VARIANTS OF THE ROC CURVE WITH APPLICATIONS TO META-ANALYSIS

VARIANTS OF THE RECEIVER OPERATING CHARACTERISTIC (ROC) CURVE WITH APPLICATIONS TO META-ANALYSIS

By

ROXANNE FILL, B.Sc.

A Thesis

Submitted to the School of Graduate Studies

in Partial Fulfilment of the Requirements

for the Degree

Master of Science

McMaster University

MASTER OF SCIENCE (2007) McMaster University

(Statistic) Hamilton, Ontario

TITLE:	Guidelines for the Partial Area under the Summary Receiver Operating Characteristic (SROC) Curve
AUTHOR:	Roxanne Fill, B.Sc
SUPERVISOR:	Professor Peter Macdonald
NUMBER OF PAGES:	xiii, 99

Abstract

The accuracy of a diagnostic test is often evaluated with the measures of sensitivity and specificity and the joint dependence between these two measures is captured by the receiver operating characteristic (ROC) curve. To combine multiple testing results from studies that are assumed to follow the same underlying probability law, a smooth summary receiver operating characteristic (SROC) curve can be fitted. Moses et al. (1993) proposed a least squares approach to fit the smooth SROC curve.

In this thesis we overview the summary measures for the ROC curve in single study data as well as the summary statistics for the SROC curves in meta-analysis. These summary statistics include, the area under the curve (AUC), Q^* statistic, area swept under the curve (ASC) and the partial area under the curve (pAUC).

Our focus, however is mainly on the partial area under the SROC curve as it is being used frequently in meta-analysis of diagnostic testing. The appeal to use the pAUC instead of the full AUC is that the partial area can be used to focus

on a clinically relevant region of the SROC curve where false positive rate (FPR) is small. Simulations and considerations for the use of the summary indices of the ROC and SROC curves are presented here.

Key Words: receiver operating characteristic (ROC) curve; summary receiver operating characteristic (SROC) curve; meta-analysis; area under the curve (AUC); Q^* statistic; area swept out by the curve (ASC); partial area under the curve (pAUC); homogeneous; heterogeneous

Acknowledgments

I would first like to thank my supervisor, Dr. Peter Macdonald, whose supervision, guidance, patience, inspiration and teaching encouraged me to make this thesis possible and enjoyable.

An appreciation and gratitude is needed for Dr. Stephen Walter whose expertise and insight on the topic helped set a path for this thesis

I would also like to thank Dr. Hoppe and Dr. Thabane for serving on my supervisory committee and giving valuable advice.

I wish to sincerely thank Dr. Roman Viveros, Dr. Angelo Canty, Dr. Aaron Childs, Dr. N. Balakrishnan, Dr. Fred Hoppe, Dr. Ernest Mead, Dr. Ron Zhu and Dr. Shui Feng for their teaching and help in my B.Sc and M.Sc studies.

I would like to thank all my friends for their friendships, love, encouragement, help and humour during the last year and a half.

Finally, I would like to thank my parents, John and Jerry, my sister, Natalka and my boyfriend, Luke, whose love and patience encouraged me to finish this thesis.

Contents

List of Tables

List of Figures

Table of Acronyms

ASC	area swept out by the curve
AUC	area under the curve
CI	confidence interval
DOR	diagnostic odds ratio
FNCH	Fisher's noncentral hypergeometric
FNR	false negative rate
FPR	false positive rate
MW	Mann-Whitney (U-statistic)
OR	odds ratio
pAUC	partial area under the curve
pAUC*	scaled partial area under the curve
ROC	receiver operating characteristic
SE	standard error
SROC	summary receiver operating characteristic
TNR	true negative rate
TPR	true positive rate
WNCH	Wallenius' noncentral hypergeometric
2D	two-dimensional
3D	three-dimensional

Chapter 1

Introduction

The pooling of receiver operating characteristic (ROC) curves, which combines evidence from independent studies examining the diagnostic value of test results on a continuous scale originated in 1990. Techniques have been described for combining sensitivities and specificities of studies on tests with separate outcomes. These techniques use different analysis for sensitivity and specificity, as well as a linear model on a logit scale relating the two measures.

The summary receiver operating characteristic (SROC) analysis is applied to data from diagnostic tests which have been pooled from multiple sources. This method is used since simple averages can produce misleading results if the data sets vary between each other in terms of size or study quality (Rutter & Gatsonis, 2001). Poorly conducted or reported studies are more likely to produce outlying results,

which skew the overall pooled data. A weighed average can be biased towards large studies or studies comprised of very similar results. It can be difficult to identify outlying data and exclude it. On the other hand, more data mean wider conclusions can be reached. The SROC analysis deals with pooled data without these pitfalls (Jones et al., 2005).

The area under an SROC curve (AUC) is often used to summarize the diagnostic performance described by an entire SROC curve. The value of the AUC index can be interpreted as the average value of the true positive rate (TPR) over all possible values of the false positive rate (FPR) between 0 and 1.

The AUC index, however, may not be a relevant measure of a diagnostic performance in some situations. In these cases, summary measures such as Q^* statistic, which is the value of TPR at the point where sensitivity equal specificity, the area swept out by the curve (ASC), and the partial area under the SROC curve (pAUC) are recommended.

The clinical applications of some diagnostic tests demand high sensitivity. For these tests, only those study points on an SROC curve that have high sensitivity values are clinically acceptable. Therefore, the AUC index, which summarizes an entire SROC curve by giving equal weight to the study points at all sensitivity levels, does not measure diagnostic performance meaningfully from a clinical perspective in such situations.

An easiest way to understand the SROC curve is to look first at the ROC curve. The ROC curve and its indices: AUC, Q^*, ASC and pAUC statistics are discussed in Chapter 2. The homogeneous logistic and normal threshold models are compared. The homogeneous logistic threshold model is shown to generate constant ORs for over all thresholds. The comparison of the AUC and Q^* statistics are also shown. The Q^* statistic does not appear to be any more useful than the AUC in the cases where clinically relevant regions are of importance. Lastly, the ASC summary measure and it's unsatisfactory behaviour for the step ROC function are evaluated.

In Chapter 3, various properties of the SROC curve are discussed. We review the AUC, Q^*, ASC, pAUC statistics and their standard error formulas proposed by Moses et al. (1993) and Walter (2002). Chapter 4 assesses the standard errors for the AUC, Q^* and partial AUC SROC indices and compares these estimates to a bootstrapping procedure. Simulations for study design with varying study sample sizes with focus on the partial area are also presented in Chapter 4. A practical example is described in Chapter 5, based on the diagnostic performance of two magnetic resonance angiography techniques: 3D gadolinium-enhanced (3D-GD) and 2D time of flight (2D-TOF) for detecting peripheral arteriosclerotic occlusive disease (Nelemans et al., 2000). Further points of discussions are considered in Chapter 6, including considerations and general observations for the use of the summary indices of the ROC and SROC curves.

Chapter 2

Single ROC Analysis

2.1 Motivation

A two-by-two table giving the probabilities of positive and negative test results for subjects with or without disease (Table 2.1) is a standard way of describing the performance of a diagnostic test. A diagnostic test is any kind of medical procedure performed to aid in the diagnostic or detection of a disease. The test can be used to calculate the probability a subject has a disease under consideration given a certain test result. The ideal diagnostic test should differentiate diseased and healthy individuals with out any errors.

There are two questions that occur when evaluating diagnostic tests. First, whenever a diagnostic test is subjected to study, the question arises as to how sure

the experimenter can be that the control subjects are indeed healthy and that all the disease subjects actually have a disease (Youden, 1950). The second question deals with two types of errors a diagnostic test may make: false positives and false negatives. When a false positive is made, the price of retesting can become quite costly, let alone the emotional shock the subject would occur. However, if a false negative is made the subjects maybe harmed if treatment is deferred until too late (Youden, 1950).

The probabilities in Table 2.1, add to one in each column. The ideal test would show true positive rate (TPR) and true negative rate (TNP) both equal to one, with false negative rate (FNR) and false positive rate (FPR) both equal to zero (Moses et al., 1993).

Table 2.1: Summary of test performance probabilities for subjects with disease and subjects without disease

		With disease	Without disease.
Test	+	TPR	FPR
Outcome	−	FNR	TNR
Sum		1	1

An alternative to this probability table would be where the entries are counts. For example, the counts could be the number of subjects rather than the test performance probabilities as shown in Table 2.2. In this table, the ideal test would be when b and c were both zero, making a equal to n_1 and d equal to n_2.

Table 2.2: Frequencies from applying the diagnostic test in n_1 subjects with disease and n_2 subjects without disease.

		With disease	Without disease
Test	+	a	c
Outcome	−	b	d
Sum		n_1	n_2

2.2 Receiver Operating Characteristic Curve

Sensitivity (or TPR) is the proportion of subjects with the disease who test positive. Specificity (or 1-FPR) is the proportion of subjects without the disease who test negative. Sensitivity and specificity is a pair of statistics that together measure the performance of a diagnostic test. The joint dependence of TPR and FPR is captured fully in the Receiver Operating Characteristic (ROC) curve (Moses et al., 1993). The ROC curve is a well established method of summarizing the performance of a diagnostic test (Walter, 2002). It indicates the relationship between TPR and FPR of the test at various thresholds used to distinguish disease cases from non-disease cases. The curve is a plot of the TPR versus the FPR. The points of the curve are obtained by sweeping the classification threshold from the most positive classification value to the most negative. In other words, an ROC curve is a path in the unit square, rising from the lower left corner, where both TPR and FPR are zero, to the upper right corner, where they are both one. Points near the lower left

corner of the ROC plot correspond to conservative thresholds and the points near the upper right corner correspond to moderate thresholds.

If a test could perfectly differentiate from disease and control, it would have a score above which the entire disease population would fall and below all non-disease scores (Delong et al., 1988). The curve would then pass through the point (0,1) on the unit square. The closer an ROC curve comes to this ideal point, the better its discriminating ability. A test with no discrimination ability will produce a curve that follows the diagonal line from (0,0) to (1,1) (Delong et al., 1988).

Although, the whole ROC curve is informative, summary indices are always helpful. For example, indices are needed when the performances of two indicators for diagnosing a particular disease are to be compared and neither of the two corresponding ROC curves dominates the other. From the ROC curves, a particular value of specificity may attain higher sensitivity than the other values, whereas at another specificity, the reverse maybe true. Therefore, it is difficult to say which indicator is better. This problem of in-comparability can be avoided if the comparison is based on an index that summarizes the ROC curve (Lee & Hsiao, 1996).

2.3 The Homogeneous Logistic Threshold Model

In the context of ROC curves for single studies, the homogeneous logistic threshold model is equivalent to assuming two equivariant logistic distributions, for true cases

and non-cases. The implication is that disease differs from non-disease only by a location shift. (Van Der Schouw et al., 1994). In the heterogeneity situation, the two distributions also differ by a scale parameter, implying different variances (Van Der Schouw et al., 1994). The logistic distribution also gives a reasonably close approximation to normally distributed test results. Therefore, comparable estimates for the ROC curve and its summary measures of normally distributed data can be derived (Van Der Schouw et al., 1994).

Assume we have two subpopulations with different logistic or normal distributions of the score variable with the same standard deviations but different means (homogeneous case). Logistic distributions are unique in giving a constant OR for any threshold (Van Der Schouw). An estimated OR will depend on what threshold is chosen. The mean area under the ROC curve (AUC) and its standard error can be generated by repeated sampling of the two selected logistic or normal distributions for both the disease and control groups. When there is no overlap between the disease and control the AUC is one, so the test is perfect.

Suppose the mean and standard deviation of the test results for a subject without the disease are μ_1 and σ_1, respectively. A monotone transformation of the test results is given by:

$$Y' = \left[\frac{Y - \mu_1}{\sigma_1}\right] \frac{\pi}{\sqrt{3}}. \tag{2.1}$$

The distribution of the test results for a non-disease individual is

$$F_1(Y) = \frac{e^Y}{1+e^Y}, \quad \text{for} \quad Y \in \Re \tag{2.2}$$

and the distribution of the transformed test results for a disease individual is

$$F_2(X) = \frac{e^{aX-\theta}}{1+e^{aX-\theta}}, \tag{2.3}$$

for $a > 0$ and $\theta > 0$. The relationship between FPR and TPR at a cutoff point, k are

$$\text{FPR}_k = Pr(Y > k) = \frac{1}{1+e^k} \tag{2.4}$$

$$\text{TPR}_k = Pr(X > k) = \frac{1}{1+e^{ak-\theta}}, \tag{2.5}$$

respectively. From equation 2.4, k is

$$k = \ln\left(\frac{1-\text{FPR}_k}{\text{FPR}_k}\right). \tag{2.6}$$

(Van Der Schouw et al., 1994).

The test results from two logistic distributions with unequal variances determine

the equation for the shape of the ROC curve as:

$$\mathrm{TPR}_k = \frac{(\mathrm{FPR}_k)^a}{\mathrm{FPR}^a + (1 - \mathrm{FPR}_k)^a e^{-\theta}}, \tag{2.7}$$

for FPR $\in (0, 1)$, $a = \frac{\sigma_1}{\sigma_2}$ and $\theta = \ln\left[\frac{\mathrm{TPR}/(1-\mathrm{TPR})}{(\mathrm{FPR}/(1-\mathrm{FPR}))^a}\right]$. In the homogeneous case a would equal one.

2.4 Area Under the ROC Curve

Summary indices associated with the ROC curve can measure the overall accuracy of a test. The most familiar index is the area under the ROC curve (AUC). The AUC is the probability that a randomly selected disease individual has a higher score on the test than a randomly selected control person. This assumes that the disease have (on average) a higher score than the non-disease.

Let X and Y denote the diagnostic marker measurements for disease and control subjects, respectively. Bamber (1975) showed that AUC = Prob($X < Y$). Since the entire ROC curve is defined within a unit squared, AUC varies between zero and one. The values of AUC close to one indicate that the marker has high diagnostic accuracy and a test with a AUC equal to one is perfectly accurate. A test with no discrimination ability would have an AUC = 0.5.

In this section we review a non-parametric approach using the Wilcoxon Mann-

Whitney (MW) two-sample U-statistic for estimating the AUC. This approach follows from estimating the ROC curve as a step-function based on empirical cumulative distribution functions. It can be shown that the AUC for the empirical ROC curve, when calculated by the trapezoidal rule, is equal to the MW statistic applied to the two sample $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ (Bamber, 1975; Hanley & McNeil, 1982). Since the MW statistic is a generalized U-statistic, statistical analysis regarding the performance of a diagnostic test can be performed by utilizing the general theory for U-statistics. The standard error (SE) of the AUC estimate can be found by using the Bamber (1976) and Hanley & McNeil (1982) approach.

Let $A(X, Y)$ denote the AUC for X and Y as computed by the trapezoidal rule, then

$$A(X, Y) = P(X < Y) + \frac{1}{2}P(X = Y). \tag{2.8}$$

Note that $P(X = Y) = 0$ when X and Y are continuous. The Wilcoxon Mann-Whitney U-statistic is defined as being the total number of (X, Y) pairs in which $X < Y$. From this definition is can be seen that $A(X, Y)$ and MW U-statistic are closely related (Bamber, 1975). Thus, if X and Y are continuous

$$\widehat{AUC}(X, Y) = \frac{1}{mn}\sum_{1}^{m}\sum_{1}^{n} S(X_i, Y_j), \tag{2.9}$$

where

$$S(X_i, Y_j) = \begin{cases} 1 & Y_j < X_i \\ \frac{1}{2} & Y_j = X_i \\ 0 & Y_j > X_i. \end{cases} \tag{2.10}$$

The standard error for the nonparametric estimate of AUC can be calculated with the formula of Hanley & McNeil (1982):

$$\mathrm{SE}(\widehat{\theta}) = \sqrt{\frac{\theta(\theta-1) + (m-1)(Q_1 - \theta^2) + (n-1)(Q_2 - \theta^2)}{mn}}, \tag{2.11}$$

where $\theta = P(X_i < Y_j)$, the observed AUC. The quantile, $Q_1 = P(X_i > Y_j, X_k > Y_j)$, is the probability that two randomly chosen disease individuals will both be ranked higher than a randomly chosen control individual. The quantile, $Q_2 = P(X_i > Y_j, X_i > Y_l)$, is similar, with the probability that one randomly chosen disease person will be ranked higher than two randomly chosen non-disease subjects. Q_1 and Q_2 can be approximated (Hanley & McNeil, 1982) and expressed as

$$\begin{aligned} \hat{Q}_1 &= \frac{\theta}{2-\theta} \\ \hat{Q}_2 &= \frac{2\theta^2}{1+\theta}. \end{aligned} \tag{2.12}$$

When these two equations are substituted into equation (2.11), the SE can be expressed at any level of θ and the values of m and n can vary until $\mathrm{SE}(\widehat{\theta})$ is sufficiently small (Hanley & McNeil, 1982). The SE's are smallest for very high (close to one)

AUC values.

An alternative to the exact computation of the variance is to determine instead the maximum of the variance over all possible continuous distributions with the same expected value of the AUC and its variance.

Let $\mathrm{SE}(\hat{\theta}_{\max})$ denote the maximum possible value of $\mathrm{SE}(\hat{\theta})$ for fixed AUC and fixed sample sizes m and n (Bamber, 1975). For each combination of X and Y, $\mathrm{SE}(\hat{\theta})$ is calculated for fixed m and n. The largest $\mathrm{SE}(\hat{\theta})$ obtained is denoted as $\mathrm{SE}(\hat{\theta}_{\max})$.

$$\mathrm{SE}(\hat{\theta}_{\max}) = \sqrt{\frac{\theta(1-\theta)}{\min\{m,n\}}} \leq \sqrt{\frac{1}{4\min\{m,n\}}} \tag{2.13}$$

The area under the ROC curve for the logistic threshold model is

$$\mathrm{AUC} = \int_0^1 \frac{(\mathrm{FPR}_k)^a}{\mathrm{FPR}^a + (1-\mathrm{FPR}_k)^a e^{-\theta}} \delta x \tag{2.14}$$

for unequal variances ($a \neq 1$) (Van Der Schouw et al., 1994), and

$$\mathrm{AUC}_{\mathrm{hom}} = \frac{1 - e^{\theta} - \theta e^{-\theta}}{(1-e^{-\theta})^2} \tag{2.15}$$

for equal variances ($a = 1$) (Van Der Schouw et al., 1994). An alternative expression

for the homogeneous AUC is

$$\mathrm{AUC}_{\mathrm{hom}} = \frac{\mathrm{OR}}{(\mathrm{OR}-1)^2}[(\mathrm{OR}-1) - \ln(\mathrm{OR})] \tag{2.16}$$

where OR is the constant odds ratio. An approximate variance for $\widehat{\mathrm{AUC}}_{\mathrm{hom}}$ is

$$\mathrm{SE}(\widehat{\mathrm{AUC}}_{\mathrm{hom}}) = \frac{\mathrm{OR}}{(\mathrm{OR}-1)^3}[(\mathrm{OR}+1)\ln \mathrm{OR} - 2(\mathrm{OR}-1)]\mathrm{SE}(\hat{a}). \tag{2.17}$$

(Walter, 2002).

2.4.1 Examples

A comparison of the logistic and normal threshold model as well as examples demonstrating the behaviours of these models will be discussed in this section. In the two examples below, we will see that the logistic threshold model is quite variable when the location parameters are further apart in the disease and non-disease cases.

For example 1, we generated $n_1 = 50$ samples from the disease group that gives diagnostic scores which follow a logistic distribution with location = 3 and scale = 1 and another $n_2 = 50$ samples from the control group which is also logistic but with location = 0 and scale = 1. The mean AUC and standard error of AUC was 0.88 and 0.034, respectively.

In example 2, we showed what happens when the groups are a bit further apart

and have varying unequal sample sizes. The location parameters are 4 with sample size $n_1 = 50$ and 0 with sample size $n_2 = 500$ for the disease group and the control group, respectively. The scale parameters were 1 for both groups. The corresponding mean AUC was 0.94 with standard error of 0.018. In this example, the AUC is larger and standard deviation is smaller when compared to example 1. We can conclude that the further apart the two distributions are the smaller the standard errors will be. When there is less or no overlap between disease and control groups the ROC curve will be close to or always "perfect" with a right-angled at the top left corner and the AUC equal to one, no matter what the actual data. Similar results follow when using the normal threshold model.

2.5 Odds Ratio

When the scores are from a logistic distribution the true odds ratio (OR) can be calculated from the parameters and the estimated OR will depend on what threshold is chosen (Van Der Schouw et al., 1994). The cumulative distribution function (*cdf*) and the probability density function (*pdf*) for a logistic distribution are

$$F(x) = \frac{1}{(1 + \exp(-(k - \mu)/\sigma))} \tag{2.18}$$

and

$$f(x) = \frac{e^{-(k-\mu)/\sigma}}{\sigma(1+e^{-}(k-\mu)/\sigma)^2}, \tag{2.19}$$

respectively. With μ denoted as the mean, $\sigma > 0$, as the scale parameter and k as the threshold or cutoff point. From the *cdf* it is easy to show that no matter what threshold is used to classify disease and control, the logistic distribution gives an OR equal to the following formula

$$\text{OR} = \exp\left(\frac{\mu_2}{\sigma_2} - \frac{\mu_1}{\sigma_1}\right). \tag{2.20}$$

When the logistic distributions for the disease cases and control cases have equal variances, the OR is constant for all true positive and false positive rates as shown in Fig. 2.1. When the scores are normally distributed in the homogeneous case the ORs vary with threshold (Fig. 2.1).

Fig. 2.1 shows that the estimated ORs are emphasized in the tails of the distributions. There is a difference in the homogeneous case when the scores come from a normal distribution or a logistic distribution. Further exploration into the homoscedastic logistic threshold model is needed to identify if the model is not robust enough to apply when the scores follow some other distributions.

From examples 1 and 2 the true odds ratios are $\exp(3-0) = 20.1$ and $\exp(4-0) = 54.6$, respectively. While the estimated OR are defined by the intersection of the

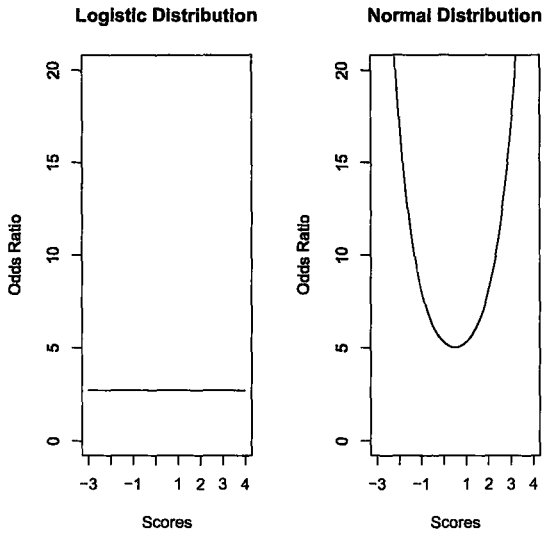

Figure 2.1: Comparison of the ORs for the two logistically distributed (right) and the two normally distributed (left) scores with means $\mu_1 = 1$ and $\mu_2 = 0$ and scale parameter $\sigma_1 = \sigma_2 = 0$.

distributions and the thresholds.

To observe the random samples from the two logistic distributions in example 1, we picked three thresholds: ($k = 1, 3$ & 5), we then set up the 2×2 tables and computed the corresponding ORs (Table 2.3). When n was small the ORs differed with varying threshold but when n was large the ORs became less sensitive to the thresholds chosen. Also, when the difference in means stayed the same and the sample size increased, the observed OR became more accurate to the true OR. In the example, when $n = 50$ the ORs bounced from $11.7 - 32.7$ with a $k = 1$ where as, when $n = 5000$ the observed ORs bounced from $18.9 - 21.8$ with the same threshold. As the sample size increased the observed ORs became more precise and

stayed closer to the true odds ratio, exp(3) = 20.1.

Table 2.3: OR values with varying thresholds of randomly generated logistic distribution with the disease group have mean = 3 and scale = 1, while the control group being the standard logistic distribution (mean = 0, scale = 1).

Sample size	Threshold	OR
50	5	0.24000
	3	8.32759
	1	17.11111
	1	23.14286
	1	18.61364
	1	32.73077
	1	11.71429
500	5	12.26374
	3	14.05115
	1	21.82232
	1	16.73057
	1	21.89079
	1	26.40650
	1	17.48782
5000	5	22.02351
	3	19.89667
	1	19.92016
	1	21.79255
	1	18.96111
	1	20.30981
	1	19.61054

2.5.1 Non-Central Hypergeometric Models

When evaluating 2 × 2 tables that had fixed row totals and a specified theoretical OR, say exp(3), the observed OR varied around that total. We simulated this

using both the Fisher's and Wallenius' noncentral hypergeometric distribution and observed that the larger the row totals and the larger the sample sizes the more accurate the observed OR would be.

The Fisher's noncentral hypergeometric (FNCH) distribution is obtained if the samples are taken independently of each other, where as the Wallenius' noncentral hypergeometric (WNCH) distribution is obtained if the sample size, n subjects are taken one by one (Fog, 2007). Each draw depends on the previous draws which would imply a competition between the individual, n subjects. FNCH would have no such dependence between draws. In this case, n would be a random variable and the Fisher's distribution is a conditional distribution which can only be determined after the experiment, when n is known. The unconditional distribution is two independent binomials for the FNCH distribution (Fog, 2007).

FNCH distribution is used mainly for statistical tests in contingency tables. The WNCH distribution is used in models of natural selection and biased sampling (Fog, 2007). The difference between FNCH and WNCH distributions are negligible when the OR is close to 1 and n is low compared to N, where N is the total number of subjects. The difference between the two distributions become meaningful when the ORs are high and n is near N. In our case, since n was randomly selected and samples were taken independently it is sufficient to use the Fisher's noncentral hypergeometric distribution when fixing the row totals and specifying an OR.

2.6 Q^* Statistic

One criticism of using the summary statistic AUC is that it depends largely on an irrelevant region. For this reason, one alternative would be to use the point of intersection on the ROC curve with the line FPR + TPR = 1, with slopes from the (0,1) corner to the (1,0) corner (Moses et al., 1993). At that intersection, sensitivity equals specificity and their common value is know as Q^*. Similar to the AUC, the Q^* is an indicator of how closely the ROC curve is to the upper-left corner (Moses et al., 1993). Moreover, the Q^* statistic is defined as a point of indifference on the ROC curve, where the probabilities of incorrect test results are equal for disease and non-disease cases. Thus, it represents the diagnostic threshold at which the probability of a correct diagnosis is constant for all subjects (Walter, 2002).

The Q^* statistic for the homoscedastic logistic threshold model is

$$Q^*_{\text{hom}} = \frac{\sqrt{\text{OR}}}{1+\sqrt{\text{OR}}}, \tag{2.21}$$

with an approximate standard error from the delta method of

$$\text{SE}(\widehat{Q^*}_{\text{hom}}) = \frac{1}{2\sqrt{\text{OR}}(\sqrt{\text{OR}}+1)^2}\text{SE}(\widehat{\text{OR}}) \tag{2.22}$$

(Water, 2003).

For the two logistically distributed scores in example 1 with the true OR of

exp(3) has Q^*_{hom} statistic equal to 0.82 and $SE(Q^*_{\text{hom}})$ equal to 0.004.

To understand the strong relationship between the AUC and Q^* statistics we used a nonparametric bootstrap of samples with replacement from the Henley & McNeil (1982) data. The AUC had a mean of 0.89 and the Q^* statistic had a mean of 0.82. The standard errors for AUC and Q^* were 0.030 and 0.033, respectively. A plot of the 1000 randomly generated AUC and Q^* statistics is shown in Fig. 2.2. Although, the AUC depends largely on an irrelevant region, the straight line indicates a high positive correlation between the two summary statistics ($r = 0.89$). Both summary measures show how closely the ROC curve is to the upper left hand corner in the unit space. To conclude, the use of the Q^* statistic over the AUC shows no increase in information in regards to a diagnostic test.

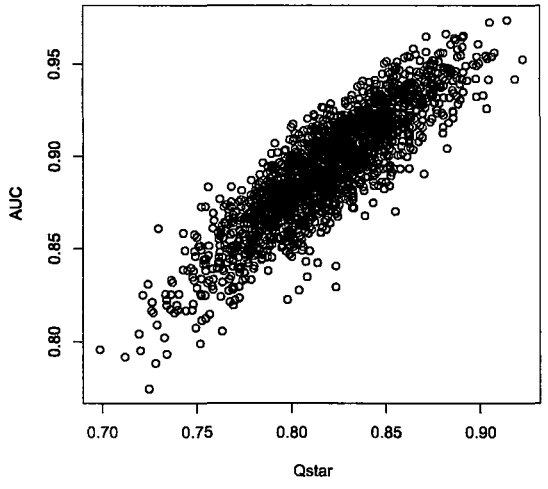

Figure 2.2: AUC vs. Q^* statistics using a nonparametric bootstrap of 1000 samples with replacement from the Henley & McNeil (1982) data.

2.7 Area Swept out by the ROC curve

Another summary index for the ROC curve, proposed by Lee & Hsiao (1996), is the area swept out by the ROC curve (ASC). They state without proof that ASC = probability of correctly diagnosing a pair of diseased *vs* non-diseased (low *vs* high) + probability of correctly diagnosing a pair of diseased *vs* non-diseased (high *vs* low), which gives a useful interpretation for this index and a justification for its use. Zhang (2004) has studied the properties of the ASC in more detail and given examples. The ASC is defined geometrically. Imagine a ray starting from the origin (0,0) to each point in the ROC curve. As the point moves from the origin to the right-uppermost point (1,1), the ray will sweep out some areas. The total of the areas swept out in this way is the ASC statistic (Lee et al., 1996). Note that if these emanating rays sweep out some regions more than once, those areas are counted repeatedly (Lee & Hsiao, 1996). In Fig. 2.3, the ROC curve consists of three line segments. Since the region B is swept out twice, the ASC in this example is $A + 2B + C$.

For a test with no diagnostic value, the ROC curve lies entirely on the diagonal and the swept out areas are zero. If the ROC curve strays from the diagonal at any point, ASC becomes positive. The maximum value ASC can attain is 1/2.

The definition of the ASC applies when a parametric ROC curve for a logit-threshold model is used or for an empirical ROC curve that is defined by a small

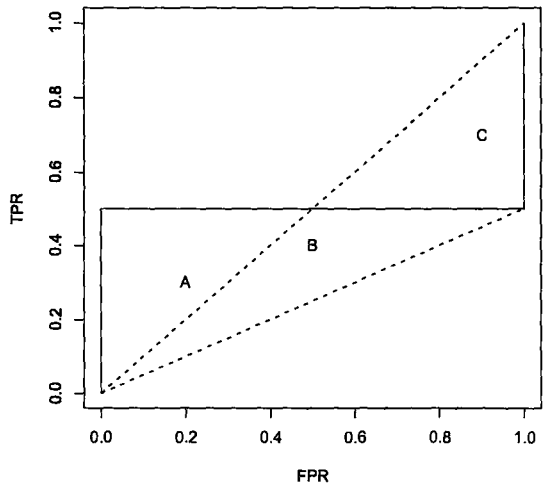

Figure 2.3: ROC curve demonstrating the calculation of the ASC index.

number of line segments. However, when we defined the empirical ROC curve as a step function with runs and rises defined by the points when sensitivities and specificities change values. The difference is most obvious when the scores in both groups are on a continuous scale so there are no tied scores within or between the groups. In this case, the ROC is a step function, stepping up when there is a score in the disease group and across when there is a score in the control group. The tip of the ray defining the ASC statistic then runs alternatively up a vertical segment and across a horizontal segment. When it sweeps across a horizontal segment, it sweeps a triangle that is half the area of the rectangle under that segment. When it sweeps up a vertical segment, it sweeps a triangle that is half the area of the rectangle to the left of the segment, so letting PQR be "area beside the curve" (i.e. to the left)

shown in Fig. 2.4. In this example the ASC = area(ΔPQR) + area(ΔPQS) + area(ΔPTS) + area(ΔPTU) + area(ΔPVU) = 1/2. This applies to any step ROC curve mo matter how many steps there are in the curve. Even with both disease and control groups large, given a relatively smooth curve the $ASC = 1/2$ (Fig. 2.5). The ASC statistic does not carry useful information when dealing with a step ROC curve.

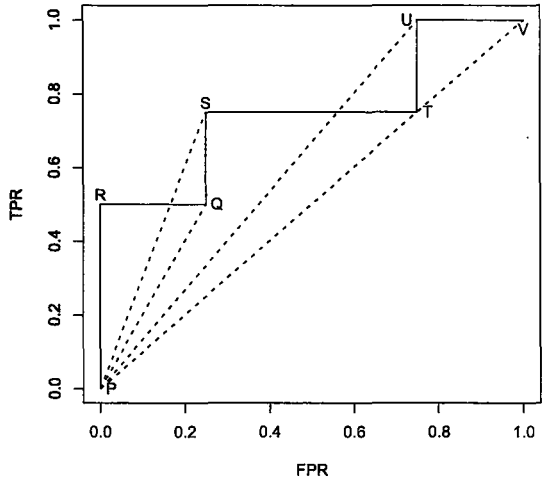

Figure 2.4: ROC curve as a step function demonstrating that the $ASC \equiv 1/2$.

2.8 Partial Area Under the ROC Curve

The partial area under the ROC curve (pAUC) is a summary measure of the ROC curve used to make statistical inference when only a region of the ROC space is of interest. The AUC summarizes across all thresholds and is the most commonly

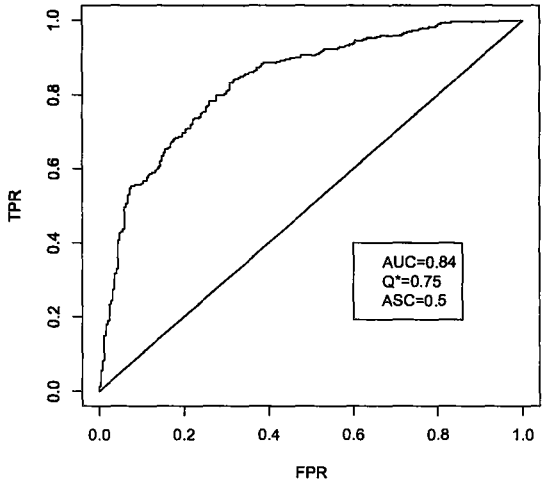

Figure 2.5: Step ROC curve for large disease and control groups.

used measure of diagnostic accuracy for quantitative tests. However, the AUC summarizes test performance over regions of the ROC space in which one would never operate (Dodd & Pepe, 2003). In diagnostic testing, it is critical to maintain a high TPR in order not to miss detecting subjects with disease. In this case, interest is in the region of the ROC curve corresponding only to acceptable high TPR values.

If the attention is focused on a limited range of FPR values, then the AUC statistic becomes irrelevant as a summary measure of the data. An alternative would be to use the pAUC summary measure. The pAUC can be thought of as the probability that a disease and control pair of test results will be correctly ranked, conditional on the disease value falling within the restricted range of the curve

(Walter, 2004).

The "smooth" partial AUC is defined as

$$\text{pAUC}(\theta) = \int_s^r \text{ROC}(\theta)\delta\theta, \tag{2.23}$$

where r and s denote the false positive rates of interest. This alternative index can be interpreted as the average TPR of the test over the restricted range of FPR values.

Selecting the interval (r, s) is an important practical issue. The choice depends on the particular setting and should depend on the cost of a false positive diagnosis as well as the benefits of a true positive (Dodd & Pepe, 2003).

The partial AUC curve, θ_r, is denoted as the trapezoid area between cutpoints r and $r+1$. The pAUC can be approximated by the area of a trapezoid;

$$\theta_r = \frac{1}{2}P(X = z_r, Y = z_r) + P(X > z_r, Y = z_r) \tag{2.24}$$

where X denotes the test result of an individual with the disease, Y denoted the test result of an individual without the disease, and z_r is an ordinal rating scale at cutpoint r $(r = 1, \ldots, s)$ (Zhang et al., 2002). The rating scale is assigned to each individual such that higher values of the rating are associated with the disease $(z_1 < \cdots < z_r < \cdots < z_s)$. For a perfect test $\text{ROC}(z_r) = 1$ for all $z_r \in (0, 1)$, and

the partial AUC is the area of the rectangle with height 1 and base $z_s - z_r$. Where z_r and z_s are the two selected cutpoints (Dodd & Pepe 2003).

The pAUC, θ_r can be estimated using the methods from Delong et al. (1988) and the Mann-Whitney U statistics (Zhang et al., 2002). Suppose there are n_r observations from Y with rating scale z_r. Then the total sample of Y can be divided into s groups $n = \sum_{r=1}^{s} n_r$. Similar, the sample size of X is, $m = \sum_{r=1}^{s} m_r$ (Zhang et al., 2002). An unbiased estimator for θ_r is

$$\hat{\theta}_r = \frac{1}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} S_r(X_i, Y_j), \tag{2.25}$$

where

$$S_r(X_i, Y_j) = \begin{cases} 1, & X_i < Y_j \text{ and } Y_j = z_r \\ \frac{1}{2}, & X_i = Y_j \text{ and } Y_j = z_r \\ 0, & \text{otherwise.} \end{cases} \tag{2.26}$$

The pAUC equals the total of each trapezoidal area between two cutpoints.

The study of partial ROC curves in medical research has increased recently. The partial ROC analysis provides more detailed information when two ROC curves cross or when interest is in a specific clinical range.

The partial AUC for $r \leq \text{FPR} \leq s$ for the homogeneous logistic threshold model

is as follows

$$\text{pAUC}_{\text{hom}} = \frac{\text{OR}}{(\text{OR}-1)^2}\left[(\text{OR}-1)(s-r) - \ln\left(\frac{1+s(\text{OR}-1)}{1+r(\text{OR}-1)}\right)\right], \tag{2.27}$$

with an approximate standard error

$$SE(\widehat{\text{pAUC}}_{\text{hom}}) \approx \left(\frac{\delta \text{pAUC}_{\text{hom}}}{\delta \text{OR}}\right) SE(\hat{\text{OR}}) \tag{2.28}$$

and

$$\frac{\delta \text{pAUC}_{\text{hom}}}{\delta \text{OR}} = \frac{1}{(\text{OR}-1)^3}[f(\text{OR}, r) - f(\text{OR}, s)], \tag{2.29}$$

where

$$f(\text{OR}, s) = \frac{[s(\text{OR}-1)(1+\text{OR}+s(\text{OR}-1)) - \ln\{1+s(\text{OR}-1)\}(\text{OR}+1)(1+s(\text{OR}-1))]}{1+s(\text{OR}-1)}. \tag{2.30}$$

(Walter, 2005).

The summary index, pAUC cannot in general attain the maximum value of one that is achievable by AUC, but instead has a maximum value of $s - r$. In order to regain the desirable property of a summary measure that ranges between 0 and 1, consider the scaled pAUC as

$$\text{pAUC}^*_{\text{hom}} = \frac{\text{pAUC}_{\text{hom}}}{s-r}, \tag{2.31}$$

where $\text{pAUC}^*_{\text{hom}}$ denotes a value scaled by the range of FPR values under considerations. The standard error of $\text{pAUC}^*_{\text{hom}}$ is

$$SE(\widehat{\text{pAUC}}^*_{\text{hom}}) = \frac{SE(\text{pAUC}_{\text{hom}})}{s - r}. \tag{2.32}$$

The pAUC and scaled pAUC in example 1 for the false positive rates ranging from 0 to 0.6 with a true OR of exp(3) is $\text{AUC}_{\text{hom}} = 0.49$ with $SE(\text{AUC}_{\text{hom}}) = 0.004$ and $\text{AUC}^*_{\text{hom}} = 0.82$ with $SE(\text{pAUC}^*_{\text{hom}}) = 0.007$, respectively.

Chapter 3

Multiple Study ROC Analysis

3.1 Motivation

To evaluate the performance of a binary scale diagnostic test, whether its binary nature comes from a true binary outcome or from a continuous outcome with a threshold applied, the result is described as a 2×2 table. From the 2×2 table, we can estimate the sensitivity and specificity which measures how accurate a binary scale diagnostic test is to detect the disease status. Since a single large size study is not easy to conduct, methods to combine the results from several independent studies are desired. Comparing to a single study, a careful structural review with rigorous meta-analysis can provide more reliable information for power analysis or sample size estimation for future studies. Some models can also be used to explore

the heterogeneity across studies.

When the response of a diagnostic test is continuous, its sensitivity and specificity are derived by dividing the outcomes at a certain threshold. Different thresholds result in different pairs of sensitivity and specificity. When combining results from different independent studies, it is assumed that there exists an underlying probability distribution and each study's results correspond to a specific threshold that determines the sensitivity (or TPR) and 1-specificity (or FPR) (Moses et al., 1993). These true positive and false positive rates are assumed to be on one common ROC curve, which is called the summary ROC (SROC) curve (Moses et al., 1993). If the underlying probability distribution is known, then we only need to estimate a few parameters in order to fit a smooth SROC curve (Moses et al., 1993). Common methods, such as maximum likelihood could be used to estimate the parameters and then the distributions. However, the underlying probability is seldom known.

The simplest method for analyzing pooled data from multiple studies is calculating sensitivities and specificities and their averages. This is valid when the same criteria for a positive result has been used in each study and each study is of similar size and quality (Jones et al., 2005). If different criteria or thresholds have been used, there will be a relationship between sensitivity and specificity across the studies. As sensitivity increases, specificity will generally drop. This is the threshold effect. The relationship between sensitivity and specificity cannot be evaluated if there is

a threshold effect across the studies. Combining such rates usually underestimates the test performance (Siadaty & Shu, 2004).

Alternatively, one may choose to extract odds ratios from each paper and then estimate the average OR across the studies. The advantage of this method is that different sensitivities and specificities can point to the same OR (homogeneous case). This means that different studies are reporting "truly different" sensitivities and specificities and that the between-study variation is not due to random noise alone, but because of the different decision thresholds chosen. Therefore, the major advantage of OR and its corresponding ROC curve, is that it provides measures of diagnostic accuracy independent from the decision criteria (Siadaty & Shu, 2004).

Occasionally, the remaining variation between studies, after utilizing OR as the summary performance measure, is still too much to be attributed to random noise. This is because the ORs may vary from study to study (heterogeneous case). Different ORs in the test performance across studies may be due to differences in study designs, subject populations, case difficulties, types of equipment, ability of rates, and dependencies of OR on the thresholds chosen (Siadaty & Shu, 2004). An SROC curve that allow for the possibility of "inconstant discrimination accuracy" would result in that of a heterogeneous SROC curve (Nelson, 1986). This means the SROC curve represents different ORs at different points.

In a single study, changing the threshold results in monotonic changes in TPR

and FPR (Henley & McNeil, 1982). In a meta-analysis, the units of analysis are separate studies. In the simplest case, each study contributes an estimate of TPR and FPR. The SROC curve is intended to represent the relationship between TPR and FPR across studies, recognizing they may have used different thresholds. In contrast to the ROC analysis, the set of (FPR, TPR) points need not necessarily yield a unique, monotonic curve (Walter, 2002).

Moses et al. (1993) proposed a least-squares approach to fit the SROC curve for combining different studies. He estimated the variances of the coefficients using the standard method for least-squares estimators. This method has been frequently used in meta-analysis literature in the last decade. Several alternative approaches have been proposed by various authors to either fit the smooth SROC curve using a hierarchical SROC model or deriving summary statistics of a diagnostic test from multiple studies. The hierarchical regression approach is a more sophisticated method that takes into account the correlation between sensitivity and 1-specificity and incorporates the intra-study and inter-study variations simultaneously (Rutter & Gatsonis, 1995). However, this method is quite computationally complex and is therefore not widely used in meta-analysis research.

In this section we introduce the notation as well as the method proposed by Moses et al. (1993) for fitting the SROC curve. This is followed by the summary measures used in SROC analysis. The summary statistics included are: area under

the SROC curve (AUC), Q^* statistic, area swept out by the curve (ASC), diagnostic odds ratio (DOR), and the partial area under the SROC curve (pAUC).

3.2 Summary ROC curve

The SROC curve is conceptually very similar to the ROC curve. However, each data point comes from a different study, not a different threshold. Diagnostic thresholds should be similar for each study, so the threshold effect does not influence the shape of the curve. The curve's shape is based entirely by the results across the studies (Jones et al., 2005). Each study produces values for sensitivities, specificities and therefore TPRs and FPRs. The SROC curve is made from the (TPR, FPR) points.

The SROC curve is placed over the points, (TPR, FPR), to form a smooth curve. The curve is calculated using a regression model (Littenberg & Moses, 1993) where TPR and FPR are transformed into logarithmic variables and graphed. A regression equation is calculated and the variables are manipulated to achieve TPR as a function of FPR. This is the equation for the SROC curve, which is then plotted over the original (TPR, FPR) points as shown in Fig. 3.1.

The curve is symmetric if the ORs do not vary between thresholds and asymmetric if the ORs vary between thresholds. Kardaun & Kardaun (1990) suggested an empirical transformation that mapped (TPR, FPR) from the ROC space, onto

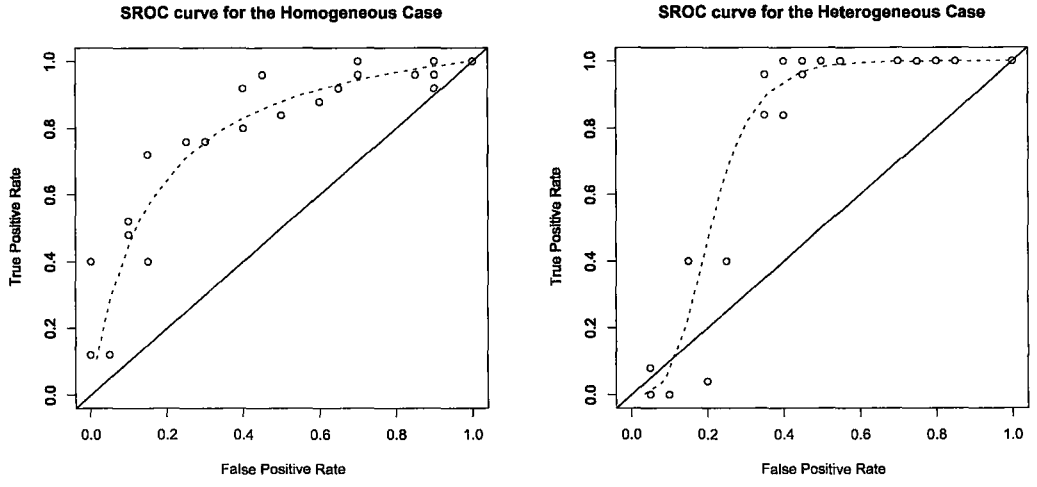

Figure 3.1: SROC curves for the homogeneous case (left) with location = 2 and scale = 0 and the heterogeneous case (right) with location = 2 and scale = 0.5.

(U, V) space, where

$$\begin{aligned} U &= \text{logit}(\text{FPR}) = \ln[\text{FPR}/(1-\text{FPR})] \\ V &= \text{logit}(\text{TPR}) = \ln[\text{TPR}/(1-\text{TPR})]. \end{aligned} \tag{3.1}$$

These definition lead to the estimates

$$\begin{aligned} \hat{U} &= \ln[(c/n_2)/(d/n_2)] = \ln[c/d] \\ \hat{V} &= \ln[(a/n_1)/(b/n_1)] = \ln[a/b], \end{aligned} \tag{3.2}$$

where a, b, c, and d are defined in Table 2.2.

If a, b, c, or d are zero, the transform involving these variables is undefined. To avoid this problem, a Cox (1970) correction was made in the rates a/n_1 and c/n_2

and resulted in the estimates (Moses et al., 1993),

$$\begin{aligned} \hat{U} &= \ln\left(\frac{c+0.5}{d+0.5}\right) \\ \hat{V} &= \ln\left(\frac{a+0.5}{b+0.5}\right). \end{aligned} \tag{3.3}$$

Under the assumption that the response of the test follows a logistic distribution, Moses et al., (1993) showed that U and V are linearly related. Let X be the test's response for a disease subject and Y be the test's response of a non-disease subject. Assume that X and Y follow logistic distributions with parameters (μ_1, s_1) and (μ_2, s_2), respectively. The probability density functions are

$$F_X(x) = \left[1 + e^{-\left(\frac{x-\mu_1}{s_1}\right)}\right]^{-1} \quad \& \quad F_Y(y) = \left[1 + e^{-\left(\frac{x-\mu_2}{s_2}\right)}\right]^{-1}. \tag{3.4}$$

It was shown that at a particular threshold k, the test had the corresponding sensitivity (TPR),

$$\text{TPR} = \text{Prob}(X > k) = 1 - F_X(k) = \left[1 + \exp\left(\frac{k-\mu_1}{s_1}\right)\right]^{-1}, \tag{3.5}$$

and 1-specificity (FPR),

$$\text{FPR} = \text{Prob}(Y > k) = 1 - F_Y(k) = \left[1 + \exp\left(\frac{k-\mu_2}{s_2}\right)\right]^{-1}. \tag{3.6}$$

(Moses et al., 1993).

Therefore, $U = (\mu_2 - k)/s_2$ and $V = (\mu_1 - k)/s_1$ are linearly related. If X and Y do not follow a logistic distribution exactly, the linear relationships might not be observed but may approximately hold (Walter, 2002). The closer the true distributions are to logistic, the closer the relationship between U and V is to linear. Some transformations can help in reaching better linearity. For example, further transformation of (U, V) into (D, S) is recommended. That is where

$$D = V - U = \ln\left(\frac{\text{TPR}}{1-\text{TPR}}\right) - \ln\left(\frac{\text{FPR}}{1-\text{FPR}}\right) \tag{3.7}$$

and

$$S = V + U = \ln\left(\frac{\text{TPR}}{1-\text{TPR}}\right) + \ln\left(\frac{\text{FPR}}{1-\text{FPR}}\right). \tag{3.8}$$

D is equivalent to the diagnostic log-odds ratio, $\ln(OR)$. It represents the odds of a positive test result among people with the disease relative to the odds of a positive test result among people without the disease. S can be looked at as a measure of the diagnostic threshold for classifying a test as positive. It has a value of zero when $\text{TPR} = 1 - \text{FPR}$ (Walter, 2002). S is positive when a threshold is used that increases sensitivity and decreases specificity and is negative when a threshold is used that decreases sensitivity and increases specificity (Irwig et al., 2006). Moses et al.,

(1993) assumed a linear relationship between D and S for all possible thresholds as

$$D = A + BS. \tag{3.9}$$

This regression equation can be fitted by standard least squares methods, assuming that D is approximately normal for a given values of S. The coefficient B represents the dependence of the test accuracy on the threshold. If $B \approx 0$, then the studies are homogeneous and can be summarized by an overall OR, noting that $A = \ln(\text{OR})$ (Walter, 2005). In this case, other approaches to combining ORs for meta-analysis can be used, for example the Mantel-Haenszel procedure (Irwig et al., 2006). If $B \neq 0$, then the studies are heterogeneous with respect to the ORs (Walter, 2002).

Reversal of the transformations (3.7) and (3.8) can be done once the regression has been fitted. When this is complete, the formulation of the relationship between TPR and FPR can be made and results in a summary ROC curve

$$TPR = \frac{\exp\left(\frac{A}{1-B}\right)\left(\frac{FPR}{1-FPR}\right)^{(1+B)/(1-B)}}{1+\exp\left(\frac{A}{1-B}\right)\left(\frac{FPR}{1-FPR}\right)^{(1+B)/(1-B)}}. \tag{3.10}$$

The SROC curve can be used to estimate TPR for each fixed value of FPR and conversely. The standard errors of the estimates can be obtained using the delta method (Gatsonis & Paliwal, 2006).

Equation (3.9) can be re-arranged as $V = \frac{A}{1-B} + \frac{1+B}{1-B}U$. Substituting U and V with the notation used in equations (3.5) and (3.6), we have the following equation:

$$\left(\frac{\mu_1 - k}{s_1}\right) = \frac{A}{1-B} + \frac{1+B}{1-B}\left(\frac{\mu_2 - k}{s_2}\right). \tag{3.11}$$

Equation (3.11) demonstrates the relation of the parameters between the two logistic distributions through the common threshold k and the true regression parameters. When A, B and (μ_2, s_2) are known, the parameters (μ_1, s_1) can be obtained according to the above equation.

The SROC curve is similar in principle to the ROC curve for a single study, except that the data points for the SROC curve are obtained from a set of studies being used for a meta-analysis (Walter, 2005). Ideally, the studies to be included would be identified through a formal search process, with inclusion and exclusion criteria, as well as other methodological requirements (Whitehead, 2002). The SROC is derived when each component from a set of studies contributes one 2×2 contingency table indicating the relationship between the true disease state (case or non-case) and the test result (positive or negative), for a single diagnostic threshold (Gatsonis & Paliwal, 2006). The SROC curve is intended to summarize the relationship between TPR and FPR across the set of studies.

3.3 Area Under the SROC Curve

There may be interest in identifying particular points on the SROC curve. It is often useful to have an overall summary measure of the curve's behaviour. For a single study the AUC is commonly used as a summary measure of the ROC curve. It indicates the overall performance of a diagnostic test in terms of its accuracy at various diagnostic thresholds that are used to discriminate cases and non-cases of disease (Henley & McNeil, 1982). A perfect test would have AUC = 1, where as a completely random test with the ROC curve lying on the main diagonal would have AUC = 0.5. In practice, most tests will lie somewhere between these two extremes. It also represents the (unweighted) average of TPR over all possible values of FPR (Van Der Schouw et al., 1994). The AUC measure is also used in meta-analysis, where each component study provides an estimate of the test sensitivity and specificity. These estimates are then combined to calculate an SROC curve which describes the relationship between test sensitivity and specificity across studies (Moses et al., 1993). The AUC for the SROC curve can be calculated as

$$\text{AUC} = \int_0^1 \frac{\exp\left(\frac{A}{1-B}\right)\left(\frac{x}{1-x}\right)^{(1+B)/(1-B)}}{1+\exp\left(\frac{A}{1-B}\right)\left(\frac{x}{1-x}\right)^{(1+B)/(1-B)}}\delta x. \tag{3.12}$$

Walter & Sinuff (2006) showed that in the homogeneous case, ($B = 0$) the AUC

can be calculated as

$$\mathrm{AUC}_{\mathrm{hom}} = \frac{\mathrm{OR}}{(\mathrm{OR}-1)^2}[(\mathrm{OR}-1) - \ln(\mathrm{OR})]. \tag{3.13}$$

In the heterogeneous case, AUC can be calculated only by using numerical integrations (Walter & Macaskill, 2004).

The AUC is calculated for SROC as for ROC. The diagnostic test is constant throughout the studies, so the AUC reflects overall performance of that test (Jones et al., 2005). The perfect test will again have an AUC of one.

The AUC can be interpreted in several different ways. First, the AUC represents the average value of TPR over all possible FPR values between 0 and 1. Second, AUC is also the probability of correctly ranking a case and non-case, based on the observed test values of these individuals (Walter, 2005). Lastly, the AUC is related to the Mann-Whitney statistic used to evaluate the significance of the differences between the sample distributions of case and non-case test values (Walter, 2005).

It can be difficult to carry out a diagnostic meta-analysis since some studies report diagnostic ORs and others provide the AUC from a ROC curve. Both OR and AUC are valid summary measures of diagnostic accuracy. However, the measures are on two different metrics, which makes the combining of studies into the meta-analysis difficult (Walter & Sinuff, 2006).

The conversion of the AUC values into OR point estimates will be achieved using

a logit-threshold model proposed by Moses et al., (1993). The logit-threshold model is $D = A + BS$ where D and S are given in equations (3.7) and (3.8), respectively. If $B \approx 0$, the general expression in (3.12) becomes

$$\mathrm{AUC}_{\mathrm{hom}} = \int_0^1 \frac{\exp(A)\left(\frac{x}{1-x}\right)}{1+\exp(A)\left(\frac{x}{1-x}\right)}\delta x. \tag{3.14}$$

We can obtain an exact solution to the AUC as shown in equation (3.13). If $A = 0$ (or OR = 1), then $AUC_{\mathrm{hom}} = \frac{1}{2}$, which is then degenerate. Equation (3.13) can be used to evaluate AUC for homogeneous studies, by using the common estimate of OR and without the need for numerical integration (Walter & Sinuff, 2006). In Walter's (2002) paper he showed that the $\mathrm{AUC}_{\mathrm{hom}}$ expressions in (3.13) and (3.14) also gives a good approximations to the AUC index even if the component studies are heterogeneous.

3.4 Q^* Statistic

The Q^* statistic is the intercept of the SROC curve at the anti-diagonal (TPR + FPR = 1) line through the unit square. Its value indicates the overall accuracy by finding where sensitivity and specificity are the same (Jones et al., 2005). The closer the curve is to the top left corner (perfect sensitivity and specificity), the better the accuracy. For symmetric curves, this value is also the point at which the curve is

closest to the ideal point, where FPR = 0 and TPR = 1 (Gatsonis & Paliwal, 2006). The anti-diagonal will cut the curve at a higher level, giving higher Q^* and a more accurate test (Jones et al., 2005).

Q^* statistic is appropriate provided high sensitivity and high specificity are equally desirable. If one is clinically more important than the other, the Q^* statistic does not address the clinical usefulness of the test (Jones et al., 2005). In this case, overall accuracy is not as relevant as overall sensitivity or specificity.

The Q^* statistic is similar to the area under the entire SROC curve as it is an indicator of how closely the SROC curve is to the north-west corner (Moses et al., 1993). The point where the line TPR + FPR = 1 intersects the SROC curve has the co-ordinates

$$\text{TPR} = \frac{\exp(A/2)}{1+\exp(A/2)} \quad \text{and} \quad \text{FPR} = \frac{1}{1+\exp(A/2)} \tag{3.15}$$

The Q^* statistic is a function of A only and the standard error for Q^* is available when the least squares model has provided the estimate of A (Moses et al., 1993). The Q^* represents the diagnostic threshold at which the probability of a correct diagnosis is constant for all subjects.

When $B \neq 0$ the SROC curve has a region where TPR < FPR , which lies below the main diagonal. In this region, the test would be predicted to performing worse than at random (Walter, 2002). For example, we can see this region near the lower

left corner when $B = 0.5$ in Fig. 3.1. If $B \geq 0$, it can be shown that the point $(\text{FPR}', \text{TPR}')$ where the SROC curve crosses the diagonal is

$$\text{FPR}' = \text{TPR}' = \frac{\exp(-A/2B)}{1 + \exp(-A/2B)} \tag{3.16}$$

If $B < 0$, there is a symmetrically opposite point in the top-right corner of the SROC space (Walter, 2002). A diagnostic test would not usually be used at such low values of TPR, so in practice the improper part of the curve where TPR < FPR is negligible.

Note that, AUC declines with increasing B and that the limit curve with $B \to 1$ passes through the common Q^* point. From (3.15) a lower bound for AUC in the curve can be formed with a given value of A

$$Q^* = \frac{\exp(A/2)}{1 + \exp(A/2)} = \frac{\sqrt{\text{OR}}}{1 + \sqrt{\text{OR}}}. \tag{3.17}$$

Equation (3.17) is the same as the TPR value in equation (3.15). Upper and lower bounds for AUC with a given value of $A > 0$ are given using the Q^* from (3.17) and the maximum value of AUC_{hom} from (3.13), respectively (Walter, 2002). This argument assumes $|B| < 1$, since $|B| > 1$ is not of practical interest.

Walter (2002) used the delta method to give an approximate standard error for $\hat{Q}^*$

$$\mathrm{SE}(\hat{Q}^*) = \frac{\sqrt{OR}}{2(\sqrt{OR}+1)^2}\mathrm{SE}(\hat{A}). \tag{3.18}$$

3.5 Area Swept Out by the SROC Curve

The SROC curve is a special case in the definition for the ASC index proposed by Lee et al. (1996). The ASC applies when a parametric ROC (or SROC) curve for a logit-threshold model is used. AUC refers to the area under the SROC curve and the ASC is the area swept out by the SROC curve. Zhang (2004) has found it useful to express the ASC in terms of the AUC since some regions of these two indices overlap.

For the heterogeneous case, when both A and B are positive, (Fig. 3.1), we can imagine a ray arising from the origin (0,0) to each point in the SROC curve. This ray can be expressed as $y = kx$, where k is the slope of the line. As the point moves from the origin to the left-uppermost corner, the slope of the line will reach its maximum at some point of the curve when the line still touches the curve except for the origin. This line is the tangent of the curve. It can be expressed as $y = k^*x$, where k^* is the slope of the tangent line. As the end of the ray moves from (0,0) to (1,1), the areas are swept out within the SROC space. Also, the areas surrounded by the tangent line, the main diagonal line and the curve may be swept out twice.

If we denote the area swept out as the ray moves from the origin up to the

tangent line as M, then the ASC when $B \neq 0$ can be expressed as

$$ASC = AUC + 2M - \tfrac{1}{2} \quad \text{when} \quad B > 0 \tag{3.19}$$

or

$$ASC = 2M + \tfrac{1}{2} - AUC \quad \text{when} \quad B < 0. \tag{3.20}$$

Zhang (2007) used the delta method to yield an approximate variance for ASC

$$\text{var}(\widehat{ASC}) = \left(\frac{\delta ASC}{\delta A}\right)^2 \text{var}(\hat{A}) + \left(\frac{\delta ASC}{\delta B}\right)^2 \text{var}(\hat{B}) + 2\left(\frac{\delta ASC}{\delta A}\right)\left(\frac{\delta ASC}{\delta B}\right)\text{cov}(\hat{A}, \hat{B}) \tag{3.21}$$

here,

$$\begin{aligned} \frac{\delta ASC}{\delta A} &= \frac{\delta AUC}{\delta A} + 2\frac{\delta M}{\delta A}, \quad \text{when} \quad B > 0, \\ \frac{\delta ASC}{\delta A} &= 2\frac{\delta M}{\delta A} - \frac{\delta AUC}{\delta A}, \quad \text{when} \quad B < 0. \end{aligned} \tag{3.22}$$

Similarly,

$$\begin{aligned} \frac{\delta ASC}{\delta B} &= \frac{\delta AUC}{\delta B} + 2\frac{\delta M}{\delta B}, \quad \text{when} \quad B > 0 \\ \frac{\delta ASC}{\delta B} &= 2\frac{\delta M}{\delta B} - \frac{\delta AUC}{\delta B}, \quad \text{when} \quad B < 0. \end{aligned} \tag{3.23}$$

In Zhang (2007) paper she showed that the values of the M index, for a fixed value of A, as B changes from negative to positive were unsteady. This resulted in non-smooth values of ASC. However, ASC decreases as B changes from negative to zero and increases as B changes from zero to positive. As $A \to \infty$, ASC $\to 0.5$, the maximum value. If the ASC is close to 0.5, the greater the probability of a correct

diagnosis when $A > 0$ and the greater the probability of a wrong diagnosis when $A < 0$. Zhang (2007) suggests that only if the AUC is large and the parameter A is positive, a large value of ASC implies a perfect test.

The ASC is affected mostly by the index M proposed by Zhang (2004). Although the AUC is symmetric with respect to $|A|$ and $|B|$, ASC lacks symmetry with respect to B for a fixed A or lacks symmetry with respect to A for fixed B. The ASC is a decreasing function of A when B is positive and an increasing function of A when B is negative. On the other hand ASC is a decreasing function of B when A is negative and an increasing function of B when A is positive. In the homogeneous case, when $B = 0$, ASC is the region between the curve and the diagonal line (Zhang, 2004).

$$\mathrm{ASC}_{\mathrm{hom}} = \mathrm{AUC}_{\mathrm{hom}} - \tfrac{1}{2}, \quad \text{for all} \quad A > 0 \tag{3.24}$$

or

$$\mathrm{ASC}_{\mathrm{hom}} = \tfrac{1}{2} - \mathrm{AUC}_{\mathrm{hom}}, \quad \text{for all} \quad A < 0. \tag{3.25}$$

The approximate variance of the ASC statistic for when $B = 0$ is

$$\mathrm{var}(\widehat{ASC}_{\mathrm{hom}}) = \frac{\exp(2A)}{[\exp(A) - 1]^6}[A(\exp(A) + 1) - 2(\exp(A) - 1)]^2 \mathrm{var}(\hat{A}) \tag{3.26}$$

(Zhang, 2007).

The basic properties of ASC in the context of the SROC curve were reviewed in

this section. The mathematical expressions of the ASC and its variance proposed by Zhang, (2004) were shown here. The ASC in the homogeneous case can provide a good approximation to heterogeneous studies with a large odds ratio. Also, the ASC and its variance are easily computed in the homogeneous case. Similar to the AUC index, ASC is related to the probability that the test will correctly rank a disease and control pair of subjects when the value of A is positive or negative. In practice, data yielding $A < 0$ are unlikely.

3.6 Diagnostic Odds Ratio

When comparing tests for the same diagnostic procedure, it can be useful to turn to the OR. Tests can differ in terms of sensitivity and specificity, which reflects as a threshold shift. Comparing ORs can help in an initial evaluation, although the total costs will depend on the relative weight attached to the false positive and false negative results. Furthermore, the diagnostic odds ratio (DOR) appears in the SROC method for meta-analysis of diagnostic tests.

The DOR offers considerable advantages in meta-analysis of diagnostic tests that combines results from different studies into summary estimates with increased precision. The approach by Moses et al. (1993) relies on the logarithmic linear regression of the DOR for a study (dependent variable, A) on an expression of the positivity threshold for that study (independent variable, B). If the regression line

has a zero slope ($B = 0$), the DOR is constant across studies. The resulting SROC curve will be symmetric and concave. In other words, study heterogeneity can be attributed to threshold differences. In the context of the DOR, the summary OR of the study under evaluation can be obtained from the intercept, A (e^A) from the regression line (Moses et al., 1993).

A method for determining if the variation is not due to random noise alone but to a study characteristic within a study. The model $D = A + BS$ can easily be expanded to a multiple linear regression model by adding one or more covariates, such as examination (X_1), subject (X_2) and study design (X_3) characteristics: $D = A + BS + B_1X_1 + B_2X_2 + B_3X_3$. The regression coefficients (B_1, B_2, B_3) are indicators of the independent effects of the corresponding covariates (X_1, X_2, X_3) on the dependant variable, $\ln(DOR)$. The magnitude of the regression coefficient of a variable represents the difference in $\ln(DOR)$ between studies with different levels of that variable, with all other variables held constant. A large regression coefficient indicates that the corresponding covariate has a large influence on diagnostic accuracy.

The DOR is another measure of the overall diagnostic power of the test. A high, $DOR > 1$ implies that the test shows good diagnostic accuracy in all subjects.

3.7 Partial Area Under the SROC Curve

The AUC has been criticized in meta-analysis for two reasons. First, the studies contributing to the meta-analysis are sometimes only observed within a limited range of FPR values. Second, even if some data are available in higher ranges of FPR, using a test with a high FPR value may be unacceptable in the clinical context (Scheidler et al., 1997). In order to adopt a new test for routine clinical use, one might restrict attention to smaller values of FPR (Walter, 2005).

To reduce the problem of the linear model in (3.9) being overly influenced by a point that is irrelevant to the area of decision making, Moses et al. (1993) suggested that one could include only those studies within a range considered clinically relevant. However, it maybe difficult to judge which areas are clinically relevant.

If the attention is indeed focused on a limited range of FPR, then the AUC statistic becomes less relevant as a summary measure of the data. A possible alternative would be to adopt the partial area under the SROC curve (pAUC) (McClish, 1989).

From the model in (3.9), the pAUC for $r \leq \text{FPR} \leq s$ is as follows

$$pAUC = \int_r^s \frac{\exp\left(\frac{A}{1-B}\right)\left(\frac{x}{1-x}\right)^{(1+B)/(1-B)}}{1+\exp\left(\frac{A}{1-B}\right)\left(\frac{x}{1-x}\right)^{(1+B)/(1-B)}}\delta x. \tag{3.27}$$

The pAUC can be interpreted as the average TPR of the test over the restricted

range of FPR values. Walter (2005) used the delta method to obtain the approximate standard error for the pAUC,

$$\begin{aligned}\widehat{\text{var}(\text{pAUC})} \approx & \left(\frac{\delta \text{AUC}}{\delta A}\right)^2 \text{var}(\hat{A}) + \left(\frac{\delta \text{AUC}}{\delta B}\right)^2 \text{var}(\hat{B}) \\ & +2\left(\frac{\delta \text{AUC}}{\delta A}\right)\left(\frac{\delta \text{AUC}}{\delta B}\right) \text{cov}(\hat{A}, \hat{B}),\end{aligned} \tag{3.28}$$

where

$$\begin{aligned}\left(\frac{\delta \text{AUC}}{\delta A}\right) &= \left(\tfrac{1}{1-B}\right) \exp\left(\tfrac{A}{1-B}\right) \int_r^s \tfrac{(x/(1-x))^p}{[1+(x/(1-x))^p \exp(A/(1-B))]^2} \delta x \\ \left(\frac{\delta \text{AUC}}{\delta B}\right) &= \left(\tfrac{1}{1-B}\right)^2 \exp\left(\tfrac{A}{1-B}\right) \int_r^s \tfrac{(x/(1-x))^p [A+2\ln(x/(1-x))]}{[1+(x/(1-x))^p \exp(A/(1-B))]^2} \delta x\end{aligned} \tag{3.29}$$

and $p = (1+B)/(1-B)$.

The scaled pAUC, denoted as pAUC*, is used as a summary measure that falls into the range 0 and 1,

$$\text{pAUC}^* = \frac{\text{pAUC}}{s-r}. \tag{3.30}$$

The pAUC* is denoted as a value scaled by the range of FPR values under consideration. The standard error for pAUC* is

$$\text{SE}(\text{pAUC}^*) = \frac{\text{SE}(\text{pAUC})}{s-r}. \tag{3.31}$$

When $B = 0$ and hence $\exp(A) = \text{OR}$ the pAUC has a closed form as shown in Section 2.8, along with its standard error. The scaled partial area under the SROC

curve, $pAUC^{*}_{hom}$ and its standard error are also shown in Section 2.8.

Irwig et al., (1995) believes that the pAUC has two drawbacks. The first, it is often difficult to judge which area is clinically relevant. Second, the method excludes (or includes) points which may be outside (inside) the area because of different study designs.

Although, the AUC, Q^{*} and ASC statistics are sufficient summary measures for describing the behaviours of the studies in an SROC curve, the pAUC summary measure is being used more to describe the relevant regions in an SROC curve. The partial area has clinical appeal in many situations, however, procedure and guidelines on the use of the pAUC in meta-analysis have not yet been produced. The next two section will focus mainly on the pAUC statistic and it's behaviours in different study designs.

Chapter 4

Assessment of Proposed Estimates and Study Designs

4.1 Motivation

The summary receiver operating characteristic (SROC) curve and its associated indices are valuable tools for the assessment of the accuracy for diagnostic tests. The area under the SROC curve is a popular summary measure of the accuracy for a test. The full area under the SROC curve, however, has been criticized because it gives equal weight to all false positive rates. Alternative indices include the area under the SROC curve in a particular range of false positive rates (partial area) and the Q^* statistic. We present an approach for computing sample sizes for the SROC

curves and their indices.

It is well known that the closer the ROC and SROC curves are to the upper left-hand corner, the more accurate the diagnostic tests. Diagnostic tests are more accurate when TPR is close to 1 and FPR is close to 0. The SROC curve obtained by a small number of studies may not lie in the upper left quadrant as would be desired. For example, small studies may produce extreme results from small population. Small studies are prone to producing outlying results and shift the overall outcome. The study size can contribute to the different sensitivity and specificity results. Heterogeneity may also contribute to the differences in sensitivity and specificity results across studies. Study designs and a small sample size, with respect to both the number of studies available for the meta-analysis, as well as the number of subjects included in each study are two possible explanations for the SROC curve not occupying the upper left-hand corner (Walter, 2002).

In this chapter we will first look at assessing the proposed estimates of the area under the SROC curve as well as the Q^* statistic. In Section 2 we will compare the approximate standard error formulas for AUC and Q^* with a bootstrapping procedure. We will follow this with a simulation comparing a variety of study sample sizes in Section 3 and 4. We will conclude with a section comparing balanced with unbalanced data sets with respects to the number of disease and control subjects within a study sample.

4.2 Bootstrap Procedure

To determine if the approximate standard errors formulas (3.28) and (3.18) are adequate for the AUC and Q^* statistics, respectively, we generated a bootstrapping procedure to estimate the sampling distributions of the two statistics. Estimated standard errors for the AUC and Q^* statistics are generated by using the variance estimator and approximated by the bootstrap methods (Table 4.1). Estimates of the standard errors generated by both Moses et al., (1993) and the bootstrap methods were similar.

Bootstrap procedures take the combined samples as a representation of the population from which the data came and creates 1000 or more bootstrapped samples by drawing, with replacement, from that pseudo-population. The means and standard deviations were calculated for each sample. The average of the statistics over all the samples are the bootstrap estimators. In this case, the bootstrap estimators after 2000 repeated samples, are the AUC and Q^* statistics. The variances of the estimated AUC and Q^* statistic provide an estimate of the sample variance for the two statistics. Since the standard error calculated by both the formulas and the bootstrap methods were similar, we will use the formulas (3.28) and (3.18) for the standard errors of AUC and Q^* statistics, respectively, for the rest of the applications in this paper.

4.3 Small vs. Large samples

In this section, we simulated 10 studies using binomial data for small and large study sample sizes, with 1000 data sets simulated for each condition. We computed the empirical variance of an SROC index by calculating $\hat{A}$ and $\hat{B}$ directly from the Moses et al., (1993) model for each data set. We estimated the areas under the SROC curve by integration, utilizing these direct estimates of A and B. We computed the variability in the 1000 estimates of $\theta = AUC$ calculated in this way. We calculated the pAUC indices holding $r = 0$ and varying the value of s.

Fig. 4.1 shows the comparison for small and large study sample sizes of the standard errors for the pAUCs or scaled pAUCs with $N = 10$, $A = 2$ and $B = -0.5,\ 0$ and 0.5. We can see that the standard errors of pAUC and scaled pAUC for $B = 0.5$ are similar for both small and large sample sizes indicating little effect with sample size differences in the estimate. This may be due to the curve having a

Table 4.1: Standard error values of the AUC and Q^* statistics with $A = 2$ and $B = -0.5,\ 0,\ 0.5$. Comparing the standard error formulas with the standard errors found by bootstrapping.

B value	Statistic	Formulas	Bootstrap
-0.5	SE(AUC)	0.03358	0.03417
	SE(Q^*)	0.02883	0.02825
0	SE(AUC)	0.03646	0.03816
	SE(Q^*)	0.02816	0.02935
0.5	SE(AUC)	0.05388	0.05394
	SE(Q^*)	0.04239	0.04232

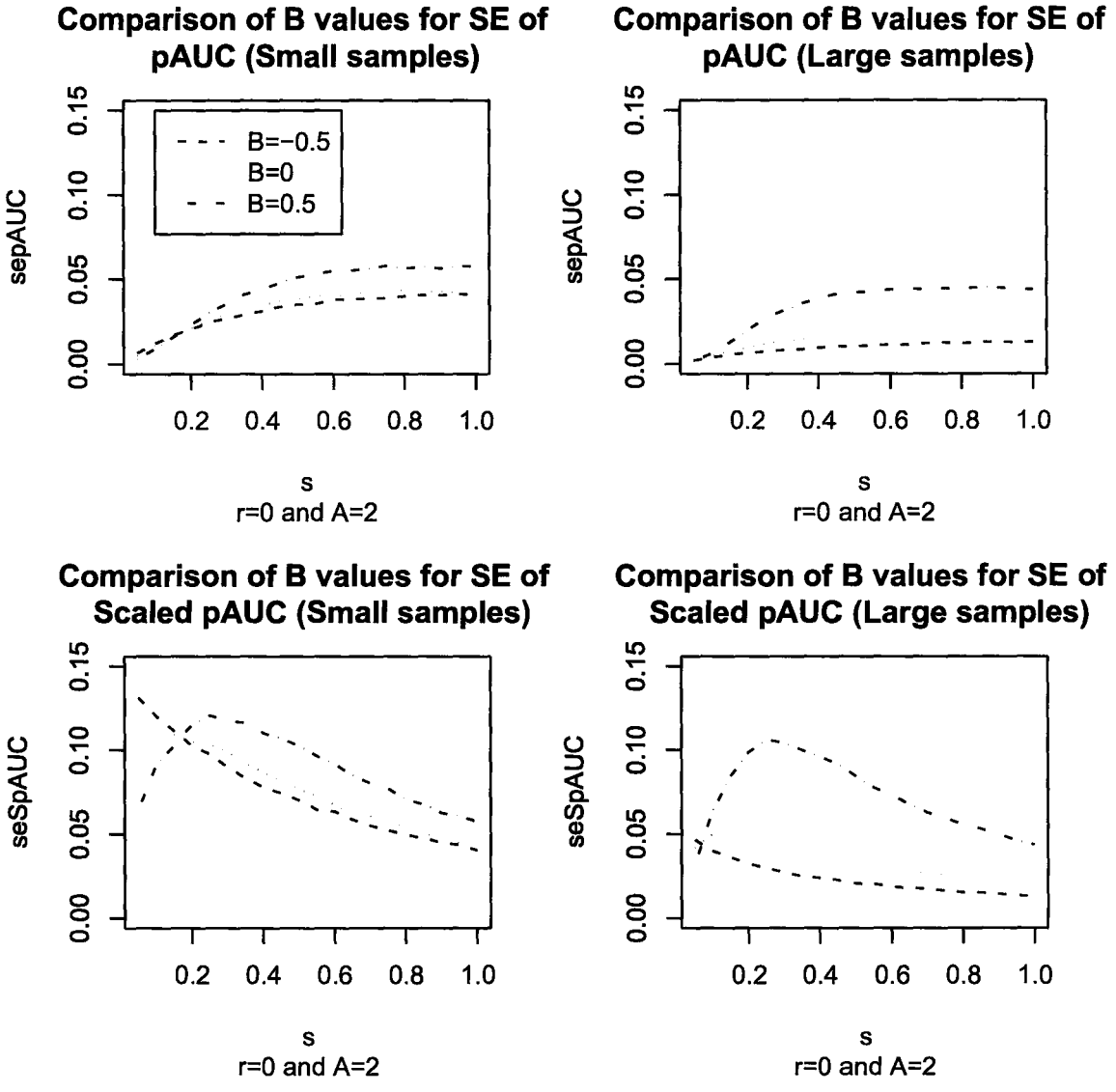

Figure 4.1: Comparison of small and large study sample sizes for the standard errors of the pAUC or scaled pAUC for $N = 10$ and $A = 2$ with varying values of B.

region where $TPR < FPR$, which lies below the main diagonal line near the bottom left-hand corner. Also, the standard errors for the pAUC and scaled pAUC with $B = 0.5$ are higher than for $B = 0$ or $B = -0.5$ in both small and large sample

sizes. If $B = -0.5$ the SROC curve has a region similar to the region given when $B = 0.5$ but in a symmetrical opposite point on the top right corner of the SROC space. A diagnostic test would not usually be used at such low values of TPR, so in practice, the improper part of the curve where $TPR < FPR$ is negligible. From this result one would want to achieve the homogeneous logistic threshold model with $B = 0$. In other words, there is a common odds ratio that underlines the N studies. To do this, the data must come from two logistic distributions with the same scale but different location parameters for the control and the disease groups, respectively. The standard errors for small sample sizes when $B = 0$ are higher than they are for larger sample sizes as one would generally expect. However, large samples sizes are seldom available when studying a diagnostic test.

When $B = 0$ or $B = -0.5$ the standard errors for the scaled partial AUC for small data are relatively the same for all scaled pAUC indices from $s = 0.2$ to 1.0 as shown in Fig. 4.1. The clinically relevant region of the SROC space, where $s = 0.05$ to 0.15, the standard errors for the scaled pAUC when $B = -0.5$ are larger then when $B = 0$. This is another indication the homogeneous model is ideal for this type of analysis when dealing with small study sample sizes. For large study sample sizes the standard errors for all scaled pAUCs are larger than when $B = -0.5$.

Fig. 4.2 compares the standard errors of the pAUC and scaled pAUC for four different samples sizes with $B = 0$ and $B = 0.5$. The four samples sizes include: all

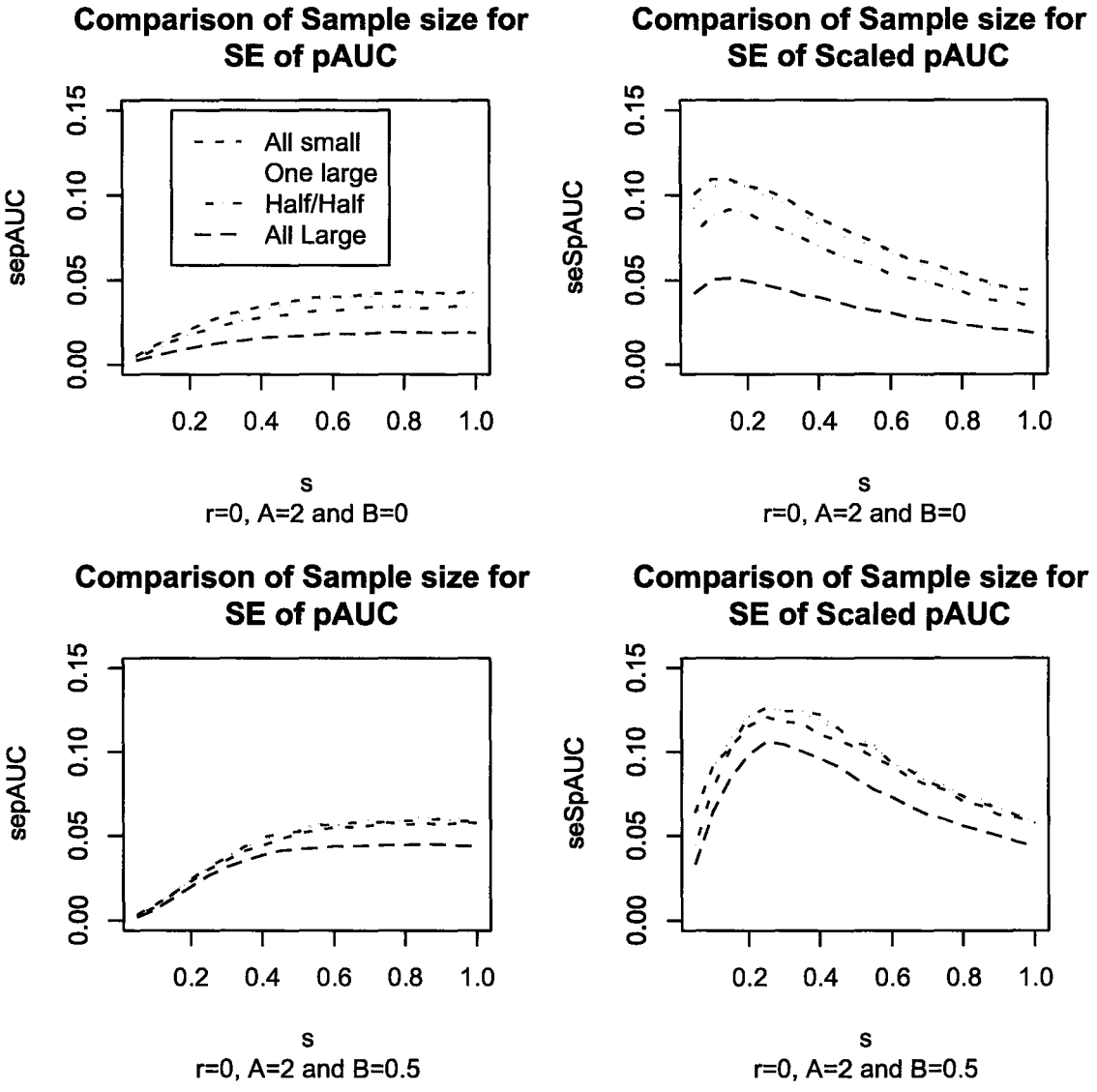

Figure 4.2: Comparison of study sample sizes for the standard errors of the pAUC or scaled pAUC for $N = 10$ and $A = 2$, $B = 0$ or $B = 0.5$.

small samples, 10% large samples, half small and half large samples and all large samples. When $B = 0$ the standard errors for pAUC and scaled pAUC decrease as the study sample sizes increase. However, when $B = 0.5$ the standard errors are

similar for 3 of the 4 sample sizes simulated. The standard errors are alike when all small samples, 10% large or half small and half large sample sizes are used. This shows that when $B > 0$ the effect of increasing the sample sizes is minimal and therefore non informative when dealing with relatively small sample sizes.

4.4 Assessment of Sample Sizes

We conducted another simulation study to assess the estimators for variance of the various SROC indices. Similar to the previous section, we simulated 10 studies using binomial data from various sets of (A, B) parameters and for various study sample sizes, with 1000 data sets simulated for each condition.

We computed the empirical variance of a summary ROC index by calculating $\hat{A}$ and $\hat{B}$ directly from the Moses model for each data set. We estimated the areas under the SROC curve by integration utilizing these direct estimates of A and B, as before. We computed the variability in the 1000 estimates of $\theta = AUC$ calculated in this way.

We considered five SROC indices: the full area under the curve ($\theta_{0,1}$); the areas under the curves in the FPR ranges of 0.0 to 0.05 ($\theta_{0,0.05}$),0.0 to 0.1 ($\theta_{0,0.1}$), 0.0 to 0.15 ($\theta_{0,0.15}$) and 0.0 to 0.2 ($\theta_{0.0,0.2}$). The simulation was done for four cases: the case when all study sample sizes are small (less than 50), when 10% of the studies

Table 4.2: SROC indices and their standard errors for the four sample size cases with $B = -0.5$, $A = 2$, $N = 10$ and $r = 0$.

Cases	s value	pAUC	SE(pAUC)	Scaled pAUC	SE(SpAUC)
I. All Small					
	0.05	0.02030	0.00658	0.40608	0.13164
	0.10	0.04696	0.01206	0.46961	0.12056
	0.15	0.07673	0.01682	0.51156	0.11211
	0.20	0.10946	0.02047	0.54742	0.10234
	1.00	0.76071	0.04064	0.76071	0.04064
II. 10% Large					
	0.05	0.02043	0.00633	0.40858	0.12659
	0.10	0.04729	0.01181	0.47292	0.11806
	0.15	0.07833	0.01619	0.52222	0.10791
	0.20	0.11151	0.01935	0.55753	0.09676
	1.00	0.75902	0.03901	0.75902	0.03901
III. 50/50					
	0.05	0.02236	0.00518	0.44727	0.10361
	0.10	0.05093	0.00905	0.50927	0.09048
	0.15	0.08243	0.01211	0.54952	0.08075
	0.20	0.11541	0.01479	0.57704	0.07393
	1.00	0.76556	0.02986	0.76556	0.02986
IV. All Large					
	0.05	0.02448	0.00234	0.48957	0.04676
	0.10	0.05513	0.00401	0.55133	0.04012
	0.15	0.08771	0.00530	0.58475	0.03598
	0.20	0.12208	0.00649	0.61038	0.03247
	1.00	0.77363	0.01290	0.77363	0.01290

are large (greater than 50), when half of the studies have small samples and half have large samples and lastly, when all sample sizes are large. The simulation was done for $B = -0.5$, 0 and 0.5, with $A = 2$, $N = 10$ and $r = 0$.

For $B = -0.5$ in Table 4.2, all pAUC and scaled pAUC statistics increase as the

Table 4.3: SROC indices and their standard errors for the four sample size cases with $B = 0$, $A = 2$, $N = 10$ and $r = 0$.

Cases	s value	pAUC	SE(pAUC)	Scaled pAUC	SE(SpAUC)
I. All Small					
	0.05	0.00976	0.00503	0.19523	0.10059
	0.10	0.02888	0.01093	0.28884	0.10931
	0.15	0.05354	0.01642	0.35696	0.10945
	0.20	0.08312	0.02107	0.41559	0.10533
	1.00	0.76928	0.04452	0.76928	0.04453
II. 10% Large					
	0.05	0.00922	0.00463	0.18446	0.09251
	0.10	0.02870	0.01048	0.28695	0.10477
	0.15	0.05241	0.01619	0.34938	0.10792
	0.20	0.08258	0.02105	0.41290	0.10523
	1.00	0.77251	0.04309	0.77251	0.04309
III. 50/50					
	0.05	0.00883	0.00385	0.17655	0.07698
	0.10	0.02727	0.00880	0.27272	0.08796
	0.15	0.05292	0.01374	0.35277	0.09161
	0.20	0.08234	0.01796	0.41171	0.08982
	1.00	0.78063	0.03387	0.78063	0.03387
IV. All Large					
	0.05	0.00818	0.00210	0.16366	0.04209
	0.10	0.02705	0.00503	0.27051	0.05027
	0.15	0.05255	0.00769	0.35030	0.05130
	0.20	0.08249	0.00991	0.41244	0.04956
	1.00	0.79241	0.01890	0.79241	0.01890

sample sizes increase. The standard errors for scaled pAUC decrease as the value of s increases. For cases I and II the standard errors do not fall with in 90 per cent confidence intervals (CI) except when $s > 0.2$. For $B = 0$ in Table 4.3, the pAUCs and scaled pAUCs for $s = 0.05$ and $s = 0.10$ decrease as the sample sizes increase. The standard errors for the scaled pAUC in this case is maximum when $s = 0.15$.

This indicates that the largest standard error falls within the 90 per cent CI for small to large sample sizes. This result shows that when $B = -0.5$ the estimate given for small samples are not as accurate when $B = 0$. When $B = 0.5$ in Table 4.4, the pAUC and scaled pAUC for $s = 0.05,\ 0.10,$ and 0.15 decrease as the sample sizes increase. The standard errors for the scaled pAUC is maximum when $s = 0.25$ (Case I: 0.121005) , which does not fall with in the 90 per cent CI region (not shown in table). However, for cases I, II, and III the standard errors for scaled pAUC fall within 90 per cent when s is between 0.05 and 0.15. For all three B values the full AUC increases as sample sizes increase.

From this simulation we can conclude that $B = 0$ and $B = 0.5$ give accurate estimates for $\theta_{0,0.05}$, $\theta_{0,0.10}$ and $\theta_{0,0.15}$ for small sample sizes. The maximum standard error for $B = 0$ falls within 10 per cent significance, whereas the maximum for $B = 0.5$ does not. For small study sample sizes, the homogeneous case, when $B = 0$, gives the most accurate estimates for the pAUC and scaled pAUC when $s \leq 0.2$.

Table 4.4: SROC indices and their standard errors for the four sample size cases with $B = 0.5$, $A = 2$, $N = 10$ and $r = 0$.

Cases	s value	pAUC	SE(pAUC)	Scaled pAUC	SE(SpAUC)
I. All Small					
	0.05	0.00468	0.00319	0.09369	0.06370
	0.10	0.01699	0.00917	0.16995	0.09174
	0.15	0.03153	0.01543	0.21023	0.10287
	0.20	0.05506	0.02305	0.27528	0.11527
	1.00	0.74457	0.05742	0.74457	0.05742
II. 10% Large					
	0.05	0.00386	0.00305	0.07720	0.06097
	0.10	0.01297	0.00876	0.12968	0.08758
	0.15	0.03035	0.01639	0.20230	0.10926
	0.20	0.05059	0.02418	0.25296	0.12090
	1.00	0.74398	0.06091	0.74398	0.06091
III. 50/50					
	0.05	0.00227	0.00222	0.04542	0.04444
	0.10	0.00990	0.00785	0.09901	0.07845
	0.15	0.02433	0.01538	0.16217	0.10255
	0.20	0.04534	0.02418	0.22668	0.12088
	1.00	0.75571	0.05865	0.75571	0.05865
IV. All Large					
	0.05	0.00188	0.00164	0.03754	0.03287
	0.10	0.00983	0.00642	0.09827	0.06420
	0.15	0.02396	0.01269	0.15971	0.08463
	0.20	0.04698	0.01972	0.23489	0.09862
	1.00	0.78014	0.04376	0.78014	0.04376

4.5 Effects of Study Design: Numbers of Control and Disease Subjects

Not all studies that estimate the SROC curve have the same number of control and disease subjects and the design may be unbalanced. For example, in single diagnostic

analysis, one may find it easier to obtain verified results for disease subjects. In other situations, the analysis may choose to test more control subjects, where the ratio of control to diseased subjects is greater than one. It is important to consider this aspect of the design in determining the appropriate sample size, of the entire size and the ratio of the numbers of control and diseased subjects. In this section we looked at the extreme cases where all study sample sizes either have twice as many disease or twice as many control subjects. We compare these two cases with a balanced design.

Lets K be the ratio of the number of control subjects (n_c) to the number of disease subjects (n_d) in the study sample.

$$K = \frac{n_c}{n_d} \tag{4.1}$$

For example, $K = 1$ means equal number of subjects with and without the disease in the study sample; $K = 0.5$ means twice as many disease subjects as control subjects in the study sample; and $K = 2$ means twice as many control subjects as disease subjects in the study.

Tables 4.5 and 4.6 show the influence of K and B on the standard errors of the partial AUC and scaled partial AUC, respectively. Here, we assume $A = 2$, giving an approximate area under the curve equal to 0.78 and the total number of studies,

Table 4.5: Effect of B and K on the estimated standard error of the pAUC indices where $\theta_{0,1} = 0.79$, $A = 2$, and $N = 10$

K	B value	$\theta_{0,1}$	$\theta_{0,0.05}$	$\theta_{0,0.1}$	$\theta_{0,0.3}$
0.5	-0.5	0.03828	0.00625	0.01133	0.02436
0.5	0	0.04400	0.00494	0.01119	0.02996
0.5	0.5	0.05868	0.00302	0.00875	0.03676
1	-0.5	0.04892	0.00767	0.01385	0.03141
1	0	0.04703	0.00569	0.01210	0.03123
1	0.5	0.05773	0.00337	0.00939	0.03541
2	-0.5	0.04814	0.00745	0.01335	0.03002
2	0	0.04451	0.00519	0.01121	0.02922
2	0.5	0.05556	0.00334	0.00905	0.03363

Table 4.6: Effect of B and K on the estimated standard error of the scaled pAUC indices where $\theta_{0,1} = 0.79$, $A = 2$, and $N = 10$

K	B value	$\theta_{0,1}$	$\theta_{0,0.05}$	$\theta_{0,0.1}$	$\theta_{0,0.3}$
0.5	-0.5	0.03828	0.12490	0.11334	0.08120
0.5	0	0.04400	0.09876	0.11191	0.09986
0.5	0.5	0.05868	0.06048	0.08746	0.12255
1.0	-0.5	0.04892	0.15340	0.13845	0.10469
1.0	0	0.04703	0.11386	0.12104	0.10411
1.0	0.5	0.05773	0.06735	0.09392	0.11804
2.0	-0.5	0.04814	0.14908	0.13346	0.10006
2.0	0	0.04451	0.10379	0.11211	0.09739
2.0	0.5	0.05556	0.06671	0.09046	0.11209

$N = 10$. When $K = 0.5$ the full AUC increases as the dependence of the test accuracy on threshold, B, increases from -0.5 to 0.5. However, when $K > 1$, B has much less of an influence. The relationship between K, B and the variance of the partial area and scaled partial area are not clear. For two of the partial areas, $\theta_{0,0.05}$ and $\theta_{0,0.1}$ the standard errors decrease with B. For $\theta_{0,0.05}$ and $\theta_{0,0.1}$ the standard

error is maximum when $B = -0.5$ and for $\theta_{0,0.3}$, a much wider range of FPR values the standard error is maximum when $B = 0.5$. The same is evident for the standard errors of the scaled pAUC.

In Fig. 4.3 we have computed the sample sizes for a range of values for K. From this figure one can note that the precision wasn't enhanced much when doubling the control group or when doubling the disease group. Therefore an equal number ($K = 1$) of subjects with and without the disease is recommended.

4.6 Remarks

We have proposed a simple method for determining the required sample size for studies of diagnostic accuracy. The method applies to studies that involve the two main indices associated with the SROC curve, namely, the area under the full SROC curve and the partial area under the curve.

These SROC indices have different applications in diagnostic accuracy studies. The area under the full curve is particularly useful in evaluating a new test or procedure. One can use it to determine if the new test had a diagnostic ability (that is, $AUC > 0.5$) and to compare its overall diagnostic ability with standard tests.

The partial area under the SROC curve is an excellent index for describing the accuracy of a test for a particular setting. This is because we can transform the value

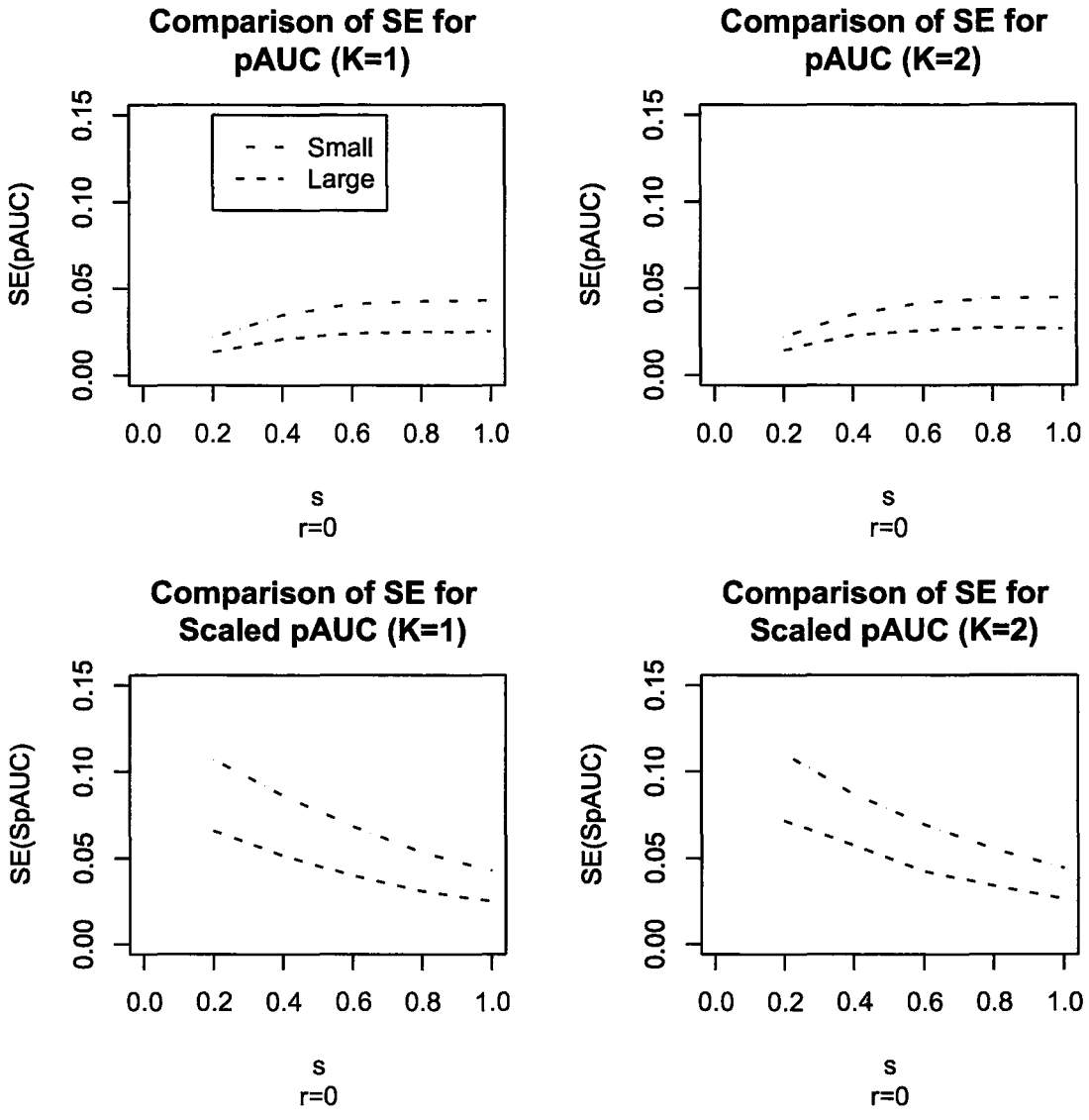

Figure 4.3: Plots of the standard errors of the pAUC and the scaled pAUC when $K = 1$ and $K = 2$. The plots show that the precision isn't enhanced much by doubling the control group.

of the partial area index to the more familiar zero to one scale for interpretation purposes. The methods described here provide unbiased estimates of the variance of this index.

We have shown that the homogeneous model would be the ideal model when dealing with the partial AUC summary statistic for small study sample sizes in a meta-analysis of diagnostic testing. When $B = 0$ there is no relationship between OR and threshold S and therefore there is no between study variation. The AUC statistic is symmetric with respects to B. The partial AUC does not posses this property and hence it will show far greater dependence on the degree of inter-study heterogeneity (Walter, 2002). This is one of many reasons why the homogeneous logistic threshold SROC model is preferred when taking the partial AUC of an SROC curve.

Walter (2002) showed that the AUC of an SROC curve and its standard error are remarkably robust against heterogeneity, which is an attractive feature for a summary measure. In contrast, the partial AUC is not robust to heterogeneity, especially when only a small portion of the SROC curve is used (Walter, 2002). The pAUC, the scaled pAUC and their standard errors depend strongly on the degree of truncation and the particular type and strength of inter-study heterogeneity.

The scaled pAUC is used to restore the numerical range of the summary measure to be between zero and one. The standard errors for the scaled pAUC when compared to the standard errors for the pAUC is much larger, indicating a lost in precision in the scaled summary measure. However, when $B = 0$, in most cases the standard errors for the scaled pAUC for region where FPR is between 0 and 0.15 are

lower than for regions where FPR is between 0.20 and 0.50. For a small number of studies in a meta-analysis, the scaled pAUC has a lower standard error than when $r = 0$ and s is between 0.05 and 0.15. The pAUC for the homogeneous model works for the clinical region in certain situations.

Chapter 5

Application

We will illustrate the usefulness of the partial area under the curve in meta-analysis by reanalyzing a meta-analysis on the diagnostic performance of two magnetic resonance angiography techniques: 3D gadolinium-enhanced (3D-GD) and 2D time of flight (2D-TOF) for detecting peripheral arteriosclerotic occlusive disease (Nelemans et al., 2000). The separate meta-regression analysis yielded an intercept of 4.13 and a slope of 0.41 for 2D-TOF. For 3D-GD, these values were, 5.93 and -0.37, respectively.

Separate summary ROC curves were constructed for studies reporting on 2D time-of-flight MR angiography and for studies reporting on 3D gadolinium-enhanced MR angiography. Fig. 5.1 shows the separate summary ROC curves for 2D-TOF and 3D-GD. Most notably for studies on 2D-TOF MR angiography, there are consid-

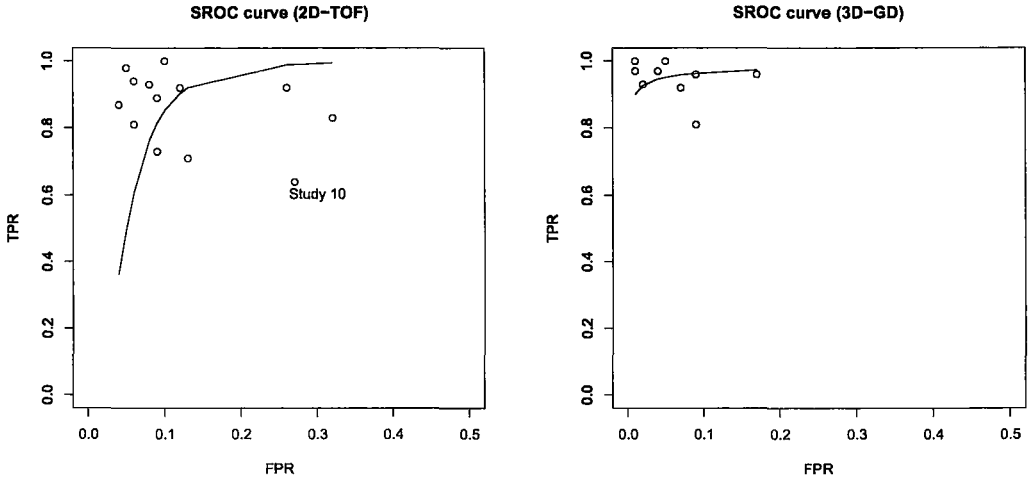

Figure 5.1: SROC curves for 2D time-of-flight (2D-TOF) MR angiography (left) and for studies reporting on 3D gadolinium-enhanced (3D-GD) MR angiography (right).

erable discrepancies between the SROC curve and the observed data points. These findings indicate that differences in the threshold for a positive examination result explain only a small part of the variation between study results. A method to explore this variation was done by Nelemans et al. (2000), where they looked at other sources of variation by adding variables to the linear regression model and compared relative diagnostic ORs. They found that about half of the between-study variation was due to four factors: (a) the intercept A; (b) the variable S, which is a measure of the leniency of the threshold for a positive examination result; (c) the MR angiographic examination type; and (d) the extent of image evaluation. Another method would be to look at eliminating the studies that have high FPR values in order to

reduce the heterogeneity among studies.

Table 5.1: Partial AUC for the on the diagnostic performance of two magnetic resonance angiography techniques: 3D gadolinium-enhanced (3D-GD) and 2D time of flight (2D-TOF).

s value	2D-TOF	Reduced 2D-TOF	3D-GD
1.00	0.94 (0.029)	0.95 (0.021)	0.98 (0.014)
0.05	0.01 (0.021)	0.02 (0.021)	0.05 (0.003)
0.10	0.04 (0.035)	0.06 (0.027)	0.09 (0.003)
0.15	0.09 (0.035)	0.10 (0.027)	0.14 (0.002)
0.20	0.14 (0.034)	0.15 (0.026)	0.19 (0.003)

The partial AUC indices, $\theta_{0,0.05}$,$\theta_{0,0.10}$, $\theta_{0,0.15}$, and $\theta_{0,0.20}$ were analyzed for the two techniques. The partial AUC, scaled partial AUC and their standard errors for 2D-TOF and 3D-GD MR angiography are found in Table 5.1 and 5.2. The interpretation of the scaled pAUC where $s = 0.05$ for the 3D-GD data is that, conditional on the FPR value being no greater than 5%, the SROC curve has achieved a partial area which is 92 per cent of its maximum potential in this restricted region. We may also consider this value in relation to the corresponding area for an informative test. For the 2D-TOF data when FPR is limited to values below 0.1, the partial AUC, $\theta_{0,0.1}$ is 0.045 (0.035). The corresponding scaled pAUC value is 0.45 (0.351). The partial area in the triangle representing the performance of a random diagnostic test up to FPR=0.1 is 0.005. To interpret these results, we note that 2D-TOF MR angiography technique is performing better than a random test in the range of

clinical interest for FPR between 0.0 and 0.05. However, there are large standard errors for the scaled pAUC in the 2D-TOF MR angiography data. These extreme standard errors for the scaled pAUC in 2D-TOF may be due to the outline point shown in Fig. 5.1. Study 10 has a smaller TPR value and higher FPR value than the 12 other studies. This outlying point may be caused by random error, differences in study methodology, population or test characteristics.

Table 5.2: Scaled partial AUC for the diagnostic performance of two magnetic resonance angiography techniques: 3D gadolinium-enhanced (3D-GD) and 2D time of flight (2D-TOF).

s value	2D-TOF	Reduced 2D-TOF	3D-GD
1.00	0.94 (0.029)	0.95 (0.021)	0.98 (0.014)
0.05	0.19 (0.426)	0.33 (0.414)	0.92 (0.053)
0.10	0.45 (0.351)	0.57 (0.272)	0.94 (0.026)
0.15	0.60 (0.235)	0.69 (0.179)	0.95 (0.017)
0.20	0.69 (0.168)	0.76 (0.128)	0.96 (0.013)

In diagnostic testing, it is critical to maintain a high TPR in order not to miss detecting subjects with disease. We analyzed the 2D-TOF data again with a restriction to the meta-analysis to include studies that have a TPR value greater than 0.70 only. This restriction would then exclude study 10 from the meta-analysis. The reduced 2D-TOF data yielded an intercept of 4.33 and a slope of 0.31. The partial AUC, scaled pAUC and their standard errors for the reduced 2D-TOF data can be found in Tables 5.1 and 5.2. When we eliminated the study with a low TPR

value we got higher values of the scaled pAUC and lower standard errors. However, when we compared the two techniques we can conclude that the 3D-GD technique was clearly superior to that of 2D-TOF MR angiography. The extreme variation in 2D-TOF MR angiography data needs to be looked into further to understand the between study variation. In this case an hierarchical SROC model which takes into account the between study variation is recommended.

Chapter 6

General Observations and Guidelines

6.1 ROC analysis

The ROC curve has been used in single study data analysis of diagnostic testing for many years. Indices for the ROC curve are available as one-number summary measures for easy interpretation of diagnostic tests. The summary measures assessed in Chapter 1 included the AUC, Q^*, ASC, and pAUC statistics. Suggestions toward these indices have been proposed in Chapter 1. In this section three observations for the use of the summary measures for a single ROC curve are presented.

Advantages of Logistic Scores

If the scores are normally distributed with different means but same scale parameters for the disease and control groups the estimated ORs will be symmetric for all rates. However, if the scores for the disease and control groups came from logistic distributions with same scale but different mean parameters then the estimated ORs will be constant for any given threshold. Also, the true odds ratio can be determined from the parameters of the logistic distribution. If the scores come from normal distributions the OR will vary between thresholds. Ideally, the scores would come from homogeneous logistic distributions, however, this is not always the case.

AUC or Q^* Statistics?

The area under the ROC curve has been criticized for its dependency on an irrelevant region where FPR is high. The Q^* statistic has been suggested as an alternative measure of the ROC curve. The point where sensitivity equals specificity is denoted as the Q^* statistic. Figure 2.2 showed an example where the AUC and Q^* statistics are highly correlated implying that the Q^* statistic carries the same information as the AUC. Note that Q^* has a simple interpretation as a function of ORs under the homogeneous logistic threshold model. We recommend using the partial area under the ROC curve as a summary measure in situations where small values of FPR are

of interest.

Disadvantages of ASC

Similar to the AUC, the area swept by the ROC curve is related to the probability that the test will correctly rank a disease and control pair of subjects. The ASC, however, is only informative for a parametric ROC curve or an empirical ROC curve defined by a small number of line segments. When a step ROC curve with no ties is generated using the runs and rises defined by the points when sensitivity and specificity changes the ASC will always equal 1/2. Therefore, when a step ROC curve is employed the ASC summary measure is of no use. Perhaps ASC could be applied to a smoothed version of the step ROC curve but that is a topic of further investigation.

6.2 SROC analysis

The SROC curve and its summary measures are being used more for meta-analysis of diagnostic testing. In particular, many researchers are using the partial area under the SROC curve as a summary measure to describe the clinically relevant region. The relevant region is when the false positive rates are low. In order to correctly utilize the pAUC statistic, guidelines must be set for proper use.

Inclusion Criteria

The inclusion criteria for a meta-analysis of diagnostic testing should always consider rejecting studies with low TPR. For example, an inclusion of studies with TPR ≥ 0.70 maybe a suitable restriction. Variations in study design, population and subject characteristics can contribute to possible low TPR values in some studies when compared with other studies analyzing the same diagnostic test. Studies with low, outlying TPR values are seldom clinically relevant.

Outlier Studies

The simulations in Section 4 showed that the best results came from a homogeneous logistic threshold model where $B = 0$. This was because the diagnostic thresholds were similar for each study in the meta-analysis. The standard errors for the homogeneous model were smaller than those for the heterogeneous model in all the indices evaluated. Several indices were compared, including the full AUC and partial AUC where s equaled 0.05, 0.10, 0.15, and 0.20 for small and large study sample sizes. The symmetric property found in the homogeneous model showed there was little between-study variation. If $B \neq 0$, then we have a heterogeneity model. In this case, the diagnostic odds ratio depends on the threshold used for a given study.

A method for achieving an SROC curve with a smaller B value could be to

eliminate studies with low TPR values. These values may be considered as outlying studies. Outlier studies are easily distinguishable in the plot of an SROC curve. For example, in the application, study 10 for 2D time-of-flight MR angiography in Nelemans' data had a lower TPR value than the other studies and when study 10 was extracted from the meta-analysis the B value decreased to 0.31 from 0.41.

Numbers of Control and Disease Subjects

In Section 4.5 some cases where disease (control) study sample sizes were twice as large as the control (disease) study sample sizes were examined. The analysis showed that the sample size between the disease and control groups for unbalanced ($K \neq 1$) or balanced ($K = 1$) studies had little difference in the SROC analysis. The estimated standard errors for AUC and pAUC summary measures were relatively the same for both balanced and unbalanced study data. Further work is needed to illustrate the trade-off between the number for subjects with and without the disease in each study for less and more extreme cases of imbalance. One situation could be to analyze the effects of the SROC indices when only one study is unbalanced.

Interpretations of pAUC

The partial AUC can be interpreted as the probability that a disease and control pair of test results will be correctly ranked, conditional on the non-case value falling within the restricted range of the curve (Dobb & Pepe, 2003). Since the pAUC in equation (3.23) is less than or equal to one, the pAUC index cannot attain the maximum value of 1 that is achievable by the AUC index, but instead has a maximum value of $s-r$ (Walter, 2005). In order to retain the desirable property that the pAUC ranges from zero to one it is recommend to use the scaled partial AUC statistic. For example the interpretation for the scaled pAUC in the application found in Section 5 for the 2D time-of-flight MR angiography data with $s = 0.15$ is, conditional on the FPR values being no greater than 15 per cent, the SROC curve has achieved a partial area which 60 per cent of its maximum potential in this restricted range. The pAUC value can be compared to the corresponding area for an informative test, where the SROC curve lies on the diagonal line. This would show that the pAUC has more or less information than if the study was selected at random with an AUC of 0.5. In the above example for an uninformative test up to a maximum FPR of 15 per cent would be 0.01125.

Chapter 7

Conclusions

Receiver operating characteristic curves and their associated indices are valuable tools for assessment of the accuracy of diagnostic tests. Chapter 1 reviewed the AUC, Q^*, ASC and pAUC summary measures. In examining these indices suggestions were made. In particular, AUC and Q^* statistics were used to determine if the ROC curve is close to the region where FPR is 0 and TPR is 1. The AUC and Q^* statistics were shown to be highly correlated ($r = 0.89$). This result indicates that AUC and Q^* express more or less the same thing so it wouldn't matter which statistic is used. The ASC is also similar to the AUC index, however the interpretation of the ASC is irrelevant when dealing with a step ROC curve. The concepts and approaches used in ROC analysis were found to be beneficial in understanding SROC analysis.

Summary receiver operating characteristic analysis is increasingly popular for

meta-analysis of diagnostic test validity. However, it is only meaningful when similar endpoints, diagnostic thresholds, study quality and test characteristics are compared. The AUC and Q^* statistics were used to compare results from different SROC analyses. Partial AUC may be used if specificity values are limited, with interpretation on an individual meta-analysis basis.

Our simulations illustrated that the partial AUC index can be used to compare diagnostic performances in SROC curves. Comparison of the AUC values is practically meaningless because, in practice, all points on the curve will not have the same clinical relevance. The partial area index is defined for sensitivity levels of clinical interest; however, there is uncertainty in the partial area index to the extent that the group of sensitivity levels are well chosen.

Although we used magnetic resonance angiography techniques for illustration purposes in this paper, the potential usefulness of the pAUC is not limited to that application but extends to the evaluation of any diagnostic test that must maintain a high sensitivity level, clinically. The partial AUC is more meaningful than the conventional AUC index in such situation because it reflects the portion of the SROC curve that is clinically relevant.

Presently, the partial AUC index is being used more in meta-analysis of diagnostic testing; however guidelines and strategies have not been established. In this thesis, we collaborated our simulation findings to incorporate broad guidelines when

using the partial AUC index in an SROC analysis.

Sometimes the sensitivity and specificity will be available for different thresholds within the same study. Depending on the predetermined diagnostic threshold and the amount of literature available, the most appropriate threshold should be chosen for the analysis. With enough literature available, it is possible to perform SROC analysis for different thresholds of the same test. The AUC, Q^* or pAUC would be used, where appropriate, to compare the accuracy of the same test for different thresholds. This requires multiple analysis which are often published separately (Jones et al., 2005).

There is a lack of understanding by clinicians of the concepts and interpretations of the partial AUC for an SROC curve. It is our desire that the guidelines on the use of the partial area under the SROC curve presented in this thesis will help clinicians with the comprehension of the pAUC index. We expect as the pAUC for the SROC curve becomes more popular and understanding grows, interpretation of the partial AUC index will become easier.

Appendix A

Source R codes for Graphs

Figure 1

```
> plot.OR<-function (tgr, filename = "plotOR.pdf")
{
    pdf(filename)
    m <- matrix(c(1, 2), 1, 2)
    layout(m)
    logisOR <- function(t, mu1, mu2 = 0, sd1 = 1, sd2 = 1) {
        (1 - plogis(t, mu1, sd1)) * plogis(t, mu2, sd2)/(plogis(t,
            mu1, sd1) * (1 - plogis(t, mu2, sd2)))
    }
    logisOR <- function(t, mu1, mu2 = 0, sd1 = 1, sd2 = 1) {
        (1 - plogis(t, mu1, sd1)) * plogis(t, mu2, sd2)/(plogis(t,
            mu1, sd1) * (1 - plogis(t, mu2, sd2)))
    }
    plot(tgr, logisOR(tgr, 1), type = "l", ylim = c(0, 20), ylab
        = "Odds Ratio", xlab = "Scores", main = "Logistic
        Distribution")
    plot(tgr, normOR(tgr, 1), type = "l", ylim = c(0, 20), ylab
        = "Odds Ratio", xlab = "Scores", main = "Normal
        Distribution")
    dev.off(dev.cur())
}
```

Figure 2

```
> roc.boot.plot
function (scored, scorec, B, filename = "rocboot.pdf") {
    pdf(filename)
    rocboot <- roc.boot(scored, scorec, B = 2000)
    pairs(rocboot)
    dev.off(dev.cur())
}
```

Figure 3

```
> ASCplot1
function (FPR, TPR, filename = "ASCplot1.pdf") {
    pdf(filename)
    dat <- cbind(c(0, 1), c(0, 1))
    dat2 <- cbind(c(0, 1), c(0, 0.5))
    plot(FPR, TPR, type = "l")
    lines(dat, lty = 2)
    lines(dat2, lty = 2)
    text(0.2, 0.3, "A")
    text(0.5, 0.4, "B")
    text(0.9, 0.7, "C")
    dev.off(dev.cur())
}
```

Figure 4

```
> ASCplot2
function (FPR, TPR, filename = "ASCplot2.pdf") {
    pdf(filename)
    plot(FPR, TPR, type = "l")
    dat <- cbind(c(0, 0.25), c(0, 0.5))
    lines(dat, lty = 2)
    text(0.025, 0, "P")
    text(0.275, 0.5, "Q")
    text(0, 0.525, "R")
    dat2 <- cbind(c(0, 0.25), c(0, 0.75))
    lines(dat2, lty = 2)
```

```
    text(0.25, 0.775, "S")
    dat3 <- cbind(c(0, 0.75), c(0, 0.75))
    lines(dat3, lty = 2)
    text(0.775, 0.75, "T")
    dat4 <- cbind(c(0, 0.75), c(0, 1))
    lines(dat4, lty = 2)
    text(0.725, 1, "U")
    dat5 <- cbind(c(0, 1), c(0, 1))
    lines(dat5, lty = 2)
    text(1, 0.975, "V")
    dev.off(dev.cur())
}
```

Figure 5

```
> ASCplot3
function (FPR, TPR, filename = "ASCplot3.pdf") {
    pdf(filename)
    dat <- cbind(c(0, 1), c(0, 1))
    plot(FPR, TPR, type = "l")
    legend(0.6, 0.1, c("AUC = 0.84", "Q* = 0.75", "ASC = 0.5"))
    lines(dat, lty = 2)
    dev.off(dev.cur())
}
```

Figure 6

```
> plot.SROChom<-function (FPR, A1, B1 = 0, m1 = 25, m2 = 20,
     filename = "plotSROChom.pdf")
{
    pdf(filename)
    TPR1.sroc <- matrix((exp(A1/(1 - B1)) * (FPR/(1 - FPR))^((1 +
        B1)/(1 - B1)))/(1 + exp(A1/(1 - B1)) * (FPR/(1 - FPR))^((1 +
        B1)/(1 - B1))))
    xx1 <- rbinomtableHet(A1, B1, FPR, m1, m2)
    plot(xx1[, 6], xx1[, 5], ylim = c(0, 1), xlim = c(0, 1),
        xlab = "False Positive Rate", ylab = "True Positive Rate",
        main = "SROC curve for the Homogeneous Case")
    lines(sort(FPR), sort(TPR1.sroc), type = "l", ylim = c(0,
```

```
        1), xlim = c(0, 1), lty = 2, col = 1)
    abline(0, 1)
    dev.off(dev.cur())
}
> plot.SROChet<-function (FPR, A2, B2, m1, m2, filename =
      "plotSROChet.pdf")
{
    pdf(filename)
    TPR2.sroc <- matrix((exp(A2/(1 - B2)) * (FPR/(1 - FPR))^((1 +
        B2)/(1 - B2)))/(1 + exp(A2/(1 - B2)) * (FPR/(1 - FPR))^((1 +
        B2)/(1 - B2))))
    xx2 <- rbinomtableHet(A2, B2, FPR, m1, m2)
    plot(xx2[, 6], xx2[, 5], ylim = c(0, 1), xlim = c(0, 1),
        xlab = "False Positive Rate", ylab = "True Positive Rate",
        main = "SROC curve for the Heterogeneous Case")
    lines(sort(FPR), sort(TPR2.sroc), type = "l", ylim = c(0,
        1), xlim = c(0, 1), lty = 2, col = 1)
    abline(0, 1)
    dev.off(dev.cur())
}
```

Figure 7

```
SEpAUC.plot2 function (xx1, xx2, xx3, xx25, xx26, xx27, filename =
"SEpAUCplot2.pdf") {
    pdf(filename)
    m <- matrix(c(1, 2, 3, 4), 2, 2)
    layout(m)
    plot(xx1[, 10], xx1[, 2], xlab = "s", ylab = "sepAUC", main =
    "Comparison of B values for SE of \n  pAUC (Small samples)",
        sub = "r=0 and A=2", type = "l", lty = 2, ylim = c(0,
            0.15))
    lines(xx1[, 10], xx2[, 2], lty = 3)
    lines(xx1[, 10], xx3[, 2], lty = 4)
    legend(0.1, 0.15, c("B=-0.5", "B=0", "B=0.5"), lty = c(2,
        3, 4))
    plot(xx1[, 10], xx1[, 4], xlab = "s", ylab = "seSpAUC", main =
    "Comparison of B values for SE of \n Scaled pAUC (Small
    samples)", sub = "r=0 and A=2", type = "l", lty = 2, ylim
```

```
    = c(0, 0.15))
    lines(xx1[, 10], xx2[, 4], lty = 3)
    lines(xx1[, 10], xx3[, 4], lty = 4)
    plot(xx25[, 10], xx25[, 2], xlab = "s", ylab = "sepAUC",
        main = "Comparison of B values for SE of \n pAUC (Large
    samples)", sub = "r=0 and A=2", type = "l", lty = 2, ylim =
    c(0, 0.15))
    lines(xx25[, 10], xx26[, 2], lty = 3)
    lines(xx25[, 10], xx27[, 2], lty = 4)
    plot(xx25[, 10], xx25[, 4], xlab = "s", ylab = "seSpAUC",
        main = "Comparison of B values for SE of \n Scaled pAUC
    (Large samples)", sub = "r=0 and A=2", type = "l", lty =
    2, ylim = c(0, 0.15))
    lines(xx25[, 10], xx26[, 4], lty = 3)
    lines(xx25[, 10], xx27[, 4], lty = 4)
    dev.off(dev.cur())
}
```

Figure 8

```
> Sample.Comp
function (xx2, xx3, xx8, xx9, xx20, xx21, xx26, xx27, filename =
"SampleComp.pdf") {
    pdf(filename)
    m <- matrix(c(1, 2, 3, 4), 2, 2)
    layout(m)
    plot(xx1[, 10], xx2[, 4], xlab = "s", ylab = "seSpAUC", main =
    "Comparison of Sample size for \n SE of Scaled pAUC",
        sub = "r=0, A=2 and B=0", type = "l", lty = 2, ylim = c(0,
            0.15))
    lines(xx1[, 10], xx8[, 4], lty = 3)
    lines(xx1[, 10], xx20$seSpAUC, lty = 4)
    lines(xx1[, 10], xx26[, 4], lty = 5)
    plot(xx1[, 10], xx3[, 4], xlab = "s", ylab = "seSpAUC", main =
    "Comparison of Sample size for \n SE of Scaled pAUC",
        sub = "r=0, A=2 and B=0.5", type = "l", lty = 2, ylim = c(0,
            0.15))
    lines(xx1[, 10], xx9[, 4], lty = 3)
    lines(xx1[, 10], xx21$seSpAUC, lty = 4)
```

```
    lines(xx1[, 10], xx27[, 4], lty = 5)
    plot(xx1[, 10], xx2[, 2], xlab = "s", ylab = "sepAUC", main =
    "Comparison of Sample size for \n SE of pAUC",
        sub = "r=0, A=2 and B=0", type = "l", lty = 2, ylim = c(0,
            0.15))
    lines(xx1[, 10], xx8[, 2], lty = 3)
    lines(xx1[, 10], xx20$sepAUC, lty = 4)
    lines(xx1[, 10], xx26[, 2], lty = 5)
    legend(0.15, 0.15, c("All small", "One large", "Half/Half",
        "All Large"), lty = c(2, 3, 4, 5))
    plot(xx1[, 10], xx3[, 2], xlab = "s", ylab = "sepAUC", main =
    "Comparison of Sample size for \n SE of pAUC",
        sub = "r=0, A=2 and B=0.5", type = "l", lty = 2, ylim = c(0,
            0.15))
    lines(xx1[, 10], xx9[, 2], lty = 3)
    lines(xx1[, 10], xx21$sepAUC, lty = 4)
    lines(xx1[, 10], xx27[, 2], lty = 5)
    dev.off(dev.cur())
}
```

Appendix B

Source R codes for ROC Analysis

```
> roc.boot
function (scored, scorec, B = 200) {
    md <- length(scored)
    mc <- length(scorec)
    t(apply(cbind(matrix(sample(scored, md * B, replace = T),
        nrow = B), matrix(sample(scorec, mc * B, replace = T),
        nrow = B)), 1, roc.stats, md = md))
}
> roc.stats
function (sall, md, ...) {
    mc <- length(sall) - md
    sind <- rep(c(0, 1), c(md, mc))
    stab <- table(sall, sind)
    stab <- stab[nrow(stab):1, ]
    tpr <- c(0, cumsum(stab[, 1]))/md
    fpr <- c(0, cumsum(stab[, 2]))/mc
    AUC <- sum(0.5 * (tpr[-1] + tpr[-length(tpr)]) * diff(fpr))
    Qi <- match(TRUE, tpr + fpr >= 1)
    Qstar <- (tpr[Qi - 1] * (fpr[Qi] - fpr[Qi - 1]) + (1 - fpr[Qi -
        1]) * (tpr[Qi] - tpr[Qi - 1]))/((fpr[Qi] - fpr[Qi - 1]) +
        (tpr[Qi] - tpr[Qi - 1]))
    c(AUC = AUC, Qstar = as.numeric(Qstar))
}
```

Appendix C

Source R codes for SROC Analysis

Random Binomial Tables

```
rbinomtableHet<-function (A, B, FPR, m1, m2)
{
    table1 <- NULL
    TPR <- matrix((exp(A/(1 - B)) * (FPR/(1 - FPR))^((1 + B)/(1 -
        B)))/(1 + exp(A/(1 - B)) * (FPR/(1 - FPR))^((1 + B)/(1 -
        B))))
    for (i in 1:length(FPR)) {
        a <- rbinom(1, m1, TPR[i])
        c <- rbinom(1, m2, FPR[i])
        table <- cbind(a = a, b = m1 - a, c = c, d = m2 - c,
            TPR1 = a/m1, FPR1 = c/m2)
        table1 <- rbind(table1, table)
    }
    table1
}
```

Moses Model

```
> mosesModel<-function (a, b, c, d)
{
    Uhat <- log((c + 0.5)/(d + 0.5))
    Vhat <- log((a + 0.5)/(b + 0.5))
    Dhat <- Vhat - Uhat
    Shat <- Vhat + Uhat
```

```
    fit <- lm(Dhat ~ Shat)
    sum.fit <- summary(fit, cor = T)
    A <- sum.fit$coef[1]
    B <- sum.fit$coef[2]
    seA <- sum.fit$coef[1, 2]
    seB <- sum.fit$coef[2, 2]
    corAB <- sum.fit$cor[2]
    covAB <- corAB * seA * seB
    list(A = A, B = B, seA = seA, seB = seB, corAB = corAB,
    covAB = covAB)
}
```

SROC indices

```
> SROC1
function (fpr, A, B, r, s, seA, seB, cov_AB)
{
    var_A <- seA^2
    var_B <- seB^2
    SROC_1 <- function(fpr) {
        tpr.roc.hom <- (exp(A/(1 - B)) * (fpr/(1 - fpr))^((1 +
            B)/(1 - B)))/(1 + exp(A/(1 - B)) * (fpr/(1 - fpr))^((1 +
            B)/(1 - B)))
    }
    x <- integrate(SROC_1, r, s, stop.on.error = FALSE)
    pAUC <- x$value
    p <- (1 + B)/(1 - B)
    integ_Af <- function(fpr) {
        Af <- ((fpr/(1 - fpr))^p)/((1 + ((fpr/(1 - fpr))^p) *
            exp(A/(1 - B)))^2)
    }
    integ_Bf <- function(fpr) {
        Bf <- (((fpr/(1 - fpr))^p) * (A + 2 * log(fpr/(1 -
        fpr))))/((1 + ((fpr/(1 - fpr))^p) * exp(A/(1 - B)))^2)
    }
    dAUC_A <- (1/(1 - B)) * exp(A/(1 - B)) * integrate(integ_Af,
        r, s, stop.on.error = FALSE)$value
    dAUC_B <- ((1/(1 - B))^2) * exp(A/(1 - B)) * integrate(integ_Bf,
        r, s, stop.on.error = FALSE)$value
```

```
    sepAUC <- sqrt((dAUC_A^2 * var_A) + (dAUC_B^2 * var_B) +
        (2 * dAUC_A * dAUC_B * cov_AB))
    ScaledpAUC <- pAUC/(s - r)
    seScaledpAUC <- sepAUC/(s - r)
    Qstar <- (exp(A/2))/(1 + exp(A/2))
    seQstar <- (sqrt(exp(A)) * seA)/(2 * (sqrt(exp(A)) + 1)^2)
    list(pAUC = pAUC, sepAUC = sepAUC, ScaledpAUC = ScaledpAUC,
        seScaledpAUC = seScaledpAUC, Qstar = Qstar, seQstar =
          seQstar)
}
```

Bootstrapping function

```
> sroc.stats<-function (FPR, A, B, m1, m2, r, s)
{
    xx <- rbinomtableHet(A, B, FPR, m1, m2)
    model <- mosesModel(xx[, 1], xx[, 2], xx[, 3], xx[, 4])
    AA <- modelSA
    BB <- modelSB
    SROC_1 <- function(fpr) {
        tpr.roc.hom <- (exp(AA/(1 - BB)) * (fpr/(1 - fpr))^((1 +
            BB)/(1 - BB)))/(1 + exp(AA/(1 - BB)) * (fpr/(1 -
            fpr))^((1 + BB)/(1 - BB)))
    }
    x <- integrate(SROC_1, r, s)
    pAUC <- x$value
    ScaledpAUC <- pAUC/(s - r)
    Qstar <- (exp(AA/2))/(1 + exp(AA/2))
    c(pAUC, ScaledpAUC, Qstar)
}

> sroc.boot<-function (FPR, Bt = 200, r, s, A, B, m1, m2)
{
    n <- length(FPR)
    t(apply(matrix(sample(FPR, n * Bt, replace = T), nrow = Bt),
        1, sroc.stats, r = r, s = s, A = A, B = B, m1 = m1, m2 = m2))
}
```

REFERENCES

1. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. J Math Psych 1975;12:387-415.

2. Cochrane reviews of diagnostic test accuracy. The Cochrane Collaboration Web site. Available at: www.cochrane.org/newslett/ccnews31-lowres/pdf. Accessed August 25, 2007.

3. Cox, D. R. Analysis of Binary Data. 1970, Spottiswoode, Ballantyne & Co Ltd. London and Colchester.

4. Delong E. R, Delong D, Clarke-Pearson D. Comparing the area under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837-845.

5. Dobb E, Pepe M. Partial AUC Estimation and Regression. Biometrics 2003;59:614-623.

6. Dobb E & Cai T. Regression Analysis for the Partial Area Under the ROC curve. Harvard University Biostatistics Working Paper Series 2006:1-39

7. Fog A. Biased Urn Theory. http://cran.r-project.org/doc/vignettes/BiasedUrn/UrnTheory.pdf, 2007.

8. Gatsonis C, Paliwal P. Meta-analysis of diagnostic and screening test accuracy evaluations: methodologic primer. AJR 2006;187:271-281.

9. Glas A, Lijmer J, Prins M, Bonsel G, Bossuyt P. The diagnostic odds ratio: a single indicator of test performance. J Clin Epidemiol 2003;56:1129-1135.

10. Henley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29-36.

11. Irwig L, Tosteson ANA, Gatsonis C, et al. Guidelines for meta-analysis evaluating diagnostic tests. Ann Intern Med 1994;120:667-676.

12. Irwig L, Macaskill P, Glasziou P, Fahey M. Meta-analytic method for diagnostic test accuracy. J Clin Epidemiol 1995;48:119-130.

13. Jones CM, Athanasiou T. Summary receiver operating characteristic curve analysis techniques in the evaluation of diagnostic tests. Ann Thorac Surg 2005;79:16-20.

14. Kardaun JW, Kardaun OJ. Comparative diagnostic performance of three radiological procedures for the detection of lumbar disk herniation. Methods Inf Med 1990;29:12-22.

15. Lee WC,Hsiao C. Alternative summary indices for the receiver operating characteristic curve. Epidemiology 1996;7:605-611.

16. Littenberg B, Moses LE. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. Med Decis Making 1993;13:313-321.

17. McClish D. Analyzing a portion of the ROC curve. Med Decs Making 1989;9:190-195.

18. Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. Stat Med 1993;12:1293-1316.

19. Nelemans PJ, Leiner T, Henrica CW, Joseph M, Peripheral Arterial Disease: Meta-analysis of the Diagnostic Performance of MR Angiography. Radiology 2000;217:105-114.

20. Nelson T. ROC curves and measure of discrimination accuracy: a reply to Swets. Psychol Bull 1986;99:128-132.

21. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med 2001;20:2865-2884.

22. Scheidler J, Hricak H, Yu K, Subak L, Segal M. Radiological evaluation of lymph node metastases in patients with cervical cancer: a meta-analysis. JAMA 1997;278:1096-1101.

23. Shapiro DE. Issues in combining independent estimates of the sensitivity and specificity of a diagnostic test. Acad Radiol 1995;2:S37-S47.

24. Siadaty MS, Shu J. Proportional odds ratio model for comparison of diagnostic test in meta-analysis. BMC Med Res Methodol 2004, 4:27.

25. Van der Schouw YT, Straatman H, Verbeek A. ROC curves and the areas under them for dichotomized test: empirical findings for logistically and normally distributed diagnostic test results. Med Decis Making 1994;14:374-381.

26. Walter SD. Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data. Stat Med 2002;21:1237-1256.

27. Walter SD. The partial area under the summary ROC curve. Stat Med 2005;24:2025-2040.

28. Walter SD, Macaskill P. SROC Curve Encyclopedia of Biopharmaceutical Statistics 2004.

29. Walter SD, Sinuff T. Studies reporting ROC curves of diagnostic and prediction data can be incorporated into meta-analyses using corresponding odds ratios. J Clin Epidemiol 2007;60:530-534.

30. Whitehead A. Meta-analysis of Controlled Clinical Trials. 2002, John Wiley & Sons Ltd.

31. Youden W. J. Index for rating diagnostic test. Cancer 1950;3(1):32-35.

32. Zhang D, Zhou XH, Freeman DH, Freeman JL. A non-parametric method for the comparison of partial areas under ROC curves and its application to large health care data sets. Stat Med 2002;21:701-715.

33. Zhang J, Properties of the Projected Length of the Curve and Area Swept out by the Curve indices for the Summary Receiver Operating Characteristic. Working Paper. 2007:1-27.