probabilities for the Delta Method and the MOVER Approach to Construct a Two-Sided 95% Confidence Interval for the Area under the ROC Curve ($n_X = 900, n_Y = 50, n_f = 49$)

A	R	Delta method		MOVER approach	
		$\sigma_Y^2 = 0.5$	$\sigma_Y^2 = 5.0$	$\sigma_Y^2 = 0.5$	$\sigma_Y^2 = 5.0$
		CV(ML, MR)% width	CV(ML, MR)% width	CV(ML, MR)% width	CV(ML, MR)% width
0.6	0.2	97.33 (0.01, 2.66) 0.50	97.63 (0.02, 2.35) 0.53	93.87 (4.58, 1.55) 0.50	93.36 (5.02, 1.62) 0.51
	0.4	96.69 (0.73, 2.58) 0.27	97.15 (0.55, 2.30) 0.33	94.73 (3.61, 1.66) 0.37	94.52 (3.77, 1.71) 0.41
	0.6	95.14 (1.95, 2.91) 0.19	95.60 (1.55, 2.85) 0.24	94.52 (2.94, 2.54) 0.20	94.33 (3.20, 2.47) 0.27
	0.8	94.52 (3.08, 2.40) 0.15	94.79 (2.88, 2.33) 0.21	94.37 (3.27, 2.36) 0.15	94.34 (3.50, 2.16) 0.21
0.7	0.2	94.20 (0.00, 5.80) 0.54	94.48 (0.00, 5.52) 0.56	94.67 (3.95, 1.38) 0.44	94.57 (4.06, 1.37) 0.45
	0.4	96.38 (0.14, 3.48) 0.29	96.52 (0.09, 3.39) 0.34	94.89 (3.72, 1.39) 0.32	94.96 (3.74, 1.30) 0.36
	0.6	95.70 (1.28, 3.02) 0.18	95.96 (0.97, 3.07) 0.23	94.71 (3.18, 2.11) 0.20	94.61 (3.37, 2.02) 0.25
	0.8	94.21 (3.28, 2.51) 0.14	94.71 (2.80, 2.49) 0.20	94.17 (3.58, 2.25) 0.14	94.18 (3.73, 2.09) 0.20
0.8	0.2	91.82 (0.00, 8.18) 0.57	92.00 (0.00, 8.00) 0.59	95.06 (3.30, 1.64) 0.34	95.05 (3.36, 1.59) 0.36
	0.4	95.58 (0.01, 4.41) 0.29	95.84 (0.00, 4.16) 0.34	94.79 (3.62, 1.59) 0.26	94.93 (3.66, 1.41) 0.29
	0.6	96.05 (0.84, 3.11) 0.16	96.22 (0.52, 3.26) 0.21	94.88 (3.24, 1.88) 0.17	94.75 (3.36, 1.89) 0.22
	0.8	94.32 (3.31, 2.37) 0.12	94.78 (2.72, 2.50) 0.18	94.23 (3.71, 2.06) 0.12	94.28 (3.77, 1.95) 0.18
0.9	0.2	89.33 (0.00, 10.67) 0.58	89.90 (0.00, 10.1) 0.60	95.14 (2.86, 2.00) 0.24	95.04 (2.91, 2.05) 0.25
	0.4	94.07 (0.00, 5.93) 0.26	94.71 (0.00, 5.29) 0.31	94.89 (3.18, 1.93) 0.18	95.08 (3.21, 1.71) 0.20
	0.6	96.39 (0.35, 3.26) 0.12	96.32 (0.24, 3.44) 0.17	94.97 (3.14, 1.89) 0.12	95.01 (3.11, 1.88) 0.16
	0.8	94.56 (3.09, 2.35) 0.08	95.08 (2.57, 2.35) 0.13	94.47 (3.49, 2.04) 0.09	94.62 (3.52, 1.86) 0.13
0.95	0.2	88.37 (0.00, 11.63) 0.58	88.89 (0.00, 11.11) 0.60	95.16 (2.65, 2.19) 0.18	95.14 (2.63, 2.23) 0.19
	0.4	93.13 (0.00, 6.87) 0.22	93.88 (0.00, 6.12) 0.28	94.91 (2.97, 2.12) 0.12	94.99 (3.00, 2.01) 0.13
	0.6	96.13 (0.18, 3.69) 0.09	96.35 (0.07, 3.58) 0.12	94.95 (3.13, 1.92) 0.08	95.06 (2.95, 1.99) 0.11
	0.8	94.68 (2.97, 2.35) 0.06	95.26 (2.35, 2.39) 0.09	94.69 (3.29, 2.02) 0.06	94.77 (3.39, 1.84) 0.09

Note: CV means coverage probability. ML means missing left coverage probability. MR means missing right coverage probability.

Table 4.4: Observed Coverage Probabilities for the Delta Method and the MOVER Approach to Construct a Two-Sided 95% Confidence Interval for the Area under the ROC Curve ($n_X = 1000, n_Y = 1000, n_f = 199$)

A	R	Delta method		MOVER approach	
		$\sigma_Y^2 = 0.5$	$\sigma_Y^2 = 5.0$	$\sigma_Y^2 = 0.5$	$\sigma_Y^2 = 5.0$
		CV(ML, MR)% width	CV(ML, MR)% width	CV(ML, MR)% width	CV(ML, MR)% width
0.6	0.2	95.60 (0.01, 4.39) 0.19	95.59 (0.01, 4.40) 0.19	94.90 (3.36, 1.74) 0.30	94.84 (3.39, 1.77) 0.30
	0.4	95.89 (1.07, 3.04) 0.08	95.83 (1.13, 3.04) 0.08	95.05 (2.85, 2.10) 0.09	94.79 (2.97, 2.24) 0.09
	0.6	95.15 (2.32, 2.53) 0.06	95.06 (2.29, 2.65) 0.06	95.07 (2.65, 2.28) 0.06	94.93 (2.66, 2.41) 0.07
	0.8	95.23 (2.33, 2.44) 0.05	94.87 (2.58, 2.55) 0.05	95.21 (2.44, 2.35) 0.05	94.92 (2.66, 2.42) 0.05
0.7	0.2	93.19 (0.00, 6.81) 0.25	93.39 (0.00, 6.61) 0.25	94.93 (3.08, 1.99) 0.27	94.86 (3.13, 2.01) 0.27
	0.4	96.19 (0.26, 3.55) 0.10	96.22 (0.26, 3.52) 0.10	94.98 (3.09, 1.93) 0.11	94.85 (3.16, 1.99) 0.11
	0.6	95.43 (1.90, 2.67) 0.06	95.30 (1.97, 2.73) 0.06	95.12 (2.82, 2.06) 0.07	95.20 (2.70, 2.10) 0.07
	0.8	95.26 (2.30, 2.44) 0.05	95.05 (2.41, 2.54) 0.05	95.29 (2.43, 2.28) 0.05	95.05 (2.66, 2.29) 0.05
0.8	0.2	92.00 (0.00, 8.00) 0.28	91.94 (0.00, 8.06) 0.28	94.97 (2.65, 2.38) 0.23	94.86 (2.75, 2.39) 0.23
	0.4	95.72 (0.02, 4.26) 0.10	95.63 (0.02, 4.35) 0.10	95.19 (2.84, 1.97) 0.11	95.01 (2.95, 2.04) 0.11
	0.6	95.55 (1.47, 2.98) 0.06	95.47 (1.46, 3.07) 0.06	95.15 (2.79, 2.06) 0.06	95.28 (2.67, 2.05) 0.06
	0.8	95.23 (2.22, 2.55) 0.04	95.16 (2.41, 2.43) 0.05	95.12 (2.62, 2.26) 0.04	95.14 (2.67, 2.19) 0.05
0.9	0.2	91.26 (0.00, 8.74) 0.27	91.23 (0.00, 8.77) 0.27	95.09 (2.55, 2.36) 0.16	94.99 (2.59, 2.42) 0.16
	0.4	94.64 (0.00, 5.36) 0.08	94.81 (0.00, 5.19) 0.08	95.30 (2.67, 2.03) 0.09	95.16 (2.81, 2.03) 0.09
	0.6	95.43 (1.05, 3.52) 0.05	95.31 (1.00, 3.69) 0.05	95.26 (2.66, 2.08) 0.05	95.11 (2.64, 2.25) 0.05
	0.8	95.10 (2.06, 2.84) 0.03	95.30 (2.16, 2.54) 0.03	95.10 (2.53, 2.37) 0.03	95.22 (2.60, 2.18) 0.03
0.95	0.2	90.90 (0.00, 9.10) 0.25	91.00 (0.00, 9.00) 0.25	95.11 (2.50, 2.39) 0.10	95.08 (2.48, 2.44) 0.10
	0.4	94.37 (0.00, 5.63) 0.06	94.48 (0.00, 5.52) 0.06	95.33 (2.57, 2.10) 0.06	95.04 (2.71, 2.25) 0.06
	0.6	95.38 (0.85, 3.77) 0.03	95.34 (0.70, 3.96) 0.03	95.21 (2.64, 2.15) 0.03	95.21 (2.52, 2.27) 0.03
	0.8	95.10 (1.95, 2.95) 0.02	95.35 (2.02, 2.63) 0.02	94.95 (2.54, 2.51) 0.02	95.32 (2.49, 2.19) 0.02

Note: CV means coverage probability. ML means missing left coverage probability. MR means missing right coverage probability.

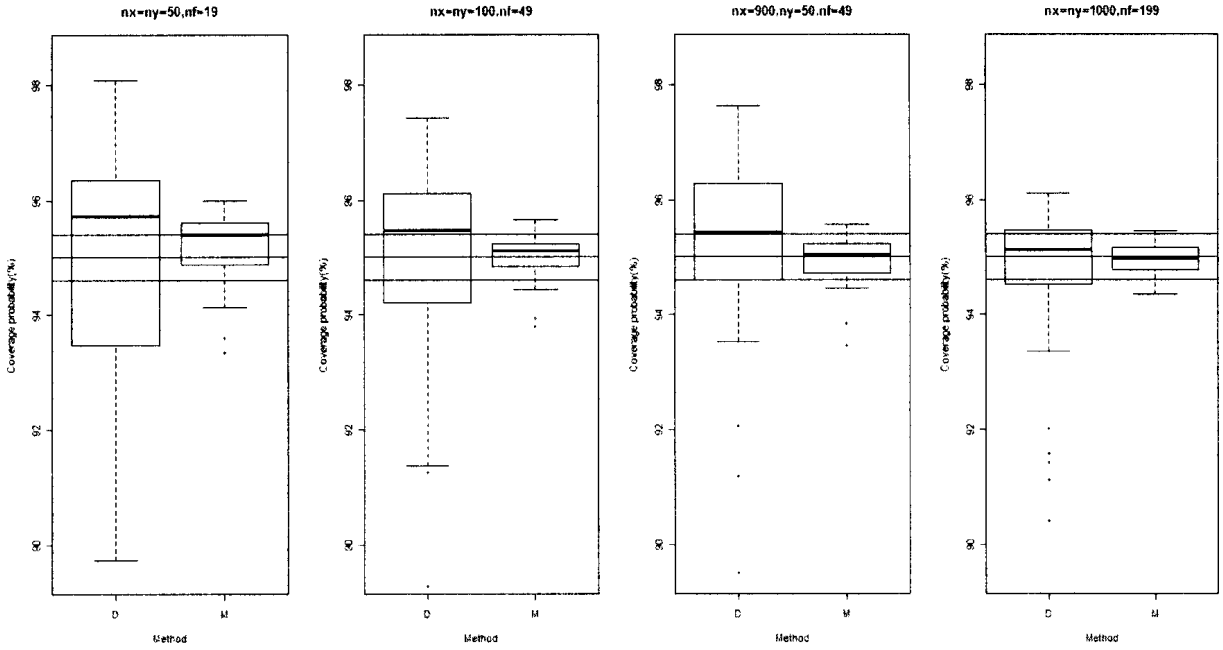

Figure 4.1: The coverage probability based on 10,000 runs for the 95% confidence intervals of the area under the ROC curve. Each boxplot was based on sample size combination and was drawn from coverage probabilities of 80 parameter combinations. Methods 'D' and 'M' represent 'Delta method' and 'MOVER approach', respectively.

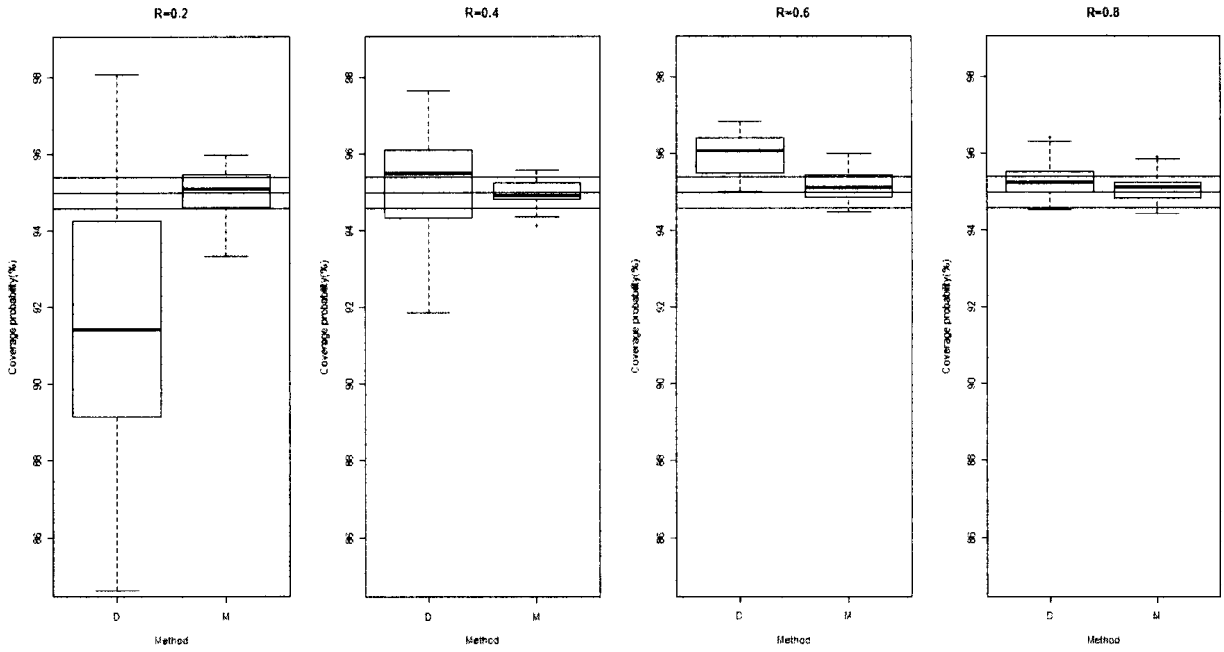

Figure 4.2: The coverage probability based on 10,000 runs for the 95% confidence intervals of the area under the ROC curve. Each boxplot was based on the reliability index and was drawn from coverage probabilities of 80 parameter combinations. Methods 'D' and 'M' represent 'Delta method' and 'MOVER approach', respectively.

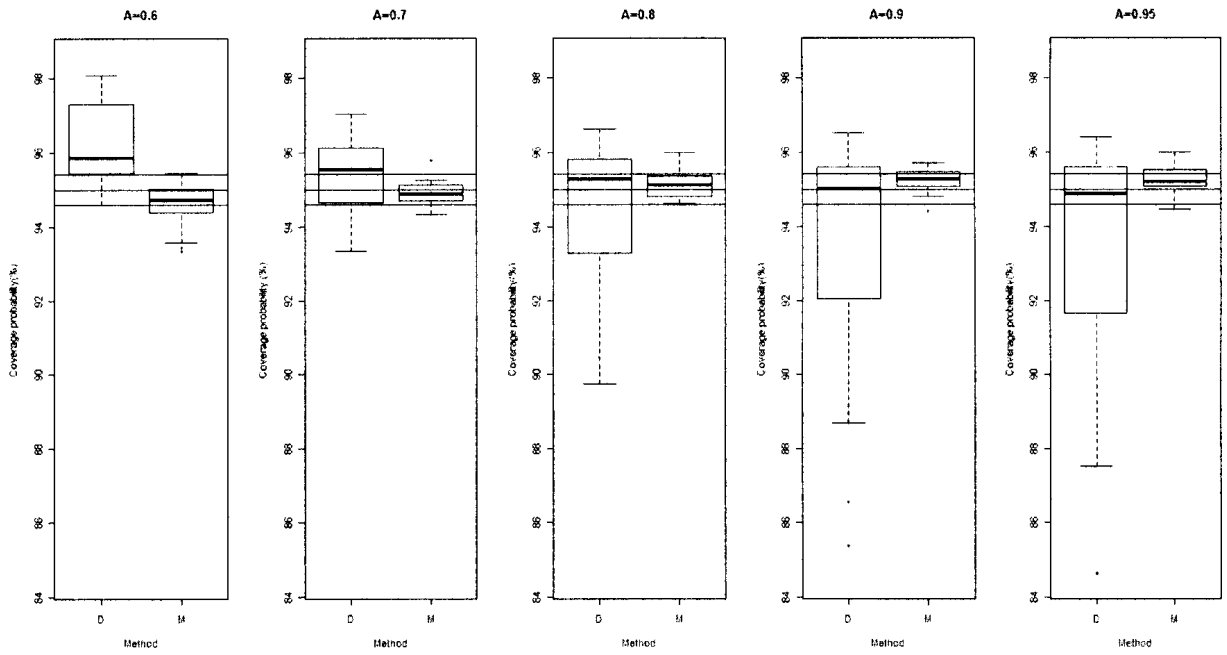

Figure 4.3: The coverage probability based on 10,000 runs for the 95% confidence intervals of the area under the ROC curve. Each boxplot was based on the area under the ROC curve and was drawn from coverage probabilities of 80 parameter combinations. Methods 'D' and 'M' represent 'Delta method' and 'MOVER approach', respectively.

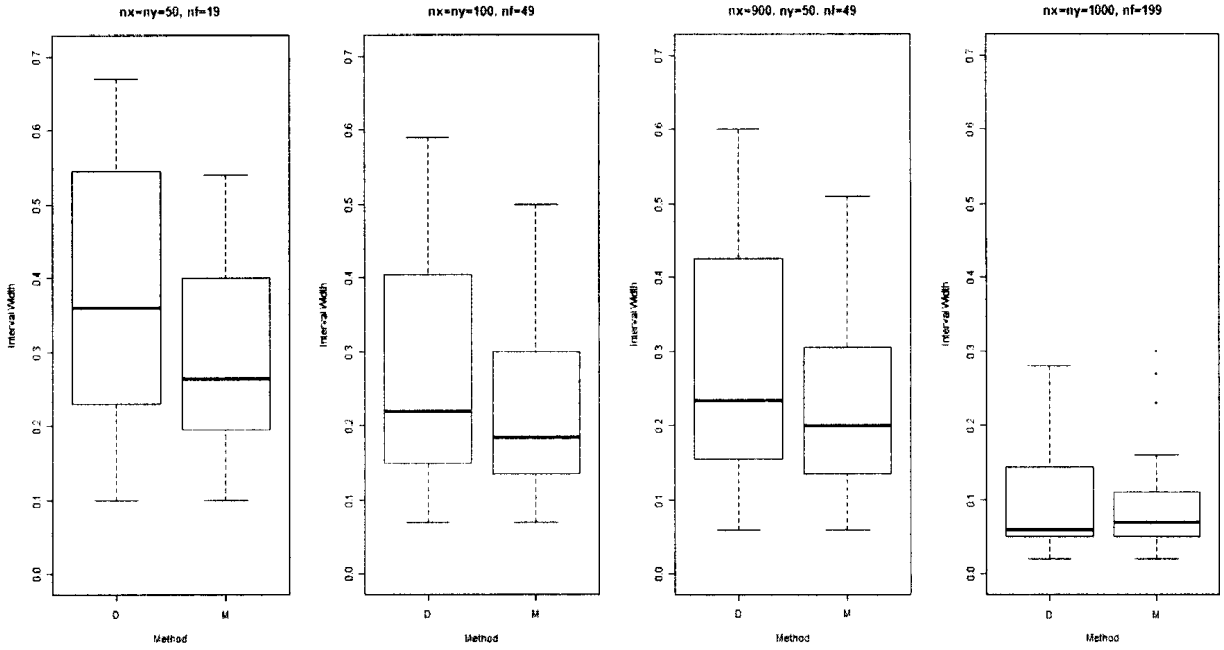

Figure 4.4: The interval width based on 10,000 runs for the 95% confidence intervals of the area under the ROC curve. Each boxplot was based on sample size combination and was drawn from coverage probabilities of 80 parameter combinations. Methods 'D' and 'M' represent 'Delta method' and 'MOVER approach', respectively.

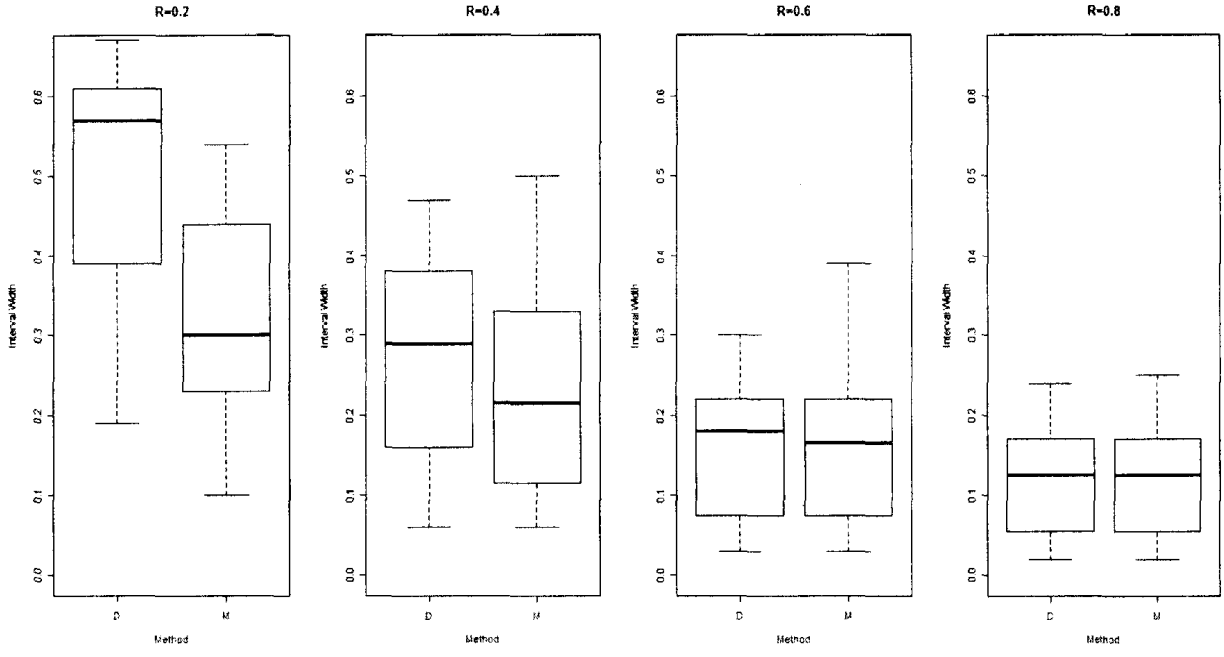

Figure 4.5: The interval width based on 10,000 runs for the 95% confidence intervals of the area under the ROC curve. Each boxplot was based on the reliability index and was drawn from coverage probabilities of 80 parameter combinations. Methods 'D' and 'M' represent 'Delta method' and 'MOVER approach', respectively.

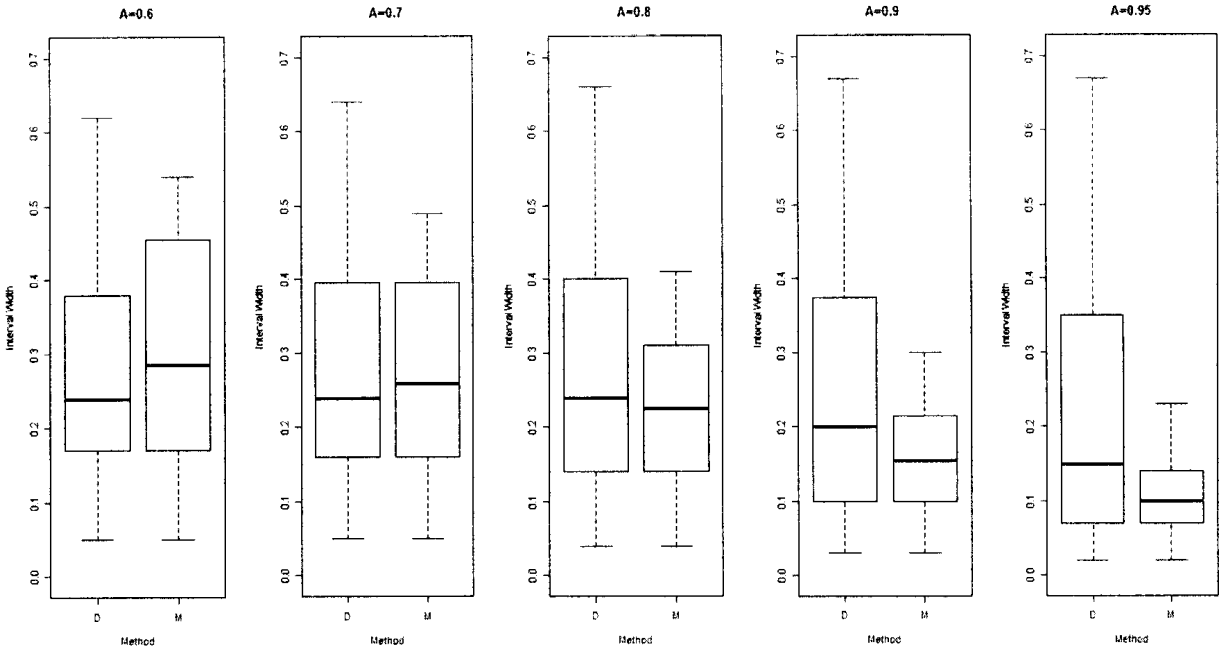

Figure 4.6: The interval width based on 10,000 runs for the 95% confidence intervals of the area under the ROC curve. Each boxplot was based on the area under the ROC curve and was drawn from coverage probabilities of 80 parameter combinations. Methods 'D' and 'M' represent 'Delta method' and 'MOVER approach', respectively.

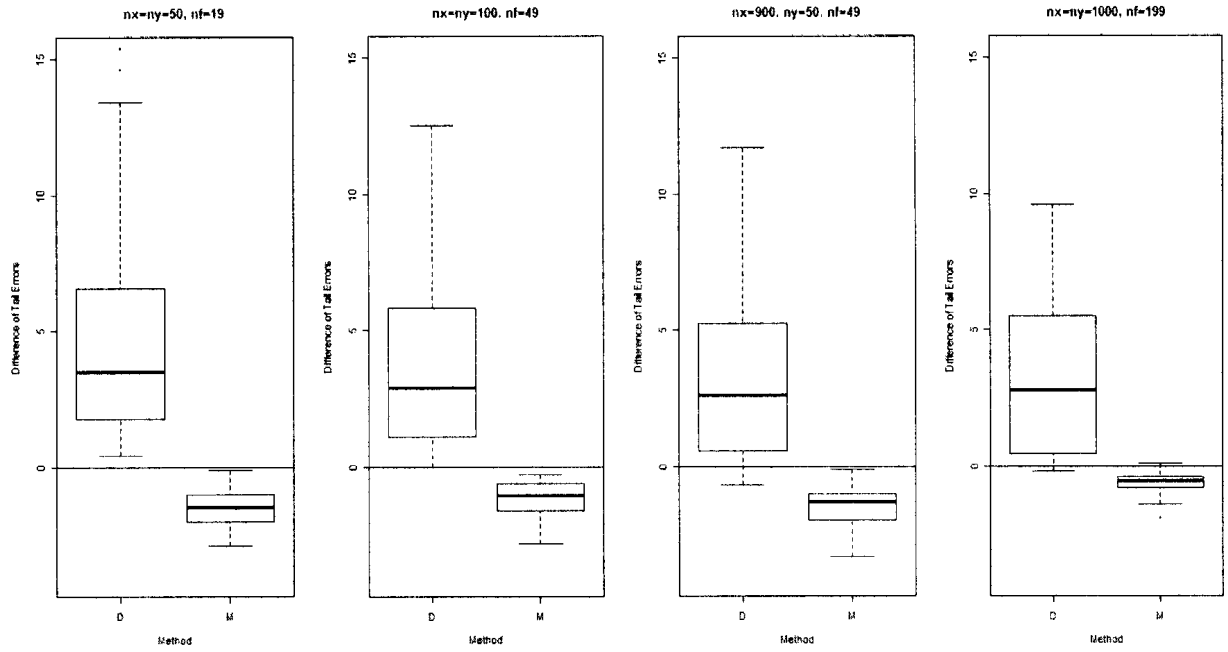

Figure 4.7: The difference of tail errors based on 10,000 runs for the 95% confidence intervals of the area under the ROC curve. Each boxplot was based on sample size combination and was drawn from coverage probabilities of 80 parameter combinations. Methods 'D' and 'M' represent 'Delta method' and 'MOVER approach', respectively.

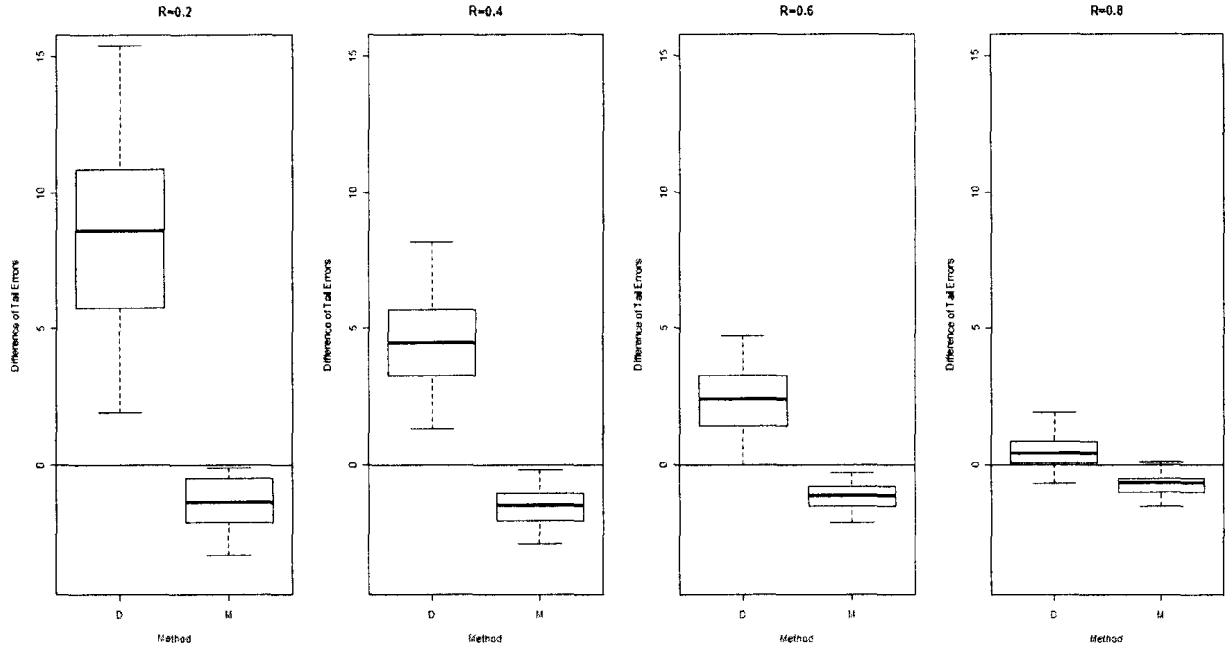

Figure 4.8: The difference of tail errors based on 10,000 runs for the 95% confidence intervals of the area under the ROC curve. Each boxplot was based on the reliability index and was drawn from coverage probabilities of 80 parameter combinations. Methods 'D' and 'M' represent 'Delta method' and 'MOVER approach', respectively.

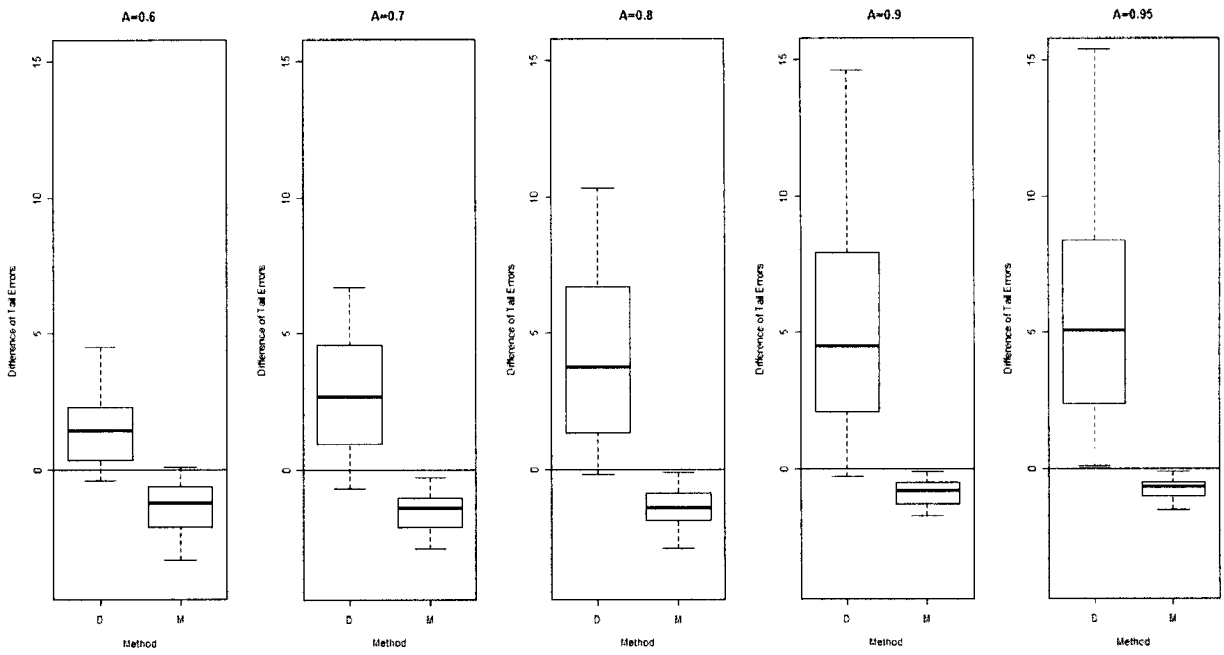

Figure 4.9: The difference of tail errors based on 10,000 runs for the 95% confidence intervals of the area under the ROC curve. Each boxplot was based on the area under the ROC curve and was drawn from coverage probabilities of 80 parameter combinations. Methods 'D' and 'M' represent 'Delta method' and 'MOVER approach', respectively.

Chapter 5

EXAMPLE

Oxygen free radicals are believed to be related to the processes of atherosclerotic coronary heart disease. Recent studies (Hoffman and Garewal, 1995) suggest that the key factor in the atherosclerotic process may be the oxidative modification of low density lipoproteins. Thiobarbituric acid reaction substances (TBARS) measure subproducts of lipid peroxidation and thus can be used as a biomarker to discriminate measurements between cardiovascular disease cases and healthy controls.

A population-based sample was randomly selected from residents of Erie and Niagara counties, New York, who were aged 35-79 years. The New York state Department of Motor Vehicles driver's license rolls were used as the sampling frame for adults between ages 35 and 64 years, while the sample for those who were aged 65-79 was randomly selected from the Health Care Financing Administration.

The cases were defined as persons with myocardial infarction. Each participant in the study provided blood sample, physical measurements, and answering a detailed questionnaire on various behavioral and physiologic patterns. Of the participants, 60 subjects had a history of cancer, 68 subjects had incomplete information on TBARS and 75 subjects were non-White. All of these subjects were excluded from the study. After the exclusion, a total of 474 White men and 494 White women were selected for the analysis.

Schisterman *et al.* (2001) found that measurements of TBARS on the raw scale were skewed, thus reciprocal of square root transformation was applied. Note that the values of biomarker for the cases become smaller than those for controls after the transformation. Therefore, when the area under the ROC curve was being estimated,

the sign of the estimator of δ needs to be reversed so that the area under the ROC curve is greater than 0.5. The following is the summary of the transformed data.

$$\overline{x} = 0.604 \qquad n_x = 928 \qquad S_x^2 = 0.0913$$

$$\overline{y} = 0.450 \qquad n_y = 40 \qquad S_y^2 = 0.0886.$$

The estimation of variance for the random measurement error in the analysis of TBARS was carried out by a reliability study on a sample of 10 participants. Seven women and three men were followed for 6 months. Twelve-hour fasting blood samples were obtained every month on the same day of each female's menstrual cycle and every month on the same calendar day for each male.

From the reliability study, the variance of the random measurement error and other summary data are obtained as follow:

$$\widehat{\sigma}_\varepsilon^2 = 0.0567 \qquad n_f = 41.$$

Hence, the data had values of $n_x = 928, n_y = 40$ and $n_f = 41$; these values were very close to the unbalanced combination $n_x = 900, n_y = 50$ and $n_f = 49$ in the simulation study.

The estimate of R was obtained by substituting estimates of the parameters in the formula of R as follow:

$$R = \frac{\sigma_X^2 + \sigma_Y^2}{\sigma_X^2 + \sigma_Y^2 + 2\sigma_\varepsilon^2},$$

which is estimated by

$$\begin{aligned} \widehat{R} &= \frac{S_x^2 + S_y^2}{S_x^2 + S_y^2 + 2\widehat{\sigma}_\varepsilon^2} \\ &= \frac{0.0913 + 0.0886}{0.0913 + 0.0886 + 2 \times 0.0567} \\ &= 0.613 \end{aligned}$$

When measurement error is considered, the estimate of A can be corrected as

$$\begin{aligned}
\widehat{A}_c &= \Phi(\widehat{\delta}) \\
&= \Phi\left(\frac{\overline{y}-\overline{x}}{\sqrt{S_x^2+S_y^2-2\widehat{\sigma}_\varepsilon^2}}\right) \\
&= \Phi\left(\frac{0.450-0.604}{\sqrt{0.0913+0.0886-2\times 0.0567}}\right) \\
&= \Phi(-0.597) \\
&= \Phi(0.597) \\
&= 0.725
\end{aligned}$$

Note that since the value of $\widehat{\delta}$ is less than 0, the sign of $\widehat{\delta}$ is reversed so that the value of $\widehat{A}_c$ is greater than 0.5. The same way will be done in the following estimations for A.

The lower and upper limits of A_c can be obtained by the Delta method and the MOVER approach, respectively.

For the Delta method, the variance for A_c is given by

$$\begin{aligned}
\widehat{\mathrm{var}}(\widehat{\delta}) = &\left[\frac{S_x^2}{n_x}+\frac{S_y^2}{n_y}\right]\times(S_x^2+S_y^2-2\widehat{\sigma}_\varepsilon^2)^{-1}+ \\
&\frac{(\overline{y}-\overline{x})^2}{4(S_x^2+S_y^2-2\widehat{\sigma}_\varepsilon^2)^3}\times\left[\frac{2S_x^4}{n_x-1}+\frac{2S_y^4}{n_y-1}+\frac{8\widehat{\sigma}_\varepsilon^4}{n_f}\right].
\end{aligned}$$

Substitute the estimates for the parameters in the above formula and get

$$\widehat{\mathrm{var}}(\widehat{\delta}) = 0.0559$$

Then, the lower and upper limits of δ is given by

$$\begin{aligned}
l_\delta &= \widehat{\delta}-z_{\alpha/2}\times\sqrt{\widehat{\mathrm{var}}(\widehat{\delta})} \\
&= -0.597-1.96\times\sqrt{0.0559} \\
&= -1.060
\end{aligned}$$

$$\begin{aligned} u_\delta &= \widehat{\delta} + z_{\alpha/2} \times \sqrt{\widehat{\text{var}}(\widehat{\delta})} \\ &= -0.597 + 1.96 \times \sqrt{0.0559} \\ &= -0.134 \end{aligned}$$

The confidence limits of A_c using the Delta method therefore are given by

$$L_{A_c} = \Phi(-u_\delta) = \Phi(0.134) = 0.553$$
$$U_{A_c} = \Phi(-l_\delta) = \Phi(1.060) = 0.856$$

That is, using the Delta method, the 95% confidence limits for A_c when considering measurement error are (0.553, 0.856).

An alternative way to calculate the confidence limits for A_c is to apply the MOVER approach. The lower and upper limits for μ_x and μ_y are

$$\begin{aligned} l_{\mu_x} &= \overline{x} - z_{\alpha/2}\sqrt{\frac{S_x^2}{n_x}} \\ &= 0.604 - 1.96 \times \sqrt{\frac{0.0913}{928}} \\ &= 0.585 \end{aligned}$$

$$\begin{aligned} u_{\mu_x} &= \overline{x} + z_{\alpha/2}\sqrt{\frac{S_x^2}{n_x}} \\ &= 0.604 + 1.96 \times \sqrt{\frac{0.0913}{928}} \\ &= 0.623 \end{aligned}$$

$$\begin{aligned} l_{\mu_y} &= \overline{y} - z_{\alpha/2}\sqrt{\frac{S_y^2}{n_y}} \\ &= 0.450 - 1.96 \times \sqrt{\frac{0.0886}{40}} \\ &= 0.358 \end{aligned}$$

$$\begin{aligned} u_{\mu_y} &= \overline{y} + z_{\alpha/2}\sqrt{\frac{S_y^2}{n_y}} \\ &= 0.450 + 1.96 \times \sqrt{\frac{0.0886}{40}} \\ &= 0.542 \end{aligned}$$

Let θ_1 denote $\mu_y - \mu_x$. Then the lower and upper limits for θ_1 are

$$\begin{aligned} l_1 &= \overline{y} - \overline{x} - \sqrt{(\overline{y} - l_{\mu_y})^2 + (u_{\mu_x} - \overline{x})^2} \\ &= 0.450 - 0.604 - \sqrt{(0.450 - 0.358)^2 + (0.623 - 0.604)^2} \\ &= -0.248 \end{aligned}$$

$$\begin{aligned} u_1 &= \overline{y} - \overline{x} + \sqrt{(u_{\mu_y} - \overline{y})^2 + (\overline{x} - l_{\mu_x})^2} \\ &= 0.450 - 0.604 + \sqrt{(0.542 - 0.450)^2 + (0.604 - 0.585)^2} \\ &= -0.060 \end{aligned}$$

For σ_x^2 and σ_y^2, since

$$\sigma_x^2 \sim \chi^2_{n_x-1} \qquad \sigma_y^2 \sim \chi^2_{n_y-1},$$

Hence, the lower and upper limits for σ_x^2 and σ_y^2 are

$$\begin{aligned} l_{\sigma_x^2} &= (n_x - 1)\frac{S_x^2}{\chi^2_{(1-\alpha/2),(n_x-1)}} \\ &= (928 - 1) \times \frac{0.0913}{1013.27} \\ &= 0.084 \end{aligned}$$

$$\begin{aligned} u_{\sigma_x^2} &= (n_x - 1)\frac{S_x^2}{\chi^2_{(\alpha/2),(n_x-1)}} \\ &= (928 - 1) \times \frac{0.0913}{844.52} \\ &= 0.100 \end{aligned}$$

$$
\begin{aligned}
l_{\sigma_y^2} &= (n_y - 1)\frac{S_y^2}{\chi^2_{(1-\alpha/2),(n_y-1)}} \\
&= (40 - 1) \times \frac{0.0886}{58.12} \\
&= 0.059
\end{aligned}
$$

$$
\begin{aligned}
u_{\sigma_y^2} &= (n_y - 1)\frac{S_y^2}{\chi^2_{(\alpha/2),(n_y-1)}} \\
&= (40 - 1) \times \frac{0.0886}{23.65} \\
&= 0.146
\end{aligned}
$$

Then the lower and upper limits for $\sigma_x^2 + \sigma_y^2$ are

$$
\begin{aligned}
l_{\sigma_x^2+\sigma_y^2} &= S_x^2 + S_y^2 - \sqrt{(S_x^2 - l_{\sigma_x^2})^2 + (S_y^2 - l_{\sigma_y^2})^2} \\
&= 0.0913 + 0.0886 - \sqrt{(0.0913 - 0.084)^2 + (0.0886 - 0.059)^2} \\
&= 0.149
\end{aligned}
$$

$$
\begin{aligned}
u_{\sigma_x^2+\sigma_y^2} &= S_x^2 + S_y^2 + \sqrt{(u_{\sigma_x^2} - S_x^2)^2 + (u_{\sigma_y^2} - S_y^2)^2} \\
&= 0.0913 + 0.0886 + \sqrt{(0.100 - 0.0913)^2 + (0.146 - 0.0886)^2} \\
&= 0.238
\end{aligned}
$$

For $\sqrt{\sigma_x^2 + \sigma_y^2 - 2\sigma_\varepsilon^2}$, which is denoted by θ_2, first calculate the confidence limits for σ_ε^2, which follows $\chi^2_{n_f}$ distribution.

$$
\begin{aligned}
l_{\sigma_\varepsilon^2} &= n_f \frac{\widehat{\sigma}_\varepsilon^2}{\chi^2_{(1-\alpha/2),n_f}} \\
&= 41 \times \frac{0.0567}{60.56} \\
&= 0.038
\end{aligned}
$$

$$
\begin{aligned}
u_{\sigma_\varepsilon^2} &= n_f \frac{\widehat{\sigma}_\varepsilon^2}{\chi^2_{\alpha/2,n_f}} \\
&= 41 \times \frac{0.0567}{25.21} \\
&= 0.092
\end{aligned}
$$

Then the lower limits for $\sigma_x^2 + \sigma_y^2 - 2\sigma_\varepsilon^2$ can be estimated as

$$\begin{aligned} l_{\sigma_x^2+\sigma_y^2-2\sigma_\varepsilon^2} &= \widehat{\sigma}_x^2 + \widehat{\sigma}_y^2 - 2\widehat{\sigma}_\varepsilon^2 - \sqrt{(\widehat{\sigma}_x^2 + \widehat{\sigma}_y^2 - l_{\sigma_x^2+\sigma_y^2})^2 + 4(u_{\sigma_\varepsilon^2} - \widehat{\sigma}_\varepsilon^2)^2} \\ &= (0.0913 + 0.0886) - 2 \times 0.0567 - \\ &\quad \sqrt{(0.0913 + 0.0886 - 0.149)^2 + 4 \times (0.092 - 0.0567)^2} \\ &= -0.010 \end{aligned}$$

Since $l_{\sigma_x^2+\sigma_y^2-2\sigma_\varepsilon^2}$ is less than 0, in order to have a defined $\sqrt{\sigma_x^2 + \sigma_y^2 - 2\sigma_\varepsilon^2}$, we replaced the negative $l_{\sigma_x^2+\sigma_y^2-2\sigma_\varepsilon^2}$ by a tiny positive value, say 0.0001, according to ?'s recommendation. The upper limit for $\sigma_x^2 + \sigma_y^2 - 2\sigma_\varepsilon^2$ is

$$\begin{aligned} u_{\sigma_x^2+\sigma_y^2-2\sigma_\varepsilon^2} &= \widehat{\sigma}_x^2 + \widehat{\sigma}_y^2 - 2\widehat{\sigma}_\varepsilon^2 + \sqrt{(u_{\sigma_x^2+\sigma_y^2} - (\sigma_x^2 + \sigma_y^2))^2 + 4(\widehat{\sigma}_\varepsilon^2 - l_{\sigma_\varepsilon^2})^2} \\ &= (0.0913 + 0.0886) - 2 \times 0.0567 + \\ &\quad \sqrt{(0.238 - (0.0913 + 0.0886))^2 + 4 \times (0.0567 - 0.038)^2} \\ &= 0.136 \end{aligned}$$

Therefore the lower and upper limits for $\theta_2 = \sqrt{\sigma_x^2 + \sigma_y^2 - 2\sigma_\varepsilon^2}$ are given by

$$l_2 = \sqrt{0.0001} = 0.010$$

$$u_2 = \sqrt{0.136} = 0.369$$

Finally, substitute the estimates and confidence limits of parameters into the equations (3.9) and (3.10), we get

$$L_\delta = -1.239$$

Since the lower limit of δ is less than 0, the sign of L_δ needs to be reversed. Therefore, the upper limit of A_c is

$$U_{A_c} = \Phi(1.239) = 0.892$$

Similarly, replacing the corresponding parameters in the formula of U_δ with the estimates, we get

$$U_\delta = -0.189$$

Then the lower limit of A_c is

$$L_{A_c} = \Phi(0.189) = 0.575$$

Therefore, when considering measurement error, the 95% confidence limits using the MOVER approach for A_c are (0.575, 0.892).

When ignoring measurement error, the estimate of A is given by

$$\begin{aligned}
\widehat{A} &= \Phi\left(\frac{-(\overline{y}-\overline{x})}{\sqrt{S_x^2+S_y^2}}\right) \\
&= \Phi\left(\frac{-(0.450-0.604)}{\sqrt{0.0886+0.0913}}\right) \\
&= \Phi(0.363) \\
&= 0.642
\end{aligned}$$

The confidence limits of $\widehat{\delta}$ can be obtained using both the Delta method and the MOVER approach.

First, using the Delta method, the variance of $\widehat{\delta}$ is given by

$$\begin{aligned}
\widehat{\mathrm{var}}(\widehat{\delta}) &= \left[\frac{S_x^2}{n_x}+\frac{S_y^2}{n_y}\right]\times(S_x^2+S_y^2)^{-1}+ \\
&\quad \frac{(\overline{y}-\overline{x})^2}{4(S_x^2+S_y^2)^3}\times\left[\frac{2S_x^4}{n_x-1}+\frac{2S_y^4}{n_y-1}\right].
\end{aligned}$$

Replacing the estimates of parameters in the above formula with the given values, then we get

$$\widehat{\mathrm{var}}(\widehat{\delta}) = 0.0133$$

Then, the lower and upper limits of A are given by

$$\begin{aligned}
L_A &= \Phi(-u_\delta) \\
&= \Phi\left[-\left(\widehat{\delta}+z_{\alpha/2}\times\sqrt{\widehat{\mathrm{var}}(\widehat{\delta})}\right)\right] \\
&= \Phi\left[-\left(-0.363+1.96\times\sqrt{0.0133}\right)\right] \\
&= \Phi(0.137) \\
&= 0.554
\end{aligned}$$

Table 5.1: The estimates and confidence intervals for the area under the ROC curve when considering measurement error (A_c) and ignoring measurement error (A)

measurement error	Methods	$\widehat{\delta}(l_\delta, u_\delta)$	$\widehat{A}(L_A, U_A)$	width for A
consider	The Delta method	-0.597 (-1.060, -0.134)	0.725 (0.553, 0.856)	0.303
	The MOVER	-0.597 (-1.239, -0.189)	0.725 (0.575, 0.892)	0.317
ignore	The Delta method	-0.363 (-0.589, -0.137)	0.642 (0.554, 0.722)	0.168
	The MOVER	-0.363 (-0.603, -0.140)	0.642 (0.556, 0.727)	0.171

$$\begin{aligned} U_A &= \Phi(-l_\delta) \\ &= \Phi\left[-\left(\widehat{\delta} - z_{\alpha/2} \times \sqrt{\widehat{\text{var}}(\widehat{\delta})}\right)\right] \\ &= \Phi\left[-\left(-0.363 - 1.96 \times \sqrt{0.0133}\right)\right] \\ &= \Phi(0.589) \\ &= 0.722 \end{aligned}$$

Hence, when ignoring measurement error, the 95% confidence limits for A using the Delta method are (0.554, 0.722).

Second, using the MOVER approach, the way to obtain the variance of $\widehat{\delta}$ when ignoring measurement error is similar to the case when considering measurement error. The only difference is that, when ignoring measurement error, θ_2 is $\sqrt{\sigma_x^2 + \sigma_y^2}$ rather than $\sqrt{\sigma_x^2 + \sigma_y^2 - 2\sigma_\varepsilon^2}$. Therefore, the estimate and confidence limits of θ_1 are -0.154 and (-0.248, -0.060), respectively. The estimate and confidence limits of $(\sigma_x^2 + \sigma_y^2)$ are 0.180 and (0.149, 0.238), respectively. Then the estimate and confidence limits of θ_2 are $\sqrt{0.180}$, which is 0.424, and $(\sqrt{0.149}, \sqrt{0.238})$, which are (0.386, 0.488), respectively. Applying equations (3.9) and (3.10), the confidence limits of δ are (-0.603, -0.140). Hence, when ignoring measurement error, the corresponding 95% confidence limits for A using the MOVER approach are (0.556, 0.727).

Overall, the estimates and confidence intervals for A_c and A are given in Table 5.1.

Comparing the above estimates and confidence intervals of A_c and A, we knew that the estimate of A is underestimated when ignoring the measurement error. The estimates of δ are the middle points of the corresponding confidence intervals obtained using the Delta method, while they are located close to the right confidence limits obtained using the MOVER approach when considering and ignoring measurement error. This provides evidence that the Delta method forced symmetry on the distribution of $(\overline{y}-\overline{x})/\sqrt{S_x^2+S_y^2-2\widehat{\sigma}_\varepsilon^2}$, but the MOVER recovered the variance estimates at the neighborhood of the confidence limits and reflected the fact of asymmetric distribution. Also, the confidence intervals of A obtained by the two methods when considering the measurement error were both wider than those when ignoring measurement error. The MOVER offered a slightly wider interval width than the Delta method does. This is consistent with the simulation results.

Chapter 6

DISCUSSION

This thesis presents two methods of constructing confidence interval for the area under the ROC curve in the presence of measurement error. The simulation study shows that the MOVER approach outperforms the Delta method in terms of coverage probability, interval width and the symmetry of tail errors, especially in the cases of relatively large measurement error, small sample size and high test accuracy.

The different performances of the two approaches can be explained by the normality assumption. The Delta method assumes that the sampling data for the parameter of interest is normally distributed and enforces symmetric variances even though the sample actually follows a skewed distribution. This leads to a poor performance in constructing confidence interval. However, the MOVER approach recovers the asymmetric variances by using the neighborhood confidence limits and thus yields a good performance.

In fact, the MOVER approach can not only be used to construct a confidence interval, where coverage probability is close to the nominal if the estimate of a parameter is asymmetrically distributed, but it can also be reduced to the traditional confidence interval if the estimate of a parameter is normally distributed. Moreover, the area under the ROC curve is actually a function of a ratio. Apart from the Delta method and the MOVER approach, Fieller's theorem (Fieller, 1954) can be used to construct the confidence interval for a ratio. However, the confidence interval for the area under the ROC curve cannot be constructed using the Fieller's theorem. This is because that Fieller's theorem requires both the numerator and denominator of the ratio to be normally distributions, yet the estimator for δ is a ratio of a normal

variate to a square root of a linear combination of chi-square distributed variates, respectively. Contrary to the normality assumption of Fieller's theorem, the MOVER approach has no such requirement, and the confidence interval of a ratio given by the MOVER can even be reduced to the confidence interval obtained by Fieller's theorem when the ratio is of two normal means (see Appendix B). This suggests that the traditional confidence interval and the interval obtained by Fieller's theorem are only the special cases of the confidence intervals given by the MOVER.

In practice, the problem of erroneously measuring data is very common in many fields, for example there are data errors due to the inaccuracy of an instrument in a hospital laboratory. Simply ignoring the measurement error usually results in underestimated estimators, low test power and even misleading conclusion. Also, the sample size of a test may be small, and the high accuracy of a test is preferred. Hence, it is necessary to get a confidence interval with a coverage probability closed to its nominal level in the cases of large measurement error, small sample size and high test accuracy, where the application of the Delta method is not appropriate. Our work shows that the MOVER is a better confidence interval construction approach in the cases mentioned above. The advantage of using the MOVER to construct confidence interval for a function of parameters is that it only requires the reliable confidence limits of the parameters rather than normality assumption. This makes the MOVER more general than the other methods that require assumptions. The disadvantage of applying the MOVER is that if the parameter of interest is a complex function of several parameters, the calculations might be complicated. However, this can be overcome by the application of computer.

In this thesis, we only discussed independently normally distributed values of a biomarker. Further studies can consider the situations involving skewed distributed values of a biomarker, small sample size data and measurement error measured by an internal reliability study.

First, consider skewed distributed values of a biomarker. In the example, the

original values of x and y were skewed. Schisterman *et al.* (2001) transformed the original values to normally distributed values $x^{-1/2}$ and $y^{-1/2}$ in order to meet the normality assumption. However, the interpretation of the area under the ROC curve, a function of $\overline{x}^{-1/2}$ and $\overline{y}^{-1/2}$, was difficult due to the transformation. To avoid the difficult interpretation, the original skewed data can be used directly. In such cases, the application of the Delta method is not appropriate, but the MOVER approach still can be used since it only requires that x and y are independently distributed and the availability of their confidence limits. When applying the MOVER approach, the confidence limits of the x and y can be obtained by the other confidence interval construction methods first (for example, a nonparametric method), then follow the MOVER approach for the confidence interval of the parameter of interest.

Second, in our simulation study, the smallest sample size is 50. According to the central limit theorem, a sample size of 50 enables the skewed data to approach normality to some extent. This limited the exploration for the generality of the MOVER approach. Therefore, a smaller sample size (for example n=20) should be investigated in further studies.

Third, measurement error measured by an internal reliability study can be considered in the further studies. In this thesis, the variance of measurement error is estimated by an external reliability study that is independent of the main study, in which the observed variances of the biomarker values for the "normal" and "abnormal" are estimated. The implied assumption of using an external reliability study is that the same parameter estimates, such as variance, can be carried over from the external study to the main study without bias (Carroll *et al.*, 1995, p.29). That is, the error distribution is the same in the subjects of both studies. If the assumption does not hold, then estimation bias can be produced when external estimates are carried over to the main study. In such case, if sufficient information in the main study can be obtained to estimate measurement error directly, an internal reliability study that uses the subset of subjects within the main study to conduct the replication can

avoid the transferred bias. Hence, an internal reliability study should be considered in further studies.

BIBLIOGRAPHY

Armstrong, B., White, E. and Saracci, R. (1992). *Principles of Exposure Measurement in Epidemiology.* New York: Oxford University Press.

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology* **12** (4), 387–415.

Boquete, H. R., Sobrado, P. G. V., Fideleff, H. L., Sequera, A. M., Giaccio, A. V., Suarez, M. G., Ruibal, G. F. and Miras, M. (2003). Evaluation of diagnostic accuracy of insulin-like growth factor (igf)-i and igf-binding protein-3 in growth hormone-deficient children and adults using roc plot analysis. *Journal of Clinical Endocrinology and Metabolism* **88** (10), 4702–4708.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology* **3**, 296–322.

Campbell, G. (1994). Advances in statistical methodology for the evaluation of diagnostic and laboratory tests. *Statistics in Medicine* **13** (5-7), 499–508.

Carmines, E. G. and Zeller, R. A. (1979). *Reliability and Validity Assessment*, vol. ser. no. 07-017, of *Sage University paper : Quantitative Applications in the Social Sciences.* Sage Publications, Beverly Hills.

Carroll, R., Ruppert, D. and Stefanski, L. (1995). *Measurement Error in Nonlinear Models.* London: Chapman and Hall.

Coffin, M. and Sukhatme, S. (1996). *Lifetime Data Models in Reliability and Survival Analysis* chapter A parametric approach to measurement errors in receiver operating characteristic studies, pp. 71–75. Dordrecht; Boston: Kluwer Academic.

Coffin, M. and Sukhatme, S. (1997). Receiver operating characteristic studies and measurement errors. *Biometrics* , **53** (3), 823–837.

Cole, S. R., Chu, H. and Greenland, S. (2006). Multiple-imputation for measurement-error correction. *International Journal of Epidemiology* , **35** (4), 1074–1081.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika* , **16** (3), 297–334.

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

Dorfman, D. D. and Alf, E. (1969). Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals-rating-method data. *Journal of Mathematical Psychology* , **6** (3), 487–496.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap.* New York : Chapman and Hall.

Erkanli, A., Sung, M., Costello, E. J. and Angold, A. (2006). Bayesian semi-parametric roc analysis. *Statistics in Medicine* **25** (22), 3905–3928.

Faraggi, D. (2000). The effect of random measurement error on receiver operating characteristic (roc) curves. *Statistics in Medicine* **19** (1), 61–70.

Ferrari, P., Friedenreich, C. and Matthews, C. E. (2007). The role of measurement error in estimating levels of physical activity. *American Journal of Epidemiology* **166** (7), 832–840.

Fieller, E. C. (1954). Some problems in interval estimation. *Journal of the Royal Statistical Society. Series B (Methodological)* **16**, 175–185.

Fletcher, R. H., Fletcher, S. W. and Wagner, E. H. (1988). *Clinical Epidemiology: The Essentials.* 2nd edition, Baltimore: Williams and Wilkins.

Fuller, W. (1987). *Measurement Error Models.* New York: John Wiley and Sons.

Goddard, M. J. and Hinberg, I. (1990). Receiver operator characteristic (roc) curves and non-normal data: an empirical study. *Statistics in Medicine* **9** (3), 325–337.

Goodenough, D., Rossmann, K. and Lusted, L. (1974). Radiographic applications of receiver operationg characteristic (roc) curve. *Diagnostic Radiology* **110**, 89–95.

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

Greiner, M., Pfeiffer, D. and Smith, R. D. (2000). Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. *Preventive Veterinary Medicine* **45** (1-2), 23–41.

Grey, D. R. and Morgan, B. J. T. (1972). Some aspects of roc curve-fitting: normal and logistic models. *Journal of Mathematical Psychology* **9** (1), 128–139.

Han, U. K. and Kim, Y. H. (1998). Determination of class ii and class iii skeletal patterns: receiver operating characteristic (roc) analysis on various cephalometric measurements. *American Journal of Orthodontics and Dentofacial Orthopedics: official publication of the American Association of Orthodontists, its constituent societies, and the American Board of Orthodontics* **113** (5), 538–545.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology* **143** (1), 29–36.

Hoffman, R. M. and Garewal, H. S. (1995). Antioxidants and the prevention of coronary heart disease. *Archives of Internal Medicine* **155** (3), 241–246.

Hsieh, F. and Turnbull, B. W. (1996). Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *Annals of Statistics* **24** (1), 25–40.

Koepsell, T. and Weiss, N. (2003). *Epidemiologic Methods - Studying the Occurrence of Illness.* Number 217, New York: Oxford.

Kuder, G. F. and Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika* **2** (3), 151–160.

Liu, K., Stamler, J. and Dyer, A. (1978). Statistical methods to assess and minimize the role of intra-invididual variability in obscuring the relationship between dietary lipids and serum cholesterol. *Journal of Chronic Diseases* **31** (6-7), 399–418.

Lloyd, C. J. (1998). Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems. *Journal of the American Statistical Association* **93** (444), 1356–1364.

Lloyd, C. J. and Yong, Z. (1999). Kernel estimators of the roc curve are better than empirical. *Statistics and Probability Letters* **44** (3), 221–228.

Lusted, L. (1968). *Introduction to Medical Decision Making.* Springfield, Illinois: Thomas.

Lusted, L. (1971*a*). Decision-making studies in patient management. *New England Journal of Medicine* **284**, 416–424.

Lusted, L. (1971*b*). Signal detectability and medical decision-making. *Science* **171**, 1217–1219.

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

Metz, C. E. (1978). Basic principles of roc analysis. *Seminars in Nuclear Medicine* **8** (4), 283–298.

Metz, C. E. (1986*a*). Roc methodology in radiologic imaging. *Investigative Radiology* **21** (9), 720–733.

Metz, C. E. (1986*b*). *Multiple Regression Analysis: Applications in the Health Sciences* chapter Statistical analysis of ROC data in evaluating diagnostic performance, pp. 365–384. New York: American Institute of Physics.

Metz, C. E., Herman, B. A. and Shen, J. . (1998). Maximum likelihood estimation of receiver operating characteristic (roc) curves from continuously-distributed data. *Statistics in Medicine* , **17** (9), 1033–1053.

Noether, G. E. (1967). *Elements of Nonparametric Statistics.* New York: Wiley.

Peterson, W., Birdsall, T. and Fox, W. (1954). The theory of signal detectability. *Institute of Radio Engineers Transactions* **PGIT-4**, 171–212.

Pham, T. and Almhana, J. (1996). The generalized gamma distribution: its hazard rate and stress-strength model. *IEEE Transactions on Rehabilitation Engineering* **44**, 392–397.

Qin, G. and Hotilovac, L. (2008). Comparison of non-parametric confidence intervals for the area under the roc curve of a continuous-scale diagnostic test. *Statistical Methods in Medical Research* **17** (2), 207–221.

Qin, G. and Zhou, X. . (2006). Empirical likelihood inference for the area under the roc curve. *Biometrics* **62** (2), 613–622.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications.* 2nd edition, New York: Wiley.

Rao, P. S. (1997). *Variance Components: Mixed Models, Methodologies and Applications.* New York: Chapman and Hall.

Reiser, B. (2000). Measuring the effectiveness of diagnostic markers in the presence of measurement error through the use of roc curves. *Statistics in Medicine* **19** (16), 2115–2129.

Schisterman, E. F., Faraggi, D., Reiser, B. and Trevisan, M. (2001). Statistical inference for the area under the receiver operating characteristic curve in the presence of random measurement error. *American Journal of Epidemiology* **154** (2), 174–179.

Shapiro, D. E. (1999). The interpretation of diagnostic tests. *Statistical Methods in Medical Research* **8** (2), 113–134.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology* **3**, 271–295.

Spiegelman, D., Zhao, B. and Kim, J. (2005). Correlated errors in biased surrogates: study designs and methods for measurement error correction. *Statistics in Medicine* **24** (11), 1657–1682.

Swets, J., ed. (1964). *Signal Detection and Recognition by Human Observers: Contemporary Readings.* New York: John Wiley and Sons, Inc.

Swets, J. (1973). The relative operating characteristic in psychology. *Science* **182** (4116), 990–1000.

Swets, J., Tanner, W. J. and Birdsall, T. (1961). Decision processes in perception. *Psychological Review* **68**, 301–340.

Swets, J. A. and Pickett, R. M. (1982). *Evaluation of Diagnostic Systems : Methods from Signal Detection Theory.* New York : Academic Press.

Tanner, W. and Swets, J. (1954). A decision-making theory of visual detection. *Psychological Review* **61**, 401–409.

Thiebaut, A. C. M., Kipnis, V., Schatzkin, A. and Freedman, L. S. (2008). The role of dietary measurement error in investigating the hypothesized link between dietary fat intake and breast cancer - a story with twists and turns. *Cancer Investigation* **26** (1), 68–73.

Thomas, D., Stram, D. and Dwyer, J. (1993). Exposure measurement error: influence on exposure-disease relationships and methods of correction. *Annual Review of Public Health* **14**, 69–93.

Tosteson, T. D., Buonaccorsi, J. P., Demidenko, E. and Wells, W. A. (2005). Measurement error and confidence intervals for roc curves. *Biometrical Journal* **47** (4), 409–416.

Van Erkel, A. R. and Pattynama, P. M. T. (1998). Receiver operating characteristic (roc) analysis: basic principles and applications in radiology. *European Journal of Radiology* **27** (2), 88–94.

Van Meter, D. and Middleton, D. (1954). Modern statistical approaches to reception in communication theory. *Institute of Radio Engineers Transactions* **PGIT-4**, 119–141.

Wald, A. (1950). *Statistical Decision Functions.* New York: John Wiley and Sons, Inc.

Wan, S. and Zhang, B. (2007). Smooth semiparametric receiver operating characteristic curves for continuous diagnostic tests. *Statistics in Medicine* **26** (12), 2565–2586.

Wieand, S., Gail, M. H., James, B. R. and James, K. L. (1989). A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika* **76**, 585–592.

Wolfe, D. A. and Hogg, R. V. (1971). On constructing statistics and reporting data. *American Statistician* **25** (4), 27–30.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer* **3**, 32–35.

Zhou, X. ., Castelluccio, P. and Zhou, C. (2005). Nonparametric estimation of roc curves in the absence of a gold standard. *Biometrics* **61** (2), 600–609+654.

Zou, G. Y. (2008). On the estimation of additive interaction by use of the four-by-two table and beyond. *American Journal of Epidemiology* **168** (2), 212–224.

Zou, G. Y. and Donner, A. (2008). Construction of confidence limits about effect measures: a general approach. *Statistics in Medicine* **27** (10), 1693–1702.

Zou, K. H., Hall, W. J. and Shapiro, D. E. (1997). Smooth non-parametric receiver operating characteristic (roc) curves for continuous diagnostic tests. *Statistics in Medicine* **16** (19), 2143–2156.

Appendix A: Derivation of variance of $\widehat{\delta}$ using the Delta method

In the measurement error case, x_i and y_j are independent normal random variables with means and variances μ_X, $\sigma_X^2 + \sigma_\varepsilon^2$ and μ_Y, $\sigma_Y^2 + \sigma_\varepsilon^2$. Consequently, $\overline{x}$ and $\overline{y}$ are independent normal random variables with means and variances μ_X, $(\sigma_X^2 + \sigma_\varepsilon^2)/n_x$ and μ_Y, $(\sigma_Y + \sigma_\varepsilon^2)/n_y$, respectively. Thus,

$$\widehat{\mu} = \overline{y} - \overline{x} \sim N\left(\mu_Y - \mu_X, \frac{\sigma_X^2 + \sigma_\varepsilon^2}{n_x} + \frac{\sigma_Y^2 + \sigma_\varepsilon^2}{n_y}\right) \tag{A1}$$

and

$$(n_x - 1)\frac{S_x^2}{\sigma_X^2 + \sigma_\varepsilon^2} \sim \chi_{n_x-1}^2, \qquad (n_y - 1)\frac{S_y^2}{\sigma_Y^2 + \sigma_\varepsilon^2} \sim \chi_{n_y-1}^2, \qquad (n_f)\frac{\widehat{\sigma}_\varepsilon^2}{\sigma_\varepsilon^2} \sim \chi_{n_y}^2,$$

are all mutually independent. Also,

$$\widehat{\sigma}^2 = \widehat{\sigma}_X^2 + \widehat{\sigma}_Y^2 = (S_x^2 - \widehat{\sigma}_\varepsilon^2) + (S_y^2 - \widehat{\sigma}_\varepsilon^2).$$

Let

$$\widehat{\delta} = \frac{\widehat{\mu}}{\widehat{\sigma}}.$$

Suppose $\widehat{\mu}$ and $\widehat{\sigma}$ are independent. Use the Delta method to obtain the approximate variance of $\widehat{\delta}$

$$\begin{aligned} \text{var}(\widehat{\delta}) &\approx \left(\frac{\partial \widehat{\delta}}{\partial \widehat{\mu}}\right)^2 \times \text{var}(\widehat{\mu}) + \left(\frac{\partial \widehat{\delta}}{\partial \widehat{\sigma}}\right)^2 \times \text{var}(\widehat{\sigma}) \\ &= \frac{1}{\widehat{\sigma}^2} \times \text{var}(\widehat{\mu}) + \frac{\widehat{\mu}^2}{\widehat{\sigma}^4} \times \text{var}(\widehat{\sigma}), \end{aligned} \tag{A2}$$

where $\text{var}(\widehat{\mu})$ is given in equation $(A1)$. $\text{var}(\widehat{\sigma})$ also can be calculated by the Delta method

$$\begin{aligned}
\text{var}(\widehat{\sigma}) &= \text{var}((\widehat{\sigma}^2)^{\frac{1}{2}}) \\
&\approx \left(\frac{\partial((\widehat{\sigma}^2)^{\frac{1}{2}})}{\partial\widehat{\sigma}^2}\right)^2 \times \text{var}(\widehat{\sigma}^2) \\
&= \frac{1}{4\widehat{\sigma}^2} \times \text{var}(\widehat{\sigma}^2) \\
&= \frac{1}{4\widehat{\sigma}^2} \times \text{var}(S_x^2 + S_y^2 - 2\widehat{\sigma}_\varepsilon^2) \\
&= \frac{1}{4\widehat{\sigma}^2} \times [\text{var}(S_x^2) + \text{var}(S_y^2) + 4\text{var}(\widehat{\sigma}_\varepsilon^2)] \\
&= \frac{1}{4\widehat{\sigma}^2} \times \left[\frac{2(\widehat{\sigma}_X^2 + \widehat{\sigma}_\varepsilon^2)^2}{n_x - 1} + \frac{2(\widehat{\sigma}_Y^2 + \widehat{\sigma}_\varepsilon^2)^2}{n_y - 1} + \frac{8(\widehat{\sigma}_\varepsilon^2)^2}{n_f}\right]. \qquad \text{(A3)}
\end{aligned}$$

Apply equations $(A1)$ and $(A3)$ to equation $(A2)$, then get

$$\begin{aligned}
\text{var}(\widehat{\delta}) \approx\ & \frac{1}{\widehat{\sigma}^2} \times \left[\frac{\widehat{\sigma}_X^2 + \widehat{\sigma}_\varepsilon^2}{n_x} + \frac{\widehat{\sigma}_Y^2 + \widehat{\sigma}_\varepsilon^2}{n_y}\right] + \\
& \frac{\widehat{\mu}^2}{4(\widehat{\sigma}^2)^3} \times \left[\frac{2(\widehat{\sigma}_X^2 + \widehat{\sigma}_\varepsilon^2)^2}{n_x - 1} + \frac{2(\widehat{\sigma}_Y^2 + \widehat{\sigma}_\varepsilon^2)}{n_y - 1} + \frac{8(\widehat{\sigma}_\varepsilon^2)^2}{n_f}\right]. \qquad \text{(A4)}
\end{aligned}$$

Substitute the estimates for the unknown parameters in equation $(A4)$, and it results in

$$\begin{aligned}
\widehat{\text{var}}(\widehat{\delta}) =\ & \left[\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}\right] \times (S_x^2 - \widehat{\sigma}_\varepsilon^2 + S_y^2 - \widehat{\sigma}_\varepsilon^2)^{-1} + \\
& \frac{(\overline{x} - \overline{y})^2}{4(S_x^2 - \widehat{\sigma}_\varepsilon^2 + S_y^2 - \widehat{\sigma}_\varepsilon^2)^3} \times \left[\frac{2S_x^4}{n_x - 1} + \frac{2S_y^4}{n_y - 1} + \frac{8\widehat{\sigma}_\varepsilon^4}{n_f}\right].
\end{aligned}$$

Appendix B: The Reduction of confidence interval for a ratio obtained by the MOVER to that obtained by the Fieller's theorem

The following is the proof of reducing the confidence interval obtained by the MOVER approach to that obtained by Fieller's theorem in the case of the ratio of two normal means.

Suppose that θ_1 and θ_2 are two normal means with confidence limits of (l_1, u_1) and (l_2, u_2), respectively. From equation (3.9), we know the lower limit of a ratio $(\lambda = \theta_1/\theta_2)$ is

$$\begin{aligned}
L_\lambda &= \frac{\hat{\theta}_1\hat{\theta}_2 - \sqrt{\hat{\theta}_1^2\hat{\theta}_2^2 - (2u_2\hat{\theta}_2 - u_2^2)(2l_1\hat{\theta}_1 - l_1^2)}}{2u_2\hat{\theta}_2 - u_2^2} \\
&= \frac{\hat{\theta}_1\hat{\theta}_2 - \sqrt{\hat{\theta}_1^2\hat{\theta}_2^2 - [\hat{\theta}_2^2 - (\hat{\theta}_2^2 - 2u_2\hat{\theta}_2 + u_2^2)][\hat{\theta}_1^2 - (\hat{\theta}_1^2 - 2l_1\hat{\theta}_1 + l_1^2)]}}{\hat{\theta}_2^2 - (\hat{\theta}_2^2 - 2u_2\hat{\theta}_2 + u_2^2)} \\
&= \frac{\hat{\theta}_1\hat{\theta}_2 - \sqrt{\hat{\theta}_1^2\hat{\theta}_2^2 - [\hat{\theta}_2^2 - (u_2 - \hat{\theta}_2)^2][\hat{\theta}_1^2 - (\hat{\theta}_1 - l_1)^2]}}{\hat{\theta}_2^2 - (u_2 - \hat{\theta}_2)^2} \\
&= \frac{\hat{\theta}_1\hat{\theta}_2 - \sqrt{\hat{\theta}_1^2\hat{\theta}_2^2 - \left[\hat{\theta}_2^2 - t^2\left(\frac{u_2-\hat{\theta}_2}{t}\right)^2\right]\left[\hat{\theta}_1^2 - t^2\left(\frac{\hat{\theta}_1-l_1}{t}\right)^2\right]}}{\hat{\theta}_2^2 - t^2\left(\frac{u_2-\hat{\theta}_2}{t}\right)^2} \\
&= \frac{\hat{\theta}_1\hat{\theta}_2 - \sqrt{\hat{\theta}_1^2\hat{\theta}_2^2 - [\hat{\theta}_2^2 - t^2\mathrm{var}(\hat{\theta}_2)][\hat{\theta}_1^2 - \mathrm{t}^2\mathrm{var}(\hat{\theta}_1)]}}{\hat{\theta}_2^2 - t^2\mathrm{var}(\theta_2^2)}.
\end{aligned} \tag{B1}$$

Similarly, the upper limit of a ratio ($\lambda = \theta_1/\theta_2$) from equation (3.10) is

$$\begin{aligned}
U_\lambda &= \frac{\hat{\theta}_1\hat{\theta}_2 + \sqrt{\hat{\theta}_1^2\hat{\theta}_2^2 - (2l_2\hat{\theta}_2 - l_2^2)(2u_1\hat{\theta}_1 - u_1^2)}}{2l_2\hat{\theta}_2 - l_2^2} \\
&= \frac{\hat{\theta}_1\hat{\theta}_2 + \sqrt{\hat{\theta}_1^2\hat{\theta}_2^2 - [\hat{\theta}_2^2 - (\hat{\theta}_2^2 - 2l_2\hat{\theta}_2 + l_2^2)][\hat{\theta}_1^2 - (\hat{\theta}_1^2 - 2u_1\hat{\theta}_1 + u_1^2)]}}{\hat{\theta}_2^2 - (\hat{\theta}_2^2 - 2l_2\hat{\theta}_2 + l_2^2)} \\
&= \frac{\hat{\theta}_1\hat{\theta}_2 + \sqrt{\hat{\theta}_1^2\hat{\theta}_2^2 - [\hat{\theta}_2^2 - (\hat{\theta}_2 - l_2)^2][\hat{\theta}_1^2 - (u_1 - \hat{\theta}_1)^2]}}{\hat{\theta}_2^2 - (\hat{\theta}_2 - l_2)^2} \\
&= \frac{\hat{\theta}_1\hat{\theta}_2 + \sqrt{\hat{\theta}_1^2\hat{\theta}_2^2 - \left[\hat{\theta}_2^2 - t^2\left(\frac{l_2-\hat{\theta}_2}{t}\right)^2\right]\left[\hat{\theta}_1^2 - t^2\left(\frac{\hat{\theta}_1-u_1}{t}\right)^2\right]}}{\hat{\theta}_2^2 - t^2\left(\frac{l_2-\hat{\theta}_2}{t}\right)^2} \\
&= \frac{\hat{\theta}_1\hat{\theta}_2 + \sqrt{\hat{\theta}_1^2\hat{\theta}_2^2 - [\hat{\theta}_2^2 - t^2\mathrm{var}(\hat{\theta}_2)][\hat{\theta}_1^2 - \mathrm{t}^2\mathrm{var}(\hat{\theta}_1)]}}{\hat{\theta}_2^2 - t^2\mathrm{var}(\theta_2^2)}. \qquad \text{(B2)}
\end{aligned}$$

On the other hand, the two roots for the equation of Fieller's theorem

$$(\hat{\theta}_1^2 - t^2\mathrm{var}(\hat{\theta}_1)) - 2\lambda(\hat{\theta}_1\hat{\theta}_2 - \mathrm{t}^2\mathrm{var}(\hat{\theta}_1\hat{\theta}_2)) + \lambda^2(\hat{\theta}_2^2 - \mathrm{t}^2\mathrm{var}(\hat{\theta}_2)) = 0$$

are given by

$$\frac{(\hat{\theta}_1\hat{\theta}_2 - t^2\mathrm{var}(\hat{\theta}_1\hat{\theta}_2)) \pm \sqrt{(\hat{\theta}_1\hat{\theta}_2 - \mathrm{t}^2\mathrm{var}(\hat{\theta}_1\hat{\theta}_2))^2 - (\hat{\theta}_2^2 - \mathrm{t}^2\mathrm{var}(\hat{\theta}_2))(\hat{\theta}_1^2 - \mathrm{t}^2\mathrm{var}(\hat{\theta}_1))}}{\hat{\theta}_2^2 - t^2\mathrm{var}(\hat{\theta}_2)}. \qquad \text{(B3)}$$

When $\hat{\theta}_1$ and $\hat{\theta}_2$ are independent (i.e. $\mathrm{var}(\hat{\theta}_1\hat{\theta}_2) = 0$), the formulae in $(B3)$ can be reduced to formulae $(B1)$ and $(B2)$, respectively. This shows that the confidence limits for the ratio of two means by the MOVER approach can be reduced to the confidence limits yielded by Fieller's theorem if the two means are normally and independently distributed.