UNIVERSITY OF CALIFORNIA,
IRVINE

On Censoring-Robust Estimation Under the Nested Case-Control Design

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Statistics

by

Michelle M. Nuño

Dissertation Committee:
Daniel L. Gillen, Chair
Associate Professor Joshua D. Grill
Professor Bin Nan
Associate Professor Zhaoxia Yu

2020

DEDICATION

To my parents, Alfonso and Patricia Nuño.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Page

LIST OF ALGORITHMS

ACKNOWLEDGMENTS

I would like to thank my advisor, Daniel L. Gillen, for all of his support throughout the entirety of my time at UCI. I feel honored to have had the opportunity to learn from someone who is so passionate about his work and about helping students to succeed.

I would also like to thank Joshua D. Grill, Bin Nan, and Zhaoxia Yu who were part of my reading committee, and to Scott Bartell and Weining Shen who were part of my advancement committee.

Parts of Chapter 2 were previously published in the *Handbook of Statistics*.

This material is based upon work supported by the Eugene Cota Robles Fellowship at the University of California, Irvine, the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1321846, and the National Institutes of Health under Grant No. R01AG053555. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors(s) and do not necessarily reflect the views of the National Science Foundation or the National Institutes of Health.

CURRICULUM VITAE

Michelle M. Nuño

EDUCATION

Doctor of Philosophy in Statistics	**August 2020**
University of California, Irvine	*Irvine, CA*
Master of Science in Statistics	**June 2017**
University of California, Irvine	*Irvine, CA*
Bachelor of Science in Mathematics	**June 2015**
University of California, Riverside	*Riverside, CA*

RESEARCH EXPERIENCE

Graduate Research Assistant	**2015–2020**
University of California, Irvine	*Irvine, California*

TEACHING EXPERIENCE

Instructor	**Summer 2018**
University of California, Irvine	*Irvine, CA*
Instructor	**Summer 2017**
University of California, Irvine	*Irvine, CA*
Teaching Assistant	**Winter 2017**
University of California, Irvine	*Irvine, CA*
Teaching Assistant	**Fall 2016**
University of California, Irvine	*Irvina, CA*

REFEREED JOURNAL PUBLICATIONS

Robust Estimation in the Nested Case-Control Design Under a Misspecified Covariate Functional Form **In Revision**

On Estimation in the Nested Case-Control Design Under Non-Proportional Hazards **In Revision**

Study Partner Types and Prediction of Cognitive Performance: Implications to Pre-Clinical Alzheimer's Trials **2019**
Alzheimer's Research and Therapy

Which MCI Patients Should Be Included in Prodromal Alzheimer's Disease Clinical Trials? **2019**
Alzheimer Disease & Associated Disorders

More than Skin Deep: Major Histocompatibility Complex (MHC)-Based Attraction Among Asian American Speed-Daters **2018**
Evolution and Human Behavior

Attitudes Toward Clinical Trials Across the Alzheimer's Disease Spectrum **2017**
Alzheimer's Research and Therapy

Alternative Sampling Designs for Time-to-Event Data with Applications to Biomarker Discovery in Alzheimer's Disease **2017**
Handbook of Statistics

On a Randomly Accelerated Particle **2017**
Involve, a Journal in Mathematics

SOFTWARE

R, SAS, C++

ABSTRACT OF THE DISSERTATION

On Censoring-Robust Estimation Under the Nested Case-Control Design

By

Michelle M. Nuño

Doctor of Philosophy in Statistics

University of California, Irvine, 2020

Daniel L. Gillen, Chair

Analysis of time-to-event data using Cox's proportional hazards model is ubiquitous in scientific research. Most commonly, a sample is taken from the population of interest and covariate information is collected on everyone. If the event of interest is rare and it is difficult or not feasible to collect full covariate information for all study participants, the nested case-control design reduces costs with minimal impact on inferential precision. However, no work has been done to investigate the performance of the nested case-control design under model mis-specification. In this dissertation we show that under model mis-specification the quantity being estimated under the nested case-control design will depend not only on the censoring distribution, but also on the number of controls sampled at each event time. This is true in the case of a binary covariate when the proportional hazards assumption is not satisfied, and in the case of a continuous covariate where the functional form is mis-specified. We propose several estimators that allow us to recover the statistic that would have been computed under the full cohort data as well as a censoring-robust estimator. We also investigate the performance of time-dependent receiver operating characteristic curves under the full cohort and nested case-control sampling scenarios. We show that if the risk score model is mis-specified, estimates of the area under the curve will also depend on the censoring distribution and we propose the use of censoring-robust risk scores that allow us to recover censoring-independent area under the curve estimates.

Chapter 1

Introduction

The work in this dissertation is motivated by the need for biomarker discovery in Alzheimer's disease (AD). It is currently estimated that 5.8 million Americans have AD [Association et al., 2020] and this number is expected to grow. AD is a neurodegenerative disease that affects not only the individual with the disease, but also the individual's friends and families due to the impact on memory and daily activities. While AD has detrimental effects on the lives of millions of people, there is still no cure and no way to prevent the disease. Existing biomarkers include the proteins amyloid-β (Aβ), total tau (T-tau), and phosphorylated tau (P-tau) (Blennow et al. [2010], Blennow [2005, 2004]). These biomarkers help identify people who likely have AD or who will progress to AD. A true diagnosis of AD, however, requires post-mortem verification. Moreover, it is difficult to distinguish early AD from other disorders involving similar symptoms (Humpel [2011]). This second point is particularly important now that AD research has shifted to earlier stages of the disease in an attempt to treat AD before irreversible damage has occurred. Discovery of new biomarkers is also important, as these could serve as targets for therapeutic treatments.

Existing biomarkers can be measured in cerebrospinal fluid (CSF) or through brain scans such as positron emission tomography (PET) scans. A tracer specific to the protein of interest is required to measure the biomarker using PET scans. Moreover, changes in protein levels in CSF often reflect changes in the brain, making CSF particularly useful when investigating AD [Humpel, 2011]. Therefore, when investigating potential biomarkers, a common approach is to first measure the biomarker in CSF, if possible. In this case, participants undergo lumbar punctures to provide a CSF sample that is then processed to measure the biomarker of interest. In AD studies, participants are often asked if they would be willing to undergo a lumbar puncture. CSF samples are then collected from participants who agree. These samples are stored and only processed as needed. However, using data collected in a survey at the University of California, Los Angeles (UCLA) Alzheimer's Disease Research Center (ADRC), we found that out of 91 respondents (22 with AD dementia, 32 with MCI, and 37 who were cognitively normal) only approximately 13% of respondents were highly willing (scores of 6 or 7 from a 7-point Likert scale) to participate in a study requiring lumbar punctures, compared to 42% who were highly willing to participate in studies requiring PET scans. It is therefore crucial that stored samples are used efficiently, motivating the need for efficient sampling designs in AD studies.

Two such designs are the nested case-control and the case-cohort designs. When the event of interest is rare, these sampling schemes allow for a large reduction in costs compared to the full cohort scenario where full covariate information on all study participants is collected. Under the nested case-control and case-cohort designs, however, we only obtain full covariate information for all participants who experience an event and a subsample of participants who do not experience an event. In the context of AD research, use of the nested case-control and case-cohort designs would greatly reduce the number of CSF samples that must be processed so that the remaining samples may be used to answer other scientific questions. While these designs provide great utility in scenarios such as the one presented here, it is important that we understand how these designs perform when model mis-specification occurs. This is the focus of the dissertation.

We begin the dissertation with a review of survival analysis including censoring, estimation of the survival function, and the Cox proportional hazards (PH) model. Under model mis-specification, the estimand corresponding to the Cox PH model, or the partial likelihood estimator, depends on the censoring distribution. That is, if the model is mis-specified, our results will depend on dropout and accrual patterns of a study, which is not usually of scientific interest and limits scientific reproducibility and replicability. In Chapter 2, we discuss the implications of model mis-specification in more detail. We also discuss censoring-robust estimators [Xu and O'Quigley, 2000, Boyd et al., 2012, Nguyen and Gillen, 2012] that can be used to remove dependence on the censoring distribution. We continue the chapter by introducing several estimators for use under the nested case-control and case-cohort sampling schemes that are meant to recover the statistic obtained using the partial likelihood estimator. We end with a comparison of the two methods.

The remaining chapters focus on the nested case-control design, as proposed by Thomas [1977], under model mis-specification and represent the methodologic contribution of the dissertation. Because the nested case-control design is often implemented using a stratified Cox PH model, we hypothesized that under model mis-specification, the estimand corresponding to the nested case-control design would also be impacted by the censoring distribution. When using the nested case-control design, one must select the number of controls to be sampled at each event time, which also alters the censoring distribution. Because of this, we suspected that the results would also be impacted by the number of controls sampled at each event time. In Chapter 3, we consider the nested case-control design under violation of the PH assumption. We show that in this case, our results will in fact be impacted by the censoring distribution and the number of sampled controls. Dependence on the number of controls makes it difficult to reproduce results, even within the same study. We therefore propose two estimators. The first estimator recovers the estimand corresponding to the partial likelihood estimator. This allows for reproducibility of results, even if a different number of controls is selected. The second is a censoring-robust estimator whose estimand does not depend on the censoring distribution even when the model is mis-specified. This estimator allows us to replicate results across studies, since they are no longer impacted by patient accrual and dropout patterns. In Chapter 4, we consider mis-specification of the functional form of a continuous covariate, which induces non-proportionality. When this occurs, our estimand again depends on the censoring distribution and, in the nested case-control setting, on the number of sampled controls. We propose an estimator that again recovers the results obtained when using the complete data set. The proposed estimator can be combined with the inverse probability of censoring weights from Chapter 3 to obtain a censoring-robust estimator. Chapters 3 and 4 focus on inference under the nested case-control design. In many disease areas, including AD, there are limited resources and there is much to learn about the disease. Application of censoring-robust estimators along with the nested case-control sampling scheme allow us to obtain censoring-robust estimates while maintaining reduction in costs afforded by the nested case-control sampling scheme.

As previously discussed, the motivation for this work is biomarker discovery in AD. In order to assess the classification performance of a potential biomarker, it is important that we have reliable statistical methods. Chapter 5 considers time-dependent receiver operating characteristic (ROC) curves and the impact of the censoring distribution on estimates of the area under the ROC curve when the model used to obtain a diagnostic risk score is mis-specified. In this case, we learn that the estimates of the area under the curve (AUC) depend on the censoring distribution due to the dependence of the coefficient estimates on the censoring distribution. This is problematic because even though the classification performance of the biomarker does not change, estimates of the AUC will be impacted by accrual and dropout patterns of a study, which are likely to differ across studies. We propose the use of censoring-robust estimators to obtain the risk scores when estimating time-dependent ROC curves.

Chapter 5 focuses on time-dependent ROC curves when full covariate information is available for all study participants. However, this is not always feasible. Therefore, in Chapter 6, we consider time-dependent ROC curves under the nested case-control sampling scheme. We present existing estimators and investigate the performance of the Cai and Zheng [2012] estimator when the model used to estimate the risk score is mis-specified. We again notice dependence on the censoring distribution and, in some cases, on the number of sampled controls. We show that when the censoring-robust estimator of Chapter 3 is used, we are able to recover more stable estimates of the AUC, regardless of the censoring distribution. The use of the censoring-robust estimators provides reliable estimates of the AUC even when the model used to estimate the risk score is mis-specified. The proposed estimators can be used to evaluate the classification performance of potential biomarkers in AD, as well as in other disease areas.

The use of the nested case-control design provides great utility when the event of interest is rare and it is difficult or expensive to collect full covariate information on all study participants. While the research in this dissertation was motivated by biomarker discovery in AD, it is important to note that these statistical methods are also applicable to other disease areas. Use of the proposed estimators allows us to obtain estimates that do not depend on the censoring distribution, while still allowing for the reduction in costs associated with the nested case-control sampling scheme.

Chapter 2

Alternative Sampling Designs for Time-to-Event Data with Applications to Biomarker Discovery in Alzheimer's Disease

2.1 Introduction

Researchers all over the world are working to find the causes of and treatments for diseases such as Alzheimer's disease (AD). However, many of these diseases are rare and hence traditional prospective studies require that scientists follow large groups of patients to observe only a small proportion that will develop the condition. For example, according to the Alzheimer's Association's 2020 report, 1 in 10 people over the age of 65 in the United States has Alzheimer's disease [Association et al., 2020], though the clinical consequences of the disease are severe. While it is believed that early signs of AD are preceded by changes in levels of the biomarkers tau and Aβ, two proteins found in cerebrospinal fluid (CSF) [Association, 2016], in order to check levels of these proteins researchers must extract CSF using lumbar punctures or spinal taps on all patients under study. It is often difficult to convince people to participate in studies that require lumbar punctures since these are often perceived to be painful. In a recent survey study conducted at the UCLA Alzheimer's Disease Research Center (ADRC), we found that only approximately 13% of participants in the (ADRC) were willing to participate in a clinical trial requiring lumbar punctures. Responses were based on a 7-point Likert scale where a score of ”1” represents ”extremely unlikely” to participate and a score of ”7” represents ”extremely likely” to participate. Participants were deemed willing to participate in a study of this nature if they responded with a score of ”6” or ”7” [Nuño et al., 2017].

In cases where a rare disease is being investigated and where exposure measurements are difficult to obtain due to logistical, monetary, and/or ethical reasons, it is appealing to find ways to reduce the number of participants required in the study. One solution is to reduce the number of controls (or people who do not develop the disease) through efficient sampling designs such as the case-cohort [Prentice, 1986] or nested case-control design [Thomas, 1977]. Case-control studies are retrospective designs in which researchers consider subjects who developed the outcome of interest and those who didn't, and then look back to compare exposure between the two groups. The Cox PH model can be used to model time-to-event data in the presence of censoring and allows for the adjustment of confounding variables. The ideas stemming from case-control studies and the Cox PH model lead us to the designs presented in this chapter. The nested case-control and the case-cohort designs are based on the idea that cases provide more information than controls do. In the presence of high censoring, we often have significantly more controls than we do cases. These designs lower the number of controls required for the analysis with relatively little loss in precision by selectively sampling only a proportion of controls. The relatively small loss of efficiency is because cases (subjects who develop the disease) provide greater statistical information relative to controls. In scenarios where there are substantially more controls than there are cases (i.e. in a rare disease setting), this proves beneficial because the total number of subjects studied can be dramatically reduced.

In this chapter, we provide a review of alternative sampling designs when scientific interest lies in the estimation of covariate associations with a (possibly) censored time-to-event endpoint. Our goal is to provide a fairly comprehensive review of the background, development and evaluation of the case-cohort and nested case-control design as well as a brief introduction to counting process notation, which will be used in the remainder of the dissertation. We begin with a brief review of censored data before moving on to the Cox PH model and its performance under model mis-specification. We then introduce time-dependent receiver operating characteristic (ROC) curves. In Section 2.6 we introduce the nested case-control design and some of its variations, which differ in how they sample from risk sets and in the use of selected controls. We also provide simulation results for the nested case-control design. Similarly, we consider the case-cohort design, its variations, as well as simulation results for the different methods. After introducing both designs, we provide an example of their implementation using data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) and compare the results obtained using the various designs. We then present alternative sampling designs whose selection of controls is based on confounding variables. Section 2.10 compares both designs in a scientific and statistical sense to assist researchers in selecting the appropriate method for their study.

2.2 A Brief Review of Survival Analysis

2.2.1 Censoring

In cohort studies it is often of interest to model the time from a fixed or designated origin to the occurrence of some event of interest (eg. death or progression of disease). A commonly encountered problem in the analysis of time-to-event data is censoring. Generally speaking, censoring occurs when the time to the event of interest is not known exactly, but is only known to have occurred within a broader time interval. Here we briefly define the most encountered forms of censoring. For further discussion on the forms of censoring that can occur, we refer the reader to the text of Klein and Moeschberger [2005].

Multiple types of censoring can exist. Perhaps the most common form of censoring that is encountered in practice is *right censoring.* This form of censoring describes the situation where subjects enter into a study at some defined origin and who are at risk for the event of interest at the time of their entry. A subject may then become right censored if their follow-up ends prior to the event of interest having occurred. Thus a subject may become right censored for a variety of reasons including: the study follow-up ending at some pre-specified time point that occurs prior to the subject's true event time (this is commonly termed *administrative right censoring*), the subject choosing to discontinue follow-up prior to the event occurring for some reason (commonly termed *random right censoring*), or another event occurring that precludes the observation of the event of interest (termed a *competing risk*). Most survival methods, including those discussed throughout the remainder of this chapter, make an assumption that the censoring time is independent of the true survival time for all subjects (conditional upon adjusted covariates that may be collected on the subjects). This critical assumption, referred to as the *non-informative censoring* assumption, is considered to be coarsening at random [Heitjan and Rubin, 1991] and is similar to the *missing at random* *assumption* in general missing data problems using the nomenclature from the seminal text of Little and Rubin [2014]. In the case of administrative right censoring, non-informative censoring is generally a reasonable assumption since the end of follow-up is pre-specified. However, for random censoring or in the presence of competing risks, careful consideration for the possibility of informative censoring must be made. In the presence of informative censoring most standard survival analysis methods, including those discussed here, can lead to biased and inconsistent parameter estimates.

Three additional types of censoring can occur in cohort studies. *Left censoring* occurs when we do not know the exact event time for subjects who experienced the event before a certain time; specifically, if the event happens before the start of the study, we may not know precisely when it happened. For example, suppose that patients of ages 65 and older are observed for the development of AD. If they develop the disease before they turn 65 years old, we will only know that it happened before they turned 65.

Interval censoring occurs when we only know that an event occurred within a specific time interval. We again refer to AD studies for an example. Patients at our ADRC are seen every year. Each year, they are given a series of tests to determine their cognitive status. If a patient is observed to have AD dementia, we do not know the exact time at which they developed dementia, only that it occurred sometime between the current visit and the previous year's visit. Such data are said to be interval censored with the intervals of length approximately 1 year.

Finally, *double censoring* occurs when there is a combination of left and right censoring. As an example, double censoring could occur if a three-year study was conducted in which patients were followed for the development of AD dementia. Patients could be enrolled regardless of whether or not they had AD dementia, but if they developed the disease before the start of the study we only know that it happened some time before, and if they become afflicted after the study has ended, researchers only know that the event occurred at some point after the end of follow-up.

2.2.2 Statistical functions of interest in time-to-event data

As stated earlier, censoring is common when we are interested in time-to-event data. Survival analysis methods seek to efficiently estimate and draw inference on statistical functions and functionals in the presence of censoring. As with uncensored data, focus lies on estimation of functionals derived from the probability density function (PDF) and cumulative distribution function (CDF). However, due to the presence of censoring, survival methods also consider estimation of the hazard function and the cumulative hazard function. While these later functions are defined for uncensored data, they are less encountered in those settings. For completeness we define each of the statistical functions most commonly of interest below:

1) <u>Probability Density Function (PDF)</u>:

$$f(t) = \lim_{\Delta t \to 0^+} \frac{1}{\Delta t} \Pr[t \leq T < t + \Delta t]$$

2) <u>Cumulative Density Function (CDF)</u>:

$$F(t) = \Pr[T \leq t] = \int_0^t f(s)ds$$

3) <u>Survival Function</u>:

$$S(t) = \Pr[T > t] = 1 - \Pr[T \leq t] = 1 - F(t) = 1 - \int_0^t f(s)ds$$

4) <u>Hazard Function</u>:

$$\lambda(t) = \lim_{\Delta t \to 0^+} \frac{1}{\Delta t} \Pr[t \leq T < t + \Delta t | T \geq t] = \frac{f(t)}{S(t)}$$

5) <u>Cumulative Hazard function</u>:

$$\Lambda(t) = \int_0^t \lambda(s) ds = -\log(S(t))$$

What is readily apparent from the above definitions and relationships is that knowing any one of the four functions allows us to recover the rest. In the presence of censoring, estimation of the hazard function is often most approachable given the conditional nature of the function. That is, we may condition upon survival up to a given time so that estimates of the hazard can be formed by utilizing data from subjects that have not been censored or known to have experienced the event prior to that time.

2.2.3 Parametric estimation of the survival distribution

One way to estimate the functions presented in the previous section is to assume a parametric distribution. Examples of commonly assumed parametric survival distributions can be found in Table 2.1.

Table 2.1: Parametric survival distributions and properties of the corresponding hazard function.

Distribution	Properties of the Hazard Function
Exponential	Constant hazard
Gamma	Monotonic hazard
Weibull	Hazard proportional to a power of time
Log-normal	Unimodal, right skewed hazard
Log-logistic	Unimodal, right skewed hazard (heavier tail than Log-normal)
Gompertz	Hazard increases exponentially with time
Generalized Gamma	Flexible (3 parameter), allowing for non-monotonicity

When we assume a parametric survival distribution and when the regularity conditions hold, we can use maximum likelihood theory to estimate the survival distribution as long as we account for censoring. Consider a sample of n subjects and suppose that $T_i \sim F_T(\cdot), C_i \sim G_C(\cdot)$ where T_i represents the event times and C_i represents the censoring times, $i = 1, \ldots, n$. Assume that the two distributions are known, and that the event and censoring times are independent. Let X_i be the observed failure time for subject i, so $X_i = \min(T_i, C_i)$ and

$$\delta_i = \begin{cases} 1 & \text{if the event was observed for subject } i \\ 0 & \text{if subject } i \text{ was censored} \end{cases}.$$

Under this censored data framework, the likelihood contribution for the i^{th} subject is

$$L_i(\lambda, x_i, \delta_i) = \begin{cases} f_T(x_i)[1 - G_c(x_i)] & \text{if } \delta_i = 1 \\ g_c(x_i)[1 - F_T(x_i)] & \text{if } \delta_i = 0 \end{cases}$$

which can be written in the following manner:

$$\begin{aligned} L_i(\lambda, x_i, \delta_i) &= \delta_i f_T(x_i)[1 - G_c(x_i)] + (1 - \delta_i)g_c(x_i)[1 - F_T(x_i)] \\ &= \{f_T(x_i)[1 - G_c(x_i)]\}^{\delta_i}\{g_c(x_i)[1 - F_T(x_i)]\}^{1-\delta_i}. \end{aligned}$$

Suppose, for example, that $T_i \sim Exp(\lambda)$. The i^{th} subject's contribution to the likelihood is:

$$L_i(\lambda, x_i, \delta_i) = [\lambda e^{-\lambda x_i}[1 - G_c(x_i)]]^{\delta_i}[g_c(x_i)e^{-\lambda x_i}]^{1-\delta_i} = \lambda^{\delta_i} e^{-\lambda x_i}[1 - G_c(x_i)]^{\delta_i}[g_c(x_i)]^{1-\delta_i}.$$

After taking the logarithm of the likelihood function and simplifying, we find that the contribution to the log-likelihood becomes $l_i(\lambda) = \delta_i \log(\lambda) - \lambda x_i + k$, where k is constant with respect to λ and the contributions to the score and the information are

$$U_i(\lambda) = \frac{\partial l_i(\lambda)}{\partial \lambda} = \frac{\delta_i}{\lambda} - x_i \;\text{ and }\; I_i(\lambda) = -\mathrm{E}\Big[\frac{\partial U_i(\lambda)}{\partial \lambda}\Big] = -\mathrm{E}\Big[-\frac{\delta_i}{\lambda^2}\Big],$$

respectively where $\mathrm{E}[\cdot]$ represents the expectation.

From this, we can use the score function to find the maximum likelihood estimator (MLE) for λ. Specifically, the MLE for λ can be obtained by solving the score equation given by

$$U(\hat{\lambda}) = \sum_{i=1}^{n} U_i(\hat{\lambda}) = 0.$$

Replacing $U_i(\hat{\lambda})$ with $\delta_i/\hat{\lambda} - x_i$, we obtain

$$\sum_{i=1}^{n}\Big(\frac{\delta_i}{\hat{\lambda}} - x_i\Big) = \frac{\sum_{i=1}^{n}\delta_i}{\hat{\lambda}} - \sum_{i=1}^{n} x_i = 0,$$

and solving for $\hat{\lambda}$ we obtain the MLE for λ as $\hat{\lambda} = (\sum_{i=1}^{n}\delta_i)\,/\,(\sum_{i=1}^{n}x_i) = \bar{\delta}/\bar{x}$.

Finally, appealing to the usual asymptotic theory for MLEs under standard regularity conditions we have that $\widehat{\lambda}\dot{\sim}\mathcal{N}\left(\lambda, I^{-1}(\lambda)\right)$ where $I(\lambda) = \sum_{i=1}^{n} I_i(\lambda)$. Note that in practice the calculation of $I(\lambda)$ requires the evaluation of the expectation of δ_i, $i = 1, \ldots, n$. Because this would require a further assumption regarding the unknown distribution of C_i, in practice $I_i(\lambda)$ is most commonly replaced with Fisher's observed information matrix and estimated by $I_i^*(\widehat{\lambda}) = \delta_i/\widehat{\lambda}^2$.

2.2.4 Non-parametric estimation of the survival distribution

As with any statistical estimation procedure, mis-specification of the parametric distribution may lead to biased and inconsistent estimates of survival. For this reason it may be desirable to estimate the survival function non-parametrically. To do this, we can rely upon the non-parametric Kaplan-Meier estimator [Kaplan and Meier, 1958], which is uniformly consistent for the survival function under the assumption of independent (or uninformative) censoring [Wang, 1987]. Below we provide a heuristic derivation of the Kaplan-Meier estimator to appeal to the reader's intuition.

First, assume that we observe information on the interval $(0, \tau_{\max}]$. To estimate the survival, we may split our interval into K intervals as such: $(0, \tau_1], (\tau_1, \tau_2], ..., (\tau_{K-1}, \tau_{\max}]$, and recall that $S(t) = \Pr[T > t]$, which in our case can be written as

$$S(t) = \Pr[T > t] = \Pr[T > t|T > \tau_{\max}] \cdot \Pr[T > \tau_{max}|T > \tau_{K-1}] \cdot ... \cdot \Pr[T > \tau_2|T > \tau_1] \cdot \Pr[T > \tau_1].$$

For illustrative purposes, consider a finite number of intervals as in Figure 2.1 where we observe three events and we let $K = 8$. Notice that we do not observe events at every interval.

In general, the conditional probability for survival to τ_j conditional upon survival up to τ_{j-1}

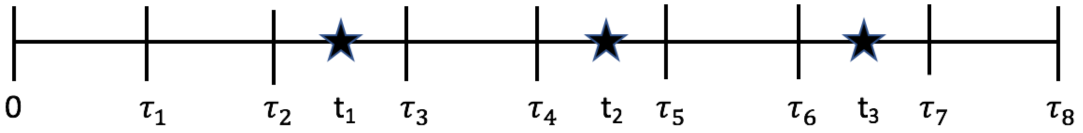


Figure 2.1: This figure presents an example in which we have split our interval into 8 smaller intervals with three events. The events are marked with a star and their corresponding event times are labeled.

may be estimated as

$$\widehat{\Pr}[T > \tau_j | T > \tau_{j-1}] = 1 - \frac{d_j}{n_j} = \frac{s_j}{n_j}$$

where d_j is the total number of events at time τ_j, s_j is the number of subjects who have not failed by time τ_j, and n_j is the total number of subjects at risk at time τ_j.

When estimating the conditional probability for a time at which no event occurred, $d_j = 0$, we find ourselves with $\widehat{\Pr}[T > \tau_j | T > \tau_{j-1}] = 1 - 0/n_j = 1$. Notice that this would be true for all intervals not including times t_1, t_2, or t_3.

Therefore in our current example,

$$\hat{S}(t) = 1 \cdot 1 \cdot \frac{s_1}{n_1} \cdot 1 \cdot \frac{s_2}{n_2} \cdot 1 \cdot \frac{s_3}{n_3} \cdot 1 = \frac{s_1}{n_1} \cdot \frac{s_2}{n_2} \cdot \frac{s_3}{n_3}.$$

Because the conditional probabilities are approximately 1 when no event has occurred, including the censoring times will not change the estimate for the survival function. This allows us to only focus on the intervals during which an event was observed. We can generalize this by letting the number of intervals grow to infinity, so that the width of each interval shrinks to zero. Then assuming that D events are observed, we obtain

$$\hat{S}(t) = 1 \cdot 1 \cdot \frac{s_1}{n_1} \cdot 1 \cdot \frac{s_2}{n_2} \cdot 1 \ldots \frac{s_D}{n_D}.$$

This leads us to the Kaplan-Meier estimator for $S(t)$:

$$\widehat{S}_{KM}(t) = \prod_{j:t_j \leq t} \Big(1 - \frac{d_j}{n_j}\Big) = \prod_{j:t_j \leq t} \Big(\frac{s_j}{n_j}\Big). \tag{2.1}$$

It should be noted that the Kaplan-Meier estimator can also be written in counting process notation as

$$\hat{S}_{KM}(t) = \prod_{s \leq t} \Big\{1 - d\hat{\Lambda}(s)\Big\}$$

where $\hat{\Lambda}(s) = \int_0^s \frac{\sum_{i=1}^n dN_i(u)}{\bar{Y}(u)}$ is the Nelson-Aalen estimator [Nelson, 1972, Aalen, 1978], $dN_i(t) = N_i(t^- + dt) - N_i(t^-)$, $N_i(t) = I(X_i \leq t, \delta_i = 1)$, $\bar{Y}(t) = \sum_{i=1}^n Y_i(t)$, and $Y_i(t) = I(X_i > t)$. Using Rebolledo's Theorem, it can be shown that the Kaplan-Meier estimator is asymptotically normally distributed [Fleming and Harrington, 2011].

Kaplan-Meier estimator as the non-parametric maximum likelihood estimator for the survival

In this section we show that the Kaplan-Meier estimator is the non-parametric maximum likelihood estimator for the survival function. We consider event times $t_1 < t_2 < \cdots < t_D$. Under the assumption of independent censoring and D events, the likelihood on the space of all survivor functions takes the following form:

$$L = \prod_{j=1}^{D} \Big\{[S(t_j^-) - S(t_j)]^{d_j} \prod_{l=1}^{m_j} S(t_{jl})\Big\}$$

where d_j is the number of events that occur at time t_j, and m_j is the number of observations that are censored in the interval $[t_j, t_{j+1})$ with censoring times $t_{j1}, \ldots, t_{jm_j}$. We denote the number of subjects at risk right before time t_j by n_j.

To find the non-parametric MLE, we must find the survival function that maximizes the likelihood. Note that the event times $t_1 < \cdots < t_D$ are discrete. Moreover, because $t_{jl} \geq t_j$, we have that $S(t_{jl})$ is maximized at $S(t_{jl}) = S(t_j)$ for $j = 1, \cdots, D$ and $l = 1, \cdots m_j$. We denote the MLE by $\widehat{S}$ with hazards $\widehat{\lambda}_1, \widehat{\lambda}_2, \ldots, \widehat{\lambda}_D$ corresponding to each event time.

We can therefore write $\widehat{S}(t_j) = \prod_{l=1}^{j}(1 - \widehat{\lambda}_l)$ and $\widehat{S}(t_j^-) = \prod_{l=1}^{j-1}(1 - \widehat{\lambda}_l)$ such that $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_D$ maximize

$$\begin{aligned} L(\vec{\lambda}) &= \prod_{j=1}^{D} \left\{ [S(t_j^-) - S(t_j)]^{d_j} \prod_{l=1}^{m_j} S(t_{jl}) \right\} \\ &= \prod_{j=1}^{D} \left[\lambda_j^{d_j} \prod_{l=1}^{j-1} (1 - \lambda_l)^{d_j} \prod_{l=1}^{j} (1 - \lambda_l)^{m_j} \right] \\ &= \prod_{j=1}^{D} \lambda_j^{d_j} (1 - \lambda_j)^{n_j - d_j}. \end{aligned}$$

We can then consider the log-likelihood,

$$\ell(\vec{\lambda}) = \sum_{j=1}^{D} d_j \log(\lambda_j) + (n_j - d_j) \log(1 - \lambda_j)$$

which yields the score function with j-th element given by

$$U_j(\vec{\lambda}) = \frac{\partial \ell(\vec{\lambda})}{\partial \lambda_j} = \frac{d_j}{\lambda_j} - \frac{n_j - d_j}{1 - \lambda_j}.$$

Setting the score function equal to zero we obtain $\hat{\lambda}_j = d_j / n_j$. Therefore, the non-parametric maximum likelihood estimator is

$$\widehat{S}(t) = \prod_{j:t_j \leq t} (1 - \hat{\lambda}_j) = \prod_{j:t_j \leq t} \left(1 - \frac{d_j}{n_j}\right),$$

which is in fact the Kaplan-Meier estimator.

Estimating the standard error of $\widehat{S}_{KM}(t)$

In order to quantify the uncertainty in $\widehat{S}_{KM}(t)$, it is necessary to derive a variance estimator. However, this is most easily done by first considering the variance of $\log \widehat{S}_{KM}(t)$. To this end, again define $\hat{\lambda}_j \equiv \frac{d_j}{n_j},\ j = 1, \ldots, D$ so that the Kaplan-Meier estimator from (2.1) is given by

$$\widehat{S}_{KM}(t) = \prod_{j:t_j \leq t} \left(1 - \hat{\lambda}_j\right),$$

and the log of the Kaplan-Meier estimator is given by

$$\log \widehat{S}_{KM}(t) = \sum_{j:t_j \leq t} \log\left(1 - \hat{\lambda}_j\right).$$

Let $\mathscr{F}(t_j)$ denote the filtration (or history) of all deaths and censoring events up to time $t_j,\ j = 1, \ldots, D$. Then conditional on $\mathscr{F}(t_j)$, the number of failures occurring in interval $[t_j, t_{j+1})$ is Binomial(λ_j, n_j) where λ_j is the probability of failure in interval $[t_j, t_{j+1})$. As such, the expectation of $\hat{\lambda}_j$ is given by

$$\mathrm{E}[\hat{\lambda}_j] = \mathrm{E}[\mathrm{E}(\hat{\lambda}_j | \mathscr{F}(t_j))] = \mathrm{E}[\lambda_j] = \lambda_j.$$

Similarly, the variance of $\hat{\lambda}_j$ is given by

$$\mathrm{Var}[\hat{\lambda}_j] = \mathrm{E}[\mathrm{Var}(\hat{\lambda}_j | \mathscr{F}(t_j))] + \mathrm{Var}[\mathrm{E}(\hat{\lambda}_j | \mathscr{F}(t_j))] = \mathrm{E}\Big[\frac{\lambda_j(1-\lambda_j)}{n_j}\Big] + \mathrm{Var}(\lambda_j) = \frac{\lambda_j(1-\lambda_i)}{n_j}.$$

We can now use the delta method to approximate the variance of $\log(1 - \hat{\lambda}_j)$ as

$$\mathrm{Var}[\log(1 - \hat{\lambda}_j)] \doteq \frac{1}{(1-\lambda_j)^2} \mathrm{Var}[\hat{\lambda}_j],$$

and hence an estimate of the variance of $\log(1 - \hat{\lambda}_j)$ is given by plugging in $\hat{\lambda}_j$ for λ_j to obtain

$$\begin{aligned}
\widehat{\mathrm{Var}}[\log(1-\hat{\lambda}_j)] &= \frac{1}{(1-\hat{\lambda}_j)^2}\cdot\widehat{\mathrm{Var}}[\hat{\lambda}_j] \\
&= \frac{1}{(1-\hat{\lambda}_j)^2}\frac{\hat{\lambda}_j(1-\hat{\lambda}_j)}{n_j} \\
&= \frac{d_j}{n_j(n_j-d_j)}.
\end{aligned}$$

To this point we have only considered the variance of the j^{th} subject's contribution to $\log \widehat{S}_{KM}(t)$. For the covariance between terms, without loss of generality, assume that $k<j$ so that

$$\mathrm{E}[\hat{\lambda}_k\hat{\lambda}_j] = \mathrm{E}[\mathrm{E}\{\hat{\lambda}_k\hat{\lambda}_j|\mathscr{F}(t_j)\}] = \mathrm{E}[\hat{\lambda}_k\mathrm{E}\{\hat{\lambda}_j|\mathscr{F}(t_j)\}] = \mathrm{E}[\hat{\lambda}_k]\lambda_j = \lambda_k\lambda_j,$$

and hence

$$\mathrm{Cov}[\hat{\lambda}_k,\hat{\lambda}_j] = \mathrm{E}[\hat{\lambda}_k\hat{\lambda}_j] - \mathrm{E}[\hat{\lambda}_k]\mathrm{E}[\hat{\lambda}_j] = \lambda_k\lambda_j - \lambda_k\lambda_j = 0.$$

From the above, the covariance between all distinct terms in $\log \widehat{S}_{KM}(t)$ is 0, yielding an estimate of $\mathrm{Var}[\log \hat{S}_{KM}(t)]$ given by

$$\widehat{\mathrm{Var}}[\log \hat{S}_{KM}(t)] = \sum_{j:t_j\le t}\widehat{\mathrm{Var}}[\log(1-\hat{\lambda}_j)] = \sum_{j:t_j\le t}\frac{d_j}{n_i(n_j-d_j)}. \tag{2.2}$$

Again applying the delta method we obtain Greenwood's formula [Greenwood et al., 1926] for the variance of the Kaplan-Meier estimator, given by

$$\widehat{\mathrm{Var}}_G[\hat{S}_{KM}(t)] = \hat{S}^2_{KM}(t)\sum_{i:t_i\le t}\frac{d_i}{n_i(n_i-d_i)}. \tag{2.3}$$

Confidence intervals for $S(t)$

It can be shown that $\hat{S}_{KM}(t) \dot{\sim} \mathcal{N}(S(t), \text{Var}[\hat{S}_{KM}(t)])$ [Fleming and Harrington, 2011]. As such, a $100 \times (1-\alpha)\%$ Wald-based confidence interval for $S(t)$ can be constructed as

$$(\hat{S}_{KM}(t) - z_{1-\frac{\alpha}{2}} \cdot \widehat{\text{SE}}_G, \hat{S}_{KM}(t) + z_{1-\frac{\alpha}{2}} \cdot \widehat{\text{SE}}_G) \tag{2.4}$$

where $z_{1-\frac{\alpha}{2}}$ denotes the $(1-\frac{\alpha}{2})$-quantile of the standard normal distribution and $\widehat{\text{SE}}_G$ is the square-root of Greenwood's variance estimator given in (2.3). However, because the support of the normal distribution is all real numbers, the interval in (2.4) can yield values outside the range $[0, 1]$. In order to fix this problem, a better approach is to first construct a confidence interval for $\log \Lambda(t)$, which has support over all real numbers, then back-transform to a confidence interval for $S(t)$. Recall from Section 2.1 that $\Lambda(t) = -\log S(t)$, so we can estimate the cumulative hazard function using $\hat{\Lambda}(t) = -\log \hat{S}_{KM}(t)$. As before we can appeal to the delta method to approximate the variance of $\hat{\Lambda}(t)$ as

$$\text{Var}[\log(-\log(\hat{S}_{KM}(t))] \doteq \frac{\text{Var}[\log(\hat{S}_{KM}(t))]}{\log^2(\hat{S}_{KM}(t))}.$$

Using the results from Section 2.2.1, we can estimate the numerator in the above expression by considering the variance estimator given in (2.2), yielding

$$\widehat{\text{Var}}[\log(-\log(\hat{S}_{KM}(t))] = \frac{\sum_{i:t_i \leq t} \frac{d_i}{n_i s_i}}{\log^2(\hat{S}_{KM}(t))}.$$

Then relying on the approximate normality of $\log \hat{\Lambda}(t)$, a $100 \times (1-\alpha)\%$ confidence interval for $\log \Lambda(t) = \log[-\log(S(t))]$ is given by

$$\left(\log\left[-\log(\hat{S}_{KM}(t))\right] - z_{1-\alpha/2} \times \widehat{\text{SE}}_{G^*}, \log\left[-\log(\hat{S}_{KM}(t))\right] + z_{1-\alpha/2} \times \widehat{\text{SE}}_{G^*}\right) \tag{2.5}$$

where $\widehat{\mathrm{SE}}_{G^*} = \sqrt{\sum_{i:t_i \leq t} \frac{d_i}{n_i s_i}} / -\log(\hat{S}_{KM}(t))$.

Finally, transforming the confidence interval limits in (2.5) to formulate a 100 × $(1-\alpha)\%$ confidence interval for $S(t)$ we obtain

$$\exp\left\{-e^{\log(\hat{\Lambda}(t)) \pm z_{1-\frac{\alpha}{2}} \times \widehat{\mathrm{SE}}_{G^*}}\right\} = \left([\hat{S}_{KM}(t)]^{\exp\left\{z_{1-\alpha/2} \times \widehat{\mathrm{SE}}_{G^*}\right\}}, [\hat{S}_{KM}(t)]^{\exp\left\{-z_{1-\alpha/2} \times \widehat{\mathrm{SE}}_{G^*}\right\}}\right). \tag{2.6}$$

Note that the confidence interval provided in (2.6) is range respecting in that both endpoints lie in the interval (0,1).

2.3 Cox Proportional Hazards Model

As discussed in Section 2.2.4, the Kaplan-Meier estimator is helpful for estimating the survival distribution. In observational studies, however, we need to include adjustment variables to account for potential differences between the subjects being observed. One way to incorporate adjustment variables when analyzing survival data is via use of the Cox PH model.

The Cox PH model focuses on the hazard function, parameterizing the hazard at time t as a function of known covariates using a multiplicative model of the form:

$$\lambda(t|z_1, ..., z_p) = \lambda_0(t) e^{\beta_1 z_1 + ... + \beta_p z_p}, \tag{2.7}$$

where the baseline hazard, $\lambda_0(t)$, is the hazard rate when all covariate values are 0. The model specification in (2.7) allows for covariates to have a multiplicative effect on the hazard function. Specifically, e^{β_k} represents the relative difference in the hazard function comparing two subpopulations that differ by one unit in z_k $(k = 1, ..., p)$, assuming all other covariates are constant. Notably, because β_k does not depend upon time, the model assumes that the effect of covariate z_k on the hazard function is constant over time.

Estimation of the parameters in the Cox PH model requires the following information:

$$x_i = \text{observed follow-up time}$$

$$\delta_i = \begin{cases} 0 & \text{if censored,} \\ 1 & \text{if an event was observed} \end{cases}$$

$$z_i = \text{covariate values.}$$

Notice that if the event is censored, x_i is the censoring time and if an event is observed, x_i is the event time.

As with the Kaplan-Meier estimator, the Cox PH model focuses on the observed event times; only utilizing information on censored subjects to compare with those of observed cases. More specifically, at each event time the covariate values of the subject who failed are compared to those of all other subjects who are still in the *risk set*, or set of subjects that have not experienced an event or been censored prior to the time of the event.

The Cox PH model avoids fully parametric assumptions about the form of (2.7) by obtaining estimates of the model parameters via maximization of the *partial likelihood*. The partial likelihood can be constructed by considering the conditional probability that a subject with specific covariate value experiences the event of interest at a particular time t, given that some event was observed at time t. More specifically, for subject j who was observed for an event at time t_j, contribution to the partial likelihood can be written as:

$$\begin{aligned}
L_j &= \Pr\{\text{subject with covariate vector } z_j \text{ fails at } t_j| \text{ some subject failed at } t_j\} \\
&= \frac{\Pr\{\text{ subject with covariate vector } z_j \text{ fails at } t_j\}}{\Pr\{\text{ some subject in } R(t_j) \text{ failed at } t_j\}} \\
&= \frac{\lambda_j(t_j)(\Delta t)}{\sum_{l \in R(t_j)} \lambda_l(t_j)(\Delta t)} \\
&= \frac{\lambda_j(t_j)}{\sum_{l \in R(t_j)} \lambda_l(t_j)},
\end{aligned}$$

where λ_j denotes the hazard function for subject j as defined by the covariate values and $R(t_j)$ represents the subjects at risk at time t_j. The last line follows because Δt does not depend on the event time.

Plugging in the model specification given in (2.7), we have that the contribution to the partial likelihood of the subj ect corresponding to the j^{th} event time is

$$L_j = \frac{\lambda_0(t_j) e^{z_j^T \beta}}{\sum_{l \in R(t_j)} \lambda_0(t_j) e^{z_l^T \beta}} = \frac{e^{z_j^T \beta}}{\sum_{l \in R(t_j)} e^{z_l^T \beta}}.$$

Therefore, the full partial likelihood incorporating all D (independent) event times is:

$$L = \prod_{j=1}^{D} L_j = \prod_{j=1}^{D} \frac{e^{z_j^T \beta}}{\sum_{l \in R(t_j)} e^{z_l^T \beta}} \tag{2.8}$$

and hence the log-partial likelihood is given by

$$\log(L) = \sum_{j=1}^{D} \log\left(\frac{e^{z_j^T \beta}}{\sum_{l \in R(t_j)} e^{z_l^T \beta}}\right) = \sum_{j=1}^{D} \left\{ z_j^T \beta - \log\left(\sum_{l \in R(t_j)} e^{z_l^T \beta}\right) \right\}. \tag{2.9}$$

Notice that the baseline hazard does not appear in the partial likelihood or in the log-partial likelihood. Because of this, the model is referred to as *semi-parametric*; the baseline hazard is non-parametric and potentially infinitely dimensional, but the risk ratio is parametric and fully specified by a finite set of parameters. The partial likelihood can also be constructed as a profile likelihood, where the baseline hazard is profiled out of the fully specified likelihood function [Breslow, 1972].

To estimate the coefficients, we maximize the partial likelihood using the partial likelihood score equation

$$U(\beta) = \frac{\partial \log(L)}{\partial \beta} = \sum_{j=1}^{D} \Big\{ z_j - \sum_{l \in R(t_j)} \frac{z_l e^{z_l^T \beta}}{\sum_{i \in R(t_j)} e^{z_i^T \beta}} \Big\}. \tag{2.10}$$

Note that the partial likelihood score function can be written in counting process notation as

$$U(\beta) = \sum_{i=1}^{n} \int_{t=0}^{\infty} \left\{ Z_i - \frac{n^{-1} \sum_{j=1}^{n} Z_j Y_j(t) \exp(Z_j \beta)}{n^{-1} \sum_{j=1}^{n} Y_j(t) \exp(Z_j \beta)} \right\} dN_i(t) = 0$$

where $Y_j(t) = I(X_j \geq t)$ and $N_i(t) = I(X_i \leq t, \delta_i = 1)$. In counting process notation, we denote the covariate values as random. We will use this notation in the following section and in the remaining chapters.

Using the Newton-Raphson method, we can solve for $\hat{\beta}$ such that $U(\hat{\beta}) = 0$, the maximum partial likelihood estimates of β. Moreover, it can be shown using Rebolledo's Theorem that $\hat{\beta} \dot{\sim} N_p(\vec{\beta_0}, \mathcal{I}^{-1}(\beta_0))$, where β_0 denotes the true value of β and $\mathcal{I}(\beta_0) = -\mathrm{E}\Big[\frac{\partial U(\beta)}{\partial \beta} \Big|_{\beta = \beta_0} \Big]$, the partial information. In practice, it is common to replace $\mathcal{I}(\beta_0)$ with the observed information given by $I(\beta_0) = \frac{\partial U(\beta)}{\partial \beta} \Big|_{\beta = \beta_0}$.

2.4 Cox Proportional Hazards Model Under Model Misspecification

When one's primary research goal is to estimate and draw inference regarding the association between a predictor of interest and response, it is important to specify the statistical model *a priori* to avoid multiple testing bias (cf. Ioannidis [2005], Gelman and Loken [2013], de Groot [2014], Motulsky [2015]). Because it is unlikely that all assumptions pertaining to an *a priori* specified model will hold when actually applied to observed data, it is critical to consider the statistical properties of estimators when underlying model assumptions are violated. When analyzing time-to-event data, use of Cox's PH model [Cox, 1972] is ubiquitous in the literature despite the fact that non-proportional hazards (NPH) effects commonly arise in clinical research. In the context of our motivating AD research, if a time-varying biomarker is only sampled at baseline, the association between the biomarker and the outcome of interest may be higher in magnitude at times local to the measurement due to within-subject changes in the biomarker that arise over time. While this would yield a NPH biomarker effect, *a priori* specification of exactly how the hazard ratio is likely to change over time is difficult, leading many researchers to simply default to Cox's PH model.

As before, let T_i, C_i and Z_i denote the true event time, censoring time, and covariate value for subject i, $i = 1, \ldots, n$, respectively. Further, denote the observed time for subject i as $X_i = \min(T_i, C_i)$. Under a PH structure, the hazard function can be written as $\lambda(t|Z) = \lambda_0(t)\exp(Z\beta)$. In results presented throughout, Z may depend on time but for ease of exposition, we omit such indexing. Using counting process notation, Cox's partial likelihood estimator can be written as the solution to

$$U(\beta) = \sum_{i=1}^{n} U_i(\beta) = \sum_{i=1}^{n} \int_{t=0}^{\infty} \left\{ Z_i - \frac{n^{-1}\sum_{j=1}^{n} Z_j Y_j(t)\exp(Z_j\beta)}{n^{-1}\sum_{j=1}^{n} Y_j(t)\exp(Z_j\beta)} \right\} dN_i(t) = 0 \qquad (2.11)$$

where $Y_j(t) = I(X_j \geq t)$ and $N_i(t) = I(X_i \leq t, \delta_i = 1)$.

Under model mis-specification and in the absence of censoring, Cox's partial likelihood estimator can be interpreted as an average covariate effect over the observed support of survival times [Xu and O'Quigley, 2000]. However, in the presence of censoring and under the standard assumption of independence between the survival and censoring time conditional upon covariates, $T \perp\!\!\!\perp C|Z$, and covariate-independent censoring, $C \perp\!\!\!\perp Z$, it has been shown [Struthers and Kalbfleisch, 1986] that if the PH assumption is invalid, the estimand consistently estimated by the solution to (2.11) depends on the censoring distribution. The practical implication of this result is that heterogeneity in accrual or drop-out patterns across studies will yield different estimates of association even though the true relative difference in hazards associated with a covariate of interest may be homogeneous. This makes replication of results difficult due to the high heterogeneity of patient accrual and retention patterns typically observed across studies.

As shown in Boyd et al. [2012], when the model is mis-specified, the estimand corresponding to the partial likelihood estimator is consistent for the solution to

$$\int_0^\infty E_Z\left(f_T(t|Z) S_C(t|Z) \left[Z - \frac{E_Z\{Z S_T(t|Z) S_C(t|Z) \exp(Z\beta)\}}{E_Z\{S_T(t|Z) S_C(t|Z) \exp(Z\beta)\}} \right] \right) dt = 0. \tag{2.12}$$

Notice the dependence on $S_C(t|Z)$, the survival function for the censoring times. When considering independent censoring (i.e. the censoring distribution does not depend on Z), (2.12) simplifies to

$$\int_0^\infty E_Z\left(f_T(t|Z) S_C(t) \left[Z - \frac{E_Z\{Z S_T(t|Z) \exp(Z\beta)\}}{E_Z\{S_T(t|Z) \exp(Z\beta)\}} \right] \right) dt = 0,$$

getting rid of the dependence on the censoring distribution in the compensator term [Boyd et al., 2012]. To obtain an estimator that is robust to the censoring distribution in this setting, Xu and O'Quigley [2000] propose reweighting the partial likelihood score function as

$$U_{XO}(\beta) = \sum_{i=1}^{n} U_{i,XO}(\beta) = \sum_{i=1}^{n} \int_{t=0}^{\infty} W_{i,XO}(t) \left\{ Z_i - \frac{n^{-1} \sum_{j=1}^{n} Z_j Y_j(t) \exp(Z_j \beta)}{n^{-1} \sum_{j=1}^{n} Y_j(t) \exp(Z_j \beta)} \right\} dN_i(t) \quad (2.13)$$

where $W_{i,XO}(t) = \hat{S}_{KM}(t)/\sum_{j=1}^{n} Y_j(t)$ and $S_{KM}(t)$ is the left-continuous Kaplan-Meier estimator of the survival function [Kaplan and Meier, 1958]. The estimand of the proposed estimator no longer depends on the censoring distribution and can be interpreted as an average covariate effect.

In some cases, the censoring times depend on the values of the covariates. For example, subjects receiving placebo in an open-label study may be more likely to drop out in search of an active treatment. To address this, Boyd et al. [2012] propose reweighting the partial likelihood by the inverse of the covariate-dependent censoring distribution in the context of randomized clinical trials. The proposed estimating equation is:

$$U_{CR}(\beta) = \sum_{i=1}^{n} \int_{t=0}^{\infty} W(t|Z = z_i) \left\{ Z_i - \frac{S_{CR}^{(1)}(\beta, t)}{S_{CR}^{(0)}(\beta, t)} \right\} dN_i(t) = 0 \quad (2.14)$$

where $S_{CR}^{(r)}(\beta, t) = n^{-1} \sum_{j=1}^{n} Z_j^r W(t|Z = z_j) Y_j(t) \exp(Z_j \beta)$, $W(t|Z = z_j) = \{\hat{S}_{C,KM}(t|Z = z_j)\}^{-1}$, and $\hat{S}_{C,KM}(t|Z = z_j)$ is the covariate-dependent left continuous Kaplan-Meier estimator. Reweighting the score function in this manner yields a censoring-robust estimator that is asymptotically equivalent to that proposed by Xu and O'Quigley [2000] when $S_C(t|Z) = S_C(t)$. This work was later extended to observational studies by Nguyen and Gillen [2012], who proposed a survival tree-based estimator. The three estimators allow us to estimate an average covariate effect when the model is mis-specified, and still yield valid results when the model is correctly specified.

2.5 Time-Dependent Receiver Operating Characteristic Curves

Oftentimes, the goal is to investigate the classification performance of a marker, or model. Receiver operating characteristic (ROC) curves are often used to evaluate the classification performance of continuous measures by considering the sensitivity and specificity of a biomarker for a wide range of thresholds. *Sensitivity* is the probability that an individual is classified as testing positive (or meeting a specific threshold) given that they have the disease, while *specificity* is the probability that an individual is classified as testing negative (or not meeting the threshold value) given that the individual does not have the disease. ROC curves are generated by plotting sensitivity vs. (1 - specificity). A common summary measure of biomarker performance is the area under the ROC curve (AUC), which provides an estimate of the probability that a randomly selected individual with the disease will be rated higher than one without the disease [Fawcett, 2006, Heagerty et al., 2000]. The higher the AUC, the better the biomarker is for classifying diseased and non-diseased individuals. One benefit of ROC curves is that the sensitivity and specificity are estimated over all possible cut points, so the results do not depend on a single cut point value. ROC curves also allow comparison of different markers, even if these are on different scales [Heagerty et al., 2000].

Classic ROC curves assume the disease status is fixed. In many cases, however, the disease status may change over time. For these scenarios, Heagerty et al. [2000] proposed the use of time-dependent ROC curves, where the sensitivity, specificity, and corresponding AUC are estimated at a particular time point. There are several ways to estimate sensitivity and specificity in the time-dependent ROC setting [Heagerty et al., 2000, Chambless and Diao, 2006, Zheng et al., 2006, Uno et al., 2007, Heagerty and Zheng, 2005]. Some methods rely on nonparametric estimation, while others use semiparametric methods to estimate the sensitivity and specificity. Several of these estimators [Heagerty and Zheng, 2005, Chambless and Diao, 2006] calculate the sensitivity and specificity using a risk score made of up several covariates, or biomarkers. While the risk score can be estimated in various ways, a common approach is to use a linear predictor based on the partial likelihood estimator [David et al., 1972] described in Section 2.3.

The use of time-dependent ROC curves allows us to investigate the classification performance at different event times and therefore also allows us to investigate the performance of the marker over time. When using time-dependent ROC curves, cases and controls can be defined in several ways, and how these are defined will impact estimation of the sensitivity and specificity. Here, we consider two commonly encountered scenarios: (1) cumulative sensitivity and dynamic specificity and (2) incident sensitivity and dynamic specificity. Cumulative sensitivity considers the probability that an individual's risk score value, B_i, exceeds some threshold value, c, given that the individual experienced an event after baseline and before time t. Dynamic specificity is the probability that an individual has a risk score value less than or equal to c conditional upon not having experienced an event up to time t [Kamarudin et al., 2017]. These can be written as $\text{Sens}_C(t, c) = P(B_i > c|T_i \leq t)$ and $\text{Spec}(t, c) = P(B_i \leq c|T_i > t)$. Incident specificity, on the other hand, is the probability that an individual has a biomarker measure above c given that they experienced an event exactly at time t and can be written as $\text{Sens}_I(t, c) = P(B_i > c|T_i = t)$. The cumulative/dynamic setting is more appropriate when there is a specific time of scientific interest at which investigators would like to see who has developed the disease and who has not. The incident/dynamic setting is more appropriate when the exact event time is known and it is of interest to investigate who has developed the disease at that time [Kamarudin et al., 2017]. While several estimators have been proposed for both scenarios, we will focus on the cumulative/dynamic estimator of Chambless and Diao [2006] and the incident/dynamic estimator of Heagerty and Zheng [2005].

The work of Heagerty and Zheng [2005] considers time-dependent ROC curves to evaluate the classification performance of risk scores made up of one or more covariates, or biomarkers. The risk scores can be obtained using a variety of models, but their manuscript focuses on the Cox PH model. Sensitivity and specificity are also estimated using the Cox PH model. When the data follow non-proportional hazards, the sensitivity is estimated as $\widehat{\text{Sens}}_{HZ}(t,c) = \sum_{k=1}^{n}[I(B_k > c)Y_k(t)\exp(B_k\hat{\gamma}(t))/\sum_{j=1}^{n} Y_j(t)\exp(B_i\hat{\gamma}(t))]$ where $Y_k(t)$ is an indicator for whether individual k is at risk at time t and $\hat{\gamma}(t)$ is the estimate for the time-dependent coefficient. The specificity is estimated using $\widehat{\text{Spec}}_{HZ}(t,c) = \frac{\sum_{k=1}^{n} I(B_k>c)Y_k(t+)}{W^R(t+)}$ where $Y_k(t+) = \lim_{\delta\to 0} Y_k(t+|\delta|)$ and $W^R(t+)$ is the number of controls (non-events) in the risk set at time t. When the data follow PH, $\hat{\gamma}(t)$ in the sensitivity is replaced by $\hat{\gamma}$, an estimate for the time-invariant effect.

Another way to estimate the sensitivity and specificity under a time-dependent ROC curve uses the methods proposed by Chambless and Diao [2006] for the cumulative/dynamic scenario. These methods also allow for assessment of a risk score, which can be obtained using various models. In their manuscript, the authors present Kaplan-Meier [Kaplan and Meier, 1958] type estimators of the sensitivity and specificity as well as regression-based estimators. The R function `AUC.cd` in the `survAUC` package implements the regression-based approach using a Cox PH model. Under this approach, the sensitivity and specificity take the form $\text{Sens}_{CD}(t,c) = \frac{E[(1-S(t|B))I(B>c)]}{E[1-S(t|B<c)]}$ and $\text{Spec}_{CD}(t,c) = \frac{E[S(t|B)I(B<c)]}{E[S(t|B)]}$. $S(t|B)$, the survival function at time t, is estimated using $\hat{S}(t|B) = \exp(-\hat{\Lambda}_0(t)\exp(\hat{\gamma}B))$ where $\hat{\Lambda}_0(t)$ is obtained using the Breslow [1972] estimator and $\hat{\gamma}$ is calculated using the Cox PH model.

Other estimators have been proposed that account for censoring using inverse probability weights. The work of Uno et al. [2007] and Hung and Chiang [2010] reweight the estimator of the sensitivity from Heagerty et al. [2000] by the inverse of the survival function for censoring, $S_c(t)$. The estimator for sensitivity is $\widehat{\text{Sens}}_{IW}(t,c) = \frac{\sum_{i=1}^{n} I(B_i>c, X_i\leq t)\delta_i/[n\hat{S}_c(X_i)]}{\sum_{i=1}^{n} I(X_i\leq t)(\delta_i/[n\hat{S}_c(X_i)])}$ and the estimator for specificity is $\widehat{\text{Spec}}_{IW}(t,c) = \frac{\sum_{i=1}^{n} I(B_i\leq c, X_i>t)}{\sum_{i=1}^{n} I(X_i>t)}$. This estimator, however, does not account for marker dependent censoring. Blanche et al. [2013] extended this method to allow for marker dependent censoring. Their proposed estimators are $\widehat{\text{Sens}}_B(t,c) = \frac{\sum_{i=1}^{n} I(B_i > c, X_i \leq t)[\delta_i/(n\hat{S}_c(X_i|B_i))]}{\sum_{i1}^{n} I(X_i \leq t)(\delta_i/(n\hat{S}_c(X_i|B_i)))}$ and $\widehat{\text{Spec}}_B(t,c) = \frac{\sum_{i=1}^{n} I(B_i \leq c, X_i > t)[1/(n\hat{S}_c(t|B_i))]}{\sum_{i1}^{n} I(X_i > t)(1/(n\hat{S}_c(t|B_i)))}$. While this estimator reweights by the inverse probability of censoring, it is based on Kaplan-Meier type estimators of the sensitivity and specificity.

Note that several of the estimators introduced in this section rely upon the use of the Cox PH model. As seen in Section 2.4, however, under model mis-specification the estimand corresponding to the partial likelihood estimator depends on the censoring distribution. Because several time-dependent ROC curve estimators rely upon use of the Cox PH model, it is important to investigate if and how model mis-specification impacts estimates of the area under the curve. We will explore this topic in more detail in Chapter 5.

2.6 Nested Case-Control Study

So far, we have introduced the partial likelihood estimator. However, use of this estimator requires that full covariate information is collected for all study participants, which is not always feasible. Under the partial likelihood estimator, cases provide more information than controls do. This motivates the idea that failure to include all controls in the analysis may lead to a relatively small loss in efficiency, which becomes especially useful in studies of rare outcomes where we have a lot more controls than we do cases. The sampling schemes for the nested case-control and case-cohort designs make use of this fact to provide efficient estimators that do not require full covariate information for all study participants. We start off by showing the influence of cases and controls under the partial likelihood estimator. In the rest of the section, we introduce the nested case-control design along with several variations.

2.6.1 Influence of Cases and Controls in the Cox Model

In this section, we motivate the nested case-control and case-cohort designs that will be the focus of the remainder of the chapter. These methods only require full covariate information on a subset of the entire cohort. While we do not consider all controls, these designs require the analysis of all cases, which is motivated by considering the contributions of subjects that do and do not have a failure time observed.

2.6.2 The partial information in the two-sample case

To examine the relative contribution of statistical information coming from cases and controls in a survival setting, we begin by analytically assessing the information contributed from cases and controls in the setting of a two-sample comparison. Specifically, we derive Fisher's Information for the Cox model when a single binary covariate, $z \in \{0, 1\}$, is considered. In this case, the score function for the Cox model given in (2.10), is given by

$$U(\beta) = \frac{\partial l(\beta)}{\partial \beta} = \sum_{j=1}^{D} \left[z_j - \frac{\sum_{l \in R(t_j)} z_l e^{z_l^T \beta}}{\sum_{l \in R(t_j)} e^{z_l^T \beta}} \right], \tag{2.15}$$

and negating the derivative of (2.15) with respect to β yields the observed information:

$$I(\beta) = -\frac{\partial U(\beta)}{\partial \beta} = \sum_{j=1}^{D} \left\{ \frac{\sum_{l \in R(t_j)} e^{z_l^T \beta} \sum_{l \in R(t_j)} z_l^2 e^{z_l^T \beta} - (\sum_{l \in R(t_j)} z_l e^{z_l^T \beta})^2}{(\sum_{l \in R(t_j)} e^{z_l^T \beta})^2} \right\}. \tag{2.16}$$

Under the null hypothesis of no covariate effect, $H_0 : \beta = 0$, we then have

$$I(0) = \sum_{j=1}^{D} \left[\frac{n_j \cdot n_{1j} - n_{1j}^2}{n_j^2} \right] = \sum_{j=1}^{D} \left[\frac{n_{1j}(n_j - n_{1j})}{n_j^2} \right] = \sum_{j=1}^{D} \left[\frac{n_{1j} n_{0j}}{n_j^2} \right]$$

where n_j, n_{1j}, $n_{0j} = n_j - n_{1j}$, denote the total number of subjects at risk, the number of subjects at risk in group 1, and the number of subjects at risk in group 0 at time t_j, respectively.

To approximate the observed information under the null hypothesis of equal survival between groups, note that the expected number of patients at risk in each group is given by

$$\mathrm{E}[n_{0j}] = N \cdot (1-\pi) \cdot S(t_j-) \cdot [1 - C(t_j-)] \quad \text{and} \quad \mathrm{E}[n_{1j}] = N \cdot \pi \cdot S(t_j-) \cdot [1 - C(t_j-)]$$

and hence $\mathrm{E}[n_j] = \mathrm{E}[n_{0j}] + \mathrm{E}[n_{1j}] = N \cdot S(t_j-) \cdot [1 - C(t_j-)]$, where N is the total number of subjects, π is the probability of being in group 1, $S(t_j-)$ is the survival function evaluated just prior to time t_j, and $C(t_j-)$ is the censoring function evaluated just prior to time t_j. Replacing each number of at risk subjects with their respective expectations, we obtain

$$\begin{aligned} I(0) &\approx \sum_{j=1}^{D} \left\{ \frac{\{N \cdot S(t_j-) \cdot [1 - C(t_j-)]\}^2 \cdot \pi - \{N \cdot S(t_j-) \cdot [1 - C(t_j-)]\}^2 \cdot \pi^2}{\{N \cdot S(t_j-) \cdot [1 - C(t_j-)]\}^2} \right\} \\ &= \sum_{j=1}^{D} [\pi \cdot (1-\pi)] = D \cdot \pi \cdot (1-\pi). \end{aligned}$$

Suppose, for example, that $\pi = \frac{1}{2}$ (as is the case for a 1:1 randomized experiment). Then

$$I(0) \approx D \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{D}{4},$$

so that the information is solely a function of the total number of events observed (cases), as opposed to the total number of subjects (cases plus controls) available for analysis.

2.6.3 An empirical assessment of the influence of cases and controls

The results of the previous section focus on the number of events observed as opposed to the number of subjects in total. This is because most of the information we are drawing from the data set comes from the cases.

One way to measure influence is through the delta-beta values calculated from the residuals. Delta-beta values compare the coefficient estimates when the model is fit using the entire data set to those where the i^{th} subject is excluded. Figure 2.2 shows the delta-beta values calculated from a Cox PH model fit to simulated data ($N = 200$ observations) against a single predictor (age). Standardized age was simulated from a Normal(0,1). Survival times were simulated from a Exponential distribution with hazard rate $0.5 \times \exp\{\log(2) \times \text{Age}\}$. Censoring times were simulated from a Uniform(0,2) distribution, resulting in 67.5% censoring.

Delta-beta values for observed events (cases) are depicted with closed triangles and values for censored observations (controls) are depicted with open circles. For visual reference, the horizontal lines in the figure are plus/minus one standard deviation of the delta-beta values. Notice that subjects for whom an event was not observed tend to have delta-beta values close to 0, while those delta-beta values that are relatively large in magnitude generally belong to cases for whom an event was observed. More specifically, 67% of cases have a delta-beta value greater than one standard deviation in magnitude. This is true for only 6.3% of controls. Though only a single empirical example, this further emphasizes the idea that a disproportionate amount of information is derived from the cases and therefore justifies the idea of not using all controls in the nested case-control and the case-cohort studies.

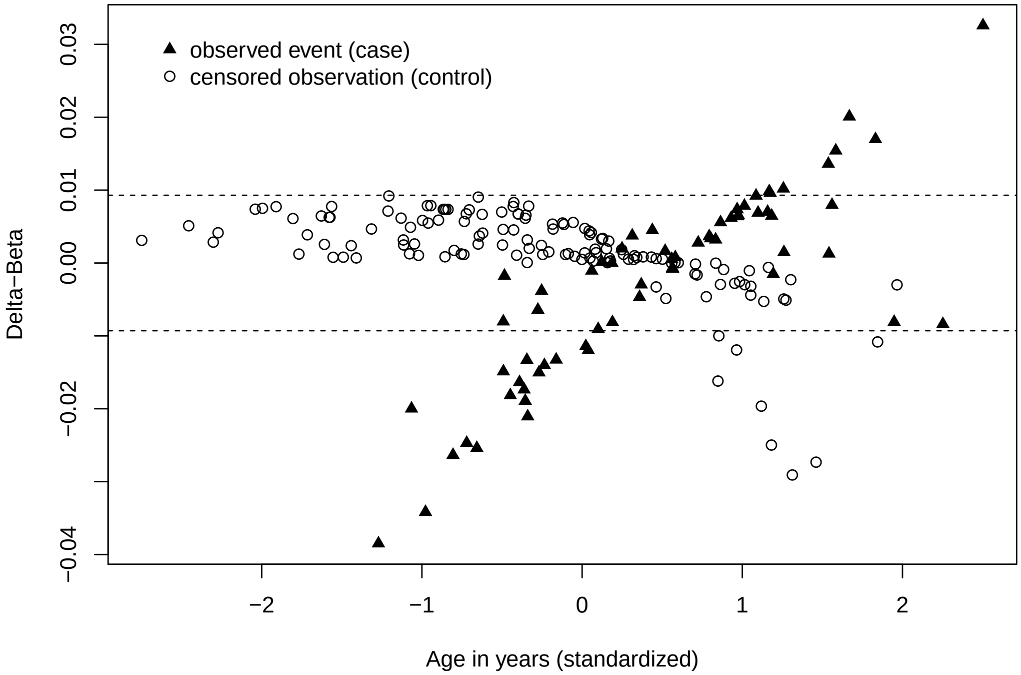


Figure 2.2: Plot of delta-beta values calculated from a Cox PH model fit to simulated data ($N = 200$ observations) against a single predictor (age). Standardized age was simulated from a Normal(0,1). Survival times were simulated from a Exponential distribution with hazard rate $0.5\times \exp\{\log(2) \times \text{Age}\}$. Censoring times were simulated from a Uniform(0,2) distribution, resulting in 67.5% censoring. Delta-beta values for observed events (cases) are depicted with closed triangles and values for censored observations (controls) are depicted with open circles. For visual reference, the horizontal lines in the figure are plus/minus one standard deviation of the delta-beta values.

2.6.4 Introduction to the nested case-control design

As seen in the previous section, under the partial likelihood estimator, cases provide more information than controls do. The nested case-control design makes use of this idea by only sampling a specified number of controls for each event. The number of controls to be sampled, M, is selected before analysis of the data. At each event time, we randomly sample M controls from the subjects who are still at risk [Thomas, 1977]. As such, the nested case-control design exploits the retrospective nature of the partial likelihood (which conditions upon observed event times) by matching cases (subjects that have experienced an event) to controls by their at-risk status at the time of the event.

Recall from (2.8) that the partial likelihood when using the full data takes the following form

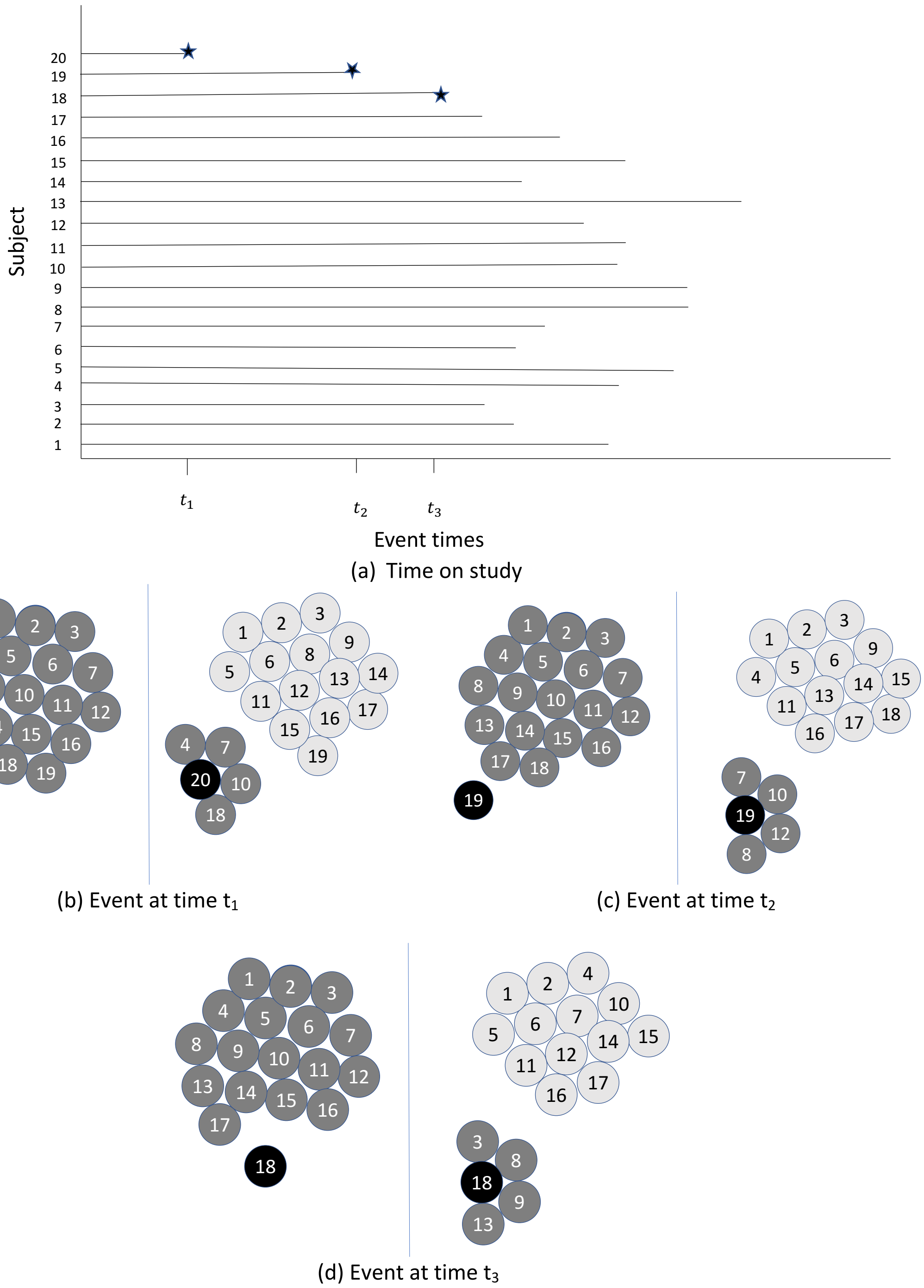
$$L = \prod_{j=1}^{D} \frac{e^{z_j^T \beta}}{\sum_{l \in R(t_j)} e^{z_l^T \beta}}.$$

Due to the matching involved in the nested case-control design, a stratified version of the partial likelihood given by

$$L_{NCC} = \prod_{j=1}^{D} \frac{e^{z_j^T \beta}}{e^{z_j^T \beta} + \sum_{l=1}^{M} e^{z_l^T \beta}} \tag{2.17}$$

is necessary for parameter estimation to appropriately account for the non-random sampling scheme. Notice that instead of summing over all subjects in the risk set, we only sum over the M controls that were randomly selected for that case. The sampled subjects may be controls for events in the future, or they may become a case themselves.

Figure 2.3 gives an example of the nested case-control design. Notice from Figure 2.3a that we have three event times (t_1, t_2, t_3). During the first event, we randomly select four controls from those who still have not had an event. In 2.3b we show the controls when using the full data (left) versus the four sampled controls (right) used with the nested case-control design: 4, 7, 10, and 18. The next event occurs at time t_2. Notice that subject 20 is no longer in the risk set since its event already occurred. We again sample four controls from those who are still at risk at time t_2. This time, we select subjects 7, 8, 10, and 12. Note that we can have controls for one event be controls for a later event. In this case we have that subjects 7 and 10 were sampled as controls for the first and the second events. The third event occurs at time t_3. The nested case-control design also allows us to sample controls who become cases at a later time. In this example, we find that subject 18, who was a control during the first



(a) Time on study

(b) Event at time t_1

(c) Event at time t_2

(d) Event at time t_3

Figure 2.3: An example of the nested case-control design. 2(a) shows the event times and subjects still at risk at each event time. 2(b)-(d) show the controls that the case would be compared to in the full cohort scenario (left) and in the nested case-control (right). Black represents the case, dark gray are the controls that the case is compared to, and light grey represents those in the risk set that are not included when using the nested case-control design.

event time, became a case.

It has been shown not only that the estimates obtained from the nested case-control design are consistent, but also that they are asymptotically normally distributed [Goldstein and Langholz, 1992].

Because the nested case-control design does not use all of the information, it is not as efficient as using the full data set. However, in many studies where covariate information may be expensive or burdensome to collect, researchers may prefer a slight loss of efficiency.

2.6.5 Equivalence of the Cox proportional hazards and conditional logistic regression model under the nested case-control design

In this section, we show the equivalence of the conditional logistic regression likelihood and Cox's stratified partial likelihood, reinforcing the connection between the nested case-control design and the usual matched case-control design. To see the connection, consider fitting a conditional logistic regression model to matched case-control data in which we assume that there are M controls for each case. Thus each matching strata contains $M+1$ observations (1 case plus M controls). For strata i, let $\vec{Y_i} = (Y_{i1}, \ldots, Y_{iM+1})$ denote the vector of binary responses for all subjects in the strata. Further note that $\sum_{j=1}^{M+1} Y_{ij} = 1$ in the case of $1:M$ matching. Then the corresponding conditional logistic regression model is given by

$$\text{logit}\left(\Pr[Y_{ij} = 1 \big| \vec{z}_{ij}]\right) = \beta_{0i} + \beta_1 z_{ij1} + \cdots + \beta_p z_{ijp} = \beta_{0i} + \vec{z}_{ij}\vec{\beta} \equiv \eta_{ij}, \tag{2.18}$$

where $\vec{z}_{ij}$ is a vector of covariate values for subject j in strata i, $j = 1, \ldots, M+1$, and $\vec{\beta}$ are the corresponding regression parameters (excluding the intercept). Further, denote

$\pi_{ij} \equiv \Pr\left[Y_{ij} = 1 \middle| \vec{z}_{ij}\right]$ and note that the model specification in (2.18) and the fact that $\sum_{j=1}^{M+1} Y_{ij} = 1$ implies

$$\pi_{ij} \equiv \Pr\left[Y_{ij} = 1 \middle| \vec{z}_{ij}\right] = \frac{e^{\eta_{ij}}}{1 + \sum_{k=1}^{M+1} e^{\eta_{ik}}}. \tag{2.19}$$

The contributions to the likelihood for the conditional logistic regression model can be formulated by noting that $\vec{Y}_i = (Y_{i1}, \ldots, Y_{iM+1})$ is a multinomial random variable and hence

$$\begin{aligned}\Pr\left[\vec{Y}_i = \vec{y} \middle| \sum_{k=1}^{M+1} y_{ik} = 1, Z_{2,i}, \vec{z}_i\right] &= \frac{\Pr\left[\vec{Y}_i = \vec{y}, \sum_{k=1}^{M+1} y_{ik} = 1 \middle| Z_{2,i}, \vec{z}_i\right]}{\sum_{\vec{y}^*:\sum y_{ik}^*=1} \Pr\left[\vec{Y}_i = \vec{y}^* \middle| Z_{2,i}, \vec{z}_i\right]} \\ &= \frac{\prod_{j=1}^{M+1} \pi_{ij}^{y_{ij}}}{\sum_{\vec{y}^*:\sum y_{ik}^*=1} \prod_{j=1}^{M+1} \pi_{ij}^{y_{ij}^*}},\end{aligned}$$

where $Z_{2,i}$ denotes all the matching covariates that were used to define the i^{th} strata. Plugging in the values of π_{ij} as given in (2.19) we obtain

$$\begin{aligned}\Pr\left[\vec{Y}_i = \vec{y} \middle| \sum_{k=1}^{M+1} y_{ik} = 1, Z_{2,i}, \vec{z}_i\right] &= \frac{\exp\left\{\sum_{j=1}^{M+1} y_{ij}\eta_{ij}/(1 + \sum_{j=1}^{M+1} e^{\eta_{ij}})\right\}}{\sum_{\vec{y}^*:\sum y_{ik}{}^*=1} \exp\left\{\sum_{j=1}^{M+1} y_{ij}^*\eta_{ij}/(1 + \sum_{j=1}^{M+1} e^{\eta_{ij}})\right\}} \\ &= \frac{\exp\left\{\sum_{j=1}^{M+1} y_{ij}\eta_{ij}\right\}}{\sum_{\vec{y}^*:\sum y_{ik}{}^*=1} \exp\left\{\sum_{j=1}^{M+1} y_{ij}^*\eta_{ij}\right\}},\end{aligned}$$

where the last equality follows because the denominators in the numerator and denominator terms do not depend on y_{ij} and hence cancel. Incorporating covariates via the linear predictor, $\eta_{ij} = \beta_{0i} + \vec{z}_{ij}\vec{\beta}$, and simplifying the result by canceling $e^{\beta_{0i}}$ from the numerator and the denominator we obtain the likelihood contribution for subject i as:

$$\begin{aligned}\Pr\left[\vec{Y}_i = \vec{y} \middle| \sum_{k=1}^{M+1} y_{ik} = 1, Z_{2,i}, \vec{z}_i\right] &= \frac{\exp\{\beta_{0i}\}\exp\left\{\sum_{j=1}^{M+1} y_{ij}\vec{z}_{ij}\vec{\beta}\right\}}{\sum_{\vec{y}^*:\sum y^*_{ik}=1}\exp\{\beta_{0i}\}\exp\left\{\sum_{j=1}^{M+1} y^*_{ij}\vec{z}_{ij}\vec{\beta}\right\}} \\ &= \frac{\exp\left\{\sum_{j=1}^{M+1} y_{ij}\vec{z}_{ij}\vec{\beta}\right\}}{\sum_{\vec{y}^*:\sum y^*_{ik}=1}\exp\left\{\sum_{j=1}^{M+1} y^*_{ij}\vec{z}_{ij}\vec{\beta}\right\}}.\end{aligned}$$

Comparing the above to the factors in (2.17) shows us that the contribution to the likelihood under conditional logistic regression model is equivalent to that of the stratified Cox PH model when we stratify by match group, since the numerator is simply the exponentiated linear predictor for the case and the denominator is the sum of exponentiated linear predictors for all subjects in the strata. To formulate the full likelihood we take the product across all independent strata (corresponding to the observed event times for the nested case-control design), with one case per stratum.

2.6.6 Variations of the Standard Nested Case-Control Design

The usual nested case-control design, as described in Section 2.6.4, randomly samples controls without replacement from the risk set at each event time. Although controls cannot be sampled more than once at each time, they may be sampled again at later event times as long as they are still at risk. Sampled controls may also become cases at later event times. In light of this, variations on the sampling strategies and estimating equations utilized in the nested case-control design have been introduced and investigated in the literature. The following sections introduce the most commonly encountered variations of the usual nested case-control design.

Inverse Probability of Sampling

We start off by introducing the estimator proposed by Samuelsen [1997]. While this estimator relies on the usual nested case-control sampling scheme, it differs from that proposed by Thomas [1977] in the estimation procedure. The standard nested case-control design is estimated using (2.17), which can be written as

$$L_{NCC} = \prod_{j=1}^{D} \frac{e^{z_j^T \beta}}{\sum_{l \in \tilde{R}(t_j)} e^{z_l^T \beta}} \tag{2.20}$$

where $\tilde{R}(t_j) = j \cup S_j$, j is the case at time t_j and S_j represents the set of sampled controls at failure time t_j. The risk sets at each event time, therefore, are made up of the case and sampled controls at each event time, and controls are only included in the risk sets for which they were were sampled as controls. The estimator proposed by Samuelsen [1997], on the other hand, includes all subjects in the nested case-control sample in all risk sets for which subjects are at risk. Each subject's contribution is reweighted by the inverse probability of ever being sampled, $p_l = \delta_l + (1 - \delta_l)[1 - \prod_{X_i < X_j}(1 - \frac{M}{n_i - 1}\delta_i)]$, where n_i is the number of subjects at risk at time t_i and M is the number of controls sampled at each event time. In this case, the pseudo-likelihood takes the form:

$$L_S(\beta) = \prod_{j=1}^{n} \left[\frac{e^{z_j^T \beta}}{\sum_{l \in R(t_j)} \frac{V_l}{p_l} e^{z_j^T \beta}} \right]^{\delta_j}$$

where $R(t_j)$ represents everyone who is at risk at time t_j in the full cohort and V_l is an indicator for whether subject l was included in the nested case-control sample either as a case or as a control. By including all controls forward and backward in time, the estimator proposed by Samuelsen [1997] allows us to gain efficiency compared to the standard estimation procedure.

Risk set sampling

Langholz and Thomas introduced three designs in their 1991 paper. These designs follow the same form as (2.20), but differ in their definition of $\tilde{R}(t_j)$.

The first design, referred to as Design I in Langholz and Thomas [1991] or "retained nested case-control sampling", randomly samples controls at each event time. Those controls are then introduced as controls forward in time, until they are no longer at risk, ie. until they have either failed or been censored. Using the notation presented in Langholz and Thomas [1991], let S_i denote the sampled controls, and R_i denote the full risk set. Their first design considers $\tilde{R}_1(t_j) = \cup_{i \leq j}[S_i \cap R(t_j)]$. In Section 2.6.8 we compare the resulting bias and efficiency of this sampling scheme, which we refer as the *Control Forward* sampling design, to the full cohort and standard nested case-control sampling schemes.

The authors also introduce a design which they refer to as Design II or "augmented nested case-control sampling". This design is motivated by the idea that we could obtain more information if we use the sampled controls forward and backward in time in all risk sets during which they are still at risk. However, because the time that subjects are at risk could be associated with the outcome of interest, this sampling scheme leads to biased estimates; those who are in the study longer have a higher probability of being sampled than those who are not. In order to avoid this problem, the authors introduce the concept of a path set. Using the notation introduced by Thomas and Langholz, a path set Q_{ij} consists of all subjects who enter the study right before time t_i (in the interval $(t_{i-1}, t_i]$) and exit right after t_j (in the interval $(t_j, t_{j+1}]$). Subjects who later become cases are not included in any of the path sets that make up their own risk set since they cannot be their own control. The controls sampled using the nested case control design provide us with random samples from the path sets. Therefore, to include the controls selected during other event times, we may randomly sample these controls by stratifying on the path sets. In this way, we consider samples of individuals with various path set lengths and we again obtain consistency [Langholz and Thomas, 1991]. Sampling is then performed using a two-step approach. The first step follows as usual; we randomly sample a set of M controls at each event time. Let P be the pooled set of controls for all event times. For each sampled control, we also have the path set that it belongs to, so each control is grouped with other controls corresponding to the same path set, Q_{ij}. At each event time we randomly sample controls from the valid path sets (in a way that is representative of the path sets' presence in the risk set) and include those controls, as well as the originally sampled controls, in the analysis.

The third design introduced in their paper selects controls in a way that reduces the number of controls sampled more than once. This design is motivated by the fact that sampling controls from those who had not been selected yet leads to inconsistent parameter estimates [Robins et al., 1986]. Under this proposed design, referred to as Design III in Langholz and Thomas [1991] we randomly sample a path set according to the probability representing that path set at each event time. We then proceed by randomly sampling a subject from the selected path set. Once a subject has been selected, that subject cannot be sampled as a control again until all controls in that path set have been sampled. In this way, we reduce the number of times the same subject is selected. The estimates generated by this design are slightly biased, and when there was an efficiency gain, improvement was very small compared to the standard nested case-control design [Langholz and Thomas, 1991]. In Section 2.6.8, we also use simulation to compare Design III (which we refer to as the *Path Sampling* design) to the full cohort and standard nested case-control sampling designs with respect to bias and efficiency.

Langholz and Thomas were rather surprised by their results. They intuitively expected that by including subjects in more risk sets than just their own, the variance of parameter estimates would naturally decrease. However, what they found, is that with Designs I and II, the variance of parameter estimates can be larger than that of the standard nested case-control design in some cases.

In the full cohort analysis, even though subjects enter many risk sets, the covariance between contributions to the partial likelihood score function is zero conditional on the filtration, or history up to the event time of each case. With the nested case-control design, however, the full history is not known since subjects are randomly sampled at each event time. Including subjects in risk sets other than the ones for which they were sampled thus introduces a covariance term because we do not have the full history of each subject. Therefore, although we are using more information at each event time, the covariance term between contributions to the score can cause the variance of parameter estimates to increase for Designs I and II. Design III also introduces a covariance term, but in this case the covariance is negative because, by design, we are trying to sample controls that are different than the ones present. Although this decreases the variance of the parameter estimates stemming from this design, because we try to sample as many unique subjects as possible, the same number of controls would require a larger number of unique subjects to be included in the overall analysis. If the goal is to reduce the number of subjects for whom complete covariate information is required (the general motivation for using a nested case-control design), the only alternative is to sample less controls per subject to match the number of controls in the usual nested case-control sampling framework, hence increasing the variance of parameter estimates. Ultimately, these opposing forces effectively cancel out and the path-sampling strategy of Design III yields parameter estimate variances similar to that of the standard nested case-control design when the total number of unique subjects that are included in the analysis are held constant between the two sampling strategies.

Biased selection of controls

We have explained that randomly sampling controls from all subjects in the risk set leads to consistent estimates. There are other practical ways of sampling controls, however, that will lead to inconsistent estimates.

One sampling scheme is introduced by Lubin and Gail [1984] as the sampling of "pure controls". Sampling of pure controls refers to sampling only subjects who have not failed due to the outcome of interest and have not been censored by the end of the study. In this way, we can never sample subjects who become cases in the future. Applying such a restriction leads to under-representation of cases in the analysis because, although cases are present in other risk sets, they are never included in risk sets other than their own. Lubin and Gail [1984] state that when the disease is rare, the bias induced by pure control sampling is very small. This can be explained by the fact that when the outcome is rare, the probability of sampling a case is also very small since there are very few cases to consider. Therefore, even if cases had been included as controls, the probability of sampling these as controls would have been very small.

Another possible sampling scheme, referred to as "case exclusion", only samples controls that do not become cases during some future interval of time. Notice that inclusion of controls in the risk set under this scenario requires that the subject has not become a case or been censored up to the current time. Moreover, it requires that they do not become a case in the specified interval, and that they are sampled as a control for the event time. Because we are sampling controls conditional upon their future, we are no longer randomly selecting subjects as controls, which induces bias.

A final control sampling scheme that might be considered is referred to as "control exclusion". In this scenario, we exclude subjects with diseases related to exposure [Lubin and Gail, 1984]. We can think about this sampling scheme in the context of competing risks. Because subjects are not included as controls if they develop another disease related to exposure, we will not record the disease of interest if it does occur. Moreover, the diseases excluded from the analysis may be precursors to the outcome of interest so removing these from the analysis will bias the results.

Although the standard nested case-control design has been shown to be consistent for the true parameters, variations of the method may induce bias. By forcing subjects into or out of risk sets, we cause over- or under-representation of these subjects in our data such as in the "pure controls" and "case exclusion" sampling designs presented. In practice, we may also find that researchers often apply exclusion criteria for study participants. "Control exclusion" shows that we should be careful when determining such criteria as this may also induce bias. Oftentimes, we find that in practice we may be forced into sampling controls in a specific way and it is important that researchers understand the effect of such sampling schemes on the reliability of their results. Moreover, understanding how bias is induced may also allow statisticians to account for such bias [Lubin and Gail, 1984].

2.6.7 Software implementation of the standard nested case-control design

The nested case-control sampling scheme can be implemented using the `Epi` package within R or via the `sttocc` command in STATA. In addition, SAS macros exist that can implement the standard nested case-control design. Within the R `Epi` package, the function `ccwc` samples the controls for each event, and arranges the data to separate each group; for each event time, the case and its sampled controls are considered to be one group. Once the data are in the correct format, we can fit a model using either conditional logistic regression or a stratified Cox PH model stratified by group. As was shown in Section 2.6.5, the two models are equivalent. The command `sttocc` within STATA samples the controls and allows for fitting in an analogous way. In the Appendix of this chapter we provide all necessary R code for implementing the data analysis presented in Section 2.8. We also present examples of STATA code to perform the nested case-control analysis and provide reference to a SAS macro that implements the nested case-control design.

2.6.8 Simulated performance of the nested case-control design

In this section, we use simulation studies to investigate consistency and relative efficiency of estimates when using the nested case-control (NCC) design compared to the full data analysis. We start with approximately 60% censoring and increase censoring to determine if and how the consistency and efficiency of the nested case-control designs change. For illustrative purposes we consider the standard nested case-control as well as the control forward and path sampling designs (Designs I and III from Langholz and Thomas (1991), respectively) using M = 1, 2, 3, and 4. We also display results that illustrate the efficiency trade-offs to be made if a simple random sample (unconditional upon disease status) of the same size as the standard nested case-control design is used.

Survival times for our simulation study were generated from a Exponential distribution having hazard rate $0.5\times \exp\{\beta Z\}$, with $Z \sim$ N(0,1) distribution. The true parameter value associated with the covariate was taken to be $\beta = \log(2.0) = 0.693$ (true HR of $e^{\beta} = 2.0$). We generated censoring times using a Uniform(0, 2.25) distribution to obtain approximately 60% censoring and took the observed time to be the minimum of the survival and censoring times. The 85% censoring scenario was generated similarly, but censoring times were drawn from a Uniform(0, 0.5) distribution.

Notice that under approximately 60% censoring (Table 2.2), all variations yield consistent estimates. However, taking a random sample from the full data is more efficient than using the nested case-control design and any of its variations. Because there are a large number of cases in this scenario, taking a random sample will still provide a sufficient number of cases. When we use a random sample with approximately 625 subjects, the variance is only approximately 64% larger than the variance using the full cohort. In this scenario, the nested case-control design, although it performs well, is not required since a random sample also yields consistent estimates and is more efficient.

Table 2.2: Simulations for the comparison of the full cohort analyses, standard nested case-control, control forward sampling, path sampling, and a simple random sample of same sample size as the standard nested case-control design. True survival times were simulated from a Exponential distribution with hazard rate 0.5× $\exp\{\beta Z\}$, with $Z \sim$ N(0,1) distribution. The true parameter value associated with the covariate was taken to be $\beta = \log(2.0) = 0.693$ (true HR of $e^{\beta} = 2.0$) and we have approximately 60% censoring. Censoring was obtained using a Uniform(0, 2.25) distribution. 10,000 simulations with 1000 subjects each were performed.

Sampling Design	Total Subjects	Exp(Coeff.)	Coeff	Empirical Variance	Relative Efficiency
Full Cohort	1000.0	2.01	0.695	0.003	1.00
(M=4)					
NCC	844.7	2.01	0.696	0.004	1.48
Control Forward	844.7	2.01	0.696	0.005	1.55
Path Sampling	893.2	2.01	0.696	0.004	1.31
Random Sample	844.7	2.01	0.696	0.003	1.19
(M=3)					
NCC	803.5	2.01	0.696	0.005	1.61
Control Forward	803.5	2.01	0.696	0.005	1.78
Path Sampling	862.4	2.01	0.696	0.004	1.39
Random Sample	803.5	2.01	0.696	0.004	1.23
(M=2)					
NCC	737.6	2.01	0.697	0.006	1.92
Control Forward	737.6	2.02	0.698	0.007	2.29
Path Sampling	807.6	2.01	0.697	0.005	1.67
Random Sample	737.6	2.01	0.696	0.004	1.36
(M=1)					
NCC	624.6	2.02	0.699	0.008	2.79
Control Forward	624.6	2.03	0.701	0.011	3.84
Path Sampling	682.8	2.02	0.698	0.007	2.43
Random Sample	624.6	2.01	0.696	0.005	1.64

In the case of a rare event, because there are less cases, taking a random sample will no longer be as efficient since we will be left with an even smaller number of cases in our analysis. This can be seen in Table 2.3, in which we present the results for approximately 85% censoring. Under this scenario, we see that the nested case-control design is more efficient than simply taking a random sample. When we use one control per case, the standard nested case-control design variance is approximately three times larger than that of the full cohort. Using the same number of subjects in total, but taking only a random sample gives us a variance more than four times larger than that of the full cohort. Similar patterns are observed when we have more than one control per case.

Table 2.3: Simulations for the comparison of the full cohort analyses, standard nested case-control, control forward sampling, path sampling, and a simple random sample of same sample size as the standard nested case-control design. True survival times were simulated from a Exponential distribution with hazard rate 0.5× $\exp\{\beta Z\}$, with $Z \sim$ N(0,1) distribution. The true parameter value associated with the covariate was taken to be $\beta = \log(2.0) = 0.693$ (true HR of $e^{\beta} = 2.0$) and we have approximately 85% censoring. Censoring was obtained using a Uniform(0, 0.5) distribution. 10,000 simulations with 1000 subjects each were performed.

Sampling Design	Total Subjects	Exp(Coeff.)	Coeff	Empirical Variance	Relative Efficiency
Full Cohort	1000.0	2.01	0.696	0.008	1.00
(M=4)					
NCC	484.2	2.02	0.699	0.012	1.50
Control Forward	484.2	2.03	0.701	0.014	1.81
Path Sampling	604.0	2.02	0.699	0.011	1.31
Random Sample	484.2	2.03	0.700	0.017	2.18
(M=3)					
NCC	419.3	2.03	0.701	0.014	1.69
Control Forward	419.3	2.04	0.704	0.017	2.16
Path Sampling	502.8	2.03	0.701	0.012	1.49
Random Sample	419.3	2.03	0.700	0.020	2.53
(M=2)					
NCC	341.8	2.04	0.704	0.016	2.03
Control Forward	341.8	2.06	0.710	0.023	2.85
Path Sampling	382.7	2.03	0.702	0.014	1.80
Random Sample	341.8	2.04	0.702	0.025	3.17
(M=1)					
NCC	248.7	2.06	0.709	0.024	3.05
Control Forward	248.7	2.09	0.716	0.037	4.58
Path Sampling	259.6	2.05	0.708	0.023	2.86
Random Sample	248.7	2.06	0.704	0.035	4.35

Moreover, notice that the standard nested case-control design performs better than the

Control Forward design. While the Path sampling design appears to do better than the nested case-control design, it should be noted that the Path sampling design requires a larger number of subjects in total as it is designed to reduce the number of controls that are included more than once. If instead of matching on the number of controls, we match on the total number of subjects included, we find that the standard design still performs better. For example, the number of subjects included for path sampling when $M = 3$ is approximately 500, which is larger than the number of subjects used for the standard nested case-control design with $M = 4$. Although the variances are equal now, the Path sampling design still requires more subjects than the standard design. Because the goal of the nested case-control design is to reduce the number of subjects required for analysis, we find that the standard design performs better than the Path sampling design.

2.7 Case-Cohort Design

2.7.1 Introduction to the case-cohort design

The case-cohort design, similar to the nested case-control design, reduces the number of controls utilized in an analysis. While the nested case-control design selects controls at the time of each event, the case-cohort design selects controls ahead of time. Before analysis or the start of the experiment, researchers and statisticians randomly select a subcohort from all patients in the study. The subcohort is based on some pre-specified proportion, call it α. Because the subcohort is randomly selected, it may include subjects who never become cases, as well as subjects who eventually become cases. The analysis then consists of the subcohort, as well as all of the cases (even if they were not sampled into the subcohort).

Recall the partial likelihood when considering the full cohort (2.8):

$$L = \prod_{j=1}^{D} \frac{e^{z_j^T \beta}}{\sum_{l \in R_j} e^{z_l^T \beta}}.$$

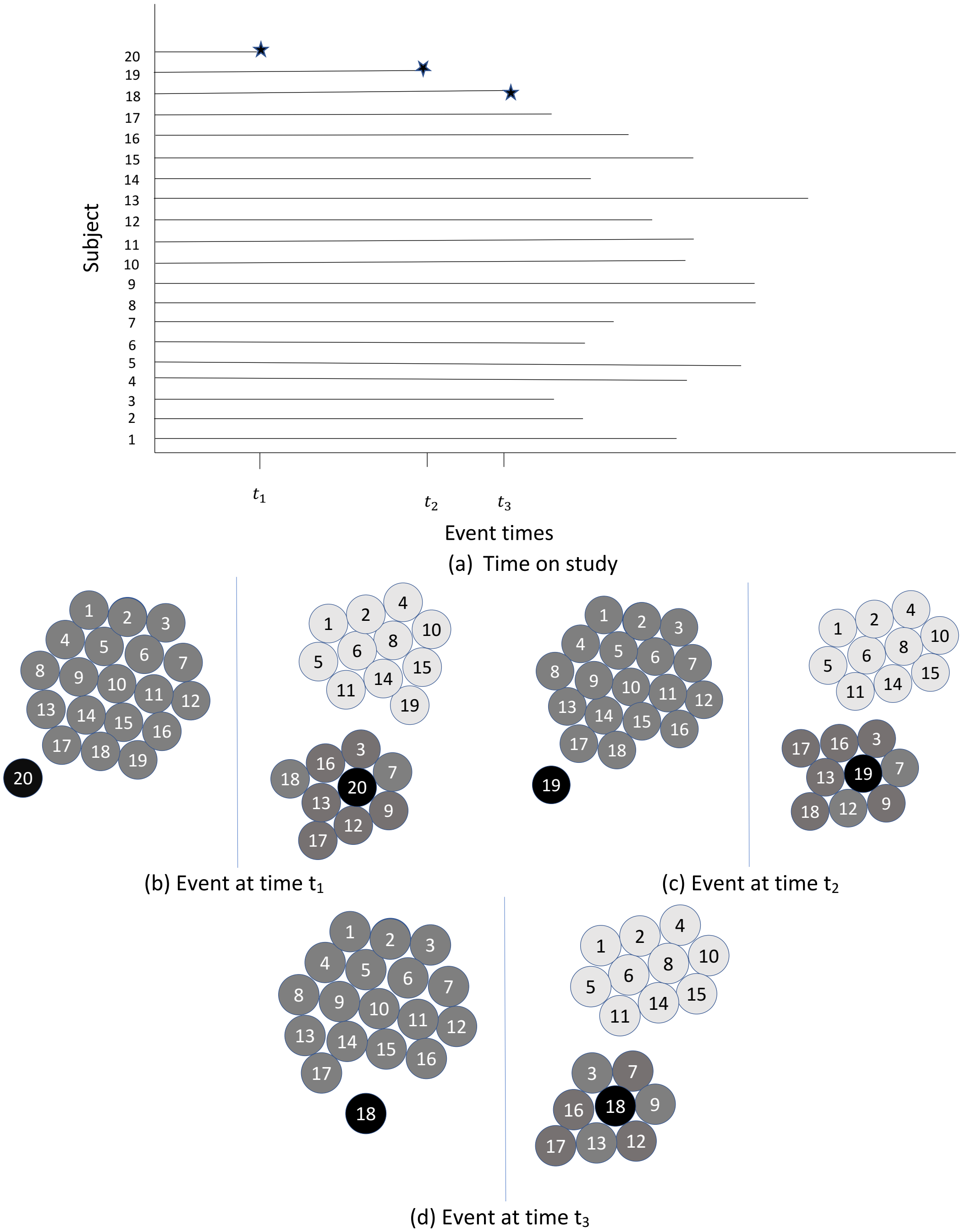
One form of the case-cohort design includes the cases that were not part of the original subcohort right before their event time [Prentice, 1986]. If the case is not in the subcohort, the psuedo-likelihood contribution for the case takes the form:

$$L_j = \frac{e^{z_j^T \beta}}{e^{z_j^T \beta} + \sum_{l \in R_{S,j}} e^{z_l^T \beta}} \tag{2.21}$$

where $R_{S,j}$ represents subjects who are still at risk at time t_j and who were also part of the selected subcohort. If the case is already in the cohort, the contribution to the psuedo-likelihood remains as usual.

Figure 2.4 illustrates an example of the sampling scheme for the case-cohort design assuming the same data as in the nested case-control example (see Figure 2.3) with a subcohort proportion, α=0.4. The selected subcohort consists of the eight subjects that subject 20 is compared to at the first failure time. Notice that the subcohort consists of subject 18, who becomes a case at time t_3. We also have that subject 19 is not in the selected subcohort, so it is not included as a control at time t_1; it is included in the analysis only at its own event time. It has been shown that the estimators based on this method are consistent. Further, using the fact that the score statistic is asymptotically normally distributed along with a Taylor series expansion, it can be shown that the estimator derived from the case-cohort design is also asymptotically normally distributed [Self and Prentice, 1988].

Another common variation of the case-cohort design follows a similar structure, but includes cases at all times during which they were in the risk set as opposed to only including them at the time of their event [Lin and Ying, 1993]. This estimator was derived for scenarios in which we have missing data, but can be used for the case-cohort design because we are



(a) Time on study

(b) Event at time t_1

(c) Event at time t_2

(d) Event at time t_3

Figure 2.4: An example of the case cohort design with $\alpha = 0.4$. We consider the same scenario as that presented in Figure 2. 3(a) shows the event times and the subjects who are still at risk at each event time. 3(b)-(d) show the controls that the case would be compared to in the full cohort scenario (left) and in the case-cohort design (right). Black represents the case, controls are in dark gray, and subjects in the risk set who are not used as controls in the case-cohort design are light gray.

”missing” data on all subjects who are neither part of the subcohort, nor cases. In the likelihood, subjects are weighted according to their subcohort status:

$$L_{SC} = \prod_{j=1}^{D} \frac{w_j e^{z_j^T \beta}}{\sum_{l \in R_{SC,j}} w_l e^{z_l^T \beta}} \tag{2.22}$$

where $R_{SC,j}$ is the set of all subjects who are part of the subcohort or who become cases and are still in the risk set at time t_j. For cases, the weight, w_j is 1, and for subcohort members, the weight is $1/\alpha$.

Notice that these weights are necessary to yield consistent parameter estimates. If one simply included cases for all risk sets during which they are risk, bias in the estimated coefficients would result because we are oversampling cases in each of the risk sets relative to the original random sample.

The design proposed by Lin and Ying [1993], assuming the example from Figure 2.4, would then include subjects 18 and 19 as controls for subject 20 at time t_1. Since subject 18, a case, is already in the subcohort, the controls for subject 19 at time t_2 would remain as they currently are in the example, as would those for subject 18 at time t_3.

The Lin and Ying [1993] estimator, which includes all subcohort members and all cases who are still at risk at each event time, belongs to a class of estimators that have been termed D-estimators [Kulich and Lin, 2004]. Additional members of the class of D-estimators and alternative designs for including cases in the analysis of case-cohort sampled data have been proposed in the literature. We will briefly discuss these in Section 2.7.2.

Various approaches have been proposed to estimate the variance of the regression parameter estimates under the case-cohort design. One of these estimators, which accounts for correlation in the score function and which may be applied to more complicated structures, is the robust variance estimator [Barlow, 1994]. It allows for estimation of the variance using a function of the delta-beta values. As previously discussed, the delta-beta values quantify the influence of each observation on the parameter estimates. Because the robust variance estimator can be easily calculated and may be implemented under various sampling schemes, it is a recommended alternative to the usual variance estimator.

2.7.2 Implementation of the case-cohort design

As stated in Section 2.7.1, there are various ways to include non-subcohort cases in the analysis of data stemming from a case-cohort design. Some of the ways in which the implementation may differ is in how subjects are weighted in the partial likelihood calculation. In this section, we discuss some of the weighting schemes available for the case-cohort design.

Weighting schemes for the case-cohort design

The case-cohort design may vary in the selection of weights used in the pseudo-likelihood. Notice that the contribution to the pseudo-likelihood in (2.21) may be written as

$$L_{j,CC} = \frac{e^{z_j^T \beta}}{w_j(t_j)e^{z_j^T \beta} + \sum_{l \in R_{s,j}} w_l(t_j) e^{z_l^T \beta}} \tag{2.23}$$

.

where w represents a weight. In (2.21), $w_j(t_j)$ and $w_l(t_j)$ are equal to one for all j and l.

In all of the weighting schemes we consider, subjects who are not part of the subcohort and who never become cases receive a weight of zero at all times. The first weighting scheme we consider places a weight equal to one for subcohort members at all times during which they are at risk. Observations who are not part of the subcohort but who become a case receive a weight of one only at their event time [Prentice, 1986]. Another weighting scheme differs

Table 2.4: This table was adopted from Barlow (1999) and presents various weighting schemes for the case-cohort design for different failure status and times.

Outcome and time	Prentice (1986)	Self and Prentice (1988)	Barlow (1994)
Non-subcohort case before failure	0	0	0
Non-subcohort case at failure	1	0	1
Subcohort case before failure	1	1	$1/\alpha$
Subcohort case at failure	1	1	1
Subcohort control	1	1	$1/\alpha$

from that of Prentice [1986] in that cases who are part of the subcohort receive a weight of one during their event time, while cases outside of the subcohort receive weights of zero even at their event time [Self and Prentice, 1988].

In a third weighting scheme, subjects who become cases receive a weight of one at their event time only. If cases are not part of the subcohort, they receive a weight of zero outside of their event time, and if they are part of the subcohort they receive a weight of $1/\alpha$ outside of their event time. For all members of the subcohort who don't experience an event, the weight is $1/\alpha$ at all times during which they are still in the risk set [Barlow, 1994].

Table 2.4, adopted from Barlow et al. [1999], summarizes the weight assignment for each of the methods described. It has been shown that the Prentice method yields results most similar to those of the full cohort, while the Self and Prentice method yields the largest differences. For large enough subcohorts, however, the three methods yield similar results [Onland-Moret et al., 2007].

Because correlation is induced when we include a case that was not part of the original subcohort, we must account for this correlation in the variance. Prentice (1986) accounts for this correlation by including a covariance term when calculating the variance of the regression parameter estimators. Another method uses bootstrapping to calculate the variance of the case-cohort method [Wacholder et al., 1989], while others have proposed the use of a jackknife variance estimator [Barlow, 1994, Lin and Ying, 1993]. Yet another estimator is based on the variance estimator in the presence of missing data in the covariate measurements [Lin and Ying, 1993]. As previously stated, Barlow [1994] proposed the robust variance estimator, which is based on the influence of each observation on $\hat{\beta}$ and is meant to handle modifications of the case-cohort design.

Alternative sampling strategies

We have considered case-cohort designs that differ by the weights placed on subjects. In this section, we consider a modification of the D-estimator [Kulich and Lin, 2004].

Barlow [1994] proposed an estimator similar to the D-estimator. He explains the estimator using an example of breast cancer data. In these data, the outcome is death due to breast cancer, but the subcohort consists of all subjects who developed breast cancer, even if they did not die from it and were not part of the subcohort [Barlow, 1994]. In doing this, Barlow allows the opportunity to study breast cancer as the outcome of interest, which may also provide important information for their study. Following a similar notation to that of Barlow [1994], this design gives a psuedo-likelihood of the following form:

$$L_{DE} = \prod_{j=1}^{D} \frac{e^{z_j^T \beta}}{e^{z_j^T \beta} + \sum_{l \in \{R_{2,j} \backslash j\}} e^{z_l^T \beta} + \frac{n(t)}{\tilde{n}(t)} \sum_{l \in \{S_j \backslash (j \cap R_{2,j})\}} e^{x_l^T \beta}} \tag{2.24}$$

where S_j represents subjects from the subcohort who are still at risk with respect to the primary outcome (death due to breast cancer). $R_{2,j}$ represents subjects who develop the second outcome (breast cancer). We also have that $n(t)$ is the number of subjects in the full cohort who are still at risk and $\tilde{n}(t)$ is the number of subjects in the subcohort who are still at risk at time t. This differs from previous methods in that we now compare the covariate values of the case with those of subcohort members who are still at risk and those of non-subcohort members who experience the second event.

In his paper, Barlow compares the estimates obtained for both of the outcomes. One consideration when looking at time-to-event data is competing risks. With death due to breast cancer, for example, subjects may die because of complications due to breast cancer, but these may not be counted as an observed event. Comparing the results with both outcomes might therefore provide greater insight into the disease. This type of design might also be useful when more than one outcome may be of interest as it provides an opportunity to analyze these data with both outcomes.

Combined doubly weighted estimator

Oftentimes, certain covariates are recorded for all subjects regardless of their case/subcohort status. These covariates may include age, gender, etc. With the case-cohort design, however, we do not use this information if subjects are neither in the subcohort nor become a case. The doubly weighted and combined doubly weighted estimators seek to use this information by including it in the analysis. These estimators implement the case-cohort design using a two-stage approach. In the first stage, we must sample subjects to be selected as part of the clinical study cohort. The second stage requires selection of these controls as subcohort members. The weights considered for these estimators are therefore based on the idea of Kulich and Lin [2004] as well as on the fact that we can use the information that is already collected on all subjects.

The doubly weighted estimator, as the name suggests, contains two sets of weights. One set of weights, referred to as the second-level weights, is involved in calculations of the subcohort sampling probabilities. The other set, referred to as the first-level weights and which also includes the subcohort sampling probabilities, is included in the psuedo-score. The variance of the doubly weighted estimator in this setting depends on the choice of second-level weights

[Kulich and Lin, 2004]. Because the variance of the doubly weighted estimator depends on the selection of the second-level weights, Kulich and Lin (2004) proposed an efficient doubly weighted estimator, which selects the second-level weights in a way that the variance no longer depends on these. One of the disadvantages of this estimator, however, is that it does not always perform well in finite sample sizes. This led to the creation of the combined doubly weighted estimator (CDW) [Kulich and Lin, 2004].

The CDW is calculated by combining the pseudo-score functions of the efficient doubly weighted estimator and that of the time-varying weights estimator introduced by Borgan et al. [2000]. By combining the two pseudo-scores, the authors ensure that the efficiency of the estimator is not lower than that of Borgan's estimator and that the doubly weighted estimator can be calculated under finite samples [Kulich and Lin, 2004].

2.7.3 Software implementation of the case-cohort design

Implementing the case-cohort design in the survival context is similar to fitting a Cox PH model with the full data. The only difference is the way in which the data is set up. One must first randomly sample the subcohort using the software selected. In R, this can be done using the function `sample` and selecting the number of subjects to be sampled. Once the subcohort has been sampled, the new data set only contains the cases and members of the subcohort. Members of the subcohort keep their usual entry time, while cases that are not in the subcohort enter the analysis immediately before their own event time. Once the data have been set up in this way, a Cox PH model can be used to fit a model with only these data. In R, a Cox PH model is fit using `coxph` and including a `Surv` object. One may also use `cch` which allows the user to select the preferred weighting scheme. As with the nested case-control design, in the Appendix of this chapter we provide all necessary R code for implementing the case-cohort data analysis presented in Section 2.8. We also present examples of STATA and SAS code to perform the case-cohort analysis.

2.7.4 Simulated performance of the case-cohort design

In this section, we discuss simulation results to investigate the consistency and efficiency of estimates when using the case-cohort design compared to that of the full analysis. These subcohort sizes were selected so that the total number of subjects is comparable to the number of subjects in the simulation study from Section 2.6.8. We also present results to compare the efficiency of these designs with that of a simple random sample of the same size as the case-cohort designs. As with the nested case-control design simulations, we start with a scenario consisting of approximately 60% censoring. These simulations were generated in the same way as those for the nested case-control design.

Notice that under approximately 60% censoring, we obtain consistent estimates with the case-cohort design as well as with a simple random sample. We find that a simple random sample performs better than most of the designs presented with the exception of the Lin and Ying [1993] estimator, with which it performs similarly. As explained with the nested case-control design, the simple random sample performs well because there are a fairly large number of cases; when we take a random sample we can still obtain a reasonable number of cases. Notice, however, that even though the other designs are less efficient than a simple random sample and the Lin and Ying estimator, all designs perform well. The greatest loss in efficiency is observed in the case with the smallest subcohort size, $\alpha = 0.35$. In this case, the Prentice [1986] and Self and Prentice [1988] methods yield a variance almost three times larger than that of the full cohort analysis, while the Lin and Ying estimator and the simple random sample yield variances less than 2 times larger.

We now consider a scenario with approximately 85% censoring. In this case, we see that a simple random sample does not perform as well compared to the other designs. We do

Table 2.5: Simulations for the comparison of the case-cohort and the full cohort analyses with a single covariate. True survival times were simulated from a Exponential distribution with hazard rate 0.5× $\exp\{\beta Z\}$, with $Z \sim$ N(0,1) distribution. The true parameter value associated with the covariate was taken to be $\beta = \log(2.0) = 0.693$ (true HR of $e^{\beta} = 2.0$) and we have approximately 60% censoring. Censoring was obtained using a Uniform(0, 2.25) distribution. 10,000 simulations with 1000 subjects each were performed.

Sampling Design	Total Subjects	Exp(Coeff.)	Coeff	Empirical Variance	Relative Efficiency
Full Cohort	1000	2.01	0.695	0.003	1.00
($\alpha = 0.725$)					
Prentice	839.4	2.01	0.696	0.004	1.35
Self & Prentice	839.4	2.01	0.697	0.004	1.36
Lin & Ying	839.4	2.01	0.696	0.003	1.14
Random sample	839.4	2.01	0.696	0.003	1.20
($\alpha = 0.65$)					
Prentice	795.6	2.01	0.696	0.004	1.49
Self & Prentice	795.6	2.01	0.698	0.004	1.50
Lin & Ying	795.6	2.01	0.696	0.003	1.20
Random sample	795.6	2.01	0.695	0.004	1.29
($\alpha = 0.55$)					
Prentice	737.1	2.01	0.696	0.005	1.72
Self & Prentice	737.1	2.02	0.698	0.005	1.75
Lin & Ying	737.1	2.01	0.696	0.004	1.29
Random sample	737.1	2.01	0.695	0.004	1.35
($\alpha = 0.35$)					
Prentice	620.2	2.02	0.697	0.008	2.69
Self & Prentice	620.2	2.03	0.702	0.008	2.78
Lin & Ying	620.2	2.01	0.697	0.005	1.69
Random sample	620.2	2.01	0.695	0.005	1.63

notice, however, that when the subcohort size is small (such as $\alpha = 0.125$), the simple random sample performs similarly to the Self and Prentice method.

The Lin and Ying estimator still performs better than the other estimators regardless of the subcohort size. The largest difference can be seen when the subcohort size becomes very small. For example, when $\alpha = 0.125$, the variance of the Lin and Ying estimator is about three times larger than that of the full cohort analysis. The Prentice estimator has a variance approximately 3.7 times larger than that of the full cohort, and the variance for the Self and

Table 2.6: Simulations for the comparison of the case-cohort and the full cohort analyses with a single covariate. True survival times were simulated from a Exponential distribution with hazard rate 0.5× $\exp\{\beta Z\}$, with $Z \sim$ N(0,1) distribution. The true parameter value associated with the covariate was taken to be $\beta = \log(2.0) = 0.693$ (true HR of $e^{\beta} = 2.0$) and we have approximately 85% censoring. Censoring was obtained using a Uniform(0, 0.5) distribution. 10,000 simulations with 1000 subjects each were performed.

Sampling	Total			Empirical	Relative
Design	Subjects	Exp(Coeff.)	Coeff	Variance	Efficiency
Full Cohort	1000	2.01	0.695	0.008	1.00
($\alpha = 0.4$)					
Prentice	481.8	2.03	0.700	0.013	1.57
Self & Prentice	481.8	2.03	0.703	0.013	1.61
Lin & Ying	481.8	2.03	0.701	0.012	1.41
Random sample	481.8	2.03	0.699	0.018	2.17
($\alpha = 0.325$)					
Prentice	417	2.03	0.699	0.014	1.74
Self & Prentice	417	2.04	0.704	0.015	1.81
Lin & Ying	417	2.03	0.701	0.013	1.54
Random sample	417	2.03	0.699	0.020	2.48
($\alpha = 0.25$)					
Prentice	352.2	2.04	0.705	0.018	2.15
Self & Prentice	352.2	2.06	0.712	0.019	2.27
Lin & Ying	352.2	2.04	0.707	0.015	1.83
Random sample	352.2	2.03	0.698	0.024	2.92
($\alpha = 0.125$)					
Prentice	244.2	2.07	0.712	0.030	3.68
Self & Prentice	244.2	2.12	0.731	0.035	4.23
Lin & Ying	244.2	2.07	0.717	0.024	2.93
Random sample	244.2	2.05	0.702	0.036	4.41

Prentice estimator is more than four times larger than that of the full cohort.

As with the nested case-control design, we find that the case-cohort design becomes useful in the case of rare events. When events are not rare, taking a simple random sample performs similarly (if not better) than the case-cohort designs. Moreover, comparing the simulation results for the nested case-control and case-cohort designs, we find that the standard nested case-control design performs similarly to the Lin and Ying estimator for the case-cohort design. Therefore, selection of the method to be used will depend on the study under consideration.

2.8 Implementation of Sampling Designs Using Data from the Alzheimer's Disease Neuroimaging Initiative (ADNI)

Data used in the preparation of this example were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see www.adni-info.org.

The study is currently in its fourth phase, ADNI 3. The first three phases include ADNI, ADNI 2 and ADNI GO, and all phases have been made possible by volunteer participants [Weiner, 2013]. The subset of the data used for this analysis only includes participants from ADNI, ADNI 2 and ADNI GO who started with mild cognitive impairment (MCI). For the purpose of this example, we considered any participants with significant memory concern, those with early MCI, and those with late MCI as belonging to the MCI diagnostic group. It should be noted that only subjects with complete demographic and biomarker data as well as more than one recorded visit were included, which left us with a total of 359 participants.

The current section presents the results based on the analysis for the full cohort as well as those for the nested case-control and the case-cohort designs. As an illustrative example, we consider estimating the association between phosphorylated tau protein, a potential biomarker for Alzheimer's disease measured in cerebral spinal fluid, and the time to progression of AD dementia. The corresponding code for the presented analyses can be found in the Appendix.

Table 2.7: This table presents disease progression for the ADNI data. The MCI category consists of participants who were severely cognitively impaired but not MCI, as well as those with early and late MCI.

Last recorded diagnosis	*Baseline diagnosis*		
	Cog. Normal	MCI	Dementia
Cog. Normal	261	18	3
MCI	112	525	96
Dementia	0	1	233

Table 2.7 presents disease progression for all participants with full demographic and biomarker data, including those who started as cognitive controls and those with AD. In this analysis, we consider the event to be progression to AD dementia. Notice that some MCI participants became part of the cognitively normal group. The reason for this is that we included severe cognitive impairment as MCI, while these were recorded as cognitively normal in their official diagnosis. Because we are only interested in the event that MCI participants progress to AD, however, we kept the official diagnosis of ”cognitively normal” as the diagnosis at their final visit.

Table 2.8: Characteristics of ADNI study participants from ADNI1, ADNI2, and ADNI GO whose baseline diagnosis was MCI. For continuous variables, we present mean (sd), and for categorical variables we have N (%).

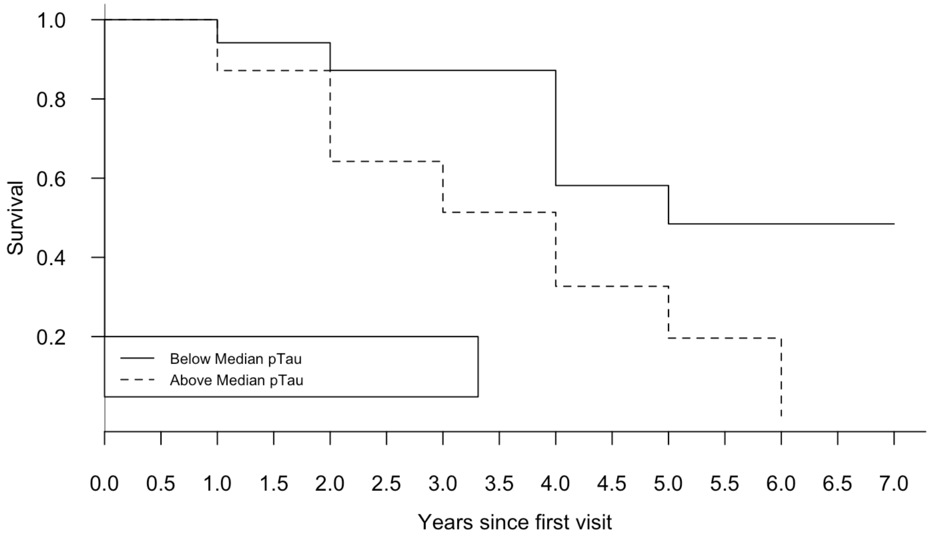
Participant Characteristics	Mean (sd) or N(%)
Race	
Caucasian	340 (94.71%)
Black	7 (1.95%)
Asian	6 (1.67%)
More than one	5 (1.39%)
Hawaiian/Other Pacific Islander	1 (0.28%)
Gender	
Male	209 (58.22%)
Female	150 (41.78%)
Education (years)	16.21 (2.76)
Age	72.73 (7.36)
APOE 4	
0 alleles	177 (49.30%)
1 allele	140 (39.00%)
2 alleles	42 (11.70%)
Median $A\beta$ at first visit (baseline pg/ml)	175.51 (52.79)
Median P-tau at first visit (baseline pg/ml)	39.16 (21.72)
Median T-tau at first visit (baseline pg/ml)	95.87 (53.21)
Progression to AD	
No	262 (72.98%)
Yes	97 (27.02%)

Table 2.8 shows the characteristics of only those subjects who were included in this study. Notice that out of the 359 participants, 97 experienced progression to AD dementia. Because many of the races only had a few subjects, we combined American Indians, Asians, Hawaiian/Pacific Islander, and subjects with more than one race into the category ”other” for this analysis. Therefore, we were left with three categories: ”Caucasian”, ”Black”, and ”other”. It should be noted that the three biomarkers were measured using various kits to account for the variation induced by the kit. The recorded measurements include individual measurements as well as the median measurement. In this table we present the median

measurements for the biomarkers amyloid beta ($A\beta$), tau protein phosphorylated at the threonine 181 (P-tau), and total tau protein (T-tau).

It has been shown that the biomarkers T-tau, P-tau, and Aβ are predictive of AD dementia in patients with MCI [Andreasen et al., 2003]. As previously noted, in this analysis we will focus on the association between the levels of P-tau and the risk of progression to AD dementia.

Figure 2.5 shows the survival curves for subjects with a P-tau level above the median (34.6) and for subjects with levels below the median. Notice that subjects with a lower P-tau level have a higher survival compared to subjects with a higher P-tau measurement, which is what we expected to see due to the association of P-tau and the development of AD dementia.



Below Median pTau	172 (0)	172 (10)	26 (20)	6 (27)	1 (27)
Above Median pTau	187 (0)	187 (24)	20 (63)	5 (69)	0 (70)
Total	359 (0)	359 (34)	46 (83)	11 (96)	1 (97)

Figure 2.5: Kaplan-Meier estimates of time to disease progression for participants with P-tau above and below the median level. Below the plot we present the total number of subjects at risk in each group at each time point, along with the the total number of events that have occurred up until that time.

Although the P-tau measurements are available for all subjects, we perform the analysis as if these measurements are not known but are measurable. Many times, for example, cerebrospinal fluid (CSF) is stored and only the samples required are analyzed. Because study participants are often unwilling to undergo lumbar punctures [Nuño et al., 2017], it would be beneficial to patients if they are not required to receive a lumbar puncture during their visit. The nested case-control and the case-cohort design are appealing in this scenario because since only about 27% of the full cohort experienced the event, we can reduce the number of participants whose full covariate information needs to be collected.

We first consider the results based on the full cohort analysis and then compare them to those obtained using the standard nested case-control and case-cohort designs. The adjustment variables, selected a priori, were age, race, gender, education, and the presence of the APOE 4 allele. These are all known to be associated with development of AD. In particular, the risk of developing AD is higher for subjects with two alleles than those with a single APOE 4 allele [Corder et al., 1993].

Table 2.9: Results for ADNI analysis using the full cohort, standard nested case-control design with three controls, and the Prentice method with $\alpha = 0.75$.

	Full Cohort		Nested case-control		Prentice	
Covariate	Est (se)	HR (95% CI)	Est (se)	HR (95% CI)	Est. (se)	HR (95% CI)
	N = 359		N = 273		N = 299	
P-tau (20 pg/ml)	0.32 (0.08)	1.38 (1.17, 1.62)	0.34 (0.09)	1.40 (1.18, 1.66)	0.30 (0.09)	1.35 (1.13, 1.60)
Age (5 years)	0.13 (0.08)	1.14 (0.98, 1.34)	0.13 (0.08)	1.13 (0.97, 1.32)	0.11 (0.09)	1.12 (0.94, 1.33)
Race						
White		1.0		1.0		1.0
Black	-0.68 (1.01)	0.51 (0.07, 3.69)	-0.40 (1.02)	0.67 (0.09, 4.98)	-0.49 (1.04)	0.61 (0.08, 4.74)
Other	-0.38 (0.72)	0.69 (0.17, 2.82)	-0.20 (0.72)	0.82 (0.20, 3.36)	-0.09 (0.75)	0.92 (0.21, 3.99)
Gender						
Male		1.0		1.0		1.0
Female	-0.29 (0.23)	0.75 (0.48, 1.16)	-0.32 (0.22)	0.73 (0.47, 1.12)	-0.31 (0.24)	0.73 (0.46, 1.17)
Education	-0.06 (0.04)	0.94 (0.87, 1.01)	-0.06 (0.04)	0.94 (0.87, 1.01)	-0.06 (0.04)	0.94 (0.86, 1.02)
APOE 4						
0 alleles		1.0		1.0		1.0
1 allele	0.28 (0.24)	1.32 (0.82, 2.12)	0.22 (0.24)	1.25 (0.78, 1.99)	0.31 (0.26)	1.37 (0.83, 2.25)
2 alleles	0.79 (0.32)	2.19 (1.18, 4.06)	0.64 (0.31)	1.90 (1.04, 3.48)	0.81 (0.35)	2.24 (1.14, 4.41)

We used three controls for the nested case-control design, giving us a total of 273 subjects for the analysis. The case-cohort design was based on $\alpha = 0.75$ which requires a total of 299 subjects for the analysis. In practice, it is difficult to obtain the same number of subjects for the nested case-control and the case-cohort designs. This is due to the random sampling of controls. In the case-cohort design, the controls are sampled ahead of time and may include subjects that later become cases. Despite having the opportunity to select the subcohort proportion, because we do not know how many cases will be sampled into the subcohort, we do not know how many subjects will be needed in total. Similarly, in the nested case-control design we sample a certain number of controls at each event time, but these controls may later become cases or may be sampled again at other event times. Once again, because we do not know how many controls are re-sampled and how many become cases in the future, we do not know exactly how many subjects will be required.

Table 2.9 presents the estimated hazard ratios based on the full cohort analysis as well as the standard nested case-control and case-cohort designs. From the full cohort, we find that comparing two populations that differ by 20 pg/ml in P-tau, the relative risk of developing AD dementia is approximately 38% (95% CI: 1.1743, 1.6150) higher for the group with the higher P-tau levels, assuming the two groups are similar with respect to the other characteristics. This is consistent with previous studies investigating the association between P-tau levels and progression to AD as well as with the survival curves presented earlier. Notice that the relative risk of developing AD dementia is approximately 30% higher for a population with one APOE 4 allele compared to a population with no APOE 4 alleles (95% CI: 0.822, 2.119) when both populations are similar with respect to the other variables. Comparing a population with two APOE 4 alleles to a population with none, we find that the relative risk of developing AD dementia is approximately two times larger for the population with both APOE 4 alleles (95% CI: 1.183, 4.062). The other estimates follow similar interpretations.

For most cases, the estimates for the nested case-control and the case-cohort designs are very similar. For example, even though the coefficient estimates associated with two APOE 4 alleles differs for the three methods, they all correspond to a hazard ratio of approximately two. The main difference is observed for the coefficients associated with race. The reason these differ so much, however, is because the ”Black” and ”other” categories each had very few subjects.

Table 2.10 presents the coefficient estimates for the full cohort and several variations of the nested case-control and the case-cohort designs. In this table, we present the estimates and standard errors along with the hazard ratios and the confidence intervals. Notice that even though the path sampling design also requires three controls per case, the total number of subjects is different than that for the other nested case-control designs. Because this design was created to reduce the number of times that each control is sampled, we naturally sample more controls than in the standard nested case-control design.

We find that all variations of the designs yield similar estimates to those of the full cohort. The control forward design appears to differ more than the other designs, while the case-cohort designs appear to yield estimates that are closest to those of the full cohort; however, these differences are almost negligible. Notice also that the standard errors for most designs are larger than those of the full cohort analysis. The smallest standard error is that of the path sampling design, which is due to the number of controls selected. Because we attempt to sample as many different controls as possible, we are nearly using the entire cohort. Notice that the standard errors are approximately the same for all three designs even though our simulation studies show that the nested case-control study is less efficient than the full cohort analysis. This example, however, represents only one random sample from the full cohort.

Although the Control Forward and Path Sampling designs were meant to increase efficiency compared to the standard nested case-control design, in their paper, Langholz and Thomas [1991] show that under most scenarios, the proposed designs are as efficient as the nested case-control design and in cases where these designs are more efficient, the increase is negligible.

Table 2.10: Estimates for the P-tau coefficient based on the ADNI analysis using the full cohort as well as variations of the nested case-control and case-cohort designs. This table presents the total number of subjects included in the analysis (N), as well as the coefficient estimates, standard error, hazard ratio, and 95% confidence interval.

Covariate	N	Est	SE	HR	(95% CI)
Full Cohort	359	0.32	0.08	1.38	(1.17, 1.62)
Std. NCC	273	0.34	0.09	1.40	(1.18, 1.66)
Control Forward	273	0.37	0.09	1.45	(1.23, 1.71)
Path Sampling	305	0.28	0.08	1.33	(1.13, 1.56)
Prentice	299	0.30	0.09	1.35	(1.13, 1.60)
Self & Prentice	299	0.29	0.09	1.33	(1.12, 1.59)
Lin & Ying	299	0.32	0.09	1.37	(1.16, 1.63)

Comparing the results from the nested case-control and the case-cohort designs, we observe that in this example, the case-cohort design provides estimates that are closer to those of the full cohort analysis. The estimates obtained from the two designs are also fairly similar. Despite some (very small) differences in the coefficient estimates, the interpretation of the results remains the same. In all designs, we find that the relative risk is approximately 40% higher comparing two populations of subjects with similar backgrounds that differ by 20 pg/ml in P-tau measurements.

In studies such as the one presented here, where covariate measurements may be difficult to obtain, the nested case-control and the case-cohort designs may be worth considering. Not only do these reduce the number of subjects whose full covariate information is required, but they also yield results similar to those of the full cohort analysis. In sections 2.10 and 2.11 , we compare and contrast the two designs and further discuss how these methods are implemented in practice.

2.9 Explicit adjustment for confounding variables using alternative sampling designs

With the nested case-control and the case-cohort designs, we can often adjust for confounding variables in the specified model. When investigators know about the existence of a possible confounding variable, they may choose to adjust for it by design. For example, if the purpose of the study is to investigate the effects of a new treatment and there are strong reasons to believe that age is a confounder, researchers may choose to randomize subjects to the new treatment and the control within certain age groups, thereby avoiding potential imbalances in age by treatment group that may have arisen by chance in the randomization process. Adjusting for confounding variables in this way proves beneficial when the functional form of the variable is not known because by matching with respect to that variable, we eliminate the need to specify a functional form. This avoids the potential for model mis-specification that can occur when adjustment is model-based.

The same idea holds when considering case-control studies. Rather than adjusting for confounding variables in the model, statisticians may decide to match on that covariate value via a matched case-control design. One should take special caution when matching on multiple variables as this could lead to very small group sizes and may leave no controls for comparison.

Variations of the nested case-control and the case-cohort designs are based on this idea of matching on confounders. This section is dedicated to discussion of these extensions, including the matched nested case-control, counter-matching, and exposure stratified case-cohort design.

2.9.1 Matching in the nested case-control design

The matched nested case-control design

In the usual nested case-control design, we have discussed that controls are randomly selected from those who are still at risk at the time of the event. In a sense, we are matching cases to controls with respect to time. The nested case-control design can also be used to adjust for confounding variables using what is known as the matched nested case-control design. This design considers subjects who are still at risk at the time of the event, and who are similar to the case with respect to some variable, usually a known confounder. When using the matched nested case-control design, we are not able to estimate the effect of the confounding variable, but we explicitly remove the potential confounding effect of that variable by forcing similarity between cases and controls [Keogh and Cox, 2014].

In the case of the matched nested case-control design, the pseudo-likelihood takes the form:

$$L_{MNCC} = \prod_{j=1}^{D} \frac{e^{z_{j,k}^T \beta}}{\sum_{l \in \tilde{R}_k(t_j)} e^{z_{l,k}^T \beta}} \tag{2.25}$$

$z_{j,k}$ tells us that the j^{th} subject is in the k^{th} strata and $\tilde{R}_k(t_j) = j \cup S_{j,k}$ where $S_{j,k}$ represents the sampled controls at time t_j. Notice that these controls are sampled from the same strata as that of the case.

The counter-matched nested case-control design

Another design extension meant to increase efficiency and which is similar to the matched nested case-control design, is known as counter-matching. Like the matched nested case-control design, counter-matching considers a certain variable when sampling the controls. Unlike the previous method, however, it seeks to draw subjects with various values for that outcome instead of restricting the sampling of controls only to those in the same strata as the case. Suppose for example that there are j strata for the variable being considered. Using the notation provided by Langholz and Clayton [1994], suppose that stratum i contains n_i subjects at risk. For the stratum in which the case is found, we will randomly sample $m_i - 1$ controls. For all other strata, we will randomly sample m_i controls. The likelihood for the counter-matched design takes the following form

$$L_{CM} = \prod_{j=1}^{D} \frac{e^{z_{j,k}^T\beta}}{e^{z_{j,k}^T\beta} + \sum_{l=1}^{m_k-1} e^{z_{l,k}^T\beta} + \sum_{g\neq k}\sum_{l=1}^{m_g} e^{z_{l,g}^T\beta}}. \tag{2.26}$$

Consider a scenario in which we have two strata (0 and 1) and are considering 1:1 matching. If the case is from stratum one, we will sample one control from stratum zero. Similarly, if the case is from stratum zero, we will randomly sample one control from stratum one. This scenario gives rise to the name ”counter-matching” [Langholz and Clayton, 1994]. One scenario in which counter-matching may prove useful is when a variable is expensive to collect, but there is an inexpensive measure that can be used as a surrogate. In this case, we use counter-matching based on the surrogate to select the controls. Once controls are selected, the ”expensive measurement” is then collected from the cases and the selected controls. This method provides large efficiency gains when the surrogate measure is moderately predictive of the expensive measure. Moreover, high specificity of the surrogate is more important than high sensitivity [Langholz and Borgan, 1995].

2.9.2 The exposure stratified case-cohort design

With the nested case-control design, we considered the matched nested case-control and the counter-matched method. The case-cohort design has a similar variation, known as the exposure stratified case-cohort design [Borgan et al., 2000]. The original case-cohort method considers a random sample from the full cohort, while the exposure stratified case-cohort groups subjects into ”exposure-related strata” when selecting the subcohort [Borgan et al., 2000]. In this design, we create L groups. From each group l, we must randomly sample m_l subjects. The subcohort is then made up of $m = \sum_{l=1}^{L} m_l$ subjects. As with the usual case-cohort design, full covariate information is collected from all of the subcohort members as well as all subjects who later become cases.

Borgan et al. [2000] consider estimators that maximize the following weighted pseudo-likelihood:

$$L_{wcc}(\beta) = \prod_{j=1}^{D} \frac{e^{z_j^T \beta} w_j(t_j)}{\sum_{k \in R_{wcc}(t_j)} e^{z_k^T \beta} w_k(t_j)}. \tag{2.27}$$

Three possible estimators, based on this pseudo-likelihood, are available for the exposure stratified case-cohort design. The first estimator considers $R_{wcc}(t_j) = S_j$, the subcohort members who are still at risk at time t_j and uses weight $w_k(t_j) = n_{s(k)}/m_{s(k)}$ where $n_{s(k)}$ is the total number of subjects in stratum $s(k)$ and $m_{s(k)}$ is the total number of subjects in stratum $s(k)$ in the subcohort.

The second estimator uses $R_{wcc}(t_j) = S_j \cup D_j$ as the members of the subcohort who are still at risk at time t_j as well as all the cases not in the subcohort who are still at risk. This estimator uses weights

$$w_k(t_j) = \begin{cases} n^0_{s(k)}/m^0_{s(k)} \text{ if k is in the subcohort but does not become a case.} \\ 1 \text{ if k becomes a case} \end{cases}$$

.

Here we have that $n^0_{s(k)}$ is the total number of non-cases in stratum $s(k)$ and $m^0_{s(k)}$ is the total number of non-cases in the subcohort that are also in stratum $s(k)$.

Notice that in the first estimator, cases outside of the subcohort are only considered at the time of their event, while in the second estimator we include cases who are still at risk (even if they are not in the original subcohort).

The third estimator uses the same weights as the first estimator. However, the difference between this estimator and the first is that if the case is not in the subcohort, that case is included in the risk set at the time of the event, but a subject from the risk set is removed from the risk set at that event time. That is, we have $R_{wcc}(t_j) = (S_j \cup j) \backslash r_j$ where r_j is a randomly selected subject. If the case is in the subcohort, the risk set remains as usual.

Because the exposure-stratified case-cohort design is a stratified version of the original case-cohort method, the asymptotic properties are similar to those of the original design. It can be shown that the estimators derived from the exposure-stratified case-cohort design are consistent for the true coefficient, and that these coefficients are asymptotically normally distributed. Moreover, the authors use a simulation study to show that the stratified case-cohort design is more efficient than the usual case-cohort design. They find that the greatest gain in efficiency is seen when there are few exposed subjects [Borgan et al., 2000].

The three variations of the exposure-stratified case-cohort design can be used with time-dependent weights as well. However, asymptotic distributions using such weights have not been derived. Furthermore, the authors discuss that more research must be conducted to determine how covariates should be stratified.

As stated with the use of counter-matching for the nested case-control design, a surrogate variable can be used to sample the subcohort. This approach may also be used with the case-cohort design [Borgan et al., 2000]. In this scenario, the strata are defined using the surrogate variable, and once the subcohort has been selected, the actual predictor of interest is collected only on the subcohort members and on those who become cases.

2.10 Nested case-control design vs. the case-cohort design

As discussed in the previous sections, the nested case-control and the case-cohort designs both reduce time for data collection, economic costs, and burden to patients. Because controls are sampled from the same sample as the cases, we also have that the cases and controls are from the same population, which is necessary for valid comparisons [Ernster, 1994]. Although these methods were designed for similar purposes, there are important differences that may make one design more appealing for a particular study. In this section, we consider both the scientific and statistical differences between the two methods.

2.10.1 Scientific Considerations

One major difference between the nested case-control and the case-cohort designs is that the nested case-control design selects controls at each event time, while the case-cohort method selects controls ahead of time. Because of this, if we are interested in multiple outcomes, the controls selected using the nested case-control for one outcome will probably not be adequate for another outcome of interest; controls selected for one outcome may not all be in the risk set at the event times for another outcome. The subcohort selected using the case-cohort design, on the other hand, may be applicable for more than one outcome because the subcohort was selected independently of event times [Wacholder, 1991]. The only difference will be the subjects who become cases and are not in the subcohort. In this case, we only need to be sure that we can obtain the complete information on these cases as well. If the analysis seeks to investigate the association of risk factors to different outcomes, no adjustments need to be made to the variance. If the goal is to compare risk factors of different diseases, however, confidence levels and intervals must be adjusted [Wacholder, 1991].

Another consideration when deciding between the nested case-control and the case-cohort methods is whether researchers are interested in continuing follow-up after the study has ended. Because the subcohort in the case-cohort design is independent of events, the same subcohort may be used after the study has concluded. As long as we have full covariate information on the subcohort and we can obtain the same information for new cases, we can conduct the analysis. For the nested case-control design, however, we must select controls at the time of the event. If it is not difficult to obtain covariate information on the new case and controls, it may not be a problem. In fact, because we would need to follow less subjects, the nested case-control design may be more favorable. If covariate information is difficult to obtain after time, however, the case-cohort design might be more appealing; since the subcohort is being followed, information on these subjects will be complete and the only remaining task is to obtain the information for new cases that are not in the subcohort [Wacholder, 1991].

Wacholder also notes that, with the case-cohort design, controls are selected more quickly. For the nested case-control design, one must wait until a case has been identified so that controls can be selected. The case-cohort design does not require waiting for events to occur, so the controls are already present. If the study must be conducted in a timely manner, the case-cohort design might be more appealing.

Because the same number of controls is selected for every event in the nested case-control design, we do not have to worry about depleting the available controls. The only scenario during which this might happen is if controls are censored before the last event(s), but that same problem would arise when conducting the full cohort and case-cohort analyses as well. With the case-cohort design, however, there may be times where many of the subcohort members become cases early on even if there are non-subcohort members who are still at risk. If this happens, there are very few, if any, subjects to compare to later cases. In this setting, the subcohort can be augmented to include more controls [Wacholder, 1991]. Although this can be fixed, this problem is not encountered using the nested case-control method.

Finally, when utilizing the case-cohort design prospectively there is also the possibility that investigators induce bias. Because researchers know ahead of time that they will need to collect full covariate information on the members of the subcohort, they might follow them more closely than non-subcohort members. If closer follow-up leads to more compliance of a proposed treatment, this may bias the results. The nested case-control design does not suffer from this potential problem because researchers do not know ahead of time who will be the cases or the controls [Langholz and Thomas, 1990].

2.10.2 Statistical Considerations

When considering the nested case-control and the case-cohort designs, one topic that comes to mind is efficiency. This problem was considered by Self and Prentice [1988] as well as by Langholz and Thomas [1990]. Although Self and Prentice (1988) did not account for repeated sampling in the nested case-control design, their results are similar to those of Langholz and Thomas [1990] in that the case-cohort design is slightly more efficient when considering a study in which all people enter at the beginning of the study and are followed until the end of the study or until they fail; that is, the only censoring that occurs is if subjects have not failed by the end of follow-up.

The difference in efficiency of the two designs can be partially explained by the use of controls in risk sets. With the nested case-control design, controls are only included in the risk set for the event time at which they were sampled. With the case-cohort design, controls belonging to the subcohort are included in all risk sets during which they are still at risk. Although controls for the nested case-control design may be included forward in time as in the Control Forward design, doing this introduces a covariance term, therefore increasing the variance [Langholz and Thomas, 1991]. We do not encounter this problem with the case-cohort design because, as explained earlier, the case-cohort design allows us to obtain the filtration of the subjects in the analysis, and we can therefore regard the observations as independent conditional upon their filtration. As shown in the simulation studies, the Lin and Ying [1993] method is strikingly more efficient than the other designs. This can be explained by the fact that the Lin and Ying method includes cases in all risk sets during which they are at risk. Although the other methods require the same number of subjects as the Lin and Ying method, they do not use non-subcohort cases outside of their event time. In this sense, the Lin and Ying method obtains more information from the cases, allowing for better efficiency.

Depending on the study under consideration, we may find that the case-cohort or the nested case-control design is more convenient. Regardless of which design is selected, however, investigators must select the control sizes carefully as these will influence results of the study.

2.11 Study Design

Throughout this chapter, we have discussed several variations of the nested case-control and the case-cohort designs. You may have noticed that, regardless of the design, there is one important decision to be made in practice. If we decide to implement the nested case-control design, we must select the number of controls that will be used for each case. Similarly, if we select the case-cohort design, we must select the size of the subcohort. The simulation studies have shown that these decisions are crucial to the efficiency of the study. This section briefly considers the selection of these control sizes. There are three possible scenarios that may provide valuable information for study planning: (1) full event-time data is available on the study cohort, (2) there is partial follow-up on the cohort, and (3) there is an external data source with a small sample size but missing covariate information and follow-up for the event.

In the first scenario, we may consider a study in which complete information is available for the full cohort (with the exception of the covariate of interest). Researchers may turn to the nested case-control and the case-cohort designs if the covariate of interest is expensive to collect, so the only missing information is the distribution of that covariate for the cases and the controls. To decide upon the size of the controls (the subcohort size or the number of controls for each case), one may make assumptions on the exposure distribution between the cases and the controls. Based on these assumptions, we may investigate the sensitivity of the results stemming from the number of controls selected through the use of simulation studies. From this, we may also consider other factors such as power to select the number of controls required for the study.

In the second scenario, we assume that we have followed a cohort for a certain amount of time, but the study has not concluded. In this case we may have complete information on all predictors with the exception of the covariate of interest. However, because the study has not concluded, we also have not seen event times for all subjects. Therefore, we must project the marginal survival to the anticipated maximal follow-up time of the study as well as make assumptions on the exposure distribution. Based on these assumptions, we may therefore perform simulation studies to observe the optimal characteristics for the study design as in the first scenario.

We encounter the third scenario when pilot data is available with follow-up for the event and covariate information on the cases and controls, but with only a small sample size. This scenario requires that we make assumptions on the exposure distribution of cases and controls. Based on this information, we may then perform simulation studies to investigate the optimal characteristics of the design with special attention to power and efficiency.

One method, designed to help with sample size/power calculation, can also be used to inform the parameter estimates and confounding variables. This method is based on a two-stage approach. During the first step, the bounds for the sample size/power calculation are established. At this time, various scenarios and potential confounders are considered for a sensitivity analysis of the potential confounders. The second stage considers "internal pilot data", or the data collected as the study progresses. Before completion of the study, one may use the data collected up to that point to update, or "refine" the estimates from stage I and to obtain estimates for the fully adjusted model as well as information on power [Haneuse et al., 2012]. With the internal pilot data, we could also perform sensitivity analyses in the usual way to investigate the number of controls or the subcohort size to be used for the fully adjusted model.

2.12 Discussion

Time-to-event studies are often time-consuming and expensive. When the disease or outcome under consideration is rare, large groups may be needed for analysis, even though most of the information being used comes from those subjects that experience the event of interest. The nested case-control and the case-cohort designs were developed for these scenarios. By considering all cases and only a fraction of the controls, both designs reduce cost to researchers and/or burden for participants.

The nested case-control design samples a specified number of controls at each event time and uses those controls only for the event at which they were sampled. Variations exist that make use of sampled controls differently, such as using them forward in time or sampling based on path sets as discussed in earlier sections. Other modifications occur with respect to how controls are sampled in order to explicitly adjust for confounding, such as in the matched nested case-control or the counter-matched design.

Alternatively, one can use the case-cohort design which samples controls ahead of time and only includes these controls and all subjects who become cases. Variations of this design consider different weighting schemes. For example, cases outside of the subcohort may receive a weight of one or zero in the denominator for the likelihood in (2.23). With the case-cohort design, we may also use the exposure stratified design which samples subjects from different strata and combines the subjects from all strata to make up the subcohort.

Although both the nested case-control and case-cohort design seek to reduce burden on subjects as well as cost and time, there are special considerations that should be made for each design. The nested case-control design, for example, selects controls at each event time while the case-cohort design selects these controls ahead of time. Because of its design, the case-cohort method also has the advantage that subjects can be followed even after the study has ended. With this, however, we also have that the case-cohort design may be subject to biased estimates if subcohort members are followed more or less rigorously than non-subcohort members.

Researchers often have limited time and funding, and it is of utmost importance to reduce patient burden when possible. The nested case-control and the case-cohort designs both assist researchers in obtaining these goals. Considering the advantages of these two methods, researchers will often find that the benefits far outweigh the disadvantages. Although neither method is universally superior to the other, researchers may find that one method is better suited for the purpose of their study.

While these designs represent excellent alternatives to traditional time-to-event cohort studies, there are many areas that can be improved. In the remaining chapters, we focus on the nested case-control design and its performance under model mis-specification and propose robust parameter estimation methods. We also investigate the performance of estimators for receiver operating characteristic (ROC) curves and the area under the curve (AUC) when a nested case-control design is used. Another problem to be studied in future work is that in the usual survival setting, we consider biomarker measurements to remain constant until the time of the next measurement. However, this is not generally the case; measurements are constantly fluctuating. Simultaneous estimation of the smooth trajectory of a biomarker over time and how it relates to the risk of an event remains an open area of research. Finally, in order to improve the efficiency of these and of existing methods when time-dependent covariates are under consideration, further work for assessing optimal control selection should be considered.

Chapter 3

On Estimation in the Nested Case-Control Design Under Non-Proportional Hazards

3.1 Introduction

Time-to-event is a common outcome in many empirical studies. Biomedical examples include modeling the time to disease progression or time to death. Typically, studies utilize a simple random sample obtained from the population of interest to estimate the time-to-event distribution, usually as a function of measured covariates. As seen in Chapter 2, however, when the outcome of interest is rare, a random sample is not the most efficient sampling method since a very small proportion of sampled subjects will experience the event. Moreover, if the covariate(s) of interest are difficult or expensive to collect, the nested case-control design may provide greater utility. The nested case-control design makes use of the fact that events provide more information than non-events when using the partial likelihood estimator [Thomas, 1977]. This sampling scheme utilizes full covariate information on all subjects who experience the event of interest (cases) and a subsample from those who do not experience the event (controls). To implement the design, at each event time we randomly sample M controls from the risk set at that time. Usually, M is selected to be between one and four, allowing for a large reduction in costs when the event of interest is rare.

The work here is motivated by research on the discovery of new biomarkers for Alzheimer's disease (AD). In this setting, potential biomarkers are generally expensive to obtain and burdensome to participants. As an example, consider phosphorylated tau (P-tau) and amyloid-β (Aβ), two proteins associated with plaques and tangles in the brain that are a hallmark of AD [Selkoe, 2001]. These proteins can be measured in the cerebrospinal fluid (CSF), which is collected via lumbar puncture. To help identify new biomarkers, we may consider investigating other proteins found in CSF. However, as previously stated, trial participants are often unwilling to undergo lumbar punctures [Nuño et al., 2017]. Therefore, when it is collected, CSF is stored and processed as needed. Applying the nested case-control design would allow for strategic use of existing CSF samples.

Use of the Cox PH model is common in the literature, even though time-varying effects often arise in clinical research. In Chapter 2, we learned that when the model is mis-specified, the estimand corresponding to the partial likelihood estimator depends on patient accrual and dropout patterns. Fortunately, several estimators have been proposed that recover an estimand that does not depend on the censoring distribution including those of Xu and O'Quigley [2000], Boyd et al. [2012] and Nguyen and Gillen [2012] when using the full cohort data. In this chapter, we show that under finite samples and a NPH covariate effect, the usual Cox estimator for the nested case-control estimates a quantity that depends on the number of controls sampled for each event time as well as on the underlying censoring distribution. In Section 3.2 we begin by considering the case of a single binary predictor of interest and build on previous work for mis-specified PH models under the full cohort (FC) or simple random sample design, extending these results to the nested case-control design setting. We then propose a new estimating equation that recovers the FC estimand, and extend it to also allow for consistent estimation of a censoring-robust estimand. The asymptotic distribution of both estimators is derived and finite sample variance estimators are provided. Finally, an extension of the proposed estimators that incorporates adjustment of confounding variables is provided at the conclusion of Section 3.2. Section 3.3 presents simulated results to illustrate the performance of the proposed estimators in finite samples and in Section 3.4 we apply the estimators to ADNI data. Section 3.5 concludes with a discussion of the scientific relevance of the methodology and avenues for future research.

3.2 Methodology

3.2.1 Model Mis-specification Under the Full Cohort

We start with a brief review of the censoring-robust estimators presented in Chapter 2.

The first estimator, proposed by Xu and O'Quigley [2000], considers the scenario in which the censoring distribution does not depend on any covariate values. In this case, the estimating equation under the partial likelihood estimator can be reweighted as

$$U_{XO}(\beta) = \sum_{i=1}^{n} U_{i,XO}(\beta) = \sum_{i=1}^{n} \int_{t=0}^{\infty} W_{i,XO}(t) \left\{ Z_i - \frac{n^{-1} \sum_{j=1}^{n} Z_j Y_j(t) \exp(Z_j \beta)}{n^{-1} \sum_{j=1}^{n} Y_j(t) \exp(Z_j \beta)} \right\} dN_i(t) \quad (3.1)$$

where $W_{i,XO}(t) = \hat{S}_{KM}(t) / \sum_{j=1}^{n} Y_j(t)$ and $\hat{S}_{KM}(t)$ is the left-continuous Kaplan-Meier estimator of the survival function [Kaplan and Meier, 1958]. The estimand of the proposed estimator no longer depends on the censoring distribution and can be interpreted as an average covariate effect.

Boyd et al. [2012] consider the conditionally independent censoring scenario in which censoring and event times are independent conditional upon the covariate value. The authors show that in this case, the partial likelihood estimator is consistent for the solution to

$$\int_0^\infty E_Z\bigg(f_T(t|Z)S_C(t|Z)\bigg[Z - \frac{E_Z\{ZS_T(t|Z)S_C(t|Z)\exp(Z\beta)\}}{E_Z\{S_T(t|Z)S_C(t|Z)\exp(Z\beta)\}}\bigg]\bigg)dt = 0, \tag{3.2}$$

which is dependent upon $S_C(t|Z)$, the covariate-specific censoring distribution. Boyd et al. [2012] propose reweighting the partial likelihood by the inverse of the covariate-specific censoring distribution where the censoring distribution depends on a single, binary covariate. The proposed estimating equation is:

$$U_{CR}(\beta) = \sum_{i=1}^{n}\int_{t=0}^{\infty} W(t|Z=z_i)\bigg\{Z_i - \frac{S_{CR}^{(1)}(\beta,t)}{S_{CR}^{(0)}(\beta,t)}\bigg\}dN_i(t) = 0 \tag{3.3}$$

where $S_{CR}^{(r)}(\beta,t) = n^{-1}\sum_{j=1}^{n} Z_j^r W(t|Z=z_j)Y_j(t)\exp(Z_j\beta)$, $W(t|Z=z_j) = \{\hat{S}_{C,KM}(t|Z=z_j)\}^{-1}$, and $\hat{S}_{C,KM}(t|Z=z_j)$ is the covariate-dependent left continuous Kaplan-Meier estimator. Reweighting the score function in this manner yields a censoring-robust estimator that is asymptotically equivalent to that proposed by Xu and O'Quigley [2000] when $S_C(t|Z) = S_C(t)$.

3.2.2 Partial Likelihood Estimator under a Nested Case-Control Design

As seen in Chapter 2, when the event of interest is rare, the nested case-control design reduces the number of subjects for whom the full covariate information is required by including all subjects who experience an event and a subsample from those who do not experience an event. At each event time, M subjects are randomly sampled from everyone who is still in the risk set at that time. Only the event and the sampled controls are included as part of the nested case-control risk set. The partial likelihood for the nested case-control design in (2.17) can be rewritten as $L_{NCC} = \prod_{i=1}^{n} \left\{ \frac{e^{z_i^T\beta}}{\sum_{j=1}^{n} \tilde{Y}_j(t_i) e^{z_j^T\beta}} \right\}^{\delta_i}$ where, as before, δ_i is an indicator for whether subject i experiences an event and $\tilde{Y}_j(t)$ is an indicator for whether subject j is in the nested case-control risk set at time t (i.e. the subject either experienced an event at time t or was sampled as a control). Written in counting process notation, the estimating equation under the nested case-control design then takes the form $U_{NCC}(\beta) = \sum_{i=1}^{n} \int_{t=0}^{\infty} \left\{ Z_i - \frac{S_{NCC}^{(1)}(\beta,t)}{S_{NCC}^{(0)}(\beta,t)} \right\} dN_i(t) = 0$ where $S_{NCC}^{(r)}(\beta,t) = n^{-1}\sum_{j=1}^{n} Z_j^r \tilde{Y}_j(t)\exp(Z_j\beta)$ and $N_i(t) = I(X_i \leq t, \delta_i = 1)$. When M controls are utilized in the nested case-control design, the cardinality of the risk set size at each event time will be $M+1$ (M controls plus the observed case) unless there are less than M potential controls at risk in the FC (in this setting, all possible controls are sampled). Recall that under a FC simple random sample, the estimating equation is given by (2.10). Thus the difference between the two estimating equations lies in the risk set size at each observed event time.

Proposition 1. *Let $f_T(t|Z)$ and $S_T(t|Z)$ denote the density and survival function for the failure times, respectively, and let $S_C(t|Z)$ denote the survival function for the censoring times. Suppose that $M/n \to a$ as $M, n \to \infty$ for some constant a, $0 < a \leq 1$. If $P(Y_i(\tau) > 0) > 0$ where τ is the maximum observed time, the partial likelihood estimator under the nested case-control design is consistent for the solution to*

$$\int_0^\infty E_Z\Bigg\{E_{Z^*|Z}\Bigg(f_T(t|Z)S_C(t|Z)\gamma(a,Z^*,t)\times\Bigg[Z-\frac{E_Z\{ZS_T(t|Z)S_C(t|Z)\exp(Z\beta)\}}{E_Z\{S_T(t|Z)S_C(t|Z)\exp(Z\beta)\}}\Bigg]\Bigg)\Bigg\}dt=0$$

where $\gamma(a,Z^*,t) = \frac{a \cdot S_T(t|Z^*)S_C(t|Z^*)}{S_T(t)S_C(t)}$ and Z^* represents the covariate values of sampled controls.

Proof. The nested case-control estimator is the solution to $U_{NCC}(\beta) = \sum_{i=1}^{n} \int_0^\infty \left\{ Z_i - \frac{S_{NCC}^{(1)}(\beta,t)}{S_{NCC}^{(0)}(\beta,t)} \right\} dN_i(t) = 0$ where $N_i(t) = I(X_i \leq t, \delta_i = 1)$ is a right continuous counting process, $S_{NCC}^{(r)}(\beta,t) = n^{-1}\sum_{j=1}^{n} Z_j^r \tilde{Y}_j(t)\exp(\beta Z_j)$ and $\tilde{Y}_j(t)$ is an indicator for whether subject j is in the nested case-control risk set at time t. Note that the expected value of $S_{NCC}^{(r)}(\beta,t)$, $r =$

0, 1 will depend on who is sampled into the nested case-control risk set. We must therefore consider the probability of experiencing an event or being sampled as a control given that the subject does not experience an event. This probability is given by:

$$E[\tilde{Y}_j(t)] = \left[\frac{M}{n(t)-1}\left\{1 - \frac{\lambda(t,Z_j)}{\sum_{k=1}^{n} Y_k(t)\lambda(t,Z_k)}\right\} + \frac{\lambda(t,Z_j)}{\sum_{k=1}^{n} Y_k(t)\lambda(t,Z_k)}\right] S_C(t|Z_j)S_T(t|Z_j), \tag{3.4}$$

where $Y_k(t)$ is an indicator for whether subject k is at risk in the FC, and $n(t)$ represents the size of the risk set in the FC at time t. If $\lim_{M,n\to\infty} M/n = a$ for some constant a where $0 < a \leq 1$, then $\lim_{M,n\to\infty} E[\tilde{Y}_j(t)] = \frac{aS_T(t|Z)S_C(t|Z)}{S_T(t)S_C(t)}$ since $\lim_{n\to\infty} \frac{\lambda(t,Z_j)}{\sum_{k=1}^{n} Y_k(t)\lambda_k(t,Z_k)} = 0$. When considering a binary covariate under the nested case-control design, events will not contribute to the score if all subjects in the risk set (case and controls) have the same covariate value. This means that contributions to the partial likelihood estimating function will depend on the values of sampled controls. We therefore consider the probability of sampling a subject with covariate value Z^* given that an event occurred. This probability is given by $\frac{MS_T(t|Z^*)S_C(t|Z^*)}{nS_T(t)S_C(t)}$. Denote $\lim_{M,n\to\infty} \frac{M\cdot S_T(t|Z^*)S_C(t|Z^*)}{nS_T(t)S_C(t)} = \frac{aS_T(t|Z^*)S_C(t|Z^*)}{S_T(t)S_C(t)}$ to be $\gamma(a, Z^*, t)$. Combining our results with the work of Struthers and Kalbfleisch [1986], we have that the nested case-control estimator is consistent for the solution to

$$\int_0^\infty E_Z\left\{E_{Z^*|Z}\left(f_T(t|Z)S_C(t|Z)\gamma(a,Z^*,t)\left[Z - \frac{E_Z\{Z\frac{aS_T(t|Z)S_C(t|Z)}{S_T(t)S_C(t)}\exp(\beta Z)\}}{E_Z\{\frac{aS_T(t|Z)S_C(t|Z)}{S_T(t)S_C(t)}\exp(\beta Z)\}}\right]\right)\right\}dt$$

$$= \int_0^\infty E_Z\left\{E_{Z^*|Z}\left(f_T(t|Z)S_C(t|Z)\gamma(a,Z^*,t)\left[Z - \frac{E_Z\{ZS_T(t|Z)S_C(t|Z)\exp(\beta Z)\}}{E_Z\{S_T(t|Z)S_C(t|Z)\exp(\beta Z)\}}\right]\right)\right\}dt = 0.$$

□

In Proposition 1, $\gamma(a, Z^*, t)$ denotes the probablity that a subject with covariate Z^* is sampled into the nested case-control risk set at failure time t. Note that if $M/n \to a$ the equation from Proposition 1 simplifies to $\int_0^\infty E_Z\left\{af_T(t|Z)S_C(t|Z)\times\left[Z - \frac{E_Z\{ZS_T(t|Z)S_C(t|Z)\exp(Z\beta)\}}{E_Z\{S_T(t|Z)S_C(t|Z)\exp(Z\beta)\}}\right]\right\}dt = 0$. In this case, the estimand for the nested case-control design is the same as that of the FC partial likelihood estimator and depends on the censoring distribution when the model is mis-specified. Under finite samples and small M (as used in practice), however, contribution to the partial score depends on the covariate values of sampled subjects. In the binary setting in particular, events for which the case and all sampled controls have the same covariate value will not contribute to the partial score and the number of times this occurs will depend on the number of controls sampled at each event time. The practical implication is that in finite samples the Cox PH model under the nested case-control sampling scheme estimates a different quantity for different values of M when the PH assumption is violated, as empirically demonstrated in Section 3.3. In this chapter, we introduce a reweighted estimating function to deal with the dependency on M and on the censoring distribution.

3.2.3 Recovering the FC Estimand: Single Binary Predictor

Consider a binary predictor of interest. In Proposition 1, the dependence of the estimand on $\gamma(a, Z^*, t)$ can be explained by the fact that under the nested case-control design, event times will not contribute to the score if all subjects in the risk set have the same covariate value. Therefore, to recover the FC estimand, we propose imputing the covariate values for subjects who were not sampled into the nested case-control design to allow for inclusion of these event times. To do this, we must estimate the number of subjects at risk with each covariate value in the FC. Let $\pi(t)$ be the proportion of subjects with covariate value $Z = 1$ at time t and let $p(t)$ be an estimator of $\pi(t)$. Multiplying $p(t)$ by the number of subjects at risk in the FC, we can obtain an estimate of the number of subjects at risk with $Z = 1$ in the FC.

Proposition 2. *Let $p(t)$ be a consistent estimator of $\pi(t)$, the true proportion of subjects with $Z = 1$ at time t, and $P(Y_i(\tau) > 0) > 0$ where τ is the maximum observed time. Further, let β_0 denote the estimand corresponding to the partial likelihood estimator based upon full* *cohort data obtained from a simple random sample. If $\hat{\beta}_{FC}$ is the solution to*

$$\tilde{U}_{FC}(\beta) = \sum_{i=1}^{n} \int_0^\infty \left\{ Z_i - \frac{n^{-1}\{\hat{n}_1(t)\exp(\beta)\}}{n^{-1}\{\hat{n}_0(t) + \hat{n}_1(t)\exp(\beta)\}} \right\} dN_i(t) = 0 \tag{3.5}$$

where $\hat{n}_0(t) = n(t) \cdot (1 - p(t))$, $\hat{n}_1(t) = n(t) \cdot p(t)$ and $n(t)$ is the number of subjects at risk in the FC at time t then $\hat{\beta}_{FC} \xrightarrow{P} \beta_0$.

The proof of Proposition 2 follows that of Proposition 3. For this result to hold, we require a consistent estimator for $\pi(t)$. Under the nested case-control design, controls are sampled at random conditional upon being at risk at that event time. As such, one can naturally borrow information from risk sets before the current event time, and use this information to impute the covariate values for subjects who were not sampled into the current nested case-control risk set. One possibility is to use a generalized additive model to smooth estimates of $\pi(t)$ at each event time. Specifically, we consider a logistic regression model of the form $\text{logit}\,(\pi(t)) = s_0 + s(t)$ where $s(t)$ is a natural cubic spline with evenly spaced knots and $\pi(t)$ is the proportion of subjects at risk at time t with covariate value $Z = 1$. At each event time, we estimate $\pi(t)$ using controls who were sampled for risk sets at or before time t. For the first two event times $p(t)$ is the proportion of subjects (including cases) in the nested case-control sample with Z = 1 in risk sets at or before the current event time.

To obtain better estimates of the proportion of subjects at risk with $Z = 1$ under the full cohort simple random sample, one may modify the nested case-control sampling scheme to allow for more controls at earlier event times. In particular, one can sample a larger number of controls at the first event time and M controls for remaining event times. At each event time, we include subjects who were previously sampled and who are still at risk in the full cohort sample.

3.2.4 Censoring-robust Estimation

In Section 3.2.3, we proposed an estimator that allows us to recover the results based on the FC. However, as stated in Section 3.2.1, under model mis-specification the estimand corresponding to the FC estimator will depend on the censoring distribution. In this section, we extend the proposed estimator to recover a censoring-robust estimand under the nested case-control design using a method analogous to that of Boyd et al. [2012].

Boyd et al. [2012] reweight the estimating function for the Cox partial likelihood estimator by the inverse of the censoring distribution to recover a censoring-robust estimand. These weights are based on estimates for the censoring distribution from the FC data. Because we do not have full covariate information on all subjects when using the nested case-control design, we cannot directly estimate the FC censoring distribution. Under the nested case-control design, however, we know the covariate values for subjects who experience an event or are sampled as controls. Therefore, we propose estimating the covariate-dependent survival for censoring as

$$\hat{S}_C(t|Z=z) = \hat{P}(C > t|Z=z) = \prod_{j:t_j \leq t} 1 - \frac{c_j(z)}{n_j(z)}, \tag{3.6}$$

where $c_j(z)$ is the number of subjects with covariate value $Z = z$ who are censored at time t_j and $n_j(z)$ is the number of subjects with $Z = z$ in the nested case-control design who are still at risk. Subjects are considered to be at risk for censoring from the time they were first included in the nested case-control sample until their observed time.

Proposition 3. *Let $p(t)$ be a consistent estimator of $\pi(t)$ and assume that $P(Y_i(\tau) > 0) > 0$, where τ is the maximum observed event time. Further, let β_{CR} denote the censoring-robust estimand based upon data obtained from a simple random sample. If $\hat{\beta}_{CR}$ is the solution to*

$$\tilde{U}_{CR}(\hat{\beta}_{CR}) = \sum_{i=1}^{n} \int_0^\infty \tilde{W}(t|Z=z_i)\left\{Z_i - \frac{n^{-1}\{\tilde{W}(t|Z=1)\hat{n}_1(t)\exp(\hat{\beta}_{CR})\}}{n^{-1}\{\tilde{W}(t|Z=0)\hat{n}_0(t) + \tilde{W}(t|Z=1)\hat{n}_1(t)\exp(\hat{\beta}_{CR})\}}\right\}dN_i(t) = 0$$

(3.7)

where $\tilde{W}(t|Z=z) = 1/\hat{S}_C(t|Z=z)$*, then* $\hat{\beta}_{CR} \xrightarrow{P} \beta_{CR}$.

Proof. To show that $\hat{\beta}_{CR} \xrightarrow{P} \beta_{CR}$ holds, we will first show that $\tilde{W}(t|Z)$ is consistent for $w(t|Z) = 1/S_C(t|Z)$. To this end, let $Y_i^*(t) = I(S_i \le t \le X_i)$ where S_i is the first time subject i entered the nested case-control design sample. Further, denote $\hat{\Lambda}_C(u) = \int_0^u \frac{J(s)d\bar{N}^*(s)}{\bar{Y}(s)}$ where $\bar{N}^*(t) = \sum_{i=1}^n N_i^*(t)$, $N_i^*(t) = I(X_i \le t, \delta_i = 0)$, $\bar{Y}(t) = \sum_{i=1}^n Y_i^*(t)$, $J(t) = I(\bar{Y}(t) > 0)$ and $d\Lambda_C^*(t) = J(t)d\Lambda_C(t)$. Then
$\frac{\hat{S}_C(t)}{S_C(t)} = 1 - \int_0^t \frac{\hat{S}_C(u-)}{S_C(u)}\{d\hat{\Lambda}_C(u) - d\Lambda_C(u)\}$(Fleming and Harrington [2011]) and

$$\begin{aligned}
\hat{S}_C(t) - S_C(t) &= -S_C(t)\int_0^t \frac{\hat{S}_C(u-)}{S_C(u)}[\{d\hat{\Lambda}_C(u) - d\Lambda_C^*(u)\} - \{d\Lambda_C(u) - d\Lambda_C^*(u)\}] \\
&= -S_C(t)\int_0^t \frac{\hat{S}_C(u-)}{S_C(u)}\left\{\frac{d\bar{N}^*(u)J(u)}{\bar{Y}(u)} - d\Lambda_C^*(u)\right\} \\
&\qquad\qquad + S_C(t)\int_0^t \frac{\hat{S}_C(u-)}{S_C(u)}\{d\Lambda_C(u) - d\Lambda_C^*(u)\} \\
&= -S_C(t)\sum_{i=1}^n \int_0^t \frac{\hat{S}_C(u-)}{S_C(u)}\left\{\frac{dN_i^*(u)J(u)}{\bar{Y}(u)} - \frac{Y_i^*(u)J(u)}{\bar{Y}(u)}d\Lambda_C(u)\right\} \\
&\qquad\qquad + S_C(t)\int_0^t \frac{\hat{S}_C(u-)}{S_C(u)}\Big\{(1 - J(u))d\Lambda_C(u)\Big\} \\
&= -S_C(t)\sum_{i=1}^n \int_0^t \frac{\hat{S}_C(u-)}{S_C(u)}\left\{\frac{J(u)}{\bar{Y}(u)}dM_i(u)\right\} \\
&\qquad\qquad + S_C(t)\int_0^t \frac{\hat{S}_C(u-)}{S_C(u)}\Big\{(1 - J(u))d\Lambda_C(u)\Big\}
\end{aligned}$$

where $dM_i(u) = dN_i^*(u) - Y_i^*(u)d\Lambda_C(u)$ and $M_i(u)$ is a martingale. This follows from the fact that controls are randomly sampled from existing risk sets. Therefore, the first term converges in probability to 0. As $n \to \infty$, $J(u) \xrightarrow{a.s.} 1$, so the second term also converges to 0. This gives us that $\hat{S}_C(t) \xrightarrow{P} S_C(t)$. Note that this proves the convergence of the marginal censoring distribution. For ease of notation, the proof of the covariate-specific censoring distribution is not shown, but the same argument holds. Therefore, $\hat{S}_C(t|Z) \xrightarrow{P} S_C(t|Z)$. Applying the continuous mapping theorem, we have that $\tilde{W}(t|Z) \xrightarrow{P} w(t|Z)$.

We now prove that our proposed estimator is consistent for the same censoring-robust estimand proposed by Boyd et al. [2012]. Define $S_{CR}^{(r)}(\beta, t) = n^{-1}\sum_{i=1}^{n} Z_i^r W(t|Z = z_i)Y_i(t)\exp(\beta Z_i)$ $(r = 0,\ 1,\ 2)$, where $Y_i(t) = I(X_i \geq t)$. Note that for a binary predictor $S_{CR}^{(0)}(\beta, t) = n^{-1}\{n_0(t)W(t|Z = 0)+n_1(t)W(t|Z = 1)\exp(\beta)\}$ and $S_{CR}^{(1)}(\beta, t) = S_{CR}^{(2)}(\beta, t) = n^{-1}\{n_1(t)W(t|Z = 1)\exp(\beta)\}$. Let $\tilde{S}_{CR}^{(0)}(\beta, t) = n^{-1}\{\hat{n}_0(t)\tilde{W}(t|Z = 0)+\hat{n}_1(t)\tilde{W}(t|Z = 1)\exp(\beta)\}$ and $\tilde{S}_{CR}^{(1)}(\beta, t) = \tilde{S}_{CR}^{(2)}(\beta, t) = n^{-1}\{\hat{n}_1(t)\tilde{W}(t|Z = 1)\exp(\beta)\}$. We define $W(t|Z = z)$ to be $\frac{1}{S_{C,KM}(t|Z=z)}$ where $S_{C,KM}(t|Z = z)$ is the Kaplan-Meier estimator for the covariate-dependent survival for censoring and $\tilde{W}(t|Z = z) = \frac{1}{\hat{S}_C(t|Z=z)}$ where $\hat{S}_C(t|Z = z)$ is defined as in (3.6). We denote the true covariate-dependent survival for censoring as $S_C(t|Z)$ and the inverse as $w(t|Z)$. Let $s_{CR}^{(r)}(\beta, t) = \lim_{n\to\infty} S_{CR}^{(r)}(\beta, t)$.

Assuming that we have selected a consistent estimator of $\pi(t)$, we have that $\hat{n}_j(t) - n_j(t) \xrightarrow{P} 0$ for $j = 0, 1$. We also have that $\hat{S}_{C,KM}(t|Z) \xrightarrow{P} S_C(t|Z)$ [Kaplan and Meier, 1958] and by continuous mapping $W(t|Z) \xrightarrow{P} w(t|Z)$. Using the fact that $\tilde{W}(t|Z)$ and $W(t|Z)$ both converge in probability to $w(t|Z)$ along with continuous mapping, we have that $\tilde{W}(t|Z) - W(t|Z) \xrightarrow{P} 0$. Applying continuous mapping again, we have that $\tilde{S}_{CR}^{(r)}(\beta, t) - S_{CR}^{(r)}(\beta, t) \xrightarrow{P} 0$. If these are both monotone and bounded, we have that $\sup_{\beta\in\mathcal{B},t\in[0,\tau]} ||\tilde{S}_{CR}^{(r)}(\beta, t) - S_{CR}^{(r)}(\beta, t)|| \xrightarrow{P} 0$. Note that $||\tilde{S}_{CR}^{(r)}(\beta, t) - s_{CR}^{(r)}(\beta, t)|| \leq ||\tilde{S}_{CR}^{(r)}(\beta, t) - S_{CR}^{(r)}(\beta, t)|| + ||S_{CR}^{(r)}(\beta, t) - s_{CR}^{(r)}(\beta, t)||$, and from Boyd et al. [2012], we have that $\sup_{\beta\in\mathcal{B},t\in[0,\tau]} ||S_{CR}^{(r)}(\beta, t) - s_{CR}^{(r)}(\beta, t)|| \xrightarrow{P} 0$. Together, these give us that $\sup_{\beta\in\mathcal{B},t\in[0,\tau]} ||\tilde{S}_{CR}^{(r)}(\beta, t) - s^{(r)}(\beta, t)|| \xrightarrow{P} 0$.

Applying the continuous mapping theorem along with our previous results, we get that $\sup_{\beta\in\mathcal{B}} ||\tilde{U}_{CR}(\beta) - U_{CR}(\beta)|| \xrightarrow{P} 0$, the estimating function corresponding to our proposed estimator and to the censoring-robust estimator of Boyd et al. [2012], respectively. Therefore, $\tilde{U}_{CR}(\beta)$ and $U_{CR}(\beta)$ converge to the same function, which, when set equal to zero, has a unique solution at β_{CR}. Therefore, $\hat{\beta}_{CR} \xrightarrow{P} \beta_{CR}$, proving Proposition 3.

To prove Proposition 2, we take $\tilde{W}(t|Z = z) = W(t|Z = z) = w(t|Z = z) = 1$. Using $\sup_{\beta\in\mathcal{B},t\in[0,\tau]} ||\tilde{S}_{CR}^{(r)}(\beta, t) - s^{(r)}(\beta, t)|| \xrightarrow{P} 0$ and following the approach of Andersen and Gill

[1982], we can show that the log partial likelihood converges to a concave function maximized at β_0, the estimand corresponding to the full cohort partial likelihood estimator. □

The finite-sample performance of the censoring-robust estimator is presented in Section 3.3.2 through the use of simulations.

3.2.5 Asymptotic Distribution and Variance Estimation

In this section we provide the asymptotic distribution of the proposed estimators along with two finite sample variance estimators.

Proposition 4. *Let $\hat{\beta}_{CR}$ denote the solution to (3.7). Suppose that $P(Y_i(\tau) > 0) > 0$, $p(t)$ is a consistent estimator of $\pi(t)$, and let β_{CR} denote the censoring-robust estimand. Then $\sqrt{n}(\hat{\beta}_{CR} - \beta_{CR}) \overset{D}{\rightarrow} N(0, A^{-1}BA^{-1})$ where $A = \lim_{n\to\infty} A_n$, $B = \lim_{n\to\infty} B_n$, with $A_n(\hat{\beta}_{CR}) = n^{-1}\sum_{i=1}^{n} \delta_i \rho(X_i)\Big(1 - \rho(X_i)\Big)$, δ_i an indicator for whether subject i experienced an event, X_i is the observed time for subject i,*

$$\rho(X_i) = \frac{\tilde{W}(X_i|Z=1)\hat{n}_1(X_i)\exp(\hat{\beta}_{CR})I_{M,1}(X_i)}{\tilde{W}(X_i|Z=0)\hat{n}_0(X_i)I_{M,0}(X_i) + \tilde{W}(X_i|Z=1)\hat{n}_1(X_i)\exp(\hat{\beta}_{CR})I_{M,1}(X_i)}, \quad (3.8)$$

and $I_{M,Z}(t)$ is an indicator for whether the original nested case-control sampling included controls from group Z at time t. Further, $B_n(\hat{\beta}_{CR}) = \sum_{j=1}^{D} \tilde{U}_j^*(\hat{\beta}_{CR})\tilde{U}_j^*(\hat{\beta}_{CR})^T$, where $t_1, t_2, \cdots, t_D$ are the unique event times, $\tilde{Y}_i(t) = 1$ if subject i is in the nested case-control *risk set at time t (otherwise, it is 0), and*

$$
\begin{aligned}
\tilde{U}_j^*(\beta) = n^{-1}\sum_{i=1}^{n}\tilde{Y}_i(t_j)\Bigg[&\delta_i\tilde{W}(t_j|Z=z_i)\bigg\{Z_i \\
&-\frac{\tilde{W}(t_j|Z=1)\hat{n}_1(t_j)\exp(\hat{\beta}_{CR})}{\tilde{W}(t_j|Z=0)\hat{n}_0(t_j)+\tilde{W}(t_j|Z=1)\hat{n}_1(t_j)\exp(\hat{\beta}_{CR})}\bigg\} \\
&-Z_i\exp(\hat{\beta}_{CR}Z_i)\frac{n(t_j)\tilde{W}(t_j|Z=z_i)}{[\sum_{i=1}^{n}\tilde{Y}_i(t_j)]\{\tilde{W}(t_j|Z=0)\hat{n}_0(t_j)+\tilde{W}(t_j|Z=1)\hat{n}_1(t_j)\exp(\hat{\beta}_{CR})\}} \\
&+\exp(\hat{\beta}_{CR}Z_i)\frac{n(t_j)\tilde{W}(t_j|Z=z_i)\tilde{W}(t_j|Z=1)\hat{n}_1(t_j)\exp(\hat{\beta}_{CR})}{[\sum_{i=1}^{n}\tilde{Y}_i(t_j)]\{\tilde{W}(t_j|Z=0)\hat{n}_0(t_j)+\tilde{W}(t_j|Z=1)\hat{n}_1(t_j)\exp(\hat{\beta}_{CR})\}^2}\Bigg].
\end{aligned}
\tag{3.9}
$$

Proof. As before, let T_i be the event time for subject i, C_i the censoring time, and let the observed time be $X_i = \min(T_i, C_i)$. $N_i(t) = I(X_i \le t, \delta_i = 1)$ is a right continuous counting process. Define $S_{CR}^{(r)}(\beta,t) = n^{-1}\sum_{i=1}^{n} Z_i^r W(t|Z=z_i)Y_i(t)\exp(\beta Z_i)$ ($r = 0,\ 1,\ 2$), where $Y_i(t) = I(X_i \ge t)$. Note that for a binary predictor $S_{CR}^{(0)}(\beta,t) = n^{-1}\{n_0(t)W(t|Z=0) + n_1(t)W(t|Z=1)\exp(\beta)\}$ and $S_{CR}^{(1)}(\beta,t) = S_{CR}^{(2)}(\beta,t) = n^{-1}\{n_1(t)W(t|Z=1)\exp(\beta)\}$. Let $\tilde{S}_{CR}^{(0)}(\beta,t) = n^{-1}\{\hat{n}_0(t)\tilde{W}(t|Z=0) + \hat{n}_1(t)\tilde{W}(t|Z=1)\exp(\beta)\}$ and $\tilde{S}_{CR}^{(1)}(\beta,t) = \tilde{S}_{CR}^{(2)}(\beta,t) = n^{-1}\{\hat{n}_1(t)\tilde{W}(t|Z=1)\exp(\beta)\}$. We define $W(t|Z=z)$ to be $\frac{1}{S_{C,KM}(t|Z=z)}$ where $S_{C,KM}(t|Z=z)$ is the Kaplan-Meier estimator for the covariate-dependent survival for censoring and $\tilde{W}(t|Z=z) = \frac{1}{\hat{S}_C(t|Z=z)}$ where $\hat{S}_C(t|Z=z)$ is defined as in (3.6). We denote the true covariate-dependent survival for censoring as $S_C(t|Z)$ and the inverse as $w(t|Z)$. Let $s_{CR}^{(r)}(\beta,t) = \lim_{n\to\infty} S_{CR}^{(r)}(\beta,t)$.

We will use Theorem 5.3 of Kalbfleisch and Prentice [2011], which implies Rebolledo's Martingale Central Limit Theorem, to derive the asymptotic distribution of our proposed estimators. This requires that there exists an open neighborhood $\mathcal{B}$ of β_{CR} and $s_{CR}^{(r)}(\beta,t)$, $r = 0,1,2$ defined on $\mathcal{B}\times[0,\tau]$ that satisfy the following: (1) $\sup_{\beta\in\mathcal{B},t\in[0,\tau]} ||\tilde{S}_{CR}^{(r)}(\beta,t) - s_{CR}^{(r)}(\beta,t)|| \xrightarrow{p} 0$ as $n \to \infty$; (2) $s_{CR}^{(0)}(\beta,t)$ is bounded away from 0 for $t \in [0,\tau]$; (3) For $r = 0,1,2$, $s_{CR}^{(r)}(\beta,t)$ is a continuous function of β uniformly in $t \in [0,\tau]$, $s_{CR}^{(1)}(\beta,t) = \frac{\partial s_{CR}^{(0)}(\beta,t)}{\partial\beta}$ and

$s_{CR}^{(2)}(\beta,t) = \frac{\partial^2 s_{CR}^{(0)}(\beta,t)}{\partial\beta^2}$; (4) $\sum(\beta,t) = \int_0^\tau \nu(\beta,u) s_{CR}^{(0)}(\beta,u)\lambda_0(u)du$ is positive definite $\forall\beta \in \mathcal{B}$; (5) Z_i is bounded $\forall t \in [0,\tau]$; (6) $\int_0^t \lambda_0(u)du < \infty$.

Assuming that $P(Y_i(\tau) > 0) > 0$ (i.e. there is positive probability that subject i is at risk over the inferential support interval) implies that conditions (2) and (6) hold. Conditions (4) and (5) are assumed. (5) together with the dominated convergence theorem ensures (3). The proof of Proposition 3 shows that condition (1) holds.

Following an analogous argument to that given in Boyd et al. [2012], it can be shown that the estimating function in (3.7) can be written as a sum over stochastic integrals of a predictable process with respect to a martingale. In our case, it is necessary to note that predictability of the nested case-control sampling holds by Goldstein and Langholz [1992]. Specifically, because controls are sampled immediately after an event through a random mechanism, the sampling processes are predictable. Moreover, the weights from our proposed estimator are also predictable because they do not use information from future event times. Thus application of Theorem 5.3 of Kalbfleisch and Prentice [2011] together with the sandwich variance of Lin and Wei [1989] and a Taylor expansion of the estimating equation about $s_{CR}^{(0)}(\beta,t)$, $s_{CR}^{(1)}(\beta,t)$, and $\lim_{n\to\infty} n^{-1}\sum_{i=1}^n \tilde{W}(t|Z = z_i)N_i(t)$ to account for uncertainty in the estimated censoring distribution and risk set sizes, implies that $\sqrt{n}(\hat{\beta}_{CR} - \beta_{CR}) \xrightarrow{D} N(0, A^{-1}BA^{-1})$ where $A = \lim_{n\to\infty} A_n(\beta_{CR})$ and $B = \lim_{n\to\infty} B_n(\beta_{CR})$, with $A_n(\beta) = n^{-1}\sum_{i=1}^n \delta_i\rho(X_i)(1-\rho(X_i))$ and $B_n(\beta) = \sum_{j=1}^D \tilde{U}_j^*(\beta)\tilde{U}_j^*(\beta)^T$. As defined in Section 3.2.5, δ_i is an indicator for whether subject i experienced an event, $\rho(X_i)$ is defined as in (3.8), and $\tilde{U}_j^*(\beta)$ is defined as in (3.9). The above argument yields the asymptotic distribution of the proposed censoring-robust estimator given in (3.7) and establishes Proposition 4. To obtain the asymptotic distribution of the full cohort estimator in (3.5), simply take $\tilde{W}(t|Z = z) = 1$. □

Note that the asymptotic results for $\hat{\beta}_{FC}$ can be obtained by setting $\tilde{W}(t|Z) = 1$. From the asymptotic results, we find that an analytic variance estimator for $\hat{\text{Var}}(\hat{\beta}_{CR})$ is given by

$$\hat{\text{Var}}(\hat{\beta}_{CR}) = n^{-1}A_n^{-1}(\hat{\beta}_{CR})B_n(\hat{\beta}_{CR})A_n^{-1}(\hat{\beta}_{CR}). \tag{3.10}$$

Another approach for estimating $\text{Var}(\hat{\beta}_{FC})$ or $\text{Var}(\hat{\beta}_{CR})$ is via a two-stage bootstrap that accounts for the nested case-control sampling scheme (see Algorithm 1). We account for the nested case-control sampling scheme by taking a bootstrap sample from the observed events. Because we consider continuous time, we add random noise to the bootstrapped events to break any ties. We then sample the controls for each event time from subjects who are at risk at that time. As before, controls are considered to be at risk from the time they were first included in the nested case-control sample until their observed time. Using the bootstrap sample, we obtain an estimate of the coefficient. We repeat this procedure B times and calculate the variance of the bootstrap estimates to obtain an estimate of $\text{Var}(\hat{\beta}_{FC})$ or $\text{Var}(\hat{\beta}_{CR})$.

3.2.6 Incorporating Adjustment Covariates

We have introduced two estimators, one to estimate the FC estimand and one to estimate the censoring-robust estimand, in the case of a single predictor. In observational studies, however, we try to adjust for potential confounding variables in our analyses. The proposed estimators can be extended for use in these scenarios by incorporating a hotdeck multiple imputation procedure [Fellegi and Holt, 1976], which we explain in detail below.

As before, our goal is to estimate the number of subjects with a given covariate structure at each event time. The estimation procedure considers the following estimating function:

$$\tilde{U}_{HD}(\beta) = \sum_{i=1}^{n} \int_{t=0}^{\infty} \tilde{W}(t|Z_1 = z_{1i}) \left\{ \vec{Z}_i - \frac{\tilde{S}_{HD}^{(1)}(\beta, t)}{\tilde{S}_{HD}^{(0)}(\beta, t)} \right\} dN_i(t) \tag{3.11}$$

Algorithm 1 Two-stage Bootstrapping

```
$B$: number of bootstrap samples
$D$: number of events
$M$: number of controls selected for each event
$d_k$, $k = 1, \cdots, D$: subjects who experience an event in the original sample
$c_{jk}$: $j^{th}$ control for event $k$, $k = 1, \cdots, D$, $j = 1, \cdots, M$
$s_k$: start time for control $c_{jk}$
$x_{jk}$: observed time for control $c_{jk}$
$t_1 < \cdots < t_D$: the ordered event times
$t_k - \Delta t$ denotes the time immediately before time $t_k$
procedure Two-stage Bootstrap
    for $b$ in $1 : B$ do
        Sample $D$ events with replacement from $\vec{d}$, call these $\vec{d^*}$ (the corresponding event times are $\vec{t^*}$)
        $\vec{d^*} \leftarrow \vec{d^*} + \text{Unif}(-a, a)$, $a$ small (remove ties)
        $s_k \leftarrow t_k - \Delta t$ (start time for control $c_{jk}$)
        for i in 1:D do
            Sample $M$ controls with replacement from subjects in $\vec{c}$ satisfying $s_k < t_i^* < x_{jk}$
        end for
        Calculate the reweighted nested case-control estimator, $\hat{\beta}_{CR}^{(b)}$ (this includes estimating the censoring distribution, if applicable)
    end for
    Variance estimate $= \frac{1}{B-1} \sum_{b=1}^{B} (\hat{\beta}^{(b)}{}_{CR} - \overline{\hat{\beta}^{(b)}}{}_{CR})^2$
end procedure
```

where $\tilde{S}_{HD}^{(r)} = n^{-1} \sum_{j=1}^{n} \frac{\tilde{W}(t|Z_1=z_{1j})\hat{n}_{z_{1j}}(t)}{\sum_{k=1}^{n} \tilde{Y}_k(t)\{z_{1k}z_{1j}+(1-z_{1k})(1-z_{1j})\}} \vec{Z}_j^r \tilde{Y}_j(t) \exp(\vec{Z}_j\beta)$, $\vec{Z}_j$ is the vector of covariates, z_{1j} is the value of the predictor of interest for subject j and $\hat{n}_{z_{1j}}(t)$ is the estimated number of subjects at risk with the same covariate value as z_{1j} at time t. To recover the FC estimator, we can set $\tilde{W}(t|Z) = 1$.

One problem that is often encountered in the nested case-control design with a binary predictor is that there may not be full covariate representation of that predictor in the risk sets. The proposed estimator allows us to estimate the number of subjects at risk in each group, allowing representation of both groups in risk sets for which the case and sampled controls all have the same covariate value. When we only have one predictor, estimating the number of subjects at risk in each group is sufficient. However, when adjusting for confounding variables we also need to impute the values for the confounders. For risk sets in which we have full covariate representation, we may simply reweight observations in the risk set as in (3.11) to represent everyone who is at risk under the FC. When we only have covariate representation for one group, we randomly sample one control from the l previous risk sets who is still at risk and has the missing covariate value. The sampled subject will have $\tilde{Y}_j(t) = 1$ and will therefore be included in the risk set. If the l closest times do not include subjects with the required covariate value, we may increase l.

3.3 Empirical performance

In this section we present simulation results to illustrate the problems associated with model mis-specification under the usual nested case-control design and to demonstrate the performance of our proposed estimators.

3.3.1 Full Cohort Estimator

We consider the performance of the standard nested case-control design and the proposed FC estimator under PH and NPH. In both scenarios, we have a total of 2,000 subjects, half with each value of the predictor. Censoring times were drawn from Exp(0.75) for subjects with $Z = 0$ and Exp(0.45) for subjects with $Z = 1$. The observed time was taken to be the smaller of the censoring and event times and observed times were truncated at time $t = 7$. In the NPH scenario, we consider two change points resulting in a hazard function of the form: $\lambda(t) = \lambda_0(t) \exp\Bigg\{ \log(0.05) \cdot Z \cdot I(t \leq 3) + \log(2) \cdot Z \cdot I(3 < t \leq 6) + \log(1) \cdot Z \cdot I(t > 6)\Bigg\}$. Under the PH scenario, we assume a hazard function given by $\lambda(t) = \lambda_0(t) \exp[\log(0.30) \cdot Z]$. In both scenarios, we have approximately 90% censoring. We sampled 60 controls at the first event time and M ($M = 1, 2, 3, 4$) at later events. The analytic variances for the reweighted estimator in Table 3.1 were calculated using (3.10) while the bootstrap variance estimates were obtained using Algorithm 1.

Table 3.1: We present results for the usual nested case-control estimator, the proposed FC estimator (Rwt. NCC), and the variance estimators based on 500 simulations. Bootstrap estimates are based on 100 bootstrap draws.

	Non-Proportional Hazards						**Proportional Hazards**					
	N	Coeff.	%	Emp.	An.	Boot	N	Coeff.	%	Emp.	An.	Boot
		Est.	Bias	Var.	Var.	Var.		Est.	Bias	Var.	Var.	Var.
FC	2000.00	-1.2815	0.00	0.0180	0.0179	–	2000.00	-1.1978	0.00	0.0263	0.0269	–
NCC												
M = 1	446.68	-1.5040	17.36	0.0810	0.0469	0.0764	378.29	-1.1986	0.07	0.0664	0.0417	0.0667
M = 2	568.56	-1.4220	10.96	0.0483	0.0281	0.0399	503.56	-1.2047	0.58	0.0456	0.0310	0.0452
M = 3	665.70	-1.4051	9.65	0.0419	0.0233	0.0310	612.54	-1.2012	0.29	0.0390	0.0285	0.0387
M = 4	749.23	-1.3788	7.60	0.0320	0.0208	0.0258	707.72	-1.1958	-0.17	0.0338	0.0276	0.0355
Rwt. NCC												
M = 1	446.68	-1.2909	0.74	0.0454	0.0589	0.0286	378.29	-1.1834	-1.20	0.0559	0.0664	0.0393
M = 2	568.56	-1.2690	-0.97	0.0320	0.0284	0.0207	503.56	-1.1950	-0.23	0.0418	0.0360	0.0340
M = 3	665.70	-1.2827	0.10	0.0283	0.0219	0.0186	612.54	-1.1935	-0.36	0.0384	0.0295	0.0319
M = 4	749.23	-1.2782	-0.25	0.0238	0.0190	0.0172	707.72	-1.1925	-0.44	0.0332	0.0271	0.0309

When the proportionality assumption does not hold (NPH in Table 3.1), we see that the coefficient estimates based on the usual nested case-control design are different than those obtained using the partial likelihood estimator with the FC. Moreover, these vary for different values of M. $M = 1$ yields a bias of approximately 17% (compared to the FC estimator) while $M = 4$ leads to a bias of approximately 7%. Applying our proposed estimator reduces the bias from 17% to less than 1% for $M = 1$ and from 7% to less than 1% for $M = 4$. In the NPH setting, the sandwich estimator is conservative for $M = 1$ but performs well for all other values of M. The bootstrap estimator performs similarly but it underestimates the variance when $M = 1$.

We consider PH to assess the robustness of the proposed estimator. Notice that in the PH setting the usual nested case-control estimator performs well, as is expected. Our proposed estimator also performs well and is more efficient than the usual nested case-control estimator. As stated in Section 3.2.3, under the usual nested case-control design, we may find that the event and all of its sampled controls have the same covariate value. If this happens, these event times will not contribute any additional information. Because the proposed estimator combines the information from previous risk sets, it allows for contribution from these cases, allowing it to be more efficient than the usual nested case-control estimator. The variance estimators again perform similarly except when $M = 1$. In this case, the sandwich estimator is slightly conservative while the bootstrap variance estimator is anti-conservative.

3.3.2 Censoring-Robust Estimator

We now consider the performance of the proposed censoring-robust estimator under the same settings as those considered in Section 3.3.1. The results in Table 3.2 are based on 500 simulations, in which the FC estimates were obtained using the partial likelihood estimator in the absence of censoring. Standardized bias was used instead of percent bias due to the small magnitude of the coefficient under NPH.

In the previous section we showed that we could reweight the nested case-control estimating equation to estimate the same quantity as the FC estimator. However, because these estimates were obtained assuming PH, the estimand will depend on the censoring distribution. In the absence of censoring under NPH, the full cohort estimates average -0.1514. Using our proposed estimator, we manage to estimate this quantity using the data available from our nested case-control sample. The sandwich estimator is slightly conservative, but the estimates obtained by our bootstrapping approach perform well regardless of the number of controls.

Under the PH setting, the usual NCC design performs well and the estimand does not depend on the censoring distribution. The proposed estimator also performs well, with a standardized bias ranging from 0.01 to 0.45. In this setting, the sandwich estimator tends

Table 3.2: Simulation results for the proposed censoring-robust estimator based on 500 simulations. Standardized bias is obtaining by dividing the bias by the standard error of the full cohort estimator in the absence of censoring. Bootstrap estimates are based on 100 bootstrap draws for each simulation.

	N	CR-NCC	Std.	Emp.	Analytic	Bootstrap
		Est.	Bias.	Var.	Est.	Est.
NPH						
FC	2000.00	-0.1514	0.00	0.0032	0.0035	–
Rwt. NCC						
M = 1	446.68	-0.1476	0.07	0.0979	0.2761	0.1198
M = 2	568.56	-0.1104	0.73	0.0786	0.1368	0.0891
M = 3	665.70	-0.1201	0.55	0.0689	0.1027	0.0811
M = 4	749.23	-0.1157	0.63	0.0653	0.0883	0.0737
PH						
FC	2000.00	-1.2002	0.00	0.0051	0.0047	–
Rwt. NCC						
M = 1	378.29	-1.1952	0.07	0.2077	0.2502	0.1476
M = 2	503.56	-1.1996	0.01	0.1745	0.1542	0.1342
M = 3	612.54	-1.1802	0.28	0.1416	0.1259	0.1202
M = 4	707.72	-1.1684	0.45	0.1270	0.1157	0.1153

to provide conservative estimates and the bootstrap estimator provides anti-conservative estimates when $M = 1$. Both estimators perform similarly for larger values of M.

3.3.3 Including Adjustment Covariates

Here we present simulation studies for the extension of the proposed estimators in which we allow for adjustment of confounding variables under the NPH setting. This setting includes 2,000 subjects with $Z_1 \sim$ Bernoulli(0.5) and $Z_2 \sim N(\mu = 2 + 2 * I(Z_1 = 1), \sigma = 1)$. Under the NPH scenario, the hazard function takes the form $\lambda(t) = \lambda_0(t) \exp\Bigg\{ \log(0.05) \cdot Z_1 \cdot I(t \leq 3) + \log(2) \cdot Z_1 \cdot I(3 < t \leq 6) + \log(1) \cdot Z_1 \cdot I(t > 6) + \log(0.85) \cdot Z_2 \Bigg\}$. Censoring times, as before, were drawn from Exp(0.75) for subjects with $Z_1 = 0$ and Exp(0.45) for subjects with $Z_1 = 1$. Observed times were truncated at $t = 7$. Generating data in this way yields approximately 90% censoring. We also considered the PH setting where the hazard function is specified as $\lambda(t) = \lambda_0(t) \exp\{\log(0.30) \cdot Z_1 + \log(0.85) \cdot Z_2\}$. We have omitted the PH

Table 3.3: 200 simulations for hotdeck imputations. We show the results for the full cohort estimator under NPH. Hotdeck imputations are performed using $l = 5$ and three imputations.

	N	$\hat{\beta}_1$	% Bias	Var($\hat{\beta}_1$)	$\hat{\beta}_2$	% Bias	Var($\hat{\beta}_2$)
FC	2000.00	-1.3170	0.00	0.0382	-0.1150	0.00	0.0051
NCC							
M = 1	427.27	-1.6829	27.78	0.1469	-0.0880	-23.48	0.0118
M = 2	549.36	-1.5459	17.38	0.0971	-0.0998	-13.28	0.0096
M = 3	650.47	-1.4922	13.30	0.0708	-0.1028	-10.67	0.0082
M = 4	734.67	-1.4806	12.43	0.0733	-0.1021	-11.21	0.0071
Rwt. NCC							
M = 1	427.27	-1.3625	3.46	0.1389	-0.0926	-19.54	0.0211
M = 2	549.36	-1.3247	0.58	0.0799	-0.1131	-1.66	0.0124
M = 3	650.47	-1.3281	0.85	0.0667	-0.1066	-7.36	0.0096
M = 4	734.67	-1.3368	1.50	0.0592	-0.1067	-7.26	0.0084

results for brevity, but results were similar to those seen in Sections 3.3.1 and 3.3.2.

Table 3.3 presents the observed coefficient estimates under the FC, the nested case-control design, and the proposed estimator for the NPH setting. In Table 3.4 we present the FC results in the absence of censoring along with the estimates based on our censoring-robust estimator and the empirical variance. As before, we present the standardized bias for the censoring-robust estimator due to the small magnitude of the coefficients in the absence of censoring. Under NPH, the usual nested case-control design estimates a different quantity as the number of controls changes (as seen earlier). The proposed estimator greatly reduces the bias for the predictor of interest and the confounding variable. Though not shown, under PH the usual nested case-control estimator and the partial likelihood estimator under the FC estimate the same quantity as was seen before.

3.4 Application to ADNI Data

We now apply the proposed estimator to ADNI data. These data were downloaded January 14, 2017. ADNI was created to assist in the advancement of AD research and allows

Table 3.4: 200 simulations for hotdeck imputations, each with three imputations. We show the results for the censoring-robust estimator under NPH. $l = 5$ hotdeck were used and each simulation consisted of three imputations. We divide the bias by the standard error of the full cohort estimator to obtain the standardized bias.

	N	$\hat{\beta}_1$	Std. Bias	Var($\hat{\beta}_1$)	$\hat{\beta}_2$	Std. Bias	Var($\hat{\beta}_2$)
FC	2000.00	-0.3323	0.00	0.0071	-0.0577	0.00	0.0008
C-R NCC							
M = 1	427.27	-0.3804	-0.57	0.4427	-0.0503	0.26	0.0683
M = 2	549.36	-0.3177	0.17	0.2449	-0.0593	-0.05	0.0371
M = 3	650.47	-0.3144	0.21	0.2393	-0.0654	-0.27	0.0356
M = 4	734.67	-0.3057	0.32	0.1883	-0.0627	-0.17	0.0258

investigators to access data from multiple sites.

The goal of this analysis is to associate APOE 4 to the risk of progression. Previous research has shown that APOE 4 is associated with an increased risk of AD [Saunders et al., 1993]. Because the APOE 4 allele is also associated with cardiovascular disease [Liu et al., 2013], we might expect that the association between this gene and progression to AD changes with time due to subjects developing cardiovascular disease. However, we do not have enough information to accurately model how this association changes with time. A common approach in scenarios like this is to fit a Cox PH model.

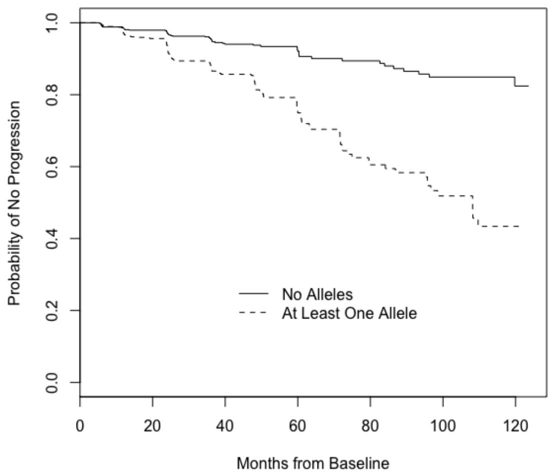
In this analysis, we define progression in terms of thresholds for the Clinical Dementia Rating Scale Sum of Boxes (CDR-SB) scores proposed by O'Bryant et al. [2008]. These thresholds were proposed to map CDR-SB scores to CD-global scores, which are often used in the staging of AD [Hughes et al., 1982]. In this study, a progressor is defined as someone with a CD-Global score of at least two after baseline, where a score of two indicates moderate dementia [O'Bryant et al., 2008]. Subjects with only one available CD-Global score are excluded from the analysis along with subjects who were censored before the first event time. Therefore, of the 1,737 subjects available in our data set, 1,643 were included in the analysis. From these subjects, 183 (11.1%) experienced an event. Time to progression was defined to be the first time (months from baseline) at which the CD-Global score increased

Table 3.5: Demographics of participants who were included in our analysis.

	0 APOE 4 alleles	>= 1 APOE 4 alleles
N	866	777
Age, mean (sd)	74.52 (7.16)	72.94 (6.95)
White, n (%)	801 (92.49)	722 (92.92)
Education, mean (sd)	16.11 (2.81)	15.71 (2.87)
Male, n (%)	477 (55.08)	434 (55.86)
CDR-SB (bl), mean (sd)	1.20 (1.60)	2.05 (1.82)
MMSE (bl), mean (sd)	27.81 (2.34)	26.58 (2.77)

to at least a score of 2.

Table 3.5 shows the demographics of study participants stratified by APOE 4 status. Participants in both groups were similar with respect to age, education, gender, race, and Mini-Mental State Exam (MMSE) scores at baseline. Subjects with at least one APOE 4 allele had higher scores in the CDR-SB at baseline. Figure 3.1 shows a Kaplan-Meier plot of the probability of no progression by APOE 4 status. This plot shows that the risk of progression is higher for subjects with at least one APOE 4 allele. As previously stated, we have reason to believe that the effect of APOE 4 on progression may change with time. Therefore, we conducted a Schoenfeld residuals test [Schoenfeld, 1982], which indicates that there is statistically significant evidence (p-value = 0.024) that APOE 4 status violates the PH assumption.



Months	0	24	48	72	96	120
No APOE ε4 allele	866(0)	660(22)	329(40)	142(48)	108(54)	25(56)
One or Two APOE ε4 alleles	777(0)	505(42)	221(82)	78(110)	43(121)	6(127)

Figure 3.1: Kaplan-Meier plot of probability of no progression by APOE 4 status. At each of the specified times, we include the number of subjects at risk (cumulative number of events). The Schoenfeld residuals test gives a p-value of 0.024 for APOE 4.

We fit a Cox PH model to all the available data using the Breslow approximation to account for ties [Breslow et al., 1974]. The censoring-robust estimates for the FC sample were obtained using the estimating equation proposed by Boyd et al. [2012]. To assess the performance of our estimator, we took 200 nested case-control samples from the FC for each value of M ($M = 1, 2, 3, 4$) and coefficient estimates were calculated assuming that complete data were only available for subjects in the nested case-control samples. At the first event time, we drew 60 controls regardless of the value of M. Using each sample, we estimated the log hazard ratio for APOE 4 status, education, and age using the usual nested case-control estimator as well as the proposed estimators. The results presented include the average estimates for the coefficients from the 200 samples as well as for the variance.

Table 3.6 presents the coefficient estimates for APOE 4. The FC model indicates that for subjects similar in age and education, the risk of progression is approximately 3.5 times

Table 3.6: Mean coefficient estimates for APOE 4 allele status along with the hazard ratio (HR). These results are based on 200 nested case-control draws. Estimates of the variance are also provided.

		Full Cohort				**Censoring-Robust**			
	N	APOE 4 (HR)	% Bias	Analytic Var	Boot. Var	APOE 4 (HR)	% Bias	Analytic Var.	Boot. Var.
Full Cohort	1643	1.2599 (3.5)	0.00	0.0271	-	1.4691 (4.4)	0.00	0.0562	-
NCC									
M = 1	379.23	1.0046 (2.7)	-20.26	0.0266	0.0663	-	-	-	-
M = 2	497.71	1.1328 (3.1)	-10.09	0.0262	0.0457	-	-	-	-
M = 3	597.20	1.1683 (3.2)	-7.27	0.0262	0.0387	-	-	-	-
M = 4	680.81	1.1943 (3.3)	-5.21	0.0261	0.0359	-	-	-	-
Rwt. NCC									
M = 1	379.23	1.2306 (3.4)	-2.33	0.0473	0.0465	1.4715 (4.4)	0.16	0.1432	0.1087
M = 2	497.71	1.2546 (3.5)	-0.42	0.0321	0.0379	1.4866 (4.4)	1.19	0.0855	0.0832
M = 3	597.20	1.2558 (3.5)	-0.33	0.0280	0.0343	1.4690 (4.4)	-0.01	0.0697	0.0749
M = 4	680.81	1.2636 (3.5)	0.29	0.0266	0.0326	1.4818 (4.4)	0.86	0.0652	0.0730

higher for subjects with at least one APOE 4 allele compared to subjects with none.

As seen in previous results, the usual nested case-control estimates differed for each value of M. The nested case-control estimates were approximately 20% lower than the FC estimate when $M = 1$. When $M = 4$, the coefficient estimates under the nested case-control design were, on average, approximately 5% lower than that of the FC. While we observed non-proportional hazards (as seen in Figure 3.1), the current example did not present an extreme case of non-proportionality. Even then, we observed considerable differences in the coefficient estimates for different values of M under the usual nested case-control design. However, we found that the estimates obtained using the proposed estimator recovered the results from the FC regardless of M. When $M = 1$, for example, we found that the mean estimate obtained from our proposed FC estimator was only 2% smaller than the FC estimate.

When we applied the censoring-robust estimator proposed by Boyd et al. [2012] to the full cohort data, we estimated that the risk of progression was approximately 4.4 times higher for subjects with at least one APOE 4 allele. The proposed censoring-robust estimator performed well for all values of M with the largest difference occurring at $M = 1$. In this case, the average estimate obtained from our proposed censoring-robust estimator was approximately unbiased. Table 3.6 also presents estimates of the variance using the two methods proposed in Section 3.2.5. We found that the bootstrap estimates and the analytic variance estimates were similar regardless of the number of controls. In general, although the same sample was utilized, the coefficient estimates provided by the proposed estimators were closer to the full cohort estimates compared to those obtained using the Cox partial likelihood estimator with the nested case-control sampling scheme, therefore supporting the use of the proposed estimator.

3.5 Discussion

Under NPH, it has been shown that the usual partial likelihood estimator is consistent for a quantity that depends on the censoring distribution. This makes it difficult to replicate results across studies simply because of possible differences in drop out and accrual patterns. We investigated this scenario when using the nested case-control design and found that in this case the finite sample estimates depend on M, the number of controls selected at each event time. Not only do we have the problem of replicability, but we also run into a problem when it comes to reproducing the results using the same sample if different values of M are used.

In this chapter we address both of the problems encountered with the nested case-control design. We propose an estimator that recovers the FC estimates as well as one whose estimand does not depend on the censoring distribution. When we have NPH, we show that these estimators recover the desired quantities. If we indeed have PH, the FC estimator not only remains consistent but is more efficient than the usual nested case-control design because it borrows information from past risk sets. Our proposed estimators can be extended to account for confounding variables using a hotdeck imputation approach. While other imputation methods can be used, we use a hotdeck imputation approach so that estimation remains feasible when there are a large number of covariates.

We proposed two variance estimation methods for the proposed estimators. As seen in Sections 3.3.1 and 3.3.2, the performance of the estimators improves as M increases. However, we found that in some cases the analytic variance estimator provides conservative estimates of the variance. Therefore, we recommend using the bootstrap approach when possible.

The nested case-control design can provide great reduction in costs when the event of interest is rare because it only requires covariate values for subjects included in the nested case-control sample. The proposed estimators ensure robustness of the results in the presence of NPH and a binary predictor of interest, and provide the added benefit that they perform well even when we truly have PH. Therefore, we recommend the use of these estimators as they are robust to model mis-specification while retaining the cost reductions afforded by the nested case-control sampling scheme. While we have focused on the nested case-control design with a binary predictor of interest, similar problems can arise under the nested case-control design when the functional form of a continuous variable is mis-specified. We discuss this in more detail in the following chapter.

Chapter 4

Robust Estimation in the Nested Case-Control Design Under a Mis-specified Covariate Functional Form

4.1 Introduction

In Chapter 3, we showed that when the proportionality assumption does not hold under the nested case-control design and a binary predictor is of interest, the expectation of the sampling distribution of the usual nested case-control estimator will depend on M, the number of controls sampled at each event time [Nuño and Gillen, 2019].

If the PH assumption holds under a correctly specified PH model, mis-specification of the functional form of a covariate in the model will induce non-proportional hazards. Therefore, based upon the results provided in Nuño and Gillen [2019], it is natural to hypothesize that if the functional form of a covariate is mis-specified in the usual nested case-control design then the expectation of the sampling distribution of the usual nested case-control estimator will depend on the parameters of the design (i.e. the number of controls sampled at each event time). It is important to note, however, that this dependence arises differently than that explored in Nuño and Gillen [2019] which considered the time-varying effect of a discrete covariate as opposed to an induced dependence on the sampling design parameters via a mis-specified model.

Returning to our motivating AD research, note that AD trials require participants to undergo various tests to help detect progression of the disease. One such examination is the Alzheimer's Disease Assessment Scale 11 (ADAS-11), which was created to evaluate cognitive and behavioral function [Rosen et al., 1984], both of which are compromised by AD. If we are interested in investigating the association of ADAS-11 and time to progression to AD, one would have to *a priori* specify the functional form of baseline ADAS-11 in order to avoid dependence of the estimand on the sampling design. Failure to do so could reasonably lead to lack of scientific reproducibility and replicability. The functional form of a continuous covariate is not obvious, however, and since interest lies in conducting inference rather than data-driven modeling we might decide to fit a first-order linear trend to relate ADAS-11 to the log-hazard for time to dementia progression. As we will see later in this chapter, the observed relationship is indeed not linear in nature. If interest lies in the association, changing the *a priori* specified model in a post-hoc fashion to fit the observed data will inflate the Type I error rate. It is therefore necessary to understand precisely how model mis-specification impacts the resulting estimator in this case and to correct, if possible, any deleterious impacts of the mis-specification.

In this chapter, we show the dependence of the estimand on the sampling proportion (which in finite samples leads to a dependence on M). We propose an estimator that recovers the estimand corresponding to the full cohort partial likelihood estimator, and in Section 4.2.3 we present the asymptotic distribution of the proposed estimator and finite-sample estimators for the variance. We include simulation studies for the proposed estimator and end with an application of the proposed estimator to data stemming from the ADNI study to investigate the association between the ADAS-11 and time to progression of AD dementia.

4.2 Methodology

4.2.1 Partial Likelihood Estimator under the usual Nested Case-Control Design

Recall that under the nested case-control design, M subjects are randomly sampled from everyone who is still at risk at each event time. In Chapter 3, we saw that if the nested case-control sampling scheme is utilized and the proportionality assumption is not satisfied, the PL estimator is consistent for the solution to

$$\int_0^\infty E_Z\left\{E_{Z^*|Z}\left(f_T(t|Z)S_c(t|Z)\gamma(a,Z^*,t)\times\left[Z-\frac{E_Z\{ZS_T(t|Z)S_C(t|Z)\exp(Z\beta)\}}{E_Z\{S_T(t|Z)S_C(t|Z)\exp(Z\beta)\}}\right]\right)\right\}dt=0 \tag{4.1}$$

where $f_T(t|Z)$ and $S_T(t|Z)$ denote the density and survival function for the failure times, respectively, and $S_C(t|Z)$ denotes the survival function for the censoring times. Moreover, $\gamma(a,Z^*,t)=\frac{a\cdot S_T(t|Z^*)S_C(t|Z^*)}{S_T(t)S_C(t)}$ represents the probability of sampling a control with covariate value Z^* if an event is observed at time t with $\lim_{M,n\to\infty} M/n=a$. As described in our previous work, in finite samples the estimates obtained using the nested case-control design will depend on M, the number of controls sampled at each event time. Under the nested case-control design, we alter the risks sets compared to those of the FC and, as a result, we also change the observed censoring distribution. As seen in Proposition 1, the censoring distribution determines the weight given to each event time and therefore influences the estimates. Note that the censoring distribution (and in turn the weighting scheme) will differ for different values of M. When the PH assumption is satisfied, the weighting scheme will not impact the estimates because the relative hazards do not vary with time. When the functional form of a predictor is mis-specified, however, we no longer satisfy the proportionality assumptions and the weights given to each event time will effect the estimates.

4.2.2 Recovering the FC Estimand for a Single Continuous Variable with Mis-specified Functional Form

As described in the previous section, when the PH model is mis-specified the expectation of the sampling distribution of the usual nested case-control estimator will depend on the number of controls sampled at each event time. This result is due to the changing censoring distribution, and hence, potentially changing covariate distributions of subjects included in the risk sets of the nested case-control design relative to the FC analysis. In order to mimic the risk-set representation of the FC, we propose imputing the values of subjects in the FC risk sets who were not included in the nested case-control sample. Because controls are randomly sampled at each event time, we can use information from previous risk sets to learn about subjects who are still at risk in the FC. Under the nested case-control design we have full covariate information for subjects sampled into the nested case-control sample. We also know the at-risk status for all subjects in the FC at each event time. We can therefore use this information to impute the covariate values for subjects in the FC who were not sampled. Using the new risk sets with the imputed values, we can obtain estimates of the coefficients.

Proposition 5. *Let $\tilde{R}_I(t)$ be the risk set including the imputed values at time t and assume that the values of covariates in $\tilde{R}_I(t)$, $\tilde{Z}_j$, are sampled from the same distribution as those* *in $R(t)$, the FC risk set at time t. Denote β_0 to be the estimand corresponding to the FC PL estimator and let $\hat{\beta}_I$ be the solution to*

$$\tilde{U}_I(\hat{\beta}_I) = \sum_{i=1}^{n} \int_0^\infty \left\{ Z_i - \frac{n^{-1} \sum_{j \in \tilde{R}_I(t)} \tilde{Z}_j \exp(\hat{\beta}_I \tilde{Z}_j)}{n^{-1} \sum_{j \in \tilde{R}_I(t)} \exp(\hat{\beta}_I \tilde{Z}_j)} \right\} dN_i(t) = 0. \tag{4.2}$$

Then $\hat{\beta}_I \xrightarrow{P} \beta_0$.

Proof: Let T_i, C_i, and $X_i = \min(T_i, C_i)$ be the event, censoring, and observed times for subject i, respectively. $N_i(t) = I(X_i \leq t, \delta_i = 1)$ is a right-continuous counting process. Define $S^{(r)}(\beta, t) = n^{-1} \sum_{i=1}^{n} Z_i^r \exp(\beta Z_i) Y_i(t)$ $(r = 0, 1, 2)$ where $Y_i(t) = I(X_i \geq t)$. Let $s^{(r)}(\beta, t) = \lim_{n\to\infty} S^{(r)}(\beta, t)$ and $\tilde{S}_I^{(r)}(\beta, t) = n^{-1} \sum_{i=1}^{n} \tilde{Z}_i^r \exp(\beta \tilde{Z}_i) Y_i(t)$ where $\tilde{Z}_i = Z_i$ if subject i was originally sampled into the nested case-control sample and the imputed value otherwise.

To establish consistency, we can show that $\sup_{\beta \in \mathcal{B}, t \in [0,\tau]} ||\tilde{S}_I^{(r)}(\beta, t) - s^{(r)}(\beta, t)|| \xrightarrow{P} 0$. We have that $s^{(r)}(\beta, t) = \mathrm{E}[S^{(r)}(\beta, t)]$ and that $||S^{(r)}(\beta, t) - s^{(r)}(\beta, t)|| \xrightarrow{P} 0$ by the strong law of large numbers. Now, suppose that $Z \sim f_Z$ and $\tilde{Z} \sim f_Z$. This gives us that $s^{(r)}(\beta, t) = \mathrm{E}[\tilde{S}_I^{(r)}(\beta, t)]$ and that $||\tilde{S}_I^{(r)}(\beta, t) - s^{(r)}(\beta, t)|| \xrightarrow{P} 0$ by the strong law of large numbers.

The proof of consistency follows from the work of Andersen and Gill [1982]. Using the fact that $\sup_{\beta \in \mathcal{B}, t \in [0,\tau]} ||\tilde{S}^{(r)}(\beta, t) - s^{(r)}(\beta, t)|| \xrightarrow{P} 0$ and under the assumption that $\tilde{Z}$ and Z are drawn from the same distribution, it is easy to show that the log partial likelihood of the proposed estimator converges in probability to a concave function maximized at β_0.

□

Note that $\tilde{R}_I(t)$ represents the same subjects as $R(t)$. The notation is used to emphasize that covariate values for subjects not in the nested case-control sample were imputed. As seen in the proof of Proposition 5, for the result to hold we need the imputed values to be drawn from the same distribution as those in the FC risk set at time t. While the covariate values can be imputed in several ways, one way to do so is via Algorithm 2. In this setting we estimate $\mu(t)$, the mean covariate value, for subjects in the risk set at each event time and calculate the mean squared error, $\sigma^2_{MSE}(t)$. To obtain $\hat{\mu}(t)$, we start by calculating the sample mean for the first event times (five in our example). The number selected here can differ and depends on the number of event times required to fit the natural spline. When fitting the natural spline, we include subjects sampled for previous event times only for the time at which they were sampled. Once we obtain $\hat{\mu}(t)$, we impute covariate values for subjects not in the nested case-control risk set (but who are still at risk in the full cohort) by randomly drawing values for the predictor of interest from a $\mathrm{N}(\hat{\mu}(t), \sigma^2_{MSE}(t))$ distribution where $\hat{\mu}(t)$ represents the estimated mean covariate value at time t.

Algorithm 2 Imputation approach for the univariate setting with a continuous predictor

```
Imputation Approach for the Univariate Setting
$D$: number of events
$t_j, j = 1, \cdots, D$: ordered event times
$R_j$: risk set at time $t_j$ under the FC
$\tilde{R}_{I,j}$: risk set at time $t_j$ including the imputed values
$M$: number of controls sampled at each event time
$s_0$: is the intercept
$s(t)$: a natural spline with evenly spaced knots
$\mu(t)$: the mean covariate value at time $t$
$z_{kc}$: predictor of interest for subject $k$ sampled at time $t_c$
procedure IMPUTATION OF THE PREDICTOR OF INTEREST
    for $j$ in $1 : D$ do
        if $j \leq 5$ (Note: 5 was selected to allow enough time points to fit the natural spline). then
            Calculate $\hat{\mu}(t_j) = \bar{z} = \frac{1}{\sum_{c=1}^{j} \sum_{k=1}^{n} \tilde{Y}_k(t_c)} \sum_{c=1}^{j} \sum_{k=1}^{n} z_{kc} \tilde{Y}(t_c)$
        else
            Fit $\hat{\mu}(t) = s_0 + s(t)$ using subjects sampled for all $t_k \leq t_j$ to obtain $\hat{\mu}(t_j)$
            $\sigma^2_{MSE}(t_j) = \frac{1}{\sum_{k=1}^{j} |\tilde{R}_I(t_k)|} \sum_{k=1}^{j} \sum_{i=1}^{|\tilde{R}_I(t_k)|} (\hat{\mu}(t_k) - z_{ik})^2$
            Sample $|R(t_k)| - \sum_{i=1}^{n} \tilde{Y}_i(t_j)$ values from $\mathrm{N}(\hat{\mu}(t_j), \sigma^2_{MSE}(t_j))$. These values, together with the original nested case-control controls, make up $\tilde{R}_I(t_j)$.
        end if
    end for
    Fit a Cox proportional hazards model using the imputed values.
end procedure
```

If the sample size is small, we can increase the number of controls sampled at the first event time to obtain better estimates of the mean covariate value at each time while not grossly impacting the overall efficiency of the nested case-control design. Moreover, all controls from previous risk sets can be used to estimate the means at each event time as long as those controls are still at risk.

Previous work also relies on imputation of subjects not sampled into the nested case-control risk sets [Nuño and Gillen, 2019]. However, in the binary case, the imputed value can only take on two values so estimating the number of subjects in each group is sufficient. When a continuous covariate is of interest, we must account for the variability in the covariate values (as is proposed in Algorithm 2) so that the imputed values are representative of the full cohort risk set.

4.2.3 Asymptotic Properties and Estimation of the Variance

In this section we provide the asymptotic properties of the proposed estimator and introduce a finite-sample variance estimator.

Proposition 6. *Let $\hat{\beta}_I$ denote the solution to (4.2). Suppose that $P(Y_i(\tau) > 0) > 0$ and let β_0 denote the estimand corresponding to the FC PL estimator. If values in $\tilde{R}_I(t_j)$ are drawn from the same conditional distribution as those in $R(t_j)$ then $\sqrt{n}(\hat{\beta}_I - \beta_0) \xrightarrow{D} N(0, A_I^{-1} B_I A_I^{-1})$ where $A_I = \lim_{n\to\infty} A_{I,n}$, $B_I = \lim_{n\to\infty} B_{I,n}$, with $A_{I,n}(\hat{\beta}) = n^{-1}\sum_{i=1}^n \delta_i \rho_I(X_i)\Big(1 - \rho_I(X_i)\Big)$, δ_i an indicator for whether subject i experienced an event, X_i the observed time for subject i, $\rho_I(X_i) = \frac{n^{-1}\sum_{j=1}^n \tilde{Y}_j(X_i) Z_j \exp(Z_j\beta_0)}{n^{-1}\sum_{j=1}^n \tilde{Y}_j(X_i)\exp(Z_j\beta_0)}$ and $\tilde{Y}_j(X_i)$ an indicator for whether subject j was originally sampled to be in the nested case-control risk set at time X_i. Further, $B_{I,n}(\beta_0) = \sum_{j=1}^D \tilde{U}_j^I(\beta_0)\tilde{U}_j^I(\beta_0)^T$, where $t_1, t_2, \cdots, t_D$ are the unique event times and*

$$\tilde{U}_j^I(\beta_0) = n^{-1}\sum_{i=1}^n \Bigg[\delta_i\Bigg\{Z_i - \frac{\sum_{k\in\tilde{R}_I(t_j)} \tilde{Z}_k \exp(\beta_0\tilde{Z}_k)}{\sum_{k\in\tilde{R}_I(t_j)} \exp(\beta_0\tilde{Z}_k)}\Bigg\} - \frac{Z_i\exp(\beta_0 Z_i)}{\sum_{k\in\tilde{R}_I(t_j)} \tilde{Z}_k \exp(\beta_0\tilde{Z}_k)} + \exp(\beta_0 Z_i)\frac{\sum_{k\in\tilde{R}_I(t_j)} \tilde{Z}_k\exp(\beta_0\tilde{Z}_k)}{\{\sum_{k\in\tilde{R}_I(t_j)}\exp(\beta_0\tilde{Z}_k)\}^2}\Bigg]. \quad (4.3)$$

Proof: Let T_i, C_i, and $X_i = \min(T_i, C_i)$ be the event, censoring, and observed times for subject i, respectively. $N_i(t) = I(X_i \leq t, \delta_i = 1)$ is a right-continuous counting process. Define $S^{(r)}(\beta, t) = n^{-1}\sum_{i=1}^n Z_i^r \exp(\beta Z_i) Y_i(t)$ $(r = 0, 1, 2)$ where $Y_i(t) = I(X_i \geq t)$. Let $s^{(r)}(\beta, t) = \lim_{n\to\infty} S^{(r)}(\beta, t)$ and $\tilde{S}_I^{(r)}(\beta, t) = n^{-1}\sum_{i=1}^n \tilde{Z}_i^r \exp(\beta\tilde{Z}_i) Y_i(t)$ where $\tilde{Z}_i = Z_i$ if subject i was originally sampled into the nested case-control sample and the imputed value otherwise. For subjects not in the original nested case-control sample, $\tilde{Z}_i \sim N(\hat{\mu}(t), \sigma_{MSE}^2(t))$ where $\hat{\mu}(t)$

is an estimate of $E[ZY(t)]$. We prove the asymptotic properties of the proposed estimator using Theorem 5.3 of (Kalbfleisch and Prentice [2011]), which implies Rebodello's theorem. This requires that there exists an open neighborhood $\mathcal{B}$ of β_0 and $s^{(r)}(\beta,t)$, $r=0,1,2$ defined on $B\times[0,\tau]$ that satisfy the following: (1) $\sup_{\beta\in\mathcal{B},t\in[0,\tau]}||\tilde{S}_I^{(r)}(\beta,t)-s^{(r)}(\beta,t)||\xrightarrow{P}0$; (2) $s^{(0)}(\beta),t)$ is bounded away from 0 for $t\in[0,\tau]$; (3) For $r=0,1,2$, $s^{(r)}(\beta,t)$ is a continuous function of β uniformly in $t\in[0,\tau]$, $s^{(1)}(\beta,t)=\frac{\partial s^{(0)}(\beta,t)}{\partial\beta}$ and $s^{(2)}(\beta,t)=\frac{\partial^2 s^{(0)}(\beta,t)}{\partial\beta^2}$; (4) $\Sigma(\beta,t)=\int_0^\tau\nu(\beta_0,u)s^{(0)}(\beta,u)\lambda_0(u)du$ is positive definite $\forall\beta\in\mathcal{B}$; (5) Z_i is bounded $\forall t\in[0,\tau]$; (6) $\lambda_0(u)du<\infty$. As in Eriksson et al. [2019], our results require that the imputed values are drawn from the same conditional distribution as the covariates for subjects in the full cohort and that missing values are missing at random. The latter is satisfied by design.

We assume that $P(Y_i(\tau)>0)>0$ (i.e. there is positive probability that subject i is at risk over the inferential support interval) which implies that conditions (2) and (6) hold. We also assume that conditions (4) and (5) hold. Condition (5) along with the dominated convergence theorem ensures that (3) is also satisfied. We have already shown that condition (1) holds when proving Proposition 5.

(4.2) is a sum over stochastic integrals of a predictable process with respect to a martingale and the predictability of the nested case-control sampling scheme holds by Goldstein and Langholz [1992]. Notice that at each event time the proposed estimator only considers controls that were sampled into risk sets up to the current time so the proposed estimator maintains predictability. Therefore, Theorem 5.3 of Kalbfleisch and Prentice [2011] with the sandwich variance estimator of Lin and Wei [1989] and a Taylor expansion of the estimating function about $s^{(0)}(\beta,t),s^{(1)}(\beta,t)$ and $\lim_{n\to\infty}n^{-1}\sum_{i=1}^n N_i(t)$ implies that $\sqrt{n}(\hat{\beta}_I-\beta_0)\xrightarrow{D}N(0,A_I^{-1}B_IA_I^{-1})$. $A_I=\lim_{n\to\infty}A_{I,n}(\beta_0)$ and $B_I=\lim_{n\to\infty}B_{I,n}(\beta_0)$ where $A_{I,n}(\beta)=n^{-1}\sum_{i=1}^n\delta_i\rho_I(X_i)\Bigg(1-\rho_I(X_i)\Bigg)$ and $B_{I,n}(\beta)=\sum_{j=1}^D\tilde{U}_j^I(\beta)\tilde{U}_j^I(\beta)^T$. δ_i an indicator for whether subject i experienced an event and $\rho_I(X_i)=\frac{n^{-1}\sum_{j=1}^n\tilde{Y}_j(X_i)Z_j\exp(Z_j\beta)}{n^{-1}\sum_{j=1}^n\tilde{Y}_j(X_i)\exp(Z_j\beta)}$ where $\tilde{Y}_j(X_i)$

is an indicator for whether subject j was originally sampled to be in the nested case-control risk set at time X_i. $\tilde{U}_j^I(\hat{\beta}_I)$ is defined as in (4.3). This provides the asymptotic distribution and establishes the consistency of the proposed estimator. $\square$

Based on the asymptotic properties of our estimator, we find that the finite-sample variance can be estimated using $\widehat{Var}(\hat{\beta}_I) = n^{-1}A_{I,n}^{-1}(\hat{\beta}_I)B_{I,n}(\hat{\beta}_I)A_{I,n}^{-1}(\hat{\beta}_I)$. Notice that A_n is the variance under the usual nested case-control design when the model is correctly specified. B_n represents the true variance and accounts for imputation of the risk sets through a Taylor expansion.

4.2.4 Incorporating Adjustment Covariates

So far we have introduced an estimator that recovers the FC results in the univariate setting. In observational studies, however, we often adjust for potential confounding variables to isolate the association of interest. As before, the imputation approach can be selected by the user, but in this manuscript we use a hotdeck multiple imputation approach [Fellegi and Holt, 1976]. Imputation of the risk sets can be accomplished as in Algorithm 3.

The estimating function in this setting takes the form $\tilde{U}_{IHD}(\beta) = \sum_{i=1}^{n} \int_{t=0}^{\infty} \left\{ \vec{Z}_i - \frac{\tilde{S}^{(1)}_{IHD}(\beta,t)}{\tilde{S}^{(0)}_{IHD}(\beta,t)} \right\} dN_i(t)$ where $\tilde{S}^{(r)}_{IHD} = n^{-1} \sum_{j \in \tilde{R}_I(t_j)} \vec{Z}_j^r \exp(\vec{Z}_j \beta)$ and $\vec{Z}_j$ is the vector of covariates. $\tilde{R}(t_j)$ represents the covariate values for the imputed risk sets which include the originally sampled subjects and the imputed subjects. For subjects who were not originally sampled into the nested case-control sample we draw values for the predictor of interest from a $N(\hat{\mu}(t), \sigma^2_{MSE}(t))$ as in the univariate setting. We match each of the imputed values to the subject from the l previous risk sets in the nested case-control sample with the closest value to the imputed value (l is selected by the user). The values of the adjustment variables of the selected subject are used to impute the values for the imputed subject. If more than one nested case-control subject can be used to impute the covariate values, we randomly sample one subject from eligible subjects. As stated earlier, l can be selected by the user. If event times are far apart, we recommend selecting a small l because neighboring event times may have drastically different risk sets. If event times are close, a larger l may be selected. In fact, if event times are close, l can be selected to include all previous risk sets. When sampling subjects to impute covariate values, however, one must ensure that the subjects selected are still at risk during the current event time.

Algorithm 3 Imputation approach for the multivariate setting with a continuous predictor

```
Imputation Approach for the Multivariate Setting
$D$: number of events
$t_j, j = 1, \cdots, D$: ordered event times
$R_j$: risk set at time $t_j$ under the FC
$\tilde{R}_{I,j}$: risk set at time $t_j$ including the imputed values and the originally sampled nested case-control sample controls
$\tilde{R}_j$: the risk set at time $t_j$ under the nested case-control design
$M$: number of controls sampled at each event time
$s(t)$: a natural spline with evenly spaced knots
$\mu(t)$: the mean covariate value at time $t$
$z_{ik1}$: predictor of interest for subject $i$ sampled in the nested case-control sample at time $t_k$
$\tilde{z}_{ik1}$: imputed value for the predictor of interest for subject $i$ at time $t_k$
$p$: the number of covariates in the model
$l$: the number of previous risk sets to consider for hotdeck imputations
procedure IMPUTATION OF CONFOUNDING VARIABLES (DONE AFTER ALGORITHM 2)
    for j in 1:D do
        for m in 1:length($\tilde{R}_I(t_j)$) do
            if $\tilde{Y}_m(t_j) \neq 1$ then Find $\min|\tilde{z}_{mj1} - z_{hj1}|$ for $h \in \cup_{k=(j-l)}^{j} \tilde{R}(t_k)$.
                Let $z_1^*$ be $z_{hj1}$ for $h \in \cup_{k=(j-l)}^{j} \tilde{R}(t_k)$ with the smallest absolute difference.
                Impute values of $\tilde{z}_{mj2}, \cdots, \tilde{z}_{mjp}$ using $z_2^*, \cdots, z_p^*$
            end if
        end for
    end for
    Fit a Cox PH model with the imputed subjects.
end procedure
```

Some covariate information, such as demographic information, may be easily available for all study participants. If this is the case, we may use this information (instead of estimating the mean covariate values) along with the hotdeck imputation method to impute covariate values that may be difficult or expensive to collect.

4.3 Empirical Performance

4.3.1 Univariate Results

We begin by presenting simulation results for the univariate setting. Table 4.1 illustrates the performance of the usual nested case-control PL estimator and our proposed estimator when the functional form is mis-specified. Values for the predictor of interest were sampled from a N(μ = 1.5, σ = 0.5) distribution. The true hazard function takes the form $\lambda(t) = \exp(\log(0.1) + \log(1.25)Z + \log(0.5)Z^2)$, failure times were drawn from Exp(rate = $\lambda(t)$), and censoring times were drawn from a Unif(0, 6) distribution. Observed event times were taken to be the minimum of the event and censoring times. Generating data in this way led to approximately 90% censoring and we included 2,000 subjects in each of the 200 simulations. We considered nested case-control samples with one to four controls per event time. We did not consider more than four controls since in practice people often use up to four controls [Ernster, 1994]. Moreover, it has been shown that using more than four controls does not provide a large benefit in terms of efficiency and power [Taylor, 1986].For each nested case-control sample, we sampled 60 controls at the first event time regardless of M. The 60 controls were only used at the first event time for the usual nested case-control PL estimator and for the proposed estimator. These 60 controls were also used to impute the covariate values of subjects not in the nested case-control sample. To illustrate the performance of the estimators under a mis-specified functional form we fit a model of the form $\lambda(t) = \lambda_0(t)\exp(\beta Z)$. The analytic variance estimates for the usual nested case-control PL estimator were obtained using the robust variance estimator while those of the proposed estimator were obtained using the estimator presented in Section 4.2.3.

	Mis-specified Model					**Correctly Specified Model**				
	N	Coeff.	% Est.	Emp.	An.	N	Coeff.	% Est.	Emp.	An.
		Est.	Bias	Var.	Var.		Est.	Bias	Var.	Var.
FC	2000.00	-1.4064	0.00	0.0167	0.0182	2000.00	0.9311	0.00	0.0239	0.0215
NCC										
M = 1	403.49	-1.6614	18.13	0.0808	0.0574	411.25	0.9272	-0.42	0.0595	0.0318
M = 2	544.65	-1.5686	11.53	0.0521	0.0371	553.62	0.9194	-1.26	0.0418	0.0254
M = 3	668.32	-1.5388	9.42	0.0324	0.0298	679.66	0.9237	-0.80	0.0332	0.0236
M = 4	778.99	-1.5308	8.85	0.0376	0.0271	791.18	0.9278	-0.35	0.0346	0.0228
Proposed Estimator										
M = 1	403.49	-1.4136	0.52	0.0776	0.1728	411.25	0.9156	-1.67	0.0675	0.1300
M = 2	544.65	-1.4065	0.01	0.0505	0.0773	553.62	0.9147	-1.76	0.0524	0.0627
M = 3	668.32	-1.4075	0.08	0.0422	0.0529	679.66	0.9185	-1.36	0.0426	0.0472
M = 4	778.99	-1.4219	1.10	0.0417	0.0437	791.18	0.9167	-1.55	0.0391	0.0395

Table 4.1: 200 simulations under a mis-specified functional form (left) and a correctly specified functional form (right). The nested case-control samples included 60 controls at the first event time, regardless of M. Empirical variance and analytic variance estimates are also provided.

Table 4.1 shows that the nested case-control PL estimator performs poorly when the model is not specified correctly and that the results obtained depend on the value of M, the number of controls sampled at each event time. The proposed estimator, however, reduces the bias relative to the full cohort estimator from approximately 18% to less than 1% when $M = 1$ and from approximately 9% to 1% when $M = 4$. The robust variance estimator for the nested case-control PL estimator tends to under estimate the variance for smaller values of M. Our proposed sandwich estimator is conservative when $M = 1$ but performs well for $M = 2$ to 4. To assess the robustness of the proposed estimator we also investigated the performance when the functional form is specified correctly. In this setting data were generated as in the first scenario, but now the true hazard function takes the form $\lambda(t) = \exp(\log(0.008) + \log(2.5)Z)$ and the failure times were drawn from Exp(rate $= \lambda(t)$). These data also had approximately 90% censoring. As seen on the right side of Table 4.1, when the model is specified correctly the nested case-control PL estimator estimates the same quantity as the FC estimator. Our proposed estimator also performs well regardless of M, yielding a bias (relative to the full cohort estimator) between 1% and 2% for all values of M. In this setting the proposed variance estimator again gives conservative estimates of the variance when $M = 1$, but performs well for $M = 2$ to 4. When the functional form is specified correctly, we observe a small loss in efficiency. However, this loss is nearly negligible and we have the added benefit that if the functional form is not specified correctly we still estimate the same quantity as that of the full cohort.

4.3.2 Multivariate Results

When using data from observational studies, it is almost always necessary to adjust for confounding variables. In Section 4.2.4 we described a hotdeck imputation approach to impute the values for the missing covariates. In this section we present simulation results when adjusting for a confounding variable. As before, we consider two scenarios- one in which the functional form of the predictor of interest is mis-specified and one in which the functional form is correctly specified. In this setting we assume that no covariate information is available for subjects not sampled into the nested case-control sample.

	N	$\hat{\beta}_1$	% Est.	Emp.	$\widehat{Var}(\hat{\beta}_1)$	$\hat{\beta}_2$	% Est.	Emp.	$\widehat{Var}(\hat{\beta}_2)$
			Bias	Var.			Bias	Var.	
Mis-specified Functional Form									
Full Cohort	2000.00	-1.4446	0.00	0.0226	0.0215	0.2753	0.00	0.0024	0.0022
NCC									
M = 1	383.68	-1.6510	14.29	0.0926	0.0685	0.2941	6.83	0.0067	0.0041
M = 2	517.73	-1.6080	11.31	0.0671	0.0442	0.2883	4.72	0.0046	0.0031
M = 3	635.58	-1.5923	10.22	0.0540	0.0372	0.2896	5.19	0.0037	0.0028
M = 4	742.26	-1.5658	8.39	0.0478	0.0331	0.2900	5.34	0.0034	0.0026
Proposed Estimator									
M = 1	383.68	-1.4964	3.59	0.1241	0.2869	0.2912	5.78	0.0086	0.0186
M = 2	517.73	-1.5055	4.22	0.0795	0.1188	0.2918	5.99	0.0063	0.0085
M = 3	635.58	-1.4596	1.04	0.0481	0.0768	0.2811	2.11	0.0045	0.0058
M = 4	742.26	-1.4707	1.81	0.0492	0.0617	0.2846	3.38	0.0042	0.0048
Correctly Specified Functional Form									
Full Cohort	2000.00	0.9152	0.00	0.0076	0.0068	-0.6847	0.00	0.0074	0.0066
NCC									
M = 1	408.57	0.9366	2.34	0.0241	0.0175	-0.7046	2.91	0.0208	0.0145
M = 2	550.38	0.9234	0.90	0.0165	0.0117	-0.6876	0.42	0.0154	0.0101
M = 3	675.37	0.9219	0.73	0.0130	0.0099	-0.6908	0.89	0.0130	0.0089
M = 4	786.52	0.9189	0.40	0.0124	0.0089	-0.6907	0.88	0.0115	0.0081
Proposed Estimator									
M = 1	408.57	0.9349	2.15	0.0301	0.0974	-0.7354	7.40	0.0360	0.0882
M = 2	550.38	0.9122	-0.33	0.0247	0.0389	-0.7007	2.34	0.0224	0.0339
M = 3	675.37	0.9081	-0.78	0.0159	0.0251	-0.6949	1.49	0.0164	0.0225
M = 4	786.52	0.9093	-0.64	0.0188	0.0201	-0.6869	0.32	0.0158	0.0178

Table 4.2: 200 simulations for hotdeck imputations, each with 3 imputations under mis-specification of the functional form (top) and a correctly specified functional form (bottom). The nested case-control samples included 60 controls at the first event time, regardless of M. Empirical and analytic variance estimates are also provided.

Table 4.2 presents the results for the multivariate setting with a mis-specified functional form. The predictor of interest and the confounding variable are distributed as $Z_1 \sim N(\mu = 1.5, \sigma = 0.5)$ and $Z_2 \sim N(\mu = 0 + 2 \cdot I(z_1 \geq 1.6), \sigma = 1.5)$, respectively. The true hazard function for this scenario takes the form $\lambda(t) = \exp(\log(0.075) + \log(1.25)Z_1 + \log(0.5)Z_1^2 + \log(1.35)Z_2)$ and failure times were drawn from Exp(rate $= \lambda(t)$). As before, censoring times were drawn from Unif(0, 6) and the observed times were the minimum of the observed and censoring times, yielding approximately 90% censoring. We sampled 60 controls at the first event time regardless of M and the model is assumed to take the form $\lambda(t) = \lambda_0(t)\exp(\beta_1 Z_1 + \beta_2 Z_2)$. We find that when the model is mis-specified, the usual nested case-control PL estimator produces biased coefficient estimates (when compared to the full cohort estimates) and the estimates obtained depend on the number of controls sampled at each event time. The proposed estimator, however, yields results similar to those of the FC PL estimator. When $M = 1$, the bias relative to the full cohort estimator is approximately 14% under the usual nested case-control PL estimator. This is reduced to approximately 4% when the proposed estimator is used. In this setting, the proposed variance estimator gives conservative estimates of the variance, but performance of the variance estimator improves as M increases. While the estimates provided by our variance estimator can be conservative, it should be noted that those provided by the robust variance estimator for the nested case-control PL estimator tend to give anti-conservative estimates of the variance. The bottom portion of Table 4.2 presents the results for the usual nested case-control PL estimator and our proposed estimator when the model is correctly specified. Data were generated as in the previous scenario, but the true hazard function takes the form $\lambda(t) = \exp(\log(0.0125) + \log(2.5)Z_1 + \log(0.5)Z_2)$, with failure times being drawn from Exp(rate $= \lambda(t)$). In this setting the fitted model takes the same form as the true data-generating mechanism. The usual nested case-control PL estimator and the proposed estimator perform similarly, both having a small bias relative to the full cohort estimator regardless of the selected M. In this setting, the proposed variance estimator is again conservative when $M = 1$, but its performance improves as M increases.

4.4 Application to ADNI Example

In this section we apply the proposed estimator to data from ADNI to investigate the association between the ADAS-11 at baseline and time to progression to AD dementia. The ADAS-11 is a cognitive test used to evaluate cognition and behavioral function, both of which are affected by AD [Rosen et al., 1984]. We had 974 participants in our analysis. These participants had ADAS-11 and CSF Aβ at baseline and did not have a diagnosis of Alzheimer's disease dementia at baseline. In our analysis, progression was defined as a stable clinical diagnosis of dementia or a diagnosis of dementia at the last visit. Based on this definition, approximately 15% of subjects experienced an event. Baseline characteristics of our sample can be found in Table 4.3. The mean age in the sample was 72.9 years, approximately 45% of participants were female, and approximately 42% of subjects had at least one APOE 4 allele.

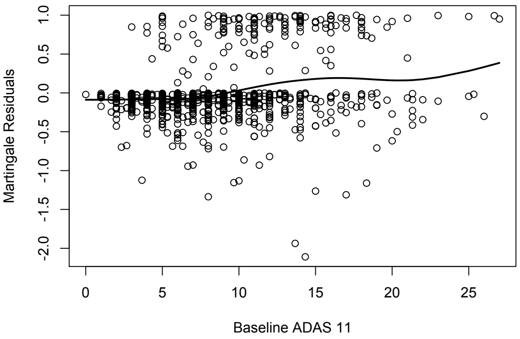


Figure 4.1: Martingale residuals plotted against baseline ADAS-11. The solid line represents a smoother.

As stated before, the goal of this analysis is to investigate the association between ADAS-11 at baseline and time to progression to AD. In this case, one may a priori specify a model that assumes a linear relationship between ADAS-11 and time to progression. Using the Martingale residuals as in Klein and Moeschberger [2005], we found that the functional form of ADAS-11 at baseline is not linear (Figure 4.1). Because the goal of the study was to investigate an association, changing the *a priori* selected model to fit the observed data could increase the Type I error rate and therefore is not recommended. Instead, we fit a first-order trend to investigate the behavior of the nested case-control design and the proposed estimator in this setting.

Characteristic	mean (sd) or n(%)
N	974
Progressors (to AD dementia)	152 (15.6%)
Age	72.9 (7.0)
Female	442 (45.4%)
White	909 (93.3%)
$\geq$ 1 APOE 4 allele	405 (41.6%)
ADAS-11	8.5 (4.6)
Mini-Mental State Examination	28.2 (1.7)
Education	16.2 (2.7)
Aβ	182.1 (53.7)

Table 4.3: Baseline demographics for subjects in our study.

We fit the Cox proportional hazards model to the entire sample to obtain the FC estimates. We then obtained 200 nested case-control samples for each value of M and applied the PL estimator and the proposed estimator as if we only had full covariate information for subjects in the nested case-control sample. We sampled 60 controls at the first event time for all nested case-control samples, regardless of M. All models were adjusted for age, education, race, the presence of at least one APOE 4 allele, gender, and baseline CSF Aβ levels. Because APOE 4 status and CSF Aβ levels would be the most difficult covariates to collect, we applied the nested case-control sampling scheme as if these measurements were not available. Demographic information, on the other hand, is easily collected for all study participants. Therefore, we assume that demographic information is available for all study participants, even if they were not sampled into the nested case-control sample. We used the hotdeck imputation method to impute values of APOE 4 and Aβ for participants not sampled into the nested case-control sample. Mahalanobis distance [Mahalanobis, 1936] was used to match subjects with missing values to sampled controls. When the covariance matrix was singular, we used Euclidean distance.

	N	ADAS-11	% Est.	Var.	APOE 4	% Est.	Var.	Aβ (HR)	% Est.	Var.
		(HR) 5 pts.	Bias	Est.	(HR)	Bias	Est.	50 pg/ml	Bias	Est.
Full Cohort	974.0	0.651 (1.92)	0	0.008	0.099 (1.10)	0.00	0.046	-0.537 (0.58)	0.00	0.010
Usual NCC										
M = 1	310.0	0.748 (2.11)	14.75	0.016	0.248 (1.28)	150.71	0.046	-0.583 (0.56)	8.57	0.020
M = 2	393.6	0.770 (2.16)	18.22	0.013	0.249 (1.28)	151.72	0.044	-0.572 (0.56)	6.52	0.010
M = 3	462.1	0.764 (2.15)	17.32	0.012	0.240 (1.27)	142.73	0.043	-0.570 (0.57)	6.15	0.010
M = 4	518.2	0.747 (2.11)	14.69	0.011	0.201 (1.22)	102.93	0.042	-0.562 (0.57)	4.66	0.010
M = 5	564.6	0.724 (2.06)	11.16	0.010	0.182 (1.20)	83.54	0.042	-0.560 (0.57)	4.28	0.010
Proposed Est.										
M = 1	310.0	0.677 (1.97)	3.99	0.064	0.159 (1.17)	61.01	0.201	-0.583 (0.56)	8.57	0.070
M = 2	393.6	0.668 (1.95)	2.61	0.032	0.103 (1.11)	3.74	0.118	-0.592 (0.55)	10.24	0.030
M = 3	462.1	0.660 (1.94)	1.37	0.023	0.119 (1.13)	19.90	0.091	-0.580 (0.56)	8.01	0.030
M = 4	518.2	0.657 (1.93)	0.92	0.019	0.092 (1.10)	-7.58	0.079	-0.575 (0.56)	7.08	0.020
M = 5	564.6	0.656 (1.93)	0.75	0.016	0.072 (1.07)	-26.97	0.073	-0.583 (0.56)	8.57	0.020

Table 4.4: Mean coefficient estimates for ADAS-11, APOE 4, and Aβ based on 200 nested case-control samples from the FC data. 60 controls were sampled at the first event time, regardless of M.

Table 4.4 presents the coefficient estimates for a difference of five points in baseline ADAS-11, our predictor of interest, as well as for APOE 4 and CSF Aβ. Under the FC PL estimator, we estimate that comparing two subpopulations that differ by five points in baseline ADAS-11, the risk of progression to AD dementia is approximately 92% higher for the group with higher ADAS-11. When we estimate the coefficients using the PL estimator and the nested case-control sampling scheme, we find that as in the simulated examples, the estimates are different than those obtained using the FC PL estimator and that these differ by the value of M. The bias relative to the full cohort estimator in this case ranges from 11% to 18% compared to the FC PL estimates. Applying our proposed estimator reduces this to between 0.75% and 4% while using the same sample sizes as the usual nested case-control design. Notice also that, as expected, the variance estimates for the proposed estimator are larger than those for the usual nested case-control PL estimator and that both are larger than those of the full cohort PL estimator.

To calculate the usual nested case-control PL estimator and the proposed estimator, we would only have to collect full covariate information for subjects in the nested case-control samples. That is, we would only have to perform genotype testing and process CSF samples for subjects who progressed to AD dementia or those who were sampled as controls. This reduces costs associated with these tests and allows us to use CSF samples to answer other questions that we may have about AD.

4.5 Discussion

It has been shown that the expectation of the sampling distribution of the usual nested case-control estimator will depend on the number of controls sampled at each event time when the PH assumption is violated. Previous work has proposed an estimator that yields the same results as those obtained using the FC data when the predictor of interest is binary [Nuño and Gillen, 2019]. In this scenario, the functional form of the covariate of interest is specified correctly, but the effect of the covariate is assumed to be constant when in reality it varies with time.

In this chapter, we consider the performance of the PL estimator under the nested case-control design when the effect of the covariate is constant over time, but the functional form is mis-specified. We again observe that the estimates obtained using the PL estimator under the nested case-control design also depend on the number of controls sampled at each event time. We therefore propose a method that estimates the same quantity as the FC PL estimator under mis-specification of the functional form, while only using the information from the usual nested case-control design. By only requiring full covariate information from the nested case-control sample, our proposed estimator maintains the reduction in costs afforded by the nested case-control design. The proposed estimator recovers the FC estimates when the model is mis-specified, both in the univariate and multivariate scenarios. When the model is specified correctly, the proposed estimator still recovers the FC estimates regardless of M. While the proposed estimator increases the bias relative to the full cohort estimator for $M = 1$ in the multivariate setting, it should be noted that M is usually larger than one in practice. Our proposed finite-sample variance estimator performs well for M greater than one but yields conservative estimates when $M = 1$.

It is known that the estimand corresponding to the FC PL estimator depends on the censoring distribution when the model is mis-specified [Struthers and Kalbfleisch, 1986, Xu and O'Quigley, 2000, Boyd et al., 2012]. In the previous chapter, we introduced an estimator for the FC censoring distribution that only requires the nested case-control sample. The estimator for the censoring distribution can also be used to reweight the estimating function to yield a censoring-robust estimator in this setting [Nuño and Gillen, 2019]. Similarly, the proposed weights for censoring-robust estimator can be applied to the Samuelsen [1997] estimator.

The nested case-control design provides great reduction in costs when the event of interest is rare. When the model is specified correctly, the nested case-control design estimates the same quantity as the FC PL estimator. If the functional form is mis-specified, however, the results obtained from the usual nested case-control estimator depend on the number of controls sampled at each event time. The proposed estimator uses the same information as the usual nested case-control design but recovers the FC results even when the functional form is mis-specified. We therefore recommend application of the proposed estimator since the estimator performs well even when the functional form is specified correctly and still affords the cost reductions offered by the nested case-control sampling scheme. When using our estimator, however, we do recommend using M larger than one (which is commonly done in practice).

Chapter 5

Censoring-Robust Time-dependent Receiver Operating Characteristic Curve Estimators

5.1 Introduction

Biomarker discovery is crucial in many fields since biomarkers play a critical role in understanding the mechanisms of disease, tracking disease development [Mayeux, 2004] and, oftentimes, testing for treatment efficacy [Strimbu and Tavel, 2010]. One area in which the discovery of biomarkers is of utter importance is in Alzheimer's disease (AD) where diagnosis of AD requires post-mortem verification. Recently, research in AD has shifted to earlier stages in an attempt to prevent the disease before it causes significant, irreversible damage. Existing biomarkers, such as the proteins amyloid beta (Aβ), total tau (T-tau) and phosphorylated tau (P-tau), help identify individuals who are more likely to develop AD [Blennow et al., 2010, Blennow, 2005, 2004]. However, it is difficult to distinguish early AD from other disorders involving similar symptoms (Humpel [2011]). Moreover, it is difficult to distinguish between symptoms of AD and those of normal aging [Denver and McClean, 2018]. The discovery of new biomarkers for AD could help not only to accurately diagnose people, but could also provide targets for therapeutic treatments, allowing us to develop treatments for AD. The importance of biomarker discovery is not unique to AD, making it even more important to have reliable methods to aid in the discovery of new biomarkers.

Receiver operating characteristic (ROC) curves are often used to evaluate the classification performance of continuous measures, such as potential biomarkers. ROC curves are defined by the sensitivity and specificity of a biomarker over a range of thresholds. *Sensitivity* is the probability that an individual is classified as testing positive (or meeting a specific threshold) given that they have the disease, while *specificity* is the probability that an individual is classified as testing negative (or not meeting the threshold value) given that the individual does not have the disease. A common summary measure of biomarker performance is the area under the ROC curve (AUC), which provides an estimate of the probability that a randomly selected individual with the disease will be rated higher than one without the disease [Fawcett, 2006].

When dealing with time-to-event data, it is common for the disease status and the risk sets to change over time. Because of this, Heagerty et al. [2000] proposed using time-dependent ROC curves in which the sensitivity, specificity, and the corresponding AUC are estimated at a particular time point. There are several ways to estimate sensitivity and specificity in the time-dependent ROC setting [Heagerty et al., 2000, Chambless and Diao, 2006, Zheng et al., 2006, Uno et al., 2007, Heagerty and Zheng, 2005]. Some methods rely on nonparametric estimation, while others use semiparametric methods to estimate the sensitivity and specificity. Several of these estimators [Heagerty and Zheng, 2005, Chambless and Diao, 2006] calculate the sensitivity and specificity using a risk score made of up several covariates, or biomarkers. While the risk score can be estimated in various ways, a common approach is to use a linear predictor based on the partial likelihood estimator [David et al., 1972].

As seen in previous chapters, the partial likelihood estimator is commonly used to estimate the hazard ratio for time-to-event data. Its popularity is partially due to the fact that the Cox proportional hazards model does not require specification of the baseline hazard [David et al., 1972]. While the Cox model is of great utility, it has been shown that when the model is mis-specified, the estimand corresponding to the partial likelihood estimator depends on the censoring distribution [Struthers and Kalbfleisch, 1986]. As seen throughout this dissertation, when the goal of a study is to conduct inference, the dependence on the censoring distribution makes it difficult to replicate results across studies. If the goal is to evaluate classification performance, dependence on the censoring distribution would make it difficult to accurately evaluate the biomarker's performance since the results will depend on dropout and accrual patterns of the current study.

The goal of this chapter is to investigate how the censoring distribution affects estimates of the AUC when the risk score used to estimate the sensitivity and specificity is obtained using coefficient estimates derived from the partial likelihood estimator and the proportionality assumption is not satisfied. We also provide a method for censoring-robust estimation of the AUC. We begin with a brief review of time-dependent ROC curves. We then discuss the impact of model mis-specification under the partial likelihood estimator, investigate the performance of the AUC when the partial likelihood estimator is used in violation of the proportionality assumption, and propose the use of censoring-robust estimators. We end the chapter with an application of the proposed methodology to Alzheimer's disease using data from ADNI.

5.2 Background

5.2.1 Time-dependent ROC Curves

ROC curves are commonly used to assess the classification performance of a continuous variable and are obtained by plotting sensitivity versus (1 - specificity). The area under the curve (AUC) is often used as a summary measure for the ROC curve and can be interpreted as the probability that a randomly selected individual with the disease will have a higher level of the marker than someone without the disease [Fawcett, 2006, Heagerty et al., 2000]. The higher the AUC, the better the biomarker is for classifying diseased and non-diseased individuals. One benefit of ROC curves is that the sensitivity and specificity are estimated over all possible cut points, so the results do not depend on a single cut point value. ROC curves also allow comparison of different markers, even if these are on different scales [Heagerty et al., 2000].

Classic ROC curves assume the disease status is fixed. In many cases, however, the disease status may change over time. For these scenarios, Heagerty et al. [2000] proposed the use of time-dependent ROC curves, where the sensitivity and specificity are estimated at a particular time point. The use of time-dependent ROC curves allows us to investigate the classification performance at different event times and therefore also allows us to investigate the performance of the marker over time. When using time-dependent ROC curves, cases and controls can be defined in several ways, and how these are defined will impact estimation of the sensitivity and specificity. In this chapter, we consider two commonly encountered scenarios: (1) cumulative sensitivity and dynamic specificity and (2) incident sensitivity and dynamic specificity. Cumulative sensitivity considers the probability that an individual's risk score exceeds some threshold value, c, given that the individual experienced an event after baseline and before time t. Dynamic specificity is the probability that an individual has a risk score value (B_i) less than or equal to c conditional upon not having experienced an event up to time t [Kamarudin et al., 2017]. These can be written as $\text{Sens}_C(t, c) = P(B_i > c|T_i \leq t)$ and $\text{Spec}(t, c) = P(B_i \leq c|T_i > t)$. Incident specificity, on the other hand, is the probability that an individual has a risk score measure above c given that they experienced an event exactly at time t and can be written as $\text{Sens}_I(t, c) = P(B_i > c|T_i = t)$. The cumulative/dynamic setting is more appropriate when there is a specific time of scientific interest at which investigators would like to see who has developed the disease and who has not. On the other hand, the incident/dynamic setting is more appropriate when the exact event time is known and it is of interest to investigate who has developed the disease at that time [Kamarudin et al., 2017]. While several estimators have been proposed for both scenarios, we will focus on the cumulative/dynamic estimator of Chambless and Diao [2006] (Section 3.3) and the incident/dynamic estimator of Zheng et al. [2006].

The work of Heagerty and Zheng [2005] considers time-dependent ROC curves to evaluate the classification performance of risk scores made up of one or more covariates, or biomarkers. The risk scores can be obtained using a variety of models, but their manuscript focuses on the Cox proportional hazards model. Sensitivity and specificity are also estimated using the Cox proportional hazards model. When the data follow non-proportional hazards, the sensitivity is estimated as $\widehat{\text{Sens}}_{HZ}(t, c) = \sum_{k=1}^{n}[I(B_k > c)Y_k(t)\exp(B_k\hat{\gamma}(t))/\sum_{j=1}^{n} Y_j(t)\exp(B_j\hat{\gamma}(t))]$ where $Y_k(t)$ is an indicator for whether individual k is at risk at time t and $\hat{\gamma}(t)$ is the estimate for the time-dependent coefficient. The specificity is estimated using $\widehat{\text{Spec}}_{HZ}(t, c) = \frac{\sum_{k=1}^{n} I(B_k > c)Y_k(t+)}{W^R(t+)}$ where $Y_k(t+) = \lim_{\delta \to 0} Y_k(t + |\delta|)$ and $W^R(t+)$ is the number of controls (non-events) in the risk set at time t. When the data follow proportional hazards, $\hat{\gamma}(t)$ in the sensitivity is replaced by $\hat{\gamma}$, an estimate for the time-invariant effect. In this manuscript, we focus on the proportional hazards estimator. While the first estimator for sensitivity is more flexible, investigators may not expect a time-varying effect, or exact specification of the change points may not be known. In this case, one may consider estimating the sensitivity using a time-invariant effect. Moreover, the widely used `CoxWeights` function in the `risksetROC` package in R implements the proportional hazards approach. Therefore, it is important that we investigate the properties of this procedure when the model is mis-specified.

Another way to estimate the sensitivity and specificity under a time-dependent ROC curve uses the methods proposed by Chambless and Diao [2006] for the cumulative/dynamic scenario. These methods also allow for assessment of a risk score, which can be obtained using various models. In their manuscript, the authors present Kaplan and Meier [1958] type estimators of the sensitivity and specificity as well as alternative estimators that use a regression-based estimator of the survival function. The R function `AUC.cd` in the `survAUC` package implements the regression-based approach using a Cox proportional hazards model. Under this approach, the sensitivity and specificity take the form $\text{Sens}_{CD}(t,c) = \frac{E[(1-S(t|B))I(B>c)]}{E[1-S(t|B<c)]}$ and $\text{Spec}_{CD}(t,c) = \frac{E[S(t|B)I(B<c)]}{E[S(t|B)]}$. $S(t|B)$, the survival function at time t, can be estimated using $\hat{S}(t|B) = \exp(-\hat{\Lambda}_0(t)\exp(-\hat{\gamma}B))$ where $\hat{\Lambda}_0(t)$ is obtained using the Breslow [1972] estimator and $\hat{\gamma}$ is calculated using the Cox proportional hazards model.

Other estimators have been proposed that account for censoring using inverse probability weights. The work of Uno et al. [2007] and Hung and Chiang [2010] reweight the estimator of the sensitivity from Heagerty et al. [2000] by the inverse of the survival function for censoring, $S_c(t)$. The estimator for sensitivity is $\widehat{\text{Sens}}_{IW}(t,c) = \frac{\sum_{i=1}^n I(B_i>c, X_i \leq t)\delta_i/[n\hat{S}_c(X_i)]}{\sum_{i=1}^n I(X_i \leq t)(\delta_i/[n\hat{S}_c(X_i)])}$ and the estimator for specificity is $\widehat{\text{Spec}}_{IW}(t,c) = \frac{\sum_{i=1}^n I(B_i \leq c, X_i>t)}{\sum_{i=1}^n I(X_i>t)}$. This estimator, however, does not account for marker dependent censoring. Blanche et al. [2013] extended this method to allow for marker dependent censoring. Their proposed estimators are $\widehat{\text{Sens}}_B(t,c) = \frac{\sum_{i=1}^n I(B_i>c, X_i \leq t)[\delta_i/(n\hat{S}_c(X_i|B_i))]}{\sum_{i1}^n I(Z_i \leq t)(\delta_i/(n\hat{S}_c(X_i|B_i)))}$ and $\widehat{\text{Spec}}_B(t,c) = \frac{\sum_{i=1}^n I(B_i \leq c, X_i>t)[1/(n\hat{S}_c(t|B_i))]}{\sum_{i1}^n I(X_i>t)(1/(n\hat{S}_c(t|B_i)))}$. While these estimators reweight by the inverse probability of no censoring, they focus on Kaplan-Meier type estimators of the sensitivity and specificity. Moreover, these articles do not show the impact of model mis-specification on estimates of the AUC under different censoring scenarios. In this manuscript, our goal is to emphasize the effect of the censoring distribution on estimates of the AUC and to propose estimators that can be quickly and easily implemented in R with existing packages. In the following section, we present the partial likelihood estimator and its behavior under model mis-specification.

5.2.2 Model Mis-specification under the Partial Likelihood Estimator

As we have seen in the previous section, various methods for time-dependent ROC curves make use of the partial likelihood estimator, either to obtain a risk score or to estimate the sensitivity and specificity. The partial likelihood estimator is commonly used when dealing with time-to-event data and does not require specification of the baseline hazard. Recall that the estimating function under the Cox proportional hazards model takes the form

$$U(\beta) = \sum_{i=1}^{n} \delta_i \left(Z_i - \frac{\sum_{j=1}^{n} Z_j Y_j(t) \exp(\beta Z_j)}{\sum_{j=1}^{n} Y_j(t) \exp(\beta Z_j)} \right).$$

However, Struthers and Kalbfleisch [1986] show that the estimand corresponding to the partial likelihood estimator is the solution to

$$\int_0^\infty E_Z \left(f_T(t|Z) S_C(t|Z) \left[Z - \frac{E_Z\{Z S_T(t|Z) S_C(t|Z) \exp(Z\beta)\}}{E_Z\{S_T(t|Z) S_C(t|Z) \exp(Z\beta)\}} \right] \right) dt = 0,$$

which depends on $S_C(t|Z)$, the covariate-specific censoring distribution. When the proportionality assumption is satisfied, the dependence on the censoring distribution does not impact the results. As seen in previous chapters, if the proportionality assumption is violated, the results will depend on the censoring distribution. That is, simply changing the accrual and dropout patterns will change the quantity estimated by the partial likelihood estimator. This makes it difficult to replicate results across studies, since the dropout and accrual patterns will likely differ. Fortunately, several estimators have been proposed that estimate the average covariate effect, which does not depend on the censoring distribution.

Xu and O'Quigley [2000] consider the scenario where $S_C(t) = S_C(t|Z)$. As shown by Boyd et al. [2012], in this case, (2.12) simplifies to $\int_0^\infty E_Z\Bigg(f_T(t|Z)S_C(t)\Bigg[Z - \frac{E_Z\{ZS_T(t|Z)\exp(Z\beta)\}}{E_Z\{S_T(t|Z)\exp(Z\beta)\}}\Bigg]\Bigg)dt = 0$. To get rid of the dependence on $S_C(t)$, Xu and O'Quigley [2000] propose using the estimating function

$$U_{XO}(\beta) = \sum_{i=1}^{n} \delta_i W_{XO}(X_i)\Bigg(Z_i - \frac{\sum_{j=1}^{n} Z_j Y_j(t)\exp(\beta Z_j)}{\sum_{j=1}^{n} Y_j(t)\exp(\beta Z_j)}\Bigg)$$

where $W_{XO}(t) = \hat{S}_{KM}(t)/\sum_{i=1}^{n} Y_i(t)$ and $S(t)$ can be estimated using the Kaplan-Meier estimator.

In many scenarios, however, the censoring distribution will depend on the covariate values. Boyd et al. [2012] consider the scenario where the censoring distribution depends on a single, binary covariate. They propose the following reweighted estimating function

$$U_{CR}(\beta) = \sum_{i=1}^{n} \delta_i/\hat{S}_{C,KM}(t|Z_i)\Bigg(Z_i - \frac{\sum_{j=1}^{n} Z_j Y_j(t)/\hat{S}_{C,KM}(t|Z_j)\exp(\beta Z_j)}{\sum_{j=1}^{n} Y_j(t)/\hat{S}_{C,KM}(t|Z_j)\exp(\beta Z_j)}\Bigg).$$

The resulting estimator again recovers a censoring-independent estimand that can be interpreted as the average covariate effect. Note that although the Boyd et al. [2012] estimator allows for a covariate-dependent censoring distribution, it can also be used when the censoring distribution does not depend on other covariates.

While the estimator proposed by Boyd et al. [2012] allows for dependence of the censoring distribution on a single covariate, the censoring distribution might depend on multiple, possibly continuous, covariates. Nguyen and Gillen [2017] proposed survival tree based estimators for these scenarios. Succinctly stated, the three estimators presented in this section allow us to recover an estimand that does not depend on the censoring distribution, therefore allowing us to replicate results across studies.

5.3 Censoring-Robust Time-dependent ROC Curves

Due to the dependence of the partial likelihood estimator on the censoring distribution and because the sensitivity and specificity can be estimated using Cox regression, we hypothesized that estimates of the AUC would also differ when the risk score model is mis-specified. While there are various ways to account for the dependence on the censoring distribution, we propose estimating the risk scores using the estimators presented in Section 5.2.2. In this chapter, we apply the work of Boyd et al. [2012], which can be used when the censoring distribution depends on a single covariate. If the censoring distribution depends on more than one covariate, the estimator proposed by Nguyen and Gillen [2017] should be used.

In Section 5.2.1 we saw that some estimators allow for time-dependent effects when estimating the time-dependent ROC curves. However, application of these estimators would require specification of the change points, which is not always obvious. Other works have proposed inverse probability weighted estimators where the sensitivity and specificity are calculated using a weighted Kaplan-Meier type estimator. The work presented in this manuscript differs from previous work in several ways. First, we explore the impact of the censoring distribution on the AUC estimates. We also propose the use of censoring-robust estimators for estimation of the risk scores, which corrects the dependence on the censoring distribution and can be easily implemented in R using existing functions.

5.3.1 Empirical Performance

In this section, we consider the empirical performance of the Chambless and Diao [2006] and Heagerty and Zheng [2005] AUC estimators when the risk score is calculated using the Cox partial likelihood estimator and the estimator proposed by Boyd et al. [2012]. We performed 2,000 simulations, each with 2,000 subjects. Two variables, Z_1 and Z_2, were used to estimate the risk score for the AUC. Half of all subjects had $z_2 = 0$ and half had $z_2 = 1$. $Z_1 \sim N(\mu = 2 + 2 * Z_2, sd = 1)$. The true hazard function takes the form $\lambda(t) = \exp\left\{ \log(0.5) + \log(0.85)Z_1 + \log(0.05) \cdot Z_2 \cdot I(t \leq 3) + \log(2) \cdot Z_2 \cdot I(3 < t \leq 6) + \log(1) \cdot Z_2 \cdot I(t > 6) \right\}$ with change points at times $t = 3$ and $t = 6$. Event times, T_i, were drawn from $Exp(\lambda(t))$. The observed times were taken to be $X_i = \min(T_i, C_i)$ and we applied different censoring distributions following the cumulative density function $F_C(c) = (c/7)^r$, with r taking values from 0.25 to 4. Under the nonproportional hazards scenario, the risk score was calculated using a linear combination of the form $\hat{\beta}_1 Z_1 + \hat{\beta}_2 Z_2$ where the coefficient estimates were obtained from a model of the form $\hat{\lambda}(t) = \hat{\lambda}_0 \exp(\hat{\beta}_1 Z_1 + \hat{\beta}_2 Z_2)$ using the partial likelihood and the censoring-robust estimator. Note that although data were generated with a time-varying effect, the risk score was calculated assuming a time-invariant effect. Because the Chambless and Diao [2006] estimator requires a training set and a test set, we generated an additional 1,000 observations to use as the test set when implementing their estimator.

Figure 5.1 presents the density curves for the different censoring distributions (left plot) along with the coefficient estimates under no censoring, the partial likelihood estimator, and the Boyd et al. [2012] estimator. Note that, as discussed in Section 5.2.2, when the model is mis-specified, the coefficient estimates obtained from the partial likelihood estimator differ depending on the censoring distribution. When censoring times are drawn using scenario 1 ($r = 0.25$), we find that the coefficient estimate corresponding to Z_2 is estimated to be approximately -1.21. When we consider censoring scenario 16 ($r = 4$), the coefficient estimate is approximately -0.7. The estimates obtained using the Boyd et al. [2012] estimator, on the other hand, do not depend on the censoring distribution. In this case, the coefficient estimate corresponding to Z_2 is approximately -0.61 regardless of the censoring distribution.

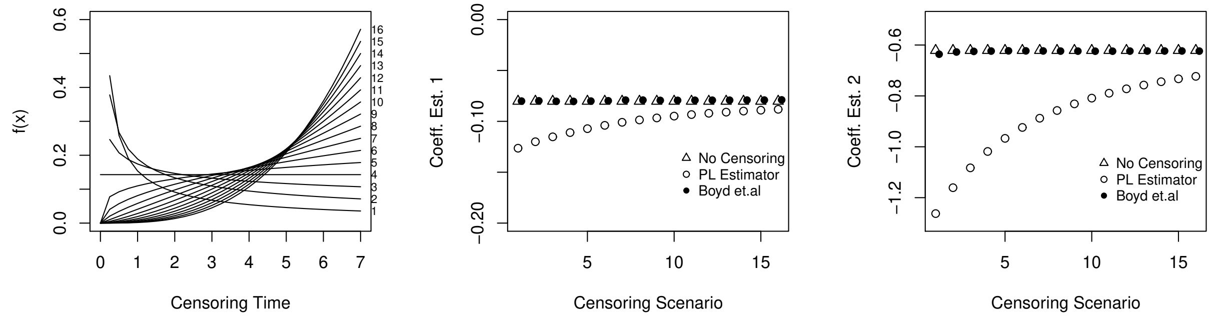


Figure 5.1: Density curves for the censoring times (left) and the corresponding estimates when there is no censoring, under the partial likelihood estimator, and the censoring-robust estimator based on 2000 simulations.

Figure 5.2 presents the estimates of the AUC for two common estimators. On the left is the estimator of Chambless and Diao [2006] and on the right is the estimator of Heagerty and Zheng [2005]. In both cases the partial likelihood estimator is used to calculate the risk score. Note that, as seen with the coefficient estimates, the estimates of the AUC differ depending on the censoring distribution when the partial likelihood estimator is used to obtain the risk score. Using the Chambless and Diao [2006] estimator under censoring scenario 1, we find that the AUC is estimated to be approximately 0.71, and in censoring scenario 16, the AUC is estimated to be approximately 0.63. Similarly, when the Heagerty and Zheng [2005] estimators are used, we find that the estimate of the AUC is approximately 0.70 under censoring scenario 1 and 0.60 under censoring scenario 16. Similar trends are observed at times $t = 4.5$ and 5.5. When the Boyd et al. [2012] estimator is used to estimate the risk score, the estimates of AUC remain at approximately 0.62 at time $t = 1.5$ regardless of the censoring distribution used under the Chambless and Diao [2006] and at approximately 0.60 under the Heagerty and Zheng [2005] estimator. The estimates of the AUC under the Boyd et al. [2012] estimated risk score are not impacted by the censoring distribution.

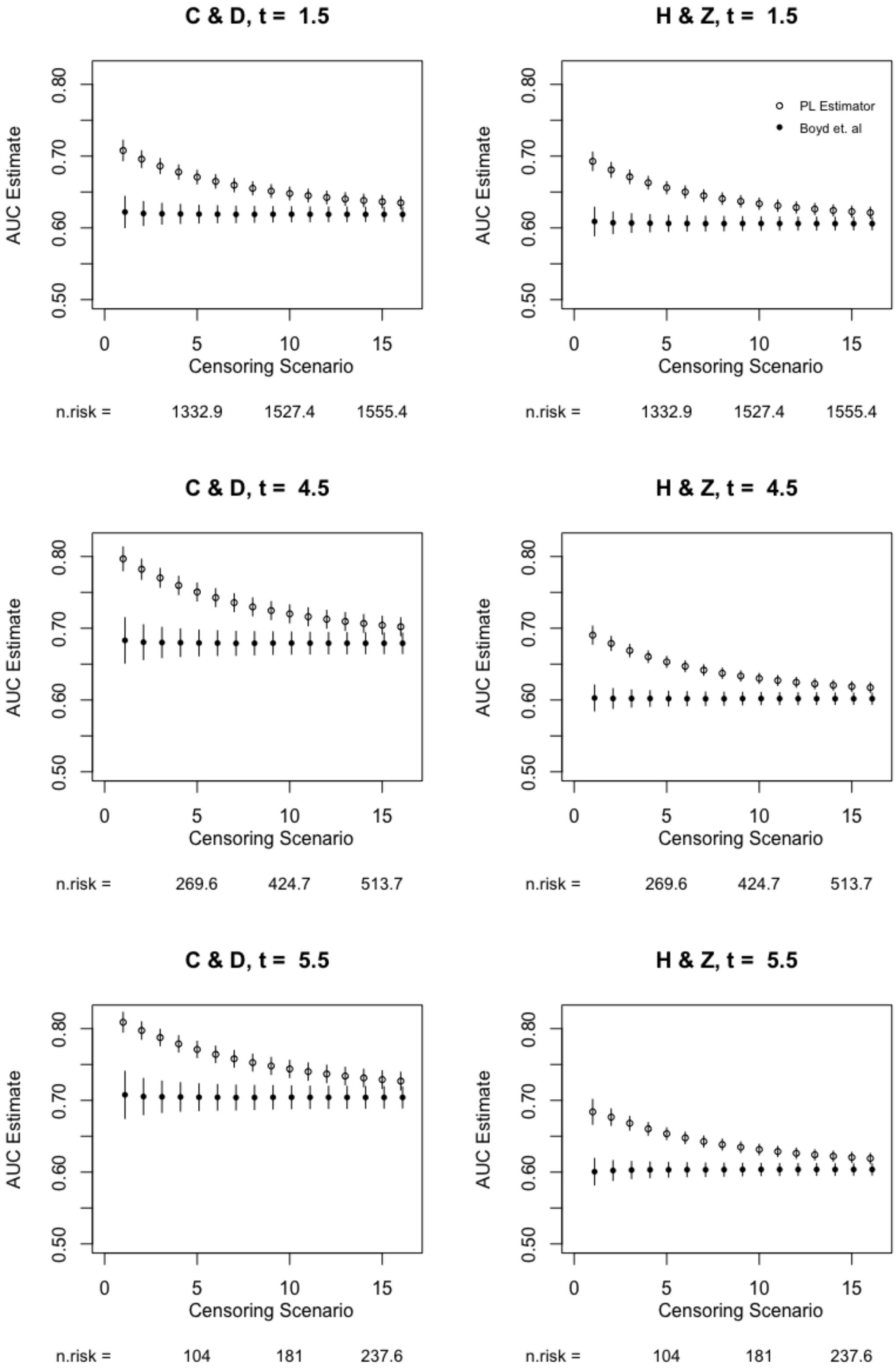


Figure 5.2: AUC estimates (± 1 s.d) based on 2000 simulations under nonproportional hazards. The number of subjects at risk under each censoring scenario can be found below each plot. Risk scores are calculated using the partial likelihood estimator and the censoring-robust estimator of Boyd et al. [2012].

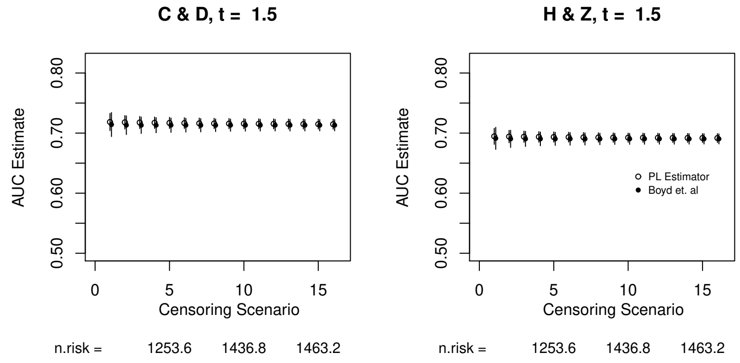


Figure 5.3: AUC estimates (± 1 s.d) based on 2000 simulations under proportional hazards. Risk scores are calculated using the partial likelihood estimator and the censoring-robust estimator of Boyd et al. [2012].

To ensure that the Boyd et al. [2012] estimator still performs well when the model is specified correctly, we considered another simulation setting, this time satisfying the proportionality assumption. The covariates and censoring times were drawn in the same way as in the non-proportional hazards setting. The hazard function now takes the form $\lambda(t) = \exp(\log(0.5) + \log(0.85)Z_1 + \log(0.30)Z_2)$.

Figure 5.3 presents the AUC estimates under the proportional hazards setting. We find that when the proportionality assumption is satisfied, both the partial likelihood estimator and the Boyd et al. [2012] estimator perform well under the methods proposed by Chambless and Diao [2006] and Heagerty and Zheng [2005]. The variability in the estimates (based on the empirical variance) of the AUC is similar regardless of which estimator is used to obtain the risk score.

5.4 Bootstrap Estimates of the Variance

In Section 5.3.1, we presented the AUC estimates along with the empirical standard deviation. In practice, we will only observe one realization, and can estimate the variance of the AUC estimates using a bootstrapping approach. To obtain bootstrap estimates, we sample with replacement from our original sample and obtain estimates of the coefficients, risk scores, and the AUC. If a training and test set are used, bootstrap samples are required for both the training set and the test set. In this section, we present estimates of the bootstrap variance of the AUC estimators of Chambless and Diao [2006] and those of Heagerty and Zheng [2005]. The simulation set-up is the same as that in Section 5.3.1. The only difference is that we ran 1,000 simulations with 1,500 subjects. The bootstrap estimates are based on 100 bootstrap samples.

Figure 5.4 presents the empirical variance and the bootstrap estimates of the variance for the AUC estimates under the partial likelihood and censoring-robust estimators. For the Chambless and Diao [2006] estimator, the bootstrap estimator is able to correctly estimate the variance. When considering the censoring-robust estimator, the bootstrap estimates tend to be slightly lower than the true variance. Even in this scenario, however, the bootstrap estimates are close to the empirical variance. Under the Heagerty and Zheng [2005] estimator of the time-dependent AUC, we find that the bootstrap estimates are very close to the empirical variance. The only exception lies in the first censoring scenario at time $t = 5.5$. In this case, the bootstrap estimates tend to be larger than the empirical variance. This is likely due to the fact that there are less people at risk, and therefore less events, at time $t = 5.5$.

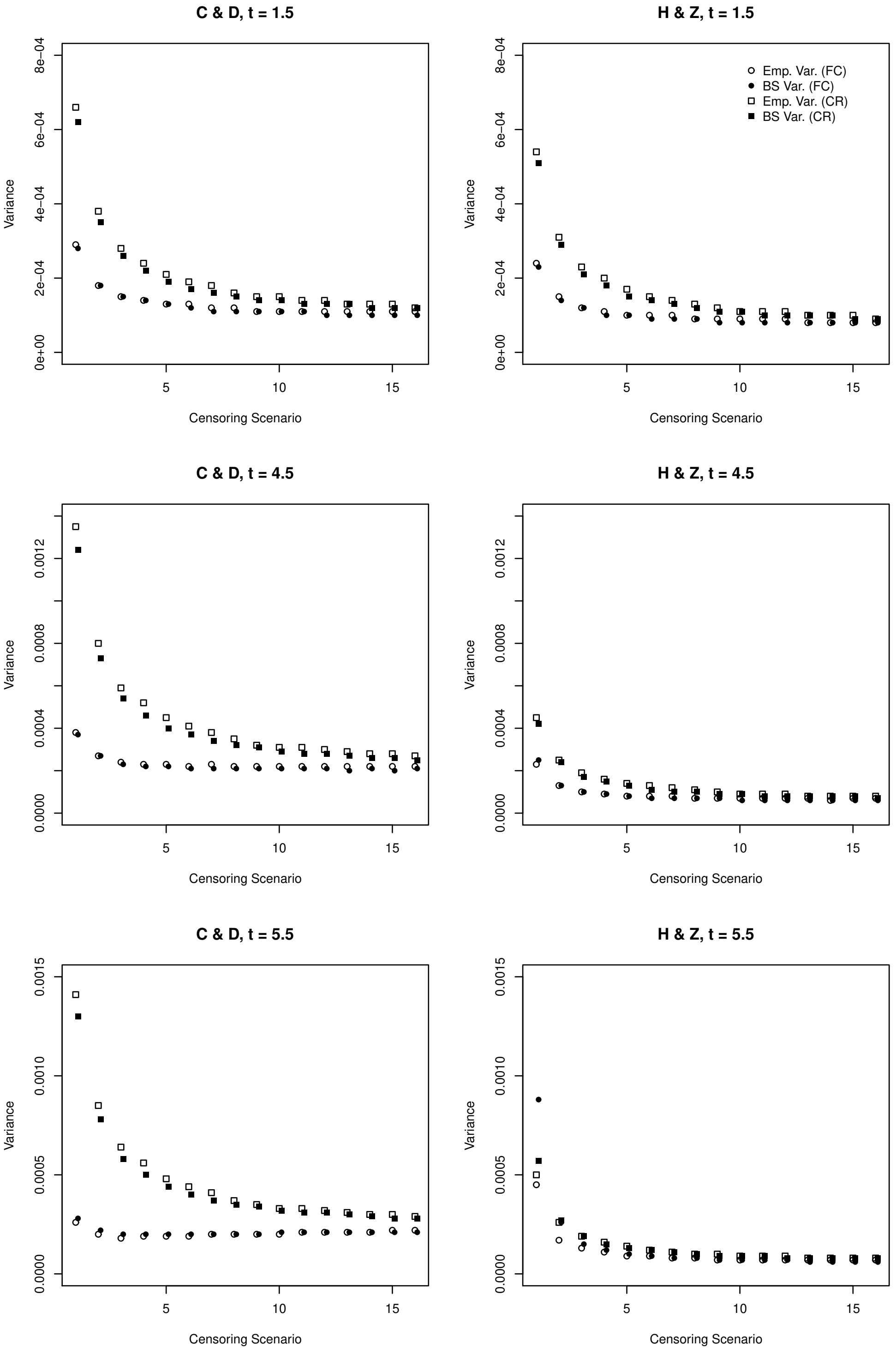


Figure 5.4: Bootstrap estimates of the variance based on 1,000 simulations each with 1,500 subjects. Results are obtained using 100 bootstrap samples for each simulation.

5.5 Evaluating the Classification Performance of Aβ and APOE 4

As discussed in previous chapters, there is currently no cure for AD. In an attempt to prevent AD, research has shifted to earlier stages during which people have few or no symptoms. In these settings, it is important to consider early biomarkers to help identify people who are likely to progress. While biomarkers such as CSF Aβ exist, in the early stages it is difficult to distinguish between symptoms of AD and those of normal aging [Denver and McClean, 2018]. Even when signs of dementia are present, it is difficult to distinguish AD from other types of dementia [Humpel, 2011]. In order to develop treatments for AD, we must find new biomarkers, or combinations of biomarkers and other risk factors, to help identify subjects who are likely to progress to AD. Moreover, the discovery of new biomarkers could provide targets for therapeutic treatments.

There are currently several known biomarkers and risk factors for AD. One biomarker is amyloid beta, or Aβ, a protein that can be measured in cerebrospinal fluid (CSF), with low levels of CSF Aβ indicating increased AD pathology [Grimmer et al., 2009]. A threshold of 192 pg/ml has been proposed as a possible cutoff to distinguish people with low levels of CSF Aβ and those not considered to have low levels of Aβ [Shaw et al., 2009]. The apolipoprotein E (APOE) 4 allele is also associated with a higher risk of developing AD. People may have one or two alleles of this gene and the risk of AD increases with the number of alleles present [Saunders et al., 1993].

In this section, we investigate the performance of CSF Aβ and APOE 4 for distinguishing people who develop AD dementia from those who do not. Aβ was included as an indicator for whether the individual had low levels of Aβ at baseline. To increase the classification performance, we also included other measures that are associated with AD. These measures included the Alzheimer's Disease Assessment Cognitive Subscale- 11 (ADAS-11) [Rosen et al., 1984] and the Mini-Mental State Exam (MMSE) [Folstein et al., 1975], two cognitive assessments. We hypothesized that there would be a time-varying effect since all measures included were those at baseline. Therefore, we developed risk scores using the partial likelihood estimator and the censoring-robust estimator of Boyd et al. [2012] to compare the estimates of the AUC.

We used data from the Alzheimer's Disease Neuroimaging Initiative (ADNI), an ongoing cohort study, and required that participants had Aβ, ADAS-11, and MMSE at baseline, as well as their APOE 4 status. To be in our study, participants could not have a diagnosis of dementia at baseline, and progression was defined as progressing from an earlier stage to a diagnosis of dementia. Table 5.1 presents the baseline demographics of participants in our study. We included 973 individuals, with a mean age of 72.92 years. Participants had an average of 16.19 years of education. Approximately 40% of participants had at least one APOE 4 allele, and the average baseline measure of CSF Aβ was 182.14 pg/ml.

	Mean (sd) or n (%)
N	973
Age	72.92 (7.95)
Education	16.19 (2.71)
Female (vs. Male)	442 (45.43%)
White (vs. non-White)	908 (93.32%)
$\geq$ 1 APOE 4 allele (vs. 0)	404 (41.52%)
Aβ	182.14 (53.76)
ADAS-11	8.52 (4.56)
MMSE	28.21 (1.73)

Table 5.1: Baseline characteristics of participants in our study.

We considered the performance of the cumulative/dynamic estimator of Chambless and Diao [2006] and the incident/dynamic estimator of Heagerty and Zheng [2005] when risk scores were obtained using the partial likelihood estimator and the censoring-robust estimator of Boyd et al. [2012]. While the simulation results for the Chambless and Diao [2006] estimator were based on the availability of a training and test set, due to the number of participants in this study, in this example we only used a training set. Along with the coefficient estimates, Table 5.2 includes AUC estimates at years one, two, three, and four.

Note that the coefficient estimates differ for the partial likelihood estimator and the censoring-robust estimator. When using the partial likelihood estimator, we estimated that the risk of progressing to AD dementia was 25% higher for people with at least one APOE 4 allele compared to those without any allele. The hazard ratio was estimated to be 1.71 under the censoring-robust estimator. The largest difference was observed when considering Aβ. The risk of developing AD dementia was estimated to be approximately 3.7 times higher for people with low Aβ at baseline compared to those who do not have low Aβ. When using the censoring-robust estimator, the hazard ratio was estimated to be 2.64. The differences in the coefficient estimates led to differences in the AUC estimates. Under the Chambless and Diao [2006] estimator, the AUC was estimated to be 0.784 when the risk score was calculated using the partial likelihood estimator and 0.721 when the censoring-robust estimator was used. At year four, the AUC was estimated to be 0.795 under the partial likelihood estimator and 0.735 under the censoring-robust estimator. Similar results were observed when using the Heagerty and Zheng [2005] estimator. At year 1, the AUC was 0.782 under the partial likelihood estimator and 0.721 under the censoring-robust estimator. At year four, the AUC estimate was 0.768 under the partial likelihood estimator and 0.715 under the censoring-robust estimator.

	Log Hazard Ratio (Hazard Ratio)				AUC Estimates			
	APOE 4	ADAS-11 (bl)	Abeta < 192	MMSE (bl)	1 year	2 years	3 years	4 years
C & D								
PLE	0.220 (1.25)	0.123 (1.13)	1.305 (3.69)	-0.096 (0.91)	0.784	0.787	0.791	0.795
CR	0.535 (1.71)	0.085 (1.09)	0.969 (2.64)	0.011 (1.01)	0.721	0.725	0.729	0.735
H & Z								
PLE	0.220 (1.25)	0.123 (1.13)	1.305 (3.69)	-0.096 (0.91)	0.782	0.780	0.777	0.768
CR	0.535 (1.71)	0.085 (1.09)	0.969 (2.64)	0.011 (1.01)	0.721	0.721	0.719	0.715

Table 5.2: AUC estimates when APOE 4, ADAS-11, Aβ eligibility, and MMSE are used to predict progression to dementia.

There is a large discrepancy when using the partial likelihood estimator and the censoring-robust estimator. As seen in Section 5.3.1, the estimates of the AUC obtained using the partial likelihood estimator differ depending on the censoring distribution while those obtained from the censoring-robust estimator do not depend on the censoring distribution. We also observe differences in the estimates of the AUC when different estimators [Chambless and Diao, 2006, Heagerty and Zheng, 2005] are applied. This is due to the definition of the sensitivity. Investigators should select the appropriate estimator based on which definition of sensitivity is more scientifically relevant.

5.6 Discussion

It has been shown that under model mis-specification, the estimand corresponding to the partial likelihood estimator will depend on the censoring distribution. Several estimators, including those of Xu and O'Quigley [2000], Boyd et al. [2012], and Nguyen and Gillen [2017] have been proposed to deal with the dependence on the censoring distribution. Most work in this area has focused on inference. However, many estimators of the AUC allow for the use of the partial likelihood estimator to obtain risk scores based on multiple biomarkers, or covariates. Therefore, it is important to consider how the censoring distribution impacts the estimates of the AUC. While Viallon and Latouche [2011] show that model mis-specification can lead to biased estimates of the AUC by comparing three censoring scenarios, the focus of their manuscript was to derive a relation between the AUC and the predictiveness curve. Our manuscript focuses on the impact of the censoring distribution on estimates of the AUC. Moreover, we propose the use of censoring-robust estimators to obtain risk scores. These risk scores allow us to recover estimates of the AUC that do not depend on the censoring distribution and can be easily implemented using `CoxWeights` and `AUC.cd` in R. Moreover, these estimators perform well even when the model is specified correctly. In Chapter 6 we focus on estimation of the AUC under the nested case-control design [Thomas, 1977] and propose censoring-robust estimators analogous to those presented in this chapter.

Chapter 6

Time-Dependent Censoring-Robust Area Under the Curve Estimators under the Nested Case-Control Sampling Scheme

6.1 Introduction

As previously discussed, it is imperative that we find new biomarkers for AD. A common approach for biomarker discovery in AD is to first measure potential biomarkers in CSF to determine if these can accurately distinguish diseased from non-diseased individuals. However, as seen in Nuño et al. [2017], study participants are often unwilling to undergo lumbar punctures, so it is difficult to obtain CSF samples on all study participants. Use of the nested case-control design would allow us to use existing CSF samples efficiently by only requiring that some of these samples are processed, while still allowing us to continue the search for new AD biomarkers. In order to identify these biomarkers, however, it is also important that we have reliable statistical methods to evaluate their performance.

In Chapter 5 we presented estimators for the AUC in the time-dependent ROC curve setting. In cases like the one presented above, however, it may not be feasible to collect full covariate information on all study participants. While our motivating example comes from AD, the problem of limited resources is one that is common in most disease areas. In light of this, several estimators have been proposed to estimate the AUC under the nested case-control design, including those of Cai and Zheng [2012] and Zhou et al. [2013]. However, as seen in Chapters 3 and 4, the partial likelihood estimator under the nested case-control design is also susceptible to dependence on the censoring distribution under model mis-specification. In Chapter 5, we showed that due to the dependence on the censoring distribution, estimates of the AUC under the full cohort partial likelihood estimator (when all the data are used) depend on the censoring distribution. We hypothesized that due to this dependence, estimates of the AUC obtained using the nested case-control sampling scheme would also depend on the censoring distribution.

In this chapter, we introduce the estimator proposed by Cai and Zheng [2012]. We investigate its performance when the model for the risk score is mis-specified and propose the use of censoring-robust estimators to allow for censoring-robust estimators of the AUC. We then show the empirical performance of the estimators through simulation studies. We end with an application to ADNI data where, as before, the goal was to evaluate the classification performance of Aβ while accounting for APOE4, ADAS-11 at baseline, and MMSE at baseline.

6.2 Background

In Chapter 5, we discussed time-dependent ROC curves to evaluate the classification performance of a continuous measure when full covariate data are available for all study participants. This measure is often either a single biomarker measure or a risk score based on several biomarkers or risk factors. As previously discussed, the risk score is often obtained using a linear combination of the biomarkers along with coefficient estimates derived from the partial likelihood estimator. There are many scenarios in which it is difficult or expensive to collect full covariate information on all study participants. When the event of interest is rare, the nested case-control design provides a reduction in costs due to the fact that it only requires full covariate information on participants who experience an event and a subsample of those who are censored. While the nested case-control design is often used to conduct inference, it can also be used to investigate the classification performance of potential biomarkers.

Several estimators have been proposed for use under the nested case-control sampling scheme. The first estimator was proposed by Cai and Zheng [2011] and uses non-parametric estimators of the sensitivity and specificity. The authors later propose a semi-parametric approach [Cai and Zheng, 2012]. In this case, the sensitivity and specificity are estimated as

$$\widehat{\text{Sens}}_{CZ}(t,c) = \frac{\hat{F}^W(c) - \hat{S}^W(t,c)}{1 - \hat{S}^W(t)} \tag{6.1}$$

and

$$\widehat{\text{Spec}}_{CZ}(t,c) = 1 - \frac{\hat{S}^W(t,c)}{\hat{S}^W(t)} \tag{6.2}$$

where $\hat{S}^W(t,c) = \frac{\sum_{i=1}^N V_i/p_i \hat{S}^W(t|B_i) I(B_i > c)}{\sum_{i=1}^N V_i/p_i}$. $\hat{S}^W(t|B) = \exp\{-\hat{\Lambda}_0^W(t)\exp(\hat{\beta}^W B)\}$ where $\hat{\Lambda}_0^W(t) = \sum_{i=1}^N \frac{V_i/p_i I(X_i \leq t)\delta_i}{\sum_{j=1}^N V_j/p_j I(X_j \geq X_i)\exp(\hat{\beta} Z_j)}$, $p_i = \delta_i + (1-\delta_i)[1 - \prod_{X_l < X_i}(1 - \frac{M}{n_l - 1}\delta_l)]$ and V_i is an indicator for whether individual i was included in the nested case-control sample either as a case or as a control.

$\hat{\beta}^W$ is estimated using the inverse probability weight estimator for the nested case-control design proposed by Samuelsen [1997]. $\hat{F}^W(c) = 1 - \hat{S}^W(0,c)$ and $S(t)$ can be estimated as $\hat{S}^W(t) = \hat{S}^W(t, -\infty)$. Note that, as before, B may be a risk score made up of a combination of several markers. This is often done when comparing the predictive abilities of different models. In the full data case, for example, Chambless and Diao [2006] consider use of the Cox PH model to estimate the coefficients to create a risk score.

Several estimators have been proposed to estimate the log hazard ratio under the nested case-control sampling scheme. In this chapter, we will focus on the estimators of the AUC proposed by Cai and Zheng [2012] with risk scores obtained using coefficient estimates as proposed by Thomas [1977], Samuelsen [1997], and Nuño and Gillen [2019]. We show that when the risk score model is mis-specified, estimates of the AUC will also depend on the censoring distribution unless the risk score model is calculated using censoring-robust estimators.

6.3 Estimating the Risk Scores

As discussed in Chapter 2, when the model is mis-specified, the estimand corresponding to the partial likelihood estimator under the full data setting [Cox, 1972] will depend on the censoring distribution. Similarly, the partial likelihood estimator under the nested case-control design depends on the censoring distribution and on the number of sampled controls, as seen in Chapters 3 and 4. In the previous chapter, we showed that when the risk score model is mis-specified, the AUC estimates will also depend on the censoring distribution. Because of this, we hypothesized that under the nested case-control design, the AUC estimates would depend on the censoring distribution under model mis-specification and that the use of censoring-robust estimators such as that proposed by Nuño and Gillen [2019] would allow for censoring-robust estimation.

Recall that the estimating equation for the estimator proposed by Thomas [1977] is

$$U_{NCC}(\beta) = \sum_{i=1}^{n} \int_{t=0}^{\infty} \left\{ Z_i - \frac{S_{NCC}^{(1)}(\beta, t)}{S_{NCC}^{(0)}(\beta, t)} \right\} dN_i(t) = 0$$

where $S_{NCC}^{(r)}(\beta, t) = n^{-1} \sum_{j=1}^{n} Z_j^r \tilde{Y}_j(t) \exp(Z_j \beta)$. The risk sets only include cases and controls who are sampled for that event time.

Samuelsen [1997] proposed including subjects in the nested case-control sample at all risk sets during which they were at risk and reweighting by the inverse probability of sampling. The estimating equation under the Samuelsen [1997] estimator takes the form

$$U_S(\beta) = \sum_{i=1}^{n} \int_{t=0}^{\infty} \left\{ Z_i - \frac{S_{IPW}^{(1)}(\beta, t)}{S_{IPW}^{(0)}(\beta, t)} \right\} dN_i(t) = 0$$

where $S_{IPW}^{(r)}(\beta, t) = n^{-1} \sum_{j=1}^{n} Z_j^r Y_j^*(t) \exp(Z_j \beta) V_j / p_j$, $Y_j^*(t)$ is an indicator for whever subject j was ever sampled into the nested case-control sample and is at risk at time t, and p_j represents the probability of ever being sampled into the nested case-control sample, either as a case or as a control.

When the model is mis-specified, we have seen that the quantity estimated depends on the censoring distribution. To deal with this dependence, a censoring-robust estimator was proposed by Nuño and Gillen [2019], for which the estimating equation takes the form

$$\tilde{U}_{HD}(\beta) = \sum_{i=1}^{n} \int_{t=0}^{\infty} \tilde{W}(t|Z_1 = z_{1i}) \left\{ \vec{Z}_i - \frac{\tilde{S}_{HD}^{(1)}(\beta, t)}{\tilde{S}_{HD}^{(0)}(\beta, t)} \right\} dN_i(t) = 0 \tag{6.3}$$

where $\tilde{S}^{(r)}_{HD} = n^{-1} \sum_{j=1}^{n} \frac{\tilde{W}(t|Z_1 = z_{1j}) \hat{n}_{z_{1j}}(t)}{\sum_{k=1}^{n} \tilde{Y}_k(t)\{z_{1k} z_{1j} + (1 - z_{1k})(1 - z_{1j})\}} \vec{Z}_j^r \tilde{Y}_j(t) \exp(\vec{Z}_j \beta)$, $\vec{Z}_j$ is the vector of covariates, z_{1j} is the value of the predictor of interest for subject j and $\hat{n}_{z_{1j}}(t)$ is the estimated number of subjects at risk with the same covariate value as z_{1j} at time t. This estimator requires imputation of the covariate values for subjects in the full cohort who were not sampled into the original nested case-control sample. While the imputation method can be selected by the user, one possible method is hotdeck imputation. Note that to recover the FC estimator, we can set $\tilde{W}(t|Z) = 1$.

In the following section, we show the performance of the estimators presented here under model mis-specification. We then show how mis-specification of the risk score model effects estimates of the AUC under the Cai and Zheng [2012] estimators.

6.4 Empirical Performance

We first present the performance of the four estimators when estimating the coefficients under a mis-specified model. The simulation studies used were the same as those in Chapter 5. This time, we performed 1,000 simulations each with 1,500 subjects (750 with $z_2 = 0$ and 750 with $z_2 = 1$). As before, the true hazard function takes the form $\lambda(t) = \exp\Bigg\{ \log(0.5) + \log(0.85) Z_1 + \log(0.05) \cdot Z_2 \cdot I(t \leq 3) + \log(2) \cdot Z_2 \cdot I(3 < t \leq 6) + \log(1) \cdot Z_2 \cdot I(t > 6) \Bigg\}$ with change points at times $t = 3$ and $t = 6$. We again applied different censoring distributions following the cumulative density function $F_C(c) = (c/7)^r$, with r taking values from 0.25 to 4. Although we had a time-varying effect, the risk score was calculated using a linear combination of the form $\hat{\beta}_1 Z_1 + \hat{\beta}_2 Z_2$ where the coefficient estimates were obtained from a model of the form $\hat{\lambda}(t) = \hat{\lambda}_0(t) \exp(\hat{\beta}_1 Z_1 + \hat{\beta}_2 Z_2)$. The Thomas, Samuelson, and the proposed estimators (full cohort and censoring-robust) from Chapter 3 were used to calculate the risk score. These risk scores were then used to estimate the time-dependent AUC.

6.4.1 Coefficient Estimates

Figure 6.1 presents the coefficient estimates under $M = 1$ to 4 controls under four different estimators. Note that we estimate a different quantity depending on the censoring distribution. Moreover, the Thomas [1977] estimator also leads to dependence on the number of controls sampled at each event time. When $M = 1$, the coefficient estimate averages approximately -1.41. When $M = 4$, the estimated log hazard ratio is approximately -1.3. If the estimators of Samuelsen [1997] and Nuño and Gillen [2019] are used, we observe less of a dependence on the number of controls sampled, but we still observe the dependence on the censoring distribution. When $r = 0.25$, the average coefficient estimate is approximately -1.25, and approximately -0.65 when $r = 4$ is used.

When the censoring-robust estimator is applied (bottom right), on the other hand, we obtain estimates that do not depend on the censoring distribution. It should be noted that the Samuelsen [1997] estimator can be reweighted to account for the censoring distribution using the weights proposed in Nuño and Gillen [2019]. Doing this also provides estimates that do not differ according to the censoring distribution.

6.5 Estimates of the area under the curve

In this section, we consider estimates of the AUC when the risk scores are obtained using the four estimators from Section 6.3. We consider estimates of the AUC at times $t = 1.5, 4.5$, and 5.5. As with the coefficient estimates, we find that estimates of the AUC also depend on the censoring distribution. If we consider the Thomas [1977] estimator, we find that when $r = 0.25$ and $M = 1$, the AUC is estimated to be approximately 0.74. When we instead have $r = 4$, the AUC is estimated to be approximately 0.65. We observe similar results under the Samuelsen [1997] estimator and the estimator proposed by Nuño and Gillen [2019]

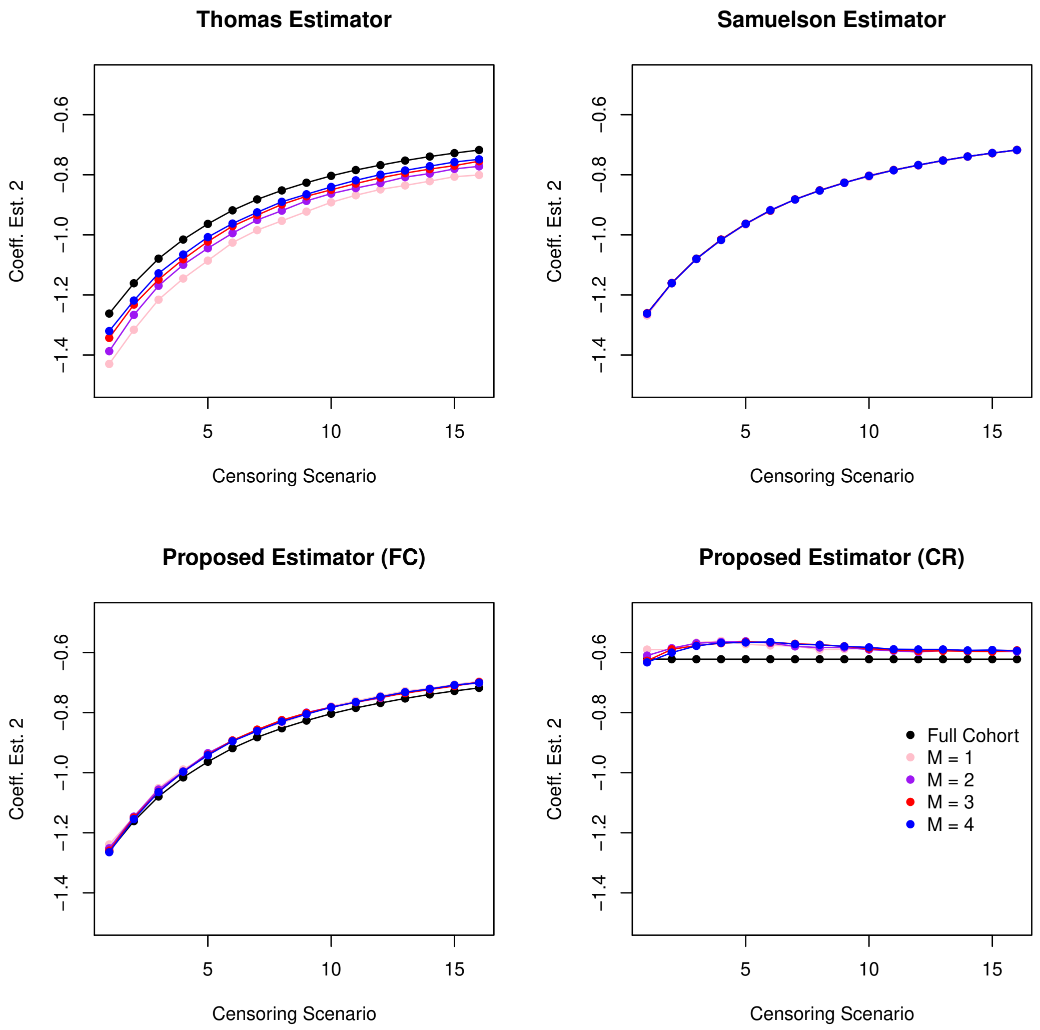


Figure 6.1: Coefficient estimates for Z_2 based on 1,000 simulations with 1,500 subjects each. These results were obtained under model mis-specification, where a time-varying effect exists, but it is assumed to be time invariant.

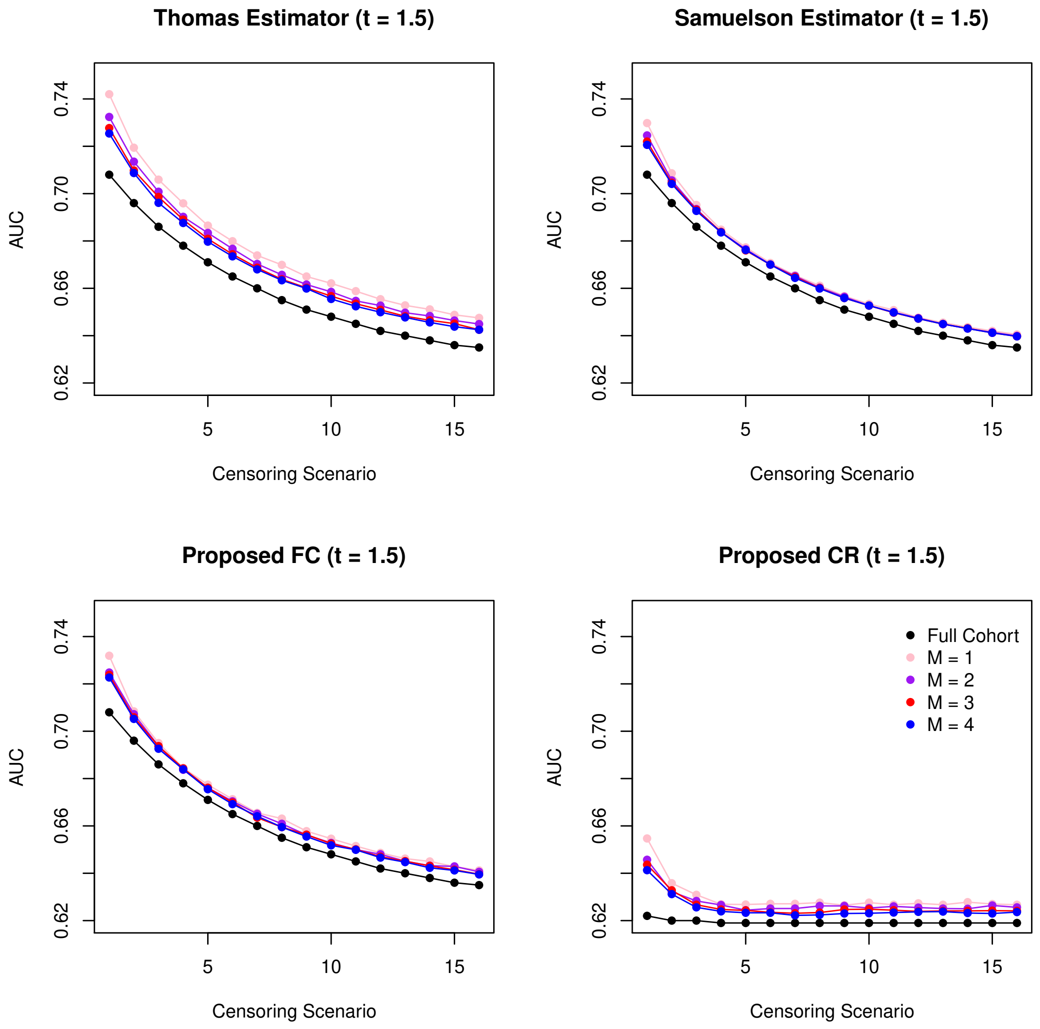


Figure 6.2: Estimates of the AUC under various nested case-control estimators and for different numbers of sampled controls based on 1,000 simulations with 1,500 subjects each.

to recover the full cohort partial likelihood estimator. Notice that because of how the risk sets are formed under the Thomas [1977] estimator, the estimates of the AUC also depend on M. Considering again $r = 0.25$, we find that when $M = 1$, the average estimated AUC is approximately 0.74. When $M = 4$, the average estimated AUC is approximately 0.725. While these differences are small, we do find that there is less dependence on M when the Samuelsen [1997] or full cohort Nuño and Gillen [2019] estimators are used. We observe similar patterns at times $t = 4.5$ and $t = 5.5$.

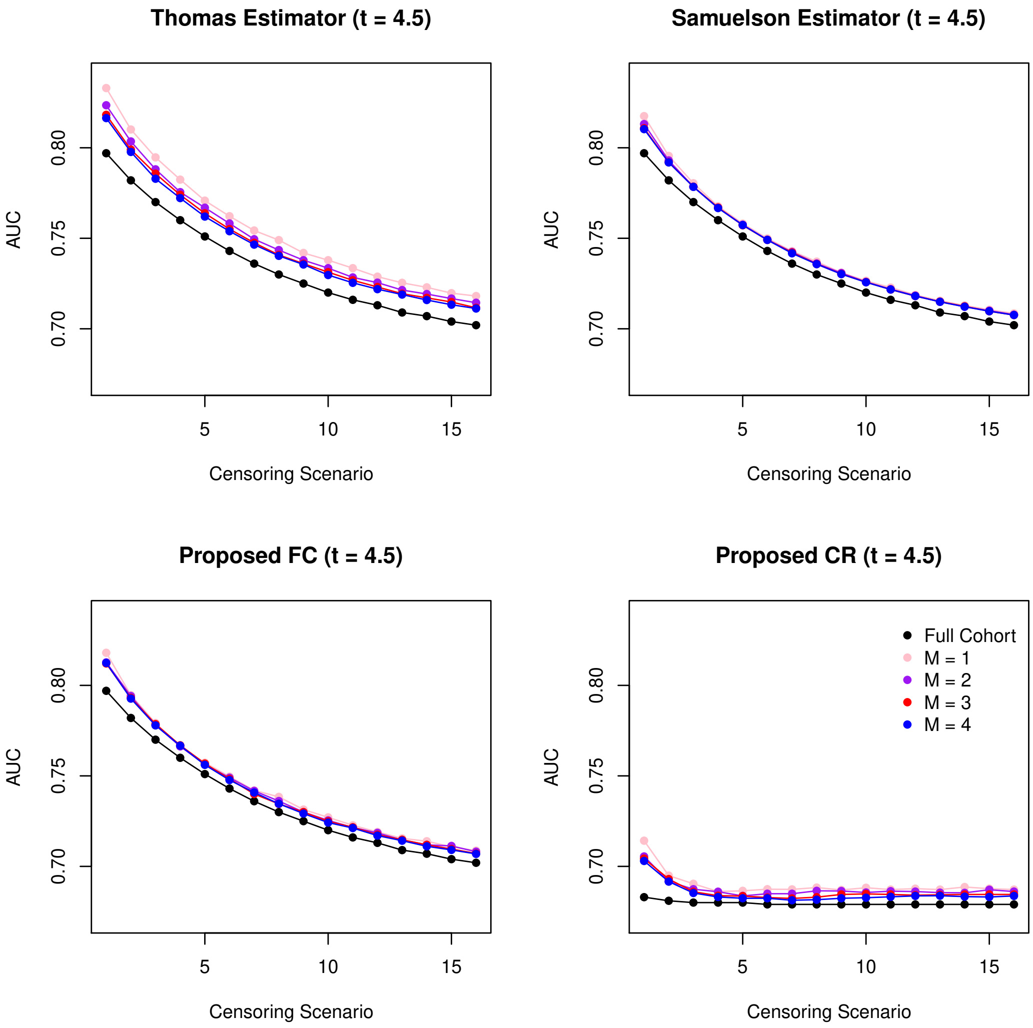


Figure 6.3: Estimates of the AUC under various nested case-control estimators and for different numbers of sampled controls based on 1,000 simulations with 1,500 subjects each.

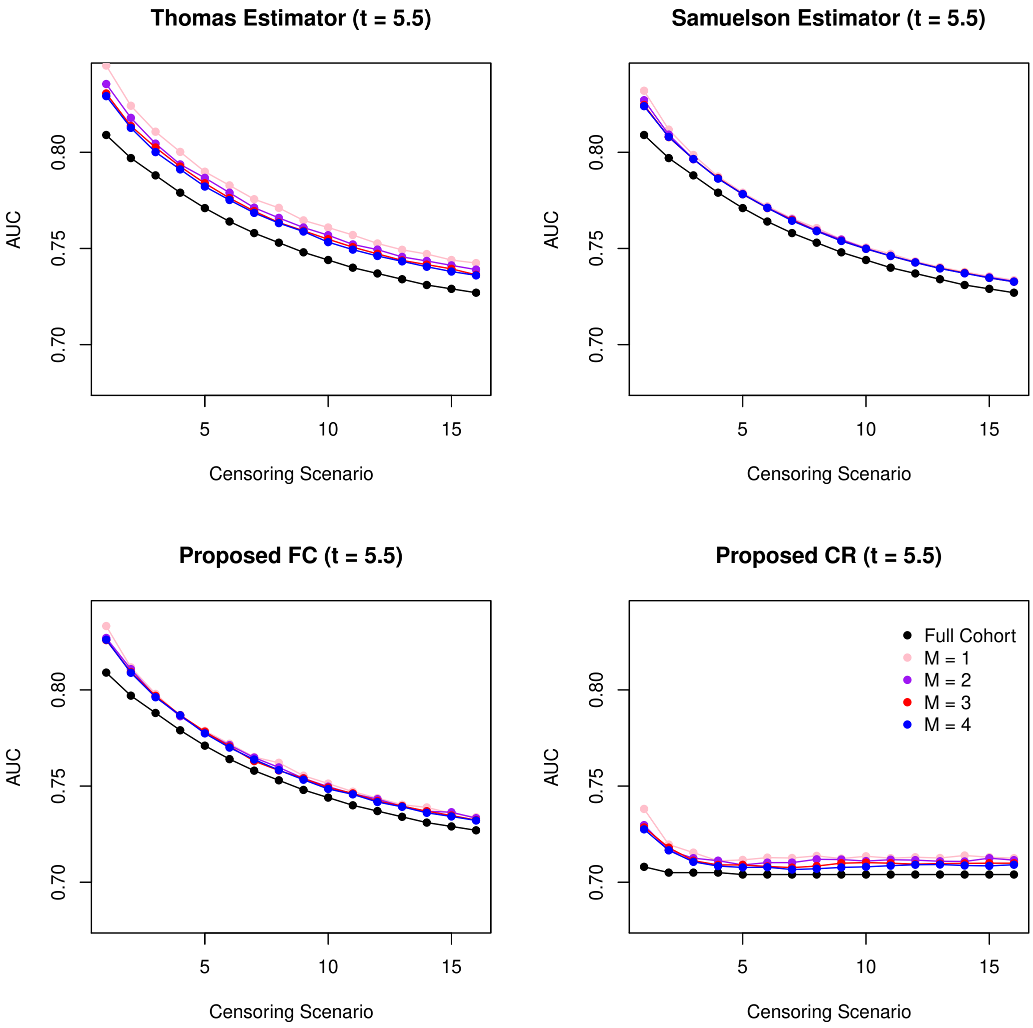


Figure 6.4: Estimates of the AUC under various nested case-control estimators and for different numbers of sampled controls based on 1,000 simulations with 1,500 subjects each.

We now consider the censoring-robust (CR) estimator at time $t = 1.5$. The estimates from the full cohort in this scenario are based on risk scores using the estimator proposed by Boyd et al. [2012]. Notice that we are able to obtain much more stable estimates when the censoring-robust estimator of Nuño and Gillen [2019] is applied to the nested case-control sample. For the first three scenarios ($r = 0.25$, 0.5, and 0.75), we find that we are not able to fully recover the estimates obtained from the full data. However, the censoring-robust estimator does provide much more stable results compared to the other estimators. The same is true when we consider estimates of the AUC at times $t = 4.5$ and 5.5.

When the model is correctly specified, however, we obtain similar estimates regardless of the censoring distribution. This holds regardless of the estimator used.

6.6 Evaluating the Classification Performance of Aβ and APOE 4 Under the Nested Case-Control Design

We again consider the ADNI example presented in Section 5.5. Recall that in AD studies, it is often difficult to obtain CSF samples, so use of the nested case-control design would limit the number of CSF samples that need to be processed. As before, we were interested in investigating the classification performance of Aβ when APOE 4, MMSE at baseline, and ADAS 11 at baseline are also available. Considering the fact that CSF samples may be limited, we used the time-dependent ROC curve estimator of Cai and Zheng [2012] along with a nested case-control sample to calculate the AUC. Risk scores were calculated using coefficient estimates obtained using the estimators of Thomas [1977], Samuelsen [1997], and Nuño and Gillen [2019]. Our data set included the same subjects as those in Table 5.1. To obtain the estimates in Table 6.1, we drew 100 nested case-control samples from the full cohort data set for each number of controls ($M = 1$, 2, 3, and 4).

We assumed that baseline MMSE and ADAS 11 were available for all participants, even if they were not included in the nested case-control sample. However, Aβ and APOE 4 status were only available for participants who experienced an event or were sampled as controls. Recall that our proposed estimator requires imputation of covariate values for subjects not in the original nested case-control sample. While the standard nested case-control method only includes participants in risk sets for which they were sampled as controls, we included all subjects in the nested case-control sample in all risk sets for which they were still at risk in the full cohort study. For the remainder of the participants (those not included in the sample), we used a hotdeck imputation approach where we matched to participants who were still at risk and had been sampled at or before the current event time. Matching was based on MMSE and ADAS 11 at baseline using Mahalanobis distance. To obtain the censoring-robust estimator, the censoring distribution was estimated as in Section 3.2.4.

As before, the largest difference between the full cohort and censoring-robust estimators was observed for Aβ. The log hazard ratio was estimated to be 1.3 under the partial likelihood estimator and 0.969 under the censoring-robust estimator of Boyd et al. [2012].

We also observed a difference in the coefficient estimates obtained using the Thomas [1977] estimator, where the log hazard ratio was estimated to be approximately 0.3 for APOE 4 compared to 0.22 under the full cohort. The coefficient estimates for ADAS 11 and MMSE did not differ by the number of controls and were similar for all estimators although the estimates obtained using the Thomas [1977] estimator did differ slightly. When the censoring-robust estimator of Nuño and Gillen [2019] was used, the estimates were much closer to those obtained using the Boyd et al. [2012] estimator.

We also noticed a large difference in the AUC estimates when using the censoring-dependent estimators compared to the censoring-robust estimators. Under the censoring-dependent estimators, the AUC was estimated to be approximately 0.80. Notice that the estimates of the AUC were similar, although slightly higher, when using the nested case-control estimators compared to the full cohort estimators. When the risk score was estimated using the censoring-robust estimators, the estimates of the AUC decreased to approximately 0.73 under the full cohort. Under the nested case-control sample, the estimates were a little larger than those obtained using the full cohort data. The largest difference was observed when $M = 1$, where the estimated AUC at four years was estimated to be 0.758 (compared to 0.735 under the full cohort). However, it should be noted that $M = 1$ is rarely used in practice.

		Estimated Log Hazard Ratio (HR)				**AUC**			
	N	APOE 4	ADAS 11 (bl)	Aβ < 192	MMSE (bl)	1 yr.	2 yrs.	3 yrs.	4 yrs.
Censoring-Dependent									
Full Cohort	973	0.220 (1.25)	0.123 (1.13)	1.305 (3.69)	-0.096 (0.91)	0.784	0.787	0.791	0.795
Thomas									
M = 1	303.71	0.332 (1.39)	0.152 (1.16)	1.303 (3.68)	-0.064 (0.94)	0.813	0.814	0.817	0.820
M = 2	403.54	0.318 (1.37)	0.148 (1.16)	1.346 (3.84)	-0.083 (0.92)	0.812	0.814	0.816	0.819
M = 3	483.84	0.318 (1.37)	0.150 (1.16)	1.294 (3.65)	-0.083 (0.92)	0.811	0.813	0.815	0.818
M = 4	546.20	0.300 (1.35)	0.142 (1.15)	1.316 (3.73)	-0.086 (0.92)	0.806	0.808	0.811	0.814
Samuelson									
M = 1	303.71	0.217 (1.24)	0.137 (1.15)	1.309 (3.70)	-0.093 (0.91)	0.800	0.802	0.805	0.809
M = 2	403.54	0.217 (1.24)	0.127 (1.14)	1.329 (3.78)	-0.101 (0.90)	0.790	0.793	0.796	0.801
M = 3	483.84	0.231 (1.26)	0.129 (1.14)	1.286 (3.62)	-0.098 (0.91)	0.790	0.792	0.796	0.800
M = 4	546.20	0.213 (1.24)	0.125 (1.13)	1.315 (3.72)	-0.097 (0.91)	0.788	0.791	0.795	0.799
Proposed									
M = 1	303.71	0.261 (1.30)	0.131 (1.14)	1.278 (3.59)	-0.101 (0.90)	0.798	0.800	0.804	0.808
M = 2	403.54	0.263 (1.30)	0.127 (1.14)	1.294 (3.65)	-0.097 (0.91)	0.791	0.793	0.797	0.801
M = 3	483.84	0.254 (1.29)	0.127 (1.14)	1.261 (3.53)	-0.097 (0.91)	0.789	0.791	0.795	0.799
M = 4	546.20	0.238 (1.27)	0.126 (1.13)	1.287 (3.62)	-0.098 (0.91)	0.789	0.792	0.796	0.800
Censoring-Robust									
Full Cohort	973.00	0.535 (1.71)	0.085 (1.09)	0.969 (2.64)	0.011 (1.01)	0.721	0.725	0.729	0.735
Proposed									
M = 1	303.71	0.519 (1.68)	0.090 (1.09)	1.07 (2.92)	-0.011 (0.99)	0.744	0.747	0.752	0.758
M = 2	403.54	0.580 (1.79)	0.086 (1.09)	0.989 (2.69)	0.019 (1.02)	0.731	0.734	0.739	0.745
M = 3	483.84	0.537 (1.71)	0.089 (1.09)	0.929 (2.53)	0.018 (1.02)	0.725	0.729	0.733	0.738
M = 4	546.20	0.543 (1.72)	0.086 (1.09)	0.956 (2.60)	0.014 (1.01)	0.726	0.730	0.734	0.740

Table 6.1: Coefficient and AUC estimates were obtained using 100 nested case-control samples from the ADNI data. Estimates were obtained using various nested case-control estimators along with the estimators proposed by Cai and Zheng [2012]. Full cohort estimates were obtained using the partial likelihood estimator (PLE) and the censoring-robust estimator of Boyd et al. [2012] with the Chambless and Diao [2006] estimator.

6.7 Discussion

Biomarker discovery is crucial in various disease areas, including in AD. In order to discover new biomarkers, however, it is important that we are able to carefully assess the classification performance of these biomarkers which requires reliable statistical methods. When determining whether a biomarker is able to contribute additional information compared to existing biomarkers, one may decide to compare the performance of several risk scores. It is common to calculate these risk scores using a linear combination of several biomarkers using coefficient estimates obtained from a survival model.

In this chapter, we found that when the risk score model is mis-specified, the AUC estimates will depend on the censoring distribution due to dependence of the coefficient estimates on the censoring distribution. This is concerning because classification performance of these biomarkers does not depend on patient accrual and dropout patterns, which often differ from one study to the next. It is therefore important that we are able to estimate the same quantity regardless of the censoring distribution. We propose the use of censoring-robust estimators when estimating the risk scores since these estimate the same quantity for the AUC, regardless of the censoring distribution.

Chapter 7

Conclusion

Throughout this dissertation, we have seen the utility of the nested case-control sampling scheme and its performance under model mis-specification. When the model is mis-specified, commonly used estimators yield results that depend not only on the censoring distribution (affecting replicability), but also on the number of sampled controls (affecting reproducibility within a study). We have proposed three estimators for conducting inference under the nested case-control sampling scheme. We have also investigated the performance of time-dependent ROC curves both in the full data and the nested case-control setting and found that when the model used to obtain the risk score is mis-specified, estimates of the AUC depend on the censoring distribution. In the nested case-control setting, they may also depend on the number of sampled controls. To fix this, we proposed estimating the risk scores using censoring-robust estimators so that estimates of classification performance are not impacted by the censoring distribution.

There are many areas in which the nested case-control design would prove useful, but which could benefit from new methodology or extensions to existing methods. In this final chapter, we discuss areas of future work. In particular, we consider extensions of joint longitudinal survival and the analysis of recurrent events under the nested case-control sampling scheme.

7.1 Future work

7.1.1 Joint Longitudinal Survival Models under the Nested Case-Control Design

In many cases, we may be interested in investigating the association between the time to an event of interest and a longitudinal variable. While measures of these variables are often changing, they are only observed (or measured) at specific time points. A common approach is to use the last observation carried forward, which assumes that these measurements remain constant until the next time they are collected. This is not often the case and assuming it is could lead to biased results [Prentice, 1982]. In these settings, a common approach is the use of joint longitudinal survival models. These models estimate the trajectory of the longitudinal marker and its association to the event of interest.

Various methods have been proposed for joint modeling of longitudinal covariates in the survival setting including the work of Tsiatis et al. [1995], Faucett and Thomas [1996], Wulfsohn and Tsiatis [1997], Dafni and Tsiatis [1998], Tsiatis and Davidian [2001] among many others. These include two-stage and likelihood based methods. In the two-stage methods, a linear mixed effects model is often used to estimate the covariate values at each event time. These estimates are then used to fit a survival model [Wu et al., 2012]. Under the likelihood approach, the coefficient estimates are obtained using the likelihoods of the longitudinal covariates as well as the likelihood of the survival data. In this approach, the longitudinal measures and the coefficient estimates are obtained simultaneously.

While there exists extensive literature about joint longitudinal survival models in the full cohort data setting, literature in the nested case-control setting is scarce. Tseng and Liu [2009] propose a joint longitudinal survival model using shared latent parameters and maximum likelihood estimation. Their proposed methodology allows for adjustment of other covariates when estimating the longitudinal trajectory. Under the nested case-control sampling scheme some covariate measures, such as demographic information, are available for all study participants regardless of whether they were sampled into the nested case-control sample. The estimator proposed by Tseng and Liu [2009] allows for inclusion of these participants in the likelihood of the survival data, therefore allowing the use of all available information. The authors propose the use of the E-M algorithm to solve for the coefficient estimates.

As we have discussed, the standard nested case-control design randomly samples a certain number of controls at each event time. Therefore, participants who are in the study longer have a higher probability of being sampled. In the joint longitudinal survival setting, this means that people with longer follow-up will be over-represented when estimating the longitudinal trajectory, which could lead to biased results. To account for oversampling of participants with longer follow-up, SAN [2013] propose a weighted likelihood approach similar to that of Tseng and Liu [2009], but which reweights contributions by the inverse probability of sampling when estimating coefficients.

While existing methods provide a great option that allows for reduction of costs even when longitudinal measurements are collected, there are several ways in which existing methods can be improved. Under the standard nested case-control design, as proposed by Thomas [1977] and used by Tseng and Liu [2009] and SAN [2013], controls are randomly sampled without replacement at each event time. However, it is common for some participants to have different numbers of measurements as well as different lengths of time between measurements. In scenarios where the nested case-control design would prove most beneficial, biospecimen samples (such as CSF) are collected and only processed if they are needed due to the high expenses associated with processing these samples. Therefore, we know how many measurements people have available for sampling, as well as the length of time between measurements. In these settings, we could benefit from optimally sampling controls to individuals likely to provide more information about the association of interest and the longitudinal trajectory. One possibility is to consider weighted sampling of controls in which controls are sampled with probability proportional to their overall contribution. During estimation of the coefficients, we would then reweight by the inverse probability of sampling.

Another consideration is the performance of joint longitudinal survival models under model mis-specification. As seen throughout this dissertation, under model mis-specification the nested case-control sampling scheme and use of the partial likelihood estimator can lead to dependence on the censoring distribution when estimating model parameters. The same problems may arise when considering joint longitudinal survival models. Therefore, we also propose to investigate the performance of these models under the nested case-control sampling scheme and to provide robust methods.

7.1.2 Recurrent Event Analysis under the Nested Case-Control Design

The methods presented throughout this dissertation have focused on a single outcome. However, in many disease areas including cancer, an individual may experience recurring events. There are a variety of ways to model recurrent event data in the full cohort setting. Commonly used models include Poisson regression, extensions of the Cox proportional hazards model [Andersen and Gill, 1982, Prentice et al., 1981], and frailty models. Under Poisson regression, we model the number of recurrent events. Another way to model recurrent events is through the use of frailty models, which include random effects, or frailties, to account for differences between individuals [Amorim and Cai, 2015]. In this setting, the full likelihood must be specified, including the distribution of the random effects. The Andersen-Gill and

Prentice et al. [1981] models are semi-parametric and do not require specification of the baseline intensity function. When using the Andersen-Gill model and the total time model of Prentice et al. [1981], individuals with more than one event provide multiple observations, where each observation starts after the individual's previous event. The Anderson-Gill model assumes that the baseline intensity is the same regardless of the event number (how many previous events an individual has had), while the model proposed by Prentice et al. [1981] stratifies by the event number to allow for different baseline intensity functions. As previously stated, one benefit of using these models is that the baseline intensity function does not need to be specified. While recurrent event analysis has been greatly studied in the full cohort setting, little work has been done for recurrent event analysis under the the nested case-control design. In this setting, Jazić et al. [2019] propose the use of joint frailty models for recurrent events subject to a terminal event. They present five designs for sampling controls in a similar fashion to the nested case-control design. These include sampling controls at the first observed recurrent event, at the terminal event, or modifications of these. Under the proposed estimator, however, one needs to fully specify the likelihood, including the distribution of the frailty terms and the baseline intensity functions.

To avoid modeling the baseline intensity functions and distribution of the frailty terms, we may consider use of a semi-parametric approach such as that proposed by Prentice et al. [1981] for the full cohort setting. Under the nested case-control design, this would require reweighting contributions to the partial likelihood by the inverse probability of sampling to account for the sampling under the nested case-control design. This is analogous to the estimator of Samuelsen [1997] when modeling a single event, or time to the first event. In their work, Jazić et al. [2019] also mention the possibility of oversampling controls that provide more information about the frailty parameters. A similar idea could be used in the proposed methodology by oversampling controls who experience more events and reweighting by the inverse probability of sampling weights.

Bibliography

Odd Aalen. Nonparametric inference for a family of counting processes. *The Annals of Statistics*, pages 701–726, 1978.

Leila DAF Amorim and Jianwen Cai. Modelling recurrent events: a tutorial for analysis in epidemiology. *International journal of epidemiology*, 44(1):324–333, 2015.

Per Kragh Andersen and Richard D Gill. Cox's regression model for counting processes: a large sample study. *The annals of statistics*, pages 1100–1120, 1982.

Niels Andreasen, Eugeen Vanmechelen, Hugo Vanderstichele, Pia Davidsson, and Kaj Blennow. Cerebrospinal fluid levels of total-tau, phospho-tau and aβ42 predicts development of alzheimer's disease in patients with mild cognitive impairment. *Acta Neurologica Scandinavica*, 107(s179):47–51, 2003.

Alzheimer's Association. 2016 alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 12(4):459–509, 2016.

Alzheimer's Association et al. 2020 alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 16(3):391–460, 2020.

William E Barlow. Robust variance estimation for the case-cohort design. *Biometrics*, pages 1064–1072, 1994.

William E Barlow, Laura Ichikawa, Dan Rosner, and Shizue Izumi. Analysis of case-cohort designs. *Journal of clinical epidemiology*, 52(12):1165–1172, 1999.

Paul Blanche, Jean-François Dartigues, and Hélène Jacqmin-Gadda. Review and comparison of roc curve estimators for a time-dependent outcome with marker-dependent censoring. *Biometrical Journal*, 55(5):687–704, 2013.

K Blennow. Csf biomarkers for mild cognitive impairment. *Journal of internal medicine*, 256(3):224–234, 2004.

Kaj Blennow. Csf biomarkers for alzheimer's disease: use in early diagnosis and evaluation of drug treatment. *Expert review of molecular diagnostics*, 5(5):661–672, 2005.

Kaj Blennow, Harald Hampel, Michael Weiner, and Henrik Zetterberg. Cerebrospinal fluid and plasma biomarkers in alzheimer disease. *Nature Reviews Neurology*, 6(3):131, 2010.

Ornulf Borgan, Bryan Langholz, Sven Ove Samuelsen, Larry Goldstein, and Janice Pogoda. Exposure stratified case-cohort designs. *Lifetime data analysis*, 6(1):39–58, 2000.

Adam P Boyd, John M Kittelson, and Daniel L Gillen. Estimation of treatment effect under non-proportional hazards and conditionally independent censoring. *Statistics in medicine*, 31(28):3504–3515, 2012.

Norman Breslow, John Crowley, et al. A large sample study of the life table and product limit estimates under random censorship. *The Annals of Statistics*, 2(3):437–453, 1974.

Norman E Breslow. Contribution to discussion of paper by dr cox. *J. Roy. Statist. Soc., Ser. B*, 34:216–217, 1972.

Tianxi Cai and Yingye Zheng. Nonparametric evaluation of biomarker accuracy under nested case-control studies. *Journal of the American Statistical Association*, 106(494):569–580, 2011.

Tianxi Cai and Yingye Zheng. Evaluating prognostic accuracy of biomarkers in nested case–control studies. *Biostatistics*, 13(1):89–100, 2012.

Lloyd E Chambless and Guoqing Diao. Estimation of time-dependent area under the roc curve for long-term risk prediction. *Statistics in medicine*, 25(20):3474–3486, 2006.

EH Corder, AM Saunders, WJ Strittmatter, DE Schmechel, PC Gaskell, GWet al Small, AD Roses, JL Haines, and Margaret A Pericak-Vance. Gene dose of apolipoprotein e type 4 allele and the risk of alzheimer's disease in late onset families. *Science*, 261(5123): 921–923, 1993.

David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

Urania G Dafni and Anastasios A Tsiatis. Evaluating surrogate markers of clinical outcome when measured with error. *Biometrics*, pages 1445–1462, 1998.

Cox R David et al. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society*, 34(2):187–220, 1972.

Adrianus Dingeman de Groot. The meaning of "significance" for different types of research [translated and annotated by eric-jan wagenmakers, denny borsboom, josine verhagen, rogier kievit, marjan bakker, angelique cramer, dora matzke, don mellenbergh, and han lj van der maas]. *Acta psychologica*, 148:188–194, 2014.

Paul Denver and Paula L McClean. Distinguishing normal brain aging from the development of alzheimer's disease: inflammation, insulin signaling and cognition. *Neural regeneration research*, 13(10):1719, 2018.

Rishi J Desai, Robert J Glynn, Shirley Wang, and Joshua J Gagne. Performance of disease risk score matching in nested case-control studies: a simulation study. *American journal of epidemiology*, page kwv269, 2016.

Frank Eriksson, Torben Martinussen, and Søren Feodor Nielsen. Large sample results for frequentist multiple imputation for cox regression with missing covariate data. *Annals of the Institute of Statistical Mathematics*, pages 1–28, 2019.

Virginia L Ernster. Nested case-control studies. *Preventive medicine*, 23(5):587–590, 1994.

Cheryl L Faucett and Duncan C Thomas. Simultaneously modelling censored survival data and repeatedly measured covariates: a gibbs sampling approach. *Statistics in medicine*, 15(15):1663–1685, 1996.

Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

Ivan P Fellegi and David Holt. A systematic approach to automatic edit and imputation. *Journal of the American Statistical association*, 71(353):17–35, 1976.

Thomas R Fleming and David P Harrington. *Counting processes and survival analysis*, volume 169. John Wiley & Sons, 2011.

Marshal F Folstein, Susan E Folstein, and Paul R McHugh. "mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *Journal of psychiatric research*, 12(3):189–198, 1975.

Andrew Gelman and Eric Loken. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*, 2013.

Larry Goldstein and Bryan Langholz. Asymptotic theory for nested case-control sampling in the cox regression model. *The Annals of Statistics*, pages 1903–1928, 1992.

Major Greenwood et al. A report on the natural duration of cancer. *A Report on the Natural Duration of Cancer.*, (33), 1926.

Timo Grimmer, Matthias Riemenschneider, Hans Förstl, Gjermund Henriksen, William E Klunk, Chester A Mathis, Tohru Shiga, Hans-Jürgen Wester, Alexander Kurz, and Alexander Drzezga. Beta amyloid in alzheimer's disease: increased deposition in brain is reflected in reduced concentration in cerebrospinal fluid. *Biological psychiatry*, 65(11):927–934, 2009.

Sebastien Haneuse, Jonathan Schildcrout, and Daniel Gillen. A two-stage strategy to accommodate general patterns of confounding in the design of observational studies. *Biostatistics*, 13(2):274–288, 2012.

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

Patrick J Heagerty, Thomas Lumley, and Margaret S Pepe. Time-dependent roc curves for censored survival data and a diagnostic marker. *Biometrics*, 56(2):337–344, 2000.

Daniel F Heitjan and Donald B Rubin. Ignorability and coarse data. *The annals of statistics*, pages 2244–2253, 1991.

Charles P Hughes, Leonard Berg, Warren Danziger, Lawrence A Coben, and Ronald L Martin. A new clinical scale for the staging of dementia. *The British journal of psychiatry*, 140(6):566–572, 1982.

Christian Humpel. Identifying and validating biomarkers for alzheimer's disease. *Trends in biotechnology*, 29(1):26–32, 2011.

Hung Hung and Chin-tsang Chiang. Optimal composite markers for time-dependent receiver operating characteristic curves with censored survival data. *Scandinavian journal of statistics*, 37(4):664–679, 2010.

John PA Ioannidis. Why most published research findings are false. *PLoS medicine*, 2(8): e124, 2005.

Ina Jazić, Sebastien Haneuse, Benjamin French, Gaëtan MacGrogan, and Virginie Rondeau. Design and analysis of nested case–control studies for recurrent events subject to a terminal event. *Statistics in medicine*, 38(22):4348–4362, 2019.

John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*, volume 360. John Wiley & Sons, 2011.

Adina Najwa Kamarudin, Trevor Cox, and Ruwanthi Kolamunnage-Dona. Time-dependent roc curve analysis in medical research: current methods and applications. *BMC medical research methodology*, 17(1):53, 2017.

Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

Ruth H Keogh and David Roxbee Cox. *Case-control studies*, volume 4. Cambridge University Press, 2014.

John P Klein and Melvin L Moeschberger. *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media, 2005.

Michal Kulich and DY Lin. Improving the efficiency of relative-risk estimation in case-cohort studies. *Journal of the American Statistical Association*, 99(467):832–844, 2004.

Bryan Langholz and Ørnulf Borgan. Counter-matching: a stratified nested case-control sampling method. *Biometrika*, 82(1):69–79, 1995.

Bryan Langholz and David Clayton. Sampling strategies in nested case-control studies. *Environmental Health Perspectives*, 102(Suppl 8):47, 1994.

Bryan Langholz and Duncan C Thomas. Nested case-control and case-cohort methods of sampling from a cohort: a critical comparison. *American Journal of Epidemiology*, 131 (1):169–176, 1990.

Bryan Langholz and Duncan C Thomas. Efficiency of cohort sampling designs: Some surprising results. *Biometrics*, pages 1563–1571, 1991.

Danyu Y Lin and Lee-Jen Wei. The robust inference for the cox proportional hazards model. *Journal of the American statistical Association*, 84(408):1074–1078, 1989.

DY Lin and Zhiling Ying. Cox regression with incomplete covariate measurements. *Journal of the American Statistical Association*, 88(424):1341–1349, 1993.

Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2014.

Chia-Chen Liu, Takahisa Kanekiyo, Huaxi Xu, and Guojun Bu. Apolipoprotein e and alzheimer disease: risk, mechanisms and therapy. *Nature Reviews Neurology*, 9(2):106, 2013.

Jay H Lubin and Mitchell H Gail. Biased selection of controls for case-control analyses of cohort studies. *Biometrics*, pages 63–75, 1984.

Prasanta Chandra Mahalanobis. On the generalized distance in statistics. National Institute of Science of India, 1936.

Richard Mayeux. Biomarkers: potential uses and limitations. *NeuroRx*, 1(2):182–188, 2004.

Harvey J Motulsky. Common misconceptions about data analysis and statistics. *British journal of pharmacology*, 172(8):2126–2132, 2015.

Wayne Nelson. Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14(4):945–966, 1972.

Vinh Q Nguyen and Daniel L Gillen. Censoring-robust estimation in observational survival studies: Assessing the relative effectiveness of vascular access type on patency among end-stage renal disease patients. *Statistics in Biosciences*, pages 1–25, 2012.

Vinh Q Nguyen and Daniel L Gillen. Censoring-robust estimation in observational survival studies: Assessing the relative effectiveness of vascular access type on patency among end-stage renal disease patients. *Statistics in Biosciences*, 9(2):406–430, 2017.

Michelle M. Nuño and Daniel L. Gillen. On estimation in the nested case-control design under non-proportional hazards. *Under Review*, 2019.

Michelle M. Nuño, Daniel L. Gillen, Kulwant K. Dosanjh, Lijie Di, Jenny Kotlerman, David Elashoff, Ringman John, and Joshua D. Grill. Attitudes towards clinical trials across the alzheimer's disease spectrum. *Alzheimer's Research & Therapy (In Press)*, 2017.

Michelle M Nuño, Daniel L Gillen, Kulwant K Dosanjh, Jenny Brook, David Elashoff, John M Ringman, and Joshua D Grill. Attitudes toward clinical trials across the alzheimer's disease spectrum. *Alzheimer's research & therapy*, 9(1):81, 2017.

Sid E O'Bryant, Stephen C Waring, C Munro Cullum, James Hall, Laura Lacritz, Paul J Massman, Philip J Lupo, Joan S Reisch, and Rachelle Doody. Staging dementia using clinical dementia rating scale sum of boxes scores: a texas alzheimer's research consortium study. *Archives of neurology*, 65(8):1091–1095, 2008.

N Charlotte Onland-Moret, Yvonne T van der Schouw, Wim Buschers, Sjoerd G Elias, Carla H van Gils, Jeroen Koerselman, Mark Roest, Diederick E Grobbee, Petra HM Peeters, et al. Analysis of case-cohort data: a comparison of different methods. *Journal of clinical epidemiology*, 60(4):350–355, 2007.

Ross L Prentice. Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*, 69(2):331–342, 1982.

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

Ross L Prentice, Benjamin J Williams, and Arthur V Peterson. On the regression analysis of multivariate failure time data. *Biometrika*, 68(2):373–379, 1981.

James M Robins, Mitchell H Gail, and Jay H Lubin. More on" biased selection of controls for case-control analyses of cohort studies". *Biometrics*, pages 293–299, 1986.

Wilma G Rosen, Richard C Mohs, and Kenneth L Davis. A new rating scale for alzheimer's disease. *The American journal of psychiatry*, 1984.

Sven Ove Samuelsen. A psudolikelihood approach to analysis of nested case-control studies. *Biometrika*, 84(2):379–394, 1997.

ELIAN CHIA HUI SAN. *Joint Modelling of Survival and Longitudinal Data under Nested Case-control Sampling.* PhD thesis, Saw Swee Hock School of Public Health National University of Singapore, 2013.

Ann M Saunders, Warren J Strittmatter, D Schmechel, PH St George-Hyslop, MA Pericak-Vance, SH Joo, and et al. Association of apolipoprotein e allele e4 with late-onset familial and sporadic alzheimer's disease. *Neurology*, 43(8):1467–1467, 1993.

David Schoenfeld. Partial residuals for the proportional hazards regression model. *Biometrika*, 69(1):239–241, 1982.

Steven G Self and Ross L Prentice. Asymptotic distribution theory and efficiency results for case-cohort studies. *The Annals of Statistics*, pages 64–81, 1988.

Dennis J Selkoe. Alzheimer's disease: genes, proteins, and therapy. *Physiological reviews*, 81(2):741–766, 2001.

Leslie M Shaw, Hugo Vanderstichele, Malgorzata Knapik-Czajka, Christopher M Clark, Paul S Aisen, Ronald C Petersen, Kaj Blennow, Holly Soares, Adam Simon, Piotr Lewczuk, et al. Cerebrospinal fluid biomarker signature in alzheimer's disease neuroimaging initiative subjects. *Annals of neurology*, 65(4):403–413, 2009.

Kyle Strimbu and Jorge A Tavel. What are biomarkers? *Current Opinion in HIV and AIDS*, 5(6):463, 2010.

Cyntha A Struthers and John D Kalbfleisch. Misspecified proportional hazard models. *Biometrika*, 73(2):363–369, 1986.

Jeremy MG Taylor. Choosing the number of controls in a matched case-control study, some sample size, power and efficiency considerations. *Statistics in medicine*, 5(1):29–36, 1986.

Duncan C Thomas. Addendum to: Methods of cohort analysis: Appraisal by application to asbestos mining. by fdk liddell, jc mcdonald and dc thomas. *Journal of the Royal Statistical Society, Series A*, 140:469–491, 1977.

Chi-hong Tseng and Mengling Liu. Joint modeling of survival data and longitudinal measurements under nested case-control sampling. *Statistics in Biopharmaceutical Research*, 1(4):415–423, 2009.

Anastasios A Tsiatis and Marie Davidian. A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika*, 88(2):447–458, 2001.

Anastasios A Tsiatis, Victor Degruttola, and Michael S Wulfsohn. Modeling the relationship of survival to longitudinal data measured with error. applications to survival and cd4 counts in patients with aids. *Journal of the American Statistical Association*, 90(429): 27–37, 1995.

Hajime Uno, Tianxi Cai, Lu Tian, and Lee-Jen Wei. Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association*, 102(478):527–537, 2007.

Vivian Viallon and Aurélien Latouche. Discrimination measures for survival outcomes: connection between the auc and the predictiveness curve. *Biometrical Journal*, 53(2):217–236, 2011.

Sholom Wacholder. Practical considerations in choosing between the case-cohort and nested case-control designs. *Epidemiology*, pages 155–158, 1991.

Sholom Wacholder, Mitchell H Gail, David Pee, and Ron Brookmeyer. Alternative variance and efficiency calculations for the case-cohort design. *Biometrika*, 76(1):117–123, 1989.

Jia-Gang Wang. A note on the uniform consistency of the kaplan-meier estimator. *The Annals of Statistics*, pages 1313–1316, 1987.

Michael W. Weiner. Welcome from the adni principal investigator. `http://adni-info.org`, 2013. Accessed: 2017-01-22.

Lang Wu, Wei Liu, Grace Y Yi, and Yangxin Huang. Analysis of longitudinal and survival data: joint modeling, inference methods, and issues. *Journal of Probability and Statistics*, 2012, 2012.

Michael S Wulfsohn and Anastasios A Tsiatis. A joint model for survival and longitudinal data measured with error. *Biometrics*, pages 330–339, 1997.

Ronghui Xu and John O'Quigley. Estimating average regression effect under non-proportional hazards. *Biostatistics*, 1(4):423–439, 2000.

Yingye Zheng, Tianxi Cai, and Ziding Feng. Application of the time-dependent roc curves for prognostic accuracy with multiple biomarkers. *Biometrics*, 62(1):279–287, 2006.

Qian M Zhou, Yingye Zheng, and Tianxi Cai. Assessment of biomarkers for risk prediction with nested case–control studies. *Clinical Trials*, 10(5):677–679, 2013.

Appendix A

Implementing ADNI Analysis in R, SAS, and STATA

This section provides the code used to perform the analysis in Section 2.8. In the following code we refer to data for MCI participants from the ADNI site with minor data cleaning.

A.1 Implementation in R

The analysis for this chapter was conducted using R, and code to recreate the analyses follows below. Note that permission from ADNI must first be obtained prior to using these data, and that the dataset referred to in the example below (MCI-ADNI.csv) was preprocessed.

```
#load the required packages
library(Epi)
library(survival)

#read in the data to be used, and set the base categories
```

```
mci.tte.surv <- (read.csv("MCI-ADNI.csv", header = TRUE))
mci.tte.surv$race <- relevel(mci.tte.surv$race, "White")
mci.tte.surv$PTGENDER <- relevel(mci.tte.surv$PTGENDER, "Male")

#fit the model for the full cohort analysis
fit.full <- coxph(Surv(obstime, Fail) ~ I(ptau_bl/20) + I(AGE/5)+ race +
PTGENDER + PTEDUCAT + factor(APOE4), data = mci.tte.surv)

summary(fit.full)



################ nested case-control ##############################
#sample 3 controls for the nested case-control
set.seed(12345)

ncc.mci.tte.surv3 <- ccwc(exit = obstime, fail = Fail, controls = 3,
    include = list(RID,ptau_bl, AGE,race, PTGENDER,PTEDUCAT, APOE4, obstime),
    data = mci.tte.surv)

#fit the model to the data from the case-control design

fit.ncc <- coxph(Surv(obstime, Fail) ~ I(ptau_bl/20) + I(AGE/5) + race
+ PTGENDER + PTEDUCAT + factor(APOE4) + strata(Set), data = ncc.mci.tte.surv3)

summary(fit.ncc)

#function to organize nested case-control data for Design I
nestedcc.forward <- function(tx.cc1){
  tx.cc1 <- tx.cc1[order(tx.cc1$Time),]
```

```
times.order <- sort(tx.cc1$Time[!duplicated(tx.cc1$Time)])

tx.cc1$Set <- NA
for(l in 1:length(times.order)){
  tx.cc1$Set[tx.cc1$Time == times.order[l]] <- l
}

controls <- tx.cc1[tx.cc1$Fail == 0,]
controls$Set1 <- controls$Set

tx.cc.x <- tx.cc1[tx.cc1$Fail == 1,]

split.tx.cc.x <- split(tx.cc.x, tx.cc.x$Set)
ncc.forward <- NA
for(i in 1:length(split.tx.cc.x)){
  temp <- split.tx.cc.x[[i]]
  temp$Set1 <- temp$Set
  temp.controls <- rbind(temp, controls)
  temp.controls$Set <- i
  temp.controls$Time <- temp$Time[1]
  ncc.forward <- rbind(ncc.forward, temp.controls)
  }

ncc.forward <- ncc.forward[-1,]
ncc.forward <- ncc.forward[which((ncc.forward$Set1 <= ncc.forward$Set)
                               & (ncc.forward$obstime >= ncc.forward$Time)),]
ncc.forward <- ncc.forward[,-which(names(ncc.forward) == "Set1")]
ncc.forward <- ncc.forward[which(!duplicated(ncc.forward[,c(1,2)])),]
```

```
  return(ncc.forward)
  }



#call function and fit model
ncc.mci.i <- nestedcc.forward(ncc.mci.tte.surv3)

fit.ncc.i <- coxph(Surv(obstime, Fail) ~ I(ptau_bl/20) + I(AGE/5) + race
                      + PTGENDER + PTEDUCAT + factor(APOE4) + strata(Set),
                      data = ncc.mci.i)
summary(fit.ncc.i)

#generate data for Design III
ncc.design3 <- function(dat, delta,n, id, numcontrols){

#set variables based on the ones in the data set
dat <- dat
dat$delta <- delta
dat$id <- id
event.times <- dat$obstime[dat$delta == 1]
u.event.times <- sort(unique(event.times))

#label index
compare <- matrix(rep(c(NA, u.event.times), each = dim(dat)[1]), byrow = FALSE,
                   nrow= n, ncol = length(u.event.times) + 1)

compare[,1] <- dat$obstime
dat$indx <- apply(compare, 1, rank)[1,] - 1
dat$status <- 1
```

```
dat$indx <- ceiling(dat$indx)

temp <- NA
controls <- NA
ncc.nr <- NA

  #sample controls for each event time
  for(i in 1:length(u.event.times)){
    set.seed(i)
    controls <- dat[which(dat$indx >= i),]

    #save the cases for this event time and remove them from controls
    case <- controls[which(controls$delta == 1 & controls$indx == i), ]
    case$Fail <- 1
    controls <- controls[-which(controls$delta == 1 & controls$indx == i),]
    controls$Fail <- 0

    #set weights for path sets
    weights <- prop.table(table(controls$indx))

    #if not the last pathset, sample from the list of available path sets
    if(weights[1] != 1){
      select.pathset <- sample(sort(unique(controls$indx[controls$indx >= i])),
       (numcontrols*dim(case)[1]), prob = weights, replace = TRUE)
        }else{
      select.pathset <- rep(unique(controls$indx), numcontrols)
      }

    #sample controls
```

```
for(k in 1:(numcontrols*dim(case)[1])){
  sample.controls <- controls$id[which((controls$indx ==
  select.pathset[k]) & (controls$status == 1))]

  #if number of controls available is greater than one, sample a control
  if(length(sample.controls) > 1){
    select.subject <- sample(sample.controls, 1)
    }
    #if number of controls available is one, take that as the control
    else if(length(sample.controls) == 1){
      select.subject <- sample.controls
      }
    #if there are no available controls, make all controls
    available for that path set
    #and select control
    else{
      dat$status[dat$indx == select.pathset[k]] <- 1
      controls$status[controls$indx == select.pathset[k]] <- 1

      sample.controls <- controls$id[which((controls$indx ==
      select.pathset[k]) & (controls$status == 1))]

      select.subject <- ifelse(length(sample.controls) >1,
      sample(sample.controls,1),
                               sample.controls)
      }

  #mark the selected control as used and add them to the set of controls
  dat$status[which(dat$id == select.subject)] <- 0
  controls$status[which(controls$id == select.subject)] <- 0
  temp <- rbind(temp, controls[which(controls$id == select.subject),])

  #if the control is repeated as a control, sample another control
  while(sum(duplicated(temp$id)) != 0
                 & (sum(is.na(match(controls$id, temp[-1,]$id))) != 0)){
    sample.controls <- controls$id[which((controls$indx == select.pathset[k])
                                   & (controls$status == 1))]

    dup <- which(duplicated(temp$id))
    temp <- temp[-dup,]

    if(length(sample.controls) > 1){
      select.subject <- sample(sample.controls, 1)
      dat$status[which(dat$id == select.subject)] <- 0
      controls$status[which(controls$id == select.subject)] <- 0
      temp <- rbind(temp, controls[which(controls$id == select.subject),])
      }
      else if(length(sample.controls) == 1){
        select.subject <- sample.controls
        dat$status[which(dat$id == select.subject)] <- 0
        controls$status[which(controls$id == select.subject)] <- 0
        temp <- rbind(temp, controls[which(controls$id == select.subject),])
      }
   }
}

#remove the NA in the first row
temp <- temp[-1,]
```

```
    #combine cases and controls and save them to data set
    case.controls <- rbind(case, temp)
    case.controls$Set <- i
    case.controls$Time <- case$obstime[1]
    ncc.nr <- rbind(ncc.nr, case.controls)
    temp <- NA
    }

  ncc.nr <- ncc.nr[-1,]
  return(ncc.nr)
  }


#generate Design III data
ncc.nr <- ncc.design3(mci.tte.surv, mci.tte.surv$Fail, dim(dat)[1],
mci.tte.surv$RID, 3)

#change the reference level
ncc.nr$PTGENDER <- relevel(ncc.nr$PTGENDER, "Male")
ncc.nr$race <- relevel(ncc.nr$race, "White")


fit.ncc.iii <- coxph(Surv(obstime, Fail) ˜ I(ptau_bl/20) + I(AGE/5) +race +
PTGENDER + PTEDUCAT + factor(APOE4) + strata(Set), data = ncc.nr)
summary(fit.ncc.iii)


################ case-cohort #####################
```

```
#sample the subcohort
set.seed(12345)
n.obs <- dim(mci.tte.surv)[1]
keep <- sample(1:n.obs, ceiling(n.obs*0.75), replace = FALSE)
mci.tte.surv$subcohort <- 0
mci.tte.surv$subcohort[keep] <- 1
mci.subcohort <- mci.tte.surv[(mci.tte.surv$subcohort == 1) |

(mci.tte.surv$Fail == 1),]

#change the reference level
mci.subcohort$race <- relevel(mci.subcohort$race, "White")
mci.subcohort$PTGENDER <- relevel(mci.subcohort$PTGENDER, "Male")

###Prentice method
fit.subcohort.p <- cch(Surv(obstime, Fail) ~ I(ptau_bl/20) + I(AGE/5) + race +
                            PTGENDER + PTEDUCAT + factor(APOE4),
                            data = mci.subcohort,
                            subcoh = ~subcohort,
                            id = ~RID, cohort.size = n.obs)
summary(fit.subcohort.p)

###Self and Prentice method
fit.subcohort.sp <- cch(Surv(obstime, Fail) ~ I(ptau_bl/20) + I(AGE/5) + race +
                            PTGENDER + PTEDUCAT + factor(APOE4),
                            data = mci.subcohort,
                             subcoh = ~subcohort, id = ~RID, cohort.size = n.obs,
                             method = "SelfPrentice")
summary(fit.subcohort.sp)
```

```
###Lin and Ying method
fit.subcohort.ly <- cch(Surv(obstime, Fail) ~ I(ptau_bl/20) + I(AGE/5) + race
                            + PTGENDER + PTEDUCAT + factor(APOE4),
                            data= mci.subcohort, subcoh = ~subcohort,
                            id = ~RID,
                            cohort.size = n.obs, method = "LinYing")

summary(fit.subcohort.ly)
```

A.1.1 Implementation in STATA

The following represents analogous code used to conduct the analysis in STATA. In STATA and SAS (next section), we only discuss the nested case-control design and the Prentice (1986) case-cohort design.

```
*read in the data
insheet using MCI-ADNI.csv

*only keep the variables needed for the analysis
keep rid age ptgender pteducat apoe4 ptau_bl obstime fail race
save MCI-ADNI2.csv

*create indicator variables for categorical
tabulate apoe4, generate(apoe)
tabulate ptgender, generate(gender)
tabulate race, gen(_race)
```

```
*divide ptau by 20
 generate ptau20 = ptau_bl/20

*divide age by 5
 generate age5 = age/5

 *define survival data
 stset obstime, failure(fail==1)

*fit model for full cohort
stcox ptau20 age5 _race1 _race3 gender1 pteducat apoe2 apoe3, efron


*****************************
* nested case-control
*****************************
set seed 12345
sttocc, n(3) nodots

*set the ncc data as survival data
stset obstime, failure(_case==1)

*fit model for nested cc
stcox ptau20 age5 _race1 _race3 gender1 pteducat apoe2 apoe3, strata(_set) efron


****************************
* case cohort
```

```
****************************
set seed 12345
gen u = runiform()

*generate 55% censoring
gen subcohort = u < 0.55

*set start time to be 0 if in subcohort
*if not in suhcohort, set start time to be right before event time
gen start = 0
replace start = obstime - 0.0001 if subcohort == 0

*set case-cohort data as survival data
stset obstime, id(rid) failure(fail==1) origin(start)

*fit model for case-cohort data
stcox ptau20 age5 _race1 _race3 gender1 pteducat apoe2 apoe3, efron
```

The `sttocc` command also allows for the option to match based on a specified variable, therefore allowing the implementation of the matched nested case-control design. We may also implement some of the case-cohort designs by setting the correct weights.

A.2 Implementation in SAS

The code to perform this analysis in SAS for the case-cohort data is the following:

```
/*read in the data */
```

```
proc import datafile="D:\HOSchapter\ADNI\MCI-ADNI-short.csv"
out= mci dbms = csv replace;
getnames = yes;

/*delete subjects who only have the baseline visit recorded */
data mci2;
set mci;
if obstime = 0 then delete;

/*fit the model to the full cohort */
proc phreg data = mci2;
class race PTGENDER APOE4;
model obstime*Fail(0) = race PTGENDER PTEDUCAT AGE APOE4 ptau20;

/*create a new variable that randomly assigns numbers from
the uniform distribution*/
data mci2;
set mci2;

unif = ranuni(12345);

/*select approximately 55% of the cohort as the subcohort */
data mci2;
set mci2;
if unif > 0.55 then subcohort = 0;
else subcohort = 1;

/*define a new variable that decides whether they are included in the analysis*/
data mci2;
```

```
set mci2;
if (subcohort = 1 or Fail = 1) then count = 1;
else count = 0;

/*delete all observations that are not in the subcohort nor become cases*/

data mci2;
set mci2;
if count = 0 then delete;

/* set start time to be 0 if in the subcohort, otherwise start
time is right before event time */
data mci2;
set mci2;
if subcohort = 1 then start = 0;
else start = obstime - 0.0001;

/* fit model to the case-cohort data */
proc phreg data = mci2;
class race PTGENDER APOE4;
model (start, obstime)*Fail(0) = race PTGENDER PTEDUCAT AGE
APOE4 ptau20/ties = efron;
run;
quit;
```

The SAS code for the nested case-control design is not included here. Instead, we refer readers to the `nCCsampling` macro, available at:

http://www.drugepi.org/wp-content/uploads/2011/04/nCC-sampling-macro.txt.

This macro generates the data for a nested case-control analysis and allows for matching on confounding variables Desai et al. [2016]. Once the data is generated, a model can be fit similarly to the case-cohort analysis but stratifying on the group or case number.