# TEST-DEPENDENT SAMPLING DESIGN AND SEMI-PARAMETRIC INFERENCE FOR THE ROC CURVE

Bethany Jablonski Horton

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics.

Chapel Hill
2014

Approved by:

Dr. Haibo Zhou

Dr. Amy Herring

Dr. Gary Koch

Dr. Alison Stuebe

Dr. Xiaofei Wang

# ABSTRACT

**BETHANY JABLONSKI HORTON: TEST-DEPENDENT SAMPLING DESIGN AND SEMI-PARAMETRIC INFERENCE FOR THE ROC CURVE (Under the direction of Dr. Haibo Zhou)**

The receiver operating characteristic (ROC) curve and area under the ROC curve (AUC) are used to describe the ability of a screening test to discriminate between diseased and non-diseased subjects. As evaluating the true disease status can be costly, researchers can increase study efficiency by allowing selection probabilities to depend on the screening test. We consider a test dependent sampling (TDS) design where TDS inclusion depends on a continuous screening test measure. Disease status is validated only for subjects in the SRS and TDS components. To improve efficiency, this sampling design incorporates three components: the simple random sample (SRS) component, TDS component, and the un-sampled subjects.

We propose semi-parametric empirical likelihood estimators for the AUC, partial AUC, and the covariate-specific ROC curve. First, the AUC estimator allows us to summarize the ability of the screening test to distinguish between diseased and non-diseased subjects. Empirical likelihood methods are used to avoid making distributional assumptions for the screening test variable. Second, the AUC estimator is adapted to estimate partial AUC when a subset of false positive rates is more clinically relevant. Third, the covariate-specific ROC curve is estimated using a binormal model for the screening test variable. Although parametric assumptions are made for the screening test, distributional assumptions are avoided for the covariates by using empirical likelihood methods. This ROC curve estimator allows us to assess the influence covariates have on the accuracy of the diagnostic test.

This cost-effective sampling design allows for a more powerful study on the same budget. Efficiency is gained in all three estimators by incorporating information from both the sampled and un-sampled portions of the population.

## ACKNOWLEDGMENTS

I would like to thank my dissertation advisor, Dr. Haibo Zhou, for his insight and encouragement throughout this process. His guidance, knowledge, and motivation have been key in all of my work. Dr. Xiaofei Wang has also been very instrumental in the development of this research and I would like to thank him for his ideas and time. I would like to thank Dr. Amy Herring for her mentoring and counsel as my academic advisor. I would also like to thank my remaining dissertation committee members Dr. Gary Koch and Dr. Alison Stuebe for their time, encouragement, and collaboration during my time at UNC.

I could not have completed this work without the unwavering support of my family. My greatest thanks goes to my husband, Josh. Without his support, I would not be where I am today. I would also like to thank my parents, Tod and Linda Jablonski, who have always encouraged me to work hard for my dreams and have had continuous confidence in me during the ups and downs of graduate school. I would like to thank my brothers, Michael and Thad, and their families who have always been there to encourage me. Also, I thank my in-laws Ed and Dianne Horton, and Rebecca and her family for their enthusiasm and encouragement. All of my family has offered nothing but unconditional love and support, which has been invaluable during this process.

I would also like to thank Melissa Hobgood, Annie Green Howard, Alison Wise, Elizabeth Koehler, Andrea Byrnes, Elena Bordonali, and Jennifer Clark. My time at UNC would not have been the same without them.

This work was supported by NIH grants R01ES021900 and T32ES007018.

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

## LIST OF ABBREVIATIONS

| | |
|---|---|
| AUC | Area under the ROC curve |
| CL | Cervical length |
| FFN | Fetal Fibronectin |
| FPR | False positive rate |
| LS-ROC | Least squares ROC curve estimator |
| MW-AUC | Mann-Whitney AUC estimator |
| NP-pAUC | Nonparametric partial AUC estimator |
| NPEL-AUC | Nonparametric empirical likelihood AUC estimator |
| NPEL-pAUC | Nonparametric empirical likelihood partial AUC estimator |
| NSCLC | Non-small cell lung cancer |
| ODS | Outcome dependent sample |
| pAUC | Partial area under the ROC curve |
| PPS | Preterm Prediction Study |
| PTB | Preterm birth |
| ROC | Receiver operating characteristic |
| SE | Standard error |
| SPEL-AUC | Semi-parametric empirical likelihood AUC estimator |
| SPEL-pAUC | Semi-parametric empirical likelihood pAUC estimator |
| SPEL-ROC | Semi-parametric empirical likelihood covariate-specific ROC curve estimator |
| SRS | Simple random sampling |
| TDS | Test-dependent sampling |
| TPR | True positive rate |

# Chapter 1

# Literature Review

## 1.1 Introduction and motivation

Using statistical tools to discriminate between different populations is beneficial in a wide variety of areas. One such tool is the receiver operating characteristic (ROC) curve, which was developed for electronic signal detection (Hanley, 1989). The diagnostic methods have been expanded to be useful in a wide variety of medical applications: from medical imaging techniques (Swets, 1979) and studying risk markers for cardiovascular disease (Yeboah et al., 2012) to using prostate-specific antigen to detect prostate cancer (Dodd and Pepe, 2003b) and applying time-dependent accuracy summaries in the setting of survival analysis models (Heagerty and Zheng, 2005). There is also a wide variety of statistical methods proposed in this area: from new summary measurements to methods of dealing with missing data in this diagnostic setting and changing the way in which subjects are sampled into the study.

Receiver operating characteristic (ROC) curves and area under the ROC curve (AUC) are summary measures used to describe the ability of a screening test to discriminate between diseased and non-diseased subjects (Bamber, 1975). As evaluating the true disease status can be costly, it is beneficial for researchers to increase study efficiency by allowing selection probabilities to depend on the screening test (Wang et al., 2012). Increased efficiency translates to cost and time savings for studies as well as decreased burden on subjects.

Consider screening for non-small-cell lung carcinoma (NSCLC) cancer recurrence. Lung cancer is the most common cause of cancer death among men and women in the world (Blanchon et al., 2006). Lung cancer is classified as either small-cell lung carcinoma (SCLC)

or non-small-cell lung carcinoma (NSCLC), of which NSCLC accounts for approximately 80% of all lung cancers. After surgical lung resection, a large proportion of stage 1 NSCLC patients have cancer recurrence within five years (Bueno et al., 2012). When surgery is used as the primary treatment for NSCLC, adjuvant chemotherapy may benefit patients who have a high risk of cancer recurrence. Identifying patients who are at high risk of cancer recurrence is important in order for treatment to be given to those who would benefit most. This is an important area of study for patients, families, and doctors when making decisions on a treatment plan.

We used data from the data from the CALGB 150807 study conducted by the Cancer and Leukemia Group B (Bueno et al., 2012). This study is a subset of patients registered in the CALGB 140202 study who have stage 1A or 1B non-small-cell lung cancer (NSCLC). Among patients in the CALGB 150807 study, 1,061 patients were not censored before 12 months and were used in this analysis. The Balcone risk score, outlined by Blanchon et al. (2006), has been developed to identify patients who are at greatest risk of cancer recurrence. The risk score is developed by considering factors such as age, gender, activity level at diagnosis, histological type, and the tumor-node-metastasis staging system.

There are many interesting questions that can be explored in this study. AUC can be used to investigate the ability of the Balcone risk score to predict cancer recurrence. Given the need for an accurate test, partial AUC (pAUC) can be used to evaluate the performance of the Balcone risk score where a specific range of FPRs or TPRs is considered. Because large FPRs are less clinically relevant, we can restrict the range of interest to FPR$\in (0, 0.3)$, for example. With the wealth of patient information available, a covariate-specific ROC would allow us to evaluate the performance of the Balcone risk score, while accounting for covariates that appear to be associated with cancer recurrence. This covariate-specific ROC estimator can then be used to identify subsets of the population where the screening test is better at distinguishing between subjects who have cancer recurrence and those who do not.

There have been many methods studied and proposed in the area of ROC and AUC analysis. Bamber (1975) proposed a nonparametric AUC estimator, which is equivalent to the Mann-Whitney U-statistic (Pepe, 2004). Parametric AUC estimators were developed by

Swets and Pickett (1982) and Hanley et al. (1983). These methods use SRS for selection into the study. Wang et al. (2012) proposed a nonparametric AUC estimator which utilizes a biased sampling design in order to target subjects who contribute more information to the study. McClish (1989) and Thompson and Zucchini (1989) introduced the idea of evaluating only part of the AUC when a subset of FPRs are of interest. The pAUC can be interpreted as the joint probability that $Y_D > Y_{\bar{D}}$ and $Y_{\bar{D}}$ fall within the FPR range of interest (Dodd and Pepe, 2003a). Estimators similar to those given above for the AUC have been proposed for the pAUC. A nonparametric pAUC estimator that uses SRS data was proposed by Dodd and Pepe (2003a) and a nonparametric pAUC estimator using a biased sampling scheme was proposed by Wang et al. (2012). Another discriminatory measure used to differentiate between two populations when data are available over time is the C statistic (Rizopoulos, 2011; Pencina et al., 2012b; Heagerty and Zheng, 2005; Antolini et al., 2005). The C statistic is a weighted average of the AUCs across multiple time points in the study. Heagerty and Zheng (2005) suggested that the time specific AUCs can be plotted over time to assess changes in accuracy across time for a time to event outcome.

Another important area of research the use of covariates in modeling ROC and in estimating AUC and pAUC. The use of covariates allows us to better understand the influence covariates have on accuracy of the screening test (Wang et al., 2013). Thompson and Zucchini (1989) proposed nonparametric direct estimation of the AUC for specific level of a categorical covariate. Wang et al. (2013) proposed ROC estimation which uses a biased sampling design and a binormal model for screening test variable. Dodd and Pepe (2003a,b) proposed using a generalized linear model framework for modeling the screening test, which can be used to estimate AUC and pAUC.

The proposed work focuses on estimation of AUC, pAUC, and a covariate-specific ROC curve using biased sampling methods and all available information, including incomplete information available for un-sampled subjects. We consider a test dependent sampling (TDS) design where TDS inclusion is dependent on a continuous screening test measure. Here, the Balcone risk score is the measure used for the biased sampling scheme. This biased sampling design incorporates a simple random sample (SRS), the TDS component, and the

un-sampled subjects as opposed to a design using only a simple random sample of the same size. The idea behind the supplemental sample is to target resources where the greatest amount of information can be attained (Zhou et al., 2002). Cancer recurrence and other covariate information are known only for those included in the SRS and TDS components. The screening test measure and other baseline information are available for all subjects in the study. Information from un-sampled subjects will also be utilized in the proposed methods. Using the biased sampling design as well as incorporating observed information from the un-sampled subjects can lead to efficiency improvements. This suggests that a smaller sample size can be used with these methods, compared to existing methods, where a larger sample size would be necessary to obtain the same level of efficiency in the estimator. A smaller sample size translates to cost savings for the study and decreased subject burden. Using this biased sampling design, we propose multiple approaches to studying these data that answer different questions, which are helpful in understanding the utility of the screening test.

## 1.2 ROC Curves and Area under the ROC curve (AUC)

### 1.2.1 Unadjusted methods

There are many ways to approach the use of data in the area of medical decision making. Methods have been proposed for a variety of types of estimators. Greenhouse and Mantel (1950) suggested that to be considered an acceptable test, a screening test should be able to correctly classify at least a pre-specified percentage of the diseased subjects and incorrectly classify no more than a set percentage who are well. Other common measures include area under the ROC curve (AUC) and partial AUC, where a particular interval of FPRs or TPRs are of interest. A three dimensional extension of ROC and AUC was proposed by Skaltsa et al. (2012), where instead of a two level outcome (diseased or not diseased) the outcome can have more levels. This is beneficial for studying diseases such as Alzheimer's disease. In this case, the disease naturally presents with a transition state that falls between normal aging and irreversible Alzheimer's disease, which can be described as mild cognitive impairment. Yu et al. (2011), Liu and Zhou (2011), and Long et al. (2011b,a) considered methods to account

for a missing outcome or diagnostic screening variable without eliminating those subjects from the analysis, limiting bias and loss in efficiency. Wang et al. (2012, 2013) suggested an alternate approach to sampling subjects in order to target those who contribute more information. With this biased sampling scheme, a smaller sample size can be used to attain estimates that are as good as or better than alternative sampling methods, such as SRS.

*ROC and AUC*

The receiver operating characteristic (ROC) curve is a tool used to display how well a screening test, $Y$, is able to indicate disease status, $D$. The ROC curve is constructed by plotting the false positive rate (FPR, $Pr\left(Y \geq c|D = 0\right)$) versus the true positive rate (TPR, $Pr\left(Y \geq c|D = 1\right)$), where c is the threshold for the screening test to indicate disease. The area under the ROC curve (AUC) is a summary measure used to determine both the importance of a difference between two populations and also describes the accuracy of discrimination performance (Bamber, 1975). Figure 1.1 shows an ROC curve with corresponding AUC. The FPR and TPR range from 0 to 1, and the AUC ranges from 0.5 to 1. An ROC curve with intercept 1 and slope 0 indicates a perfect screening test that correctly identifies disease status in every subject. An ROC curve with intercept 0 and slope 1, creating a $45^o$ line, indicates a screening test that is essentially as good as flipping a coin. A screening test with an ROC curve that falls above the $45^o$ line indicates some level of ability of the screening test to discriminate between diseased and non-diseased subjects.

Another summary measure is the partial AUC (pAUC), shown in Figure 1.2. The pAUC restricts the FPR (or TPR) to a range that is more clinically relevant. McClish (1989) and Thompson and Zucchini (1989) introduced the idea of evaluating only part of the AUC for certain FPR intervals that are of interest. The pAUC can be interpreted as the joint probability that $Y_D > Y_{\bar{D}}$ and $Y_{\bar{D}}$ fall within the FPR range of interest (Dodd and Pepe, 2003a). There are downsides that must be considered when using the pAUC. The standard error of the pAUC estimator increases and there is a loss in precision when a major restriction is made on FPR (Walter, 2005).

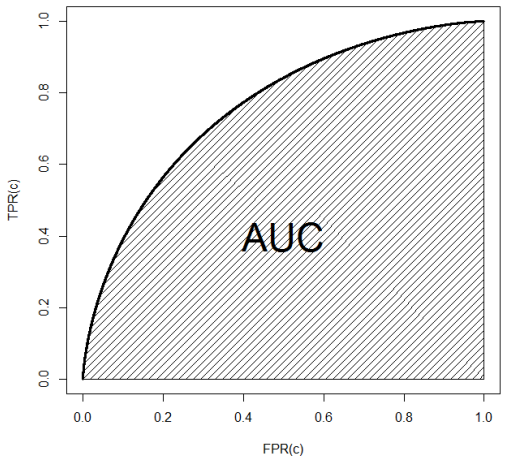


Figure 1.1: Area under the ROC curve

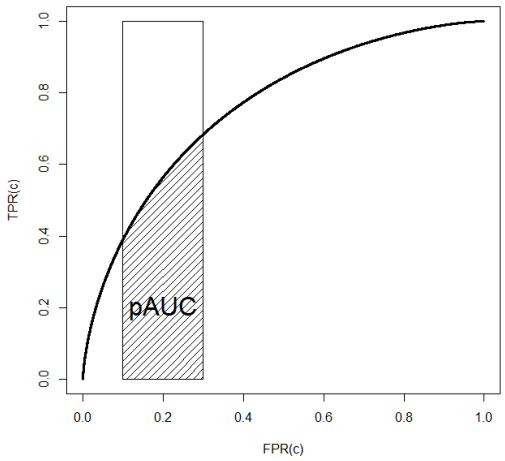


Figure 1.2: Partial area under the ROC curve

*AUC and pAUC estimators*

A binormal model for estimating AUC was proposed by Swets and Pickett (1982) and Hanley et al. (1983). They compared the binormal model for estimating AUC to the nonparametric AUC estimator. Swets and Pickett (1982) suggested that the method assuming a binormal model for the screening test variable is superior to the nonparametric estimator because with the binormal model, the estimator is less affected by location and spread of points that define the ROC. The area under the empirical ROC curve is given by $\hat{A}^{SRS} = \frac{1}{n_D n_{\bar{D}}} \sum_{j=1}^{n_{\bar{D}}} \sum_{i=1}^{n_D} \left[ I\left(Y_{Di} > Y_{\bar{D}j}\right) + \frac{1}{2} I\left(Y_{Di} = Y_{\bar{D}j}\right) \right]$, which is the Mann-Whitney U-statistic (Bamber, 1975). Both of these approaches to estimating the AUC use data that are sampled from the population with SRS.

Dodd and Pepe (2003a) extended this AUC estimator in the SRS setting for pAUC. The proposed pAUC estimator restricts the FPR (or TPR) and is given by

$$\hat{A}_t^{SRS} = \frac{1}{n_D n_{\bar{D}}} \sum_{i=1}^{n} \sum_{j=1}^{n} D_i \left(1 - D_j\right) I\left(Y_i > Y_j, Y_j \in (q_0, q_1)\right)$$

where $q_0 = FPR^{-1}(t_1)$ and $q_1 = FPR^{-1}(t_0)$. This estimator is nonparametric and shows great improvements compared to other estimators, such as being more robust while losing

only moderate efficiency compared to parametric estimators. For the FPR component of the estimator, they found that using the estimated quantiles instead of the true quantiles gave improved efficiency for estimating pAUC (Dodd and Pepe, 2003a).

An empirical likelihood method for estimating AUC was proposed by Qin and Zhou (2006). This estimator showed improved small sample properties compared to assuming a normal approximation. A confidence interval for the AUC was also developed. The empirical likelihood methods made it possible to obtain estimates for parameters without specifying a distribution for the screening test. To obtain confidence intervals, they showed that their proposed AUC estimator followed a scaled chi-square distribution, giving asymptotically correct coverage probability. Although these methods were derived for a SRS, the methods can be extended to account for a stratified sampling design. McNeil et al. (1984) developed methods when a fixed FPR or TPR are of interest. These methods assumed normality of the screening test variable.

*TDS methods*

Wang et al. (2012) proposed estimators for both AUC and pAUC that improve efficiency by using a biased sampling design. These estimators are nonparametric and show improvement over the simple random sampling setting when using the standard AUC estimator and the pAUC estimator proposed by Bamber (1975) and Dodd and Pepe (2003a), respectively. The form of these estimators is similar to that of the SRS estimators, but weights are incorporated that account for the biased sampling design. The AUC and pAUC estimators proposed by Wang et al. (2012) are given by:

$$\begin{aligned}
\hat{A}^{TDS} &= \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j D_i (1-D_i) I(Y_i > Y_j)}{\sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j D_i (1-D_i)} \\
\hat{A}_t{}^{TDS} &= \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j D_i (1-D_i) I\left(Y_i > Y_j, Y_j \in F\hat{P}R\right)}{\sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j D_i (1-D_i)}
\end{aligned}$$

where the false positive rate is estimated by $F\hat{P}R_j = \frac{\sum_i \hat{p}_i(1-D_i)I(Y_i>Y_j)}{\sum_i \hat{p}_i(1-D_i)}$. In Wang et al. (2012), the TDS methods described for AUC and pAUC were used in evaluating the survival benefit of celecoxib, a COX-2 inhibitor, for patients with positive COX-2 expression. COX-2

is a protein that is over-expressed with lung cancer. Its intensity ranges from 0 to 10, and it stratified into three groups to obtain the TDS portion of the sample: negative (COX-2 $< 2$), moderate ($2 \leq$ COX-2 $< 4$), and positive (COX-2 $\geq 4$). Preliminary data showed that the proportions of patients falling into these categories were approximately 60%, 13%, and 27%, respectively. In order to study the relationship between COX-2 value and survival, a range of COX-2 values needs to be seen. Because treating and tracking outcomes for subjects is costly, a sample is usually taken in order to complete the study on a fixed budget. The TDS method for sampling was implemented in order to select enough subjects with moderate and positive COX-2 to study this relationship. Define $D = 1$ as patients who survive less than 6 years and $D = 0$ otherwise. Targeting a small range for the FPR can be important as false positive results add increased cost and burden on subjects. With this in mind, the FPR interval of interest was $(0, 0.1)$. More details in the biased sampling component for this estimator are given later in the Outcome-Dependent-Sampling portion of the literature review.

### 1.2.2 Covariate adjusted methods

Methods have been developed which consider the effect of covariate information on ROC curves. This can be accomplished in many ways, such as estimating the covariate effect on the screening test, directly estimating the AUC, and directly estimating the covariate specific ROC curve. Tosteson and Begg (1988) proposed modeling the effect of covariates on the screening test, $Y$. Here, a distribution function was assigned for $Y$, and the resulting covariate effect on the ROC curve was calculated. There are limitations here, as model misspecification can lead to erroneous results. Thompson and Zucchini (1989) and Dodd and Pepe (2003a) proposed directly estimating the AUC, and Dodd and Pepe (2003a) proposed directly estimating pAUC, while accounting for covariates. Methods for directly estimating the survival function or the ROC curve were proposed by Pepe (1997, 2000), Cai and Pepe (2002), and Wang et al. (2013). Generalized linear modeling methods were used in Pepe (1997, 2000). These results were extended to a semi-parametric approach by Cai and Pepe (2002).

Thompson and Zucchini (1989) proposed that estimation can be completed by specifying a distribution function for $Y$ or can be completed nonparametrically using the Wilcoxon statistic: $\hat{AUC}_k = n_{D,k}^{-1} n_{\bar{D},k}^{-1} \sum_i^{n_{D,k}} \sum_i^{n_{\bar{D},k}} \{I[Y_j < Y_i] + 0.5I[Yi = Y_j]\}$, where $k = 1, \ldots, K$ denotes the covariate level. Thompson and Zucchini (1989) also proposed an analysis of variance (ANOVA) approach for modeling to compare the means of an accuracy index for different combinations of variables. In this setting, images are read by multiple people and these results are compared to see how ratings compare between readers. The model is given by $Y_{ijk} = \mu + \alpha_i + b_j + (ab)_{ij} + c + e_{ijk}$, where $\mu + \alpha_i$ represents the mean level of Y for the i-th combination of the variables. The variable $b_j$ is a random variable, allowing for variation between image readers. Zheng and Heagerty (2007) proposed a semi-parametric estimate of the survival function of the screening test over time. The ROC is constructed from this estimated survival function, and AUC can be assessed over time. The added component of following a subject's screening test variable over time allows for the ability to assess diagnostic accuracy at different intervals of time between measurement and diagnosis.

Methods to estimate the AUC and pAUC while adjusting for covariates provide useful model interpretations for both discrete and continuous covariates (Dodd and Pepe, 2003a,b). These methods are semi-parametric and take advantage of generalized linear model framework. Dodd and Pepe (2003b) define the covariate specific AUC as $Pr\left(Y_i^D > Y_j^{\bar{D}} | X_i^D, X_j^{\bar{D}}\right) = \theta_{ij}$. The regression model is given by $g(\theta_{ij}) = \mathbf{X}_{ij}^T \boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is a vector of parameters and $g$ is a monotone increasing link function. The proposed estimating function is given by $S_N(\beta) = \sum_i^{n_D} \sum_j^{n_{\bar{D}}} \frac{\partial \theta_{ij}}{\partial \boldsymbol{\beta}} v(\theta_{ij})^{-1} (U_{ij} - \theta_{ij}) \equiv \sum_i^{n_D} \sum_j^{n_{\bar{D}}} S_{ij}(\boldsymbol{\beta})$. Dodd and Pepe (2003a) propose the covariate-specific pAUC given by

$$AUC_X(t_0, t_1) = Pr\left(Y^D > Y^{\bar{D}}, Y^{\bar{D}} \in (q_0, q_1) | X\right).$$

The general model is given by $AUC_X(t_0, t_1) = g\left(X^T \beta\right)$ for a specified link function $g$. For the pAUC setting the estimating equation is given by

$$V_{n_D, n_{\bar{D}}}(\beta) = \sum_i^{n_D} \sum_j^{n_{\bar{D}}} \frac{\partial \theta_X}{\partial \beta} v(\theta_X)^{-1} \left(V_{ij}^{(q_0, q_1)} - \theta_X\right) = 0$$

where $V_{ij}^{(q_0,q_1)} = I\left(Y_i^D > Y_j^{\bar{D}}, Y_j^{\bar{D}} \in (q_1, q_0)\right)$. When using the logit link, exponentiated model parameters can be interpreted as AUC or pAUC odds. For a binary covariate, the exponentiated model parameter can be interpreted as the ratio of AUC or pAUC odds between the two levels of that covariate. For a continuous covariate, the exponentiated model parameter can be used to describe how AUC or pAUC changes for diseased and not diseased subjects as that covariate changes. Dodd and Pepe (2003b) used their proposed AUC methods to study the ability of the distortion product otoacoustic emission (DPOAE) device in assessing impaired hearing. The DPOAE device is used at three different frequencies and three intensity settings, creating nine combinations of settings. The severity of hearing loss is also of interest in this setting. A behavioral test where subjects indicate the point at which a sound is audible is the gold standard in assessing hearing loss. The model used here is given by $log\left(\frac{AUC}{1-AUC}\right) = \beta_0 + \beta_1 intensity + \beta_2 frequency + \beta_3 severity$. Results from this analysis showed that DPOAE is able to discriminate between severely impaired ears and normal ears better than mildly impaired and normal ears, which is not surprising. Also, stimuli with lower intensities achieved greater accuracy. Dodd and Pepe (2003a) considered the ability of prostate-specific antigen (PSA) to diagnose prostate cancer. The data came from the $\alpha$-Tocopheraol and $\beta$-Carotene Study (ATBC). Serum samples were collected and stored at baseline and three years later. Adjusting for time was important here, especially because the time from measurement to diagnosis varied greatly and it was expected that PSA levels taken close to the time of diagnosis would be more predictive. Clinical evidence showed a relationship between PSA levels and prostate cancer. Two methods of quantifying PSA were considered, total PSA and the ratio of free to total PSA. The comparison of these two methods was incorporated into the model. Ultimately, 240 subjects in the study were diagnosed with prostate cancer during the eight year study follow-up period. Serum samples were age matched for 237 non-prostate diagnosed subjects who were sampled for comparison. They considered FPR values in $(0, 0.4)$. The model was given by:

$$log\left[\frac{AUC\,(0,0.4)}{0.4 - AUC\,(0,0.4)}\right] = \beta_0 + \beta_1 \text{test} + \beta_2 \text{time} + \beta_3 \text{test} * \text{time}$$

The results showed that PSA accuracy improved when subjects were measured at times closer to the time of diagnosis. Total PSA appeared to be a better diagnostic tool for prostate cancer than the ratio.

Pepe (1997) proposed a regression design that directly modeled covariate effects on the ROC curve. Denote $D$ the binary indicator of disease status, $Y$ the non binary diagnostic test, $Z$ the factors that potentially influence test accuracy, and $X$ the vector of covariates. The ROC curve associated with $Z$ for a logistic model is given by $ROC_Z(t) = \frac{\exp\{\alpha_0(t)+\boldsymbol{X\beta}\}}{1+\exp\{\alpha_0(t)+\boldsymbol{X\beta}\}}$, where $\alpha_0(t)$ is a monotone function from $(0,1)$ to $(-\infty,\infty)$, and $t$ denotes the false positive rate. No distributional assumptions are made for $Y$; assumptions are made only for the relationship between diseased and non diseased subjects through the ROC curve model. This approach allows for examining the influence covariates have on the accuracy of a diagnostic test in discriminating disease status. This method was applied to radiology data, the same used in Thompson and Zucchini (1989), where images were constructed and then evaluated by three readers. Here, there were 50 each of diseased and non diseased images and the readers classified their evaluation of each image with an ordinal scale from 1 to 5. After data collection, the $4^{th}$ and $5^{th}$ categories were collapsed due to sparse data. A logistic type regression was fit to the data with the model $ROC_Z(t) = \frac{\exp\{\alpha_0(t)+\beta_1X_1+\beta_2X_2+\beta_3X_3\}}{1+\exp\{\alpha_0(t)+\beta_1X_1+\beta_2X_2+\beta_3X_3\}}$, where $X_i$ corresponds to the evaluation made by the $i^{th}$ reader. This technique allowed for comparisons between readers, such as reader 3 rating images systematically lower than the other two readers.

An ROC curve estimator was proposed by Wang et al. (2013), which uses test dependent sampling (TDS), a biased sampling design. With this method, portions of the population are oversampled to gain efficiency. A binormal model is assumed for the screening test, $Y$, such that $Y = \beta_0 + \beta_D D + \beta_X^T X + \beta_{DX}^T DX_D + \sigma(D)\epsilon$, where $\epsilon \sim \mathcal{N}(0,1)$ and $\sigma(D) = \sigma_1 I[D=1] + \sigma_0[D=0]$. No distributional assumptions are needed for the covariates due to the use of empirical likelihood methods. Estimates of the survival function were estimated and combined for a covariate-specific ROC curve, given by: $ROC_X(t) = S_{1X}\left(S_{1X}^{-1}(t)\right) = \Phi\left(\frac{\beta_D+\beta_{DX}^T X_D+\sigma_0\Phi^{-1}(t)}{\sigma_1}\right)$. This method was applied to data from a study evaluating the prognostic value of COX-2 for survival of patients with lung cancer, the same used in Wang

et al. (2012). Covariates such as age and gender were included in the model. These covariates not only have a relationship with cancer survival, but also on the COX-2 expression, so ability to include these covariates in the study is valuable. Simulation studies showed that the estimator will perform fairly well under mild misspecification of the binormal model.

*Other summary measures*

Another discriminatory measure used to differentiate between two populations is the C statistic (Rizopoulos, 2011; Pencina et al., 2012b; Heagerty and Zheng, 2005; Antolini et al., 2005). In situations such as survival analysis settings, time dependent ROC and AUC can be used to evaluate the screening test over time. Rizopoulos (2011) described the C statistic as a summary of the screening test variable over the study period. This weighted average of the AUCs is given by $C = \frac{\int AUC_t Pr(T_i^* > t) dt}{\int Pr(T_i^* > t) dt}$, where $Pr\left(T_i^* > t\right)$ is the marginal survival probability. The marginal survival probability takes into account censoring, since all time points will not contribute equally. Heagerty and Zheng (2005) suggested that the time specific AUCs can be plotted over time to assess a change in accuracy across time. While the methods proposed by Heagerty and Zheng (2005) and Rizopoulos (2011) are semiparametric, the methods proposed by Antolini et al. (2005) are non-parametric making them more robust.

*Missing data*

Instead of focusing of new types of summary measures for discrimination between populations, Yu et al. (2011), Liu and Zhou (2011), and Long et al. (2011b,a) considered the common issue of missing data. Long et al. (2011b,a) developed methods for missing screening test values. Loss of efficiency and, depending on the type of missingness, bias may be introduced when only including subjects with complete data. Both missing at random and missing not at random scenarios are considered. This AUC estimator was shown to work well under model misspecification. Nonparametric imputation procedures were used in developing methods to analyze ROC when the biomarker for screening was missing. In this case, other auxiliary variables were present and were used in imputing the main biomarker of interest. Instead of the absence of the screening test variable, Yu et al. (2011) and Liu and Zhou (2011) developed methods where the gold standard (verification of disease status) was missing. Yu et al. (2011)

combined multiple continuous tests together as a composite test to discriminate between two populations. It was assumed that the test values are binormal, and a Bayesian latent disease model was used, along with an MCMC algorithm for computation. Glomerular filtration rate (GFR) is a measurement associated with the ability of the kidneys to filter. A reduced GFR is a marker for chronic kidney disease. Chronic kidney disease is defined as kidney damage or GFR measurement below a set threshold. There are different ways to estimate the GFR and no true gold standard in defining chronic kidney disease. Methods proposed by Yu et al. (2011) were used to assess the optimal way of measuring GFR to diagnose chronic kidney disease. Liu and Zhou (2011) developed a semiparametric ROC curve estimator where the gold standard is missing for a subset of the study population. The missing at random assumption was assumed in the development of these estimators. Weighted estimating equations were used to account for the missing gold standard for a subset of the subjects.

*Optimal threshold for Y*

Another interesting and important research topic in the area of ROC and AUC is choosing the best threshold value of the screening test to indicate disease. Different approaches can be considered in finding the optimal threshold for the screening test and in assessing improvement in the screening tests available. Molanes-López and Letón (2011) used empirical likelihood methods to assess the most appropriate cut-off value for the diagnostic test using the Youden index. The nonparametric empirical likelihood methods were compared to a newly developed parametric methods. Simulation studies showed that the nonparametric method was competitive with other parametric methods and was superior. Pencina et al. (2012a) evaluated the improvement in population discrimination using AUC. These methods assume multivariate normality and use a linear discriminant analysis. The measures under study reduce to a function of Mahalanobis distance, which helps to describe the magnitude of improvement in estimation. Let $M^2_{p+q} = \delta^T \Sigma^{-1} \delta$ denote the Mahalanobis distance for $p+q$ cases. The first of the three estimators proposed was assessing the change in AUC, $\Delta AUC$. This estimator reduces to $\Delta AUC = \Phi\left(\sqrt{\frac{M^2_{p+q}}{2}}\right) - \Phi\left(\sqrt{\frac{M^2_p}{2}}\right)$.

To evaluate competing events, Zheng et al. (2012) proposed a method that evaluates

the predictive accuracy of a marker for each type of event. Huang et al. (2011) proposed a nonparametric procedure that optimizes the linear combination of diagnostic tests to maximize the AUC. Combining these diagnostics provides a combined score that can used to estimate the AUC estimator. Four methods were considered in estimating AUC, including cross-validation, bootstrap, sigmoid function smoothing, and approximated cross-validation for variable selection. As these methods can be very computationally intensive, the cross-validation methods are strongly suggested, in an effort to reduce computational cost.

ROC and AUC are discriminatory measures that can be used when the outcome has two levels, such as diseased versus not diseased. When the outcome has more than two levels, other approaches need to be considered. Skaltsa et al. (2012) developed methods to assess the optimum threshold for this diagnostic setting. Consider Alzheimer's disease. This disease naturally presents with a transition state that can be described as mild cognitive impairment, which falls between normal aging and irreversible Alzheimer's disease. A three-dimensional classification plot is constructed, which is similar to the ROC curve. Volume under the surface of this three-dimensional plot gives a measure of accuracy that is similar to AUC. Different weights can be used in the estimator that involve the cost of evaluation to aid in finding the optimum threshold. Disease prevalence and classification cost are incorporated here, both of which need to be considered when finding the cut-off for the optimal test.

## 1.3 Outcome dependent sampling

### 1.3.1 Methods for binary and discrete outcomes

Many study designs exist to help assess the relationship between disease and exposure. Prospective studies are one such example, but these studies tend to be time consuming and expensive. When a rare disease or event is of interest, the study population would have to be quite large to observe enough subjects with disease in order to assess a relationship. Time is also an important factor in this type of study design, especially if time from exposure to observing the event is large. Retrospective studies are another way of studying disease and exposure relationships.

The case-control sampling design is a very beneficial design when exploring relationships between covariates and a dichotomous disease (or outcome) status. Logistic regression is a tool used to explore the relationship between the dichotomous disease outcome and multiple variables of interest. The logistic model is expressed as $log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_j x_j = \mathbf{x}\boldsymbol{\beta}$ where $p = Pr(y = 1|\mathbf{x})$ (Prentice and Pyke, 1979). This tool extends the ability to study the relationship between disease and exposure status to include other covariates of interest. In situations where subjects are selected using biased sampling methods, the sampling design must be accounted for, and basic logistic regression cannot be used.

When there is interest in studying a rare disease or exposure, taking a random sample of the population may not provide a sufficient number of diseased or exposed subjects to obtain model estimates and understanding of the rare disease or exposure of interest. White (1982) suggested a two-stage approach where disease and exposure status are found for a large sample in the first stage, which can be time consuming and costly. This design offers improvement over a one stage design because some groups would contribute small cells in the stratum-specific table. In the second stage, covariate information is found for only a stratified subsample of the first stage subjects. Sub-sampling in the second stage is accomplished by separating subjects into four groups based on their disease and exposure status. Rare diseases or exposures will have more representation in the study by using this approach. Ascertaining exposure status and covariates can be expensive and also invasive for subjects, which can be problematic since exposure is established in the first stage. Similar to the two-stage case control design described above is the case-cohort design. Prentice (1986) suggested a two stage design where the disease status is identified in the first stage. One or more subjects without disease are then matched to a diseased subject, and a random sample is then selected from the entire cohort, or study population, in the second stage. From here, covariate information is ascertained for the selected cohort and case subjects.

*Multi-level outcome and exposure*

Discrete choice analysis was discussed in Manski and McFadden (1981). Instead of the binary classification for outcome and exposure, the classification is generalized so that there

can be many options of outcome and exposure indication. This structure is beneficial in situations where instead of looking at non-diseased versus diseased, it may be more informative to consider non diseased and multiple levels of disease severity. Hsieh et al. (1985) and Scott and Wild (1986) developed a conditional maximum likelihood method for choice-based sampling. Estimators for response probabilities (the probability of illness given exposure and other covariates) were proposed. Breslow and Cain (1988) combined the two-stage sampling framework with the conditional maximum likelihood developed by Hsieh et al. (1985) for choice-based sampling. This modified logistic regression method adjusts for potential biased caused by oversampling certain groups in the second stage. Fears and Brown (1986) used a maximum likelihood estimation method that included stratum specific terms for the case-control setting. This work was improved upon by Scott and Wild (1991) where the method was made more computationally reasonable.

### 1.3.2 Methods for continuous outcomes

Zhou et al. (2002) proposed a semiparametric empirical likelihood method using outcome dependent sampling design for continuous outcomes. The information available for sampled subjects is a continuous outcomes variable, $Y$, and a vector of covariates, $X$. For the continuous outcome $Y$, consider splitting the variable into $K$ mutually exclusive intervals, $C_k = (a_{k-1}, a_k]$, $k = 1, ..., K$, where $a_0 = -\infty < a_1 < a_2 < \cdots < a_K = \infty$. The sampling components consist of a SRS portion and an ODS portion. For the ODS portion, sampling is targeted within the $K$ intervals. The available data is given by $\{Y_{ki}, X_{ki}\}$, where $k$ indexes the sampling group ($k = 0$ indicates the SRS sampling component) and $i$ indexes the subject within the $k^{th}$ sampling group. The sample size is $n = n_0 + n_1 + \cdots + n_K$. Denote the density of $Y$ given $X$ as $f_{Y|X}(y|x, \beta)$ where $\beta$ is the regression coefficient of interest. The cumulative distribution of $X$ is given by $G_X$ and the density function is given by $g_X$. The likelihood of

the observed data is given by

$$\begin{aligned} L(\beta, G_X) &= \left\{ \prod_{i=0}^{n_0} f_\beta(y_{0i}|x_{0i})\, g_X(x_{0i}) \right\} \times \left[ \prod_{k=1}^{K} \prod_{j=0}^{n_k} f_\beta(y_{kj}, x_{kj}|y_{kj} \in C_k) \right] \\ &= \left\{ \prod f_\beta(y_{0i}|x_{0i}) \times \prod_{k=1}^{K} \prod_{j=0}^{n_k} \frac{f_\beta(y_{kj}|x_{ki})}{F(a_k|x_{kj}) - F(a_{k-1}|x_{kj})} \right\} \\ &\quad \times \left\{ \prod g_X(x_{0i}) \times \prod_{k=1}^{K} \prod_{j=0}^{n_k} \frac{F(a_k|x_{kj}) - F(a_{k-1}|x_{kj})}{F(a_k) - F(a_{k-1})} \right\} \\ &= L_1(\beta) \times L_2(\beta, G_X). \end{aligned}$$

Without loss of generality, let $K = 3$. In order to get an estimate for $\beta$ they first estimate $G_X$. The distribution of $X$ is not defined. By fixing $\beta$, the empirical likelihood function of $G_X$ can be found over all observed values of $X$. Denote $\pi_1 = F(a_1)$, $\pi_3 = \bar{F}(a_2)$, and $p_i = g_X(w_i)$ where $(w_1, \ldots, w_n) = (x_{01}, \ldots, x_{0n_0}, x_{11}, \ldots, x_{1n_1}, x_{31}, \ldots, x_{3n_3})$. The portion of the likelihood that involves $g_X(w_i)$ is given by $L_2(\beta, G_X) = L_2(\beta, \{p_i\}) \propto \prod_{i=1}^{n} p_i \pi_1^{-n_1} \pi_3^{-n_3}$. To find $\{\hat{p}_i\}$ that maximizes $L_2$, they considered the following constraints:

$$\left\{ p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i \{F(a_1|w_i) - \pi_1\} = 0, \sum_{i=1}^{n} p_i \left\{\bar{F}(a_2|w_i) - \pi_3\right\} = 0 \right\}$$

From Qin and Lawless (1994), a unique maximum exists for $\{p_i\}$ under the above constraints if 0 is inside the convex hull of points $\{F(a_1|w_1) - \pi_1\}, \ldots, \{F(a_1|w_n) - \pi_1\}$ and $\left\{\bar{F}(a_2|w_1) - \pi_3\right\}, \ldots, \left\{\bar{F}(a_2|w_n) - \pi_3\right\}$. The maximum over $\{p_i\}$ can be found by incorporating Lagrange multipliers:

$$\begin{aligned} H &= \log L_2(\beta, \{p_i\}) + \rho \left( 1 - \sum_{i=1}^{n} p_i \right) + n\lambda_1 \sum_{i=1}^{n} p_i \{F(a_1|w_i) - \pi_1\} \\ &\quad + n\lambda_3 \sum_{i=1}^{n} p_i \left\{\bar{F}(a_2|w_i) - \pi_3\right\} \end{aligned}$$

From here, they find that $\rho = n$ and $\hat{p}_i = \frac{1}{n}\left[1 + \lambda_1 \{F(a_1|w_i) - \pi_1\} + \lambda_3 \left\{\bar{F}(a_2|w_i) - \pi_3\right\}\right]^{-1}$.

Plugging this estimate $\{\hat{p}_i\}$ into the original likelihood function gives an empirical log likelihood. An iterative procedure such as the Newton-Raphson algorithm can be used to obtain estimates for $\beta$. This method proposed by Zhou et al. (2002) gives a method of estimation that is not only efficient, but no distributional assumptions need to be made for the covariates of interest.

Zhou et al. (2007) proposed the Horvitz-Thompson approach with inverse-probability weights for the sampling design described in Zhou et al. (2002). The parameter $\beta$ is estimated without specifying $G(X)$, where $\beta$ is the vector of regression coefficients that links the exposure and outcome, $X$ and $Y$, respectively. This approach requires knowledge of the sampling probabilities, unlike in Zhou et al. (2002). If all $N$ subjects are observed, the log-likelihood is given by $\sum_{i=1}^{N} \log P(y_i|x_i;\beta)$. The log-likelihood is estimated by weighting the observed subjects with the inverse of their second-stage selection probability. The inverse probability weighted estimator, $\hat{\beta}_{IPW}$, is found by solving $\frac{1}{N}\sum_k \sum_{i\in C_k} \frac{1}{p_k} \frac{\frac{\partial}{\partial\beta} P_\beta(y_i|x_i)}{P_\beta(y_i|x_i)} = 0$, where $\hat{p}_k = \frac{n_k}{N_k}$. Simulation results show that for evaluating the linear relationship between a continuous exposure and continuous outcome, using the ODS technique described is more efficient than a simple random sample. This method reduces the number of subjects needed to obtain the same level of accuracy compared to a SRS design. This weighted method is only available if the weights are known or can be estimated, which can be difficult in some circumstances. Gains in efficiency are found when the continuous outcome, $Y$, is stratified into a large number of groups, defined by $C_k$.

*Inclusion of non-validation data*

Chatterjee et al. (2003) and Weaver and Zhou (2005) suggested methods for an outcome dependent sampling design that allowed for the utilization of the un-sampled portion of the data. Consider the sampling design detailed by Zhou et al. (2002). Let $Y$ be a continuous outcome variable of interest that is partitioned into $K$ mutually exclusive intervals. The $k^{th}$ stratum is given by $C_k = (a_{k-1}, a_k]$ where $a_0 = -\infty < a_1 < a_2 < \cdots < a_K = \infty$ and $k = 1, ..., K$. Let the SRS and ODS components be referred to as the validation set, indexed by $V$. The un-sampled portion of the population is referred to as the non-validation set,

indexed by $\bar{V}$. The outcome variable must be known for all subjects in the population in order to use the ODS design. However, the covariates of interest, $X$, are only ascertained for the validation set. A sampling indicator is created to distinguish between the validation and non-validation portions where $R_i = 1$ if $X_i$ is observed and $R_i = 0$ if $X_i$ is not observed. The likelihood for the validation set where the variable information is complete is given by

$$L_V(\theta, G_X) = \left[\prod_{i \in V} f(Y_i|X_i;\theta)\right]\left[\prod_{i \in V} dG_X(X_i)\right]\left[\prod_{k=1}^{K} \pi_k(\theta, G_X)^{-n_k}\right] \tag{1.1}$$

where $\pi_k(\theta, G_X) = \int P_k(x;\theta)\, dG_X(x)$ and $P_k(x;\theta)\, dG_X(x) = \int_{C_k} f(y|x;\theta)\, dy$. Here $P_k(x;\theta)$ and $\pi_k(\theta, G_X)$ are the conditional and marginal probabilities that $Y$ is in the $k^{th}$ stratum. Consider the use of all validation and non-validation subjects. The stratum sizes for the non-validation set are calculated by taking the total number of subjects in $C_k$ and subtracting the number of validation subjects whose outcome variable fall within $C_k$, given by $n_{\bar{V},k} = N_k - n_{0,k} - n_k$. The stratum size follows a multinomial law where $Pr\left(\left\{n_{\bar{V},k}\right\}\right) = \frac{(N-n_0)!}{\prod_{k=1}^{K}(N_k - n_{0,k})!}\prod_{k=1}^{K}\left[\pi_k(\theta, G_X)\right]^{N_k - n_{0,k}}$. Non-validation contribution to the likelihood is given by

$$\prod_{k=1}^{K}\prod_{j \in \bar{V}_k} \frac{f_Y(Y_j;\theta)}{\pi_k(\theta, G_X)}, \tag{1.2}$$

where $f_Y(Y_j;\theta) = \int f(Y_j|u;\theta)\, dG_X(u)$. The full likelihood is found by combining (1.1) and (1.2), and reduces to

$$L_F(\theta, G_X) \propto \left[\prod_{i \in V} f(Y_i|X_i;\theta)\right] \times \left[\prod_{i \in V} dG_X(X_i)\right] \times \left[\prod_{j \in \bar{V}} f(Y_i;\theta)\right]. \tag{1.3}$$

The maximum estimated likelihood estimator (MELE) developed by (Weaver and Zhou, 2005) is similar to the pseudoscore estimator (PSE) proposed by Chatterjee et al. (2003). The semiparametric estimator proposed by Chatterjee et al. (2003) relaxes the assumption that all subjects have a positive probability of selection. By not requiring all subjects to have a positive selection probability, it is possible to create a sampling framework for a

case or control only design. Although much of the results are given for a discrete outcome, exploring these methods for a continuous outcome is possible and would require nonparametric regression methods. These two methods (MELE and PSE) were compared and found to be more efficient than a semiparametric maximum likelihood estimator. The MELE is found by replacing the unspecified marginal distribution function (such as $G_X$ in (1.3) above) with a consistent estimator. The resulting likelihood incorporates the observed selection probability $p_k^{'} = \frac{(n_{0,k}+n_k)}{N_k}$. To use the PSE estimator, function $q_\theta(X_i) = \sum_{l=1}^{K} p_l^{'} P_l(X_i;\theta)$ is substituted for the observed selection probability ($p_k^{'}$).

*Restricted maximum likelihood*

Song et al. (2009) proposed an estimation method using the ODS design described in Weaver and Zhou (2005). Empirical likelihood methods similar to those used in Zhou et al. (2002) and Weaver and Zhou (2005) were incorporated in developing the likelihood, which gives

$$\hat{g}_i = \left\{ n - \sum_{j \in \bar{V}} \frac{f(Y_j|X_i;\theta)}{\sum_{k \in V} \hat{g}_k f(Y_j|X_j;\theta)} \right\}^{-1}, \tag{1.4}$$

where $g(\cdot)$ is the probability density function for $X$. It is noted that the number of constraints increases as the sample size increases, and the mixed Newton method for estimation is suggested with the following steps:

1. Begin with initial estimates $\theta^0$ and $g_i^0$, $i \in V$.

2. Insert $\theta^0$ and $g_i^0$ in to the right hand side of the score equations given above (1.4) and solve the equations iteratively using the fixed-point algorithm until it converges, calling the solution $g_i^c$.

3. Take $g_i^c$ from the second step and plug into the likelihood to maximize the parametric likelihood using Newton's method to update $\theta^c$.

4. Repeat the second and third steps until the proposed convergence criteria is met.

Efficiency gains seen in simulation studies depends heavily on the proportion of subjects sampled in the tails for the validation set.

*OADS*

Outcome and auxiliary dependent sampling is a biased method of sampling where a subject's probability of selection depends on both the outcome and an auxiliary variable. A semiparametric empirical likelihood method was proposed by Wang et al. (2009). These methods were applied to a lung cancer biomarker study where it was seen that subjects with epidurmal growth factor receptor (EGFR) mutations responded to EGFR inhibitor drugs differently than those without the mutation. Because of the expense in testing for the EGFR mutation, the predicted probability of a subject having the EGFR mutation was used as the auxiliary variable for the ODS design. Logistic regression was used to obtain the predicted probability with the model incorporating patient record information, including variables known to be associated with EGFR mutation. The supplementary sample was made up of two groups: those who responded to the inhibitor ($Y$, outcome) and those who did not respond to the inhibitor but had predicted probability above a set threshold ($W$, auxiliary). The likelihood was developed using a generalized linear model with known link function. Misspecification of the distribution of the covariates, $G(\boldsymbol{X}|w = k)$, will lead to inconsistent results. Because of this, empirical likelihood methods were used to estimate the distribution function of $\boldsymbol{X}$, similar to the approach described for Zhou et al. (2002). Simulation studies suggest that the OADS design shows gains in efficiency when there is a moderate to high correlation between the biomarker and the auxiliary variable. For rare disease, this method improves efficiency compared to SRS.

*ROC and AUC*

Wang et al. (2012) used the ODS framework to estimate area under the receiver operating characteristic curve. Consider a disease outcome $D$ and a screening test $Y$, where it is of interest to measure how well this screening test is able to indicate a subject's disease status. In this situation, the biased sampling design is completed using the screening test, which is known, instead of the disease status outcome, which is unknown. For this reason, instead

of an outcome dependent sample, a test-dependent sampling design is discussed. Similar to Zhou et al. (2002), the empirical likelihood method was used to obtain an estimate for $p_i = f(Y_i, D_i)$. An ROC curve estimator was proposed by Wang et al. (2013) that incorporates other covariate information. A binormal model is assumed for the screening test, $Y$. No distributional assumptions are needed for the covariates due to the use of empirical likelihood methods.

## 1.4 Proposed research

### 1.4.1 AUC using test-dependent sampling

Motivated by the need to improve efficiency, a semi-parametric AUC estimator is proposed that incorporates the test dependent sampling design and the of inclusion of information from the un-sampled portion of the population. The TDS sampling design has three components: the SRS component, TDS component, and non-validation set (un-sampled subjects). The subjects sampled in the SRS and TDS components combined make up the validation set, indexed by $V$, where true disease status is validated. The remainder of the population not selected for sampling makes up the non-validation set, indexed by $\bar{V}$. The disease status, $D$, is only ascertained for subjects who are in the validation set. These components are defined by

$$\begin{array}{ccc}
\text{SRS component} & (D_{0j}, Y_{0j}) & j = 1, ..., n_0 \\
\text{TDS}_1 \text{ component} & (D_{1j}, Y_{1j} | Y_{1j} \in C_1) & j = 1, ..., n_1 \\
\vdots & \vdots & \vdots \\
\text{TDS}_K \text{ component} & (D_{Kj}, Y_{Kj} | Y_{1j} \in C_K) & j = 1, ..., n_K \\
\text{Non-validation component} & (Y_{\bar{v}j} | i \neq (0, 1, \ldots, K)) & j = 1, ..., n_{\bar{V}}
\end{array}$$

where $n_{\bar{v}} = N - n_0 - \sum_{k=1}^{K} n_k$. The portion of the likelihood for the sampled subjects is given by $L_V$ and the portion of the likelihood for the un-sampled portion of the subjects is given

by $L_{\bar{V}}$. These likelihood components are given below.

$$\begin{aligned} L_V(f_D) &= \prod_{k=1}^{K}\prod_{j=1}^{n_k} Pr(Y_{kj} \in C_k)^{-n_k} \times \prod_{k=0}^{K} f(Y_{ij}|D_{kj})\, Pr(D_{kj} = d) \\ L_{\bar{V}}(f_D) &= \frac{(N-n_0)!}{\prod_{k=1}^{K}(N_k - n_{0,k})!} \times \prod_{j=1}^{n_{\bar{V}}}\sum_{d=0}^{1} f\left(Y_{\bar{V}j}|D_{\bar{V}j} = d\right) Pr\left(D_{\bar{V},j} = d\right) \\ &\quad \times \prod_{k=1}^{K} Pr(Y_{kj} \in C_k)^{-n_k}. \end{aligned} \tag{1.5}$$

The full likelihood is given by $L(f_D) = L_V(f_D) \times L_{\bar{V}}(f_D)$ where the distribution of the screening test conditional on disease status, $f(Y_{ij}|D_{ij} = d)$, is not specified. Empirical likelihood inference is used to estimate the distribution of the screening test, $Y$, conditional on disease status. This is desirable as model misspecification can introduce problems for full or semi parametric estimators. The Newton-Raphson algorithm is used to obtain estimates for model parameters, which are then used to estimate the expected disease status. The estimated expected disease status is necessary in this design because true disease status is missing for subjects in the non-validation component. The proposed AUC estimator is given by

$$\hat{A}^{P}_{V,\bar{V}} = \frac{\sum_{l\neq l'}^{N} D^*_l\left(1 - D^*_{l'}\right) I(Y_l > Y_{l'})}{\sum_{l\neq l'}^{N} D^*_l\left(1 - D^*_{l'}\right)},$$

$$\text{where } D^*_l = \begin{cases} D_l & \text{if } l \in V \\ \widehat{E(D_l)} = \frac{\hat{p}}{\hat{p} + e^{\hat{\alpha}+\hat{\beta}y_l}(1-\hat{p})} & \text{if } l \in \bar{V} \end{cases}. \tag{1.6}$$

Simulation studies show that the proposed AUC estimator improves efficiency over other current methods, including a SRS only estimator proposed by Bamber (1975) and a method proposed by Wang et al. (2012) that utilizes the TDS design without incorporating information from the un-sampled portion of the population.

### 1.4.2 Partial AUC using test-dependent sampling

It may be more clinically relevant to evaluate the utility of a screening test for a subset of false positive rates. Using the likelihood development in (1.5), then we use empirical likelihood methods to estimate pAUC without making assumptions of the distribution of the screening test variable. The proposed pAUC estimator is given by

$$\hat{A}^{P}_{t:V,\bar{V}} = \frac{\sum_{l\neq l'}^{N} D_l^* \left(1 - D_{l'}^*\right) I\left(Y_l > Y_{l'}, \widehat{FPR}_{l'} \in (t_0, t_1)\right)}{\sum_{l\neq l'}^{N} D_l^* \left(1 - D_{l'}^*\right)},$$

$$\text{where } \widehat{FPR}_{l'} = \frac{\sum_{l}^{N} \left(1 - D_l^*\right) I\left(Y_l > Y_{l'}\right)}{\sum_{l}^{N} \left(1 - D_l^*\right)}$$

$$\text{and } D_l^* = \begin{cases} D_l & \text{if } l \in V \\ \widehat{E\left(D_l\right)} = \frac{\hat{p}}{\hat{p} + e^{\hat{\alpha} + \hat{\beta} y_l}\left(1 - \hat{p}\right)} & \text{if } l \in \bar{V} \end{cases}. \tag{1.7}$$

Simulation studies show that efficiency is improved for the proposed pAUC estimator compared to other current methods, including a SRS only estimator proposed by Dodd and Pepe (2003a) and a method proposed by Wang et al. (2012) that utilizes the TDS design without including information from the un-sampled portion of the population.

### 1.4.3 Covariate-specific ROC curve estimation using test-dependent sampling

In evaluating a ability of the screening test to assign disease or outcome status correctly, it may be beneficial to also evaluate covariate effects on diagnostic accuracy. We propose a semi-parametric covariate-specific ROC curve estimator which incorporates a test-dependent sampling design and inclusion of un-sampled subjects. A binormal model is used to describe the relationship between the screening test, the disease status, and covariates. The sampling components are the same as described for the AUC estimator. The screening test model is given by

$$Y = \mathbf{X}\boldsymbol{\beta} + \sigma\left(D\right)\epsilon = \beta_0 + \beta_1 D + \beta_2 X + \beta_3 X D + \sigma\left(D\right)\epsilon \tag{1.8}$$

where $\epsilon \sim \mathcal{N}(0,1)$ and $\sigma(D) = \sigma_1 D + \sigma_0(1-D)$. The likelihood is given by

$$\begin{aligned} L(G,p,\boldsymbol{\beta},\sigma) \quad \propto \quad & \prod_{i=0}^{3}\prod_{j=1}^{n_i} f(Y_{ij}|X_{ij},D_{ij})\, g(X_{ij}|D_{ij}=d)\, h(D_{ij}) \\ & \times \prod_{j=1}^{n_{\bar{V}}}\sum_{d=0}^{1} f\left(Y_{\bar{V}j}|X_{\bar{V}j},D_{\bar{V}j}\right) g\left(X_{\bar{V}j}|D_{\bar{V}j}\right) Pr\left(D_{\bar{V},j}=d\right) dX. \end{aligned} \quad (1.9)$$

Although the screening test distribution is specified, empirical likelihood methods are used in maximizing the likelihood without specifying the distribution function for the covariates.

## 1.5 Outline of dissertation

In Chapter 2, notation and data structure are defined and AUC estimation is explored. We develop a semi-parametric AUC estimator by using empirical likelihood methods. Simulation results and analysis of data from a lung cancer study and the Preterm Prediction Study show that the proposed AUC estimator is unbiased and more efficient than other AUC estimators compared.

In Chapter 3, partial AUC is considered. Similar to the AUC estimator proposed in Chapter 2, a semi-parametric empirical likelihood estimator for the pAUC is proposed. Simulation results and analysis of data from a lung cancer study and the Preterm Prediction Study show that the proposed pAUC estimator is unbiased and more efficient than other pAUC estimators compared.

In Chapter 4, the covariate-specific ROC curve estimator is proposed. This semi-parametric estimator uses empirical likelihood methods to estimate the ROC curve without making assumptions on the distribution of the covariates. Simulation studies show gains in efficiency compared to current ROC curve estimators due to the TDS design and inclusion of un-sampled subjects. Analysis of data for a lung cancer study and the Preterm Prediction Study show the utility of this ROC curve estimator by showing that the screening test is more effective for some covariate values than others.

# Chapter 2

# AUC under Test-Dependent Sampling

## 2.1 Introduction

The receiver operating characteristic (ROC) curve and area under the ROC curve (AUC) are summary measures used to describe the ability of a screening test to discriminate between diseased and non-diseased subjects (Bamber, 1975). As evaluating the true disease status can be costly, it is important for researchers to increase study efficiency by allowing selection probabilities to depend on the screening test (Wang et al., 2012, 2013). Increased efficiency translates to cost and time savings for studies as well as decreased burden on subjects. We propose a semi-parametric AUC estimator which incorporates a test-dependent sampling design and inclusion of un-sampled subjects. Simulation studies show that the proposed AUC estimator is unbiased and improves efficiency compared to estimators using a simple random sample (SRS) design and those that use only information from the sampled subjects.

Our research is motivated by the study of non-small-cell lung cancer where we identify patients at high risk of cancer recurrence in order to adjust treatment plans to be most effective. Using data from the Cancer and Leukemia Group B (CALGB) 150807 study, we evaluate the ability of the Balcone risk score to identify patients who are at greatest risk of non-small-cell lung cancer (NSCLC) recurrence by estimating the AUC. The ROC curve is a tool used to graphically display the ability of the Balcone risk score to identify patients who survive beyond 12 months by plotting the false positive rate (FPR, $Pr\left(Y \geq c|D=0\right)$) against the true positive rate (TPR, $Pr\left(Y \geq c|D=1\right)$). The AUC is a summary measure used to describe the ability of the Balcone risk score to discriminate between patients who

survive beyond 12 months and those who do not. When surgery is used as the primary treatment for NSCLC, adjuvant chemotherapy may benefit patients who have a high risk of cancer recurrence. Identifying patients who are at high risk of cancer recurrence is important in order for treatment to be given to those who would benefit most. This is an important area of study for patients, families, and doctors when making decisions on a treatment plan. The proposed methods are especially beneficial considering the length of time this type of study will follow patients and cost of following a large number of subjects in this setting.

We also consider data from the Preterm Prediction Study in evaluating the utility of the proposed AUC estimator (Goldenberg et al., 1996). Preterm birth (PTB), defined as delivery at less than 37 weeks of gestation, contributes to neonatal morbidity and mortality. This is an important area of study due to the negative impact of spontaneous PTB on maternal and child health outcomes. Knowing the fetal fibronectin (FFN) measurement will not change the incidence of spontaneous PTB, but it can affect the treatment plan. We evaluate the ability of FFN to predict spontaneous PTB by estimating the AUC.

Previous research has explored multiple approaches to estimating AUC. A non-parametric AUC estimator proposed by Bamber (1975) is equivalent to the Mann-Whitney U-statistic (Pepe, 2004). Although the Bamber (1975) estimator is unbiased, the normal approximation-based Mann-Whitney confidence interval has low coverage accuracy for high values of the AUC when sample sizes for diseased and non-diseased subjects are small and unequal (Qin and Zhou, 2006). A binormal model for estimating AUC was proposed by Swets and Pickett (1982) and Hanley et al. (1983) where the estimator is less affected by the location and spread of points that define the ROC. For both of these approaches, subjects are selected by SRS from the population. An empirical likelihood method was proposed by Qin and Zhou (2006) to estimate AUC, making it possible to obtain estimates for parameters without specifying a distribution for the screening test. Qin and Zhou (2006) showed that their AUC estimator followed a scaled chi-square distribution, giving asymptotically correct coverage probability and a reliable alternative approach for constructing the confidence interval of the AUC. Although these methods were derived for SRS, the methods can be extended to account for a stratified sampling scheme. Under a test-dependent sampling (TDS) design, Wang et al.

(2012) proposed a non-parametric estimator for the AUC. This sampling design includes SRS and TDS components and improves efficiency over SRS-only designs. The TDS design is related to outcome-dependent sampling (ODS). Zhou et al. (2002) used the ODS design and empirical likelihood methods in regression modeling to develop parameter estimates where inclusion in the sample depends on a continuous outcome variable. Weaver and Zhou (2005) developed semi-parametric estimators for regression coefficients using the ODS framework which utilize incomplete information for the un-sampled portion of the population. In the design proposed by Weaver and Zhou (2005), the outcome which is used to develop the ODS is observed for all subjects but covariates, which are observed for subjects selected in the sample, are missing for the un-sampled portion of the population.

We propose the use of a test-dependent sampling (TDS) design in which TDS inclusion depends on the continuous screening test measure, such as the Balcone risk score. The TDS design incorporates an SRS component, a TDS component, and the remaining un-sampled portion of the population. The TDS design allows investigators to over-sample subjects from specified ranges of the screening test variable, allowing for a concentration of resources where there is the greatest amount of information. All data are available for subjects sampled in the study, but only the screening test value is available for the un-sampled portion of the population. Wang et al. (2012) showed that the TDS design yields more efficient estimates of AUC than the SRS design. We show that the efficiency of AUC estimation under the TDS design can be further improved by utilizing information from both the sample and un-sampled subjects.

Many screening tests define results in dichotomous terms, "positive" for a test value above a threshold and "negative" for a test value below the threshold. Verification bias is a concern for studies where all subjects whose test result is "positive" for the outcome have their disease status verified (Pepe, 2004) but among subjects who test "negative" for the outcome, either none or a subset of this group have their disease status verified. For the TDS design, subjects are selected conditional on their test result, but only a subset of subjects within groups are selected for ascertaining disease status. Accordingly, verification bias exists in the TDS design (Wang et al., 2012, 2013). In the proposed design, all subjects are included in estimating

the AUC. Although disease status is missing for subjects in the un-sampled portion of the population, the inclusion of these subjects eliminates the bias typically associated with the TDS design.

This chapter is organized as follows. In Section 2.2, we introduce existing AUC estimators, propose use of empirical likelihood methods to develop a semi-parametric estimator for AUC. In Section 2.3, asymptotic properties of the proposed AUC estimator are studied. In Section 2.4, we use simulation studies to compare the proposed estimator with existing methods. In Section 2.5, we analyze data from the lung cancer study. In Section 2.6, we use the proposed method to analyze data from the Preterm Prediction Study. We conclude with a discussion in Section 2.7.

## 2.2 Semi-parametric empirical likelihood AUC (SPEL-AUC) estimation

### 2.2.1 Notation and data structure for the SPEL-AUC

Consider a continuous test variable, $Y$, and a binary disease indicator, $D$. The distribution of $Y$ can be divided into $K$ mutually exclusive intervals defined by $C_k = (a_{k-1}, a_k]$ where $k = 1, \ldots, K$. The sample size within each of the $C_k$ intervals can be different. The TDS is made up of three components: the SRS component, TDS component, and non-validation set (un-sampled subjects). The subjects sampled in the SRS and TDS components combined make up the validation set, indexed by $V$, where true disease status is validated. The remainder of the population not selected for ascertainment of disease status makes up the non-validation set, indexed by $\bar{V}$. The sample size of the validation set is given by $n_V = n_0 + \sum_{k=1}^{K} n_k$, where $n_0$ is the sample size from the SRS component and $n_k$ is the sample size for the $k^{th}$ TDS component interval, $k = 1, \ldots, K$. The size of the non-validation set is given by $n_{\bar{V}} = N - n_V$. To define subscripts, $i$ indexes the sampling group where $i = \{0, 1, ..., K, \bar{V}\}$ and $j = \{1, ..., n_i\}$ denotes the individual in the $i^{th}$ sampling group. The test variable, $Y$, is observed for all subjects in the dataset. The disease status, $D$, is only ascertained for subjects who are in the

validation set. The data framework is given by

$$\begin{array}{lll} \text{SRS component} & (D_{0j}, Y_{0j}) & j = 1, ..., n_0 \\ \text{TDS}_1 \text{ component} & (D_{1j}, Y_{1j} | Y_{1j} \in C_1) & j = 1, ..., n_1 \\ \vdots & \vdots & \vdots \\ \text{TDS}_K \text{ component} & (D_{Kj}, Y_{Kj} | Y_{1j} \in C_K) & j = 1, ..., n_K \\ \text{Non-validation component} & \left(Y_{\bar{v}j} | i = \bar{V}\right) & j = 1, ..., n_{\bar{V}}. \end{array} \tag{2.1}$$

### 2.2.2 Existing AUC estimators

Two existing nonparametric AUC estimators are included in the simulation study to compare with the proposed AUC estimator. These estimators are

1) the SRS only estimator (MW-AUC) proposed by Bamber (1975), denoted $\hat{A}_V^{SRS}$, and

2) the empirical likelihood estimator (NPEL-AUC) proposed by Wang et al. (2012), denoted $\hat{A}_V^{TDS}$.

First, we introduce the MW-AUC. This estimator utilizes a SRS design and is equivalent to the Mann-Whitney U-statistic (Pepe, 2004), given by

$$\hat{A}_V^{SRS} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} D_i \left(1 - D_i\right) I \left(Y_i > Y_j\right)}{\sum_{i=1}^{n} \sum_{j=1}^{n} D_i \left(1 - D_i\right)}. \tag{2.2}$$

The second estimator under comparison, NPEL-AUC, was proposed by Wang et al. (2012) for a TDS design where TDS inclusion depends on the test variable, Y. Empirical likelihood methods were used to avoid making distributional assumptions on the screening test, Y. The data structure is similar to the structure described in Section 2.2.1, except that non-validation data are not included in the NPEL-AUC. The sample size within each of the $C_k$ intervals is equal. The sample size is given by $n = n_0 + \sum_{k=1}^{K} n_k$, where $n_0$ is the sample size from the SRS component and $n_k = \frac{n - n_0}{K}$ is the sample size for the $k^{th}$ TDS interval, $k = 1, \ldots, K$. The estimator is given by

$$\hat{A}_V^{TDS} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \hat{p}_i \hat{p}_j D_i \left(1 - D_i\right) I \left(Y_i > Y_j\right)}{\sum_{i=1}^{n} \sum_{j=1}^{n} \hat{p}_i \hat{p}_j D_i \left(1 - D_i\right)}, \tag{2.3}$$

where $\hat{p}_i = \left[n_0 + \sum_{k=1}^{K} \frac{n_k}{\hat{\theta}_k} I\left(y_i \in C_k\right)\right]^{-1}$. The biased sampling scheme is accounted for by incorporating the empirical probability masses $p_i$ and $p_j$ in the numerator and denominator. The MW-AUC and NPEL-AUC use the same number of subjects to estimate the AUC and differ in their allocation of subjects, where the MW-AUC uses only SRS to generate the sample.

### 2.2.3 Semi-parametric empirical likelihood approach

Denote $f_{Y,D}\left(Y_{ij}, D_{ij}\right)$ the joint distribution of disease status $(D)$ and screening test $(Y)$, $f_Y\left(Y_{ij}\right)$ the marginal distribution of $Y$, and $f_{Y,D}\left(Y_{ij}, D_{ij} | Y_{ij} \in C_k\right)$ the distribution of $Y$ and $D$ conditional on $Y_{ij} \in C_k$, $k = (1, 2, 3)$. For $k = \{1, 2, 3\}$, let $N_k$ be the stratum size for the $k^{th}$ strata in the population, $N_k = n_{0,k} + n_k + n_{\bar{V},k}$, where $n_{0,k} = \sum_{j=1}^{n_i} I(Y_{0j} \in C_k)$, $n_{\bar{V},k} = \sum_{j=1}^{n_i} I(Y_{\bar{V}j} \in C_k)$, and $n_k$ is predetermined. The likelihood for the validation data can be written as

$$\begin{aligned} L_V\left(f_D\right) &= \prod_{j=1}^{n_0} f\left(Y_{0j}, D_{0j}\right) \times \prod_{k=1}^{K} \prod_{j=1}^{n_k} f\left(Y_{kj} D_{kj} | Y_{kj} \in C_k\right) \\ &= \prod_{k=1}^{K} \prod_{j=1}^{n_k} Pr\left(Y_{kj} \in C_k\right)^{-n_k} \times \prod_{k=0}^{K} f\left(Y_{ij} | D_{kj}\right) Pr\left(D_{kj} = d\right). \end{aligned} \tag{2.4}$$

The non-validation portion of the likelihood takes into account the missing data, where the disease status is unknown, and is given by

$$\begin{aligned} L_{\bar{V}}\left(f_D\right) &= \frac{(N - n_0)!}{\prod_{k=1}^{K}\left(N_k - n_{0,k}\right)!} \prod_{k=1}^{K} Pr\left(Y_{kj} \in C_k\right)^{N_k - n_{0,k}} \prod_{\substack{j=1 \\ Y \in C_k}}^{n_{\bar{V}}} \frac{f\left(Y_{\bar{V}j}\right)}{Pr\left(Y_{kj} \in C_k\right)} \\ &= \frac{(N - n_0)!}{\prod_{k=1}^{K}\left(N_k - n_{0,k}\right)!} \times \prod_{j=1}^{n_{\bar{V}}} \sum_{d=0}^{1} f\left(Y_{\bar{V}j} | D_{\bar{V}j} = d\right) Pr\left(D_{\bar{V},j} = d\right) \\ &\quad \times \prod_{k=1}^{K} Pr\left(Y_{kj} \in C_k\right)^{-n_k}. \end{aligned} \tag{2.5}$$

The full likelihood is found by combining the validation and non-validation portions of the likelihood, 2.4 and 2.5, given by

$$\begin{aligned} L\left(\{q_{ij}\},\{r_{ij}\},p\right) &= L_V\left(f_D\right) \times L_{\bar{V}}\left(f_D\right) \\ &\propto \prod_{\substack{k,j\in V \\ D=1}} f\left(Y_{kj}|D_{kj}=1\right) Pr\left(D_{ij}=1\right) \prod_{\substack{k,j\in V \\ D=0}} f\left(Y_{kj}|D_{kj}=0\right) Pr\left(D_{kj}=0\right) \\ &\quad \times \prod_{j=1}^{n_{\bar{V}}} \sum_{d=0}^{1} f\left(Y_{\bar{V}j}|D_{\bar{V}j}=d\right) Pr\left(D_{\bar{V}j}=d\right) \\ &= \prod_{\substack{k,j\in V \\ D=1}} q_{kj}p \times \prod_{\substack{k,j\in V \\ D=0}} r_{kj}\left(1-p\right) \times \prod_{j=1}^{n_{\bar{V}}} \left[q_{\bar{V}j}p + r_{\bar{V}j}\left(1-p\right)\right], \end{aligned} \tag{2.6}$$

where $p = Pr\left(D=1\right)$, $q_{ij} = f\left(Y_{ij}|D_{ij}=1\right)$, and $r_{ij} = f\left(Y_{ij}|D_{ij}=0\right)$. We propose to non-parametrically estimate $\{q_{ij}\}$ and $\{r_{ij}\}$ in (2.6). An interesting constraint between $\{q_{ij}\}$ and $\{r_{ij}\}$, given by $\frac{r_{ij}}{q_{ij}} = e^{\alpha+\beta y_{ij}}$, was developed by Qin and Zhang (1997, 2003). To see this, consider the standard logistic regression model where $Pr\left(D=1|Y\right) = \frac{e^{m^*(Y)\boldsymbol{\alpha}}}{1+e^{m^*(Y)\boldsymbol{\alpha}}} = \psi\left(Y\right)$. The Bayes' rule gives $f\left(Y|D=1\right) = \frac{f(Y)Pr(D=1|Y)}{Pr(D=1)} = \frac{f(Y)\psi(Y)}{p}$. Similarly, $f\left(Y|D=0\right) = \frac{f(Y)(1-\psi(Y))}{1-p}$. Consider the ratio

$$\begin{aligned} \frac{r_{ij}}{q_{ij}} &= \frac{f\left(Y|D=0\right)}{f\left(Y|D=1\right)} \\ &= \left[\frac{f\left(Y\right)\left(1-\psi\left(Y\right)\right)}{1-p}\right] \times \left[\frac{p}{f\left(Y\right)\psi\left(Y\right)}\right] \\ &= \frac{p}{1-p}\left(\frac{1}{1+e^{m^*(Y)\boldsymbol{\alpha}}}\right)\left(\frac{e^{m^*(Y)\boldsymbol{\alpha}}}{1+e^{m^*(Y)\boldsymbol{\alpha}}}\right) \\ &= e^{m(Y)\boldsymbol{\alpha}}, \end{aligned}$$

which implies that $r_{ij} = q_{ij}e^{m(\mathbf{X}_{ij})\boldsymbol{\alpha}}$. Let $m\left(\mathbf{X}_{ij}\right)\boldsymbol{\alpha} = \alpha + \beta Y$. Applying this constraint to the log-likelihood gives

$$l\left(\{q_{ij}\},p,\alpha,\beta\right) \propto \sum_{ij} \ln q_{ij} + n_{V,D=0}\alpha + \beta \sum_{\substack{i,j\in V \\ D=0}} y_{ij} + \sum_{j=1}^{n_{\bar{V}}} \ln\left[p + e^{\alpha+\beta y_{\bar{V}j}}\left(1-p\right)\right] \tag{2.7}$$

Without loss of generality, consider partitioning the screening test variable into three

mutually exclusive intervals: $C_1 = (-\infty, a_1]$, $C_2 = (a_1, a_2]$, and $C_3 = (a_2, \infty)$. The TDS consists of an SRS of size $n_0$, a TDS component of size $n_1 + n_2 + n_3$, and the non-validation set of size $N - n_V$, where $n_V = n_0 + n_1 + n_2 + n_3$. Subjects are eligible to be sampled for the TDS component based on their screening test result, Y. For example, if a subject's test result is less than or equal to $a_1$ and $n_1 > 0$, the probability of being selected in the TDS component for $C_1$ is greater than zero.

To develop the proposed SPEL-AUC estimator, we first estimate $(\{q_{ij}\}, \alpha, \beta)$ using empirical likelihood methods outlined below. We then estimate the expected value of disease for subjects in the non-validation set, whose true disease status is missing, using $(\{\hat{q}_{ij}\}, \hat{\alpha}, \hat{\beta})$ and use this expected disease status estimate for those in the non-validation set to construct the proposed SPEL-AUC estimator.

Estimation of $(\{q_{ij}\}, \alpha, \beta)$ is done using the profile likelihood approach by first fixing $(\alpha, \beta)$ and obtaining $\{q_{ij}\}$ from a constrained likelihood function. Specifically, we estimate $\{q_{ij}\} = f(Y_{ij}|D_{ij} = 1)$ by maximizing a constrained likelihood function of (2.6) under the following constraints:

$$\left\{ q_{ij} \geq 0, \sum_{i \in (0,1,2,3,\bar{V})} \sum_{j=1}^{n_i} q_{ij} = 1, \sum_{i \in (0,1,2,3,\bar{V})} \sum_{j=1}^{n_i} q_{ij} \left\{ e^{\alpha + \beta y_{ij}} - 1 \right\} = 0 \right\}. \tag{2.8}$$

A unique maximum for $\{q_{ij}\}$ exists under the constraints given in (2.8) if 1 is inside the convex hull of points $e^{\alpha+\beta y_{ij}}$ for all $(i, j)$ (Owen, 1988, 1990; Qin and Lawless, 1994). The Lagrange multiplier method can be used to derive the maximum over $\{\hat{q}_{ij}\}$. The logarithm of the constrained likelihood is as follows:

$$\begin{aligned} H \quad \propto \quad & \sum_{ij} \ln q_{ij} + n_{v,D=0}\alpha + \beta \sum_{\substack{i,j \in V \\ D=1}} y_{ij} + \sum_{j=1}^{n_{\bar{V}}} \ln \left[ p + e^{\alpha + \beta y_{\bar{V}j}} (1-p) \right] \\ & + \lambda_1 \left( 1 - \sum_{ij} q_{ij} \right) + N\lambda_2 \sum_{ij} q_{ij} \left\{ e^{\alpha+\beta y_{ij}} - 1 \right\}. \end{aligned} \tag{2.9}$$

Estimates $\{\hat{q}_{ij}\}$ and $\hat{\lambda}_1$ are found by taking the derivative of H with respect to $q_{ij}$ and setting the derivative equal to zero. The derivative of H is given by $\frac{\partial H}{\partial q_{ij}} = \frac{1}{q_{ij}} - \lambda_1 +$

$N\lambda_2\left\{e^{\alpha+\beta y_{ij}}-1\right\}$. The estimate $\hat{\lambda}_1$ is found by evaluating $\sum_{ij} q_{ij}\frac{\partial H}{\partial q_{ij}} = N-\lambda_1\sum_{ij} q_{ij} + N\lambda_2\sum_{ij} q_{ij}\left\{e^{\alpha+\beta y_{ij}}-1\right\}=0$. By setting the derivative of H equal to zero and solving for $q_{ij}$, we have

$$\hat{q}_{ij} = \frac{1}{N}\left[1-\lambda_2\left(e^{\alpha+\beta y_{ij}}-1\right)\right]^{-1}, \tag{2.10}$$

for $i \in (0,1,2,3,\bar{V})$ and $j \in (1,\ldots,n_i)$.

*Profile log-likelihood*

The empirical profile log-likelihood is obtained by plugging the estimates $\hat{q}_{ij}$ given by (2.10) into (2.7). Denoting $pl(\xi)$ as the natural logarithm of the empirical profile likelihood, we have

$$\begin{aligned} pl(\xi) \quad \propto \quad & -\sum_{ij}\ln\left[1-\lambda_2\left(e^{\alpha+\beta y_{ij}}-1\right)\right] + n_{V,D=0}\alpha + \beta\sum_{\substack{i,j\in V\\ D=0}} Y_{ij} \\ & + \sum_{j=1}^{n_{\bar{V}}}\ln\left[p+e^{\alpha+\beta Y_{\bar{V}j}}(1-p)\right]. \end{aligned} \tag{2.11}$$

The Newton-Raphson algorithm can be used to obtain $\hat{\xi}$, where $\xi=(\alpha,\beta,\lambda_2)$. These estimators, $\hat{\xi}$, are used in estimating the expected disease status. Disease status is unknown for the non-validation portion of the population. For these subjects, an estimate of the expected disease status is used in place of the true disease status in the SPEL-AUC. The expected value of disease is given by

$$\begin{aligned} E(D_l) &= 1 * Pr(D_l=1|Y_l) \\ &= \frac{f(D_l=1,Y_l)}{f(Y_l)} \\ &= \frac{f(Y_l|D_l=1)\,Pr(D_l=1)}{f(Y_l|D_l=1)\,Pr(D_l=1)+f(Y_l|D_l=0)\,Pr(D_l=0)} \\ &= \frac{q_l p}{q_l p + r_l(1-p)} \\ &= \frac{p}{p+e^{\alpha+\beta y_{ij}}(1-p)}. \end{aligned} \tag{2.12}$$

An estimate of the expected disease status is found by plugging estimators $\hat{\xi}$ and $\hat{p}$ into (2.12). This estimate of expected disease status is given by

$$\widehat{E(D_l)} = \frac{\hat{p}}{\hat{p} + e^{\hat{\alpha}+\hat{\beta}y_l}(1-\hat{p})}. \tag{2.13}$$

The SPEL-AUC uses the information from both the validation and non-validation portions of the population. Estimated expected disease, given by (2.13), is used for non-validation subjects where the true disease status is missing. Let $l = 1, ..., N$ index the entire population. The SPEL-AUC is given by

$$\hat{A}^P_{V,\bar{V}} = \frac{\sum_{l\neq l'}^{N} D_l^* \left(1 - D_{l'}^*\right) I\left(Y_l > Y_{l'}\right)}{\sum_{l\neq l'}^{N} D_l^* \left(1 - D_{l'}^*\right)},$$
$$\text{where } D_l^* = \begin{cases} D_l & \text{if } l \in V \\ \widehat{E(D_l)} = \frac{\hat{p}}{\hat{p}+e^{\hat{\alpha}+\hat{\beta}y_l}(1-\hat{p})} & \text{if } l \in \bar{V} \end{cases}. \tag{2.14}$$

## 2.3 Asymptotic properties of the SPEL-AUC

The asymptotic properties of the proposed SPEL-AUC estimator are established in Theorem 1. Detail of the proof is provided in the appendix, including asymptotic results for the components that make up the proposed SPEL-AUC estimator. Consider the U-process $U_N(A,\eta) = R_N(A,\eta) - E(R_N(A,\eta))$. Let $R_N(A,\eta) = \frac{1}{N^2}$ $\times \sum_{i\neq j} D_i' \left(1 - D_j'\right)(I_{ij} - A)$ where $I_{ij} = I(Y_i > Y_j, Y_j \in (t_0, t_1))$ and $D_l' = \begin{cases} D_l & \text{if } l \in V \\ \widehat{E(D_l)} = \frac{p}{p+e^{\alpha+\beta y_l}(1-p)} & \text{if } l \in \bar{V} \end{cases}$. Using this U-process, we show that

$$\sqrt{N}\left(\hat{A}^P_{V,\bar{V}} - A\right) = -\left\{\frac{\partial E\left[R_N(A,\eta)\right]}{\partial A}\right\}^{-1} \sum_{i\in\left(0,1,2,3,\bar{V}\right)} \rho_i n_i^{-1/2} \sum_{j=1}^{n_i} Q_{ij},$$

where $Q_{ij}(\eta) = E\left(R_{(ij)(ij)'} + R_{(ij)'(ij)}\right) + \rho_i^{-1} \frac{\partial E R_N(A,\eta)}{\partial p}\left[\frac{-1}{n_0}\frac{\partial^2 l_{srs}(p)}{p^2}\right]^{-1} P_{0j} I(i=0)$ $+ \frac{\partial E R_N(A,\eta)}{\partial \xi}\left[\frac{-1}{N}\frac{\partial^2 pl(\xi)}{\partial \xi_i \partial \xi_{i'}}\right]^{-1} \mathbf{H}_{ij}(\eta)$.

***Theorem 1:*** Under general regularity conditions,

$$\sqrt{N}\left(\hat{A}^{P}_{V,\bar{V}} - A\right) \xrightarrow{d} \mathcal{N}(0, \Sigma), \tag{2.15}$$

where $\Sigma = \left[\frac{\partial E[R_N(A,\eta)]}{\partial A}\right]^{-2} \sum_{i\in\left(0,1,2,3,\bar{V}\right)} \rho_i var\left(Q_{ij}\right)$.

We propose a variance estimator using the above asymptotic variance by replacing the large sample quantities in $\Sigma$ with their corresponding finite sample quantities. Specifically, we have

$$\hat{\Sigma} = \left[\frac{\partial E\left[R_N\left(A,\eta\right)\right]}{\partial A}\right]^{-2} \sum_{i\in\left(0,1,2,3,\bar{V}\right)} \hat{\rho}_i var\left(\hat{Q}_{ij}\right). \tag{2.16}$$

## 2.4 Simulation study

We evaluate the behavior of the SPEL-AUC under various situations to examine its behavior. The simulation studies were conducted using R version 2.14. The data were generated under the model

$$Y = \beta_0 + D\beta_1 + \epsilon,$$

where $D = 1$ for diseased subjects and $D = 0$ for non-diseased subjects. For the following simulations, we generated data where $\epsilon \sim \mathcal{N}(0,1)$ and $D \sim Bernoulli\left(0.3\right)$. The population size used in simulations is $N = 2000$ and the distribution of $Y$ is partitioned into three mutually exclusive sets given by $C_1 = (-\infty, a_1]$, $C_2 = (a_1, a_2]$, and $C_3 = (a_2, \infty)$. In the following simulations, we consider the impact of 1) varying the cut points used to partition the range of screening test values, $\{a_1, a_2\} = \{\mu_Y - \alpha\sigma_Y, \mu_Y + \alpha\sigma_Y\}$; 2) varying overall sample size, $n_V$; 3) varying the proportion of SRS to TDS component sizes, $\frac{n_0}{n_V}$; and 4) varying model parameter $\beta_1$. The variations under consideration are: cut-point parameter ($\alpha$) 1 and 1.5; validation sample size ($n_V$) 120, 240, and 360; proportion of SRS subjects among validation set ($\frac{n_0}{n_V}$) 0.5 and 0.75; and model parameter for disease status ($\beta_1$) 0.5 and 1. For simulation results, the estimated means, standard errors, mean of the standard error estimates, and 95%

nominal coverage probabilities for an estimator are obtained from 1000 independent runs.

*Sample allocation for TDS*

The SPEL-AUC uses the TDS design to target subjects on both tails of the distribution. This sample consists of the following three components:

1) SRS component of size $n_0$,

2) TDS component sample of size $n_1 + n_2 + n_3$ where $n_1$ subjects are sampled such that $Y_{1j} \in C_1 = (-\infty, a_1]$, $n_3$ subjects are sampled such that $Y_{3j} \in C_3 = (a_2, \infty)$, and $n_1$=$n_3$, and

3) non-validation component of size $N - n_V$, comprised of all subjects not sampled into the SRS and TDS components.

Because the TDS component only over-samples from the tails, $n_2 = 0$ where $Y_{2j} \in C_2 = (a_1, a_2]$. Other sample allocations were considered and results are presented in the last two lines of Table 2.2. The two-tailed allocation for the SPEL-AUC was chosen due to its consistent estimation of the AUC and consistent reduction in standard error compared to other sample allocations. For the NPEL-AUC, proposed by Wang et al. (2012), the sample consists of the SRS and TDS components, except that in the TDS component, the subjects are allocated equally across the three intervals, such that $Y_{ij} \in C_i$ and $n_1 = n_2 = n_3$.

*Estimators to be compared*

The SPEL-AUC, $\hat{A}^P_{V,\bar{V}}$ in (2.14), is compared to three estimators in the simulation studies. These estimators are given below. Specifically, under each setting, we compare the following four estimators.

1) MW-AUC: the SRS only estimator (Bamber, 1975), denoted by $\hat{A}^{SRS}_V$, is given by (2.2).

2) SPEL-AUC(SRS): the SRS with validation and non-validation data estimator, denoted $\hat{A}^{SRS}_{V,\bar{V}}$, has the same form as the proposed estimator given in (2.14). The difference between this estimator and the proposed estimator is the sampling design. This gives a comparison of the SRS and TDS methods while incorporating non-validation data.

3) NPEL-AUC: the TDS data only estimator (Wang et al., 2012), denoted $\hat{A}_V^{TDS}$, is given by (2.3).

4) SPEL-AUC: the proposed TDS with validation and non-validation data estimator, denoted $\hat{A}^P_{V,\bar{V}}$, is given by (2.14).

*Results*

*Unbiasedness* Simulation results are summarized in Tables 2.1 through 2.3. All four AUC estimators yield unbiased estimates. To illustrate this, we simulated data using multiple allocations, sample sizes, and cut-points. Tables 2.1 and 2.3 show that the averages of all AUC estimators are close to or equal to the true value. Other allocation schemes were considered for the SPEL-AUC in Table 2.2 by varying the proportion of subjects allocated to the SRS component ($\frac{n_0}{n_V} = (0.5, 0.75)$), the cut-point defining the TDS component intervals ($\alpha = (1, 1.5)$ where $\{a_1, a_2\} = \{\mu_Y - \alpha\sigma_Y, \mu_Y + \alpha\sigma_Y\}$), and the TDS component allocation (one tail, both tails, and three intervals). In all of these variations, the SPEL-AUC continues to be unbiased.

*Efficiency* Efficiency-wise, the proposed SPEL-AUC, $\hat{A}^P_{V,\bar{V}}$, is the most efficient among all compared. The NPEL-AUC, $\hat{A}_V^{TDS}$, is more efficient than the SRS estimators (SPEL-AUC(SRS) and MW-AUC) and the MW-AUC is the least efficient among those compared. This supports the idea that both the use of the TDS design and inclusion of non-validation subjects create a more efficient alternative to the SRS design and validation-only estimators while sampling the same number of subjects. For the SPEL-AUC, efficiency is similar when the cut-points, $\mu_Y \pm \alpha\sigma_Y$, are further from the mean ($\alpha = 1$ versus $\alpha = 1.5$) but performance of the asymptotic standard error estimator ($\hat{SE}$) is improved for $\alpha = 1$. The asymptotic standard error estimator is obtained using the proposed asymptotic variance estimator in (2.16), by substituting finite sample quantities for large sample quantities. For example, consider the results in Table 2.2 for $n_V = 360$, $\frac{n_0}{n_V} = 0.75$, and allocation (270,45,0,45,1640). For both $\alpha = 1$ and $\alpha = 1.5$ SE=0.024 with $\hat{SE} = 0.23$ for $\alpha = 1$ compared to $\hat{SE} = 0.22$ for $\alpha = 1.5$. We show a similar result when comparing the proportion of SRS subjects within the TDS sample for $\frac{n_0}{n_V} = 0.75$ versus 0.5. In Table 2.2, for $n_V = 360$ and $\alpha = 1$,

we have SE=0.024 for both $\frac{n_0}{n_V} = 0.75$ and $\frac{n_0}{n_V} = 0.5$ but the asymptotic standard error is underestimated in the case of $\frac{n_0}{n_V} = 0.5$ ($\hat{SE} = 0.21$ versus $\hat{SE} = 0.23$ when $\frac{n_0}{n_V} = 0.75$). Table 2.2 shows that the estimated standard errors are very close to their true values. The last two lines of Table 2.2 show that for the SPEL-AUC, allocation to only one tail gives results that are less efficient than the two-tailed allocation.

*Robustness* The SPEL-AUC does not require model specification for the screening test, Y. To explore the SPEL-AUC's robustness, simulation studies were generated using both Normal and Chi-squared distributions for the screening test. Simulation results reported in Table 2.3 show that across multiple sample sizes (120, 240, and 360) the simulation study gives similar results. All estimators are unbiased and the SPEL-AUC is more efficient than the MW-AUC, SPEL-AUC(SRS), and NPEL-AUC estimators.

## 2.5 Analysis of the lung cancer study data

We used the SPEL-AUC to analyze non-small-cell lung cancer (NSCLC) data from the CALGB 150807 study conducted by the Cancer and Leukemia Group B (Bueno et al., 2012). This study is a subset of patients registered in the CALGB 140202 study who have stage 1A or 1B non-small-cell lung cancer (NSCLC), have not received preoperative chemotherapy or radiation, and are not missing histological, demographic, clinical, and follow-up information of interest. Among patients in the CALGB 150807 study, 1,061 patients were not censored before 12 months and were used in this analysis.

Lung cancer is the most common cause of cancer death among men and women in the world (Blanchon et al., 2006). Lung cancer is classified as either small-cell lung carcinoma (SCLC) or non-small-cell lung carcinoma (NSCLC), of which NSCLC accounts for approximately 80% of all lung cancers. After surgical lung resection, a large proportion of stage 1 NSCLC patients have cancer recurrence within five years (Bueno et al., 2012). When surgery is used as the primary treatment for NSCLC, adjuvant chemotherapy may benefit patients who have a high risk of cancer recurrence. Identifying patients who are at high risk of cancer recurrence is important in order for treatment to be given to those who would benefit most. This is

an important area of study for patients, families, and doctors when making decisions on a treatment plan.

The Balcone risk score, outlined by Blanchon et al. (2006), has been developed to identify patients who are at greatest risk of cancer recurrence. To select the variables that are included in the scoring algorithm, Blanchon et al. (2006) used a Cox model to identify variables that were independently associated with mortality. The associated variables were then weighted to create the Balcone risk index. The components of the risk score are given by: age (>70 years, 1 point); sex (male, 1 point); performance status at diagnosis (reduced activity, 3 points; active >50%, 5 points; inactive >50%, 8 points; and total incapacity, 10 points); histological type (large-cell carcinoma, 2 points); and tumour-node-metastasis (TNM) staging system (IIA or IIB, 3 points; IIIA or IIIB, 6 points; and IV, 8 points). For the data used in this analysis, the Balcone risk score ranges from 0 to 15.

The goal of this analysis is to summarize the ability of the Balcone risk score to distinguish between patients who survive beyond 12 months and those who do not. This will allow us to evaluate the benefit of using this risk score to identify patients at a higher risk of early cancer recurrence. The outcome of interest is survival beyond 12 months and the screening test is the Balcone risk score. Although all information is available for these patients, we illustrate the utility of the proposed AUC estimator by sampling from the study data and evaluating the estimated AUC. Table 2.4 gives descriptive statistics for the Balcone risk score, stratified by survival at 12 months. Cut-points for the TDS component were defined by $\alpha = 1$ standard deviations from the mean of the Balcone risk score. A sample size of $n_V = 360$ was used for all estimators compared and details on sample allocation for each estimator are given in Table 2.5. The MW-AUC has an SRS of size $n = 360$. The SPEL-AUC(SRS) allocates 100% of the sample to the SRS component and utilizes incomplete data from the non-validation set. The NPEL-AUC allocates 50% of the validation sample to the SRS component and the remaining 50% are allocated equally between the three intervals, $C_i$, such that $n_1 = n_2 = n_3 = 60$. The SPEL-AUC allocates 75% of the sample to the SRS component and samples the remaining 25% from the tails, where $n_1 = n_3 = 45$ and $n_2 = 0$, while utilizing incomplete data from the non-validation set. Because all data has been obtained for these patients, we can apply the

MW-AUC estimator on the full data (1,061 subjects) and we obtain an AUC of 0.657. Using a sample size of $n = 360$, the AUC estimates and estimates of the standard error for the methods compared are given by: MW-AUC 0.680 (0.051), SPEL-AUC(SRS) 0.670 (0.050), NPEL-AUC 0.681 (0.046), and SPEL-AUC 0.724 (0.044). All estimators compared have an estimated AUC that is similar to the best estimate we have for the true pAUC, which is 0.657, found by using the complete data. We can also see that the proposed SPEL-AUC estimator has the smallest estimated variance among the four AUC estimators compared.

## 2.6 Analysis of the Preterm Prediction Study data

We used the SPEL-AUC to analyze data from the Preterm Prediction Study, a multi-center prospective study designed to study spontaneous preterm birth (Goldenberg et al., 1996). The Maternal Fetal Medicine Units Network of the National Institute of Child Health and Human Development carried out this study using ten centers to recruit women. There were 3073 women recruited between October 1992 and July 1994. Measurements were collected every two weeks from 22 to 30 weeks' gestation. Among the 3073 women recruited, 3001 had valid measurements of interest for this analysis.

PTB, defined as delivery at less than 37 weeks of gestation, contributes to neonatal morbidity and mortality which increases as gestational age decreases (McCormick, 1985). Bastek and Elovitz (2013) combined results from multiple studies on this topic to gain a better understanding the relationship between biomarkers and PTB. The results were not definitive for most biomarkers with the exception of fetal fibronectin (FFN).

Fetal fibronectin (FFN) is a protein that is produced by the fetal membrane. Knowing the FFN measurement will not change the incidence of spontaneous PTB but it will effect the treatment plan. Deshpande et al. (2013) found that FFN has moderate accuracy in predicting PTB. Although many studies are concerned with the ability to predict spontaneous PTB, Bastek and Elovitz (2013) suggest that the ability to predict those who will not have spontaneous PTB is also valuable. Because FFN typically has a high negative predictive value (proportion of true negatives over all who test negative), a negative FFN test is widely used

in clinical practice to send patients home. Measurable levels of FFN are considered to be abnormal between 20 and 37 weeks' gestation. Lockwood et al. (1991) show that in 588 FFN samples from uncomplicated pregnancies, a higher percentage of subjects were positive for FFN (level above 0.05 $\mu$g/mL) before 22 and after 37 weeks' gestation compared to between 22 and 37 weeks' gestation. For example the percentage of cervical samples with positive FFN for <22, 22 to 37 and >37 weeks' gestation were 24%, 4%, and 32%, respectively. This is an important area of study due to the negative effects of spontaneous PTB on maternal and child health outcomes.

In our analysis, the outcome of interest is spontaneous PTB at less than 37 weeks' gestation and the screening test considered in FFN. Table 2.6 gives descriptive statistics for FFN, stratified by spontaneous PTB. Because the standard deviation is large compared to the mean, the cut-points for the TDS component were defined using $\alpha = 0.15$. A sample size of $n_V = 360$ was used for all estimators compared and details on sample allocation for each estimator are given in Table 2.7. The MW-AUC has an SRS of size $n = 360$. The SPEL-AUC(SRS) allocates 100% of the sample to the SRS component and utilizes incomplete data from the non-validation set. The NPEL-AUC allocates 50% of the validation sample to the SRS component and the remaing 50% are allocated equally between the three intervals, $C_i$, such that $n_1 = n_2 = n_3 = 60$. The SPEL-AUC allocates 75% of the sample to the SRS component and samples the remaining 25% from the tails, where $n_1 = n_3 = 45$ and $n_2 = 0$, while utilizing incomplete data from the non-validation set. Because the outcome of spontaneous PTB is known for all patients, we can apply the MW-AUC estimator on the full data (3,001 patients) and we obtain an AUC of 0.576. The estimates from each estimator are: MW-AUC 0.522, SPEL-AUC(SRS) 0.536, NPEL-AUC 0.575, and SPEL-AUC 0.591. These results show that all estimators accurately estimate the AUC.

## 2.7 Discussion

We have proposed a semi-parametric estimator for area under the ROC curve (AUC), which allows us to summarize the ability of a screening test to discern between diseased and

non-diseased subjects. This estimator incorporates a TDS design and includes both validation and non-validation data. The use of empirical likelihood methods allows us to estimate the AUC without specifying a distribution for the screening test. We establish the asymptotic properties of the proposed estimator under general regularity conditions and show that this estimator has good finite sample properties.

The proposed design is motivated by the need to improve efficiency in estimating AUC. Ascertaining true disease status can be costly and invasive for subjects. Although disease status is missing for the non-validation set, the proposed estimator takes advantage of all information available for a larger number of subjects than the validation-only estimators. Although all estimators are unbiased, the proposed estimator was shown to be the most efficient, compared to three competing AUC estimators. This suggests that to obtain the same variability less subjects would be needed when the proposed method is used, reducing study cost and subject burden. These results support the idea that both the use of the TDS design and inclusion of non-validation subjects create a more efficient alternative to the SRS design and the validation-only estimators while sampling the same number of subjects. For example, with a sample size of 240, the standard error of the proposed estimator is 0.03 compared to 0.034 and 0.035 in the competing methods (Table 2.1). In this case, we have an estimated standard error of 0.029 which gives coverage proportion of 0.945. The proposed method is also robust in its ability to estimate AUC under varying distributions. Simulation studies show that when the screening test is simulated from a Chi-squared distribution, the AUC estimators are unbiased and the SPEL-AUC continues to be the most efficient AUC estimator under comparison.

Care should be taken when choosing the sample allocation and cut-points for the test-dependent sample. If data are skewed with a long tail the standard error could be quite large, as in the case of the Preterm Prediction Study and simulation studies where screening test data are generated from a Chi-squared distribution. The proportion of subjects selected for the SRS component and the TDS allocation of subjects within the intervals defined by cut-points will largely depend on the characteristics of the data used.

Table 2.1: Comparison of SPEL-AUC and competing methods

| | $n_V = 120$ | | $n_V = 240$ | | $n_V = 360$ | |
|---|---|---|---|---|---|---|
| Method | Mean | SE | Mean | SE | Mean | SE |
| $Y_{D=1} \sim \mathcal{N}(1,1)$, $Y_{D=0} \sim \mathcal{N}(0,1)$ and True AUC= 0.7601 | | | | | | |
| $\hat{A}_V^{SRS}$ | 0.761 | 0.049 | 0.758 | 0.035 | 0.760 | 0.028 |
| $\hat{A}_{V,\bar{V}}^{SRS}$ | 0.761 | 0.049 | 0.760 | 0.034 | 0.760 | 0.026 |
| $\hat{A}_V^{TDS}$ | 0.762 | 0.048 | 0.759 | 0.034 | 0.759 | 0.027 |
| $\hat{A}_{V,\bar{V}}^{P}$ | 0.765 | 0.045 | 0.762 | 0.030 | 0.761 | 0.024 |
| $Y_{D=1} \sim \mathcal{N}(0.5,1)$, $Y_{D=0} \sim \mathcal{N}(0,1)$ and True AUC= 0.6380 | | | | | | |
| $\hat{A}_V^{SRS}$ | 0.639 | 0.055 | 0.635 | 0.040 | 0.638 | 0.033 |
| $\hat{A}_{V,\bar{V}}^{SRS}$ | 0.639 | 0.055 | 0.638 | 0.038 | 0.637 | 0.031 |
| $\hat{A}_V^{TDS}$ | 0.637 | 0.051 | 0.640 | 0.037 | 0.637 | 0.030 |
| $\hat{A}_{V,\bar{V}}^{P}$ | 0.637 | 0.045 | 0.639 | 0.033 | 0.637 | 0.027 |

Cutpoints for the TDS component are defined by $(a_1, a_2) = (\mu_Y - \alpha\sigma_Y, \mu_Y + \alpha\sigma_Y)$ where $\alpha = 1$. $\hat{A}_V^{SRS}$ denotes the MW-AUC which uses SRS; $\hat{A}_{V,\bar{V}}^{SRS}$ denotes the SPEL-AUC(SRS) which uses SRS and utilizes information for both validation and non-validation data; $\hat{A}_V^{TDS}$ denotes the NPEL-AUC which uses TDS; and $\hat{A}_{V,\bar{V}}^{P}$ denotes the proposed SPEL-AUC which uses TDS and utilizes information from both validation and non-validation data. All estimators sample the same number of subjects.

Table 2.2: Asymptotic results for SPEL-AUC

| $n_V$ | $\frac{n_0}{n_V}$ | $\alpha$ | $(n_0, n_1, n_2, n_3, n_{\bar{V}})$ | Mean | SE | $\widehat{SE}$ | CP |
|---|---|---|---|---|---|---|---|
| | | | $n_1 = n_3$ and $n_2 = 0$ | | | | |
| 120 | 0.75 | 1 | $(90, 15, 0, 15, 1880)$ | 0.765 | 0.045 | 0.043 | 0.936 |
| 240 | 0.75 | | $(180, 30, 0, 30, 1760)$ | 0.762 | 0.030 | 0.029 | 0.945 |
| 360 | 0.75 | | $(270, 45, 0, 45, 1640)$ | 0.761 | 0.024 | 0.023 | 0.938 |
| | | | $n_1 = n_2 = n_3$ | | | | |
| 120 | 0.75 | 1 | $(90, 10, 10, 10, 1880)$ | 0.764 | 0.047 | 0.045 | 0.921 |
| 240 | 0.75 | | $(180, 20, 20, 20, 1760)$ | 0.761 | 0.032 | 0.030 | 0.937 |
| 360 | 0.75 | | $(270, 30, 30, 30, 1640)$ | 0.760 | 0.025 | 0.024 | 0.946 |
| | | | $n_1 = n_3$ and $n_2 = 0$ | | | | |
| 360 | 0.5 | 1 | $(180, 90, 0, 90, 1640)$ | 0.761 | 0.024 | 0.021 | 0.922 |
| 360 | 0.75 | 1.5 | $(270, 45, 0, 45, 1640)$ | 0.761 | 0.024 | 0.022 | 0.931 |
| | | | $n_1 \neq n_3$ and $n_2 = 0$ | | | | |
| 360 | 0.75 | 1 | $(270, 90, 0, 0, 1640)$ | 0.760 | 0.025 | 0.024 | 0.944 |
| 360 | 0.75 | 1 | $(270, 0, 0, 90, 1640)$ | 0.761 | 0.026 | 0.022 | 0.901 |

The true AUC is 0.7601. Screening test data is simulated assuming $Y_{D=1} \sim \mathcal{N}(1,1)$ and $Y_{D=0} \sim \mathcal{N}(0,1)$. The fraction $\frac{n_0}{n_V}$ is the proportion of subjects allocated to the SRS component out of the total number of validation subjects sampled. Cutpoints for the TDS component are defined by $(a_1, a_2) = (\mu_Y - \alpha\sigma_Y, \mu_Y + \alpha\sigma_Y)$. The sample allocation, $(n_0, n_1, n_2, n_3, n_{\bar{V}})$, gives the number of subjects allocated to the SRS component, three intervals of the TDS component, and the non-validation set, respectively.

Table 2.3: Comparison of SPEL-AUC and competing methods for Chi-Squared Distributed Data

| | $n_V = 120$ | | $n_V = 240$ | | $n_V = 360$ | |
|---|---|---|---|---|---|---|
| Method | Mean | SE | Mean | SE | Mean | SE |
| $\hat{A}_V^{SRS}$ | 0.802 | 0.044 | 0.802 | 0.032 | 0.802 | 0.026 |
| $\hat{A}_{V,\bar{V}}^{SRS}$ | 0.790 | 0.044 | 0.790 | 0.032 | 0.790 | 0.025 |
| $\hat{A}_V^{TDS}$ | 0.804 | 0.048 | 0.802 | 0.034 | 0.803 | 0.027 |
| $\hat{A}_{V,\bar{V}}^{P}$ | 0.790 | 0.041 | 0.789 | 0.029 | 0.789 | 0.023 |

The true AUC is 0.8023. Screening test data is simulated assuming $Y_{D=1} \sim \chi(4,3)$ and $Y_{D=0} \sim \chi(3)$. Cutpoints for TDS component are defined by $(a_1, a_2) = (\mu_Y - \alpha\sigma_Y, \mu_Y + \alpha\sigma_Y)$ where $\alpha = 1$.

Table 2.4: Descriptive statistics for the Balcone risk score

| | N | Minimum | Q1 | Median | Q3 | Maximum |
|---|---|---|---|---|---|---|
| Overall | 1076 | 0 | 1 | 1 | 3 | 10 |
| Survival beyond 12 months | 965 | 0 | 1 | 1 | 3 | 10 |
| Survival less than 12 months | 111 | 0 | 1 | 2 | 5 | 9 |

Table 2.5: Sample allocation for the non-small-cell lung cancer data

| Component | MW-AUC | SPEL-AUC(SRS) | NPEL-AUC | SPEL-AUC |
|---|---|---|---|---|
| SRS | 360 | 360 | 180 | 270 |
| TDS $(n_1, n_2, n_3)$ | $(0, 0, 0)$ | $(0, 0, 0)$ | $(60, 60, 60)$ | $(45, 0, 45)$ |
| non-validation | 0 | 701 | 0 | 701 |

Table 2.6: Descriptive statistics for FFN

| | N | Minimum | Q1 | Median | Q3 | Maximum | Mean | St.Dev. |
|---|---|---|---|---|---|---|---|---|
| Spontaneous PTB | 309 | 0 | 0.88 | 4.48 | 17.08 | 924.56 | 43.64 | 123.86 |
| Not PTB | 2692 | 0 | 0.28 | 2.61 | 7.22 | 2151.44 | 13.35 | 83.73 |

Table 2.7: Sample allocation for the Preterm Prediction Study

| Component | MW-AUC | SPEL-AUC(SRS) | NPEL-AUC | SPEL-AUC |
|---|---|---|---|---|
| SRS | 360 | 360 | 180 | 270 |
| TDS $(n_1, n_2, n_3)$ | $(0, 0, 0)$ | $(0, 0, 0)$ | $(60, 60, 60)$ | $(45, 0, 45)$ |
| non-validation | 0 | 2641 | 0 | 2641 |

# Chapter 3

# Partial AUC under Test-Dependent Sampling

## 3.1 Introduction

The receiver operating characteristic (ROC) curve and area under the ROC curve (AUC) are summary measures used to describe the ability of a screening test to discriminate between diseased and non-diseased subjects (Bamber, 1975). When it is clinically relevant to consider only a subset of false positive rates (FPR) or true positive rates (TPR), the partial AUC (pAUC) is another summary measure that should be considered. As evaluating the true disease status can be costly, it is important for researchers to increase study efficiency by allowing selection probabilities to depend on the screening test (Wang et al., 2012). Increased efficiency translates to cost for studies as well as decreased burden on subjects. We propose a semi-parametric pAUC estimator which incorporates a test-dependent sampling design and inclusion of un-sampled subjects. Simulation studies show that the proposed pAUC estimator is unbiased and improves efficiency compared to estimators using a simple random sample (SRS) design and those that use only information from the sampled subjects.

Using data from the Cancer and Leukemia Group B (CALGB) 150807 study, we evaluate the ability of the Balcone risk score to identify patients who are at greatest risk of non-small-cell lung cancer (NSCLC) recurrence by estimating the AUC while restricting the FPR to be within the interval $(0.1, 0.3)$ (Bueno et al., 2012). The ROC curve is a tool used to graphically display the ability of the Balcone risk score to identify patients who survive beyond 12 months by plotting the false positive rate (FPR, $Pr\left(Y \geq c|D = 0\right)$) against the true positive rate (TPR, $Pr\left(Y \geq c|D = 1\right)$). The pAUC is a summary measure used to describe the ability

of the Balcone risk score to discriminate between patients who survive beyond 12 months and those who do not while restricting the FPRs (or TPRs) to an interval that is clinically relevant. When surgery is used as the primary treatment for NSCLC, adjuvant chemotherapy may benefit patients who have a high risk of cancer recurrence. Identifying patients who are at high risk of cancer recurrence is important in order for treatment to be given to those who would benefit most. At the same time, we want a screening test that minimizes the false positives to reduce the number of patients that are subjected to potentially dangerous treatments that are unnecessary. The proposed methods are especially beneficial considering the length of time this type of study will follow patients and the cost of following a large number of subjects in this setting.

We also evaluate data from the Preterm Prediction Study to assess the utility of fetal fibronectin (fFN) in predicting spontaneous preterm birth while restricting the FPRs (or TPRs) to an interval that is clinically relevant (Goldenberg et al., 1996). Preterm birth (PTB), defined as delivery at less than 37 weeks of gestation, contributes to neonatal morbidity and mortality. The prevalence of adverse events increases as gestational age decreases (McCormick, 1985). This is an important area of study due to the negative impact of spontaneous PTB on maternal and child health outcomes. Knowing the fFN measurement will not change the incidence of spontaneous PTB, but it will affect the treatment plan.

Previous research has explored multiple approaches to estimating pAUC. McClish (1989) and Thompson and Zucchini (1989) introduced the idea of evaluating only part of the AUC for a specified FPR interval. A nonparametric pAUC estimator proposed by Dodd and Pepe (2003a), which is similar to the AUC estimator proposed by Bamber (1975), incorporates a restriction in the numerator for the FPR interval of interest. The estimator proposed by Dodd and Pepe (2003a) uses an SRS design for sampling subjects. Wang et al. (2012) proposed a nonparametric pAUC estimator using a test-dependent sampling (TDS) design. This estimator uses empirical likelihood methods to avoid making assumptions on the distribution of the screening test, Y. Weights are incorporated into the estimator to account for the biased sampling design. The TDS design is related to outcome-dependent sampling (ODS). Zhou et al. (2002) used the ODS design and empirical likelihood methods in regression modeling

to develop parameter estimates in which inclusion in the sample depends on a continuous outcome variable. Weaver and Zhou (2005) developed a semi-parametric estimator for regression coefficients using the ODS framework, which utilizes incomplete information for the un-sampled portion of the population. In the design proposed by Weaver and Zhou (2005), the outcome used to develop the ODS is observed for all subjects, but covariates are observed for subjects selected in the sample.

We propose the use of a test-dependent sampling (TDS) design where TDS inclusion depends on the continuous screening test measure fFN. The TDS design incorporates an SRS component, a TDS component, and the remaining un-sampled portion of the population. The TDS design allows investigators to over-sample subjects from specified ranges of the screening test variable, allowing for a concentration of resources where there is the greatest amount of information. All data are available for subjects sampled in the study, but only the screening test value is available for the un-sampled portion of the population. The proposed method gives improved estimates of the pAUC which are more efficient than those given by Dodd and Pepe (2003a) and Wang et al. (2012).

Verification bias is associated with the TDS design described by Wang et al. (2012) in that both designs generate bias. Many screening tests define results in dichotomous terms, "positive" for a test value above a threshold and "negative" for a test value below the threshold. Verification bias is a concern, mostly for cohort studies, where all subjects whose test result is "positive" for the outcome have their disease status verified (Pepe, 2004) but among subjects who test "negative" for the outcome, either none or a subset of this group have their disease status verified. For the TDS design described by Wang et al. (2012), subjects are selected conditional on their test result, but only a subset of subjects within groups are selected. In the proposed design, all subjects are included in estimating the AUC. Although disease status is missing for subjects in the un-sampled portion of the population, the inclusion of these subjects eliminates the bias typically associated with the TDS design.

This chapter is organized as follows. In Section 3.2, we introduce existing pAUC estimators, propose use of empirical likelihood methods to develop a semi-parametric estimator for pAUC. In Section 3.3, asymptotic properties of the proposed pAUC estimator are explored.

In Section 3.4, we use simulation studies to compare the proposed estimator with existing methods. In Section 3.5, we analyze data from the lung cancer study. In Section 3.6, we use the proposed method to analyze data from the Preterm Prediction Study. We conclude with a discussion in Section 3.7.

## 3.2 Semi-parametric empirical likelihood pAUC (SPEL-pAUC) estimation

### 3.2.1 Notation and data structure

Consider a continuous test variable, $Y$, and binary disease indicator, $D$. The distribution of $Y$ can be divided into $K$ mutually exclusive intervals defined by $C_k = (a_{k-1}, a_k]$ where $k = 1, \ldots, K$. The sample size within each of the $C_k$ intervals may be different. The TDS is made up of three components: the SRS component, TDS component, and the non-validation set (un-sampled subjects). This set of subjects sampled in the SRS and TDS components combined make up the validation set, indexed by $V$, where the true disease status is validated. The remainder of the population not selected for sampling makes up the non-validation set, indexed by $\bar{V}$. The sample size of the validation set is given by $n_V = n_0 + \sum_{k=1}^{K} n_k$, where $n_0$ is the sample size from the SRS component and $n_k$ is the sample size for the $k^{th}$ TDS interval, $k = 1, \ldots, K$. The size of the non-validation set is given by $n_{\bar{V}} = N - n_V$. To define subscripts, $i$ indexes the sampling group where $i = \left(0, 1, ..., K, \bar{V}\right)$ and $j = \{1, ..., n_i\}$ denotes the individual in the $i^{th}$ sampling group. The test variable, $Y$, is observed for all subjects in the dataset. The disease status, $D$, is only ascertained for subjects who are in the validation set. The data framework is given by

$$\begin{array}{lll} \text{SRS component} & (D_{0j}, Y_{0j}) & j = 1, ..., n_0 \\ \text{TDS}_1 \text{ component} & (D_{1j}, Y_{1j} | Y_{1j} \in C_1) & j = 1, ..., n_1 \\ \vdots & \vdots & \vdots \\ \text{TDS}_K \text{ component} & (D_{Kj}, Y_{Kj} | Y_{1j} \in C_K) & j = 1, ..., n_K \\ \text{Non-validation component} & (Y_{\bar{v}j} | i \neq (0, 1, \ldots, K)) & j = 1, ..., n_{\bar{V}}. \end{array} \tag{3.1}$$

### 3.2.2 Existing pAUC estimators

Two existing pAUC estimators are included in the simulation study to compare with the proposed pAUC estimator. These estimators are

1) the SRS only estimator (NP-pAUC) proposed by Dodd and Pepe (2003a), denoted $\hat{A}_{t:V}^{SRS}$, and

2) the nonparametric empirical likelihood estimator (NPEL-pAUC) proposed by Wang et al. (2012), denoted $\hat{A}_{t:V}^{TDS}$.

First, we introduce the NP-pAUC proposed by Dodd and Pepe (2003a). This nonparametric estimator utilizes an SRS structure and is given by

$$\hat{A}_{t:V}^{SRS} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} D_i \left(1 - D_i\right) I \left(Y_i > Y_j, Y_j \in (q_0, q_1)\right)}{\sum_{i=1}^{n} \sum_{j=1}^{n} D_i \left(1 - D_i\right)}. \tag{3.2}$$

where $q_0 = F_{Y|D=0}^{-1}\left(1 - t_1\right) = FPR^{-1}\left(t_1\right)$ and $q_1 = F_{Y|D=0}^{-1}\left(1 - t_0\right) = FPR^{-1}\left(t_0\right)$.

The second estimator under comparison, NPEL-pAUC, was proposed by Wang et al. (2012) and uses a TDS design, incorporating the SRS and TDS components where TDS inclusion depends on the continuous screening test variable, Y. Empirical likelihood methods were used to avoid making distributional assumptions on the screening test, Y. The data structure is similar to the structure described in Section 3.2.1, except that non-validation data are not included in the NPEL-pAUC. The sample size is given by $n = n_0 + \sum_{k=1}^{K} n_k$, where $n_0$ is the sample size from the SRS component and $n_k = \frac{n - n_0}{K}$ is the sample size for the $k^{th}$ TDS interval, $k = 1, \ldots, K$. This pAUC estimator incorporates a restriction for the FPR interval of interest, $(t_0, t_1)$, given by $\widehat{FPR}_j = \frac{\sum_i \hat{p}_i (1 - D_i) I(Y_i > Y_j)}{\sum_i \hat{p}_i (1 - D_i)}$. The pAUC estimator is given by

$$\hat{A}_{t:V}^{TDS} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \hat{p}_i \hat{p}_j D_i \left(1 - D_i\right) I \left(Y_i > Y_j, \widehat{FPR}_j \in (t_0, t_1)\right)}{\sum_{i=1}^{n} \sum_{j=1}^{n} \hat{p}_i \hat{p}_j D_i \left(1 - D_i\right)}, \tag{3.3}$$

where $\hat{p}_i = \left[n_0 + \sum_{k=1}^{K} \frac{n_k}{\hat{\theta}_k} I \left(y_i \in C_k\right)\right]^{-1}$. The biased sampling scheme is accounted for by incorporating weights $p_i$ and $p_j$ in the numerator and denominator. Empirical likelihood

methods were used in developing the estimator proposed by Wang et al. (2012).

### 3.2.3 Semi-parametric empirical likelihood approach

To develop the likelihood, denote $f_{Y,D}\left(Y_{ij}, D_{ij}\right)$ the joint distribution of disease status and screening test. The marginal distribution of the screening test variable is given by $f_Y\left(Y_{ij}\right)$. The distribution of disease status and screening test, conditional on $Y_{ij}$ falling in the interval $C_k$, is given by $f_{Y,D}\left(Y_{ij}, D_{ij}|Y_{ij} \in C_k\right)$, $k = (1, \ldots, K)$. Consider stratum sizes in the population $N_k = n_{0,k} + n_k + n_{\bar{V},k}$, where $n_{i,k} = \sum_{j=1}^{n_i} I(Y_i \in C_k)$ for $i = \left\{0, \bar{V}\right\}$ and $k = \{1, \ldots, K\}$. The validation portion of the likelihood is given by

$$\begin{aligned} L_V\left(f_D\right) &= \prod_{j=1}^{n_0} f\left(Y_{0j}, D_{0j}\right) \times \prod_{k=1}^{K} \prod_{j=1}^{n_k} f\left(Y_{kj} D_{kj} | Y_{kj} \in C_k\right) \\ &= \prod_{k=1}^{K} \prod_{j=1}^{n_k} Pr\left(Y_{kj} \in C_k\right)^{-n_k} \times \prod_{k=0}^{K} f\left(Y_{ij}|D_{kj}\right) Pr\left(D_{kj} = d\right). \end{aligned} \tag{3.4}$$

The non-validation portion of the likelihood takes into account the missing data, where the disease status is unknown, and is given by

$$\begin{aligned} L_{\bar{V}}\left(f_D\right) &= \frac{(N - n_0)!}{\prod_{k=1}^{K}\left(N_k - n_{0,k}\right)!} \prod_{k=1}^{K} Pr\left(Y_{kj} \in C_k\right)^{N_k - n_{0,k}} \prod_{\substack{j=1 \\ Y \in C_k}}^{n_{\bar{V}}} \frac{f\left(Y_{\bar{V}j}\right)}{Pr\left(Y_{kj} \in C_k\right)} \\ &= \frac{(N - n_0)!}{\prod_{k=1}^{K}\left(N_k - n_{0,k}\right)!} \times \prod_{j=1}^{n_{\bar{V}}} \sum_{d=0}^{1} f\left(Y_{\bar{V}j}|D_{\bar{V}j} = d\right) Pr\left(D_{\bar{V},j} = d\right) \\ &\quad \times \prod_{k=1}^{K} Pr\left(Y_{kj} \in C_k\right)^{-n_k}. \end{aligned} \tag{3.5}$$

The full likelihood is found by combining the validation and non-validation portions of the likelihood, (3.4) and (3.5), given by

$$\begin{aligned} L\left(\{q_{ij}\},\{r_{ij}\},p\right) &= L_V\left(f_D\right) \times L_{\bar{V}}\left(f_D\right) \\ &\propto \prod_{\substack{k,j\in V \\ D=1}} f\left(Y_{kj}|D_{kj}=1\right) Pr\left(D_{ij}=1\right) \prod_{\substack{k,j\in V \\ D=0}} f\left(Y_{kj}|D_{kj}=0\right) Pr\left(D_{kj}=0\right) \\ &\quad \times \prod_{j=1}^{n_{\bar{V}}} \sum_{d=0}^{1} f\left(Y_{\bar{V}j}|D_{\bar{V}j}=d\right) Pr\left(D_{\bar{V}j}=d\right) \\ &= \prod_{\substack{k,j\in V \\ D=1}} q_{kj} p \prod_{\substack{k,j\in V \\ D=0}} r_{kj}\left(1-p\right) \prod_{j=1}^{n_{\bar{V}}} \left[q_{\bar{V}j}p + r_{\bar{V}j}\left(1-p\right)\right], \end{aligned} \tag{3.6}$$

where $p = Pr\left(D=1\right)$, $q_{ij} = f\left(Y_{ij}|D_{ij}=1\right)$, and $r_{ij} = f\left(Y_{ij}|D_{ij}=0\right)$. The probability of having disease, $p$, is estimated from the SRS component of the TDS. Due to the relationship between $q_{ij}$ and $r_{ij}$ and the need to decrease computational load, Qin and Zhang (1997, 2003) propose a constraint given by $\frac{r_{ij}}{q_{ij}} = e^{\alpha+\beta y_{ij}}$. Consider the standard logistic regression model where $Pr\left(D=1|Y\right) = \frac{e^{m^*(Y)\boldsymbol{\alpha}}}{1+e^{m^*(Y)\boldsymbol{\alpha}}} = \psi\left(Y\right)$. The Bayes' rule gives $f\left(Y|D=1\right) = \frac{f(Y)Pr(D=1|Y)}{Pr(D=1)} = \frac{f(Y)\psi(Y)}{p}$. Similarly, $f\left(Y|D=0\right) = \frac{f(Y)(1-\psi(Y))}{1-p}$. Consider the ratio

$$\begin{aligned} \frac{r_{ij}}{q_{ij}} &= \frac{f\left(Y|D=0\right)}{f\left(Y|D=1\right)} \\ &= \left[\frac{f\left(Y\right)\left(1-\psi\left(Y\right)\right)}{1-p}\right] \times \left[\frac{p}{f\left(Y\right)\psi\left(Y\right)}\right] \\ &= \frac{p}{1-p}\left(\frac{1}{1+e^{m^*(Y)\boldsymbol{\alpha}}}\right)\left(\frac{e^{m^*(Y)\boldsymbol{\alpha}}}{1+e^{m^*(Y)\boldsymbol{\alpha}}}\right) \\ &= e^{m(Y)\boldsymbol{\alpha}}, \end{aligned}$$

which implies that $r_{ij} = q_{ij}e^{m(\mathbf{X}_{ij})\boldsymbol{\alpha}}$. Let $m\left(\mathbf{X}_{ij}\right)\boldsymbol{\alpha} = \alpha + \beta Y$. Applying this constraint to the log-likelihood gives

$$l\left(\{q_{ij}\},p,\alpha,\beta\right) \propto \sum_{ij} \ln q_{ij} + n_{V,D=0}\alpha + \beta \sum_{\substack{i,j\in V \\ D=0}} y_{ij} + \sum_{j=1}^{n_{\bar{V}}} \ln\left[p + e^{\alpha+\beta y_{\bar{V}j}}\left(1-p\right)\right] \tag{3.7}$$

Without loss of generality, the continuous screening test variable is partitioned into three

mutually exclusive intervals: $C_1 = (-\infty, a_1]$, $C_2 = (a_1, a_2]$, and $C_3 = (a_2, \infty)$. The TDS consists of an SRS of size $n_0$, a TDS component of size $n_1 + n_2 + n_3$, and the non-validation set of size $N - n_V$, where $n_V = n_0 + n_1 + n_2 + n_3$. Subjects are eligible to be sampled for the TDS groups based on their screening test result. For example, if a subject's test result is less than or equal to $a_1$ and $n_1 > 0$, the probability of being selected in the TDS component for $C_1$ is greater than zero.

In order to develop the semi-parametric empirical likelihood pAUC estimator (SPEL-pAUC), we first need to obtain estimates for $\{q_{ij}\}$, $\alpha$, and $\beta$. We accomplish this by using empirical likelihood methods outlined below. Once we have these parameter estimates, we estimate the expected value of disease for subjects in the non-validation set and we estimate the false positive rate for all subjects. We can then use this expected disease status and false positive rate estimates in the SPEL-pAUC for the non-validation subjects where the true disease status is missing.

To obtain estimates for $\alpha$ and $\beta$ the profile likelihood must be constructed. The distribution of the screening test conditional on disease status, $f(Y_{ij}|D_{ij} = 1)$, is not known or assumed. A robust estimator for the pAUC can be constructed without making these distributional assumptions by fixing $\alpha$ and $\beta$ and obtaining the empirical likelihood function of $F(Y_{ij}|D_{ij} = 1)$, with support at the observed values of $Y$. To maximize the likelihood, we estimate $\{\hat{q}_{ij}\} = f(Y_{ij}|D_{ij} = 1)$ under the following constraints:

$$\left\{ q_{ij} \geq 0, \sum_{i \in (0,1,2,3,\bar{V})} \sum_{j=1}^{n_i} q_{ij} = 1, \sum_{i \in (0,1,2,3,\bar{V})} \sum_{j=1}^{n_i} q_{ij} \left\{ e^{\alpha + \beta y_{ij}} - 1 \right\} = 0 \right\}. \tag{3.8}$$

A unique maximum for $\{q_{ij}\}$ exists under the constraints given in (3.8) if 1 is inside the convex hull of points $e^{\alpha+\beta y_{ij}}$ for all $(i, j)$ (Owen, 1988, 1990; Qin and Lawless, 1994). Lagrange

multipliers, $\lambda_1$ and $\lambda_2$, are used to derive the maximum over $\{\hat{q}_{ij}\}$. Consider the function

$$\begin{aligned} H \quad \propto \quad & \sum_{ij} \ln q_{ij} + n_{v,D=0}\alpha + \beta \sum_{\substack{i,j \in V \\ D=1}} y_{ij} + \sum_{j=1}^{n_{\bar{V}}} \ln \left[ p + e^{\alpha+\beta y_{\bar{V}j}} (1-p) \right] \\ & + \lambda_1 \left( 1 - \sum_{ij} q_{ij} \right) + N\lambda_2 \sum_{ij} q_{ij} \left\{ e^{\alpha+\beta y_{ij}} - 1 \right\}. \end{aligned} \tag{3.9}$$

Estimates $\{\hat{q}_{ij}\}$ and $\hat{\lambda}_1$ are found by taking the derivative of H with respect to $q_{ij}$, where H is given by (3.9), and setting the derivative equal to zero. The derivative of H is given by $\frac{\partial H}{\partial q_{ij}} = \frac{1}{q_{ij}} - \lambda_1 + N\lambda_2 \left\{ e^{\alpha+\beta y_{ij}} - 1 \right\}$. The estimate $\hat{\lambda}_1$ is found by evaluating $\sum_{ij} q_{ij} \frac{\partial H}{\partial q_{ij}} = N - \lambda_1 \sum_{ij} q_{ij} + N\lambda_2 \sum_{ij} q_{ij} \left\{ e^{\alpha+\beta y_{ij}} - 1 \right\} = 0$. By setting the derivative of H equal zero and solving for $q_{ij}$, we have

$$\hat{q}_{ij} = \frac{1}{N} \left[ 1 - \lambda_2 \left( e^{\alpha+\beta y_{ij}} - 1 \right) \right]^{-1}, \tag{3.10}$$

for $i \in (0, 1, 2, 3, \bar{V})$ and $j \in (1, \ldots, n_i)$.

*Profile log-likelihood*

The empirical profile log-likelihood is obtained by plugging the estimates $\hat{q}_{ij}$, given in (3.10), into (3.7). Denoting $pl(\xi)$ as the natural logarithm of the empirical profile likelihood, we have

$$\begin{aligned} pl(\xi) \quad \propto \quad & -\sum_{ij} \ln \left[ 1 - \lambda_2 \left( e^{\alpha+\beta y_{ij}} - 1 \right) \right] + n_{V,D=0}\alpha + \beta \sum_{\substack{i,j \in V \\ D=1}} y_{ij} \\ & + \sum_{j=1}^{n_{\bar{V}}} \ln \left[ p + e^{\alpha+\beta y_{\bar{V}j}} (1-p) \right]. \end{aligned} \tag{3.11}$$

The Newton-Raphson algorithm can be used to obtain $\hat{\xi}$, where $\xi = (\alpha, \beta, \lambda_2)$. These estimators, $\hat{\xi}$, are used in estimating the expected disease status. Disease status is unknown for the non-validation portion of the population. For these subjects, an estimate for the expected disease status is used in place a true disease status in the SPEL-pAUC. The expected value

of disease is given by

$$\begin{aligned} E(D_l) &= 1 * Pr(D_l = 1|Y_l) \\ &= \frac{f(D_l = 1, Y_l)}{f(Y_l)} \\ &= \frac{f(Y_l|D_l = 1) Pr(D_l = 1)}{f(Y_l|D_l = 1) Pr(D_l = 1) + f(Y_l|D_l = 0) Pr(D_l = 0)} \\ &= \frac{q_l p}{q_l p + r_l (1 - p)} \\ &= \frac{p}{p + e^{\alpha + \beta y_{ij}} (1 - p)}, \end{aligned} \tag{3.12}$$

since $r_l = q_l e^{\alpha + \beta y_l}$. An estimate of the expected disease status is found by plugging estimators $\hat{\xi}$ and $\hat{p}$ into (3.12). This estimate of expected disease status is given by

$$\widehat{E(D_l)} = \frac{\hat{p}}{\hat{p} + e^{\hat{\alpha} + \hat{\beta} y_l} (1 - \hat{p})}. \tag{3.13}$$

The pAUC measures the area under the ROC curve where we are interested only in the region where the FPR falls within $(t_0, t_1)$, such that $0 < t_0 < t_1 < 1$. This restriction is accounted for in the estimator by incorporating an estimate of the FPR along with the chosen $t_0$ and $t_1$ bounds. The false positive rate is estimated by

$$\widehat{FPR}_{l'} = \frac{\sum_l^N (1 - D_l^*) I(Y_l > Y_{l'})}{\sum_l^N (1 - D_l^*)}. \tag{3.14}$$

The SPEL-pAUC uses the information from both the validation and non-validation portions of the population. Estimated expected disease, give by (3.13), is used for non-validation subjects where true disease status is missing. Let $l = 1, ..., N$ index the entire population.

The SPEL-pAUC is given by

$$\hat{A}^{P}_{t:V,\bar{V}} = \frac{\sum_{l\neq l'}^{N} D_l^* \left(1-D_{l'}^*\right) I\left(Y_l > Y_{l'}, \widehat{FPR}_{l'} \in (t_0,t_1)\right)}{\sum_{l\neq l'}^{N} D_l^* \left(1-D_{l'}^*\right)},$$

$$\text{where } \widehat{FPR}_{l'} = \frac{\sum_{l}^{N} \left(1-D_l^*\right) I\left(Y_l > Y_{l'}\right)}{\sum_{l}^{N} \left(1-D_l^*\right)}$$

$$\text{and } D_l^* = \begin{cases} D_l & \text{if } l \in V \\ \widehat{E\left(D_l\right)} = \frac{\hat{p}}{\hat{p}+e^{\hat{\alpha}+\hat{\beta}y_l}(1-\hat{p})} & \text{if } l \in \bar{V} \end{cases}. \tag{3.15}$$

### 3.3 Asymptotic properties of the SPEL-pAUC

In this section we give the asymptotic properties of the SPEL-pAUC. Theorem 1 establishes the asymptotic normality of the SPEL-pAUC. Further detail is provided in the appendix, including asymptotic results for the components that make up the SPEL-pAUC.

Consider the U-process $U_N\left(A_t,\eta\right) = R_N\left(A_t,\eta\right) - E\left(R_N\left(A_t,\eta\right)\right)$. Let $R_N\left(A_t,\eta\right) = \frac{1}{N^2}$ $\times \sum_{i\neq j} D_i' \left(1-D_j'\right)\left(I_{t:ij} - A_t\right)$ where $I_{t:ij} = I\left(Y_i > Y_j, Y_j \in (t_0,t_1)\right)$ and $D_l' = \begin{cases} D_l & \text{if } l \in V \\ \widehat{E\left(D_l\right)} = \frac{p}{p+e^{\alpha+\beta y_l}(1-p)} & \text{if } l \in \bar{V} \end{cases}$. Using this U-process, we show that

$$\sqrt{N}\left(\hat{A}^{P}_{t:V,\bar{V}} - A_t\right) = -\left\{\frac{\partial E\left[R_N\left(A_t,\eta\right)\right]}{\partial A_t}\right\}^{-1} \sum_{i\in\left(0,1,2,3,\bar{V}\right)} \rho_i n_i^{-1/2} \sum_{j=1}^{n_i} Q_{ij},$$

where $Q_{ij}\left(\eta\right) = E\left(R_{(ij)(ij)'} + R_{(ij)'(ij)}\right) + \rho_i^{-1} \frac{\partial E R_N(A_t,\eta)}{\partial p}\left[\frac{-1}{n_0}\frac{\partial^2 l_{srs}(p)}{p^2}\right]^{-1} P_{0j} I\left(i=0\right)$ $+ \frac{\partial E R_N(A_t,\eta)}{\partial \xi}\left[\frac{-1}{N}\frac{\partial^2 pl(\xi)}{\partial \xi_i \partial \xi_{i'}}\right]^{-1} \mathbf{H}_{ij}\left(\eta\right)$.

***Theorem 1:*** Under general regularity conditions,

$$\sqrt{N}\left(\hat{A}^{P}_{t:V,\bar{V}} - A_t\right) \xrightarrow{d} \mathcal{N}(0,\Sigma_t), \tag{3.16}$$

where $\Sigma = \left[\frac{\partial E[R_N(A_t,\eta)]}{\partial A_t}\right]^{-2} \sum_{i\in\left(0,1,2,3,\bar{V}\right)} \rho_i var\left(Q_{ij}\right)$.

The asymptotic variance estimator

$$\hat{\Sigma}_t = \left[\frac{\partial E\left[R_N\left(A_t,\eta\right)\right]}{\partial A_t}\right]^{-2} \sum_{i\in\left(0,1,2,3,\bar{V}\right)} \hat{\rho}_i var\left(\hat{Q}_{ij}\right) \tag{3.17}$$

is obtained by replacing the large sample quantities in $\Sigma_t$ with their corresponding finite sample quantities.

### 3.3.1 Alternative estimation of the variance of the SPEL-pAUC

Standard error estimates were generated using the bootstrap method (Efron and Tibshirani, 1993). The following algorithm was applied to each of 1,000 iterations of the simulation.

1) From the generated sample population, we drew B independent bootstrap samples $s_1, s_2, ..., s_B$ of size $n_V$ with replacement, following the proposed sampling design in Section 3.2.

2) For each bootstrap sample we then computed the SPEL-pAUC, denoted $\hat{\theta}_b$ and given in (3.15), resulting in B values of the SPEL-pAUC.

3) The average of these bootstrap ROC estimators is given by $\bar{\hat{\theta}}_b = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b$ with variance $\hat{V}\left(\hat{\theta}_b\right) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}_b - \bar{\hat{\theta}}_b\right)^2$ and standard deviation given by $\sqrt{\hat{V}\left(\hat{\theta}_b\right)}$.

4) We repeated these steps 1,000 times in generating the simulation results.

The $\widehat{SE}$ displayed in the simulation results was found by taking the average of the standard deviation estimates across 1,000 independent iterations of the simulation.

## 3.4 Simulation study

We evaluate the behavior of the SPEL-pAUC under many situations to better examine its robustness. The simulation studies were conducted using R version 2.14. The data were generated under the model $Y = \beta_0 + D\beta_1 + \epsilon$, where $D = 1$ for diseased subjects and $D = 0$ for non-diseased subjects. For the following simulations, we generate data where $\epsilon \sim \mathcal{N}(0,1)$ and $D \sim Bernoulli\,(0.3)$. The population size used in simulations is $N = 2000$ and the

distribution of $Y$ is partitioned into three mutually exclusive sets given by $C_1 = (-\infty, \hat{q}_0]$, $C_2 = (\hat{q}_0, \hat{q}_1]$, and $C_3 = (\hat{q}_1, \infty)$. The values $\hat{q}_0$ and $\hat{q}_1$ are the empirical quantiles of the SRS component in the TDS data. Here, $\hat{q}_1$ corresponds to $(1 - t_1) \times 100\%$ of the SRS component data falling below $\hat{q}_1$. In the following simulations, we consider the impact of 1) varying overall sample size, $n_V$; 2) varying proportion of SRS to TDS component sizes, $\frac{n_0}{n_V}$; and 3) varying model parameter $\beta_1$. The variations under consideration are: validation sample size ($n_V$) 120, 240, and 360; proportion of SRS subjects among validation set ($\frac{n_0}{n_V}$) 0.25, 0.5, and 0.75; and FPR intervals. For simulation results, the estimated means, standard errors, mean of the standard error estimates, and 95% nominal coverage probabilities for an estimator are obtained from 1000 independent runs. Estimated standard errors were obtaining with bootstrapping, using 50 replicate samples.

*Sample allocation for TDS*

The SPEL-pAUC uses the TDS design to target subjects on the left tail of the distribution. This sample consists of the following three components:

1) the SRS component of size $n_0$,

2) the TDS component of size $n - n_0$ where all subjects are sampled from the left-most interval, depending on the FPR interval of interest, and $\frac{n_0}{n_V} = 0.5$. For example, if we are interested in FPR $\in (0, 0.1)$, then $n_1 = 0$ subjects are sampled such that $Y_{1j} \in C_1$, $n_2 = \frac{1}{2}n_V$ subjects are sampled such that $Y_{2j} \in C_2$, and $n_3 = 0$ subjects are sampled such that $Y_{3j} \in C_3$. Whereas, if we are interested in FPR $\in (0.1, 0.2)$, then $n_1 = \frac{1}{2}n_V$ subjects are sampled such that $Y_{1j} \in C_1$, $n_2 = 0$ subjects are sampled such that $Y_{2j} \in C_2$, and $n_3 = 0$ subjects are sampled such that $Y_{3j} \in C_3$, and

3) the non-validation component of size $N - n_V$, comprised of all subjects not sampled into the SRS and TDS components.

For the NPEL-pAUC, proposed by Wang et al. (2012), the sample consists of the same SRS and TDS components, except that in the TDS component, the subjects are allocated equally

across all intervals. For interest in FPR $\in (0, 0.1)$, $n_1 = 0$ and $n_2 = n_3 = \frac{1}{2}(n - n_0)$. For interest in FPR $\in (0, 0.1)$, $n_1 n_2 = n_3 = \frac{1}{3}(n - n_0)$.

*Estimators to be compared*

The proposed estimator, SPEL-pAUC in (3.15), is compared to three estimators in the simulation studies. These estimators are given below. Specifically, under each setting, we compare the following four estimators.

1) NP-pAUC: the SRS only estimator Dodd and Pepe (2003a), denoted by $\hat{A}^{SRS}_{t:V}$, is given by (3.2).

2) SPEL-pAUC(SRS): the SRS with validation and non-validation data estimator, denoted $\hat{A}^{SRS}_{t:V,\bar{V}}$, has the same form as the SPEL-pAUC in (3.15). The difference between this estimator and the SPEL-pAUC is the sampling scheme. This gives a comparison of the SRS and TDS designs while incorporating non-validation data.

3) NPEL-pAUC: the TDS data only estimator (Wang et al., 2012), denoted $\hat{A}^{TDS}_{t:V}$, is given by (3.3).

4) SPEL-pAUC: the proposed TDS with validation and non-validation data estimator, denoted $\hat{A}^{P}_{t:V,\bar{V}}$, is given by (3.15).

*Results*

*Unbiasedness* All four AUC estimators yield unbiased estimates. To illustrate this, we simulated data using multiple allocations, sample sizes, and FPR intervals. Tables 3.1 and 3.3 show that the average of all AUC estimators are close to or equal to the true value. Other allocation schemes were considered for the SPEL-pAUC in Table 3.2 by varying the proportion of subjects allocated to the SRS component ($\frac{n_0}{n_V} = (0.25, 0.5, 0.75)$) and the TDS component allocation. In all of the allocations considered, the SPEL-pAUC continues to be unbiased.

*Efficiency* The SPEL-pAUC is the most efficient among pAUC estimators compared and the NP-pAUC is the least efficient among those compared. This supports the idea that both the use of the TDS design and inclusion of non-validation subjects create a more efficient

alternative to the SRS design and validation-only estimators while sampling the same number of subjects. In most cases for the chosen allocation ($\frac{n_0}{n_V} = 0.5$ and allocation to left-most interval), shown in Table 3.2, the bootstrapped standard errors are equal to the standard error of the estimator. For FPR$\in (0, 0.2)$, for example, the first three lines of results show that for sample sizes 120, 240, and 360, the bootstrapped standard errors are 0.014, 0.010, 0.008, respectively, which equals the standard error of the SPEL-pAUC estimator. For FPR$\in (0.1, 0.5)$, the first line gives results for $n_V = 120$ with a bootstrapped standard error 0.027 compared to the standard error of 0.029. As the sample size increases, the bootstrapped standard error does a better job of estimating the standard error of the SPEL-AUC. Table 3.2 shows that the standard error is only slightly effected for FPR$\in (0, 0.2)$ when alternate allocations are considered. Whereas for FPR$\in (0.1, 0.5)$ the chosen allocation is the smallest standard error at 0.015 for $n_V = 360$, compared to 0.016, 0.017, and 0.018 for alternate allocations.

*Robustness* The SPEL-pAUC does not require model specification for the screening test, Y. To explore the SPEL-pAUC's robustness, simulation studies were generated using both Normal and Chi-squared distributions for the screening test. Simulation results reported in Table 3.3 show that the simulation study gives similar results when the screening test is generated using a Chi-squared distribution, as far as unbiasedness of the estimators and efficiency of the SPEL-pAUC.

## 3.5 Analysis of the lung cancer study data

We used the SPEL-pAUC to analyze non-small-cell lung cancer (NSCLC) data from the CALGB 150807 study (Bueno et al., 2012) conducted by the Cancer and Leukemia Group B. This study is a subset of patients registered in the CALGB 140202 study who have stage 1A or 1B non-small-cell lung cancer (NSCLC), have not received preoperative chemotherapy or radiation, and are not missing histological, demographic, clinical, and follow-up information of interest. Among patients in the CALGB 150807 study, 1,061 patients were not censored before 12 months and were used in this analysis.

Lung cancer is the most common cause of cancer death among men and women in the world (Blanchon et al., 2006). Lung cancer is classified as either small-cell lung carcinoma (SCLC) or non-small-cell lung carcinoma (NSCLC), of which NSCLC accounts for approximately 80% of all lung cancers. After surgical lung resection, a large proportion of stage 1 NSCLC patients have cancer recurrence within five years (Bueno et al., 2012). When surgery is used as the primary treatment for NSCLC, adjuvant chemotherapy may benefit patients who have a high risk of cancer recurrence. Identifying patients who are at high risk of cancer recurrence is important in order for treatment to be given to those who would benefit most. This is an important area of study for patients, families, and doctors when making decisions on a treatment plan.

The Balcone risk score, outlined by Blanchon et al. (2006), has been developed to identify patients who are at greatest risk of cancer recurrence. To select the variables that are included in the scoring algorithm, Blanchon et al. (2006) used a Cox model to identify variables that were independently associated with mortality. The associated variables were then weighted to create the Balcone risk index. The components of the risk score are given by: age (>70 years, 1 point); sex (male, 1 point); performance status at diagnosis (reduced activity, 3 points; active >50%, 5 points; inactive >50%, 8 points; and total incapacity, 10 points); histological type (large-cell carcinoma, 2 points); and tumour-node-metastasis (TNM) staging system (IIA or IIB, 3 points; IIIA or IIIB, 6 points; and IV, 8 points). For the data used in this analysis, the Balcone risk score ranges from 0 to 15.

The goal of this analysis is to summarize the ability of the Balcone risk score to distinguish between patients who survive beyond 12 months and those who do not when the FPR falls within $(0.1, 0.3)$. This FPR range allows us to assess the utility of the SPEL-pAUC when the FPR is low. Clinicians may be less interested in the Balcone risk score at higher FPR values because this would suggest that patients are being falsely identified as high risk for cancer recurrence, leading to treatment plans that may be dangerous and unnecessary. The outcome of interest is survival beyond 12 months and the screening test is the Balcone risk score. Although all information is available for these patients, we illustrate the utility of the proposed AUC estimator by sampling from the study data and evaluating the estimated

AUC. Table 3.4 gives descriptive statistics for the Balcone risk score, stratified by survival at 12 months. A sample size of $n_V = 360$ was used for all estimators compared and details on sample allocation for each estimator are given in Table 3.5. The NP-pAUC has an SRS of size $n = 360$. The SPEL-pAUC(SRS) allocates 100% of the sample to the SRS component and utilizes incomplete data from the non-validation set. The NPEL-pAUC allocates 50% of the validation sample to the SRS component and the remaing 50% are allocated equally between the three intervals, $C_i$, such that $n_1 = n_2 = n_3 = 60$. The SPEL-pAUC allocates 50% of the sample to the SRS component and samples the remaining 50% to the left tail, where $n_1 = 180$ and $n_2 = n_3 = 0$, while utilizing incomplete data from the non-validation set. Because survival at 12 months is known for all subjects, we use the NP-pAUC estimator to evaluate the pAUC using complete information, given by 0.150. The pAUC estimates are: NP-pAUC 0.132, SPEL-pAUC(SRS) 0.158, NPEL-pAUC 0.149, and SPEL-pAUC 0.147. All estimators compared have an estimated pAUC that is similar to the best estimate we have for the true pAUC, which is 0.15, found by using the complete data.

## 3.6 Analysis of the Preterm Prediction Study data

We used the SPEL-pAUC to analyze data from the Preterm Prediction Study, a multi-center prospective study designed to study spontaneous preterm birth (Goldenberg et al., 1996). The Maternal Fetal Medicine Units Network of the National Institute of Child Health and Human Development carried out this study using ten centers to recruit women. There were 3073 women recruited between October 1992 and July 1994. Measurements were collected every two weeks from 22 to 30 weeks' gestation. Among the 3073 women recruited, 3001 had valid measurements of interest for this analysis.

PTB is defined as delivery at less than 37 weeks of gestation, contributes to neonatal morbidity and mortality which increases as gestational age decreases (McCormick, 1985). Bastek and Elovitz (2013) combined results from multiple studies on this topic to gain a better understanding the relationship between biomarkers and PTB. The results were not definitive for most biomarkers with the exception of fetal fibronectin (FFN).

Fetal fibronectin (FFN) is a protein that is produced by the fetal membrane. Knowing the FFN will not change the incidence of spontaneous PTB but it will affect the treatment plan. Deshpande et al. (2013) found that FFN has moderate accuracy in predicting PTB. Although many studies are concerned with the ability to predict spontaneous PTB, Bastek and Elovitz (2013) suggest that the ability to predict those who will not have spontaneous PTB is also valuable. Because FFN typically has a high negative predictive value (proportion of true negatives over all who test negative), a negative FFN test is widely used in clinical practice to send patients home. Measurable levels of FFN are considered to be abnormal between 20 and 37 weeks' gestation. Lockwood et al. (1991) show that in 588 FFN samples from uncomplicated pregnancies, a higher percentage of subjects were positive for FFN (level above 0.05 $\mu$g/mL) before 22 and after 37 weeks' gestation compared to between 22 and 37 weeks' gestation. For example the percentage of cervical samples with positive FFN for <22, 22 to 37 and >37 weeks' gestation were 24%, 4%, and 32%, respectively. This is an important area of study due to the negative effects of spontaneous PTB on maternal and child health outcomes.

In our analysis, the outcome of interest is spontaneous PTB at less than 37 weeks' gestation and the screening test considered in FFN. Values of the screening test that are associated with high FPRs are not of interest clinically and very low FPRs may not be realistic. Because of this, we are interested in estimating the pAUC where we restrict the FPR interval to $(0.1, 0.5)$. Table 3.6 gives descriptive statistics for FFN, stratified by spontaneous PTB. A sample size of $n_V = 360$ was used for all estimators compared and details on sample allocation for each estimator are given in Table 3.7. The NP-pAUC has an SRS of size $n = 360$. The SPEL-pAUC(SRS) allocates 100% of the sample to the SRS component and utilizes incomplete data from the non-validation set. The NPEL-pAUC allocates 50% of the validation sample to the SRS component and the remaing 50% are allocated equally between the three intervals, $C_i$, such that $n_1 = n_2 = n_3 = 60$. The SPEL-pAUC allocates 50% of the sample to the SRS component and samples the remaining 50% to the left tail, where $n_1 = 180$ and $n_2 = n_3 = 0$, while utilizing incomplete data from the non-validation set. Because the outcome of spontaneous PTB is known for all subjects, we use the NP-pAUC estimator to

evaluate the pAUC using complete information, given by 0.177. The pAUC estimates are: NP-pAUC 0.189, SPEL-pAUC(SRS) 0.133, NPEL-pAUC 0.157, and SPEL-pAUC 0.126. All estimators compared have an estimated pAUC that is similar to the best estimate we have for the true pAUC, which is 0.177, found by using the complete data.

## 3.7 Discussion

We have proposed a semi-parametric estimator for the partial area under the ROC curve (pAUC), which allows us to summarize the ability of a screening test to discern between diseased and non-diseased subjects while restricting the FPR to a range that is clinically relevant. This estimator incorporates a TDS design and includes both validation and non-validation data. The use of empirical likelihood methods allows us to estimate the pAUC without specifying a distribution for the screening test. We use bootstrapping to estimate the standard error and simulation studies show good coverage probabilities when using the standard error estimated using bootstrap methods.

The proposed design is motivated by the need to improve efficiency in estimating pAUC. Ascertaining true disease status can be costly and invasive for subjects. Although disease status is missing for the non-validation set, the proposed estimator takes advantage of all information available for a larger number of subjects than the validation-only estimators. Although all estimators are unbiased, the proposed estimator was shown to be the most efficient, compared to three competing pAUC estimators. This suggests that to obtain the same variability less subjects would be needed when the SPEL-pAUC is used, reducing study cost and subject burden. These results support the idea that both the use of the TDS design and inclusion of non-validation subjects create a more efficient alternative to the SRS design and the validation-only estimators while sampling the same number of subjects. For example, with a sample size of 120 and FPR interval of $(0, 0.02)$, the standard error of the proposed estimator is 0.014 compared to 0.018 and 0.016 in the competing methods (Table 3.1). In this case, we have an estimated standard error of 0.014 which gives coverage proportion of 0.952. The SPEL-pAUC is also robust in its ability to estimate pAUC under varying

distributions. Simulation studies show that when the screening test is simulated from a Chi-squared distribution, the pAUC estimators are unbiased and the SPEL-pAUC continues to be the most efficient pAUC estimator under comparison.

Table 3.1: Comparison of SPEL-pAUC and competing methods

| | $n_V = 120$ | | $n_V = 240$ | | $n_V = 360$ | |
|---|---|---|---|---|---|---|
| Method | Mean | SE | Mean | SE | Mean | SE |
| FPR$\in (0, 0.2)$ and True pAUC=0.0727 | | | | | | |
| $\hat{A}^{SRS}_{t:\bar{V}}$ | 0.076 | 0.018 | 0.074 | 0.013 | 0.074 | 0.010 |
| $\hat{A}^{SRS}_{t:V,\bar{V}}$ | 0.074 | 0.016 | 0.073 | 0.011 | 0.073 | 0.008 |
| $\hat{A}^{TDS}_{t:\bar{V}}$ | 0.074 | 0.016 | 0.075 | 0.012 | 0.074 | 0.009 |
| $\hat{A}^{P}_{t:V,\bar{V}}$ | 0.074 | 0.014 | 0.073 | 0.010 | 0.073 | 0.008 |
| FPR$\in (0.1, 0.5)$ and True pAUC=0.2646 | | | | | | |
| $\hat{A}^{SRS}_{t:\bar{V}}$ | 0.265 | 0.033 | 0.263 | 0.023 | 0.265 | 0.019 |
| $\hat{A}^{SRS}_{t:V,\bar{V}}$ | 0.266 | 0.030 | 0.265 | 0.021 | 0.264 | 0.016 |
| $\hat{A}^{TDS}_{t:\bar{V}}$ | 0.266 | 0.030 | 0.263 | 0.022 | 0.264 | 0.018 |
| $\hat{A}^{P}_{t:V,\bar{V}}$ | 0.266 | 0.029 | 0.265 | 0.018 | 0.263 | 0.015 |

Screening test data is simulated assuming $Y_{D=1} \sim \mathcal{N}(1, 1)$ and $Y_{D=0} \sim \mathcal{N}(0, 1)$. $\hat{A}^{SRS}_{t:\bar{V}}$ denotes the NP-pAUC which uses SRS; $\hat{A}^{SRS}_{t:V,\bar{V}}$ denotes the SPEL-pAUC(SRS) which uses SRS and utilizes information for both validation and non-validation data; $\hat{A}^{TDS}_{t:\bar{V}}$ denotes the NPEL-pAUC which uses TDS; and $\hat{A}^{P}_{t:V,\bar{V}}$ denotes the proposed SPEL-pAUC which uses TDS and utilizes information from both validation and non-validation data. All estimators sample the same number of subjects.

Table 3.2: Properties of the SPEL-pAUC

| $n_V$ | $\frac{n_0}{n_V}$ | $(n_0, n_1, n_2, n_3, n_{\bar{V}})$ | Mean | SE | $\widehat{SE}$ | CP |
|---|---|---|---|---|---|---|
| FPR$\in (0, 0.2)$ and True pAUC=0.0727 | | | | | | |
| | | $n_1 = n_3 = 0$ | | | | |
| 120 | 0.5 | $(60, 0, 60, 0, 1880)$ | 0.074 | 0.014 | 0.014 | 0.952 |
| 240 | 0.5 | $(120, 0, 120, 0, 1760)$ | 0.073 | 0.010 | 0.010 | 0.944 |
| 360 | 0.5 | $(180, 0, 180, 0, 1640)$ | 0.073 | 0.008 | 0.008 | 0.950 |
| | | $n_1 = n_3 = 0$ | | | | |
| 360 | 0.25 | $(90, 0, 270, 0, 1640)$ | 0.072 | 0.009 | 0.010 | 0.951 |
| 360 | 0.75 | $(270, 0, 90, 0, 1640)$ | 0.073 | 0.008 | 0.008 | 0.948 |
| | | $n_1 = n_2 = n_3$ | | | | |
| 360 | 0.5 | $(180, 0, 90, 90, 1640)$ | 0.073 | 0.008 | 0.008 | 0.950 |
| FPR$\in (0.1, 0.5)$ and True pAUC=0.2646 | | | | | | |
| | | $n_2 = n_3 = 0$ | | | | |
| 120 | 0.5 | $(60, 60, 0, 0, 1880)$ | 0.265 | 0.029 | 0.027 | 0.926 |
| 240 | 0.5 | $(120, 120, 0, 0, 1760)$ | 0.265 | 0.018 | 0.018 | 0.953 |
| 360 | 0.5 | $(180, 180, 0, 0, 1640)$ | 0.263 | 0.015 | 0.016 | 0.955 |
| | | $n_2 = n_3 = 0$ | | | | |
| 360 | 0.25 | $(90, 270, 0, 0, 1640)$ | 0.263 | 0.018 | 0.018 | 0.957 |
| 360 | 0.75 | $(270, 90, 0, 0, 1640)$ | 0.265 | 0.016 | 0.017 | 0.962 |
| | | $n_1 = n_2 = n_3$ | | | | |
| 360 | 0.5 | $(180, 60, 60, 60, 1640)$ | 0.265 | 0.017 | 0.016 | 0.937 |

Screening test data is simulated assuming $Y_{D=1} \sim \mathcal{N}(1, 1)$ and $Y_{D=0} \sim \mathcal{N}(0, 1)$. The fraction $\frac{n_0}{n_V}$ is the proportion of subjects allocated to the SRS component out of the total number of validation subjects sampled. The sample allocation, $(n_0, n_1, n_2, n_3, n_{\bar{V}})$, gives the number of subjects allocated to the SRS component, three intervals of the TDS component, and the non-validation set, respectively.

Table 3.3: Comparison of SPEL-pAUC and competing methods for Chi-Squared Distributed Data

| | FPR$\in (0, 0.2)$ | | FPR$\in (0.1, 0.5)$ | |
|---|---|---|---|---|
| Method | Mean | SE | Mean | SE |
| $\hat{A}^{SRS}_{t:V}$ | 0.092 | 0.018 | 0.291 | 0.032 |
| $\hat{A}^{SRS}_{t:V,\bar{V}}$ | 0.096 | 0.016 | 0.286 | 0.026 |
| $\hat{A}^{TDS}_{t:V}$ | 0.092 | 0.017 | 0.293 | 0.029 |
| $\hat{A}^{P}_{t:V,\bar{V}}$ | 0.101 | 0.016 | 0.307 | 0.024 |

Screening test data is simulated assuming $Y_{D=1} \sim \chi(4, 3)$ and $Y_{D=0} \sim \chi(3)$ and $n_V = 120$. The true pAUC for FPR$\in (0, 0.2)$ is 0.0901 and for FPR$\in (0.1, 0.5)$ is 0.2921.

Table 3.4: Descriptive statistics for the Balcone risk score

| | N | Minimum | Q1 | Median | Q3 | Maximum |
|---|---|---|---|---|---|---|
| Overall | 1076 | 0 | 1 | 1 | 3 | 10 |
| Survival beyond 12 months | 965 | 0 | 1 | 1 | 3 | 10 |
| Survival less than 12 months | 111 | 0 | 1 | 2 | 5 | 9 |

Table 3.5: Sample allocation for the non-small-cell lung cancer data

| Component | MW-AUC | SPEL-AUC(SRS) | NPEL-AUC | SPEL-AUC |
|---|---|---|---|---|
| SRS | 360 | 360 | 180 | 180 |
| TDS $(n_1, n_2, n_3)$ | $(0, 0, 0)$ | $(0, 0, 0)$ | $(60, 60, 60)$ | $(180, 0, 0)$ |
| non-validation | 0 | 701 | 0 | 701 |

Table 3.6: Descriptive statistics for FFN

| | N | Minimum | Q1 | Median | Q3 | Maximum |
|---|---|---|---|---|---|---|
| Spontaneous PTB | 309 | 0 | 0.88 | 4.48 | 17.08 | 924.56 |
| Not PTB | 2692 | 0 | 0.28 | 2.61 | 7.22 | 2151.44 |

Table 3.7: Sample allocation for the Preterm Prediction Study

| Component | NP-pAUC | SPEL-pAUC(SRS) | NPEL-pAUC | SPEL-pAUC |
|---|---|---|---|---|
| SRS | 360 | 360 | 180 | 180 |
| TDS $(n_1, n_2, n_3)$ | $(0, 0, 0)$ | $(0, 0, 0)$ | $(60, 60, 60)$ | $(180, 0, 0)$ |
| non-validation | 0 | 2641 | 0 | 2641 |

The FPR interval of interest is $(0.1, 0.5)$.

# Chapter 4

# Covariate-specific ROC Curve under Test-Dependent Sampling

## 4.1 Introduction

The receiver operating characteristic (ROC) curve is a summary measure used to describe the ability of a screening test to discriminate between diseased and non-diseased subjects (Bamber, 1975). Consider a screening test, Y, and disease or outcome, D, where D=1 indicates presence of the disease and D=0 indicates no disease. The ROC curve is constructed by plotting the false positive rate (FPR, $Pr\left(Y \geq c|D=0\right)$) against the true positive rate (TPR, $Pr\left(Y \geq c|D=1\right)$), where c is the threshold for the screening test to indicate disease. As evaluating the true disease status can be costly, it is important for researchers to increase study efficiency by allowing selection probabilities to depend on the screening test (Wang et al., 2012). Increased efficiency translates to cost and time savings for studies as well as decreased burden on subjects. Incorporating covariates into the ROC curve estimator allows for evaluation of the utility of the screening test for different subsets of a population. We propose a semi-parametric covariate-specific ROC curve estimator, which incorporates a test-dependent sampling design and inclusion of un-sampled subjects. Simulation studies show that the proposed ROC curve estimator is unbiased and improves efficiency compared to estimators using a simple random sample (SRS) design and those that use only information from the sampled subjects.

Using data from the Cancer and Leukemia Group B (CALGB) 150807 study, we evaluate the ability of the Balcone risk score to identify patients who are at greatest risk of non-small-cell lung cancer (NSCLC) recurrence by estimating the covariate-specific ROC curve (Bueno et al., 2012). When surgery is used as the primary treatment for NSCLC, adjuvant

chemotherapy may benefit patients who have a high risk of cancer recurrence. Identifying patients who are at high risk of cancer recurrence is important in order for treatment to be given to those who would benefit most. We can evaluate the utility of the Balcone risk score in predicting survival at 12 months for specific values of a covariate. Including covariates, such as age and gender, in the model will help of identify subsets of the population where the screening test is more effective at predicting survival at 12 months. This is an important area of study for patients, families, and doctors when making decisions on a treatment plan. The proposed methods are especially beneficial considering the length of time this type of study will follow patients and the cost of following a large number of subjects in this setting.

We also use data from the Preterm Prediction Study to evaluate the ability of fetal fibronectin (FFN) to predict spontaneous preterm birth by estimating the covariate-specific ROC curve, while incorporating information from un-sampled subjects and including cervical length, maternal age, and previous PTB (Goldenberg et al., 1996). Preterm birth (PTB), defined as delivery at less than 37 weeks of gestation, contributes to neonatal morbidity and mortality. The prevalence of adverse events increases as gestational age decreases (McCormick, 1985). We can consider covariates, such a cervical length, when estimating the ROC curve to identify subset of the study population in which FFN is a better predictor of PTB. This is an important area of study due to the negative impact of spontaneous PTB on maternal and child health outcomes. Knowing the FFN measurement will not change the incidence of spontaneous PTB, but it will affect the treatment plan. The use of covariates allows us to better understand the influence covariates have on accuracy of this screening test (Wang et al., 2013).

Methods have been developed which consider the effect of covariate information on ROC curves and summary measures area under the ROC curve (AUC) and partial AUC (pAUC). Thompson and Zucchini (1989) and Dodd and Pepe (2003a) proposed direct estimation of the AUC, and Dodd and Pepe (2003a) proposed direct estimation of the pAUC, while accounting for covariates. Pepe (2000) and Cai and Pepe (2002) used generalized linear modeling methods to estimate the covariate-adjusted ROC curve. An alternative to direct ROC curve estimation is to model the screening test variable as a function of covariates and the disease

status. This approach has been used by Tosteson and Begg (1988) and Wang et al. (2013). These methods make parametric assumptions for the screening test. Wang et al. (2013) proposed the use of test-dependent sampling (TDS) in which inclusion in the sample depends on the continuous screening test measure. The TDS design is related to outcome-dependent sampling (ODS). Zhou et al. (2002) used the ODS design and empirical likelihood methods in regression modeling to develop parameter estimates where inclusion in the sample depends on a continuous outcome variable. Weaver and Zhou (2005) developed a semi-parametric estimator for regression coefficients using the ODS framework, which utilizes incomplete information for the un-sampled portion of the population. In the design proposed by Weaver and Zhou (2005), the outcome used to develop the ODS is observed for all subjects, but covariates are missing for the un-sampled portion of the population.

We propose the use of a test-dependent sampling (TDS) design in which TDS inclusion depends on the continuous screening test measure. The TDS design incorporates an SRS component, a TDS component, and the remaining un-sampled portion of the population. The TDS design allows investigators to over-sample subjects from specified ranges of the screening test variable, allowing for a concentration of resources where there is the greatest amount of information. All data are available for subjects sampled in the study, but only the screening test value and covariates are available for the un-sampled portion of the population. The proposed method gives improved estimates of the covariate-specific ROC curve which are more efficient than methods which utilize only the sampled subjects.

This chapter is organized as follows. In Section 4.2, we introduce an alternative covariate-specific ROC curve estimator and propose use of empirical likelihood methods to develop a semi-parametric estimator for the covariate-specific ROC curve. In Section 4.3, a variance estimator for the proposed covariate-specific ROC curve estimator is described. In Section 4.4, we use simulation studies to compare the proposed estimator with competing methods. In Section 4.5, we analyze data from the lung cancer study. In Section 4.6, we use the proposed method to analyze data from the Preterm Prediction Study. We conclude with a discussion in Section 4.7.

## 4.2 Semi-parametric empirical likelihood ROC curve (SPEL-ROC) estimation

### 4.2.1 Notation and data structure

Consider a continuous test variable, $Y$, a vector of covariates, $\mathbf{X}$, and binary disease indicator, $D$. The distribution of $Y$ can be divided into $K$ mutually exclusive intervals defined by $C_k = (a_{k-1}, a_k]$ where $k = 1, \ldots, K$. The sample size within each of the $C_k$ intervals may be different. The TDS is made up of 3 components: the SRS component, TDS component, and the non-validation set (un-sampled subjects). The subjects sampled in the SRS and TDS components combined make up the validation set, indexed by $V$, where the true disease status is validated. The remainder of the population not selected for sampling makes up the non-validation set, indexed by $\bar{V}$. The sample size of the validation set is given by $n_V = n_0 + \sum_{k=1}^{K} n_k$, where $n_0$ is the sample size from the SRS component and $n_k$ is the sample size for the $k^{th}$ TDS interval, $k = 1, \ldots, K$. The size of the non-validation set is given by $n_{\bar{V}} = N - n_V$. To define subscripts, $i$ indexes the sampling group where $i = \left(0, 1, ..., K, \bar{V}\right)$ and $j = \{1, ..., n_i\}$ denotes the individual in the $i^{th}$ sampling group. The test variable, $Y$, and covariates, $\mathbf{X}$, are observed for all subjects in the dataset. The disease status, $D$, is only ascertained for subjects who are in the validation set. The data framework is given by

$$\begin{array}{ccc}
\text{SRS component} & (D_{0j}, Y_{0j}, \mathbf{X}_{0j}) & j = 1, ..., n_0 \\
\text{TDS}_1 \text{ component} & (D_{1j}, Y_{1j}, \mathbf{X}_{1j} | Y_{1j} \in C_1) & j = 1, ..., n_1 \\
\vdots & \vdots & \vdots \\
\text{TDS}_K \text{ component} & (D_{Kj}, Y_{Kj}, \mathbf{X}_{Kj} | Y_{Kj} \in C_K) & j = 1, ..., n_K \\
\text{Non-validation component} & \left(Y_{\bar{v}j}, \mathbf{X}_{\bar{v}j} | i = \bar{V}\right) & j = 1, ..., n_{\bar{V}}.
\end{array} \tag{4.1}$$

### 4.2.2 Alternative ROC curve estimator

The alternative covariate-specific ROC estimator (LS-ROC) included in the simulation studies uses a binormal ROC model, discussed in Pepe (2004) and Wang et al. (2013), and

samples subjects with an SRS design. The model is given by

$$Y = \mathbf{X}\boldsymbol{\beta} + \sigma(D)\,\epsilon = \beta_0 + \beta_1 D + \beta_2 \mathbf{X} + \beta_3 \mathbf{X} D + \sigma(D)\,\epsilon \tag{4.2}$$

where $\epsilon \sim \mathcal{N}(0,1)$ and $\sigma(D) = \sigma_1 D + \sigma_0(1-D)$. Parameter estimates are found using the ordinary least squares method where $\hat{\boldsymbol{\beta}} = \left[\mathbf{X}^T\mathbf{X}\right]^{-1}\mathbf{X}^T y$ and the variances, $\sigma_1^2$ and $\sigma_0^2$, are found by calculating the sample variance for the respective outcome groups, $D = 1$ and $D = 0$. The covariate-specific survival function is estimated for the diseased and non-diseased groups for a chosen covariate value, giving $\hat{S}_{1X} = \Phi\left(\frac{\hat{\mu}_{1X}-c}{\sqrt{\hat{v}_1}}\right)$ and $\hat{S}_{0X} = \Phi\left(\frac{\hat{\mu}_{0X}-c}{\sqrt{\hat{v}_0}}\right)$ where $\hat{\mu}_{1X} = \hat{\beta}_0 + \hat{\beta}_D + \hat{\beta}_X^T\mathbf{X} + \hat{\beta}_{DX}^T\mathbf{X}D$ and $\hat{\mu}_{0X} = \hat{\beta}_0 + \hat{\beta}_X^T\mathbf{X}$. The ROC curve is then estimated by

$$\widehat{ROC}(t) = \hat{S}_{1X}\left(\hat{S}_{0X}^{-1}(t)\right). \tag{4.3}$$

The covariate-specific ROC curve can be generated by plotting $\hat{S}_{1X}(t)$ against $\hat{S}_{0X}(t)$ for specified covariate values, $\mathbf{X}$.

### 4.2.3 Semi-parametric empirical likelihood approach

To develop the likelihood, denote $f(Y_{ij}|\mathbf{X}_{ij}, D_{ij})$ the distribution of the screening test conditional on covariates and disease status. Denote $g(\mathbf{X}_{ij}|D_{ij} = d)$ the distribution of the covariates conditional on disease status. The joint distribution of disease status, covariates, and screening test, conditional on $Y_{ij}$ falling in the interval $C_k$ is given by $f(Y_{ij}, D_{ij}, \mathbf{X}_{ij}|Y_{ij} \in C_k)$, $k = (1,\ldots,K)$. Consider stratum sizes in the population $N_k = n_{0,k} + n_k + n_{\bar{V},k}$, where $n_{i,k} = \sum_{j=1}^{n_i} I(Y_i \in C_k)$ for $i = \left\{0, \bar{V}\right\}$ and $k = \{1,\ldots,K\}$. We use a binormal model to describe the relationship between the screening test and the disease status and covariates, given by

$$Y = \mathbf{X}\boldsymbol{\beta} + \sigma(D)\,\epsilon = \beta_0 + \beta_1 D + \beta_2 \mathbf{X} + \beta_3 \mathbf{X} D + \sigma(D)\,\epsilon \tag{4.4}$$

where $\epsilon \sim \mathcal{N}(0,1)$ and $\sigma(D) = \sigma_1 D + \sigma_0(1-D)$ (Pepe, 2004; Wang et al., 2013). Let $\boldsymbol{\sigma} = (\sigma_1, \sigma_0)$.

The validation portion of the likelihood is given by

$$\begin{aligned}
&L_V(G, p, \boldsymbol{\beta}, \boldsymbol{\sigma}) \\
&= \prod_{j=1}^{n_0} f(Y_{0j}, X_{0j}, D_{0j}) \times \prod_{k=1}^{K} \prod_{j=1}^{n_k} f(Y_{kj}, X_{kj}, D_{kj} | Y_{kj} \in C_k) \\
&= \prod_{k=1}^{K} Pr(Y_{kj} \in C_k)^{-n_k} \prod_{i=0}^{K} \prod_{j=1}^{n_i} f(Y_{ij} | X_{ij}, D_{ij} = d)\, g(X_{ij} | D_{ij} = d)\, Pr(D_{ij} = d). \qquad (4.5)
\end{aligned}$$

The non-validation portion of the likelihood takes into account the missing data, where the disease status is unknown, and is given by

$$\begin{aligned}
&L_{\bar{V}}(G, p, \boldsymbol{\beta}, \boldsymbol{\sigma}) \\
&= \frac{(N-n_0)!}{\prod_{k=1}^{K}(N_k - n_{0,k})!} \prod_{k=1}^{K} Pr(Y_{kj} \in C_k)^{N_k - n_{0,k}} \prod_{\substack{j=1 \\ Y \in C_k}}^{n_{\bar{V}}} \frac{f\left(Y_{\bar{V}j}, \mathbf{X}_{\bar{V}j}\right)}{Pr(Y_{kj} \in C_k)} \\
&= \frac{(N-n_0)!}{\prod_{k=1}^{K}(N_k - n_{0,k})!} \prod_{k=1}^{K} Pr(Y_{kj} \in C_k)^{n_k} \\
&\quad \times \prod_{j=1}^{n_{\bar{V}}} \sum_{d=0}^{1} f\left(Y_{\bar{V}j} | \mathbf{X}_{\bar{V}j}, D_{\bar{V}j} = d\right) g\left(\mathbf{X}_{\bar{V}j} | D_{\bar{V}j} = d\right) Pr\left(D_{\bar{V}j} = d\right). \qquad (4.6)
\end{aligned}$$

The full likelihood is found by combining the validation and non-validation portions of the

likelihood, (4.5) and (4.6), given by

$$
\begin{aligned}
&L\left(\{q_{ij}\},\{r_{ij}\},p,\boldsymbol{\beta},\boldsymbol{\sigma}\right) = L_V\left(G,p,\boldsymbol{\beta},\boldsymbol{\sigma}\right) \times L_{\bar{V}}\left(G,p,\boldsymbol{\beta},\boldsymbol{\sigma}\right) \\
&\propto \prod_{i,j\in V} f\left(Y_{ij}|\mathbf{X}_{ij},D_{ij}\right) g\left(\mathbf{X}_{ij}|D_{ij}\right) Pr\left(D_{ij}=d\right) \\
&\quad \times \prod_{j=1}^{n_{\bar{V}}} \sum_{d=0}^{1} f\left(Y_{\bar{V}j}|\mathbf{X}_{\bar{V}j},D_{\bar{V}j}=d\right) g\left(\mathbf{X}_{\bar{V}j}|D_{\bar{V}j}=d\right) Pr\left(D_{\bar{V}j}=d\right) \\
&= \prod_{\substack{i,j\in V \\ D=1}} f\left(Y_{ij}|\mathbf{X}_{ij},D_{ij}=1\right) q_{ij}p \times \prod_{\substack{i,j\in V \\ D=0}} f\left(Y_{ij}|\mathbf{X}_{ij},D_{ij}=0\right) r_{ij}\left(1-p\right) \\
&\quad \times \prod_{j=1}^{n_{\bar{V}}} \left[ f\left(Y_{\bar{V}j}|\mathbf{X}_{\bar{V}j},D_{\bar{V}j}=1\right) q_{\bar{V}j}p + f\left(Y_{\bar{V}j}|\mathbf{X}_{\bar{V}j},D_{\bar{V}j}=0\right) r_{\bar{V}j}\left(1-p\right) \right], \quad (4.7)
\end{aligned}
$$

where $p = Pr\left(D=1\right)$, $q_{ij} = g\left(X_{ij}|D_{ij}=1\right)$, and $r_{ij} = g\left(X_{ij}|D_{ij}=0\right)$. The probability of having disease, $p$, is estimated from the SRS component of the TDS. Due to the relationship between $q_{ij}$ and $r_{ij}$ and the need to decrease computational load, Qin and Zhang (1997, 2003) propose a constraint given by $\frac{r_{ij}}{q_{ij}} = e^{m(\mathbf{X}_{ij})\boldsymbol{\alpha}}$. Consider the standard logistic regression model where $Pr\left(D=1|Y\right) = \frac{e^{m^*(Y)\boldsymbol{\alpha}}}{1+e^{m^*(Y)\boldsymbol{\alpha}}} = \psi\left(Y\right)$. The Bayes' rule gives $f\left(Y|D=1\right) = \frac{f(Y)Pr(D=1|Y)}{Pr(D=1)} = \frac{f(Y)\psi(Y)}{p}$. Similarly, $f\left(Y|D=0\right) = \frac{f(Y)(1-\psi(Y))}{1-p}$. Consider the ratio

$$
\begin{aligned}
\frac{r_{ij}}{q_{ij}} &= \frac{f\left(Y|D=0\right)}{f\left(Y|D=1\right)} \\
&= \left[\frac{f\left(Y\right)\left(1-\psi\left(Y\right)\right)}{1-p}\right] \times \left[\frac{p}{f\left(Y\right)\psi\left(Y\right)}\right] \\
&= \frac{p}{1-p}\left(\frac{1}{1+e^{m^*(Y)\boldsymbol{\alpha}}}\right)\left(\frac{e^{m^*(Y)\boldsymbol{\alpha}}}{1+e^{m^*(Y)\boldsymbol{\alpha}}}\right) \\
&= e^{m(Y)\boldsymbol{\alpha}},
\end{aligned}
$$

which implies that $r_{ij} = q_{ij} e^{m(\mathbf{X}_{ij})\boldsymbol{\alpha}}$. Applying this constraint to the log-likelihood gives

$$
\begin{aligned}
& l\left(\{q_{ij}\}, p, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}\right) \\
& \propto \sum_{ij} \ln q_{ij} + \sum_{ij \in V} f\left(Y_{ij} | \mathbf{X}_{ij}, D_{ij} = d\right) + \sum_{\substack{i,j \in V \\ D=0}} m\left(\mathbf{X}_{ij}\right) \boldsymbol{\alpha} \\
& \quad + \sum_{j=1}^{n_{\bar{V}}} \ln \left[ f\left(Y_{\bar{V}j} | \mathbf{X}_{\bar{V}j}, D_{\bar{V}j} = 1\right) p + f\left(Y_{\bar{V}j} | \mathbf{X}_{\bar{V}j}, D_{\bar{V}j} = 0\right) e^{m(\mathbf{X}_{ij})\boldsymbol{\alpha}} (1-p) \right]. \quad (4.8)
\end{aligned}
$$

Without loss of generality, consider partitioning the screening test variable into three mutually exclusive intervals: $C_1 = (-\infty, a_1]$, $C_2 = (a_1, a_2]$, and $C_3 = (a_2, \infty)$. The TDS consists of an SRS component of size $n_0$, a TDS component of size $n_1 + n_2 + n_3$, and the non-validation set of size $N - n_V$, where $n_V = n_0 + n_1 + n_2 + n_3$. Subjects are eligible to be sampled for the TDS component based on their screening test result, Y. For example, if a subject's test result is less than or equal to $a_1$ and $n_1 > 0$, the probability of being selected in the TDS component for $C_1$ is greater than zero.

In order to develop the semi-parametric empirical likelihood ROC curve estimator (SPEL-ROC), we first need to obtain estimates for $\{q_{ij}\}$, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\sigma}$. We estimate $\{q_{ij}\}$ using empirical likelihood methods outlined below. Parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\sigma}$ are estimated using the Newton-Raphson algorithm. Then we can use these parameter estimates in the ROC curve equation, in (4.3), to obtain the SPEL-ROC.

To obtain estimates for $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\sigma}$ the profile likelihood must be constructed. The distribution of the covariates conditional on disease status, $g\left(\mathbf{X}_{ij} | D_{ij} = 1\right)$, is not known or assumed. A robust estimator for the covariate-specific ROC curve can be constructed without making these distributional assumptions by fixing $\boldsymbol{\alpha}$ and then obtaining the empirical likelihood function of $G\left(\mathbf{X}_{ij} | D_{ij} = 1\right)$, with support at the observed values of $\mathbf{X}$. To maximize $L\left(\{q_{ij}\}, p, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}\right)$, we estimate $\{\hat{q}_{ij}\} = g\left(\mathbf{X}_{ij} | D_{ij} = 1\right)$ under the following constraints:

$$
\left\{ q_{ij} \geq 0, \sum_{i \in (0,1,2,3,\bar{V})} \sum_{j=1}^{n_i} q_{ij} = 1, \sum_{i \in (0,1,2,3,\bar{V})} \sum_{j=1}^{n_i} q_{ij} \left\{ e^{m(\mathbf{X}_{ij})\alpha} - 1 \right\} = 0 \right\}. \quad (4.9)
$$

A unique maximum for $\{q_{ij}\}$ exists under the constraints given in (4.9) if 1 is inside

the convex hull of points $e^{m(\mathbf{X}_{ij})\boldsymbol{\alpha}}$ for all $(i,j)$ (Owen, 1988, 1990; Qin and Lawless, 1994). Lagrange multipliers, $\lambda_1$ and $\lambda_2$, are used to derive the maximum over $\{\hat{q}_{ij}\}$. Consider the function

$$\begin{aligned} H &= l\left(\{q_{ij}\}, p, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}\right) + \lambda_1 \left(1 - \sum_{ij} q_{ij}\right) + n\lambda_2 \sum_{ij} q_{ij} \left\{e^{m(\mathbf{X}_{ij})\alpha} - 1\right\} \\ &\propto \sum_{ij} \ln q_{ij} + \lambda_1 \left(1 - \sum_{ij} q_{ij}\right) + N\lambda_2 \sum_{ij} q_{ij} \left\{e^{m(\mathbf{X}_{ij})\alpha} - 1\right\}. \end{aligned} \tag{4.10}$$

Estimates $\{\hat{q}_{ij}\}$ and $\hat{\lambda}_1$ are found by taking the derivative of H with respect to $q_{ij}$, where H is given by (4.10), and setting the derivative equal to zero. The derivative of H is given by $\frac{\partial H}{\partial q_{ij}} = \frac{1}{q_{ij}} - \lambda_1 + N\lambda_2 \left\{e^{m(\mathbf{X}_{ij})\boldsymbol{\alpha}} - 1\right\}$. The estimate $\hat{\lambda}_1$ is found by evaluating $\sum_{ij} q_{ij} \frac{\partial H}{\partial q_{ij}} = N - \lambda_1 \sum_{ij} q_{ij} + N\lambda_2 \sum_{ij} q_{ij} \left\{e^{m(\mathbf{X}_{ij})\boldsymbol{\alpha}} - 1\right\} = 0$ and solving for $\lambda_1$. By setting the derivative of H equal zero and solving for $q_{ij}$, we have

$$\hat{q}_{ij} = \frac{1}{N}\left[1 - \lambda_2 \left(e^{m(\mathbf{X}_{ij})\boldsymbol{\alpha}} - 1\right)\right]^{-1} \tag{4.11}$$

for $i \in (0, 1, 2, 3, \bar{V})$ and $j \in (1, \ldots, n_i)$.

*Profile log-likelihood*

The empirical profile log-likelihood is obtained by plugging the estimates $\hat{q}_{ij}$ given in (4.11) into the log-likelihood, given in (4.8). Define $\xi = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}, \lambda_2)$. Denoting $pl\left(\xi\right)$ as the natural

logarithm of the empirical profile likelihood, we have

$$
\begin{aligned}
& pl\left(\xi\right) \\
& \propto -\sum_{ij}\ln\left[1-\lambda_2\left(e^{m(\mathbf{X}_{ij})\boldsymbol{\alpha}}-1\right)\right]+\sum_{ij\in V}\ln f\left(Y_{ij}|\mathbf{X}_{ij},D_{ij}=d\right)+\sum_{\substack{i,j\in V\\ D=0}}m\left(\mathbf{X}_{ij}\right)\boldsymbol{\alpha} \\
& \quad +\sum_{j=1}^{n_{\bar{V}}}\ln\left[\ f\left(Y_{\bar{V}j}|\mathbf{X}_{\bar{V}j},D_{\bar{V}j}=1\right)p+f\left(Y_{\bar{V}j}|\mathbf{X}_{\bar{V}j},D_{\bar{V}j}=0\right)e^{m(\mathbf{X}_{\bar{V}j})\boldsymbol{\alpha}}\left(1-p\right)\ \right] \\
& \propto -\sum_{ij}\ln\left[1-\lambda_2\left(e^{m(\mathbf{X}_{ij})\boldsymbol{\alpha}}-1\right)\right]\qquad +\sum_{\substack{i,j\in V\\ D=0}}m\left(\mathbf{X}_{ij}\right)\boldsymbol{\alpha}-\frac{1}{2}n_{v,D=1}\ln\sigma_1^2 \\
& \quad -\sum_{\substack{i,j\in V\\ D=1}}\left[\frac{1}{2\sigma_1^2}\left(Y_{ij}-\mathbf{X}_{ij}\beta\right)^2\right]-\sum_{\substack{i,j\in V\\ D=0}}\left[\frac{1}{2\sigma_0^2}\left(Y_{ij}-\mathbf{X}_{ij}\beta\right)^2\right]-\frac{1}{2}n_{v,D=0}\ln\sigma_0^2 \\
& \quad +\sum_{j=1}^{n_{\bar{V}}}\ln\left[(2\pi)^{-1/2}\sigma_1^{-1}\exp\left\{\frac{-1}{2\sigma_1^2}\left(Y_{\bar{V}j}-\mathbf{X}_{\bar{V}j,D=1}\beta\right)^2\right\}p\right. \\
& \qquad\qquad \left.+(2\pi)^{-1/2}\sigma_0^{-1}\exp\left\{\frac{-1}{2\sigma_0^2}\left(Y_{\bar{V}j}-\mathbf{X}_{\bar{V}j,D=0}\beta\right)^2\right\}e^{m(\mathbf{X}_{\bar{V}j})\boldsymbol{\alpha}}\left(1-p\right)\right]. \qquad (4.12)
\end{aligned}
$$

The Newton-Raphson algorithm can be used to obtain $\hat{\xi}$.

The ROC curve is given by $ROC = S_1\left(S_0^{-1}\left(t\right)\right)$ and can be viewed by plotting $S_1\left(t\right)$ against $S_0\left(t\right)$ for all possible screening test thresholds t, where $S_d$ is the survival function of the screen test for subjects with disease outcome $d$. For the SPEL-ROC, the screening test is assumed to be normally distributed and the survival function can be estimated by

$$\hat{S}_D\left(t\right)=\Phi\left(\frac{\mathbf{X}\hat{\boldsymbol{\beta}}-t}{\hat{\sigma}\left(D\right)}\right), \qquad (4.13)$$

where $\Phi$ denotes the standard normal cumulative distribution function, $\mathbf{X}\boldsymbol{\beta}=\hat{\beta}_0+\hat{\beta}_1 D+\hat{\beta}_2 X+\hat{\beta}_3 XD$, and $\hat{\sigma}\left(D\right)=\hat{\sigma}_1 D+\hat{\sigma}_0\left(1-D\right)$. The estimated ROC curve is seen by plotting $\hat{S}_1\left(t\right)$ against $\hat{S}_0\left(t\right)$ for all possible screening test values, $t$, and it is estimated by

$$\widehat{ROC}_{V,\bar{V}}^{P}\left(t\right)=\hat{S}_1\left(\hat{S}_0^{-1}\left(t\right)\right). \qquad (4.14)$$

## 4.3 Variance estimation of the SPEL-ROC

Standard error estimates were generated using the bootstrap method (Efron and Tibshirani, 1993). The following algorithm was applied to each of 1,000 iterations of the simulation.

1) From the generated sample population, we drew B independent bootstrap samples $s_1, s_2, ..., s_B$ of size $n_V$ with replacement, following the proposed sampling design in Section 4.2.

2) For each bootstrap sample we then computed the SPEL-ROC, denoted $\hat{\theta}_b$ and given in (4.14), resulting in B values of the SPEL-ROC.

3) The average of these bootstrap ROC estimators is given by $\bar{\hat{\theta}}_b = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b$ with variance $\hat{V}\left(\hat{\theta}_b\right) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}_b - \bar{\hat{\theta}}_b\right)^2$ and standard deviation given by $\sqrt{\hat{V}\left(\hat{\theta}_b\right)}$.

4) We repeated these steps 1,000 times in generating the simulation results.

The $\widehat{SE}$ displayed in the simulation results was found by taking the average of the standard deviation estimates across 1,000 independent iterations of the simulation.

## 4.4 Simulation study

We evaluate the behavior of the SPEL-ROC under many situations to better examine its robustness. The simulation studies were conducted using R version 2.14. The data were generated under the binormal model given by $Y = \mathbf{X}\boldsymbol{\beta} + \sigma(D)\,\epsilon = \beta_0 + \beta_D D + \beta_X X + \beta_{XD} XD + \sigma(D)\,\epsilon$, where $\epsilon \sim \mathcal{N}(0,1)$, $X \sim \mathcal{N}(0,2)$, $\sigma(D) = \sigma_1 D + \sigma 0\,(1-D)$, and $D \sim Bernoulli\,(0.3)$ where $D = 1$ for diseased subjects and $D = 0$ for non-diseased subjects. Model parameters are given by: $\beta_0 = 0.5$, $\beta_D = 1$, $\beta_X = 0.5$, $\beta_{XD} = -0.3$, $\sigma_1^2 = 1$, and $\sigma_0^2 = 2$. The population size used in simulations is $N = 1000$ and the distribution of $Y$ is partitioned into three mutually exclusive sets given by $C_1 = (-\infty, a_1]$, $C_2 = (a_1, a_2]$, and $C_3 = (a_2, \infty)$. In the following simulations, we consider the impact of 1) varying the sample size, $n_V$ and 2) varying the proportion of SRS to TDS component sizes, $\frac{n_0}{n_V}$. The variations under consideration are: validation sample size ($n_V$) 120, 240, and 360; and proportion of

SRS subjects among validation set ($\frac{n_0}{n_V}$) 0.25, 0.5, and 0.75. For simulation results, the estimated ROC curve, standard errors, average of the standard error estimates, and 95% nominal coverage probabilities for an estimator are obtained from 1000 independent runs. Estimated standard errors were obtaining with bootstrapping, using 50 replicate samples.

*Sample allocation for TDS*

The SPEL-ROC uses the TDS design to target subjects in all three intervals of the screening test distribution with sampling fraction $\frac{n_0}{n_V} = 0.5$. This sample consists of the following three components:

1) SRS component of size $n_0$,

2) TDS component of size $n_1 + n_2 + n_3$ where $n_1$ subjects are sampled such that $Y_{1j} \in C_1 = (-\infty, a_1]$, $n_2$ subjects are sampled such that $Y_{2j} \in C_2 = (a_1, a_2]$, $n_3$ subjects are sampled such that $Y_{3j} \in C_3 = (a_2, \infty)$, and $n_1 = n_2 = n_3$, and

3) non-validation component of size $N - n_V$, comprised of all subjects not sampled into the SRS and TDS components.

*Estimators to be compared*

The SPEL-ROC, given by $\widehat{ROC}^{P}_{V,\bar{V}}$ in (4.14), is compared to two estimators in the simulation study. Under each setting, we compare the following three estimators.

1) LS-ROC: the SRS data only estimator, denoted $\widehat{ROC}^{SRS}_{V}(t)$, given in Section 4.2.2.

2) SPEL-ROC(SRS): the SRS with validation and non-validation data estimator, denoted $\widehat{ROC}^{SRS}_{V,\bar{V}}$, has the same form as the SPEL-ROC given in (4.14). The difference between the SPEL-ROC(SRS) and the SPEL-ROC is the sampling scheme. Inclusion of this estimator gives a comparison of the SRS and TDS designs while incorporating both validation and non-validation data.

3) SPEL-ROC: the proposed TDS with validation and non-validation data estimator, denoted $\widehat{ROC}^{P}_{V,\bar{V}}$, is given by (4.14).

*Results*

*Unbiasedness* All three ROC curve estimators yield unbiased estimates. To illustrate this, we simulated data using multiple allocations and sample sizes. Table 4.1 shows that the average of each ROC curve estimator is close to or equal to the true value. Other allocation schemes were considered for the SPEL-ROC in Table 4.3 by varying the proportion of subjects allocated to the SRS component ($\frac{n_0}{n_V} = (0.25, 0.5, 0.75)$). In all allocations and sample sizes, the SPEL-ROC continues to be unbiased.

*Efficiency* The SPEL-ROC and SPEL-ROC(SRS) are more efficient than the LS-ROC. For smaller sample sizes, $n_V = 120$ in Table 4.1, the SPEL-ROC has smaller SE for lower FPR values and the SPEL-ROC(SRS) has a smaller SE for a FPR of 0.5 or greater. With a larger sample size, $n_V = 360$, Table 4.1 shows that while the SPEL-ROC has smaller SE for low FPRs (0.3 or less), and the SEs for this SPEL-ROC and the SPEL-ROC(SRS) are the same when the FPR exceeds 0.3 For example, consider $n_V = 120$ and FPR= $\{0.1, 0.3\}$, the SPEL-ROC has SEs of 0.057 and 0.069, respectively, compared to 0.060 and 0.070 for the SPEL-ROC(SRS). Whereas, for $n_V = 360$ and FPR= $\{0.1, 0.3\}$, the SPEL-ROC has SEs of 0.038 and 0.045, respectively, compared to 0.042 and 0.047 for the SPEL-ROC(SRS) and for FPR= $\{0.5, 0.7, 0.9\}$ both SPEL-ROC and SPEL-ROC(SRS) have SEs 0.031, 0.014, 0.002, respectively. This supports the idea that both the inclusion of non-validation subjects creates a more efficient alternative to the SRS design using the ordinary least squares method to estimate the model parameters while sampling the same number of subjects. The simulation studies suggest that for smaller values of the FPR, less than 0.5, the SPEL-ROC is the most efficient out of the three ROC estimators. The bootstrapped standard errors estimated the SE well, with coverage probabilities close to the nominal level of 0.95. For both $n_V = 120$ and $n_V = 360$, the SE and $\hat{SE}$ are equal or very close to the same value. The greatest difference can be seen for $n_V = 360$ and $\frac{n_0}{n_V} = 0.25$, for a FPR of 0.3 we have SE of 0.048 and $\hat{SE}$ slightly underestimates the SE, given by 0.045. In Table 4.3, for $n_V = 360$ and $\frac{n_0}{n_V} = 0.5$, the coverage probability declines (0.950, 0.941, 0.922, 0.910, 0.891) as the FPR increases (0.1, 0.3, 0.5, 0.7, 0.9). This trend was observed for all allocations considered in Table 4.3. Table 4.2 shows results similar to Table 4.1 in terms of efficiency of the covariate specific ROC

curve estimators. Table 4.2 and Figure 4.1 show the utility of covariate-specific ROC curve estimator by identifying covariate values where the screening test is more effective.

## 4.5 Analysis of the lung cancer study data

We used the SPEL-ROC to analyze non-small-cell lung cancer (NSCLC) data from the CALGB 150807 study (Bueno et al., 2012) conducted by the Cancer and Leukemia Group B. This study is a subset of patients registered in the CALGB 140202 study who have stage 1A or 1B non-small-cell lung cancer (NSCLC), have not received preoperative chemotherapy or radiation, and are not missing histological, demographic, clinical, and follow-up information of interest. Among patients in the CALGB 150807 study, 1,061 patients were not censored before 12 months and were used in this analysis.

Lung cancer is the most common cause of cancer death among men and women in the world (Blanchon et al., 2006). Lung cancer is classified as either small-cell lung carcinoma (SCLC) or non-small-cell lung carcinoma (NSCLC), of which NSCLC accounts for approximately 80% of all lung cancers. After surgical lung resection, a large proportion of stage 1 NSCLC patients have cancer recurrence within five years (Bueno et al., 2012). When surgery is used as the primary treatment for NSCLC, adjuvant chemotherapy may benefit patients who have a high risk of cancer recurrence. Identifying patients who are at high risk of cancer recurrence is important in order for treatment to be given to those who would benefit most. This is an important area of study for patients, families, and doctors when making decisions on a treatment plan.

The Balcone risk score, outlined by Blanchon et al. (2006), has been developed to identify patients who are at greatest risk of cancer recurrence. To select the variables that are included in the scoring algorithm, Blanchon et al. (2006) used a Cox model to identify variables that were independently associated with mortality. The associated variables were then weighted to create the Balcone risk index. The components of the risk score are given by: age (>70 years, 1 point); sex (male, 1 point); performance status at diagnosis (reduced activity, 3 points; active >50%, 5 points; inactive >50%, 8 points; and total incapacity, 10 points); histological

type (large-cell carcinoma, 2 points); and tumour-node-metastasis (TNM) staging system (IIA or IIB, 3 points; IIIA or IIIB, 6 points; and IV, 8 points). For the data used in this analysis, the Balcone risk score ranges from 0 to 15.

The goal of this analysis is to summarize the ability of the Balcone risk score to distinguish between patients who survive beyond 12 months and those who do not. This will allow us to evaluate the benefit of using the Balcone risk score to identify patients at a higher risk of early cancer recurrence. The outcome of interest is survival beyond 12 months and the screening test is the Balcone risk score. The ROC curve is estimated at different ages to see if the Balcone score is a more effective screening test at certain ages. Although all information is available for these patients, we illustrate the utility of the proposed ROC curve estimator by sampling from the study data and evaluating the estimated ROC curve by age. Table 4.4 gives descriptive statistics for the Balcone risk score and patient age, stratified by survival at 12 months. Cut-points for the TDS component were defined by $\alpha = 1$ standard deviations from the mean of the Balcone risk score. A sample size of $n_V = 360$ was used for all estimators compared and details on sample allocation for each estimator are given in Table 4.8. The LS-ROC has an SRS of size $n = 360$. The SPEL-ROC(SRS) allocates 100% of the sample to the SRS component and utilizes incomplete data from the non-validation set. The SPEL-ROC allocates 50% of the validation sample to the SRS component and samples the remaining 50% are allocated equally between the three intervals, $C_i$, such that $n_1 = n_2 = n_3 = 60$, while utilizing incomplete data from the non-validation set.

Table 4.6 and Figure 4.2 show the results from the analysis of the lung cancer study data. We have modeled the Balcone risk score using information for the disease status and the patient age. These results show that the Balcone risk score is a more effective screening test for survival at 1 year for younger ages compared to older ages.

## 4.6 Analysis of the Preterm Prediction Study data

We used the SPEL-ROC to analyze data from the Preterm Prediction Study, a multi-center prospective study designed to study spontaneous preterm birth (Goldenberg et al.,

1996). The Maternal Fetal Medicine Units Network of the National Institute of Child Health and Human Development carried out this study using ten centers to recruit women. There were 3073 women recruited between October 1992 and July 1994. Measurements were collected every two weeks from 22 to 30 weeks' gestation. Among the 3073 women recruited, 3001 had valid measurements of interest for this analysis.

PTB is defined as delivery at less than 37 weeks of gestation and contributes to neonatal morbidity and mortality which increases as gestational age decreases (McCormick, 1985). Bastek and Elovitz (2013) combined results from multiple studies on this topic to gain a better understanding the relationship between biomarkers and PTB. The results were not definitive for most biomarkers with the exception of fetal fibronectin (FFN).

Fetal fibronectin (FFN) is a protein that is produced by the fetal membrane. Knowing the FFN measurement will not change the incidence of spontaneous PTB but it will effect the treatment plan. Other variables are important in studying the ability of FFN to distinguish between PTB and not-PTB. Cervical length (CL) has been shown to be associated with PTB, so a model that allows us to observe the estimated ROC curve at different CL measurements may lead to a better understanding of population subgroups where FFN functions as a better screening test. Other variables, such as maternal age and previous PTB, are also of interest in this setting. Deshpande et al. (2013) found that FFN has moderate accuracy in predicting PTB. Although many studies are concerned with the ability to predict spontaneous PTB, Bastek and Elovitz (2013) suggest that the ability to predict those who will not have spontaneous PTB is also valuable. Because FFN typically has a high negative predictive value (proportion of true negatives over all who test negative), a negative FFN test is widely used in clinical practice to send patients home. Measurable levels of FFN are considered to be abnormal between 20 and 37 weeks' gestation. Lockwood et al. (1991) show that in 588 FFN samples from uncomplicated pregnancies, a higher percentage of subjects were positive for FFN (level above 0.05 $\mu$g/mL) before 22 and after 37 weeks' gestation compared to between 22 and 37 weeks' gestation. For example the percentage of cervical samples with positive FFN for $<22$, 22 to 37 and $>37$ weeks' gestation were 24%, 4%, and 32%, respectively. This is an important area of study due to the negative effects of spontaneous PTB on maternal

and child health outcomes.

In our analysis, the outcome of interest is spontaneous PTB at less than 37 weeks' gestation and the screening test considered in FFN. We include cervical length as a covariate in the model. Because the standard deviation is large compared to the mean, the cut-points for the TDS component were defined using $\alpha = 0.15$. A sample size of $n_V = 360$ was used for all estimators compared. The LS-ROC has an SRS of size $n = 360$. The SPEL-ROC(SRS) allocates 100% of the sample to the SRS component and utilizes incomplete data from the non-validation set. The SPEL-ROC allocates 50% of the validation sample to the SRS component and the remaing 50% are allocated equally between the three intervals, $C_i$, such that $n_1 = n_2 = n_3 = 60$, while utilizing incomplete data from the non-validation set. Table 4.9 and Figure 4.3 show the results from the analysis of the Preterm Prediction Study data. We have modeled the screening test using information for the disease status and the cervical length. These results show that the screening test, FFN, is a more effective at screening for spontaneous PTB when the cervical length is shorter.

## 4.7 Discussion

We have proposed a semi-parametric estimator for the covariate-specific ROC curve, which allows us to summarize the ability of a screening test to discern between diseased and non-diseased subjects for specified covariate values. This estimator incorporates a TDS design and includes both validation and non-validation data. Although a binormal distribution is assumed for the screening test variable, the use of empirical likelihood methods allows us to estimate the ROC without specifying a distribution for the covariates.

The proposed design is motivated by the need to improve efficiency in estimating the ROC curve and to develop methods to observe the utility of the screening test at different covariate values. Ascertaining true disease status can be costly and invasive for subjects. Although disease status is missing for the non-validation set, the proposed estimator takes advantage of all information available for a larger number of subjects than the validation-only estimators. Although all estimators are unbiased, the proposed estimator was shown to be

the most efficient for low FPR values, compared to two competing ROC curve estimators. It is unclear if the either the SPEL-ROC or SPEL-ROC(SRS) is superior at higher FPR values, in terms of the standard error. These results suggest that to obtain the same variability less subjects would be needed when the non-validation subjects are used to estimate the ROC curve, reducing study cost and subject burden. These results support the idea that inclusion of non-validation subjects creates a more efficient alternative to the validation-only estimators while sampling the same number of subjects. For example, with a sample size of 360 and FPR of 0.1, the standard error of the proposed estimator is 0.038 compared to 0.080 and 0.060 for the LS-ROC and SPEL-ROC(SRS), respectively (in Table 4.1). In contrast, with a sample size of 360 and FPR of 0.7, the standard error of the proposed estimator is 0.014 compared to 0.016 and 0.014 for the LS-ROC and SPEL-ROC(SRS), respectively.

More simulation studies are needed to explore the robustness of the SPEL-ROC. Non-normally distributed data should be simulated for both the screening test and the covariates. The empirical likelihood methods allow for ROC curve estimation without specification of the distribution of the covariates. Simulating covariates using non-normal distributions will allow us to see how well the ROC curve estimator works for different types of data. Simulating the screening test variable using non-normal distributions will allow us to see how well the SPEL-ROC performs when the distribution of the screening test is mis-specified.

Table 4.1: Comparison of SPEL-ROC and competing methods

| | | | $\widehat{ROC}_V^{SRS}$ | | $\widehat{ROC}_{V,\bar{V}}^{SRS}$ | | $\widehat{ROC}_{V,\bar{V}}^{P}$ | |
|---|---|---|---|---|---|---|---|---|
| $n_V$ | FPR | True ROC | Mean | SE | Mean | SE | Mean | SE |
| 120 | 0.1 | 0.208 | 0.212 | 0.080 | 0.200 | 0.060 | 0.200 | 0.057 |
| | 0.3 | 0.602 | 0.600 | 0.092 | 0.582 | 0.070 | 0.581 | 0.069 |
| | 0.5 | 0.841 | 0.838 | 0.063 | 0.826 | 0.046 | 0.824 | 0.048 |
| | 0.7 | 0.959 | 0.955 | 0.029 | 0.952 | 0.021 | 0.950 | 0.023 |
| | 0.9 | 0.998 | 0.996 | 0.005 | 0.996 | 0.006 | 0.996 | 0.004 |
| 360 | 0.1 | 0.208 | 0.212 | 0.048 | 0.208 | 0.042 | 0.207 | 0.038 |
| | 0.3 | 0.602 | 0.605 | 0.054 | 0.600 | 0.047 | 0.599 | 0.045 |
| | 0.5 | 0.841 | 0.842 | 0.036 | 0.839 | 0.031 | 0.839 | 0.031 |
| | 0.7 | 0.959 | 0.958 | 0.016 | 0.958 | 0.014 | 0.958 | 0.014 |
| | 0.9 | 0.998 | 0.997 | 0.002 | 0.997 | 0.002 | 0.997 | 0.002 |

Cutpoints for the TDS component are defined by $(a_1, a_2) = (\mu_Y - \alpha\sigma_Y, \mu_Y + \alpha\sigma_Y)$ where $\alpha = 1$. $\widehat{ROC}_V^{SRS}$ denotes the LS-ROC which uses SRS; $\widehat{ROC}_{V,\bar{V}}^{SRS}$ denotes the SPEL-ROC(SRS) which uses SRS and utilizes information for both validation and non-validation data; and $\widehat{ROC}_{V,\bar{V}}^{P}$ denotes the proposed SPEL-ROC which uses TDS and utilizes information from both validation and non-validation data. All estimators sample the same number of subjects. ROC curve estimates are calculated for the population average of the covariate.

Table 4.2: SPEL-ROC for specific covariate values

| | | | $\widehat{ROC}_V^{SRS}$ | | $\widehat{ROC}_{V,\bar{V}}^{SRS}$ | | $\widehat{ROC}_{V,\bar{V}}^{P}$ | |
|---|---|---|---|---|---|---|---|---|
| | FPR | True ROC | Mean | SE | Mean | SE | Mean | SE |
| X = population average of the covariate | | | | | | | | |
| | 0.1 | 0.208 | 0.212 | 0.048 | 0.208 | 0.042 | 0.207 | 0.038 |
| | 0.3 | 0.602 | 0.605 | 0.054 | 0.600 | 0.047 | 0.599 | 0.045 |
| | 0.5 | 0.841 | 0.842 | 0.036 | 0.839 | 0.031 | 0.839 | 0.031 |
| | 0.7 | 0.959 | 0.958 | 0.016 | 0.958 | 0.014 | 0.958 | 0.014 |
| | 0.9 | 0.998 | 0.997 | 0.002 | 0.997 | 0.002 | 0.997 | 0.002 |
| $X = -0.5$ | 0.1 | 0.246 | 0.260 | 0.054 | 0.257 | 0.049 | 0.255 | 0.043 |
| | 0.3 | 0.649 | 0.664 | 0.054 | 0.662 | 0.050 | 0.658 | 0.046 |
| | 0.5 | 0.870 | 0.877 | 0.033 | 0.876 | 0.030 | 0.874 | 0.029 |
| | 0.7 | 0.969 | 0.970 | 0.013 | 0.971 | 0.011 | 0.970 | 0.011 |
| | 0.9 | 0.998 | 0.998 | 0.002 | 0.998 | 0.001 | 0.998 | 0.001 |
| $X = 1$ | 0.1 | 0.144 | 0.136 | 0.043 | 0.137 | 0.039 | 0.137 | 0.037 |
| | 0.3 | 0.503 | 0.485 | 0.066 | 0.489 | 0.062 | 0.487 | 0.059 |
| | 0.5 | 0.773 | 0.758 | 0.052 | 0.762 | 0.049 | 0.759 | 0.048 |
| | 0.7 | 0.932 | 0.924 | 0.027 | 0.926 | 0.024 | 0.925 | 0.024 |
| | 0.9 | 0.995 | 0.993 | 0.005 | 0.994 | 0.004 | 0.993 | 0.004 |

Cutpoints for the TDS component are defined by $(a_1, a_2) = (\mu_Y - \alpha\sigma_Y, \mu_Y + \alpha\sigma_Y)$ where $\alpha = 1$.

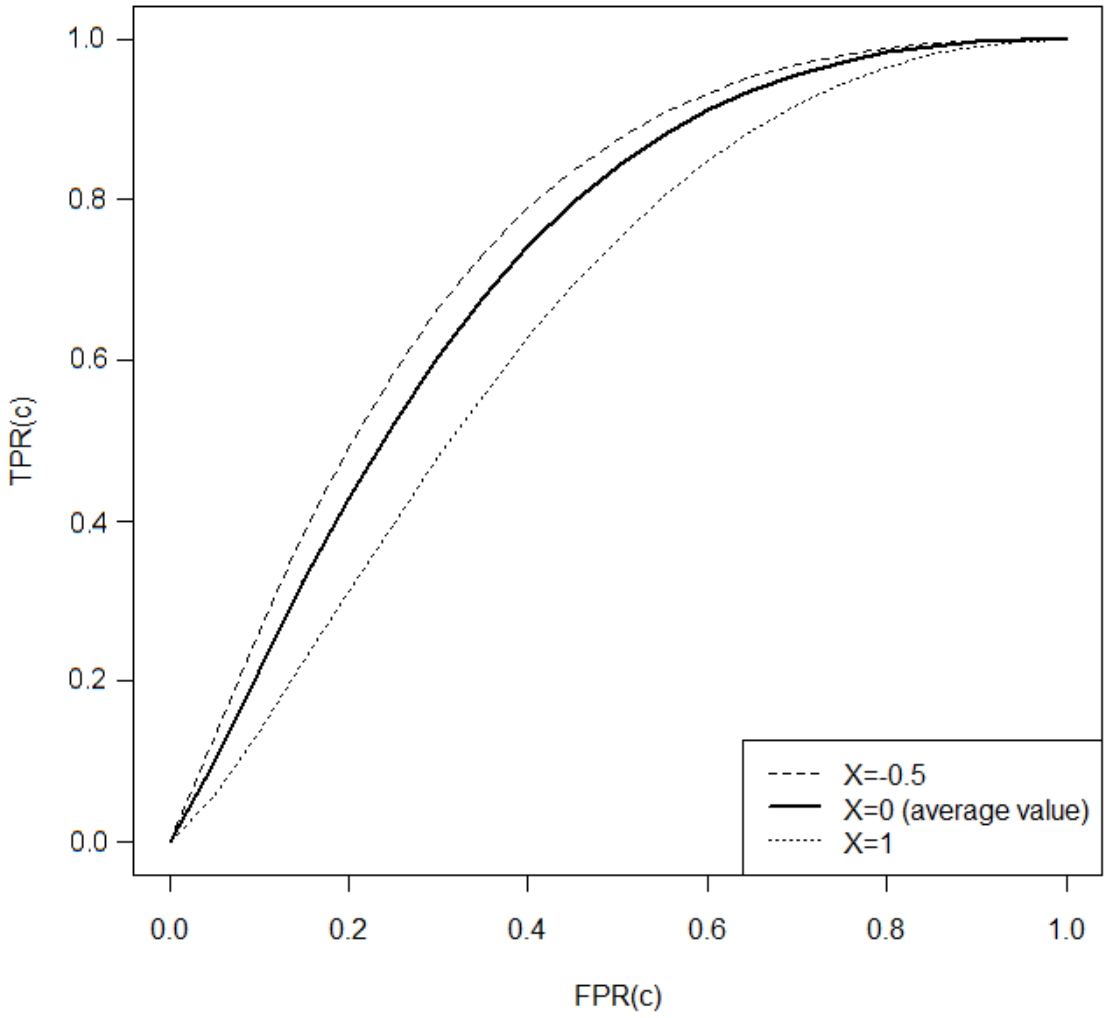


Figure 4.1: Covariate-specific ROC Curve

The SPEL-ROC is calculated with a sample size of $n_V = 360$ and allocation given by $(n_0, n_1, n_2, n_3, n_{\bar{V}}) = (180, 60, 60, 60, 640)$. The parameter c indicates all possible values of the screening test, Y. Cutpoints for the TDS component are defined by $(a_1, a_2) = (\mu_Y - \alpha\sigma_Y, \mu_Y + \alpha\sigma_Y)$ where $\alpha = 1$.

Table 4.3: Properties of the SPEL-ROC

| $(n_0, n_1, n_2, n_3, n_{\bar{V}})$ | | FPR 0.1 | 0.3 | 0.5 | 0.7 | 0.9 |
|---|---|---|---|---|---|---|
| | $n_V = 120$, $\frac{n_0}{n_V} = 0.5$ | | | | | |
| (60,20,20,20,880) | $\widehat{ROC}^P_{V,\bar{V}}$ | 0.200 | 0.581 | 0.824 | 0.950 | 0.996 |
| | SE | 0.057 | 0.068 | 0.048 | 0.023 | 0.004 |
| | $\widehat{SE}$ | 0.059 | 0.070 | 0.050 | 0.025 | 0.006 |
| | 95% CP | 0.942 | 0.936 | 0.938 | 0.929 | 0.941 |
| | $n_V = 360$, $\frac{n_0}{n_V} = 0.5$ | | | | | |
| (180,60,60,60,640) | $\widehat{ROC}^P_{V,\bar{V}}$ | 0.207 | 0.599 | 0.839 | 0.958 | 0.997 |
| | SE | 0.038 | 0.045 | 0.031 | 0.014 | 0.002 |
| | $\widehat{SE}$ | 0.039 | 0.045 | 0.030 | 0.012 | 0.002 |
| | 95% CP | 0.950 | 0.941 | 0.922 | 0.910 | 0.891 |
| | $n_V = 360$, $\frac{n_0}{n_V} = 0.25$ | | | | | |
| (90,90,90,90,640) | $\widehat{ROC}^P_{V,\bar{V}}$ | 0.208 | 0.600 | 0.840 | 0.958 | 0.997 |
| | SE | 0.040 | 0.048 | 0.033 | 0.014 | 0.002 |
| | $\widehat{SE}$ | 0.039 | 0.045 | 0.030 | 0.012 | 0.002 |
| | 95% CP | 0.930 | 0.934 | 0.913 | 0.892 | 0.869 |
| | $n_V = 360$, $\frac{n_0}{n_V} = 0.75$ | | | | | |
| (270,30,30,30,640) | $\widehat{ROC}^P_{V,\bar{V}}$ | 0.209 | 0.600 | 0.839 | 0.958 | 0.997 |
| | SE | 0.041 | 0.048 | 0.032 | 0.014 | 0.002 |
| | $\widehat{SE}$ | 0.041 | 0.046 | 0.030 | 0.013 | 0.002 |
| | 95% CP | 0.940 | 0.931 | 0.924 | 0.913 | 0.909 |

The fraction $\frac{n_0}{n_V}$ is the proportion of subjects allocated to the SRS component out of the total number of validation subjects sampled. Cutpoints for the TDS component are defined by $(a_1, a_2) = (\mu_Y - \alpha\sigma_Y, \mu_Y + \alpha\sigma_Y)$ where $\alpha = 1$. The sample allocation, $(n_0, n_1, n_2, n_3, n_{\bar{V}})$, gives the number of subjects allocated to the SRS component, three intervals of the TDS component, and the non-validation set, respectively. The standard error estimators, $\hat{SE}$, are found by bootstrapping.

Table 4.4: Descriptive statistics for the Balcone risk score and age of patients

| | N | Minimum | Q1 | Median | Q3 | Maximum |
|---|---|---|---|---|---|---|
| | | | Balcone risk score (FFN) | | | |
| Overall | 1061 | 0 | 1 | 1 | 3 | 10 |
| Survival beyond 12 months | 965 | 0 | 1 | 1 | 3 | 10 |
| Survival less than 12 months | 111 | 0 | 1 | 2 | 5 | 9 |
| | | | Age | | | |
| Overall | 1061 | 23.9 | 61.9 | 69.3 | 75.5 | 95.1 |
| Survival beyond 12 months | 965 | 23.9 | 61.4 | 68.9 | 75.0 | 95.1 |
| Survival less than 12 months | 111 | 43.2 | 66.2 | 72.0 | 76.8 | 87.5 |

Table 4.5: Sample allocation for the non-small-cell lung cancer data

| Component | MW-AUC | SPEL-AUC(SRS) | NPEL-AUC | SPEL-AUC |
|---|---|---|---|---|
| SRS | 360 | 360 | 180 | 180 |
| TDS $(n_1, n_2, n_3)$ | $(0, 0, 0)$ | $(0, 0, 0)$ | $(60, 60, 60)$ | $(60, 60, 60)$ |
| non-validation | 0 | 701 | 0 | 701 |

Table 4.6: Lung Cancer Study: Comparison of Covariate-specific ROC Curve Estimators

| | FPR | $\widehat{ROC}_V^{SRS}$ | $\widehat{ROC}_{V,\bar{V}}^{SRS}$ | $\widehat{ROC}_{V,\bar{V}}^{P}$ |
|---|---|---|---|---|
| 68.3 Years old (study average) | | | | |
| | 0.1 | 0.330 | 0.229 | 0.432 |
| | 0.3 | 0.551 | 0.474 | 0.701 |
| | 0.5 | 0.699 | 0.657 | 0.844 |
| | 0.7 | 0.820 | 0.809 | 0.932 |
| | 0.9 | 0.931 | 0.939 | 0.986 |
| 45 Years old | 0.1 | 0.352 | 0.442 | 0.489 |
| | 0.3 | 0.575 | 0.704 | 0.748 |
| | 0.5 | 0.720 | 0.841 | 0.875 |
| | 0.7 | 0.835 | 0.929 | 0.949 |
| | 0.9 | 0.939 | 0.984 | 0.990 |
| 85 Years old | 0.1 | 0.317 | 0.137 | 0.400 |
| | 0.3 | 0.537 | 0.338 | 0.671 |
| | 0.5 | 0.686 | 0.521 | 0.823 |
| | 0.7 | 0.810 | 0.700 | 0.921 |
| | 0.9 | 0.926 | 0.885 | 0.982 |

All estimators are calculated with a sample size of $n_V = 360$. For the SPEL-ROC, cutpoints for the TDS component are defined by $(a_1, a_2) = (\mu_Y - \alpha\sigma_Y, \mu_Y + \alpha\sigma_Y)$ where $\alpha = 1$.

Table 4.7: Preterm Prediction Study: Descriptive Statistics for FFN and Cervical Length

| | N | Minimum | Q1 | Median | Q3 | Maximum |
|---|---|---|---|---|---|---|
| | | | Fetal fibronectin (FFN) | | | |
| Overall | 2987 | 0 | 0.37 | 2.80 | 7.68 | 2151.44 |
| Spontaneous PTB | 308 | 0 | 0.88 | 4.47 | 16.22 | 926.55 |
| Not PTB | 2679 | 0 | 0.28 | 2.60 | 7.20 | 2151.44 |
| | | | Cervical length (CL) | | | |
| Overall | 2987 | 0 | 30 | 35 | 40 | 70 |
| Spontaneous PTB | 308 | 0 | 26 | 33 | 38 | 58 |
| Not PTB | 2679 | 0 | 31 | 35 | 40 | 70 |

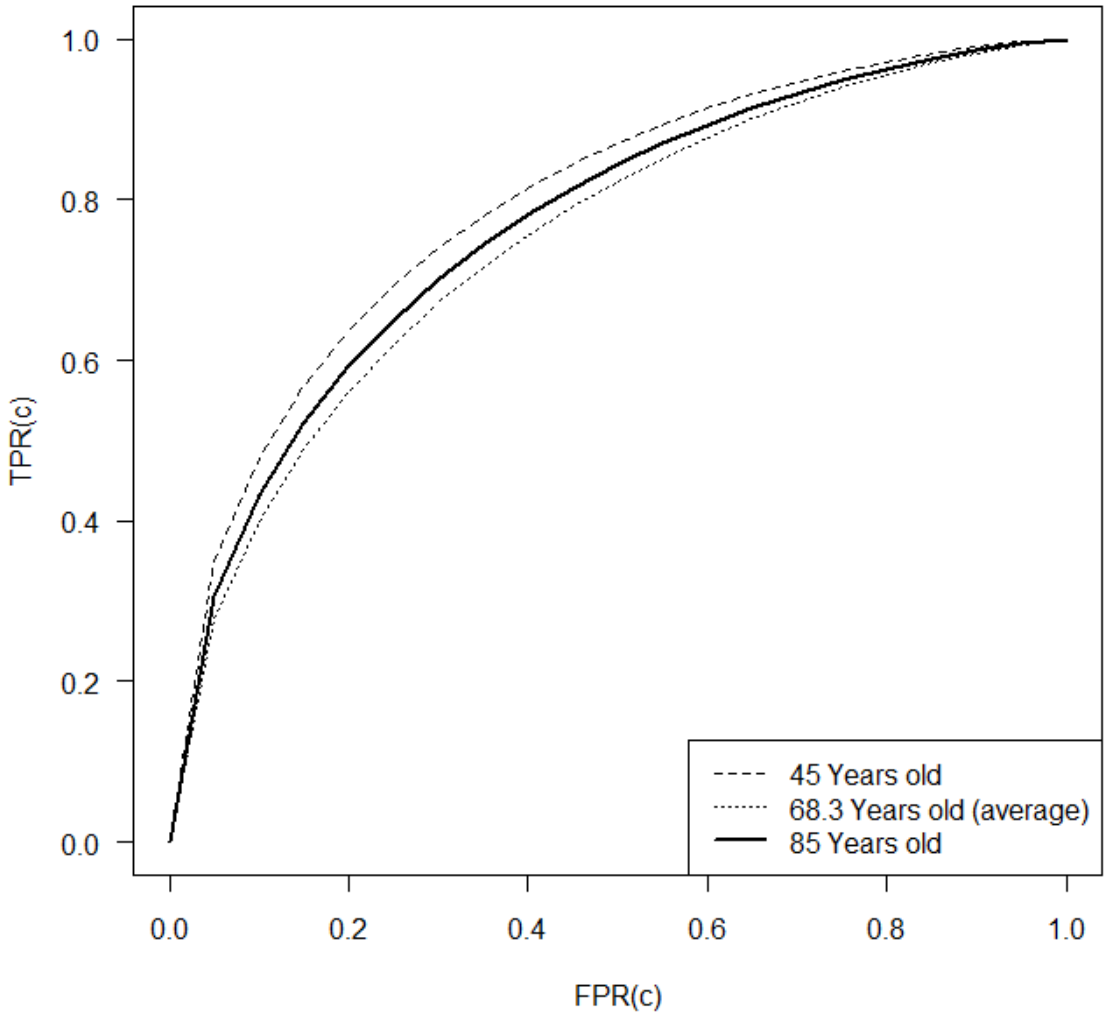


Figure 4.2: Lung Cancer Study: SPEL-ROC by Age

The SPEL-ROC is calculated with a sample size of $n_V = 360$ and allocation given by $(n_0, n_1, n_2, n_3, n_{\bar{V}}) = (180, 60, 60, 60, 701)$. Cutpoints for the TDS component are defined by $(a_1, a_2) = (\mu_Y - \alpha\sigma_Y, \mu_Y + \alpha\sigma_Y)$ where $\alpha = 1$.

Table 4.8: Sample allocation for the Preterm Prediction Study

| Component | LS-ROC | SPEL-ROC(SRS) | SPEL-ROC |
|---|---|---|---|
| SRS | 360 | 360 | 180 |
| TDS $(n_1, n_2, n_3)$ | $(0, 0, 0)$ | $(0, 0, 0)$ | $(60, 60, 60)$ |
| non-validation | 0 | 2627 | 2627 |

Table 4.9: Preterm Prediction Study: Comparison of Covariate-specific ROC Curve Estimators

| | FPR | $\widehat{ROC}_V^{SRS}$ | $\widehat{ROC}_{V,\bar{V}}^{SRS}$ | $\widehat{ROC}_{V,\bar{V}}^{P}$ |
|---|---|---|---|---|
| CL = 35.2mm (population average) | | | | |
| | 0.1 | 0.153 | 0.200 | 0.188 |
| | 0.3 | 0.370 | 0.391 | 0.378 |
| | 0.5 | 0.558 | 0.547 | 0.536 |
| | 0.7 | 0.734 | 0.697 | 0.687 |
| | 0.9 | 0.906 | 0.861 | 0.856 |
| $CL = 15mm$ | 0.1 | 0.442 | 0.446 | 0.304 |
| | 0.3 | 0.708 | 0.668 | 0.525 |
| | 0.5 | 0.847 | 0.797 | 0.678 |
| | 0.7 | 0.934 | 0.890 | 0.805 |
| | 0.9 | 0.986 | 0.964 | 0.925 |
| $CL = 25mm$ | 0.1 | 0.281 | 0.312 | 0.242 |
| | 0.3 | 0.545 | 0.533 | 0.452 |
| | 0.5 | 0.723 | 0.684 | 0.609 |
| | 0.7 | 0.857 | 0.809 | 0.751 |
| | 0.9 | 0.961 | 0.926 | 0.895 |
| $CL = 40mm$ | 0.1 | 0.109 | 0.154 | 0.165 |
| | 0.3 | 0.264 | 0.328 | 0.345 |
| | 0.5 | 0.475 | 0.480 | 0.500 |
| | 0.7 | 0.662 | 0.635 | 0.655 |
| | 0.9 | 0.866 | 0.820 | 0.835 |

All estimators are calculated with a sample size of $n_V = 360$. For the SPEL-ROC, cutpoints for the TDS component are defined by $(a_1, a_2) = (\mu_Y - \alpha\sigma_Y, \mu_Y + \alpha\sigma_Y)$ where $\alpha = 1$.

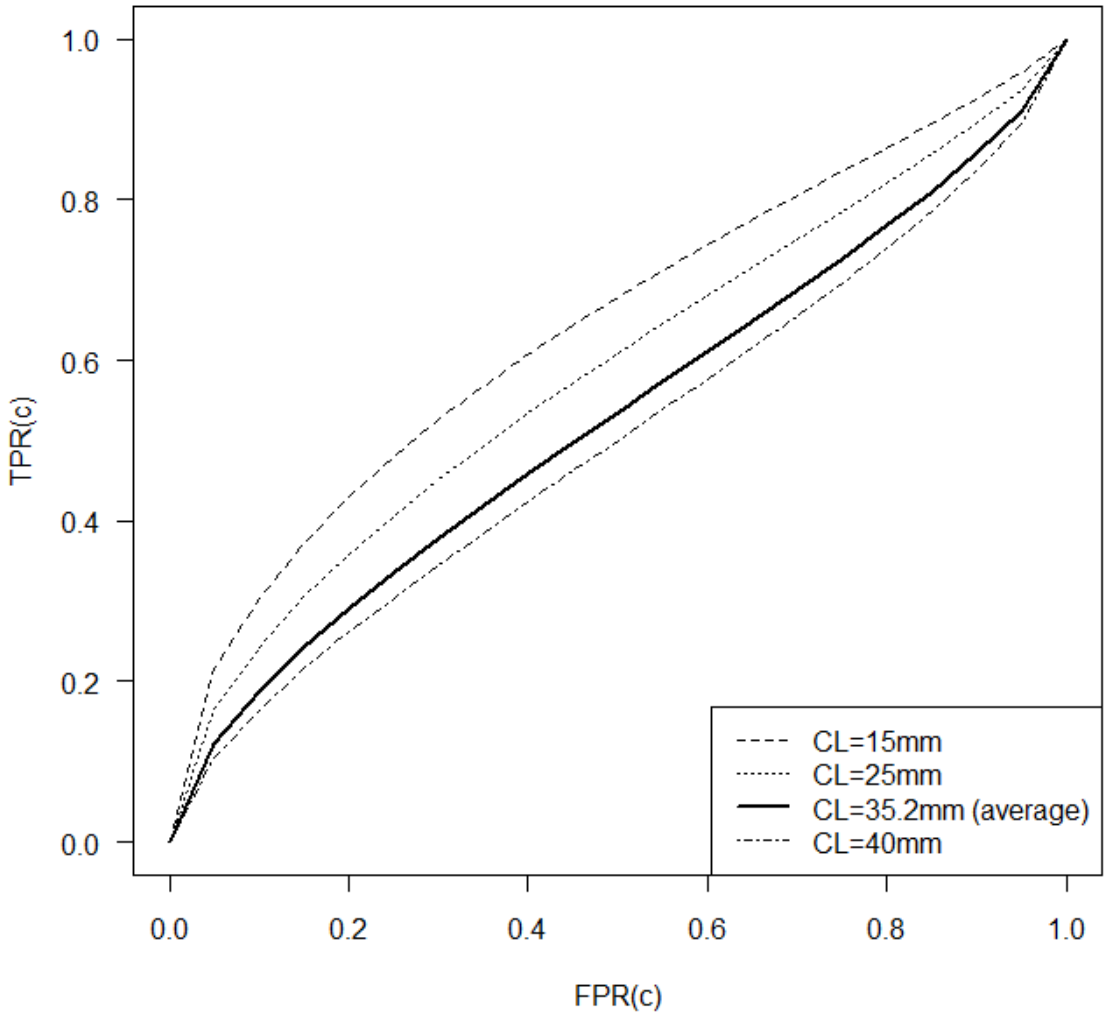


Figure 4.3: Preterm Prediction Study: SPEL-ROC by Cervical Length

The SPEL-ROC is calculated with a sample size of $n_V = 360$ and allocation given by $(n_0, n_1, n_2, n_3, n_{\bar{V}}) = (180, 60, 60, 60, 2627)$. CL indicates cervical length (mm) and c indicates all possible values of the screening test, FFN. Cutpoints for the TDS component are defined by $(a_1, a_2) = (\mu_Y - \alpha\sigma_Y, \mu_Y + \alpha\sigma_Y)$ where $\alpha = 1$.

# Chapter 5

# Conclusions

In this dissertation, we have used semi-parametric empirical likelihood methods to develop estimators for the area under the ROC curve (AUC), partial AUC, and the covariate-specific ROC curve. These tools help us study the ability of a screening test to discern between diseased and non-diseased populations. We use a test-dependent sampling (TDS) design where TDS inclusion depends on the continuous screening test measure. The TDS design incorporates an SRS component, a TDS component, and the remaining un-sampled portion of the population in which disease status is not validated. The TDS design allows investigators to over-sample subjects from specified ranges of the screening test variable, allowing for a concentration of resources where there is the greatest amount of information. This sampling design is particularly useful in studies similar to the lung cancer study described by Bueno et al. (2012), where the true disease status is expensive to ascertain due to the length of time needed to observe the outcome or the invasive procedures needed to validate disease status.

In Chapters 2 and 3, we developed semi-parametric empirical likelihood estimators for the area under the ROC curve (AUC) and the partial AUC, respectively. No distributional assumptions are made for the screening test and the disease status is not validated for the non-validation subjects in the un-sampled portion of the population. Simulation studies show that both the use of the TDS design and the inclusion of non-validation subjects give a more efficient alternative to the SRS designs and validation-only estimators when sampling the same number of subjects. Although all estimators compared are unbiased, the proposed AUC and pAUC estimators are shown to be the most efficient, compared to competing estimators. This suggests that to obtain the same variability, fewer subjects would be needed when the

proposed method is used, reducing study cost and subject burden. Analysis of data from the lung cancer study and the Preterm Prediction Study show that the proposed AUC and pAUC estimators are effective tools for discerning between the two outcome populations.

In Chapter 4, we developed a semi-parametric estimator for the covariate-specific ROC curve. Empirical likelihood methods were used to estimate the ROC curve without making assumptions on the distribution of the covariates. Normality of the screening test variable is assumed for this estimator. Simulation studies support the idea that inclusion of the un-sampled subjects, in which the disease status is not validated, improves efficiency compared to the estimator which uses only sampled subjects will full data available. Analysis of data from the lung cancer study and the Preterm Prediction Study show the utility of the covariate-specific ROC curve estimator by showing that by including covariates in the model, the screening test is more effective at discerning disease status for certain subsets of the population.

The TDS design is shown to improve efficiency of the proposed estimators compared to current estimators both in simulation studies and in analyzing data from the lung cancer study described by Bueno et al. (2012) and data from the Preterm Prediction Study (Goldenberg et al., 1996). The proposed estimators utilize empirical likelihood methods to reduce the distributional assumptions needed to estimate AUC, pAUC, and the covariate-specific ROC curve. Future research for the covariate-specific ROC curve will include relaxation of distributional assumptions of the screening test variable.

# APPENDIX A

## Asymptotic results for the SPEL-AUC

In this section we develop the asymptotic distribution of the SPEL-AUC, given by $\hat{A}^P_{V,\bar{V}}$ in (2.14). First we show the asymptotic distribution of $\eta = (p, \alpha, \beta, \lambda_2)$ in Section A.1. Next, we derive the asymptotic distribution of $\hat{R}_N(A, \eta)$ in Section A.2, which is used to find the asymptotic distribution of the proposed estimator, $\hat{A}^P_{V,\bar{V}}$.

### A.1 Asymptotic properties of $\eta = (p, \beta, \alpha, \lambda)$

#### A.1.1 Asymptotic distribution for $\xi = (\alpha, \beta, \lambda_2)$

The Newton-Raphson algorithm was used to construct estimators for $\xi = (\alpha, \beta, \lambda_2)$. The estimator for $p$ was constructed independently of $\xi$. Consider the profile log-likelihood function

$$\begin{aligned} pl(\xi) \quad \propto \quad & -\sum_{ij} \ln\left[1 - \lambda_2\left(e^{\alpha+\beta y_{ij}} - 1\right)\right] + n_{V,D=0}\alpha + \beta \sum_{\substack{i,j\in V \\ D=0}} Y_{ij} \\ & + \sum_{j=1}^{n_{\bar{V}}} \ln\left[p + e^{\alpha+\beta Y_{\bar{V}j}}(1-p)\right]. \end{aligned} \tag{A.1}$$

Define $\mathbf{H}_{ij}(\eta)$ such that the derivative of the profile likelihood (A.1) is given by $\frac{\partial pl(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}} = \sum_{l=1}^{N} \mathbf{H}_{ij}(\eta)$. As $N \to \infty$, $\frac{n_i}{N} \to \rho_i$ for $i = \left(0, 1, 2, 3, \bar{V}\right)$. Consider the Taylor expansion of

$\frac{\partial pl(\xi)}{\partial \xi}|_{\xi=\hat{\xi}}$ at $\xi$:

$$\begin{aligned}
\frac{\partial pl\left(\xi\right)}{\partial \xi}|_{\xi=\hat{\xi}} &= 0 = \frac{\partial pl\left(\xi\right)}{\partial \xi} + \left(\hat{\xi}-\xi\right)\frac{\partial^2 pl\left(\xi\right)}{\partial \xi_l \partial \xi_{l'}} + o_p(1) \\
\Rightarrow N^{1/2}\left(\hat{\xi}-\xi\right) &= N^{-1/2}\left[\frac{-1}{N}\frac{\partial^2 pl\left(\xi\right)}{\partial \xi_l \partial \xi_{l'}}\right]^{-1}\frac{\partial pl\left(\xi\right)}{\partial \xi} + o_p(1) \\
&= N^{-1/2}\left[\frac{-1}{N}\frac{\partial^2 pl\left(\xi\right)}{\partial \xi_l \partial \xi_{l'}}\right]^{-1}\sum_{i\in\left(0,1,2,3,\bar{V}\right)}\sum_{j=1}^{n_i}\mathbf{H}_{ij}\left(\eta\right) + o_p(1) \\
&= \sum_{i\in\left(0,1,2,3,\bar{V}\right)}\rho_i^{1/2} n_i^{-1/2}\left[\frac{-1}{N}\frac{\partial^2 pl\left(\xi\right)}{\partial \xi_l \partial \xi_{l'}}\right]^{-1}\sum_{j=1}^{n_i}\mathbf{H}_{ij}\left(\eta\right) + o_p(1). \quad \text{(A.2)}
\end{aligned}$$

***Lemma 1:*** Applying the central limit theorem to each term $n_i^{-1/2}\left[\frac{-1}{N}\frac{\partial^2 pl(\xi)}{\partial \xi_i \partial \xi_{i'}}\right]^{-1}\sum_{j=1}^{n_i}\mathbf{H}_{ij}\left(\eta\right)$ in (A.2), we have

$$\sqrt{N}\left(\hat{\xi}-\xi\right) \xrightarrow{d} \mathcal{N}(0,\psi_\xi), \quad \text{(A.3)}$$

where $\psi_\xi = \sum_{i\in\left(0,1,2,3,\bar{V}\right)}\rho_i^{1/2} var\left(\left[\frac{-1}{N}\frac{\partial^2 pl(\xi)}{\partial \xi_l \partial \xi_{l'}}\right]^{-1}\mathbf{H}_{ij}\left(\eta\right)\right)$.

### A.1.2 Asymptotic properties of $p$

The estimate of $p$ is found using the SRS portion of the test-dependent sample. Consider the likelihood and log-likelihood for the subjects sampled in the SRS portion, given by

$$\begin{aligned}
L_{SRS}\left(f_D\right) &= \prod_{j=1}^{n_0} f\left(Y_{0j}, D_{0j}\right) \\
&= \prod_{j:D=1} f\left(Y_{0j}|D_{0j}=1\right)p \times \prod_{j:D=0} f\left(Y_{0j}|D_{0j}=0\right)(1-p) \\
l_{SRS}\left(f_D\right) &= \sum_{j:D=1}\left\{\ln f\left(Y_{0j}|D_{0j}=1\right)+\ln p\right\} + \sum_{j:D=0}\left\{\ln f\left(Y_{0j}|D_{0j}=0\right)+\ln\left(1-p\right)\right\} \\
&\propto n_{0,D=1}\ln p + n_{0,D=0}\ln\left(1-p\right). \quad \text{(A.4)}
\end{aligned}$$

Define $P_{0j}$ such that $\frac{dl_{SRS}(p)}{dp} = \sum_{j=1}^{n_0} P_{0j}\left(p\right)$. Consider the Taylor expansion of $\frac{dl_{SRS}(p)}{dp}|_{p=\hat{p}}$

at $p$:

$$\begin{aligned}
\frac{dl_{SRS}(p)}{dp}|_{p=\hat{p}} = 0 &= \frac{dl_{SRS}(p)}{dp} + (\hat{p}-p)\frac{d^2 l_{SRS}(p)}{dp^2} + o_p(1) \\
&= \sum_{j=1}^{n_0} P_{0j} + (\hat{p}-p)\sum_{j=1}^{n_0}\frac{d}{dp}P_{0j} + o_p(1) \\
\Rightarrow \qquad n_0^{1/2}(\hat{p}-p) &= n_0^{-1/2}\left[\frac{-1}{n_0}\frac{\partial^2 l_{srs}(p)}{p^2}\right]^{-1}\sum_{l=1}^{n_0} P_{0j} + o_p(1) \qquad \text{(A.5)}
\end{aligned}$$

***Lemma 2:*** Applying the central limit theorem to A.5, we have

$$\sqrt{N}(\hat{p}-p) \xrightarrow{d} \mathcal{N}(0,\psi_p), \qquad \text{(A.6)}$$

where $\psi_p = \sum_{j=1}^{n_0}\rho_i^{1/2} var\left(\left[\frac{-1}{n_0}\frac{\partial^2 l_{srs}(p)}{p^2}\right]^{-1} P_{0j}\right)$.

## A.2 Asymptotic properties of $\hat{R}_N(A,\eta)$

In developing the asymptotic distribution of $\hat{A}^P_{V,\bar{V}}$, we assume that the U-statistic, $U_N(A,\eta) = R_N(A,\eta) - E(R_N(A,\eta))$ is stochastically equicontinuous. In this section we derive the asymptotic distribution of $R_N(A,\eta)$ and use this to develop the asymptotic distribution of the SPEL-AUC, given by $\hat{A}^P_{V,\bar{V}}$. Suppose $A$ is the true AUC value and $\eta = (p,\alpha,\beta,\lambda_2)$. To show asymptotic normality of $\hat{A}^P_{V,\bar{V}}$, consider the two-sample U-process, $U_N(A,\eta) = R_N(A,\eta) - E(R_N(A,\eta))$, where $R_N(A,\eta) = \frac{1}{N^2}\sum_{i\neq j} D'_i\left(1-D'_j\right)(I_{ij}-A)$, $I_{ij} = I(Y_i > Y_j)$, and

$$D'_l = \begin{cases} D_l & \text{if } l \in V \\ \widehat{E(D_l)} = \frac{p}{p+e^{\alpha+\beta y_l}(1-p)} & \text{if } l \in \bar{V} \end{cases}.$$

Now, consider the difference between the proposed AUC estimator and the true AUC. Next, we use this difference to solve for $R_N(A,\eta)$, given by:

$$\begin{aligned}
\hat{A}^P_{V,\bar{V}} - A &= \frac{\sum_{i\neq j} D'_i\left(1-D'_j\right)I_{ij}}{\sum_{i\neq j} D'_i\left(1-D'_j\right)} - A \\
&= \frac{N^2 R_N(A,\eta)}{\sum_{i\neq j} D'_i\left(1-D'_j\right)}
\end{aligned}$$

$$\Rightarrow R_N(A,\eta) = \frac{1}{N^2}\sum_{i\neq j} D_i'\left(1-D_j'\right)(I_{ij}-A)$$
$$= \left(\hat{A}^P_{V,\bar{V}} - A\right) \times \frac{1}{N^2}\sum_{i\neq j} D_i'\left(1-D_j'\right). \quad (A.7)$$

Let $R_{ij} = D_i'\left(1-D_j'\right)(I_{ij}-A)$. From *Asymptotic Statistics* (Chapter 12, van der Vaart), we know that

$$\hat{R}_N(A,\eta) = \frac{1}{N}\sum_i^N E\left[R_{ij}|Y_i,D_i\right] + \frac{1}{N}\sum_j^N E\left[R_{ij}|Y_j,D_j\right]$$
$$= \frac{1}{N}\sum_i^N E\left[R_{ij}|Y_i,D_i\right] + \frac{1}{N}\sum_i^N E\left[R_{ji}|Y_i,D_i\right]$$
$$= \frac{1}{N}\sum_i^N E\left[R_{ij}+R_{ji}|Y_i,D_i\right]. \quad (A.8)$$

***Lemma 3:*** From *Asymptotic Statistics* (Chapter 12, van der Vaart), if $E\left[\left(D_i'\left(1-D_j'\right)(I_{ij}-A)\right)^2\right] < \infty$ then $\sqrt{N}\left(R_N(A,\eta) - E\left[R_N(A,\eta)\right] - \hat{R}_N(A,\eta)\right) \xrightarrow{p} 0$ and

$$\sqrt{N}\left(R_N(A,\eta) - E\left[R_N(A,\eta)\right]\right) \xrightarrow{d} \mathcal{N}(0,\Sigma) \quad (A.9)$$

where $\Sigma = 4Var\left(R_{ij}\right)$.

Ultimately, we want to know the asymptotic distribution of $\sqrt{N}\left(\hat{A}^P_{V,\bar{V}} - A\right)$. This can be accomplished by considering the Taylor expansion of $\sqrt{N}R_N\left(\hat{A}^P_{V,\bar{V}},\hat{\eta}\right)$ at $(A,\eta)$ and using the asymptotic distribution of $R_N(A,\eta)$, given in (A.9). We expand $\sqrt{N}R_N\left(\hat{A}^P_{V,\bar{V}},\hat{\eta}\right)$ at $(A,\eta)$. First, note that $\sqrt{N}R_N\left(\hat{A}^P_{V,\bar{V}},\hat{\eta}\right) = \left(\hat{A}^P_{V,\bar{V}} - \hat{A}^P_{V,\bar{V}}\right) \times \sum_{i\neq j} D_i'\left(1-D_j'\right) = 0$. This

Taylor expansion is given by:

$$\begin{aligned}
&\sqrt{N}R_N\left(\hat{A}^P_{V,\bar{V}},\hat{\eta}\right)\\
&\quad=\sqrt{N}R_N\left(\hat{A}^P_{V,\bar{V}},\hat{\eta}\right)\pm\sqrt{N}R_N\left(A,\eta\right)\pm\sqrt{N}E\left[R_N\left(A,\eta\right)\right]\pm\sqrt{N}E\left[R_N\left(\hat{A}^P_{V,\bar{V}},\hat{\eta}\right)\right]\\
&\quad=\sqrt{N}\left\{R_N\left(\hat{A}^P_{V,\bar{V}},\hat{\eta}\right)-E\left[R_N\left(\hat{A}^P_{V,\bar{V}},\hat{\eta}\right)\right]\right\}-\sqrt{N}\left\{R_N\left(A,\eta\right)-E\left[R_N\left(A,\eta\right)\right]\right\}\\
&\qquad+\sqrt{N}E\left[R_N\left(\hat{A}^P_{V,\bar{V}},\hat{\eta}\right)\right]-\sqrt{N}E\left[R_N\left(A,\eta\right)\right]+\sqrt{N}R_N\left(A,\eta\right)\\
&\quad=\sqrt{N}E\left[R_N\left(\hat{A}^P_{V,\bar{V}},\hat{\eta}\right)\right]-\sqrt{N}E\left[R_N\left(A,\eta\right)\right]+\sqrt{N}R_N\left(A,\eta\right).
\end{aligned}\tag{A.10}$$

Since we assume that the U-process, $U_N\left(A,\eta\right)=\sqrt{n}\left\{R_N\left(A,\eta\right)-E\left[R_N\left(A,\eta\right)\right]\right\}$ is equicontinuous, then

$$\begin{aligned}
&\sqrt{N}\left\{R_N\left(\hat{A}^P_{V,\bar{V}},\hat{\eta}\right)-E\left[R_N\left(\hat{A}^P_{V,\bar{V}},\hat{\eta}\right)\right]\right\}\to 0 \text{ and}\\
&\sqrt{N}\left\{R_N\left(A,\eta\right)-E\left[R_N\left(A,\eta\right)\right]\right\}\to 0.
\end{aligned}$$

Now that we have simplified $\sqrt{N}R_N\left(\hat{A}^P_{V,\bar{V}},\hat{\eta}\right)$, we can expand $E\left[R_N\left(\hat{A}^P_{V,\bar{V}},\hat{\eta}\right)\right]$ at $(A,\eta)$ within (A.10), which gives us:

$$\begin{aligned}
&\sqrt{N}R_N\left(\hat{A}^P_{V,\bar{V}},\hat{\eta}\right)\\
&\quad=\sqrt{N}E\left[R_N\left(\hat{A}^P_{V,\bar{V}},\hat{\eta}\right)\right]-\sqrt{N}E\left[R_N\left(A,\eta\right)\right]+\sqrt{N}R_N\left(A,\eta\right)\\
&\quad=\sqrt{N}\left\{\,E\left[R_N\left(A,\eta\right)\right]+\frac{\partial E\left[R_N\left(A,\eta\right)\right]}{\partial A}\left(\hat{A}^P_{V,\bar{V}}-A\right)\right.\\
&\qquad\left.+\left[\frac{\partial E\left[R_N\left(A,\eta\right)\right]}{\partial \eta}\right]^T_{1x4}\left(\hat{\eta}^P-\eta\right)_{4x1}\right\}\\
&\qquad-\sqrt{N}E\left[R_N\left(A,\eta\right)\right]+\sqrt{N}R_N\left(A,\eta\right)+o_p\left(1\right)\\
&\quad=\sqrt{N}\left\{\frac{\partial E\left[R_N\left(A,\eta\right)\right]}{\partial A}\left(\hat{A}^P_{V,\bar{V}}-A\right)+\left[\frac{\partial E\left[R_N\left(A,\eta\right)\right]}{\partial \eta}\right]^T_{1x4}\left(\hat{\eta}^P-\eta\right)_{4x1}\right\}\\
&\qquad+\sqrt{N}R_N\left(A,\eta\right)+o_p\left(1\right).
\end{aligned}\tag{A.11}$$

To derive the asymptotic distribution of $\sqrt{N}\left(\hat{A}^P_{V,\bar{V}}-A\right)$, we solve for $\left(\hat{A}^P_{V,\bar{V}}-A\right)$ in the final expression of $\sqrt{N}R_N\left(\hat{A}^P_{V,\bar{V}},\hat{\eta}\right)$ in (A.11). Then, we use (A.2) and (A.5) to estimate the

asymptotic variance. We express $\sqrt{N}\left(\hat{A}^{P}_{V,\bar{V}} - A\right)$ as:

$$\begin{aligned}
&\sqrt{N}\left(\hat{A}^{P}_{V,\bar{V}} - A\right) \\
&\quad = -\left\{\frac{\partial E\left[R_N\left(A,\eta\right)\right]}{\partial A}\right\}^{-1}\sqrt{N}\left[R_N\left(A,\eta\right) + \left[\frac{\partial E\left[R_N\left(A,\eta\right)\right]}{\partial \eta}\right]^{T}_{1x4}\left(\hat{\eta}-\eta\right)_{4x1}\right] \\
&\quad = -\left\{\frac{\partial E\left[R_N\left(A,\eta\right)\right]}{\partial A}\right\}^{-1}\sqrt{N} \\
&\qquad \times\left[R_N\left(A,\eta\right) + \frac{\partial E\left[R_N\left(A,\eta\right)\right]}{\partial p}\left(\hat{p}-p\right) + \left[\frac{\partial E\left[R_N\left(A,\eta\right)\right]}{\partial \xi}\right]^{T}_{1x3}\left(\hat{\xi}-\xi\right)_{3x1}\right] \\
&\quad = -\left\{\frac{\partial E\left[R_N\left(A,\eta\right)\right]}{\partial A}\right\}^{-1}\sum_{i\in\left(0,1,2,3,\bar{V}\right)}\rho_i n_i^{-1/2}\sum_{j=1}^{n_i}Q_{ij}.
\end{aligned}$$

We can re-express $R_{ll'} = R_{(ij)(ij)'}$ so that the above express can be written as a double sum. Define $Q_{ij}$ as:

$$\begin{aligned}
Q_{ij}\left(\eta\right) &= E\left(R_{(ij)(ij)'} + R_{(ij)'(ij)}\right) + \rho_i^{-1}\frac{\partial E R_N\left(A,\eta\right)}{\partial p}\left[\frac{-1}{n_0}\frac{\partial^2 l_{srs}\left(p\right)}{p^2}\right]^{-1}P_{0j}I\left(i=0\right) \\
&\qquad + \frac{\partial E R_N\left(A,\eta\right)}{\partial \xi}\left[\frac{-1}{N}\frac{\partial^2 pl\left(\xi\right)}{\partial\xi_i\partial\xi_{i'}}\right]^{-1}\mathbf{H}_{ij}\left(\eta\right) \\
\text{and } \hat{Q}_{ij} &= \hat{E}\left(R_{(ij)(ij)'} + R_{(ij)'(ij)}\right) + \hat{\rho}^{-1}\hat{E}\frac{\partial R_N\left(A,\eta\right)}{\partial p}\left[\frac{1}{n_0}\frac{\partial^2 l_{srs}\left(p\right)}{p^2}\right]|^{-1}_{p=\hat{p}}P_{0j}\left(\hat{p}\right) \\
&\qquad + \hat{E}\frac{\partial R_N\left(A,\eta\right)}{\partial \xi}\left[\frac{-1}{N}\frac{\partial^2 pl\left(\xi\right)}{\partial\xi_i\partial\xi_{i'}}\right]|^{-1}_{\xi=\hat{\xi}}\mathbf{H}_l\left(\hat{\xi}\right).
\end{aligned}$$

# APPENDIX B

## Asymptotic results for the SPEL-pAUC

In this section we develop the asymptotic distribution of the SPEL-pAUC, given by $\hat{A}^P_{t:V,\bar{V}}$ in (3.15). First we show the asymptotic distribution of $\eta = (p, \alpha, \beta, \lambda_2)$ in section B.1. Next, we derive the asymptotic distribution of $\hat{R}_N(A_t, \eta)$ in section B.2, which is used to find the asymptotic distribution of the proposed estimator, $\hat{A}^P_{t:V,\bar{V}}$.

### B.1 Asymptotic distribution for $\eta = (p, \beta, \alpha, \lambda)$

#### B.1.1 Asymptotic distribution for $\xi = (\alpha, \beta, \lambda_2)$

The Newton-Raphson algorithm was used to construct estimators for $\xi = (\alpha, \beta, \lambda_2)$. The estimator for $p$ was constructed independently of $\xi$. Consider the profile log-likelihood function

$$\begin{aligned} pl(\xi) \quad \propto \quad & -\sum_{ij} \ln\left[1 - \lambda_2\left(e^{\alpha+\beta y_{ij}} - 1\right)\right] + n_{V,D=0}\alpha + \beta \sum_{\substack{i,j \in V \\ D=0}} Y_{ij} \\ & + \sum_{j=1}^{n_{\bar{V}}} \ln\left[p + e^{\alpha+\beta Y_{\bar{V}j}}(1-p)\right]. \end{aligned} \tag{B.1}$$

Define $\mathbf{H}_{ij}(\eta)$ such that the derivative of the profile likelihood (B.1) is given by $\frac{\partial pl(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}} = \sum_{l=1}^{N} \mathbf{H}_{ij}(\eta)$. As $N \to \infty$, $\frac{n_i}{N} \to \rho_i$ for $i = \left(0, 1, 2, 3, \bar{V}\right)$. Consider the Taylor expansion of

$\frac{\partial pl(\xi)}{\partial \xi}|_{\xi=\hat{\xi}}$ at $\xi$:

$$\begin{aligned}
\frac{\partial pl\left(\xi\right)}{\partial \xi}|_{\xi=\hat{\xi}} &= 0 = \frac{\partial pl\left(\xi\right)}{\partial \xi} + \left(\hat{\xi}-\xi\right)\frac{\partial^2 pl\left(\xi\right)}{\partial \xi_l \partial \xi_{l'}} + o_p(1) \\
\Rightarrow N^{1/2}\left(\hat{\xi}-\xi\right) &= N^{-1/2}\left[\frac{-1}{N}\frac{\partial^2 pl\left(\xi\right)}{\partial \xi_l \partial \xi_{l'}}\right]^{-1}\frac{\partial pl\left(\xi\right)}{\partial \xi} + o_p(1) \\
&= N^{-1/2}\left[\frac{-1}{N}\frac{\partial^2 pl\left(\xi\right)}{\partial \xi_l \partial \xi_{l'}}\right]^{-1}\sum_{i\in\left(0,1,2,3,\bar{V}\right)}\sum_{j=1}^{n_i}\mathbf{H}_{ij}\left(\eta\right) + o_p(1) \\
&= \sum_{i\in\left(0,1,2,3,\bar{V}\right)}\rho_i^{1/2} n_i^{-1/2}\left[\frac{-1}{N}\frac{\partial^2 pl\left(\xi\right)}{\partial \xi_l \partial \xi_{l'}}\right]^{-1}\sum_{j=1}^{n_i}\mathbf{H}_{ij}\left(\eta\right) + o_p(1). \quad \text{(B.2)}
\end{aligned}$$

***Lemma 1:*** Applying the central limit theorem to each term $n_i^{-1/2}\left[\frac{-1}{N}\frac{\partial^2 pl(\xi)}{\partial \xi_i \partial \xi_{i'}}\right]^{-1}\sum_{j=1}^{n_i}\mathbf{H}_{ij}\left(\eta\right)$ in (A.2), we have

$$\sqrt{N}\left(\hat{\xi}-\xi\right) \xrightarrow{d} \mathcal{N}(0,\psi_\xi), \quad \text{(B.3)}$$

where $\psi_\xi = \sum_{i\in\left(0,1,2,3,\bar{V}\right)}\rho_i^{1/2} var\left(\left[\frac{-1}{N}\frac{\partial^2 pl(\xi)}{\partial \xi_l \partial \xi_{l'}}\right]^{-1}\mathbf{H}_{ij}\left(\eta\right)\right)$.

### B.1.2 Asymptotic distribution for $p$

The estimate of $p$ is found using the SRS portion of the test-dependent sample. Consider the likelihood and log-likelihood for the subjects sampled in the SRS portion, given by

$$\begin{aligned}
L_{SRS}\left(f_D\right) &= \prod_{j=1}^{n_0} f\left(Y_{0j}, D_{0j}\right) \\
&= \prod_{j:D=1} f\left(Y_{0j}|D_{0j}=1\right)p \times \prod_{j:D=0} f\left(Y_{0j}|D_{0j}=0\right)(1-p) \\
l_{SRS}\left(f_D\right) &= \sum_{j:D=1}\left\{\ln f\left(Y_{0j}|D_{0j}=1\right)+\ln p\right\} + \sum_{j:D=0}\left\{\ln f\left(Y_{0j}|D_{0j}=0\right)+\ln\left(1-p\right)\right\} \\
&\propto n_{0,D=1}\ln p + n_{0,D=0}\ln\left(1-p\right). \quad \text{(B.4)}
\end{aligned}$$

Define $P_{0j}$ such that $\frac{dl_{SRS}(p)}{dp} = \sum_{j=1}^{n_0} P_{0j}\left(p\right)$. Consider the Taylor expansion of $\frac{dl_{SRS}(p)}{dp}|_{p=\hat{p}}$

at $p$:

$$\begin{aligned}
\frac{dl_{SRS}(p)}{dp}|_{p=\hat{p}} = 0 &= \frac{dl_{SRS}(p)}{dp} + (\hat{p}-p)\frac{d^2 l_{SRS}(p)}{dp^2} + o_p(1) \\
&= \sum_{j=1}^{n_0} P_{0j} + (\hat{p}-p)\sum_{j=1}^{n_0}\frac{d}{dp}P_{0j} + o_p(1) \\
\Rightarrow \qquad n_0^{1/2}(\hat{p}-p) &= n_0^{-1/2}\left[\frac{-1}{n_0}\frac{\partial^2 l_{srs}(p)}{p^2}\right]^{-1}\sum_{l=1}^{n_0} P_{0j} + o_p(1) \qquad \text{(B.5)}
\end{aligned}$$

***Lemma 2:*** Applying the central limit theorem to A.5, we have

$$\sqrt{N}(\hat{p}-p) \xrightarrow{d} \mathcal{N}(0,\psi_p), \qquad \text{(B.6)}$$

where $\psi_p = \sum_{j=1}^{n_0}\rho_i^{1/2} var\left(\left[\frac{-1}{n_0}\frac{\partial^2 l_{srs}(p)}{p^2}\right]^{-1} P_{0j}\right)$.

## B.2 Asymptotic distribution of $\hat{R}_N(A_t,\eta)$

In developing the asymptotic distribution of $\hat{A}^P_{t:V,\bar{V}}$, we assume that the U-statistic, $U_N(A_t,\eta) = R_N(A_t,\eta) - E(R_N(A_t,\eta))$ is stochastically equicontinuous. In this section we derive the asymptotic distribution of $R_N(A_t,\eta)$ and use this to develop the asymptotic distribution of the SPEL-pAUC, given by $\hat{A}^P_{t:V,\bar{V}}$. Suppose $A_t$ is the true AUC value and $\eta = (p,\alpha,\beta,\lambda_2)$. To show asymptotic normality of $\hat{A}^P_{t:V,\bar{V}}$, consider the two-sample U-process, $U_N(A_t,\eta) = R_N(A_t,\eta) - E(R_N(A_t,\eta))$, where $R_N(A_t,\eta) = \frac{1}{N^2}\sum_{i\neq j} D'_i\left(1-D'_j\right)(I_{t:ij} - A_t)$, $I_{t:ij} = I(Y_i > Y_j, Y_j \in (t_0,t_1))$, and $D'_l = \begin{cases} D_l & \text{if } l \in V \\ \widehat{E(D_l)} = \frac{p}{p+e^{\alpha+\beta y_l}(1-p)} & \text{if } l \in \bar{V} \end{cases}$.

Now, consider the difference between the proposed pAUC estimator and the true pAUC. Next, we use this difference to solve for $R_N(A_t,\eta)$, given by:

$$\begin{aligned}
\hat{A}^P_{t:V,\bar{V}} - A_t &= \frac{\sum_{i\neq j} D'_i\left(1-D'_j\right) I_{t:ij}}{\sum_{i\neq j} D'_i\left(1-D'_j\right)} - A_t \\
&= \frac{N^2 R_N(A_t,\eta)}{\sum_{i\neq j} D'_i\left(1-D'_j\right)}
\end{aligned}$$

$$\begin{aligned}
\Rightarrow R_N(A_t,\eta) &= \frac{1}{N^2}\sum_{i\neq j} D_i'\left(1-D_j'\right)\left(I_{t:ij}-A_t\right) \\
&= \left(\hat{A}^P_{t:V,\bar{V}}-A_t\right)\times\frac{1}{N^2}\sum_{i\neq j} D_i'\left(1-D_j'\right). \qquad \text{(B.7)}
\end{aligned}$$

Let $R_{ij} = D_i'\left(1-D_j'\right)\left(I_{t:ij}-A_t\right)$. From *Asymptotic Statistics* (Chapter 12, van der Vaart), we know that

$$\begin{aligned}
\hat{R}_N(A_t,\eta) &= \frac{1}{N}\sum_i^N E\left[R_{ij}|Y_i,D_i\right]+\frac{1}{N}\sum_j^N E\left[R_{ij}|Y_j,D_j\right] \\
&= \frac{1}{N}\sum_i^N E\left[R_{ij}|Y_i,D_i\right]+\frac{1}{N}\sum_i^N E\left[R_{ji}|Y_i,D_i\right] \\
&= \frac{1}{N}\sum_i^N E\left[R_{ij}+R_{ji}|Y_i,D_i\right]. \qquad \text{(B.8)}
\end{aligned}$$

***Lemma 3:*** From *Asymptotic Statistics* (Chapter 12, van der Vaart), if $E\left[\left(D_i'\left(1-D_j'\right)\left(I_{t:ij}-A_t\right)\right)^2\right]<\infty$ then $\sqrt{N}\left(R_N(A_t,\eta)-E\left[R_N(A_t,\eta)\right]-\hat{R}_N(A_t,\eta)\right)\xrightarrow{p}0$. Consequently,

$$\sqrt{N}\left(R_N(A_t,\eta)-E\left[R_N(A_t,\eta)\right]\right)\xrightarrow{d}\mathcal{N}(0,\Sigma) \qquad \text{(B.9)}$$

where $\Sigma = 4Var\left(R_{ij}\right)$.

Ultimately, we want to know the asymptotic distribution of $\sqrt{N}\left(\hat{A}^P_{t:V,\bar{V}}-A_t\right)$. This can be accomplished by considering the Taylor expansion of $\sqrt{N}R_N\left(\hat{A}^P_{t:V,\bar{V}},\hat{\eta}\right)$ at $(A_t,\eta)$ and using the asymptotic distribution of $R_N(A_t,\eta)$, given in ((B.9)). We expand $\sqrt{N}R_N\left(\hat{A}^P_{t:V,\bar{V}},\hat{\eta}\right)$ at $(A_t,\eta)$. First, note that $\sqrt{N}R_N\left(\hat{A}^P_{t:V,\bar{V}},\hat{\eta}\right)=\left(\hat{A}^P_{t:V,\bar{V}}-\hat{A}^P_{t:V,\bar{V}}\right)\times\sum_{i\neq j}D_i'\left(1-D_j'\right)=0$.

This Taylor expansion is given by:

$$\begin{aligned}
&\sqrt{N}R_N\left(\hat{A}^P_{t:V,\bar{V}},\hat{\eta}\right)\\
&\quad=\sqrt{N}R_N\left(\hat{A}^P_{t:V,\bar{V}},\hat{\eta}\right)\pm\sqrt{N}R_N\left(A_t,\eta\right)\pm\sqrt{N}E\left[R_N\left(A_t,\eta\right)\right]\pm\sqrt{N}E\left[R_N\left(\hat{A}^P_{t:V,\bar{V}},\hat{\eta}\right)\right]\\
&\quad=\sqrt{N}\left\{R_N\left(\hat{A}^P_{t:V,\bar{V}},\hat{\eta}\right)-E\left[R_N\left(\hat{A}^P_{t:V,\bar{V}},\hat{\eta}\right)\right]\right\}-\sqrt{N}\left\{R_N\left(A_t,\eta\right)-E\left[R_N\left(A_t,\eta\right)\right]\right\}\\
&\qquad+\sqrt{N}E\left[R_N\left(\hat{A}^P_{t:V,\bar{V}},\hat{\eta}\right)\right]-\sqrt{N}E\left[R_N\left(A_t,\eta\right)\right]+\sqrt{N}R_N\left(A_t,\eta\right)\\
&\quad=\sqrt{N}E\left[R_N\left(\hat{A}^P_{t:V,\bar{V}},\hat{\eta}\right)\right]-\sqrt{N}E\left[R_N\left(A_t,\eta\right)\right]+\sqrt{N}R_N\left(A_t,\eta\right).
\end{aligned}\tag{B.10}$$

Since we assume that the U-process, $U_N\left(A_t,\eta\right)=\sqrt{n}\left\{R_N\left(A_t,\eta\right)-E\left[R_N\left(A_t,\eta\right)\right]\right\}$ is equicontinuous, then

$$\begin{aligned}
&\sqrt{N}\left\{R_N\left(\hat{A}^P_{t:V,\bar{V}},\hat{\eta}\right)-E\left[R_N\left(\hat{A}^P_{t:V,\bar{V}},\hat{\eta}\right)\right]\right\}\to 0 \text{ and}\\
&\sqrt{N}\left\{R_N\left(A_t,\eta\right)-E\left[R_N\left(A_t,\eta\right)\right]\right\}\to 0.
\end{aligned}$$

Now that we have simplified $\sqrt{N}R_N\left(\hat{A}^P_{t:V,\bar{V}},\hat{\eta}\right)$, we can expand $E\left[R_N\left(\hat{A}^P_{t:V,\bar{V}},\hat{\eta}\right)\right]$ at $\left(A_t,\eta\right)$ within (1.23), which gives us:

$$\begin{aligned}
&\sqrt{N}R_N\left(\hat{A}^P_{t:V,\bar{V}},\hat{\eta}\right)\\
&\quad=\sqrt{N}E\left[R_N\left(\hat{A}^P_{t:V,\bar{V}},\hat{\eta}\right)\right]-\sqrt{N}E\left[R_N\left(A_t,\eta\right)\right]+\sqrt{N}R_N\left(A_t,\eta\right)\\
&\quad=\sqrt{N}\left\{\ E\left[R_N\left(A_t,\eta\right)\right]+\frac{\partial E\left[R_N\left(A_t,\eta\right)\right]}{\partial A_t}\left(\hat{A}^P_{t:V,\bar{V}}-A_t\right)\right.\\
&\qquad\left.+\left[\frac{\partial E\left[R_N\left(A_t,\eta\right)\right]}{\partial \eta}\right]^T_{1x4}\left(\hat{\eta}^P-\eta\right)_{4x1}\ \right\}\\
&\qquad-\sqrt{N}E\left[R_N\left(A_t,\eta\right)\right]+\sqrt{N}R_N\left(A_t,\eta\right)+o_p\left(1\right)\\
&\quad=\sqrt{N}\left\{\frac{\partial E\left[R_N\left(A_t,\eta\right)\right]}{\partial A}\left(\hat{A}^P_{t:V,\bar{V}}-A_t\right)+\left[\frac{\partial E\left[R_N\left(A_t,\eta\right)\right]}{\partial \eta}\right]^T_{1x4}\left(\hat{\eta}^P-\eta\right)_{4x1}\right\}\\
&\qquad+\sqrt{N}R_N\left(A_t,\eta\right)+o_p\left(1\right).
\end{aligned}\tag{B.11}$$

To derive the asymptotic distribution of $\sqrt{N}\left(\hat{A}^P_{t:V,\bar{V}}-A_t\right)$, we solve for $\left(\hat{A}^P_{t:V,\bar{V}}-A_t\right)$ in the final expression of $\sqrt{N}R_N\left(\hat{A}^P_{t:V,\bar{V}},\hat{\eta}\right)$ in (B.11). Then, we use (B.2) and (B.5) to estimate

the asymptotic variance. We express $\sqrt{N}\left(\hat{A}^{P}_{t:V,\bar{V}} - A_t\right)$ as:

$$\begin{aligned}
&\sqrt{N}\left(\hat{A}^{P}_{t:V,\bar{V}} - A_t\right) \\
&\quad = -\left\{\frac{\partial E\left[R_N\left(A_t,\eta\right)\right]}{\partial A}\right\}^{-1}\sqrt{N}\left[R_N\left(A_t,\eta\right) + \left[\frac{\partial E\left[R_N\left(A_t,\eta\right)\right]}{\partial \eta}\right]^{T}_{1x4}\left(\hat{\eta}-\eta\right)_{4x1}\right] \\
&\quad = -\left\{\frac{\partial E\left[R_N\left(A_t,\eta\right)\right]}{\partial A}\right\}^{-1}\sqrt{N} \\
&\qquad \times\left[R_N\left(A_t,\eta\right) + \frac{\partial E\left[R_N\left(A_t,\eta\right)\right]}{\partial p}\left(\hat{p}-p\right) + \left[\frac{\partial E\left[R_N\left(A_t,\eta\right)\right]}{\partial \xi}\right]^{T}_{1x3}\left(\hat{\xi}-\xi\right)_{3x1}\right] \\
&\quad = -\left\{\frac{\partial E\left[R_N\left(A_t,\eta\right)\right]}{\partial A}\right\}^{-1}\sum_{i\in\left(0,1,2,3,\bar{V}\right)}\rho_i n_i^{-1/2}\sum_{j=1}^{n_i}Q_{ij}.
\end{aligned}$$

We can re-express $R_{ll'} = R_{(ij)(ij)'}$ so that the above express can be written as a double sum. Define $Q_{ij}$ as:

$$\begin{aligned}
Q_{ij}\left(\eta\right) &= E\left(R_{(ij)(ij)'} + R_{(ij)'(ij)}\right) + \rho_i^{-1}\frac{\partial E R_N\left(A,\eta\right)}{\partial p}\left[\frac{-1}{n_0}\frac{\partial^2 l_{srs}\left(p\right)}{p^2}\right]^{-1}P_{0j}I\left(i=0\right) \\
&\qquad + \frac{\partial E R_N\left(A,\eta\right)}{\partial \xi}\left[\frac{-1}{N}\frac{\partial^2 pl\left(\xi\right)}{\partial\xi_i\partial\xi_{i'}}\right]^{-1}\mathbf{H}_{ij}\left(\eta\right) \\
\text{and } \hat{Q}_{ij} &= \hat{E}\left(R_{(ij)(ij)'} + R_{(ij)'(ij)}\right) + \hat{\rho}^{-1}\hat{E}\frac{\partial R_N\left(A,\eta\right)}{\partial p}\left[\frac{1}{n_0}\frac{\partial^2 l_{srs}\left(p\right)}{p^2}\right]|^{-1}_{p=\hat{p}}P_{0j}\left(\hat{p}\right) \\
&\qquad + \hat{E}\frac{\partial R_N\left(A,\eta\right)}{\partial \xi}\left[\frac{-1}{N}\frac{\partial^2 pl\left(\xi\right)}{\partial\xi_i\partial\xi_{i'}}\right]|^{-1}_{\xi=\hat{\xi}}\mathbf{H}_l\left(\hat{\xi}\right).
\end{aligned}$$

# BIBLIOGRAPHY


Antolini, L., Boracchi, P., and Biganzoli, E. (2005). A time-dependent discrimination index for survival data. *Statistics in medicine*, 24(24):3927–3944.

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

Bastek, J. A. and Elovitz, M. A. (2013). The role and challenges of biomarkers in spontaneous preterm birth and preeclampsia. *Fertility and sterility*, 99(4):1117–1123.

Blanchon, F., Grivaux, M., Asselain, B., Lebas, F.-X., Orlando, J.-P., Piquet, J., and Zureik, M. (2006). 4-year mortality in patients with non-small-cell lung cancer: development and validation of a prognostic index. *The lancet oncology*, 7(10):829–836.

Breslow, N. and Cain, K. (1988). Logistic regression for two-stage case-control data. *Biometrika*, 75(1):11–20.

Bueno, R., Wang, X., Richards, W. G., Harpole, D. H., and Kratzke, R. (2012). Validation of molecular prognostic tests in nsclc: A companion study to calgb 140202. *CALGB 150807 Protocol*.

Cai, T. and Pepe, M. (2002). Semiparametric receiver operating characteristic analysis to evaluate biomarkers for disease. *Journal of the American Statistical Association*, 97(460):1099–1107.

Chatterjee, N., Chen, Y., and Breslow, N. (2003). A pseudoscore estimator for regression problems with two-phase sampling. *Journal of the American Statistical Association*, 98(461):158–168.

Deshpande, S., van Asselt, A., Tomini, F., Armstrong, N., Allen, A., Noake, C., Khan, K., Severens, J., Kleijnen, J., and Westwood, M. (2013). Rapid fetal fibronectin testing to predict preterm birth in women with symptoms of premature labour: a systematic review and cost analysis. *Health Technol Assess*, 17(40):1–138.

Dodd, L. and Pepe, M. (2003a). Partial auc estimation and regression. *Biometrics*, 59(3):614–623.

Dodd, L. and Pepe, M. (2003b). Semiparametric regression for the area under the receiver operating characteristic curve. *Journal of the American Statistical Association*, 98(462):409–417.

Efron, B. and Tibshirani, R. (1993). *An introduction to the bootstrap*, volume 57. CRC press.

Fears, T. and Brown, C. (1986). Logistic regression methods for retrospective case-control studies using complex sampling procedures. *Biometrics*, pages 955–960.

Goldenberg, R. L., Mercer, B. M., MEIS, P. L., Copper, R. L., Das, A., McNELLIS, D., et al. (1996). The preterm prediction study: fetal fibronectin testing and spontaneous preterm birth. *Obstetrics & Gynecology*, 87(5, Part 1):643–648.

Greenhouse, S. and Mantel, N. (1950). The evaluation of diagnostic tests. *Biometrics*, 6(4):399–412.

Hanley, J. (1989). Receiver operating characteristic (roc) methodology: the state of the art. *Critical reviews in diagnostic imaging*, 29(3):307–335.

Hanley, J., McNeil, B., et al. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148(3):839–843.

Heagerty, P. and Zheng, Y. (2005). Survival model predictive accuracy and roc curves. *Biometrics*, 61(1):92–105.

Hsieh, D., Manski, C., and McFadden, D. (1985). Estimation of response probabilities from augmented retrospective observations. *Journal of the American Statistical Association*, pages 651–662.

Huang, X., Qin, G., and Fang, Y. (2011). Optimal combinations of diagnostic tests based on auc. *Biometrics*, 67(2):568–576.

Liu, D. and Zhou, X. (2011). Semiparametric estimation of the covariate-specific roc curve in presence of ignorable verification bias. *Biometrics*, 67(3):906–916.

Lockwood, C. J., Senyei, A. E., Dische, M. R., Casal, D., Shah, K. D., Thung, S. N., Jones, L., Deligdisgh, L., and Garite, T. J. (1991). Fetal fibronectin in cervical and vaginal secretions as a predictor of preterm delivery. *New England Journal of Medicine*, 325(10):669–674.

Long, Q., Zhang, X., and Hsu, C. (2011a). Nonparametric multiple imputation for receiver operating characteristics analysis when some biomarker values are missing at random. *Statistics in medicine*, 30(26):3149–3161.

Long, Q., Zhang, X., and Johnson, B. (2011b). Robust estimation of area under roc curve using auxiliary variables in the presence of missing biomarker values. *Biometrics*, 67(2):559–567.

Manski, C. and McFadden, D. (1981). *Structural analysis of discrete data with econometric applications*, volume 11. MIT press Cambridge, MA.

McClish, D. (1989). Analyzing a portion of the roc curve. *Medical Decision Making*, 9(3):190–195.

McCormick, M. C. (1985). The contribution of low birth weight to infant mortality and childhood morbidity. *The New England Journal of Medicine*, 312(2):82–90.

McNeil, B., Hanley, J., et al. (1984). Statistical approaches to the analysis of receiver operating characteristic (roc) curves. *Medical decision making: an international journal of the Society for Medical Decision Making*, 4(2):137.

Molanes-López, E. and Letón, E. (2011). Inference of the youden index and associated threshold using empirical likelihood for quantiles. *Statistics in Medicine*, 30(19):2467–2480.

Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2):237–249.

Owen, A. (1990). Empirical likelihood ratio confidence regions. *The Annals of Statistics*, 18(1):90–120.

Pencina, M., D'Agostino Sr., R., and Demler, O. (2012a). Novel metrics for evaluating improvement in discrimination: net reclassification and integrated discrimination improvement for normal variables and nested models. *Statistics in medicine*, 31:101–113.

Pencina, M., D'Agostino Sr., R., and Song, L. (2012b). Quantifying discrimination of framingham risk functions with different survival c statistics. *Statistics in Medicine*, 31:1543–1553.

Pepe, M. (1997). A regression modelling framework for receiver operating characteristic curves in medical diagnostic testing. *Biometrika*, 84(3):595–608.

Pepe, M. (2000). An interpretation for the roc curve and inference using glm procedures. *Biometrics*, 56(2):352–359.

Pepe, M. (2004). *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press, USA.

Prentice, R. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*, 73(1):1–11.

Prentice, R. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, 66(3):403–411.

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

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics*, 22(1):300–325.

Qin, J. and Zhang, B. (1997). A goodness-of-fit test for logistic regression models based on case-control data. *Biometrika*, 84(3):609–618.

Qin, J. and Zhang, B. (2003). Using logistic regression procedures for estimating receiver operating characteristic curves. *Biometrika*, 90(3):585–596.

Rizopoulos, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*, 67(3):819–829.

Scott, A. and Wild, C. (1986). Fitting logistic models under case-control or choice based sampling. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 170–182.

Scott, A. and Wild, C. (1991). Fitting logistic regression models in stratified case-control studies. *Biometrics*, 31:497–510.

Skaltsa, K., Jover, L., Fuster, D., and Carrasco, J. (2012). Optimum threshold estimation based on cost function in a multistate diagnostic setting. *Statistics in Medicine*, pages 1098–1109.

Song, R., Zhou, H., and Kosorok, M. (2009). A note on semiparametric efficient inference for two-stage outcome-dependent sampling with a continuous outcome. *Biometrika*, 96(1):221–228.

Swets, J. and Pickett, R. (1982). *Evaluation of diagnostic systems: Methods from signal detection theory*. Academic Press New York.

Swets, J. A. (1979). Roc analysis applied to the evaluation of medical imaging techniques. *Investigative radiology*, 14(2):109–121.

Thompson, M. and Zucchini, W. (1989). On the statistical analysis of roc curves. *Statistics in Medicine*, 8(10):1277–1290.

Tosteson, A. and Begg, C. (1988). A general regression methodology for roc curve estimation. *Medical Decision Making*, 8(3):204–215.

Walter, S. (2005). The partial area under the summary roc curve. *Statistics in medicine*, 24(13):2025–2040.

Wang, X., Ma, J., George, S., and Zhou, H. (2012). Estimation of auc or partial auc under test-result-dependent sampling. *Statistics in biopharmaceutical research*, 4(4):313–323.

Wang, X., Ma, J., and George, S. L. (2013). Roc curve estimation under test-result-dependent sampling. *Biostatistics*, 14(1):160–172.

Wang, X., Wu, Y., and Zhou, H. (2009). Outcome-and auxiliary-dependent subsampling and its statistical inference. *Journal of biopharmaceutical statistics*, 19(6):1132–1150.

Weaver, M. and Zhou, H. (2005). An estimated likelihood method for continuous outcome regression models with outcome-dependent sampling. *Journal of the American Statistical Association*, 100(470):459–469.

White, J. (1982). A two stage design for the study of the relationship between a rare exposure and a rare disease. *American Journal of Epidemiology*, 115(1):119–128.

Yeboah, J., McClelland, R., Polonsky, T., Burke, G., Sibley, C., OLeary, D., Carr, J., Goff, D., Greenland, P., and Herrington, D. (2012). Comparison of novel risk markers for improvement in cardiovascular risk assessment in intermediate-risk individualsrisk markers in cardiovascular risk assessment. *JAMA: The Journal of the American Medical Association*, 308(8):788–795.

Yu, B., Zhou, C., and Bandinelli, S. (2011). Combining multiple continuous tests for the diagnosis of kidney impairment in the absence of a gold standard. *Statistics in medicine*, 30(14):1712–1721.

Zheng, Y., Cai, T., Jin, Y., and Feng, Z. (2012). Evaluating prognostic accuracy of biomarkers under competing risk. *Biometrics*, 68:388–396.

Zheng, Y. and Heagerty, P. (2007). Prospective accuracy for longitudinal markers. *Biometrics*, 63(2):332–341.

Zhou, H., Chen, J., Rissanen, T., Korrick, S., Hu, H., Salonen, J., and Longnecker, M. (2007). Outcome-dependent sampling: An efficient sampling and inference procedure for studies with a continuous outcome. *Epidemiology*, 18(4):461.

Zhou, H., Weaver, M., Qin, J., Longnecker, M., and Wang, M. (2002). A semiparametric empirical likelihood method for data from an outcome-dependent sampling scheme with a continuous outcome. *Biometrics*, 58(2):413–421.