Estimation of ROC Curve with Complex Survey Data

by Wenliang Yao

M.S. in Statistics, May 2008, The George Washington University

A dissertation submitted to

The Faculty of
The Columbian College of Arts and Sciences
of The George Washington University
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

January 31, 2013

Dissertation directed by

Zhaohai Li
Professor of Statistics and Biostatistics

Barry I. Graubard
Senior Investigator, National Cancer Institute

The Columbian College of Arts and Sciences of The George Washington University certifies that Wenliang Yao has passed the Final Examination for the degree of Doctor of Philosophy as of November 30, 2012. This is the final and approved form of the dissertation.

Estimation of ROC Curve with Complex Survey Data

Wenliang Yao

Dissertation Research Committee:

Zhaohai Li, Professor of Statistics and Biostatistics;
Dissertation Co-Director

Barry I. Graubard, Senior Investigator, National Cancer Institute; Dissertation Co-Director

Michael D. Larsen, Associate Professor of Statistics;
Committee Member

Qing Pan, Assistant Professor of Statistics;
Committee Member

Acknowledgments

I would like to take this opportunity to express my gratitude to those who helped me through out during my Ph.D study and research.

First and foremost I would like to express my deepest appreciation to my co-advisors Professor Zhaohai Li and Dr. Barry I. Graubard for their insightful guidance, support, and encouragement throughout the research in the past two years. Their direction, generous time and persistent help have made this dissertation possible.

I am greatly thankful to my committee members: Professor Michael D. Larsen, Professor Qing Pan, Professor Yinglei Lai, Professor Joseph L. Gastwirth, and Dr. Hormuzd Katki for their time and support. I want to express my sincere appreciation to all the professors whose courses I have taken to create a foundation for my current and future research.

I would like to thank my friends/advisors who ever helped me and encouraged me in change of my career path to the field of statistics. Also thanks to the Department of Statistics for providing me this opportunity.

Finally, I want to thank my parents, my wife Xia Ma and my son Hanchi Yao for their never-ending love and unconditional support. They are always there and cheer me up. I dedicate this thesis to my parents, my wife and my son.

Abstract of Dissertation

Estimation of ROC Curve with Complex Survey Data

Receiver Operating Characteristic (ROC) curve analysis has gained an increased interest in past decades. It has been widely used to evaluate the performance of diagnostic tests. The area under the ROC curve (denoted by AUC) is the most commonly used summary index of a ROC curve. A larger AUC value for a diagnostic test usually means that the test has better discriminating ability between diseased and non-diseased populations. Both parametric and nonparametric methods have been developed to estimate and compare AUCs. However, these methods are standardly used for simple random sample, not complex samples.

In surveys, complex sample designs with cluster sampling are commonly implemented. The Hispanic Health and Nutrition Examination Survey (HHANES) was conducted to assess the health and nutritional status the population of Hispanic individuals aged 6 months to 74 years in specific areas in U.S. by using a multistage, stratified, probability design with complex weight calculation. Analyses without accounting for weighting and clustering effect that is induced by the complex survey sampling can be biased. Thus, standard statistical methods of estimation of AUC for the population and its variance are not applicable to complex survey data.

In this dissertation, we propose an extension of the nonparametric method in the estimation of the population AUC that accounts for sample weighting under differing complex survey designs. We provide and study the accuracy of a jackknife method, along with balanced repeated replication (BRR), for variance estimation of our proposed estimator of AUC. We also discuss informative sample designs where the

selection probabilities are related to the parameter of interest, so that the standard analyses that ignore the sample weights can be seriously biased. Finally, our proposed methods are then applied to the Mexican-American portion of the HHANES to compare the classification accuracy of three predictors for overweight/obese using measured BMI as a gold standard.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Literature Review

1.1 Sensitivity and Specificity

Diagnosis of a disease in medical has often been made easier to report by a binary result, positive or negative, after compared to an assumed "gold standard". The performance of the diagnostic test is frequently assessed by its discriminative ability between alternative status of health (Zweig and Campbell, 1993).

Sensitivity and specificity are perhaps the most widely used parameters to characterize the diagnostic accuracy. Sensitivity is also called true positive rate (TPR), i.e., the probability that the test result is positive if the individual has disease. Specificity is defined by the probability the test result is negative if the individual does not have disease, which is often called true negative rate (TNR) or 1 - false positive rate (FPR).

Let T denote a binary test result, where $T+$ for a positive test result, $T-$ for a negative test result, and D be the actual disease status where D=1 if the disease is present and $D = 0$ otherwise. Suppose that c is the value of the threshold in a particular classification rule, and let Y be the continuous test results. Then the sensitivity and specificity for threshold c can be defined as followings:

$$\text{Sensitivity} = TPR(c) = P(T+|D=1) = P(Y \geq c|D=1)$$

$$\text{Specificity} = TNR(c) = P(T-|D=0) = P(Y < c|D=0)$$

In practice, the sensitivity and specificity are usually used together as joint measures of performance because, in general, decreasing the c threshold value will increase the sensitivity (TPR) and decrease the specificity (TNR). However the choice of an optimal c is generally unknown in advance and may depend on practical considerations.

1.2 Receiver Operating Characteristic (ROC) curve

The Receiver Operating Characteristic (ROC) curve allows jointly displaying both sensitivity and specificity, defined as a plot of sensitivity, or true positive rate (TPR) vs. 1 - specificity, or false positive rate (FPR) across all possible threshold values. Each point on the plot is generated by a different decision threshold c. The ROC curve is

$$(FPR(c), TPR(c)), c \in (-\infty, +\infty). \tag{1.1}$$

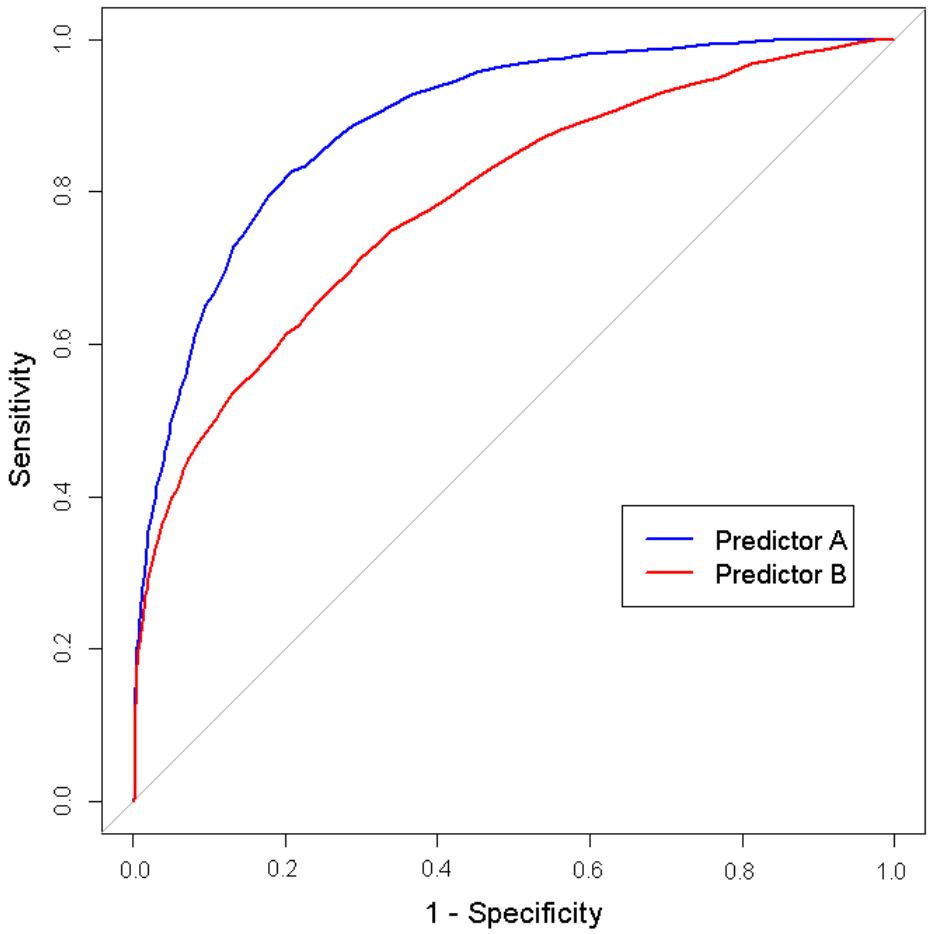

Figure 1.1: ROC curves

The ROC also can be rewritten as (Pepe, 2003)

$$(t, ROC(t), t \in (0, 1), \tag{1.2}$$

where the ROC(.) function maps t to TPR(c), and c is the threshold corresponding to FPR(c) = t.

Figure 1.1 shows an example of ROC curves for comparing the performance of predictors A and B for a diagnostic test. Some features can be observed from the Figure 1.1. First, sensitivity is nondecreasing as the specificity decreases. In

other words, the ROC curve is a nondecreasing function. Secondly, the higher the ROC curve is in the quadrant [0,1] $\times$ [0,1] indicates the better of its capability to discriminate between subjects with and without the disease of interest. Another feature is that the ROC curve provides a direct visual comparison of two or more diagnostic tests at all possible decision thresholds. Also the ROC curve serves as a useful tool for finding the optimal cut-off point to minimize misclassification of disease status.

1.3 Area Under the ROC Curve (AUC)

It is often convenient to reduce an ROC curve to a single quantitative measure. Among various summary indices of ROC curve, the area under the ROC curve (AUC) is the most commonly used approach. The area under the ROC curve is formally defined as

$$AUC = \int_0^1 ROC(t)dt \tag{1.3}$$

Observe that the AUC is bounded between 0.5 and 1. A larger AUC value usually presents a better discriminating ability of a diagnostic procedure. The closer AUC is to 1, in other words, the closer an ROC curve comes to the point (0,1), the better the overall diagnostic performance of the test. Conversely, when AUC is 0.5, i.e., an ROC curve that follows the diagonal line, the diagnostic test has the discriminative ability that is no better than tossing a coin.

Let SD and $S\bar{D}$ be the survival functions for Y in the diseased and non-diseased populations respectively: $S_D(y) = P(Y \geq y|D = 1)$ and $S_{\bar{D}}(y) = P(Y \geq y|D = 0)$. Let $c = S_{\bar{D}}^{-1}(t)$, i.e., c is the threshold corresponding to the false positive rate, $P(Y \geq c|D = 0) = t$, and the corresponding true positive rate is $P(Y \geq c|D = 1) = S_D$.

Then, the ROC curve can be expressed as (Pepe, 2003)

$$ROC(t) = S_D(c) = S_D(S_{\bar{D}}^{-1}(t)), t \in (0,1) \tag{1.4}$$

Let $Y_D$and $Y_{\bar{D}}$ be independent and randomly selected test results from diseased and non-diseased population respectively. Using the formula (1.4), we can have

$$AUC = P(Y_D > Y_{\bar{D}}). \tag{1.5}$$

Proof:

$$\begin{aligned}
AUC &= \int_0^1 ROC(t)dt \\
&= \int_0^1 S_D(S_{\bar{D}}^{-1}(t))dt \\
&= \int_\infty^{-\infty} S_D(y)dS_{\bar{D}}(y) \\
&= \int_{-\infty}^{\infty} P(Y_D > y)f_{\bar{D}}(y)dy \\
&= \int_{-\infty}^{\infty} P(Y_D > y, Y_{\bar{D}} = y)dy \\
&= P(Y_D > Y_{\bar{D}}).
\end{aligned}$$

where $y = S_{\bar{D}}^{-1}(t)$, and $f_{\bar{D}}$ is probability density function for $Y_{\bar{D}}$. The step from line 4 to line 5 is based on independence of Y_D and $Y_{\bar{D}}$.

The area under an ROC curve can be interpreted in different ways: (a)as the probability that a randomly chosen diseased subject is correctly ranked higher than a randomly chosen non-diseased subject (Bamber, 1975; Hanley and McNeil, 1982); (b) as the average value of specificity for all possible values of sensitivity, (Metz, 1978, 1998), and (c) as the average values of true positive rate over all possible values of false positive rate (Pepe, 2003).

1.4 Estimation of ROC Curve

Various statistical methods for estimating the ROC curves for continuous test results have been proposed. These methods can be summarized into three approaches: parametric, nonparametric, and semiparametric. In this dissertation, we are interested in the nonparametric method.

For large sample sizes and continuous test results, the nonparametric estimate of the ROC curve may be preferred since it passes through all observed points and provides unbiased estimates of sensitivity, specificity, and AUC (Zweig and Campbell, 1993). However, when the variables take a small number of discrete values, the area under the empirical ROC curve estimated by the trapezoidal rule tends to underestimate true area (Hanley and McNeil, 1982; Swet and Pickett, 1982).

Without making any distributional assumptions for the test values, the ROC curve can be fitted empirically. As shown in Figure 1.2, the empirical ROC curve is fitted by connecting the points (TPR(c), FPR(c)) for each c by a step function (Hanley and McNeil, 1982; Zweig and Campbell, 1993).

Let X_i denote the test values of i-th diseased subject $i = \{1, 2, \cdots, m\}$, Y_j denote the test values of j-th nondiseased subject $j = \{1, 2, \cdots, n\}$, and C_1 and C_2 denote the diseased and nondiseased groups in the population respectively. The sensitivity and specificity can be computed as

$$\mathsf{Sensitivity} = p(X_i \geq c \mid X_i \in C_1) = \frac{1}{m}\sum_{i}^{m} I(X_i \geq c) \tag{1.6}$$

$$\mathsf{Specificity} = p(Y_j < c \mid Y_i \in C_2) = \frac{1}{n}\sum_{i}^{n} I(Y_i < c) \tag{1.7}$$

Using the trapezoidal rule, Bamber (1975) was the first to show that the area

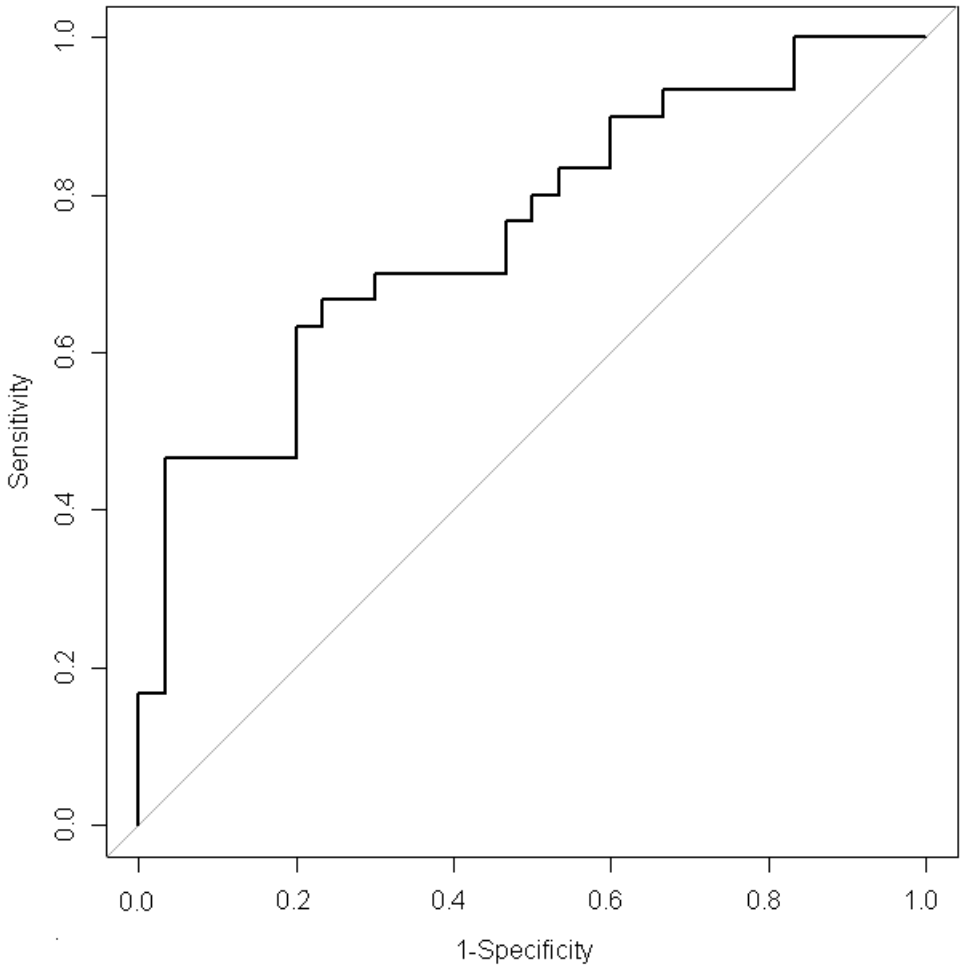

Figure 1.2: Empirical ROC curve

under the empirical ROC curve is equivalent to the quantity obtained when one performs the Mann-Whitney version of the Wilcoxon two-sample rank sum statistic. This can be expressed as the average over the kernel ψ as

$$\hat{\theta}_s = \frac{1}{mn} \sum_{j=1}^{n} \sum_{i=1}^{m} \psi(X_i, Y_j), \tag{1.8}$$

where

$$\psi(X,Y) = \begin{cases} 1 & Y < X \\ 1/2 & Y = X \\ 0 & Y > X \end{cases}$$

Thus the AUC is the sum of the proportion of pairs for which the test results from diseased individuals exceed the test results from normal individuals, and half the proportion of ties. It can be viewed as a measure based on pairwise comparison between two classes. Note that $\hat{\theta}$ is an unbiased estimate of true AUC, θ, where $E(\hat{\theta}_s) = \theta = E(\psi(X,Y)) = P(Y < X) + \frac{1}{2}P(X = Y) = P(Y < X)$, since $P(X = Y) = 0$ for continuous test results.

This nonparametric method for estimating the AUC has been used throughout the literature. The asymptotic expression of its variance has been proposed by various authors (Dorfman and Alf, 1969; Bamber 1975; Hanley and McNeil 1982, 1984; DeLong et al., 1988; Wieand et al., 1989; Cai and Pepe, 2002; Pepe 2004a). One way to express the variance was presented by Hanley and McNeil (1982) that depends on two distribution-specific quantities, θ_1 and θ_2 :

$$\mathbf{var}(\hat{\theta}_s) \quad = \quad \frac{\theta(1-\theta) + (m-1)(\theta_1 - \theta^2) + (n-1)(\theta_2 - \theta^2)}{mn} \tag{1.9}$$

where $\theta_1 = E[\psi(X_i, Y_j)\psi(X_s, Y_j)] - \theta^2, \quad i \neq s$ and
$\theta_2 = E[\psi(X_i, Y_j)\psi(X_i, Y_t)] - \theta^2, \quad j \neq t.$

Proof:

$$\begin{aligned}
\mathbf{var}(\hat{\theta}_s) &= \frac{1}{(mn)^2}\mathbf{var}\left[\sum_{j=1}^{n}\sum_{i=1}^{m}\psi(X_i,Y_j)\right] \\
&= \frac{1}{(mn)^2}\left(\sum_{j=1}^{n}\sum_{i=1}^{m}\mathbf{var}[\psi(X_i,Y_j)] + \sum_{j=1}^{n}\sum_{s\neq i,s=1}^{m-1}\sum_{i=1}^{m}\mathbf{cov}[\psi(X_i,Y_j),\psi(X_s,Y_j)]\right. \\
&\quad \left. + \sum_{t\neq j,t=1}^{n-1}\sum_{j=1}^{n}\sum_{i=1}^{m}\mathbf{cov}[\psi(X_i,Y_j),\psi(X_i,Y_t)]\right) \\
&= \frac{1}{(mn)^2}\left(mn\theta(1-\theta) + nm(m-1)(\theta_1-\theta^2) + mn(n-1)(\theta_2-\theta^2)\right) \\
&= \frac{\theta(1-\theta)+(m-1)(\theta_1-\theta^2)+(n-1)(\theta_2-\theta^2)}{mn}
\end{aligned}$$

DeLong *et al*. (1988) has shown an alternative methodology using structural components method (Sen 1960) which exploits the properties of the Mann-Whitney statistic. They defined X-components and Y-components as

$$V_{10}(X_i) = \frac{1}{n}\sum_{j=1}^{n}\psi(X_i,Y_j), \tag{1.10}$$

and

$$V_{01}(Y_j) = \frac{1}{m}\sum_{i=1}^{m}\psi(X_i,Y_j). \tag{1.11}$$

Let S_{10} and S_{01} be the estimated covariance estimates of $V_{10}(X)$ and $V_{01}(Y)$ components respectively.

$$S_{10} = \frac{1}{m-1}\sum_{i=1}^{m}(V_{10}(X_i)-\hat{\theta})^2, \tag{1.12}$$

and

$$S_{01} = \frac{1}{n-1}\sum_{j=1}^{n}(V_{01}(X_j)-\hat{\theta})^2. \tag{1.13}$$

Then the estimated variance of AUC is

$$S = \frac{1}{m} S_{10} + \frac{1}{n} S_{01}. \tag{1.14}$$

This approach turns out to be equivalent to the variance estimation by using jackknife method (DeLong *et al*. 1988).

1.5 Motivation

Both parametric and nonparametric methods have been developed to estimate and compare the AUCs. However, these methods are standardly used for simple random sampling, not complex sampling. To analyze sample survey data, one may need to take into account other aspects of sample design into the analysis, for example sample weights. The sample weights incorporate the differential sampling rates and may also incorporate adjustments for nonresponse and poststratification The sample weight of a sampled individual estimates the number of people in the population that the sampled individual represents. When the sample weights are correlated with the measurements of interest, in our case the test results, then ignoring the sample weights in an analysis can lead to substantial bias (Korn and Graubard, 1999).

Note that in Delong's method (1988), the test results are assumed to be independent. In the presence of clustered data, the estimation and inference of the accuracy of diagnostic tests should account for potential intracluster correlation among test results. Obuchowski(1997) proposed a method for estimating the area under the ROC curve by applying Delong et al. (1988)'s structural components approach to ROC curve estimation but extended their method to clustered ROC data using the ideas of Rao and Scott (1992). However, Obuchowski's method did not account for other

aspects of sample design, such as sampling weights and multistage cluster sampling.

Davis and Flegal (2003) provided a model-based approach using sample weighted pseudo-likelihood to test the difference between two ROC curves estimated from a complex survey sample data, the National Health and Nutrition Examination Survey III (NHANES III). However, the sampling weights were not properly incorporated in the estimate of AUC because they used weights based on single inclusion probabilities when they should have used weights based on pairwise joint inclusion probabilities among the pairs of observations.

Complex survey data such as the Hispanic Health and Nutrition Examination Survey (HHANES) has stratified multistage-cluster sample designs with sample weighting that can inflate variances and alter the expectations of test statistics (National Center for Health Statistics, 1985) and cluster sampling that can also inflate variances. Thus, the standard statistical methods of estimation of AUC and its variance estimate are not applicable to data from complex samples.

In this dissertation, we propose an extension of the nonparametric method with considering sample weights for estimation of the population AUC and its variance under different survey designs that range from simple random samples to multistage stratified cluster samples. We also provide and study the accuracy of a jackknife method, along with balanced repeated replication (BRR), for variance estimation of our proposed estimator of AUC.

Chapter 2

Estimation of AUC for Simple Random Sampling and Stratified Simple Random Sampling

In this chapter we consider simple random sample and stratified simple random sample designs; two popularly used sample designs. Let the target population consist of elements (e.g., individual subjects) and for these designs the units that are sampled, called sample units, are the individual elements. In contrast in Chapter 3, we consider sample units that can consist of multiple elements or clusters of elements. Simple random sample and stratified simple random sample designs of population elements are types of single-stage sample designs because the elements are directly sampled without any further sampling such as subsample. In Chapter 3 we will consider multi-stage sampling of sample units consisting of clusters of population elements.

2.1 Simple Random sampling

2.1.1 Introduction

Simple random sampling (SRS) without replacement is the simplest of the methods of probability sampling. Although the method is rarely used in sampling human populations, it is useful to start with it in the presentation of the sampling theory of surveys as its understanding is helpful in studying other and more complex sample designs.

The simple random sampling is sometime also called random sampling method where an equal probability of inclusion in the sample is obtained for each subject that sampled from the population. A simple random sample is drawn subject by subject from a finite population of subjects. If the number of subjects in the finite population is T, the probability of selecting any subject in the population at the first draw is $1/T$. The probability of selecting any unit from among the available subjects at second draw is $1/(T-1)$ for without replacement sampling, or $1/T$ if using for with replacement sampling.

Sampling with replacement is entirely possible but seldom used in practice. As a rule, sampling with replacement is usually less precise than the sampling without replacement because of the possibility of sampling the same person more than one time in with replacement sampling resulting in less information in sample for a given sample size than without replacement sampling(Cochran, 1959). When the sampling fraction (t/T) is small, so that the probability that same subject is selected twice is low, sampling with replacement is nearly equivalent to sampling without replacement. In this dissertation, all sampling designs will refer to sampling without replacement unless otherwise specified.

2.1.2 The Inclusion Probabilities

The inclusion of a given subject i in a sample s is a random event indicated by the random variable I_i defined as

$$I_i = \begin{cases} 1 & \text{if } i \in s \\ 0 & \text{if not} \end{cases} \tag{2.1}$$

Suppose sample s is drawn by SRS from a finite population U, and let the size of sample s be t, and the size of population be T. Because of the equal inclusion probability of each subject sampled from the population for SRS, the probability that subject i will be included in a sample s, denoted π_i , is

$$\pi_i = P(i \in s) = P(I_i = 1) = \sum_{s \ni i} p(s) = \begin{pmatrix} T-1 \\ t-1 \end{pmatrix} / \begin{pmatrix} T \\ t \end{pmatrix} = \frac{t}{T} \tag{2.2}$$

Where $s \ni i$ denotes that the sum is over those samples s that contain i. The joint inclusion probability that two subjects i and j will be included in sample s, is denoted as π_{ij}:

$$\pi_{ij} = P(i\&j \in s) = P(I_i I_j = 1) = \sum_{s \ni i\&j} p(s) = \begin{pmatrix} T-2 \\ t-2 \end{pmatrix} / \begin{pmatrix} T \\ t \end{pmatrix} = \frac{t(t-1)}{T(T-1)} \tag{2.3}$$

For SRS, we have $\pi_{ij} = \pi_{ji}$ for all i, j, and when $i = j$,

$$\pi_{ii} = P(I_i^2 = 1) = P(I_i = 1) = \pi_i \tag{2.4}$$

Let the sample weight (w) be the inverse of inclusion probability, $w_i = 1/\pi_i$, and a joint sample weight is defined as the inverse of joint inclusion p robability, $w_{ij} = 1/\pi_{ij}$. Note that

$$E(I_i) = \pi_i = 1/w_i$$

and

$$E(I_{ij}) = \pi_{ij} = 1/w_{ij}$$

.

2.1.3 Estimation of AUC for SRS

Let U be the target finite population, containing M diseased subjects labeled as $\{X_1, X_2, \cdots, X_M\}$, and N nondiseased subjects labeled as $\{Y_1, Y_2, \cdots, Y_N\}$, so the size of the population is $T = M + N$, and let C_1 and C_2 denote diseased and nondiseased subpopulations respectively. Suppose sample s is drawn from the finite population U by SRS, $s = \{x_1, \cdots, x_m, y_1, \cdots, y_n\}$, containing the size of m diseased subpopulation denoted as c_1 and the size of n nondiseased subpopulation denoted as c_2, so that the size of $s = c_1 \bigcup c_2$ is $t = m + n$.

Our interest is to estimate the population AUC for the finite population U given by incorporating the sample weights the estimator of population AUC is as following:

$$\hat{\theta}_U = \frac{1}{\hat{M}\hat{N}} \sum_{j=1}^{T} \sum_{i=1}^{T} w_{ij} I_{ij} \delta_i (1 - \delta_j) \psi(X_i, Y_j) \tag{2.5}$$

where $\hat{M}$ and $\hat{N}$ are the estimators of M ($E[\hat{M}] = M$) and $N(E[\hat{N} = N])$,

w_{ij} are the sample weights, equal to $\frac{T(T-1)}{t(t-1)}$, I_{ij} is the indicator of including subject i and j in the sample, δ is the indicator that it is equal to 1 if the subject

belongs to diseased subpopulation C_1, otherwise 0.

In practice, the population size T is known before sampling from census data. If we know the disease prevalence (denoted as d, $d = M/T$) for the finite population, then the number of diseased subject, $M = Td$ and non-diseased subject, $N = T - M$. However, d, M and N are usually unknown. After sampling and collecting the data we are able to check the disease status for each sampled individuals. The disease prevalence (d) can be estimated from the sample, as $\hat{d} = m/t$, so that M and N can be estimated as

$$\hat{M} = \hat{d}T = \frac{m}{t}T \tag{2.6}$$

$$\hat{N} = T - \hat{M} = \frac{n}{t}T \tag{2.7}$$

In the formula 2.5, both numerator and denominator are random variables, therefore the estimator of AUC is a ratio estimator. While both numerator and denominator are unbiased estimators of their corresponding components, the estimator of AUC is a biased estimator but asymptotically consistent to the population AUC. The following shows the numerator is unbiased. Let $\mathbf{A}$) be the expected numerator in formula 2.5.

$$\begin{aligned}
\mathbf{A} &= \mathbf{E}\left[\sum_{j=1}^{T}\sum_{i=1}^{T} w_{ij} I_{ij} \delta_i (1-\delta_j)\psi(X_i, Y_j)\right] \\
&= \sum_{j=1}^{T}\sum_{i=1}^{T} \psi(X_i, Y_j)\mathbf{E}[w_{ij}]\mathbf{E}[I_{ij}]\mathbf{E}[\delta_i(1-\delta_j)] \\
&= \sum_{j=1}^{T}\sum_{i=1}^{T} \theta \frac{T(T-1)}{t(t-1)} \frac{t(t-1)}{T(T-1)} \frac{M}{T} \frac{N}{T} \\
&= MN\theta
\end{aligned}$$

Please note that, the estimator of sample AUC ($\hat{\theta}_s$) used in standard case in

Chapter 1 is an unbiased estimator. When randomly choosing any two subjects from the sample s, total $t(t-1)$ pairs combination can be obtained, but only mn pairs are valid pairwise comparisons between diseased and nondiseased subjects which is taken into account with the using the indicator δ in formula (2.5).

2.1.4 Variance Estimation

We use a jackknife method of variance estimation throughout this dissertation. The jackknife method is a broadly used subsample replication technique for variance estimation. The basic idea of jackknife method is to calculate the "estimates of the parameter of interest from each of several subsample of the parent sample and then estimate the variance of the parent sample estimator from the variability between the subsample estimates(Wolter, 2003).

We suppose that the sample can be divided into k random groups and all groups are mutually independent. Let $\hat{\theta}_U$ be the estimator of population AUC, θ, and $\hat{\theta}_{(-i)}$ be the estimator of the same functional form as $\hat{\theta}_U$, but computed from the reduced sample by omitting the i-th group in the sample. When calculating $\hat{\theta}_{(-i)}$, the sample weights of the remaining data are multiplied by a factor of $k/(k-1)$ so that it will be reasonable estimator of θ.

The jackknife variance estimator of $\hat{\theta}_U$ is defined by (Korn and Graubard, 1999).

$$\begin{aligned} \widehat{V}_{JK}(\hat{\theta}_U) &= \frac{k-1}{k}\sum_{i=1}^{k}(\hat{\theta}_{(-i)} - \hat{\theta}_U)^2 \\ &= \frac{1}{k(k-1)}\sum_{i=1}^{k}(\hat{\theta}_i - \hat{\theta}_U)^2 \end{aligned} \tag{2.8}$$

where $\hat{\theta}_i$ are the pseudovalues, defined as $\hat{\theta}_i = k\hat{\theta} - (k-1)\hat{\theta}_{(-i)}$.

Wolter (1985) introduced an alternative jackknife variance estimator that one

can replace $\hat{\theta}_U$ in (2.8) with the mean of the $\hat{\theta}_{(-i)}$, denoted as $\hat{\bar{\theta}}$.

$$\widehat{V}_{JK2}(\hat{\theta}_U) \quad = \quad \frac{k-1}{k}\sum_{i=1}^{k}(\hat{\theta}_{(-i)} - \hat{\bar{\theta}})^2 \tag{2.9}$$

The latter form (2.9) is considered a less conservative estimator since

$$\widehat{V}_{JK}(\hat{\theta}_U) = \widehat{V}_{JK2}(\hat{\theta}_U) + (\hat{\theta}_U - \hat{\bar{\theta}})^2/(k-1) \tag{2.10}$$

We prefer the conservative form (2.8) in this dissertation.

For simple random sampling, one may treat each independent individual as a single group, so that the number of independent groups is equal to the sample size, $k = t$. Then the variance estimator of $\hat{\theta}_U$ is

$$\widehat{V}_{JK}(\hat{\theta}_U) = \frac{t-1}{t}\sum_{i=1}^{t}(\hat{\theta}_{(i)} - \hat{\theta}_U)^2 \tag{2.11}$$

In the nonsurvey setting, the sample is considered as two strata, diseased and nondiseased. When pseudovalues are computed for diseased stratum and non-diseased stratum separately, then the jackknife variance estimator is equivalent to the Delong's variance estimator by using the structural components method (DeLong, 1988).

2.1.5 Simulation

Monte Carlo simulation studies are conducted to assess the the performance of the proposed weighted estimator for the population AUC for SRS and its variance estimators (weighted estimators), as compared to the conventional approach (unweighted estimators).

We generate a fixed size of finite population $T = 200{,}000$. The disease prevalence

in the finite population (d) is controlled at 5% and 30%, respectively. The population AUC (θ) is set at 0.95, 0.9, 0.8, 0.7, 0.6, or 0.55 respectively. Both diseased and nondiseased subjects' test results are randomly generated as normally distributed with mean of 5 for diseased subjects, mean of nondiseased subjects determined by the value of population AUC, and variance equal to 1. Different sample sizes t= 250, 500, or 1000 are randomly drawn from the finite population without replacement resulting in different inclusion probabilities. For each combination of disease prevalence, sample size, and AUC, 2000 simulations were conducted where the finite population was regenerated for each simulation.

When reporting the simulation results, we use these labels throughout this dissertation. The label "1" and "2" are used to indicate unweighted estimators and weighted estimators respectively. "RelBias" stands for the relative bias, ie., RelBias1=$\theta/\hat{\theta}1$, RelBias2=$\theta/\hat{\theta}2$. "RelSE(JK)" is for relative jackknife(JK) standard error(SE), ie., RelSE(JK1)=SE(JK1)/$\hat{\theta}1$, RelSE(JK2)=SE(JK2)/$\hat{\theta}2$. "Bias(SE)" is the absolute bias of SE, ie., Bias(SE1)=|SE(JK1)-SE(EMP1)|, Bias(SE2)=|SE(JK2)-SE(EMP2)|, where SE(EMP) denotes for SE for empirical variance estimate. "RMSE" is for the square root of mean square error(MSE), RMSE1=$\sqrt{MSE1}$, RMSE2=$\sqrt{MSE2}$.

Table 2.1: Simulation results for SRS, $d = 0.05$

θ	t	RelBias1*	RelBias2*	SE(DL)	SE(JK1)	RelSE(JK1)	SE(JK2)	RelSE(JK2)	SE(EMP1)	Bias(SE1)	SE(EMP2)	Bias(SE2)	RMSE1	RMSE2
0.95	1000	1.0005	0.9995	0.0145	0.0147	0.0155	0.0147	0.0155	0.0144	0.0003	0.0144	0.0003	0.0144	0.0144
	500	0.9997	0.9977	0.0208	0.0212	0.0223	0.0213	0.0223	0.0210	0.0002	0.0210	0.0002	0.0210	0.0211
	250	0.9997	0.9960	0.0298	0.0313	0.0329	0.0314	0.0329	0.0291	0.0022	0.0292	0.0022	0.0291	0.0295
0.9	1000	0.9999	0.9989	0.0219	0.0221	0.0246	0.0221	0.0246	0.0216	0.0005	0.0216	0.0005	0.0216	0.0217
	500	0.9997	0.9977	0.0314	0.0321	0.0356	0.0321	0.0356	0.0311	0.0010	0.0311	0.0010	0.0311	0.0312
	250	1.0013	0.9973	0.0460	0.0483	0.0538	0.0485	0.0538	0.0465	0.0019	0.0466	0.0019	0.0465	0.0467
0.8	1000	0.9996	0.9988	0.0316	0.0319	0.0399	0.0320	0.0399	0.0306	0.0013	0.0307	0.0013	0.0306	0.0307
	500	1.0018	0.9986	0.0460	0.0469	0.0588	0.0470	0.0587	0.0449	0.0021	0.0449	0.0021	0.0449	0.0450
	250	0.9964	0.9924	0.0654	0.0685	0.0854	0.0688	0.0854	0.0656	0.0030	0.0658	0.0030	0.0656	0.0661
0.7	1000	1.0016	0.9996	0.0379	0.0383	0.0548	0.0383	0.0547	0.0384	0.0002	0.0385	0.0002	0.0384	0.0385
	500	0.9979	0.9969	0.0546	0.0558	0.0795	0.0559	0.0796	0.0535	0.0022	0.0536	0.0022	0.0536	0.0537
	250	0.9990	0.9950	0.0777	0.0815	0.1163	0.0818	0.1163	0.0764	0.0051	0.0767	0.0051	0.0764	0.0767
0.6	1000	0.9978	0.9968	0.0414	0.0417	0.0694	0.0418	0.0694	0.0408	0.0010	0.0408	0.0010	0.0408	0.0408
	500	0.9964	0.9944	0.0583	0.0595	0.0988	0.0596	0.0988	0.0561	0.0034	0.0562	0.0034	0.0561	0.0563
	250	0.9801	0.9762	0.0819	0.0858	0.1401	0.0861	0.1401	0.0725	0.0133	0.0727	0.0134	0.0735	0.0742
0.55	1000	0.9913	0.9903	0.0414	0.0418	0.0754	0.0419	0.0754	0.0359	0.0059	0.0359	0.0059	0.0362	0.0364
	500	0.9793	0.9774	0.0588	0.0600	0.1068	0.0601	0.1068	0.0447	0.0153	0.0448	0.0153	0.0462	0.0466
	250	0.9522	0.9484	0.0829	0.0868	0.1503	0.0872	0.1503	0.0580	0.0288	0.0582	0.0289	0.0642	0.0655

Note: * 1 = unweighted estimator, 2 = weighted estimator. RelBias1=$\theta/\hat{\theta}1$, RelBias2=$\theta/\hat{\theta}2$, RelSE(JK1)=SE(JK1)/$\hat{\theta}1$, RelSE(JK2)=SE(JK2)/$\hat{\theta}2$, Bias(SE1)=|SE(JK1)-SE(EMP1)|, Bias(SE2)=|SE(JK2)-SE(EMP2)|.

Table 2.1 summarize the simulation results for disease prevalence at 0.05. It shows the unweighted relative bias (RelBias1) is less bias (closer to 1) than weighted relative bias (RelBias2). When comparing jackknife standard errors, as expected that the weighted jackknife standard error are slightly larger than unweighted jackknife standard error because sample weight usually inflates the variance estimation, and similarly for comparing the empirical standard errors and square root of MSEs, the weighted ones are slightly larger than the unweighted ones. With the increasing sample size, the differences between weighted and unweighted of relative biases, jackknife standard errors, and empirical standard errors bias are decreasing, and they are identical at 4th decimal place when sample size reaches to $t = 1000$. However when comparing the relative jackknife standard errors, there are little difference between weighted and unweighted standard errors, and similarly for the bias of standard errors. These results indicates that the weighted estimator performs almost as good as the unweighted estiamtor in a relative scale.

We also compare the unweighted standard errors by using Delong's method (SE(DL)) with the unweighted jackknife standard errors (SE(JK1)). As expected, Delong's standard errors are consistently smaller than jackknife standard errors because as discussed in previous section, Delong's method employed the structural components approach and treated diseased and nondiseased populations as two strata, effectively conditioning on the disease and nondisease sample sizes, which treats these sample sizes as fixed, when in fact they are random.

When disease prevalence increases from $d = 0.05$ to 0.3, there is an increase in the number of decreased and nondiseased subject pairs. Thus with larger numbers of pairs the variance estimates in Table 2.2 tend to smaller. In general the pattern of the results within each of Tables 2.1 and 2.2 are similar.

Table 2.2: Simulation results for SRS, $d = 0.3$

θ	t	RelBias1*	RelBias2*	SE(DL)	SE(JK1)	RelSE(JK1)	SE(JK2)	RelSE(JK2)	SE(EMP1)	Bias(SE1)	SE(EMP2)	Bias(SE2)	RMSE1	RMSE2
0.95	1000	1.0000	0.9990	0.0068	0.0068	0.0072	0.0068	0.0072	0.0069	0.0001	0.0069	0.0001	0.0069	0.0070
	500	1.0000	0.9980	0.0097	0.0097	0.0102	0.0097	0.0102	0.0098	0.0001	0.0098	0.0001	0.0098	0.0100
	250	0.9996	0.9956	0.0138	0.0137	0.0145	0.0139	0.0145	0.0140	0.0003	0.0141	0.0002	0.0140	0.0147
0.9	1000	1.0000	0.9990	0.0104	0.0104	0.0115	0.0104	0.0115	0.0100	0.0004	0.0100	0.0004	0.0100	0.0101
	500	1.0000	0.9980	0.0147	0.0147	0.0163	0.0147	0.0164	0.0150	0.0003	0.0150	0.0003	0.0150	0.0152
	250	0.9998	0.9958	0.0209	0.0209	0.0232	0.0210	0.0232	0.0209	0.0000	0.0209	0.0001	0.0209	0.0213
0.8	1000	1.0004	0.9994	0.0150	0.0150	0.0188	0.0150	0.0188	0.0149	0.0001	0.0149	0.0001	0.0149	0.0149
	500	0.9995	0.9975	0.0213	0.0213	0.0266	0.0214	0.0266	0.0208	0.0004	0.0209	0.0005	0.0208	0.0210
	250	1.0008	0.9968	0.0304	0.0303	0.0379	0.0305	0.0380	0.0298	0.0005	0.0299	0.0006	0.0298	0.0300
0.7	1000	0.9999	0.9989	0.0179	0.0179	0.0255	0.0179	0.0256	0.0179	0.0001	0.0180	0.0000	0.0179	0.0180
	500	0.9992	0.9972	0.0253	0.0253	0.0361	0.0254	0.0361	0.0252	0.0001	0.0252	0.0001	0.0252	0.0253
	250	1.0016	0.9976	0.0361	0.0360	0.0515	0.0363	0.0517	0.0368	0.0009	0.0370	0.0007	0.0369	0.0370
0.6	1000	1.0005	0.9995	0.0195	0.0195	0.0325	0.0195	0.0325	0.0193	0.0001	0.0193	0.0002	0.0193	0.0193
	500	0.9989	0.9969	0.0276	0.0275	0.0458	0.0276	0.0459	0.0278	0.0003	0.0278	0.0002	0.0278	0.0279
	250	1.0016	0.9976	0.0393	0.0391	0.0653	0.0394	0.0656	0.0390	0.0002	0.0391	0.0003	0.0390	0.0391
0.55	1000	0.9986	0.9976	0.0198	0.0198	0.0359	0.0198	0.0360	0.0194	0.0004	0.0194	0.0004	0.0194	0.0195
	500	0.9981	0.9961	0.0281	0.0281	0.0509	0.0282	0.0510	0.0268	0.0013	0.0269	0.0013	0.0268	0.0269
	250	0.9909	0.9869	0.0397	0.0396	0.0713	0.0399	0.0716	0.0347	0.0049	0.0348	0.0051	0.0351	0.0356

Note: * 1 = unweighted estimator, 2 = weighted estimator. RelBias1=$\theta/\hat{\theta}1$, RelBias2=$\theta/\hat{\theta}2$, RelSE(JK1)=SE(JK1)$/\hat{\theta}1$, RelSE(JK2)=SE(JK2)$/\hat{\theta}2$, Bias(SE1)=|SE(JK1)-SE(EMP1)|, Bias(SE2)=|SE(JK2)-SE(EMP2)|.

2.2 Stratified Simple Random Sampling

2.2.1 Introduction

In practice, populations can be very heterogeneous with respect disease prevalence or accuracy of the test because of behaviors such smoking or drug that can chemically or biologically interfere with the test results. Also estimation may be required for certain subgroups of the population, for example racial groups of age groups. For both considerations the simple random sample design may be less efficient than stratified sampling.

Suppose it is possible to divide the population into subgroups based on some certain characteristics such as gender, race, income, or geographical locations. If these subgroups are disjoint and comprise the whole of population, they can be treated as sampling strata. When strata have been determined, independent random sample can be drawn from each stratum with specified sample sizes that can produce more efficient estimates than a simple random sample of the entire population. This sample design called stratified simple random sampling (SSRS). Stratified sampling is a very common procedure in sample surveys. It offers several advantages over simple random sampling.

1. In survey, practical aspects related to nonresponse rates, measurement problems and auxiliary information may greatly differ across different strata. To increase the efficiency of the estimation, one may take these factors into consideration when choosing the sample sizes from each stratum.

2. For administrative reasons, the survey organization may have been divided by geographic regions. It is convenient to make each region as a stratum.

3. As indicated above, stratification can help to achieve higher precision in the estimation of characteristics of the population.

4. Because it can provide higher precision, a stratified sample often requires a smaller sample, which saves money.

There are also some disadvantages:

1. The main disadvantage is that it may require more administrative effort than a simple random sample.

2. The stratified sampling technique may be misapplied if the population cannot be exhaustively partitioned properly into disjoint subgroups.

2.2.2 Estimation of AUC for SSRS

Suppose sample s is randomly selected by a SSRS design. Let $U_1, \cdots, U_H$ denote the population within the H strata, where $\bigcup_{h=1}^{H} U_h = U$. Each stratum population U_h has two subpopulations, diseased and nondiseased, $U_h = C_{h1} \bigcup C_{h2}$, containing N_h and M_h diseased and nondiseased subjects respectively. The size of U_h is denoted by $T_h = N_h + M_h$, and $\sum_{h=1}^{H} T_h = T$. Suppose a sample (s_h) size of t_h is randomly selected from U_h. The total sample is $s = \bigcup_{h=1}^{H} s_h$, $s_h = c_{h1} \bigcup c_{h2}$, and the size of s is $t = \sum_{h=1}^{H} t_h$, where $t_h = n_h + m_h$.

The subjects distribution of the test results may differ across strata. The comparison of diseased and non-diseased subjects can be made either within the same stratum or across the strata. Let $\theta_{h,h}$ be the AUC for the diseased subjects in h stratum and the non-diseased subjects in h stratum, and $\hat{\theta}_{h,h}$ be the estimator of $\theta_{h,h}$.

$$\hat{\theta}_{hh'} = \frac{1}{\hat{N}_h \hat{M}_{h'}} \sum_{j=1}^{T_{h'}} \sum_{i=1}^{T_h} w_{(hi,h'j)} I_{(hi,h'j)} \delta_{hi} (1 - \delta_{h'j}) \psi(X_{hi}, Y_{h'j}) \tag{2.12}$$

Then the AUC can be described as a matrix

$$\theta_{(H \times H)} = \begin{pmatrix} \theta_{1,1} & \theta_{1,2} & ... & \theta_{1,H} \\ \theta_{2,1} & \theta_{2,2} & ... & \theta_{2,H} \\ ... & ... & ... & ... \\ \theta_{H,1} & \theta_{H,2} & ... & \theta_{H,H} \end{pmatrix} \tag{2.13}$$

The finite population parameter for the AUC that is written using the stratification structure of the finite population is,

$$\theta_U = \frac{1}{NM} \sum_{h=1}^{H} \sum_{h'=1}^{H} \sum_{j=1}^{T_{h'}} \sum_{i=1}^{T_h} \delta_{hi} (1 - \delta_{h'j}) \psi(X_{hi}, Y_{h'j}) \tag{2.14}$$

And the estimator of AUC for the finite population for SSRS is

$$\hat{\theta}_U = \frac{1}{\hat{N}\hat{M}} \sum_{h=1}^{H} \sum_{h'=1}^{H} \sum_{j=1}^{T_{h'}} \sum_{i=1}^{T_h} w_{(hi,h'j)} I_{(hi,h'j)} \delta_{hi} (1 - \delta_{h'j}) \psi(X_{hi}, Y_{h'j}) \tag{2.15}$$

where $w_{(hi,h'j)}$ is $\frac{T_h(T_h-1)}{t_h(t_h-1)}$ if $h = h'$, or $\frac{T_h}{t_h}\frac{T_{h'}}{t_{h'}}$, if $h \neq h'$.

The above formula can be rewritten as (Kondratovich, 2000)

$$\hat{\theta}_U = \sum_{h=1}^{H} \sum_{h'=1}^{H} \hat{p}_h \hat{\theta}_{h,h'} \hat{q}_{h'} = \hat{p}^T \hat{\theta}_{(H \times H)} \hat{q} \tag{2.16}$$

where $\hat{p}$ and $\hat{q}$ are (H $\times$ 1) vectors, $\hat{p}^T = (\hat{p}_1, \cdots, \hat{p}_H)$ and $\hat{q}^T = (\hat{q}_1, \cdots, \hat{q}_H)$, $\hat{p}_h = \hat{M}_h/\hat{M}, \hat{q}_{h'} = \hat{N}_{h'}/\hat{N}$.

Again, the estimator of AUC for the finite population is a ratio estimator, the denominator and numerator are unbiased to each corresponding component. The

following is to show the unbiasedness of the numerator in formula 2.15 (a similar approach for the proof of the unbiasedness of the numerator in formula 2.16). Let **A** be the expected numerator in formula 2.16.

Proof:

$$\begin{aligned}
\mathbf{A} &= \mathbf{E}\left[\sum_{h=1}^{H}\sum_{h'=1}^{H}\sum_{j=1}^{T_{h'}}\sum_{i=1}^{T_h} w_{(hi,h'j)} I_{(hi,h'j)} \delta_{hi}(1-\delta_{h'j})\psi(X_{hi}, Y_{h'j})\right] \\
&= \sum_{h=1}^{H}\sum_{h'=1}^{H}\sum_{j=1}^{T_{h'}}\sum_{i=1}^{T_h} \mathbf{E}[w_{(hi,h'j)}]\mathbf{E}[I_{(hi,h'j)}]\mathbf{E}[\delta_{hi}(1-\delta_{h'j})]\mathbf{E}[\psi(X_{hi}, Y_{h'j})] \\
&= \sum_{h=1}^{H}\sum_{h'=1}^{H}\sum_{j=1}^{T_{h'}}\sum_{i=1}^{T_h} \frac{M_{h'}}{T_{h'}}\frac{N_h}{T_h}\theta \\
&= \theta[(N_1M_1 + N_1M_2 + \cdots + N_1M_h) + \cdots + (N_hM_1 + N_hM_2 + \cdots + N_hM_h)] \\
&= NM\theta
\end{aligned}$$

The above approach (formula 2.15) allows pairwise comparison of diseased and nondiseased subjects across different strata. Another possible approach is to restrict the comparison within each stratum, and then simply combine all counts from each stratum and dividing total number of pairs. The strata AUCs combined estimator for finite population AUC can be defined as

$$\hat{\theta}_U = \frac{1}{\sum_{h=1}^{H}\hat{N}_h\hat{M}_h}\sum_{h=1}^{H}\sum_{j=1}^{T_h}\sum_{i=1}^{T_h} w_{(hi,hj)} I_{(hi,hj)} \delta_{hi}(1-\delta_{hj})\psi(X_{hi}, Y_{hj}) \tag{2.17}$$

Clearly, $\sum_{h=1}^{H}\hat{N}_h\hat{M}_h \neq \hat{N}\hat{M}$, and $\sum_{h=1}^{H}\hat{N}_h\hat{M}_h$ is only a subset of $\hat{N}\hat{M}$. So this approach is relative simple but ignore the difference of AUCs across strata so loss of some information. In this dissertation we use the general version formula 2.15.

2.2.3 Variance Estimation

The jackknife variance estimator for simple random sampling has a stratified sampling version (Wolter 2003). Random groups of approximately equal sizes are formed independently within in each stratum. After removing i-th sampled independent group from h-th stratum, the sample weights of the remaining data in the h-th stratum are multiplied by a factor of $k_h/(k_h - 1)$ without changing the sample weights of the observations in the other strata. Again we can consider each sampled subject as a single independent group in each stratum. Then the jackknife variance estimator of population AUC can be defined as

$$\widehat{V}_{JK}(\hat{\theta}_U) = \sum_{h=1}^{H}[\frac{t_h - 1}{t_h}\sum_{i=1}^{t_h}(\hat{\theta}_{(-hi)} - \hat{\theta}_U)^2] \tag{2.18}$$

where the $\hat{\theta}_{(-hi)}$ is the estimators of the same functional form as $\hat{\theta}_U$, but computed from the reduced sample by omitting the i-th subjects in h stratum.

2.2.4 Simulation

A finite population of size of $T = 200000$ is generated with $H = 8$ strata. The size of each stratum is set to equal. In the context of stratified sampling, we assume that the stratum AUCs (θ_h) vary across different strata. For example, when we fix the population AUC at $\theta = 0.95$, the stratum AUCs (θ_h) are randomly assigned between 0.975 to 0.925, i.e., $\theta_h = \theta \pm 0.025$. Similar settings are made when we fix population AUC at other values, $\theta = 0.9$, 0.8, 0.7, 0.6, or 0.55, respectively. The population prevalence is fixed at 0.05. Similarly to Section 2.1.5, both diseased and nondiseased subjects' test results are randomly generated from normal distributions for each stratum separately, with variances 1. The mean of diseased subjects is set at 5 for each stratum, and the mean of nondiseased subjects for each stratum is

determined by the value of stratum specific AUCs.

The sampling in different strata are performed independently. To make the sample sizes of this simulation comparable to the sample size for the simple random sampling simulation in 2.1.5, we use sample size of $t_h = 30$, 60, or 120 from each stratum, so that the total sample size is $t = 240$, 480, or 960, respectively.

As discussed in the previous section, stratification can help to achieve more efficient estimation. While the variability of population AUC has been increased, the standard errors in Table 2.3 from stratified simple random sampling are not enlarged as compared to the corresponding standard errors in Table 2.1 from simple random sampling.

Similar to the previous findings in Table 2.1 and Table 2.2, Table 2.3 shows the weighted standard errors (SE(JK2)) are consistently larger than unweighted standard errors (SE(JK1)) since sample weights usually inflate the variance estimation. With increased sample size, the difference between unweighted and weighted standard errors becomes smaller, and the jackknife for variance estimates converge to the empirical variances with large sample sizes. When comparing the standard errors in relative scale - relative standard errors and bias of standard errors, there is little difference between weighted ones and unweighted ones.

Table 2.3: Simulation results for SSRS, $d = 0.05$

θ	t	RelBias1*	RelBias2*	SE(DL)	SE(JK1)	RelSE(JK1)	SE(JK2)	RelSE(JK2)	SE(EMP1)	Bias(SE1)	SE(EMP2)	Bias(SE2)	RMSE1	RMSE2
0.95	960	1.0016	1.0035	0.0136	0.0149	0.0157	0.0149	0.0157	0.0146	0.0003	0.0146	0.0003	0.0146	0.0150
	480	1.0041	1.0075	0.0174	0.0209	0.0221	0.0209	0.0222	0.0202	0.0007	0.0202	0.0007	0.0206	0.0214
	240	1.0075	1.0138	0.0214	0.0314	0.0333	0.0315	0.0336	0.0296	0.0018	0.0297	0.0018	0.0304	0.0324
0.9	960	0.9989	0.9979	0.0207	0.0226	0.0251	0.0226	0.0251	0.0223	0.0002	0.0224	0.0002	0.0224	0.0224
	480	1.0043	1.0063	0.0266	0.0321	0.0358	0.0322	0.0360	0.0326	0.0004	0.0326	0.0004	0.0328	0.0331
	240	1.0070	1.0108	0.0325	0.0486	0.0543	0.0488	0.0548	0.0456	0.0030	0.0458	0.0030	0.0460	0.0468
0.8	960	1.0020	1.0018	0.0300	0.0327	0.0410	0.0328	0.0410	0.0319	0.0009	0.0319	0.0009	0.0319	0.0319
	480	1.0049	1.0082	0.0392	0.0475	0.0597	0.0476	0.0600	0.0469	0.0006	0.0470	0.0007	0.0470	0.0474
	240	0.9910	0.9877	0.0478	0.0703	0.0871	0.0706	0.0872	0.0678	0.0025	0.0681	0.0025	0.0682	0.0688
0.7	960	1.0005	0.9989	0.0356	0.0388	0.0554	0.0388	0.0554	0.0375	0.0013	0.0375	0.0013	0.0375	0.0375
	480	1.0038	0.9958	0.0465	0.0562	0.0806	0.0563	0.0801	0.0543	0.0019	0.0544	0.0019	0.0544	0.0545
	240	0.9922	0.9881	0.0560	0.0831	0.1178	0.0835	0.1179	0.0769	0.0062	0.0773	0.0062	0.0771	0.0777
0.6	960	0.9959	0.9948	0.0387	0.0421	0.0699	0.0422	0.0699	0.0412	0.0009	0.0413	0.0009	0.0413	0.0414
	480	0.9928	0.9924	0.0499	0.0602	0.0997	0.0604	0.0998	0.0555	0.0047	0.0556	0.0047	0.0557	0.0558
	240	0.9841	0.9799	0.0600	0.0885	0.1452	0.0889	0.1452	0.0710	0.0175	0.0713	0.0176	0.0717	0.0724
0.55	960	0.9949	0.9939	0.0391	0.0427	0.0772	0.0427	0.0772	0.0358	0.0069	0.0358	0.0069	0.0359	0.0360
	480	0.9902	0.9843	0.0502	0.0605	0.1090	0.0606	0.1085	0.0456	0.0150	0.0457	0.0150	0.0459	0.0465
	240	0.9805	0.9749	0.0593	0.0883	0.1574	0.0887	0.1572	0.0589	0.0294	0.0592	0.0295	0.0599	0.0609

Note: * 1 = unweighted estimator, 2 = weighted estimator. RelBias1=$\theta/\hat{\theta}1$, RelBias2=$\theta/\hat{\theta}2$, RelSE(JK1)=SE(JK1)/$\hat{\theta}1$, RelSE(JK2)=SE(JK2)/$\hat{\theta}2$, Bias(SE1)=|SE(JK1)-SE(EMP1)|, Bias(SE2)=|SE(JK2)-SE(EMP2)|.

Chapter 3

Estimation of AUC for Multi-Stage Cluster Sampling

3.1 Introduction

It was pointed out in Chapter 2, the sampling units could be population elements or sets of population elements. A set of elements is called a cluster. For example for household surveys clusters consist of subjects (elements) who live in the same geographic area such as a county or city. When the sampling unit is a cluster, the procedure of sampling these units and collecting information from the elements in the sampled clusters is called cluster sampling or single-stage cluster sampling. Multi-stage cluster sampling is a more complex form of cluster sampling. In this case, the population is hierarchically partitioned into disjoint clusters of population, ie., the largest clusters are the first stage clusters, e.g., counties or cities, which are partitioned into second stage clusters, e.g., census tracks, and so forth depending

on the number of stages. In multi-stage cluster sampling a random sample of first clusters are selected and then from the sample first-stage clusters a random sample of second stage clusters are selected and so forth until the last stage sample units are selected. In many sample surveys such as household surveys or institutional surveys, direct population element sampling is not practical and multi-stage cluster sampling is used, where the first-stage clusters consist of population subjects that live in the same geographic area or the clusters consist of hospitals for patient populations. Some of advantages and disadvantages of multi-stage cluster sampling are:

1. The population elements or subjects may be scattered over wide geographic area. One major benefit for multi-stage sampling is that it lowers the travel costs for interviewers to conduct personal interviews among clusters of sampled subjects living near each other. Also, it is more efficient to supervise the field work among geographically-based clusters of interviewed subjects than direct element samples that are geographically scattered, which could result in minimizing the nonresponse rates and measurement errors.

2. Another reason for using multi-stage sampling is that a sampling frame may not exist or it is impossible to form a list of all the units in the target population. For example in national household interview survey, there is no list of all households, therefore in geographically-based multi-stage cluster sampling sequential frame construction can be used, say, by geographic areas from counties, districts, streets, then to households, which can result in tremendous effort savings and reduced effort.

3. However, typically, a multi-stage cluster samples result in less precise estimates than a simple random sample of the same size due to intraclass correlation, i.e, correlation of the observations within clusters. However, often this less precision

is offset by the advantages in 1) and 2) described above.

Here are some terms often seen in cluster sampling:

1. Primary sampling units: The clusters sampled at the first stage of multi-stage cluster sampling are called primary sampling units, abbreviated as PSUs.

2. Secondary sampling units (SSUs) refers the sampling units to be used at the second stage sampling, which could be elements or cluster of elements.

3. Ultimate sampling units (USUs) refers the sampling units in the last-stage sampling.

4. Penultimate sampling units refers to those in the next to the last stage.

The basic weighted estimation technique used in multi-stage sampling is similar to that used with single-stage sampling in that a sample weighted estimator is used where the weights are the inverse of single inclusion probabilities of the sampled subjects. The two-stage cluster sample design here captures much of the complexities of three or more stage cluster sampling while allowing for more manageable notation. Therefore in this chapter, we focus our attention on two-stage cluster sampling.

3.2 Two-Stage Cluster Sampling

3.2.1 Introduction

For two-stage cluster sampling (TSCS), the general procedure is first to select a number of clusters then to choose a specified number of elements from each selected

cluster. The main advantage of two-stage sampling is more flexible than one stage sampling, but it adds some complexities, for example whether the size of clusters are equal or not, what is the intracluster correlation (ICC), how to balance between "cost efficiency" and "variance efficiency", etc.

One may consider different cluster sampling strategies for different situations. For example, an equal probability sampling design i.e., self-weighted sample design or an unequal probability sample design. The simple random sampling of population elements discussed in Chapter 2 is an equal probability sample design. When the size of clusters is equal, it reasonable to employ the simple random sampling of the clusters and then equal sample size simple random sampling in the second stage. However when the cluster sizes are unequal, an unequal probability sampling design - probability proportional-to-size (population size of the cluster) sampling (PPS) of the clusters with equal sample size random sampling in the second stage is beneficial because it will approximately or more closely result in equal sample weighting of sampled subjects. Equal sample weighting will usually result in estimators that have smaller variances than estimators from sample with unequal sample weighting (Korn and Graubard, 1999). Also operationally for geographically based cluster sampling with PPS first stage sampling with equal sample sizes at the second stage assures approximately equal workloads of interviews of study subjects within sampled clusters.

Two-stage cluster PPS sampling technique was used in our simulation studies when the cluster sizes were set to be unequal.

3.2.2 Intracluster Correlation

The intraclass correlation has been regarded within the framework of analysis of variance (ANOVA), and random effects models. Consider the following random effects model

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \tag{3.1}$$

where Y_{ij} is the jth observation in the ith cluster, μ is the overall mean of all the observations in the population, α_i is the cluster-specific random effect that measures the difference between the i-th cluster mean and the population mean. And ϵ_{ij} is the individual-specific residual error. In this model, the α_i are assumed to be identically distributed with mean 0 and variance σ^2_α, and the ϵ_{ij} are assumed to be identically distributed with mean 0 and variance σ^2_ϵ. The α_i, ϵ_{ij} are independent. Then the variance of Y_{ij} is given by $\sigma^2_Y = \sigma^2_\alpha + \sigma^2_\epsilon$.

The population intraclass correlation (ICC) can be defined as the correlation coefficient between any two observations Y_{ij} and Y_{ik} in the same cluster. In this framework,

$$\begin{aligned}
\rho &= \frac{\mathbf{cov}(Y_{ij}, Y_{ik})}{\sqrt{\mathbf{var}(Y_{ij})\mathbf{var}(Y_{ik})}} \\
&= \frac{E(Y_{ij} - \mu)(Y_{ik} - \mu)}{\sigma^2_Y} \\
&= \frac{E(\alpha_i + \epsilon_{ij})(\alpha_i + \epsilon_{ik})}{\sigma^2_Y} \\
&= \frac{E(\alpha_i^2)}{\sigma^2_Y} \\
&= \frac{\sigma^2_\alpha}{(\sigma^2_\alpha + \sigma^2_\epsilon)}
\end{aligned} \tag{3.2}$$

Steps 3 and 4 above, follow from the α_i being independent of ϵ_{ij} and ϵ_{ik}, and ϵ_{ij} being independent to ϵ_{ik}, and the expected values of α_i, ϵ_{ij} and ϵ_{ik} are equal to 0.

Theoretically, values of ρ range from 0 to 1. A very small value for ρ implies that the within-cluster variance is much greater than the between-cluster variance, and a ρ of 0 indicates that observations within a cluster are uncorrelated. As the within-cluster variance (σ_ϵ^2) approaches 0, ρ approaches 1. When $\rho = 1$, all observations within a cluster are identical. In that case the effective sample size is reduced to the number of clusters.

3.2.3 PPS Cluster Sampling Design

PPS cluster sampling have been discussed in a various sampling books (Cochran 1959, Levy and Lemeshow 1991, Korn and Graubard 1999). The basically PPS cluster sampling is when the clusters are assigned a measure of size, usually the population size of the cluster, and the clusters are randomly sampled with probabilities that are proportional to their measure of size. We will consider two stage cluster sampling where the first stage is PPS sampling of the clusters and then a fixed common number of subjects are sampled from each sampled cluster. As mentioned earlier in this chapter this type of sampling results in each sampled subject having same probability of being sampled.

Steps in applying PPS cluster sampling is illustrated in following example - a size of population $T = 10000$ with 20 clusters (Table 3.1):

1. List all clusters (Column "Cluster") and their populations (Column "Size").

Table 3.1: Illustration of PPS cluster sampling technique

Cluster	Size	CumSize	ClstSmpd	Prob1	SubjSmpd	Prob2	Weight
1	498	498					
2	159	657					
3	548	1205					
4	817	2022	1516	40.9 %	50	6.1 %	40
5	282	2304					
6	747	3051					
7	527	3578	3516	26.4 %	50	9.5 %	40
8	351	3929					
9	717	4646					
10	752	5398					
11	418	5816	5516	20.9 %	50	12.0 %	40
12	246	6062					
13	480	6542					
14	333	6875					
15	869	7744	7516	43.5 %	50	5.8 %	40
16	419	8163					
17	837	9000					
18	448	9448					
19	237	9685	9516	11.9 %	50	21.1 %	40
20	315	10000					

2. Calculate the cumulative population size(Column “CumSize”). The last number in this column is the total population, 10000.

3. Determine the number of clusters (”ClstSmp”) that will be sampled, for example selecting 5 clusters from total 20 clusters.

4. Determine sampling interval (SI) by dividing the total population by the number of clusters to be sampled, SI = 10000/5 = 2000.

5. Choose a random number between 1 and the SI to get random start (RS). In this example, the RS is 1516.

6. Calculate the following series: RS; RS + SI; RS + 2*SI; RS + 3*SI; RS + 4*SI.

7. The clusters selected are those for which the cumulative population (Column “CumSize”) contains the numbers in the series we calculated. In this example, the selected clusters are numbers 4, 7, 11, 15 and 19 (Column “Cluster sampled”).

8. A fixed number of subjects (”SubjSmp”) will be randomly drawn from each sampled cluster. In this example, 50 subjects are taken.

The probability of each cluster being sampled (Column “Prob 1”) and the probability of each subject being sampled in each selected cluster (Column “Prob 2”) can be calculated as following:

$$Prob1 = Clusterpopulation * NumberofClusters/TotalPopulation$$

$$Prob2 = Numberofsubjectstobesampledineachcluster/ClusterPopulation$$

The overall weight of a subject being sampled from the population is the inverse of the probability of selection.

$$W = 1/(prob1 * prob2)$$

Overall weight for each selected subject is same, it is 40 in this example (Table 3.1).

PPS sampling technique is used extensively in survey practice. It is easily implemented, and can reduce variance of the survey estimates.

3.2.4 Estimation of AUC for TSCS

Let $U_1, \cdots, U_K$ denote K clusters, $g = 1, 2, \cdots, K$, where $\bigcup_{g=1}^{K} U_g = U$. The size of U_g is denoted by T_g. Suppose k clusters, $g = 1, 2, \cdots, k$, is randomly taken from population, then a sample of size t_g is randomly selected from T_g. The complete sample is $s = \bigcup_{g=1}^{k} s_g$, and the size of s is t. Briefly, the following symbols refer to the whole population or sample:

$s = \{x_1, \cdots, x_m, y_1, \cdots, y_n\}$, containing the size of m diseased subpopulation denoted as c_1 and the size of n nondiseased subpopulation denoted as c_2, so that the size of s is $t = m + n$.

The joint inclusion probability for selecting both i-th and j-th subjects is product of the probability of cluster(s) being sampled and the probability of each subject being sampled in each selected cluster. Note that two subjects could be selected from same cluster or different clusters, which leads to different joint inclusion probabilities. Suppose i-th and j-th subjects are selected from g cluster and g' cluster respectively,

	Population	Sample
Number of clusters	K	k
Size of g-th cluster for diseased subjects	M_g	m_g
Size of g-th cluster for nondiseased subjects	N_g	n_g
Size of g-th cluster	$T_g = M_g + N_g$	$t_g = n_g + m_g$
Total	$T = \sum_{g=1}^{K} T_g$	$t = \sum_{g=1}^{k} t_g$

and $g \neq g'$, then π_{ij} (Please also see formulae 2.2 and 2.3):

$$\pi_{ij} = P(I_{gg'} = 1)P(I_{ij} = 1 \mid (g, g') \in s) = \frac{k(k-1)}{K(K-1)} \frac{t_g t_{g'}}{T_g T_{g'}} \tag{3.3}$$

If $g = g'$, then $P(I_{gg'} = 1) = P(I_{g=1})$ (Please see formula 2.4), and π_{ij}:

$$\pi_{ij} = P(I_{g=1})P(I_{ij} = 1 \mid g \in s) = \frac{k}{K} \frac{t_g(t_g - 1)}{T_g(T_g - 1)} \tag{3.4}$$

The finite population parameter for the AUC that is written using the cluster structure of the finite population is

$$\theta_U = \frac{1}{MN} \sum_{g=1}^{K} \sum_{g'=1}^{K} \sum_{j=1}^{T_{g'}} \sum_{i=1}^{T_g} \delta_{gi}(1 - \delta_{g'j})\psi(X_{gi}, Y_{g'j}) \tag{3.5}$$

Similar to simple random sampling, the weighted estimator for population AUC for two-stage cluster sampling can be formed as

$$\hat{\theta}_U = \frac{1}{\hat{M}\hat{N}} \sum_{g=1}^{K} \sum_{g'=1}^{K} \sum_{j=1}^{T_{g'}} \sum_{i=1}^{T_g} w_{(gi,g'j)} I_{(gi,g'j)} \delta_{gi}(1 - \delta_{g'j})\psi(X_{gi}, Y_{g'j}) \tag{3.6}$$

where

$$w_{(gi,g'j)} = \begin{cases} \frac{K}{k}\frac{T_g(T_g-1)}{t_g(t_g-1)} & \text{if } g = g' \\ \frac{K(K-1)}{k(k-1)}\frac{T_g T_{g'}}{t_g t_{g'}} & \text{if } g \neq g' \end{cases}.$$

The proof of the unbiasedness of the numerator in the formula 3.6 is similar to the previous proof of formula 2.16. Let $\mathbf{A}$ be the expected numerator in formula 3.6.

Proof

$$\begin{aligned} \mathbf{A} &= \mathbf{E}\left[\sum_{g=1}^{K}\sum_{g'=1}^{K}\sum_{j=1}^{T_{g'}}\sum_{i=1}^{T_g} w_{(gi,g'j)} I_{(gi,g'j)}\delta_{gi}(1-\delta_{g'j})\psi(X_{gi}, Y_{g'j})\right] \\ &= \sum_{g=1}^{K}\sum_{g'=1}^{K}\sum_{j=1}^{T_{g'}}\sum_{i=1}^{T_g} \mathbf{E}[w_{(gi,g'j)}]\mathbf{E}[I_{(gi,g'j)}]\mathbf{E}[\delta_{gi}(1-\delta_{g'j})]\mathbf{E}[\psi(X_{gi}, Y_{g'j})] \\ &= \sum_{g=1}^{K}\sum_{g'=1}^{K}\sum_{j=1}^{T_{g'}}\sum_{i=1}^{T_g} \frac{M_{g'}}{T_{g'}}\frac{N_g}{T_g}\theta \\ &= [(N_1M_1 + N_1M_2 + \cdots + N_1M_g) + \cdots + (N_gM_1 + N_gM_2 + \cdots + N_gM_g)]\theta \\ &= NM\theta \end{aligned}$$

At step 3, $\mathbf{E}[w_{(gi,g'j)}]$ and $\mathbf{E}[I_{(gi,g'j)}]$ cancel out.

3.2.5 Variance Estimation

As it has been discussed in Chapter 2.1.4, to define the jackknife variance estimator, the data are first divided into disjoint random groups, and these groups should be mutually independent. For cluster sampling, a single PSU or a set of PSUs can be removed each time. When a single PSU is removed in calculating $\hat{\theta}_{(g)}$ multiplication of a factor of $k/(k-1)$ is needed. The jackknife variance estimator of weighted AUC

for two-stage cluster sampling design is defined by

$$\widehat{V}_{JK}(\hat{\theta}_U) = \frac{k-1}{k}\sum_{g=1}^{k}(\hat{\theta}_{(-g)} - \hat{\theta}_U)^2 \tag{3.7}$$

where the $\hat{\theta}_{(-g)}$ is the estimator of the same functional form as $\hat{\theta}_U$, but computed from the reduced sample by omitting the g-th cluster in the sample.

3.3 Simulation

3.3.1 For TSCS with Equal Cluster Size

Simulation studies are performed to evaluate the performance of the proposed approach for TSCS sampling design under different situations - with equal or unequal size of clusters, with or without intracluster correlation, etc.

We first consider a finite population that is composed of 100 independent clusters with equal cluster size, and the population size is $T = 200000$. The disease prevalence is $d = 0.3$ and the population AUC is controlled at different values, $\theta = 0.95$, 0.9, 0.8, 0.7, 0.6, or 0.55, respectively. We introduce the different intracluster correlations, $\rho =$ 0, 0.2, or 0.4, to the simulated population separately, by using a method via random effect model (Graubard and Korn, 1996; Korn and Graubard, 1999). Specifically, the diseased and nondiseased subjects are generated as following:

$$X_{gi} = \mu_X + \alpha_g + \epsilon_{gi},$$

$$Y_{gj} = \mu_Y + \alpha_g + \epsilon_{gj},$$

where $g = 1, \cdots, K, \quad i = 1, \cdots, M_g$ and $j = 1, \cdots, N_g, \quad \mu_X$ and μ_Y are the means of diseased and nondiseased subjects. Both ϵ_{gi} and ϵ_{gj}, as well as α_g, are normally distributed at mean 0. The variance of α_g is σ_α^2 and the variance of ϵ_{gi} and ϵ_{gj} are same, that is σ_ϵ^2. In this simulation, μ_X is fixed at 5 while μ_Y varies that is determined by the value of θ, and σ_ϵ^2 is fixed at 1 and σ_α^2 is determined by the value of ρ.

For TSCS with equal cluster size, simple random sampling technique is used in the sampling. In the first stage, 30 of 100 clusters are randomly selected, then a size of $t_g = 8$, 16, or 32 is sampled from each selected cluster so that the total sample size t varies from 240, 480, to 960 respectively.

Table 3.2, Table 3.3 and Table 3.4 summarize the respective simulation results from 1000 replicates, for TSCS with equal size of clusters and intracluster correlations $\rho = 0$, 0.2, or 0.4, respectively. Similar patterns in the results can be observed from these three tables. The weighted jackknife standard errors and weighted square root of MSE are consistently larger than unweighted jackknife standard errors and unweighted square root of MSE. However, with increasing sample sizes, (1) the difference between the weighted and unweighted standard errors becomes smaller; (2) both the weighted and unweighted jackknife standard errors approach the weighted and unweighted empirical standard errors, as the biases of standard errors decrease; (3) A reduction in the relative bias of the estimated of population AUC can be observed; (4) the difference between the weighted square root of MSE and unweighted square root of MSE becomes smaller.

When ICC, ρ, increases from 0, 0.2, to 0.4, both weighted and unweighted standard errors increase as well as for RMSE, because positive ICC is known to introduce extra variability and this variability increase with increasing ICC (Korn and Graubard, 1999).

In Table 3.3 and Table 3.4, "SE(JK1.rc)" presents the unweighted jackknife

standard errors computed from randomly grouped clusters, so that the ICC = 0. Results show that the standard errors from these random clusters ("SE(JK1.rc)") are smaller than standard errors with positive ICC ("SE(JK1)"), which confirms that the ICC plays a role in the extra variability in the estimation of the AUC. The results in Table 3.3 and Table 3.4 are consistent to the results in Table 3.2.

Table 3.2: Simulation results for TSCS with equal cluster size, $\rho = 0$

θ	t^*	RelBias1*	RelBias2*	SE(JK1)	RelSE(JK1)	SE(JK2)	RelSE(JK2)	SE(EMP1)	Bias (SE1)	SE(EMP2)	Bias (SE2)	RMSE1	RMSE2
0.95	960	1.0004	0.9959	0.0157	0.0165	0.0160	0.0168	0.0155	0.0002	0.0157	0.0004	0.0155	0.0161
	480	0.9991	0.9894	0.0225	0.0237	0.0235	0.0245	0.0214	0.0011	0.0218	0.0017	0.0214	0.0241
	240	0.9996	0.9796	0.0338	0.0356	0.0366	0.0377	0.0316	0.0022	0.0330	0.0036	0.0316	0.0384
0.9	960	0.9993	0.9948	0.0235	0.0261	0.0238	0.0263	0.0232	0.0004	0.0234	0.0004	0.0232	0.0239
	480	1.0018	0.9922	0.0344	0.0383	0.0354	0.0390	0.0342	0.0002	0.0346	0.0008	0.0342	0.0353
	240	1.0008	0.9811	0.0503	0.0559	0.0529	0.0576	0.0476	0.0028	0.0490	0.0038	0.0476	0.0520
0.8	960	1.0013	0.9968	0.0343	0.0430	0.0346	0.0431	0.0336	0.0007	0.0338	0.0009	0.0336	0.0339
	480	1.0010	0.9915	0.0493	0.0617	0.0503	0.0623	0.0482	0.0012	0.0488	0.0015	0.0482	0.0492
	240	0.9986	0.9803	0.0722	0.0901	0.0749	0.0918	0.0668	0.0054	0.0683	0.0066	0.0668	0.0702
0.7	960	1.0010	0.9964	0.0406	0.0581	0.0409	0.0583	0.0400	0.0006	0.0402	0.0007	0.0400	0.0403
	480	1.0038	0.9941	0.0588	0.0843	0.0598	0.0850	0.0566	0.0022	0.0572	0.0026	0.0566	0.0574
	240	0.9952	0.9769	0.0844	0.1200	0.0878	0.1225	0.0782	0.0062	0.0822	0.0055	0.0783	0.0839
0.6	960	0.9992	0.9954	0.0440	0.0732	0.0446	0.0740	0.0416	0.0024	0.0427	0.0019	0.0416	0.0428
	480	0.9971	0.9936	0.0619	0.1029	0.0643	0.1064	0.0562	0.0057	0.0633	0.0010	0.0562	0.0634
	240	0.9792	0.9786	0.0891	0.1455	0.0958	0.1563	0.0714	0.0177	0.0888	0.0070	0.0725	0.0898
0.55	960	0.9896	0.9966	0.0430	0.0774	0.0454	0.0823	0.0377	0.0053	0.0460	0.0006	0.0381	0.0461
	480	0.9712	0.9907	0.0616	0.1088	0.0659	0.1188	0.0468	0.0148	0.0646	0.0013	0.0495	0.0648
	240	0.9466	0.9864	0.0892	0.1535	0.0985	0.1766	0.0598	0.0294	0.0911	0.0074	0.0673	0.0914

Note: * 1 = unweighted estimator, 2 = weighted estimator. RelBias1=$\theta/\hat{\theta}1$, RelBias2=$\theta/\hat{\theta}2$, RelSE(JK1)=SE(JK1)/$\hat{\theta}1$, RelSE(JK2)=SE(JK2)/$\hat{\theta}2$, Bias(SE1)=|SE(JK1)-SE(EMP1)|, Bias(SE2)=|SE(JK2)-SE(EMP2)|.

Table 3.3: Simulation results for TSCS with equal cluster size, $\rho = 0.2$

θ	t	RelBias1*	RelBias2*	SE(JK1.rc)	SE(JK1)	RelSE(JK1)	SE(JK2)	RelSE(JK2)	SE(EMP1)	Bias (SE1)	SE(EMP2)	Bias (SE2)	RMSE1	RMSE2
0.95	960	0.9981	0.9951	0.0148	0.0170	0.0178	0.0182	0.0190	0.0166	0.0004	0.0175	0.0007	0.0167	0.0181
	480	0.9996	0.9916	0.0229	0.0240	0.0252	0.0259	0.0270	0.0232	0.0008	0.0245	0.0014	0.0232	0.0258
	240	0.9979	0.9797	0.0340	0.0332	0.0349	0.0370	0.0381	0.0313	0.0019	0.0335	0.0035	0.0314	0.0388
0.9	960	0.9973	0.9951	0.0223	0.0256	0.0283	0.0269	0.0298	0.0244	0.0011	0.0255	0.0014	0.0246	0.0259
	480	0.9977	0.9905	0.0347	0.0349	0.0387	0.0369	0.0406	0.0344	0.0005	0.0359	0.0010	0.0344	0.0369
	240	0.9982	0.9805	0.0510	0.0507	0.0563	0.0544	0.0593	0.0482	0.0025	0.0508	0.0037	0.0482	0.0538
0.8	960	0.9965	0.9948	0.0318	0.0354	0.0441	0.0367	0.0456	0.0344	0.0010	0.0354	0.0013	0.0345	0.0356
	480	0.9963	0.9882	0.0490	0.0499	0.0621	0.0519	0.0641	0.0478	0.0021	0.0492	0.0027	0.0479	0.0501
	240	1.0036	0.9841	0.0751	0.0736	0.0924	0.0775	0.0953	0.0680	0.0056	0.0710	0.0065	0.0681	0.0722
0.7	960	0.9975	0.9956	0.0382	0.0412	0.0587	0.0423	0.0601	0.0401	0.0011	0.0409	0.0013	0.0402	0.0410
	480	0.9937	0.9865	0.0582	0.0593	0.0841	0.0614	0.0865	0.0566	0.0026	0.0582	0.0031	0.0568	0.0590
	240	0.9961	0.9798	0.0871	0.0848	0.1206	0.0894	0.1252	0.0792	0.0055	0.0834	0.0060	0.0793	0.0847
0.6	960	1.0001	0.9977	0.0411	0.0443	0.0738	0.0456	0.0759	0.0427	0.0016	0.0439	0.0017	0.0427	0.0440
	480	0.9935	0.9901	0.0614	0.0621	0.1028	0.0654	0.1079	0.0566	0.0055	0.0629	0.0024	0.0567	0.0632
	240	0.9763	0.9823	0.0897	0.0891	0.1450	0.0968	0.1585	0.0744	0.0147	0.0964	0.0004	0.0758	0.0971
0.55	960	0.9878	0.9937	0.0409	0.0433	0.0778	0.0459	0.0829	0.0369	0.0064	0.0444	0.0015	0.0375	0.0446
	480	0.9683	0.9852	0.0612	0.0618	0.1089	0.0668	0.1197	0.0482	0.0136	0.0657	0.0012	0.0515	0.0662
	240	0.9399	0.9800	0.0905	0.0877	0.1499	0.0976	0.1739	0.0621	0.0257	0.0952	0.0024	0.0713	0.0958

Note: * 1 = unweighted estimator, 2 = weighted estimator. RelBias1=$\theta/\hat{\theta}1$, RelBias2=$\theta/\hat{\theta}2$, RelSE(JK1)=SE(JK1)/$\hat{\theta}1$, RelSE(JK2)=SE(JK2)/$\hat{\theta}2$, Bias(SE1)=|SE(JK1)-SE(EMP1)|, Bias(SE2)=|SE(JK2)-SE(EMP2)|.

Table 3.4: Simulation results for TSCS with equal cluster size, $\rho = 0.4$

θ	t*	RelBias1*	RelBias2*	SE(JK1.rc)	SE(JK1)	RelSE(JK1)	SE(JK2)	RelSE(JK2)	SE(EMP1)	Bias (SE1)	SE(EMP2)	Bias (SE2)	RMSE1	RMSE2
0.95	960	0.9966	0.9950	0.0136	0.0210	0.0221	0.0232	0.0243	0.0196	0.0014	0.0213	0.0019	0.0199	0.0219
	480	0.9967	0.9899	0.0209	0.0264	0.0277	0.0293	0.0305	0.0252	0.0012	0.0273	0.0020	0.0254	0.0290
	240	0.9966	0.9794	0.0330	0.0361	0.0378	0.0409	0.0421	0.0334	0.0027	0.0367	0.0042	0.0335	0.0418
0.9	960	0.9944	0.9946	0.0203	0.0303	0.0335	0.0330	0.0364	0.0291	0.0012	0.0314	0.0016	0.0296	0.0318
	480	0.9966	0.9917	0.0315	0.0392	0.0435	0.0427	0.0470	0.0375	0.0017	0.0403	0.0024	0.0377	0.0410
	240	0.9947	0.9792	0.0503	0.0527	0.0583	0.0578	0.0629	0.0476	0.0051	0.0511	0.0068	0.0479	0.0545
0.8	960	0.9936	0.9957	0.0293	0.0402	0.0499	0.0428	0.0533	0.0387	0.0015	0.0409	0.0020	0.0390	0.0410
	480	0.9934	0.9902	0.0456	0.0533	0.0662	0.0569	0.0704	0.0506	0.0028	0.0533	0.0036	0.0508	0.0538
	240	0.9942	0.9799	0.0725	0.0742	0.0923	0.0797	0.0976	0.0685	0.0058	0.0725	0.0072	0.0686	0.0744
0.7	960	0.9910	0.9926	0.0345	0.0439	0.0621	0.0461	0.0654	0.0420	0.0019	0.0439	0.0023	0.0425	0.0442
	480	0.9943	0.9910	0.0538	0.0611	0.0868	0.0643	0.0910	0.0594	0.0017	0.0622	0.0021	0.0596	0.0625
	240	0.9875	0.9736	0.0849	0.0852	0.1201	0.0911	0.1267	0.0804	0.0048	0.0854	0.0056	0.0809	0.0875
0.6	960	0.9932	0.9931	0.0380	0.0445	0.0737	0.0465	0.0769	0.0440	0.0005	0.0458	0.0007	0.0442	0.0460
	480	0.9928	0.9923	0.0570	0.0634	0.1048	0.0677	0.1120	0.0567	0.0067	0.0643	0.0034	0.0569	0.0644
	240	0.9755	0.9757	0.0891	0.0903	0.1469	0.0995	0.1618	0.0733	0.0170	0.0906	0.0088	0.0749	0.0919
0.55	960	0.9898	0.9966	0.0376	0.0435	0.0783	0.0467	0.0847	0.0370	0.0065	0.0447	0.0020	0.0374	0.0447
	480	0.9707	0.9889	0.0566	0.0622	0.1098	0.0682	0.1227	0.0466	0.0157	0.0645	0.0037	0.0495	0.0648
	240	0.9412	0.9767	0.0900	0.0890	0.1522	0.1005	0.1784	0.0610	0.0280	0.0933	0.0072	0.0700	0.0942

Note: * 1 = unweighted estimator, 2 = weighted estimator. RelBias1=$\theta/\hat{\theta}1$, RelBias2=$\theta/\hat{\theta}2$, RelSE(JK1)=SE(JK1)/$\hat{\theta}1$, RelSE(JK2)=SE(JK2)/$\hat{\theta}2$, Bias(SE1)=|SE(JK1)-SE(EMP1)|, Bias(SE2)=|SE(JK2)-SE(EMP2)|.

3.3.2 For TSCS with Unequal Cluster Size

Now we allow the size of the clusters to be unequal for TSCS. A simulated population size is $T = 200000$, containing 100 clusters with cluster size varying randomly from 1500 - 3500. Simultaneously, ICC is fixed at $\rho = 0$ or 0.2, respectively. Since the cluster sizes are different, PPS cluster sampling technique is employed in the sampling design.

Table 3.5, Table 3.6 summarize the respective simulation results based on 1000 replicates, for intracluster correlations $\rho = 0$ or 0.2, separately. Overall, results for TSCS with unequal cluster sizes are consistent with the results for TSCS with equal cluster sizes.

Weighted jackknife standard errors and weighted RMSE are consistently larger than unweighted jackknife standard errors and unweighted RMSE. With increase of the sample size, the weighted standard errors tend to coverage to unweighted standard errors, the difference between weighted and unweighted bias standard errors becomes smaller. And the standard errors are increasing with increase of ICC from 0 to 0.2.

Survey data with unequal size of clusters may increase the variation as compared to the case where there are equal cluster sizes. In this simulation setting for unequal size of clusters, we utilize PPS sampling technique, but we do not observe enlarged standard errors in Table 3.5 and Table 3.6 as compared to corresponding standard errors in Table 3.2 and Table 3.3 for TSCS with equal size of clusters.

Table 3.5: Simulation results for TSCS with unequal cluster size, $\rho = 0$

θ	t	RelBias1*	RelBias2*	SE(JK1)	RelSE(JK1)	SE(JK2)	RelSE(JK2)	SE(EMP1)	Bias (SE1)	SE(EMP2)	Bias (SE2)	RMSE1	RMSE2
0.95	960	1.0008	0.9967	0.0157	0.0166	0.0161	0.0169	0.0158	0.0001	0.0160	0.0001	0.0159	0.0163
	480	0.9993	0.9908	0.0227	0.0239	0.0237	0.0247	0.0220	0.0007	0.0225	0.0012	0.0220	0.0242
	240	0.9988	0.9798	0.0330	0.0347	0.0357	0.0369	0.0313	0.0018	0.0325	0.0032	0.0313	0.0380
0.9	960	0.9990	0.9949	0.0235	0.0260	0.0237	0.0262	0.0235	0.0000	0.0236	0.0001	0.0235	0.0241
	480	1.0007	0.9916	0.0343	0.0381	0.0351	0.0387	0.0325	0.0017	0.0330	0.0021	0.0325	0.0339
	240	1.0006	0.9814	0.0504	0.0561	0.0529	0.0577	0.0477	0.0028	0.0491	0.0038	0.0477	0.0520
0.8	960	1.0012	0.9971	0.0342	0.0428	0.0346	0.0431	0.0345	0.0003	0.0347	0.0001	0.0345	0.0347
	480	0.9988	0.9906	0.0490	0.0612	0.0499	0.0618	0.0474	0.0016	0.0480	0.0019	0.0474	0.0486
	240	0.9977	0.9789	0.0718	0.0896	0.0744	0.0911	0.0693	0.0025	0.0711	0.0033	0.0693	0.0732
0.7	960	0.9986	0.9932	0.0405	0.0577	0.0407	0.0578	0.0399	0.0006	0.0401	0.0006	0.0399	0.0404
	480	1.0016	0.9913	0.0584	0.0835	0.0593	0.0840	0.0564	0.0020	0.0573	0.0020	0.0564	0.0577
	240	1.0020	0.9823	0.0844	0.1208	0.0879	0.1234	0.0816	0.0028	0.0851	0.0028	0.0816	0.0860
0.6	960	1.0004	0.9970	0.0436	0.0728	0.0443	0.0737	0.0431	0.0005	0.0442	0.0001	0.0431	0.0443
	480	0.9939	0.9903	0.0616	0.1021	0.0641	0.1058	0.0555	0.0062	0.0620	0.0021	0.0556	0.0622
	240	0.9770	0.9780	0.0891	0.1451	0.0959	0.1563	0.0734	0.0157	0.0915	0.0044	0.0747	0.0925
0.55	960	0.9867	0.9924	0.0431	0.0773	0.0452	0.0815	0.0365	0.0066	0.0442	0.0010	0.0373	0.0444
	480	0.9718	0.9907	0.0614	0.1084	0.0657	0.1183	0.0467	0.0147	0.0639	0.0018	0.0493	0.0641
	240	0.9403	0.9823	0.0885	0.1512	0.0971	0.1734	0.0626	0.0259	0.0954	0.0017	0.0716	0.0959

Note: * 1 = unweighted estimator, 2 = weighted estimator. RelBias1=$\theta/\hat{\theta}1$, RelBias2=$\theta/\hat{\theta}2$, RelSE(JK1)=SE(JK1)/$\hat{\theta}1$, RelSE(JK2)=SE(JK2)/$\hat{\theta}2$, Bias(SE1)=|SE(JK1)-SE(EMP1)|, Bias(SE2)=|SE(JK2)-SE(EMP2)|.

Table 3.6: Simulation results for TSCS with unequal cluster size, $\rho = 0.2$

θ	t	RelBias1*	RelBias2*	SE(JK1.rc)	SE(JK1)	RelSE(JK1)	SE(JK2)	RelSE(JK2)	SE(EMP1)	Bias (SE1)	SE(EMP2)	Bias (SE2)	RMSE1	RMSE2
0.95	960	0.9981	0.9956	0.0147	0.0171	0.0180	0.0183	0.0192	0.0166	0.0005	0.0174	0.0168	0.0166	0.0179
	480	0.9986	0.9908	0.0232	0.0233	0.0245	0.0252	0.0263	0.0228	0.0005	0.0242	0.0235	0.0228	0.0257
	240	0.9987	0.9807	0.0338	0.0338	0.0355	0.0375	0.0387	0.0319	0.0018	0.0341	0.0337	0.0320	0.0389
0.9	960	0.9972	0.9954	0.0222	0.0253	0.0280	0.0266	0.0294	0.0244	0.0010	0.0253	0.0262	0.0245	0.0257
	480	0.9978	0.9908	0.0349	0.0349	0.0387	0.0368	0.0406	0.0343	0.0007	0.0358	0.0366	0.0343	0.0368
	240	0.9977	0.9802	0.0508	0.0507	0.0562	0.0544	0.0593	0.0488	0.0019	0.0511	0.0539	0.0488	0.0543
0.8	960	0.9963	0.9950	0.0321	0.0354	0.0441	0.0367	0.0456	0.0340	0.0014	0.0349	0.0430	0.0341	0.0351
	480	0.9977	0.9912	0.0503	0.0498	0.0621	0.0517	0.0641	0.0489	0.0010	0.0504	0.0599	0.0489	0.0509
	240	0.9969	0.9797	0.0732	0.0726	0.0905	0.0764	0.0936	0.0715	0.0011	0.0741	0.0877	0.0715	0.0759
0.7	960	0.9962	0.9947	0.0380	0.0411	0.0584	0.0421	0.0599	0.0394	0.0016	0.0402	0.0572	0.0395	0.0404
	480	0.9979	0.9911	0.0593	0.0587	0.0836	0.0605	0.0856	0.0573	0.0014	0.0586	0.0820	0.0573	0.0590
	240	0.9966	0.9801	0.0862	0.0849	0.1209	0.0896	0.1255	0.0836	0.0014	0.0876	0.1206	0.0836	0.0888
0.6	960	0.9966	0.9945	0.0409	0.0438	0.0728	0.0450	0.0746	0.0420	0.0018	0.0431	0.0736	0.0420	0.0432
	480	0.9948	0.9917	0.0621	0.0622	0.1031	0.0654	0.1081	0.0569	0.0053	0.0629	0.1037	0.0570	0.0631
	240	0.9761	0.9801	0.0898	0.0892	0.1450	0.0967	0.1580	0.0747	0.0144	0.0951	0.1519	0.0761	0.0959
0.55	960	0.9885	0.9943	0.0405	0.0431	0.0775	0.0456	0.0825	0.0363	0.0069	0.0438	0.0805	0.0368	0.0439
	480	0.9726	0.9916	0.0614	0.0615	0.1088	0.0666	0.1200	0.0459	0.0156	0.0637	0.1165	0.0485	0.0638
	240	0.9390	0.9802	0.0897	0.0885	0.1510	0.0984	0.1753	0.0635	0.0250	0.0967	0.1717	0.0729	0.0973

Note: * 1 = unweighted estimator, 2 = weighted estimator. RelBias1=$\theta/\hat{\theta}1$, RelBias2=$\theta/\hat{\theta}2$, RelSE(JK1)=SE(JK1)/$\hat{\theta}1$, RelSE(JK2)=SE(JK2)/$\hat{\theta}2$, Bias(SE1)=|SE(JK1)-SE(EMP1)|, Bias(SE2)=|SE(JK2)-SE(EMP2)|.

Chapter 4

Estimation of AUC for Stratified Multi-Stage Cluster Sampling

4.1 Introduction

In stratified sampling,there is a variate, e.g., geographic region, known for all of the sample units, e.g., PSUs consisting of counties or cities. This variate is used to partition the sample units into categories called strata where a random sample of the sample units is drawn from each stratum. There are several purposes for using stratified sampling including increasing efficiency of the sample design by choosing strata so the measurements of interest are less variable within strata than between strata. Another purpose is to control the sample sizes for strata to assure that substantial numbers of measurements are taken for subjects that have certain characteristics such being from a minority racial group. In contrast, cluster sampling is mainly used to reduce costs of conducting the survey even though it can increase the variability of

the estimates due to intracluster correlation of the observations.

In terms of cost and efficiency, stratified multi-stage cluster sampling design combines the advantages of both stratification and multiple stage cluster sampling. This is the major reason why stratified multi-stage cluster sampling is so popular in survey, particularly for many large-scale national household surveys.

In this chapter we want to evaluate the performance of our methods for stratified multistage cluster sample designs. As pointed out previously, multistage sampling beyond two-stage does not add difficulties in the estimation if the same basic sampling design is applied independently within each PSU. In this chapter, we will mainly focus on stratified two-stage cluster sampling (STSCS). We shall give the formulae for the estimation of weighted AUC, and for the jackknife and balanced repeated replication methods for variance estimation for STSCS. We compare simulation results between the jackknife method and balanced repeated replication methods and also provide simulation results under informative sampling design for STSCS.

4.2 Estimation of AUC for STSCS

Suppose the finite population has H strata, each stratum has K clusters such as $U = \bigcup_{h=1}^{H} U_h$, where $U_h = \bigcup_{g=1}^{K} U_{hg}$, U_{hg} is the group population for g-th cluster within h stratum. The size of U_{hg} is denoted by $T_{hg} = N_{hg} + M_{hg}$, and $\sum_{g=1}^{K} T_{hg} = T_h$, $\sum_{h=1}^{H} T_h = T$. Two-stage cluster sampling are performed independently for each stratum, so the sampling procedure for each stratum is same as the procedure for two-stage cluster sampling that was described in previous chapter. Let $s = \bigcup_{h=1}^{H} s_h$, where $s_h = \bigcup_{g=1}^{K} s_{hg}$, and the size of s_{hg} is denoted by $t_{hg} = n_{hg} + m_{hg}$, and $\sum_{g=1}^{K} t_{hg} = t_h$,

$\sum_{h=1}^{H} t_h = t$.

Briefly, the following symbols refer to the whole population or sample:

	Population	Sample
Number of strata	H	H
Number of clusters for h-th stratum	h_K	h_k
Diseased in g-th cluster h-th stratum	M_{hg}	m_{hg}
Nondiseased in g-th cluster h-th stratum	N_{hg}	n_{hg}
Size of g-th cluster in h-th stratum	$T_{hg} = M_{hg} + N_{hg}$	$t_{hg} = n_{hg} + m_{hg}$
Diseased total in h-th stratum	$M_h = \sum_{g=1}^{K} M_{hg}$	$m_h = \sum_{g=1}^{k} m_{hg}$
Nondiseased total in h-th stratum	$N_h = \sum_{g=1}^{K} N_{hg}$	$n_h = \sum_{g=1}^{k} n_{hg}$
Total in h-th stratum	$T_h = \sum_{g=1}^{K} T_{hg}$	$t_h = \sum_{g=1}^{k} t_{hg}$
Diseased total	$M = \sum_{h=1}^{H} M_h$	$m = \sum_{h=1}^{H} m_h$
Nondiseased total	$N = \sum_{h=1}^{H} N_h$	$n = \sum_{h=1}^{H} n_h$
Total	$T = \sum_{h=1}^{H} T_h$	$t = \sum_{h=1}^{H} t_h$

The joint inclusion probability for selecting diseased subject and nondiseased subject for STSCS is similar to the one for TSCS which is the product of the probability of cluster(s) being sampled and the probability of each subject being sampled in each selected cluster. But now it is more complicated since we allow diseased subjects and nondiseased subjects to be compared not only within each stratum (same situation as for TSCS) but also across strata. Suppose hgi-th subject and $h'g'j$-th subject are selected from g-th cluster within h-th stratum and g'-th cluster within h'-th stratum respectively. If hgi-th and $h'g'j$-th subjects are drawn from different clusters but within same stratum, i.e., $g \neq g'$ and $h = h'$, then $\pi_{(hgi,h'g'j)}$ can be

defined as(see formulae 2.2 and 2.3):

$$\begin{aligned} \pi_{(hgi,h'g'j)} &= P(I_{(hg,hg')}=1)P(I_{(hgi,h'gj)}=1 \mid (hg,hg') \in s) && (4.1)\\ &= \frac{k(k-1)}{K(K-1)}\frac{t_{hg}t_{hg'}}{T_{hg}T_{hg'}} && (4.2) \end{aligned}$$

If $g = g'$ and $h = h'$, then $P(I_{(hg,h'g')}=1) = P(I_{hg}=1)$ (see formula 2.4), and

$$\begin{aligned} \pi_{(hgi,h'g'j)} &= P(I_{hg}=1)P(I_{(hgi,hgj)}=1 \mid hg \in s) && (4.3)\\ &= \frac{k}{K}\frac{t_{hg}(t_{hg}-1)}{T_{hg}(T_{hg}-1)} && (4.4) \end{aligned}$$

If $h \neq h'$, then

$$\begin{aligned} \pi_{(hgi,h'g'j)} &= P(I_{(hg,h'g')}=1)P(I_{(hgi,h'g'j)}=1 \mid (hg,h'g') \in s) && (4.5)\\ &= \frac{k^2}{K^2}\frac{t_{hg}t_{h'g'}}{T_{hg}T_{h'g'}} && (4.6) \end{aligned}$$

Similarly to stratified simple random sampling, for stratified two-stage cluster sampling, the population parameter for the AUC that is written using the stratification and cluster structure of the finite population is

$$\theta_U = \frac{1}{\hat{M}\hat{N}} \sum_{h=1}^{H}\sum_{h'=1}^{H}\sum_{g=1}^{K}\sum_{g'=1}^{K}\sum_{j=1}^{\hat{T}_{h'g'}}\sum_{i=1}^{\hat{T}_{hg}} \psi(X_{gi}, Y_{g'j}) \quad (4.7)$$

and the weighted estimator for population AUC can be defined as

$$
\begin{aligned}
\hat{\theta} &= \frac{1}{\hat{N}\hat{M}} \sum_{h=1}^{H}\sum_{h'=1}^{H}\sum_{g=1}^{K}\sum_{g'=1}^{K}\sum_{j=1}^{\hat{T}_{h'g'}}\sum_{i=1}^{\hat{T}_{hg}} w_{(hgi,h'g'j)} I_{(hgi,h'g'j)} \\
&\quad \delta_{hgi}(1-\delta_{h'g'j})\psi(X_{hgi}, Y_{h'g'j}) \\
&= \sum_{h=1}^{H}\sum_{h'=1}^{H} \hat{p}_h \hat{\theta}_{(h,h')} \hat{q}_{h'} \\
&= \hat{p}^T \hat{\theta}_{(L\times L)} \hat{q}
\end{aligned}
\tag{4.8}
$$

where

$$
w_{(hgi,h'g'j)} = \begin{cases} \frac{K}{k}\frac{T_{hg}(T_{hg}-1)}{t_{hg}(t_{hg}-1)} & \text{if } h = h' \text{ and } g = g' \\ \frac{K(K-1)}{k(k-1)}\frac{T_{hg}T_{hg'}}{t_{hg}t_{hg'}} & \text{if } h = h' \text{ and } g \neq g' \\ \frac{K^2 T_{hg}T_{h'g'}}{k^2 t_{hg}t_{h'g'}} & \text{if } h \neq h' \end{cases}
$$

$\hat{p}$ and $\hat{q}$ are (H $\times$ 1) vectors, $\hat{p}^T = \hat{p}_1, \cdots, \hat{p}_H$ and $\hat{q}^T = \hat{q}_1, \cdots, \hat{q}_H$, $\hat{p}_h = \hat{M}_h/\hat{M}, \hat{q}_{h'} = \hat{N}_{h'}/\hat{N}$.

The estimator of weighted AUC for the finite population is the average of the sum of the kernel function ($\psi(.)$) over three levels, i.e., subjects level, clusters level, and strata level.

4.3 Jackknife Method

Similar to unstratified cluster sampling, for stratified multistage cluster sampling, the jackknife estimates can be calculated by removing each PSU then the variance of population AUC is computed between the jackknife estimates. After removing

g-th independent cluster from h-th stratum, the sample weights of the remaining data in the g-th cluster are multiplied by a factor of $h_k/(h_k - 1)$, without changing the sample weights of the observations in the other strata. The formula of jackknife variance estimator for stratified multistage sampling is:

$$\widehat{V}_{JK}(\hat{\theta}_U) = \sum_{h=1}^{H}[\frac{h_k - 1}{h_k}\sum_{g=1}^{h_k}(\hat{\theta}_{(-hg)} - \hat{\theta}_U)^2] \tag{4.9}$$

where the $\hat{\theta}_{(-hg)}$ is the estimator of the same functional form as $\hat{\theta}_U$, but computed from the reduced sample by omitting the g-th cluster in h stratum.

4.4 Balanced Repeated Replications for Variance Estimation

If the sample design can be approximately by or has sample design of two PSUs are randomly sampled from each stratum, then as an alternative to the jackknife method, balanced repeated replications (BRR) can be utilized for variance estimation for our proposed estimator of weighted AUC. In section 4.5.2, we conduct simulation studies to compare the jackknife and BRR methods for variance estimation.

The balanced repeated replications (BRR) or balanced half-samples for variance estimation technique has been used to estimate the variances of linear and nonlinear statistics for complex survey designs since the late 1950s and early 1960s (For a detailed discussion, please see Wolter, 2003). McCarthy (1966, 1969) introduced and developed the mathematical formula based on Plackett-Burman experimental designs (Plackett and Burman, 1946). The BRR method was typically developed

for sample designs where only two PSUs are selected from each stratum due to cost consideration. Replicates are formed by choosing one of the two sampled PSUs, including all the observation sampled from the PSU, from each stratum to form a sample that is approximately half of the original sample size. If there are H strata in the sample, 2^H replicates can be generated. Note that if H is large, this method may be infeasible. To minimize the computations, the PSUs are chosen from each stratum according to a appropriate sized Hadamard matrix. The rows of Hadamard matrix are mutually orthogonal to ensure that for linear statistics such as unweighted mean that the BRR variance estimates corresponds to usual stratified variance estimator, so called orthogonal balanced (hence the name of technique). The variance estimator is obtained by computing the variability across the half-sample replicates.

To proceed with the BRR variance estimation, a L $\times$ L Hadamard matrix is chosen so that L is the next multiple of 4 greater than the number of sampling strata H of the sample design. The first column of the matrix is dropped, which maintains balance in the replicates, and each and the L-1 column identify which PSUs from each stratum are used to produce each L replicates. Let $\hat{\theta}_l$ be the weighted estimate of l-th replicate. The final weights would be the BRR weights multiplied by sample weights. The BRR weights are zero weighted in one half of the sample and double-weighted in the other half determined by the entry of Hadamard matrix.

Based on the formula provided by Wolter (2003), the unbiased BRR variance estimator for weighted AUC can be formed as:

$$\widehat{V}_{BRR}(\hat{\theta}_U) = \frac{1}{L}\sum_{l=1}^{L}(\hat{\theta}_l - \hat{\theta}_U)^2 \tag{4.10}$$

in other words the estimate for variance of weighted AUC is the average of $(\hat{\theta}_l - \hat{\theta}_U)^2$.

Fay's method is a generalization of BRR (Fay, 1989). Fay's idea is to use the

weights of 0.5 and 1.5 instead of 0 or 2 multiplied by each sample weight accounting for the replicates using half samples within each stratum. More generally, weights can be k and 2 - k, for 0 $\leq$k$<$ 1. BRR is the special case, where $k = 0$. The variance estimate is then $\widehat{V}_{BRR}/(1-k)^2$, where $\widehat{V}_{BRR}$ is the estimate given by the BRR formula above. In this dissertation, only the typical BRR method (the weights of 0 and 2) is used.

4.5 Domain ROC estimation

In many surveys, we are interested in making estimates not only for the whole population U, but also for specific groups in the population. The individuals in a group usually share some common characteristics that can be based on age, sex, race or some other variables, for example, all female individuals, old adults (70+ years), or a low-income group in the population. These groups are called domains.

Let the domain of interest be denoted U_d, where $U_d \subset U$. And let ξ be a domain indicator to indicate whether or not an individual belongs to the specific domain, so the estimator of a domain specific AUC can be extended by adding the domain indicator variable. For example, under STSCS sample design the estimator of domain specific AUC for the finite population can be extended from the formula 4.8 as following:

$$\begin{aligned}\hat{\theta} \;=\; & \frac{1}{\hat{N}\hat{M}} \sum_{h=1}^{H}\sum_{h'=1}^{H}\sum_{g=1}^{K}\sum_{g'=1}^{K}\sum_{j=1}^{T_{h'g'}}\sum_{i=1}^{T_{hg}} w_{(hgi,h'g'j)} I_{(hgi,h'g'j)} \\ & \delta_{hgi}(1-\delta_{h'g'j})\xi_{hgi}\xi_{h'g'j}\psi(X_{hgi}, Y_{h'g'j})\end{aligned} \tag{4.11}$$

Note that sample weights are unchanged, and population structure remains same in the domain estimation (4.11). We still can use the Jackknife method for variance estimation, and the same formula can be applied as is discussed for each sample design.

4.6 Simulation

Simulation studies are carried out to assess the performance of our proposed method for stratified multi-stage cluster samples. The scenario of STSCS with unequal size of strata and unequal size of clusters is considered in this section. We study the accuracy of jackknife method along with the BRR method for variance estimation of our proposed estimator of weighted AUC. We also consider informative sampling designs compared to the noninformative sampling designs.

4.6.1 Simulation results for STSCS with Unequal Size of Stratum and Cluster

A size of finite population $T = 200000$ is generated with $H = 8$ strata, and stratum sizes varying from 1500 to 3500. Each stratum is composed of 100 unequal size of clusters. We assume that the stratum-specific AUCs (θ_h) randomly vary across the strata in the range of $\theta_h = \theta \pm 0.1$. The population AUC ($\theta$) is chosen from 0.9 to 0.6 by 0.1. ICCs are varied from $\rho = 0$ to $\rho = 0.2$, respectively.

Since the sampling in each stratum is done independently, different sampling techniques could be utilized to different strata. In this section, we applied PPS cluster sampling in all strata. The total sample size drawn from the population was, $t =$

400, 800, or 1200, respectively. At first stage, 10 of 100 clusters are selected for each stratum. The sample size for each stratum (t_h) is determined by the population size for each stratum (T_h), i.e., t_h the sample sizes for stratum is proportional to T_h, which results in sampled cluster sizes varying from 20 to 80. Then a fixed size of t_{hg} is selected from each sampled cluster within each stratum. The diseased and nondiseased subjects for each stratum are generated independently and similarly to TSCS in chapter 3.3.1.

Table 4.1 shows the simulation results for STSCS with unequal size of strata , unequal size of clusters, and $\rho = 0$. Similarly, after accounting for sample weights, the weighted jackknife standard errors and weighted RMSEs are inflated, as compared to unweighted jackknife standard errors and unweighted RMSEs, and the difference between weighted and unweighted bias of standard errors decreases with increasing sample size (t).

When $\rho = 0.2$ (Table 4.2), as expected the weighted and unweighted standard errors become larger relative to the standard errors from Table 4.1.

If we compare the results in Table 4.1 and Table 4.2 to the results from Table 3.4 and Table 3.5 for two-stage cluster sampling with unequal size of clusters , the standard errors from Table 4.1 and Table 4.2 are smaller. This confirms that stratified sampling technique can reduce the variance of the estimates of AUC.

Table 4.1: Simulation results for STSCS, $T_h \neq T_{h'}$, $T_{hg} \neq T_{hg'}$, and ρ =0

θ	t	RelBias1*	RelBias2*	SE(JK1)	RelSE(JK1)	SE(JK2)	RelSE(JK2)	SE(EMP1)	Bias(SE1)	SE(EMP2)	Bias(SE2)	RMSE1	RMSE2
0.9	1200	0.9998	0.9950	0.0097	0.0108	0.0099	0.0109	0.0095	0.0003	0.0095	0.0003	0.0095	0.0106
	800	0.9993	0.9912	0.0122	0.0135	0.0125	0.0137	0.0114	0.0008	0.0116	0.0009	0.0114	0.0141
	400	0.9992	0.9807	0.0191	0.0212	0.0200	0.0218	0.0168	0.0023	0.0173	0.0026	0.0168	0.0248
0.8	1200	0.9998	0.9950	0.0140	0.0175	0.0142	0.0176	0.0138	0.0003	0.0138	0.0003	0.0138	0.0144
	800	0.9992	0.9911	0.0175	0.0218	0.0178	0.0220	0.0165	0.0010	0.0167	0.0011	0.0165	0.0182
	400	0.9994	0.9810	0.0273	0.0342	0.0283	0.0346	0.0244	0.0029	0.0250	0.0032	0.0244	0.0295
0.7	1200	0.9998	0.9949	0.0166	0.0238	0.0168	0.0238	0.0164	0.0002	0.0165	0.0002	0.0164	0.0169
	800	0.9989	0.9909	0.0207	0.0296	0.0210	0.0298	0.0194	0.0014	0.0196	0.0014	0.0194	0.0206
	400	0.9979	0.9809	0.0319	0.0455	0.0333	0.0466	0.0285	0.0034	0.0299	0.0034	0.0285	0.0329
0.6	1200	0.9969	0.9950	0.0178	0.0295	0.0182	0.0302	0.0169	0.0009	0.0180	0.0002	0.0170	0.0183
	800	0.9925	0.9907	0.0219	0.0362	0.0229	0.0377	0.0197	0.0022	0.0214	0.0014	0.0202	0.0221
	400	0.9765	0.9810	0.0327	0.0532	0.0361	0.0590	0.0261	0.0066	0.0325	0.0036	0.0298	0.0345

Note: * 1 = unweighted estimator, 2 = weighted estimator. RelBias1=$\theta/\hat{\theta}1$, RelBias2=$\theta/\hat{\theta}2$, RelSE(JK1)=SE(JK1)/$\hat{\theta}1$, RelSE(JK2)=SE(JK2)/$\hat{\theta}2$, Bias(SE1)=|SE(JK1)-SE(EMP1)|, Bias(SE2)=|SE(JK2)-SE(EMP2)|.

Table 4.2: Simulation results for STSCS, $T_h \neq T_{h'}$, $T_{hg} \neq T_{hg'}$, and ρ =0.2

θ	t	RelBias1*	RelBias2*	SE(JK1)	RelSE(JK1)	SE(JK2)	RelSE(JK2)	SE(EMP1)	Bias(SE1)	SE(EMP2)	Bias(SE2)	RMSE1	RMSE2
0.9	1200	1.0026	1.0067	0.0114	0.0126	0.0119	0.0133	0.0107	0.0007	0.0110	0.0008	0.0109	0.0126
	800	1.0054	1.0115	0.0142	0.0158	0.0149	0.0167	0.0134	0.0008	0.0140	0.0008	0.0143	0.0174
	400	1.0061	1.0190	0.0215	0.0241	0.0228	0.0259	0.0179	0.0036	0.0195	0.0034	0.0187	0.0257
0.8	1200	1.0041	1.0082	0.0149	0.0187	0.0154	0.0194	0.0143	0.0006	0.0147	0.0007	0.0147	0.0161
	800	1.0063	1.0145	0.0187	0.0235	0.0194	0.0246	0.0183	0.0004	0.0188	0.0006	0.0190	0.0220
	400	1.0058	1.0232	0.0286	0.0359	0.0299	0.0383	0.0243	0.0042	0.0252	0.0048	0.0248	0.0310
0.7	1200	1.0036	1.0068	0.0171	0.0245	0.0175	0.0252	0.0166	0.0004	0.0170	0.0006	0.0168	0.0176
	800	1.0075	1.0142	0.0212	0.0305	0.0219	0.0317	0.0208	0.0004	0.0215	0.0004	0.0215	0.0236
	400	1.0074	1.0185	0.0322	0.0463	0.0342	0.0497	0.0286	0.0036	0.0309	0.0033	0.0290	0.0334
0.6	1200	1.0070	1.0082	0.0185	0.0310	0.0192	0.0322	0.0175	0.0010	0.0182	0.0010	0.0180	0.0188
	800	1.0139	1.0176	0.0229	0.0387	0.0241	0.0409	0.0202	0.0027	0.0231	0.0011	0.0218	0.0253
	400	0.9751	0.9772	0.0329	0.0535	0.0365	0.0595	0.0265	0.0065	0.0330	0.0035	0.0306	0.0358

Note: * 1 = unweighted estimator, 2 = weighted estimator. RelBias1=$\theta/\hat{\theta}1$, RelBias2=$\theta/\hat{\theta}2$, RelSE(JK1)=SE(JK1)/$\hat{\theta}1$, RelSE(JK2)=SE(JK2)/$\hat{\theta}2$, Bias(SE1)=|SE(JK1)-SE(EMP1)|, Bias(SE2)=|SE(JK2)-SE(EMP2)|.

4.6.2 Jackknife and Balanced Repeated Replication

Monto Carlo simulation is performed to compare the accuracy of jackknife method and balanced repeated replication (BRR) method for variance estimation of our proposed estimator for the scenario of STSCS with unequal size of strata, unequal size of clusters, and ICCs differing from $\rho = 0$ to $\rho = 0.2$.

The set up of the population dataset and the sampling design were almost identical to previous section 4.5.1, except that only two clusters are sampled from each stratum.

From Tables 4.3 and 4.4, we can make similar conclusions to what concluded from Tables 4.1 and 4.2. In addition, we can see that the standard errors and the biases of standard errors from jackknife method are consistently slightly smaller than the standard errors and the biases of standard errors from BRR method. When the sample size increases to $t = 1200$, the biases of standard errors from BRR tend to close to the biases of standard errors from jackknife method. Overall in this scenario, jackknife method for variance estimation performs better than BRR variance estimation, because the jackknife standard errors are closer to empirical standard errors than the BRR method. Note that there are total 8 strata and 2 PSU per stratum are selected so the relative limited degree freedom may increase more variability for BRR method as compared to jackknife method.

Table 4.3: Simulation results for the comparison of jackknife method and BRR method for variance estimates, $T_h \neq T_{h'}$, $T_{hg} \neq T_{hg'}$, and $\rho = 0$

θ	t	RelBias1*	RelBias2*	SE(BRR1)	SE(BRR2)	SE(JK1)	SE(JK2)	SE(EMP1)	Bias(BRR1)	Bias(JK1)	SE(EMP2)	Bias(BRR2)	Bias(JK2)	RMSE1	RMSE2
0.9	1200	0.9998	0.9951	0.0098	0.0111	0.0098	0.0104	0.0097	0.0001	0.0001	0.0099	0.0012	0.0005	0.0097	0.0108
	800	1.0003	0.9923	0.0123	0.0150	0.0123	0.0134	0.0119	0.0004	0.0004	0.0123	0.0026	0.0011	0.0119	0.0142
	400	0.9996	0.9813	0.0188	0.0277	0.0183	0.0219	0.0170	0.0018	0.0012	0.0187	0.0090	0.0032	0.0170	0.0254
0.8	1200	1.0007	0.9960	0.0142	0.0150	0.0142	0.0147	0.0135	0.0007	0.0007	0.0137	0.0013	0.0010	0.0135	0.0141
	800	0.9995	0.9915	0.0173	0.0194	0.0173	0.0183	0.0172	0.0001	0.0001	0.0175	0.0019	0.0008	0.0172	0.0188
	400	1.0010	0.9833	0.0261	0.0334	0.0256	0.0293	0.0251	0.0010	0.0005	0.0269	0.0065	0.0025	0.0251	0.0301
0.7	1200	0.9994	0.9948	0.0166	0.0175	0.0166	0.0172	0.0163	0.0003	0.0003	0.0165	0.0010	0.0007	0.0163	0.0169
	800	0.9999	0.9920	0.0204	0.0225	0.0204	0.0218	0.0203	0.0001	0.0001	0.0207	0.0018	0.0011	0.0203	0.0215
	400	0.9993	0.9829	0.0295	0.0362	0.0285	0.0340	0.0287	0.0007	0.0002	0.0310	0.0052	0.0030	0.0287	0.0333
0.6	1200	0.9962	0.9943	0.0175	0.0187	0.0168	0.0184	0.0167	0.0008	0.0001	0.0179	0.0008	0.0006	0.0169	0.0182
	800	0.9930	0.9927	0.0226	0.0243	0.0207	0.0239	0.0197	0.0030	0.0011	0.0224	0.0019	0.0015	0.0201	0.0229
	400	0.9781	0.9832	0.0365	0.0388	0.0294	0.0366	0.0254	0.0111	0.0040	0.0331	0.0057	0.0035	0.0287	0.0347

Note: * 1 = unweighted estimator, 2 = weighted estimator. RelBias1=$\theta/\hat{\theta}1$, RelBias2=$\theta/\hat{\theta}2$, Bias(BRR1)=|SE(BRR1)-SE(EMP1)|, Bias(BRR2)=|SE(BRR2)-SE(EMP2)|, Bias(JK1)=|SE(JK1)-SE(EMP1)|, Bias(JK2)=|SE(JK2)-SE(EMP2)|..

Table 4.4: Simulation results for the comparison of jackknife method and BRR method for variance estimates, $T_h \neq T_{h'}$, $T_{hg} \neq T_{hg'}$, and $\rho = 0.2$

θ	t	RelBias1*	RelBias2*	SE(BRR1)	SE(BRR2)	SE(JK1)	SE(JK2)	SE(EMP1)	Bias(BRR1)	Bias(JK1)	SE(EMP2)	Bias(BRR2)	Bias(JK2)	RMSE1	RMSE2
0.9	1200	1.0121	1.0069	0.0176	0.0318	0.0131	0.0192	0.0118	0.0057	0.0012	0.0160	0.0158	0.0032	0.0160	0.0171
	800	1.0142	1.0172	0.0187	0.0348	0.0147	0.0211	0.0136	0.0051	0.0011	0.0175	0.0173	0.0036	0.0185	0.0231
	400	1.0151	1.0214	0.0241	0.0481	0.0208	0.0300	0.0189	0.0052	0.0019	0.0245	0.0236	0.0055	0.0231	0.0309
0.8	1200	1.0146	1.0083	0.0206	0.0339	0.0167	0.0221	0.0152	0.0054	0.0014	0.0186	0.0154	0.0035	0.0191	0.0197
	800	1.0154	1.0214	0.0236	0.0388	0.0202	0.0262	0.0180	0.0056	0.0022	0.0216	0.0173	0.0047	0.0217	0.0273
	400	1.0174	1.0225	0.0313	0.0522	0.0284	0.0373	0.0254	0.0059	0.0030	0.0305	0.0218	0.0069	0.0289	0.0352
0.7	1200	1.0122	1.0076	0.0205	0.0293	0.0184	0.0223	0.0172	0.0033	0.0012	0.0194	0.0099	0.0029	0.0191	0.0201
	800	1.0131	1.0164	0.0238	0.0341	0.0219	0.0268	0.0200	0.0037	0.0018	0.0225	0.0116	0.0043	0.0220	0.0252
	400	1.0148	1.0212	0.0343	0.0501	0.0310	0.0405	0.0272	0.0071	0.0037	0.0312	0.0188	0.0093	0.0291	0.0344
0.6	1200	1.0046	0.9964	0.0201	0.0240	0.0178	0.0216	0.0166	0.0035	0.0012	0.0190	0.0051	0.0026	0.0168	0.0191
	800	0.9863	1.0063	0.0251	0.0296	0.0217	0.0272	0.0197	0.0054	0.0020	0.0238	0.0058	0.0034	0.0214	0.0241
	400	0.9822	1.0098	0.0398	0.0467	0.0307	0.0417	0.0251	0.0147	0.0056	0.0351	0.0116	0.0066	0.0274	0.0356

Note: * 1 = unweighted estimator, 2 = weighted estimator. RelBias1=$\theta/\hat{\theta}1$, RelBias2=$\theta/\hat{\theta}2$, Bias(BRR1)=|SE(BRR1)-SE(EMP1)|, Bias(BRR2)=|SE(BRR2)-SE(EMP2)|, Bias(JK1)=|SE(JK1)-SE(EMP1)|, Bias(JK2)=|SE(JK2)-SE(EMP2)|.

4.6.3 Informative Sampling

The sample designs from surveys result in informative sample weights. In this situation the selection probabilities are related to the AUC i.e., related to the level of sensitivity and specificity. The traditional unweighted analyses will ignore the informativeness of the sample weights which can lead to biased estimation of the AUC. The following simulation study is carried out to examine the effect of informative sample weights.

The simulated dataset is similar to the dataset in 4.5.1, but we simplify the STSCS survey data by choosing equal size of strata and equal size clusters. The size of the finite population is $T = 200000$, and the population contains 8 strata, each stratum has 100 clusters. The population AUC (θ) is fixed from 0.9 to 0.6 by 0.1. Again the stratum AUCs (θ_h) are randomly chosen across the strata in the range of $\theta_h = \theta \pm 0.1$. ICCs are from $\rho = 0$ to $\rho = 0.2$. The diseased and nondiseased subjects for each stratum are generated independently and similarly to TSCS in chapter 3.3.1.

Total sample size varies from $t = 400$ to 800, respectively. The sample size for each stratum t_h is depended on the stratum-specific AUCs θ_h .

If the stratum-specific AUCs θ_h are ordered from high to low, we arbitrarily set the selection probabilities (sp) for each stratum differently, sp = c(0.25, 0.25, 0.1, 0.1, 0.1, 0.1, 0.05, 0.05) for the 8 strata. Strata with high AUCs have larger sample sizes. Sampling in each stratum is performed independently by using simple random sampling.

Table 4.5 summarizes the results for informative survey sampling for STSCS with equal strata size and equal cluster size. Table 4.5 shows that the unweighted

estimates of AUC are badly biased, which illustrates the critical role of sample weights in survey data analysis for informative survey sampling. Because of the heavy weights, we can see the weighted standard errors are much larger than the unweighted standard errors.

Table 4.5: Simulation results for informative sampling for STSCS with $T_h = T_{h'}$, $T_{hg} = T_{hg'}$, and ρ = c(0, 0.2)

ρ	θ	t	RelBias1*	RelBias2*	SE(JK1)	RelSE(JK1)	SE(JK2)	RelSE(JK1)	SE(EMP1)	Bias (SE1)	SE(EMP2)	Bias (SE2)	RMSE1	RMSE2
0	0.9	800	0.9647	0.9977	0.0104	0.0111	0.0231	0.0256	0.0080	0.0024	0.0199	0.0032	0.0339	0.0200
		400	0.9641	0.9984	0.0132	0.0142	0.0342	0.0379	0.0109	0.0023	0.0323	0.0019	0.0352	0.0323
	0.8	800	0.9597	0.9982	0.0160	0.0192	0.0259	0.0323	0.0153	0.0007	0.0238	0.0021	0.0369	0.0239
		400	0.9613	0.9979	0.0214	0.0258	0.0408	0.0509	0.0225	0.0011	0.0376	0.0032	0.0393	0.0376
	0.7	800	0.9542	0.9961	0.0191	0.0260	0.0278	0.0395	0.0217	0.0026	0.0304	0.0026	0.0400	0.0305
		400	0.9558	0.9963	0.0265	0.0361	0.0399	0.0568	0.0256	0.0009	0.0351	0.0048	0.0412	0.0352
	0.6	800	0.9457	0.9957	0.0209	0.0329	0.0283	0.0469	0.0230	0.0022	0.0304	0.0022	0.0415	0.0305
		400	0.9492	0.9952	0.0293	0.0464	0.0414	0.0687	0.0318	0.0025	0.0410	0.0005	0.0452	0.0411
0.2	0.9	800	0.9839	1.0063	0.0127	0.0139	0.0245	0.0273	0.0106	0.0022	0.0217	0.0028	0.0181	0.0224
		400	0.9811	1.0073	0.0153	0.0167	0.0356	0.0399	0.0142	0.0011	0.0321	0.0035	0.0224	0.0327
	0.8	800	0.9793	1.0084	0.0170	0.0209	0.0264	0.0332	0.0172	0.0001	0.0243	0.0020	0.0241	0.0252
		400	0.9750	1.0102	0.0236	0.0288	0.0385	0.0486	0.0218	0.0018	0.0347	0.0038	0.0299	0.0357
	0.7	800	0.9805	1.0116	0.0198	0.0278	0.0279	0.0403	0.0181	0.0017	0.0276	0.0002	0.0229	0.0287
		400	0.9720	1.0120	0.0277	0.0384	0.0399	0.0577	0.0292	0.0015	0.0371	0.0028	0.0355	0.0380
	0.6	800	0.9657	1.0110	0.0207	0.0334	0.0285	0.0480	0.0234	0.0026	0.0308	0.0023	0.0316	0.0315
		400	0.9666	1.0189	0.0289	0.0465	0.0407	0.0692	0.0272	0.0016	0.0314	0.0093	0.0343	0.0333

Note: * 1 = unweighted estimator, 2 = weighted estimator. RelBias1=$\theta/\hat{\theta}1$, RelBias2=$\theta/\hat{\theta}2$, Bias(BRR1)=|SE(BRR1)-SE(EMP1)|, Bias(BRR2)=|SE(BRR2)-SE(EMP2)|, Bias(JK1)=|SE(JK1)-SE(EMP1)|, Bias(JK2)=|SE(JK2)-SE(EMP2)|.

Chapter 5

Application

5.1 Hispanic Health and Nutrition Examination Survey

We apply our proposed methods to the data from the Mexican-American portion of the Hispanic Health and Nutrition Examination Survey (HHANES).

The HHANES was conducted by National Center for Health Statistics (NCHS), between 1982 and 1984 to assess the health and nutritional status of Hispanic individuals aged 6 months to 74 years in specific area of the U.S. A stratified four- stage cluster sampling design was used (Gonzalez, 1985; NCHS, 1985):

Stage 1. Primary sampling units, PSUs, refers to counties or small groups of contiguous counties;

Stage 2. Area segments within PSUs, refers to a city block or group of blocks in urban areas, or geographic subareas in rural areas;

Stage 3. Households within area segment;

Stage 4. Individuals within households.

The PSUs and segment stage within the sampled PSUs were stratified prior to selection. For variance estimation the sample design is approximated by 8 pseudo-strata each with 2 pseudo-PSUs within each pseudo-stratum. We treated the pseudo-strata as sampling strata and pseudo-PSUs as sampled PSUs in the data analysis.

5.2 Gold Standard and Proposed Predictors

In this application, we are interested in testing the prediction of overweight/obese status of an individual using different variables related to body weight. The prevalence of excess body weight has been growing globally, with more than 1 billion adults estimated to be either overweight or obese (World Health Organization 2012). The increase overweight or obesity individuals has been observed across all age groups.

Overweight/obesity is commonly determined by using weight and height to calculate a score called the body mass index (BMI) which is equal to body weight in kilograms divided by height in meters squared. BMI is an indirect measure of body fat but considered as a reasonable indicator for body fatness for most people (Flegal and Graubard, 2009). One should be aware of its limitation, particularly when in fact the weight is due to muscle mass and heavier bone structure rather than excess fat, for example for many athletes.

An adult who has a BMI of 25 or higher is considered overweight/obese. overweight/obesity is determined differently for children (aged 2 to 18 years old) by using age- and sex- specific growth charter given by Cole et.al (Cole et al. 2000). Other methods of estimating body fat and body fat distribution include measurements of

skinfold thickness, waist circumference, and calculation of waist-to-hip circumference ratios, etc.

In the HHANES survey, we consider the measured BMI as a "gold" standard to determine the status of overweight/obese. Then self-reported BMI, triceps skinfold and subscapluar skinfold are compared with respect to their capability to predict overweight/obese based on measured BMI.

5.3 Joint Inclusion Probability

As described in section 5.1, the HHANES has a stratified multistage cluster sample design. The probabilities of selection of an individual within a sampled household are given by HHANES, which are depended on age:

1. 75% for from 6 months to 19 years,
2. 50% for 20 to 44 years and
3. 100% for 45 to 74 years

Thus, the putative inclusion conditional probabilities within the sampled family are $\pi_{s|hij} = 0.75$, 0.5, or 1.0 for the sampled subject s from the sampled family j in sampled PSU i within stratum h aged from 2 to 19, 20 to 44 and 45 to 74 years respectively.

After excluding those individuals younger than 2 years old due to undetermined overweight/obese status, there are 2677 sampled families with 7083 sampled individuals for data analysis. The final sample weight for each sampled individual (w_{hijs}) is

given by HHANES, which takes into account several features of the survey: the inclusion probability at each stage of sampling, and nonresponse and poststratification adjustments. Therefore, the sample weight for an individual $hijs$ may be not equal to the inverse of the inclusion probability, i.e.

$$w_{hijs} \neq (\pi_{hij}\pi_{s|hij})^{-1} \tag{5.1}$$

Under the assumption of Possion sampling of households, we approximate the family inclusion probabilities with

$$\pi^*_{hij} = \left(\frac{1}{m_{hij}} \sum_{s=1}^{m_{hij}} w_{hijs}\pi_{s|hij} \right)^{-1} \tag{5.2}$$

where m_{hij} is the size of the sampled household, and the joint inclusion probabilities for sampled families hij and hik

$$\pi^*_{hijk} = \pi^*_{hij}\pi^*_{hik} \tag{5.3}$$

We approximated the sampling of individuals within households by Poisson sampling with inclusion probabilities

$$\pi^*_{s|hij} = \frac{1}{(w_{hijs})\pi^*_{hij}} \tag{5.4}$$

We approximate the within family sampling by Poisson sampling so that given the household has been sampled, the joint inclusion probabilities for sampling individual s and t is

$$\pi^*_{st|hij} = \pi^*_{s|hij}\pi^*_{t|hij} \tag{5.5}$$

The joint sample weight can be obtained from the inverse of the product of joint inclusion probabilities within the cluster and the inclusion probabilities of the sampled clusters, i.e.,

$$w^*_{(hijs,hijt)} = (\pi^*_{st|hij}\pi^*_{hij})^{-1} \tag{5.6}$$

5.4 Results

We compare the prediction capability for three predictors for overweight/obese status. Table 5.1 presents both unweighted estimators and weighted estimators. It shows that after taking into account of sample weights, the self-reported BMI ("BMIsf") has the highest estimated AUC values among three predictors, which implies that the self-reported BMI appears to be the best predictor of overweight/obese, although unweighted AUC ("AUC1") for subscapular skinfold ("SubScap") is slightly higher than that of self-reported BMI. The triceps skinfold ("Triceps") has a lowest prediction power in this application. Table 5.1 also shows that the weighted jackknife standard errors are slightly larger than unweighted jackknife standard errors due to sample weights incorporated into the analysis.

Table 5.1: Estimate of AUC for three predictors of overweight/obese

Predictor	AUC1*	SE(JK1)	AUC2*	SE(JK2)
BMIsf	0.897	4.94E-03	0.911	5.59E-03
SubScap	0.899	3.57E-03	0.893	3.64E-03
Triceps	0.812	5.12E-03	0.789	5.51E-03

* Note: 1=unweighted estimator; 2=weighted estimator.

Overall the results across the unweighted and weighted estimators are very close, which indicates the sample weighting is not very informative in this example. However there is essentially little loss in precision by using the weighted estimation.

Figure 5.1 shows the ROC curve for self-reported BMI, subscapular skinfold and triceps skinfold. The ROC curve for subscapular skinfold lies completely above the curve for triceps skinfold, so that the results support that subscapular skinfold is a better predictor of overweight/obese than triceps skinfold.

Upon examination of the figure, the self-reported BMI is not an absolutely better predictor, although it has a higher estimated AUC as compared to other two predictors. At lower end of the ROC curve, the self-reported BMI is under both curves for subscapular skinfold and triceps skinfold, and in its middle to upper end of the ROC curve it is below the ROC curve for subscapular skinfold. However, we can see from Figure 5.1, self-reported BMI is a better predictor in the range that sensitivity above 50% and simultaneously specificity above 70 %, which is often the range of interest. Figure 5.1 suggests that self-reported BMI is the best predictor for overweight/obesity in this application.

The domain ROC analysis for survey designs can be extended by our proposed approach by adding an indicator to indicate whether or not each subject belongs to the specific domain (see section 4.5). We applied this extended formula to HHANES data to estimate AUC and its variance for domains of female and male who were 20 - 74 years old to see how prediction of overweight/obese status for the three predictors differ by gender. Interestingly, all three predictors of overweight/obese for female domain have better classification than the male domain, for both weighted and unweighted estimators , i.e., AUCs in female domain are larger than these in male domain, and the variances of these three predictors for both weighted and unweighted AUCs in female domain are also smaller (Table 5.2). In both adult male and female

Table 5.2: Estimate of AUC in adult male and female domains

	Male $\geq$ 20 yrs				Female $\geq$ 20 yrs			
Predictor	AUC1*	SE(JK1)	AUC2*	SE(JK2)	AUC1*	SE(JK1)	AUC2*	SE(JK2)
BMIsf	0.9348	6.68E-03	0.9362	6.80E-03	0.9645	4.10E-03	0.9652	4.19E-03
SubScap	0.8748	8.99E-03	0.8749	9.19E-03	0.8911	7.46E-03	0.8908	7.64E-03
Triceps	0.7714	1.18E-02	0.7609	1.27E-02	0.8706	8.31E-03	0.8653	8.53E-03

* Note: 1=unweighted estimator; 2=weighted estimator.

domains, the self-reported BMI is the best predictor of overweight/obese among the three predictors in HHANES.

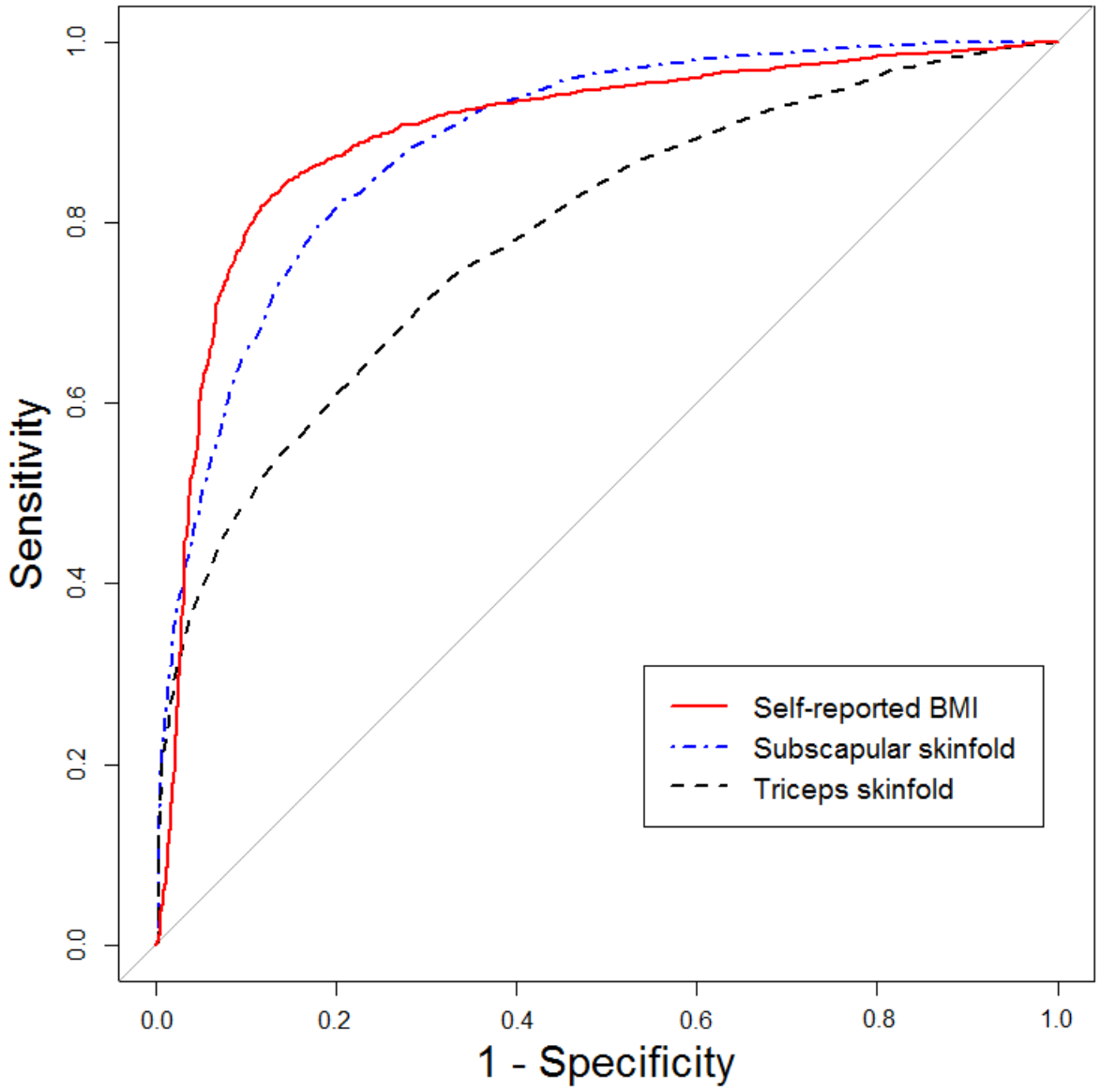

Figure 5.1: ROC curves of Self-reported BMI, Subscapular skinfold and Triceps skinfold

Chapter 6

Discussion

6.1 Summary of the weighted estimator of AUC

In surveys research surveys, complex sample designs with sample weighting and stratification and cluster sampling are commonly implemented. Analyses that do not account for weighting and clustering effects that are induced by the complex survey sampling can be biased. In this dissertation, we propose an extension of the nonparametric method for estimation of the population AUC that accounts for sample weighting in the estimation under different survey designs. We also provide and study the accuracy of a jackknife method, along with balanced repeated replication (BRR) for variance estimation of our proposed estimator of AUC.

The estimators of the population AUC under different survey designs are asymptotic unbiased estimators. From the extended nonparametric formula, we show that the proposed AUC estimator is a ratio estimator where both denominator and numerator are unbiased estimators of their corresponding components. The bias of the estimator decreases with increasing of sample size. Simulation results indicate that our approach provides consistent estimates of AUC and its standard error.

Sample weights play an essential role in survey data analysis. Survey sampling are often complex and with informative sample weighting. When the sample weights are related to the observations of interest, the traditional unweighted analysis can be seriously biased. For the estimation based on paired subjects, such as AUC, the joint sample weights are required, which can be complicated to obtain for complex survey designs. Joint sample weights depend how pairs of subjects are selected in the sample, e.g., for cluster sample designs the joint sample weights can depend on whether the two subjects are from the same sampled cluster or different sampled clusters. While the sample weights for each subject may be provided in most survey datasets, It is unusual for a survey to provide joint sample weights for pairs of subjects. Note that the sample weight for each subject provided by a survey may not only reflect the inclusion probability, but also incorporates oversampling, non-response and post-stratification adjustments. As was indicated in our application to the HHANES data, assumptions have to be made to derive the joint sampling weights from the sample design information given in the survey documentation.

In this dissertation, we applied replication methods for variance estimation. Both jackknife method and BRR method for variance estimation are broadly used in survey data analysis. BRR method was developed typically for sample designs with a large number of sample strata and only two PSUs per stratum (although there are some versions of the BRR method in the literature that allow for more PSUs sampled per stratum). Our simulation results show that the performance of jackknife method is more accurate for finite samples than BRR methods, i.e., as compared to the standard error from BRR variance estimator, the standard error from jackknife variance estimator is closer to the empirical variance estimator.

6.2 Limitation in the use of AUC

It has been widely used for ROC curve analysis in medical science to evaluate the accuracy of diagnostic tests under the gold standard. Some advantages of ROC curve has been discussed in Chapter 1, but it should be used with caution due to some limitations.

First, when comparing multiple diagnostic tests, a comparison only based on AUC value can be misleading since the predictors' curves can cross each other. A decision should be determined by specific thresholds - the interest range of sensitivities and specificities. In this case, one may be more interested in partial AUC (PAUC) instead of global AUC.

Second, it is essential to realize the difference of the role of AUC in classification and calibration. AUC is a measure of the extent of classification of a diagnostic test, i.e., it represents how well the test can separate diseased and nondiseased subjects. In a diagnostic test, the outcome is already determined but unknown to the investigator, and the discrimination is often measured by AUC for comparing with a more expensive or invasive gold standard. Calibration is usually discussed for a model-based assessment for risk prediction, which is a measure of how well predicted probabilities of an event match with actual observed event (Hosmer and Lemeshow 1980). In a prognostic model, the outcome has not yet developed or determined at the time of prediction. A prediction model based on test results and other factors would be considered well calibrated if the number of individuals in a sample or population who eventually develop disease is close to the number expected to develop disease using the predicted risks of disease from the prediction model to calculate the expected numbers. However, the prediction model would be a perfect discriminator as long as the predicted values for diseased subjects are all higher than the predicted

values for nondiseased subjects, regardless of how well the model is calibrated. AUC describes the probability of rank order of diseased and nondiseased subjects, rather than actual predicted probability of the risk of developing the disease of interest. Perfect discrimination does not necessary mean perfect calibration. It is often seen in clinical that a person with high test value in prognostic test does not necessarily become a case.

Third, AUC is a rank method which ignores the distribution of the values of the test results. For two individuals with test values that are relatively close, say 4.9 and 5.1, and assume 5 is the cutoff point, so that they would be separated into diseased and nondiseased groups, but the difference in the values of the tests might not be a meaningful difference with respect to predicting the probabilities of developing disease. In addition, the difference in test results between two individuals who have low risk (say test results of 1.0 vs. 1.2) may have little impact on the AUC as compared to two individuals who have moderate risk versus high risk (say 2.0 vs. 4.5) as long as the differences in rank are similar. However the risk prediction for these two situations can be quite different resulting different treatment decisions in a clinical setting based on the predicted risks.

Pepe (2004b) showed that an odds ratio (OR) associated with a unit increase as large as 3.0 would have little improvement in the classification. However, a novel biomarker that has an OR 3.0 may have a clinically important difference in risk prediction. Recently some new measures (Pencina et.al 2008, Steyerberg et.al 2012) have been introduced that are sensitive for detecting the improvement when a new biomarker is added into a set of predictors, such as net classification improvement (NBI), integrated discrimination improvement (IDI), net benefit (NB), and weighted net classification improvement (wNBI). Those novel measures offer some correction or reclassification among non-cases, and have been demonstrated to be better statistical

tools for biomarker assessment as compared with AUC. However, they are essentially a function of the rank and not the predicted risks. For comparing a predictive model, the measure of both calibration and discrimination, such as the Hosmer-Lemeshow test (Hosmer and Lemeshow 1980) and AUC estimation, should be assessed using prospective cohort studies. However, some basic approaches such as estimating ORs and relative risks are still useful to estimate the change of risk. The clinical usefulness of a prognostic model or a new biomarker in prediction model can depend on both calibration and discrimination, as well as the cost effectiveness of biomarkers.

6.3 Future Research

In this dissertation, we primarily used jackknife methods and balanced repeated replication (BRR) as an alternative method for variance estimation. Both replication methods are conceptually simple and easy to program, but they can be computational intensive for surveys with large numbers of cluster sample units. For example in this dissertation we used HHANES data and considered household as a cluster. Due to a large number clusters in dataset and the computing all of the pairwise inclusion probabilities, it took a long time to complete the data analysis. In future work, we could consider other more computationally efficient approaches for variance estimation, such as Taylor linearization variance estimation. The theory of linearization methods is well developed, but their application to estimating the variance of our nonparametric estimators of the AUC of the ROC curve for various complex sample designs have not been studied.

For stratified sample designs, our proposed approach of estimation of AUC allows pairwise comparison of diseased and nondiseased subjects across different strata. Another possible approach is to restrict the comparisons to being within each stratum

to compute a separate AUC for each stratum, and then combine the stratum-specific AUC estimates across the strata. So this approach is less computationally intensive but there is loss of some information across strata. Future research could compare both approaches for estimating the population AUC based on bias and efficiency.

Further work is needed to extend the estimation of the AUC for the ROC curves derived from a parametric and semi-parametric disease prediction models when the data used to fit models and/or apply the models from complex samples. For example, logistic regression is a commonly used parametric model in discriminant analysis that predicts a subject's disease status from set of covariates that are associated with disease. The strategy is to choose a cutoff value for the predicted probability of being a case and classifying observations as cases if their predicted probability is above the cutoff; otherwise, they are classified as controls. The accuracy of classification can then be assessed by comparing the model-predicted status of an observation with whether it is actually a case or a control. These parametric models may be estimated from data from complex samples as well as the data to estimated the AUC of the resulting the ROC curve. Modifications to take account of the complex sample designs of the data will involve developing appropriate estimation methods for the AUC of the ROC curve and their variances that incorporate the sample weights and the stratification and cluster sampling.

REFERENCES

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

Cai T, Pepe MS. (2002). Semi-parametric ROC analysis to evaluate biomarkers for disease. Journal of the American *Statistical Association*, 97, 1099-1107.

Cochran WG, (1977). *Sampling Techniques*. New York: John Wiley and Sons, Inc.

Cole TJ, Bellizzi MC, Flegal KM, Dietz WH. (2000). Establishing a Standard Definition for Child Overweight and ObesityWorldwide: International Survey.*British Medical Journal*, 320, 12401243. 405

Davis WW, Flegal KM. (2003). Comparing the Areas Under Receiver Operating Characteristic Curves for Cluster Designs. *Joint Statistical Meetings*: 1182-1189.

DeLong ER, DeLong DM, Clarke-Pearson DL. 1988.Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: a Nonparametric Approach. *Biometrics* 44 (3) (September): 837-45.

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

Fay RE. (1989).Theory and Application of Replicate Weighting for Variance Calculation. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 212-217.

Flegal KM, Graubard BI. (2009). Estimates of excess deaths associated with body mass index and other anthropometric variables. *The American Journal of Clinical Nutrition*, 89: 12139.

Gonzalez JF, Ezzati T, Lago J, Waksberg J. (1985). Estimation in the Southwest Components of the Hispanic Health and Nutrition Examination Survey. *Proceedings of the Section on Survey Research Methods*, Alexandria, VA: American Statistical Association, pp. 233237. 405, 407

Graubard BI, Korn EL. (1996) Modelling the sampling design in the analysis of health surveys. *Statistical Methods in Medicine Research*, 5: 263-281.

Hanley JA, McNeil BJ. (1982).The Meaning and Use of the Area Under a Receiver Operating (ROC) Curvel Characteristic. *Radiology* 143: 29-36.

Hosmer DW, Lemeshow S. (1980). A goodness-of-fit test for the multiple logistic regression model. *Communications in Statistics*. A10:1043-1069.

Kondratovich M. (2003). Matched Receiver Operating Characteristics (ROC) ANALYSIS AND PROPENSITY SCORES. *Proceedings of the 2000 Joint Statistical Meeting, Biopharmaceutical Section*

Korn EL, Graubard BI. (1999). *Analysis of Health Surveys*. New York: John Wiley and Sons, Inc.

Korn EL, Graubard BI. (2003).Estimating Variance Components by Using Survey Data. *Journal of the Royal Statistical Society*: 65 (1): 175-190.

Levy PS, Lemeshow S. (1991). *Sampling of Populations: Methods and Applications*. New York: John Wiley and Sons.

McCarthy PJ. (1966). Replication: An Approach to the Analysis of Data from Complex Surveys. *Vital and Health Statistics*, ser. 2, no. 14, Washington D.C.: National Center for Health Statistics

McCarthy PJ. (1969). Pseudo-replication: Half samples. *Review of the International Statistical Institute*. 37 (3), 239-264

McNeil, B. J., and J. A. Hanley. (1984). Statistical Approaches to the Analysis of Receiver Operating Characteristic (ROC) Curves. *Medical Decision Making* 4 (2) (January 1): 137-150.

Metz CE. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, 8:283.

Metz CE. (1998). Herman BA, Shen J-H. Maximum-likelihood estimation of ROC curves from continuously-distributed data. *Statistics in Medicine*, 17:1033.

National Center for Health Statistics (1985).Plan and Operation of the Hispanic Health and Nutrition Examination Survey 19821984. *Vital and Health Statistics,* 1(19). 405

Obuchowski NA. (1997). Nonparametric Analysis of Clustered ROC Curve Data. *Biometrics* 53 (2) (June): 567-78.

Pencina MJ, D'Agostino Sr RB, D'Agostino Jr RB, Vasan RS. (2008). Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in Medicine*, 27:157-72.

Pepe MS (2003). *The Statistical Evaluation of Medical Tetsts for Classification and Prediction*, Oxford: Oxford University Press.

Pepe MS, Cai T. (2004a). The Analysis of Placement Values for Evaluating Discriminatory Measures. *Biometrics* 60 (2) (June): 528-35.

Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. (2004b). Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *American Journal of Epidemiology* 159:882- 890.

Plackett RL, Burman JP. (1946) The Design of Optimum Multifactorial Experiments. *Biometrika* 33 (4), 305-25

Rao JNK, Scott AJ. (1992). A simple method for the analysis of clustered binary data. *Biometrics*, 15, 385-397.

Sen PK. (1960). On some convergence properties of U-statistics. *Calcutta Statistical Association Bulletin*, 10, 1-18.

Steyerberg EW, Pencina MJ, Lingsma HF, Kattan MW, Vickers AJ, Van Calster B. 2012. Assessing the incremental value of diagnostic and prognostic markers: a review and illustration. *European Journal of Clinical Investigation*, 42(2):216-28

Swets JA, Pickett RM. (1982). *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*. Academic Press, New York.

Wieand S, Gail MH, James BR, James KL. (1989). A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika*, 76:585.

Wolter KM. (2003). *Induction to Variance Estimation*. Springer.

World Health Organization, Obesity and Overweight (2012). *http://www.who.int/mediacentre/factshe*

Zhou XH. (1996). A Nonparametric Maximum Likelihood Estimator for the Receiver Operating Characteristic Curve Area in the Presence of Verification Bias. *Biometrics* 52: 299-305.

Zweig MH, Campbell G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, 39:561.