UNIVERSITY OF CALIFORNIA,
IRVINE

Efficient Sampling Designs and Robust Inference for Time-to-Event Data

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Statistics

by

Mikaela K. Nishida

Dissertation Committee:
Chancellor's Professor Daniel L. Gillen, Chair
Assistant Professor Michelle M. Nuño, Chair
Assistant Professor Tianchen Qian

2026

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to start by thanking the members of my dissertation committee, Tianchen Qian and Josh Grill, for taking the time to serve on my committee and for providing invaluable feedback on this dissertation. An additional thank you to Zhaoxia Yu for serving on my advancement committee. I am grateful for all of your insights throughout this journey.

Next, I would like to thank the members of the SMART and Grillen labs for being such a supportive group. You have given me a safe space to grow as a researcher and taught me so much along the way. An additional thanks to Josh for welcoming me into the Grillen lab and for giving me the opportunity to participate in true scientific collaboration.

To my friends in the department, thank you for all of the support you've given me throughout these past five years. You were such a significant part of this experience. We may have started off as "classmates," but I am very thankful that our paths crossed and that I get to call you friends now.

To Dan and Michelle, thank you for everything. I could not have done it without your patience, guidance, and encouragement throughout this journey. You are great researchers but also *good* ones, and I still feel incredibly lucky to have had the opportunity to work with the both of you.

Lastly, I'd like to thank my family and my partner, Russell, for always giving me a place to call home. None of this would have been possible without your unconditional love and support.

This dissertation is based upon work funded by the National Institutes of Health's National Institute on Aging (Grant No. P30 AG066519-03, RF1 AG075107, and T32 AG00096-42). 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 Institutes of Health.

VITA

Mikaela K. Nishida

EDUCATION

Doctor of Philosophy in Statistics — **August 2026**
University of California, Irvine — *Irvine, CA*

Master of Science in Statistics — **March 2024**
University of California, Irvine — *Irvine, CA*

Bachelor of Arts in Mathematics — **May 2021**
Pomona College — *Claremont, CA*

RESEARCH EXPERIENCE

Graduate Student Researcher — **Fall 2021 – Summer 2026**
University of California, Irvine — *Irvine, California*

NIA T32 Training Grant Fellow — **Summer 2025 – Summer 2026**
University of California, Irvine — *Irvine, California*

NIA Diversity Supplement Fellow — **Spring 2023 – Spring 2025**
University of California, Irvine — *Irvine, California*

TEACHING EXPERIENCE

Teaching Assistant — **Fall 2021 – Spring 2023**
University of California, Irvine — *Irvine, CA*

Teaching Assistant — **Summer 2022**
ISI-BUDS — *Irvine, CA*

REFEREED JOURNAL PUBLICATIONS

On the Analysis of Multiple Event Survival Data Under Non-Proportional Hazards — **In preparation**

Nested Case-Control Sampling for Marginal Estimation in Multiple Event Settings — **In revision**

Modified Nested Case-Control Designs for Assessing Effect Modification in Underrepresented Populations — **In revision**

Recruitment Incentives Used to Increase Willingness to Participate in Preclinical Alzheimer's Disease Trials in a Diverse Community-Based Population — **In preparation**

Impacts of Informant Replacement in Two Industry-Sponsored Alzheimer's Disease Clinical Trials — **2024**
Alzheimer's & Dementia: Translational Research & Clinical Interventions

Effects of Informant Replacement in Alzheimer's Disease Clinical Trials — **2023**
Alzheimer's & Dementia: Translational Research & Clinical Interventions

ABSTRACT OF THE DISSERTATION

Efficient Sampling Designs and Robust Inference for Time-to-Event Data

By

Mikaela K. Nishida

Doctor of Philosophy in Statistics

University of California, Irvine, 2026

Chancellor's Professor Daniel L. Gillen, Chair

The proportional hazards (PH) model is widely used in time-to-event analyses, including within the nested case-control (NCC) design, an efficient sampling approach often used to reduce data collection costs in rare disease settings. By including all events and a subsample of controls, the NCC design maximizes information while reducing the number of subjects requiring full covariate information. Since the NCC design uses fewer controls compared to the full cohort analysis, however, it also reduces the number of subjects from underrepresented groups included in the sample. This further exacerbates the issue of underrepresentation in clinical research and diminishes our ability to assess potential effect modification in these subpopulations. Additionally, while the NCC design has been extensively studied for univariate outcomes, its application to multiple event data remains limited.

In this dissertation, we present novel statistical methods to bridge existing gaps in current methodology for efficient sampling and robust estimation for time-to-event data. Specifically, we propose an extension of the NCC design that allows for oversampling of subpopulations, thereby maximizing precision for estimated covariate effects in these groups. The resulting estimation framework yields asymptotically valid inference in both marginal and stratified analyses while maximizing efficiency in underrepresented subpopulations. We also explore NCC sampling frameworks for multiple event settings and propose an event-specific stratified

sampling design with corresponding estimation methods for marginal parameter inference. We consider the efficiency of various proposed frameworks for sampling controls, highlighting differences in performance between models assuming common and event-specific baseline hazards across recurrent events. Finally, we transition to the full cohort setting, where we assess robustness of standard proportional hazards estimators under violation of key model assumptions in the multiple event setting. We show that the estimand corresponding to the partial likelihood is dependent upon the event-specific censoring and survival distributions. The result is that commonly implemented diagnostics for multiple event survival can be shown to be misleading in situations where the baseline hazard varies by event type. The work in this dissertation is motivated by current challenges in biomarker development in the setting of Alzheimer's disease, and applications to this area are presented.

Chapter 1

Introduction

Biomarkers play a critical role in the early diagnosis of disease and can serve as targets for disease interventions (Van Dyck et al., 2023). As such, biomarker discovery and validation are of primary scientific interest in many disease settings. An example of this occurs in Alzheimer's disease (AD), a neurodegenerative disease that affects memory, thinking, and behavior. In AD, amyloid beta (Aβ) and phosphorylated tau (p-tau) are two protein biomarkers that have become key to early diagnosis and serve as targets for interventions. As Aβ and p-tau are not perfect discriminators of disease, biomarker discovery remains a primary objective in AD research. Diagnostic aids of neurologic change are often derived from cerebrospinal fluid (CSF), the liquid surrounding the brain and spinal cord. CSF is most commonly obtained via a lumbar puncture, an invasive procedure in which the practitioner inserts a hollow needle into the space surrounding the patient's spinal column in the lower back. Many participants are unwilling to undergo a lumbar puncture or participate in studies that require them, so difficulty in measuring candidate biomarkers is a barrier in biomarker discovery (Witbracht et al., 2020). When CSF samples are collected, they are frozen and stored for later use in analyses. Since thawing of samples can cause specimen degradation over time, the efficient use of these samples is critical to preserving this rare resource. As such, new research on

efficient statistical sampling designs is crucial to maximizing information while reducing the total number of CSF samples that are processed. This is even more critical when assessing biomarker utility among underrepresented populations, including lower socioeconomic status and non-non-Hispanic White populations.

Efficient statistical sampling designs such as the nested case-control (NCC) design accelerate scientific knowledge while minimizing sample size requirements. The NCC design minimizes the amount of information required on patients by only requiring biomarker information for patients that experience disease progression (cases), and a randomly sampled subset of patients that have yet to progress (controls) while maintaining high precision for assessing the utility of a candidate biomarker with a reduced overall sample size relative to utilizing all controls (Nuño and Gillen, 2017). Despite its advantages, deficiencies in this method limit its use in many disease settings. First, the NCC design may limit or eliminate our ability to quantify the utility of biomarkers in underrepresented subpopulations by only sampling a fraction of the available controls from these groups, leading to less information on biomarkers among these populations. In AD research, racial and ethnic representation is an ongoing challenge, with study cohorts often being primarily comprised of highly educated, non-Hispanic, White participants (Raman et al., 2021). Second, the progression between disease stages such as the progression from cognitively normal to mild cognitive impairment (MCI) and the progression from MCI to dementia represent censored time-to-event outcomes, and the use of multiple time-to-event outcomes can increase precision for assessing biomarker utility. Little research has considered extensions of the NCC design to the setting of multiple time-to-event outcomes, despite its potential utility. The objective of the work in this dissertation is to fill in these gaps by developing novel statistical theory and methods to increase efficiency while maximizing available information on underrepresented populations to discover and validate new biomarkers for time-to-event outcomes. We hope that these methods will be applied to biomarker discovery, assessment, and validation to address growing health disparities among underrepresented populations in the diagnosis and

treatment of AD.

1.1 Dissertation Overview

The work in this dissertation begins in Chapter 2, where we provide a review of concepts that are foundational to the work that is presented in the remaining chapters. These include a review of time-to-event data, the Cox proportional hazards model, the NCC design, and multiple event survival analysis. The goals of this dissertation are achieved through a series of specific aims that are directed at contributing to the statistical methodology available to increase efficiency and robustness while maximizing information on underrepresented populations.

The first aim of this dissertation is to extend the NCC design to support weighted sampling for assessing effect modification in underrepresented populations and is motivated by the need to maximize available information on underrepresented subpopulations to estimate subpopulation-specific biomarker effects. Chapter 3 introduces a weighted NCC sampling design that allows for oversampling of subpopulations, thereby maximizing precision for estimated covariate effects in these groups. We provide corrected estimation and inferential methods, demonstrate the performance of our method in simulation, and apply our proposed method to data from the NIH-funded National Alzheimer's Coordinating Center (NACC) where we investigate potentially differential associations between Aβ and the risk of AD between Black and non-Black populations.

Similarly motivated by current gaps in the NCC design literature, the second aim of this dissertation is to develop a framework for using the NCC design to analyze multiple event survival data. In Chapter 4, we explore two different sampling frameworks for multiple events which are both motivated by the risk set formulations under models assuming common and event-specific baseline hazards. We propose an event-specific stratified sampling design with

corresponding estimation methods for marginal parameter inference. We demonstrate that stratified event-specific sampling performs similarly or better than pooled sampling with respect to efficiency in the multiple event setting. In addition to evaluating the performance of our proposed estimators in simulation, we apply our sampling method to data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) to estimate the association between p-tau and disease progression.

In addition to the work in efficient sampling designs, this dissertation aims to develop methods that are robust to model misspecification. For example, the Cox model assumes that the hazards comparing two groups are proportional at all observed times, and when this assumption is violated, it has been demonstrated that the estimate is dependent on a nuisance parameter dictated by participant accrual and dropout patterns. In the univariate setting, estimators that remove dependence on the censoring distribution have been developed to assuage this issue, but the statistical implications of non-proportional hazards (NPH) have been relatively understudied in the multiple event setting where methods analogous to univariate PH models are often used. When estimating a treatment effect in the recurrent event setting, estimates from event-specific PH models are dependent on participant accrual and dropout patterns as well as the baseline hazard, which can change across recurrence despite a common underlying treatment effect. The last aim of this dissertation pivots to assess the statistical properties of standard PH estimators when analyzing multiple event data under a violation of the PH assumption. This work is presented in Chapter 5, where we introduce the theoretical challenges that can arise when estimating event-specific treatment effects in the setting of recurrent events. In addition to this, we conclude that standard PH methods are not valid under NPH for multiple event data and result in an estimate that is not replicable or interpretable.

Chapter 2

Background

2.1 Review of Time-to-Event Data

2.1.1 Censoring

In many studies, primary scientific interest lies in analyzing the association between different factors and the time until an event of interest occurs. Some examples of this include the time to disease onset, time to hospitalization or death, and time to the occurrence or resolution of a symptom. Ideally, study participants would be followed until the event of interest is observed, but this is often unfeasible due to time constraints, leading to a key feature in time-to-event data called *censoring*. Censoring occurs when the exact time-to-event data is unknown or incomplete for some individuals. There are various types of censoring including right censoring, left censoring, and interval censoring. *Left censoring* arises when the event of interest occurs before some time, C_l, when the participant is first observed in the study. *Interval censoring* refers to the case where the event of interest is known to occur within an interval but the exact time is unknown.

Right censoring takes place when the event of interest occurs after some time C_r, but the

exact time is unknown due to reasons that may include loss to follow-up, study completion, or death due to unrelated causes. Instead, the last time at which the participant was observed at risk of experiencing the event, C_r, is recorded. We note that interval-censored data are sometimes treated as right-censored in analyses, though specialized methods exist to handle interval censored data. A few types of right censoring exist. For example, generalized Type I censoring, also known as "administrative censoring," refers to right censoring that occurs as a result of the end of the study. More specifically, only events that occur before the prespecified study end, τ, are observed. Similarly, Type II censoring refers to censoring that occurs when a study ends after the r^{th} event. Lastly, random censoring encompasses censoring due to unplanned causes such as a withdrawal from the study or death for unrelated reasons. A key assumption underlying most survival analysis methods is that the censoring times and event times are statistically independent conditional on the covariates included in the model. As such, one must ensure that causes of censoring are truly random and not informative missingness. In this dissertation, we will focus generally on methodology for right-censored data.

2.1.2 Useful Statistical Functions

We start by defining a few useful statistical functions that are commonly of interest in survival analysis. Let T denote the random variable representing the survival time of interest (e.g. time to death, time to progression of a disease, etc.).

For some time t, the **Probability Density Function (PDF)** is:

$$f(t) = \lim_{\Delta t \to 0+} \frac{1}{\Delta t} P(t \leq T < t + \Delta t),$$

and the **Cumulative Distribution Function (CDF)** is:

$$F(t) = P(T \leq t) = \int_0^t f(s) ds.$$

In the survival analysis context, it can often be easier to think about the probability of surviving past a certain time, so we define the **Survival Function** as:

$$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t) = 1 - \int_0^t f(s)ds.$$

Since censored observations are often present, many methods in survival analysis will instead consider the hazard function and cumulative hazard functions, which take censored observations into account by conditioning on survival up until a certain time. Specifically, the **Hazard Function** is defined as:

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

and can be interpreted as the instantaneous event rate at time t conditional on survival until time t. The **Cumulative Hazard Function** is defined as:

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

The Kaplan-Meier estimator of the survival distribution, $S(t)$, is:

$$\widehat{S}_{KM}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

where d_i denotes the total number of events occurring at time t_i, and n_i denotes the number of individuals who have not yet experienced the event or been censored prior to time t_i (Kaplan and Meier, 1958).

2.2 The Cox Proportional Hazards Model

When analyzing right-censored, time-to-event outcomes, it is common to use the semi-parametric Proportional Hazards (PH) model proposed by Cox (1972) to quantify the relationship between the hazard function and covariates. To introduce the model, consider the setting of right-censored time-to-event data, where T_i, C_i, and $Z_i = [z_{i1}, \ldots, z_{ip}]^T$ denote the true event time, censoring time, and $p \times 1$-vector of covariates for subject i $(i = 1, \ldots, n)$, respectively. Further define $X_i = \min(T_i,\ C_i)$ as the observed time and $\delta_i = I(X_i = T_i)$ as the event indicator for subject i. Under the Cox PH model, the event times are modeled by the covariates via the hazard function:

$$\lambda_i(t) = \lambda_0(t) \exp\{\beta^T Z_i\}$$

where $\lambda_0(t)$ denotes the baseline hazard function and $\beta = [\beta_1, \ldots, \beta_p]$ is the $1 \times p$-vector of regression parameters representing the corresponding log-hazard ratios associated with covariate vector Z. The PH model is a semi-parametric model due to the completely unspecified baseline hazard, which is treated as a nuisance parameter in the calculation of the partial likelihood.

As mentioned previously, the hazard function conditions on survival up until a specific time, so an important concept that will be referenced moving forward is the "risk set." The risk set at time X_i, denoted by R_i, is the set of subjects that have not yet experienced the event of interest and have not been censored prior to time X_i, or $R_i = \{j : X_j \geq X_i\}$. The Cox PH model can be constructed by considering the conditional probability that a subject with covariate value Z experiences the event of interest at time t, given some subject in the risk set experiences the event at time t. Specifically, subject i's contribution to the partial

likelihood is:

$$
\begin{aligned}
L_i &= P(\text{subj. with } Z_i \text{ experiences an event at } X_i \mid \\
&\qquad\qquad \text{any subj. experienced an event at } X_i) \\
&= \frac{P(\text{subj. with } Z_i \text{ experiences an event at } X_i)}{P(\text{any subj. in the risk set experienced an event at } X_i)} \\
&= \frac{\lambda_i(X_i)\ \Delta t}{\sum_{j \in R_i} \lambda_j(X_i)\ \Delta t} \\
&= \frac{\lambda_0(X_i) \exp\{\beta^T Z_i\}}{\sum_{j \in R_i} \lambda_0(X_i) \exp\{\beta^T Z_j\}} \\
&= \frac{\exp\{\beta^T Z_i\}}{\sum_{j \in R_i} \exp\{\beta^T Z_j\}}
\end{aligned}
$$

Subjects who experience an event contribute directly to the likelihood, but subjects who do not experience an event only contribute to the likelihood through their presence in the risk set. It follows that the full partial likelihood is:

$$L_{FC} = \prod_{i:\delta_i=1} L_i,$$

and the corresponding log-partial likelihood is:

$$
\begin{aligned}
\log(L_{FC}) = \sum_{i:\delta_i=1} \log(L_i) &= \sum_{i=1}^{n} \delta_i \log\left(\frac{\exp\{\beta^T Z_i\}}{\sum_{j \in R_i} \exp\{\beta^T Z_j\}} \right) \\
&= \sum_{i=1}^{n} \delta_i \left\{ \beta^T Z_i - \log\left(\sum_{j \in R_i} \exp\{\beta^T Z_j\} \right) \right\}.
\end{aligned}
$$

The parameters from the Cox PH model are estimated by maximizing the partial likelihood, or equivalently by solving the estimating equation:

$$\mathcal{U}_{FC}(\beta) = \frac{\partial \log(L_{FC})}{\partial \beta} = \sum_{i=1}^{n} \delta_i \left[Z_i - \frac{\sum_{j \in R_i} Z_j \exp\{\beta^T Z_j\}}{\sum_{j \in R_i} \exp\{\beta^T Z_j\}} \right] \equiv 0. \tag{2.1}$$

Coefficients from this model denote the log relative risk comparing subpopulations differing by one unit in the corresponding covariate. Equivalently, $\exp\{\widehat{\beta}_k\}$ is the estimated relative risk comparing two subpopulations differing in z_k by one unit but holding all other covariates constant. Under a series of regularity assumptions, $\widehat{\beta}_{FC} \dot{\sim} N(\beta_0, \mathcal{I}_1^{-1}(\beta_0))$, where β_0 is the true value of β and $\mathcal{I}(\beta_0)$ can be estimated by $I(\widehat{\beta}_{FC}) = -\frac{\partial \mathcal{U}_{FC}(\beta)}{\partial \beta}|_{\beta=\widehat{\beta}_{FC}}$. We discuss this result in more detail below.

Derivation of the Profile Likelihood

The Cox PH model partial likelihood is actually considered a "profile likelihood" due to the "profiling out" of the baseline hazard. We start deriving the profile likelihood by considering the joint density of (X, δ, Z) at (X_i, δ_i, Z_i). We have that $T|Z \sim F_T(\cdot|Z)$, $C|Z \sim G_C(\cdot|Z)$, and $Z \sim H$. If an event is observed for subject i, or $\delta_i = 1$, then we know that the event time occurred before the censoring time, or $X_i = T_i$ and $X_i < C_i$. It follows that the contribution to the likelihood from this subject would be $f_T(X_i|Z_i)\{1 - G_C(X_i|Z_i)\}$. Similarly, if $\delta_i = 0$, then the censoring time occurred before the event time, or $X_i = C_i$ and $X_i < T_i$, and subject i's contribution to the likelihood would be $g_C(X_i|Z_i)\{1 - F_T(X_i|Z_i)\}$. When putting everything together, the assumption of the independence of censoring and event times conditional on covariates allows all terms relating to G_C and H to be omitted. It follows that the likelihood contribution from the i^{th} subject is:

$$
\begin{aligned}
L_i(\beta, \lambda_0) &= f_T(X_i|Z_i)^{\delta_i}\{1 - F_T(X_i|Z_i)\}^{(1-\delta_i)} \\
&= \{\lambda_i(X_i|Z_i)S(X_i|Z_i)\}^{\delta_i}\{S(X_i|Z_i)\}^{(1-\delta_i)} \\
&= \lambda_i(X_i|Z_i)^{\delta_i} S(X_i|Z_i) \\
\text{(under the Cox PH model)} \qquad &= [\lambda_0(X_i)\exp\{\beta^T Z_i\}]^{\delta_i} \exp\left\{-\int_0^{X_i} \lambda_i(s)ds\right\} \\
&= [\lambda_0(X_i)\exp\{\beta^T Z_i\}]^{\delta_i} \exp\left\{-\Lambda_0(X_i)\exp\{\beta^T Z_i\}\right\}
\end{aligned}
$$

Denoting $\lambda_0^{(i)}$ as the point mass of λ_0 at time X_i and discretizing the cumulative baseline hazard such that $\Lambda_0(X_i) = \sum_{j=1}^{n} I(X_j \leq X_i)\lambda_0^{(i)}$, we can further define the full log-likelihood as:

$$\begin{aligned}
\ell_{FC}(\beta, \lambda_0^{(1)}, \ldots, \lambda_0^{(n)}) &= \sum_{i=1}^{n}\left[\delta_i \log(\lambda_0^{(i)}) + \delta_i \beta^T Z_i - \left[\sum_{j=1}^{n} I(X_j \leq X_i)\lambda_0^{(i)}\right] \exp\{\beta^T Z_i\}\right] \\
&= \sum_{i=1}^{n}\left[\delta_i \log(\lambda_0^{(i)}) + \delta_i \beta^T Z_i - \lambda_0^{(i)} \sum_{j=1}^{n} I(X_j \geq X_i) \exp\{\beta^T Z_j\}\right]
\end{aligned}$$

Since we are primarily interested in the log-hazard ratios, β, we can form the profile likelihood by "profiling" out the baseline hazard. That is, we first fix β and maximize $\ell_{FC}(\beta, \lambda_0^{(1)}, \ldots, \lambda_0^{(n)})$ with respect to the respective baseline hazard parameters $(\lambda_0^{(1)}, \ldots, \lambda_0^{(n)})$.

$$\begin{aligned}
&\frac{\partial \ell}{\partial \lambda_0^{(i)}} = \frac{\delta_i}{\lambda_0^{(i)}} - \sum_{j=1}^{n} I(X_j \geq X_i) \exp\{\beta^T Z_j\} \equiv 0 \\
&\Rightarrow \lambda_0^{(i)} = \frac{\delta_i}{\sum_{j=1}^{n} I(X_j \geq X_i) \exp\{\beta^T Z_j\}}
\end{aligned}$$

Then using these values of the baseline hazards, we can obtain the profile log-likelihood for β:

$$p\ell(\beta) = \sum_{i=1}^{n}\left[\delta_i \beta^T Z_i - \delta_i \log \sum_{j=1}^{n} I(X_j \geq X_i) \exp\{\beta^T Z_j\} - \delta_i\right]$$

which is equivalent to the log-partial likelihood after dropping a constant term.

Asymptotic Normality of the Cox Model

To demonstrate asymptotic normality of the Cox PH estimator, we typically appeal to Rebolledo's Central Limit Theorem, also known as the Martingale Central Limit Theorem. We refer the reader to Kalbfleisch and Prentice (2002) for more information. We start by noting that the estimating equation for the Cox PH model (Equation 2.1) can also be written

using counting process notation:

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

where $N_i(t) = I(X_i \leq t, \delta_i = 1)$ and $Y_i(t) = I(X_i \geq t)$ denote the right-continuous counting process and left-continuous at-risk process for the i^{th} individual, respectively. As mentioned previously, parameter estimates from the Cox PH model are asymptotically normal under the regularity conditions listed below.

Regularity conditions:

1. $\{Y_i(\cdot), N_i(\cdot), Z_i\}$ are independent and identically distributed (i.i.d.).

2. Z_i is bounded with probability 1.

3. $\Lambda_0(\tau) < \infty$.

4. $P\{Y_i(t) = 1\} > 0$ for all $t \in (0, \tau]$.

5. $I(\beta_0)$ is positive-definite.

To prove asymptotic normality of this estimator, we start by demonstrating asymptotic normality of the score function. That is, $n^{-1/2}\, \mathcal{U}(\beta_0) \xrightarrow{d} N(0, I(\beta_0))$.

Proof. For the simplicity of notation for this proof, let $S^{(r)}(\beta, t) = \sum_{j=1}^{n} Y_j(t)\, Z_j^{\otimes r} \exp\{\beta^T Z_j\}$, and $\bar{Z}(t;\beta) = \frac{S^{(1)}(\beta,t)}{S^{(0)}(\beta,t)}$. We start by defining the score process as $\mathcal{U}(t;\beta) = \sum_{i=1}^{n} \int_0^t \{Z_i - \bar{Z}(s;\beta)\} dN_i(s) \equiv 0$, where $N_i(s)$ represents the counting process for subject i's event experience. We can re-express the score process as a martingale. Specifically, $\mathcal{U}(t;\beta) = \sum_{i=1}^{n} \int_0^t \{Z_i - \bar{Z}(s;\beta)\} dM_i(s;\beta) + \sum_{i=1}^{n} \int_0^t \{Z_i - \bar{Z}(s;\beta)\} Y_i(s) \exp\{\beta^T Z_i\} \lambda_0(s) ds$ where the first term is a Martingale and the second term can be shown to equal 0. We can express the normalized score process, evaluated at $\beta = \beta_0$, as $n^{-1/2}\mathcal{U}(t;\beta_0) = n^{-1/2} \sum_{i=1}^{n} \int_0^t \{Z_i -$

$\bar{Z}(s;\beta_0)\}dM_i(s;\beta_0)$, which is a martingale transform because $n^{-1/2}\{Z_i - \bar{Z}(t;\beta_0)\}$ is $\mathcal{F}_t$-predictable. Since $n^{-1/2}\mathcal{U}(t;\beta_0)$ is a martingale process, we have that $E[n^{-1/2}\mathcal{U}(t;\beta_0)] = 0$, and $E[n^{-1/2}\mathcal{U}(s;\beta_0)\{n^{-1/2}\mathcal{U}(t;\beta_0) - n^{-1/2}\mathcal{U}(s;\beta_0)\}] = 0$ for $s \leq t$.

In order to appeal to Rebolledo's Central Limit Theorem, we must verify convergence of the variation process and the Lindeberg condition:

1. Convergence of Variation Process: $\langle n^{-1/2}\mathcal{U}(t;\beta_0), n^{-1/2}\mathcal{U}(t;\beta_0)\rangle \xrightarrow{p} \Sigma(t)$ for a positive-definite $\Sigma(t)$.

2. Lindeberg Condition: For all $\epsilon > 0$, $\langle n^{-1/2}\mathcal{U}_{\ell\epsilon}(t;\beta_0), n^{-1/2}\mathcal{U}_{\ell\epsilon}(t;\beta_0)\rangle = n^{-1}\sum_{i=1}^n \int_0^t \{Z_{i\ell} - \bar{Z}_\ell(s;\beta_0)\}^2 I\{n^{-1/2}|Z_{i\ell} - \bar{Z}_\ell(s;\beta_0)| > \epsilon\} Y_i(s)\exp\{\beta_0^T Z_i\}\lambda_0(s)ds \xrightarrow{p} 0$.

For the convergence of the variation process, we have that $\langle n^{-1/2}\mathcal{U}(t;\beta_0), n^{-1/2}\mathcal{U}(t;\beta_0)\rangle = n^{-1}\sum_{i=1}^n \int_0^t \{Z_i - \bar{Z}(s;\beta_0)\}^{\otimes 2} Y_i(s)\exp\{\beta_0^T Z_i\}\lambda_0(s)ds$. Using the Dominated Convergence Theorem and the Law of Large Numbers, it can be shown that $\langle n^{-1/2}\mathcal{U}(t;\beta_0), n^{-1/2}\mathcal{U}(t;\beta_0)\rangle \xrightarrow{p} \Sigma(t)$, where $\Sigma(t) = I(\beta_0)$.

For the Lindeberg condition, we start by defining $n^{-1/2}\mathcal{U}_{\ell\epsilon}(t;\beta_0) = n^{-1/2}\sum_{i=1}^n \int_0^t \{Z_{i\ell} - \bar{Z}_\ell(s;\beta_0)\} I\{n^{-1/2}|Z_{i\ell} - \bar{Z}_\ell(s;\beta_0)| > \epsilon\} dM_i(s;\beta_0)$ for $\ell = 1,\ldots,p$. Then, we need to demonstrate that the variation process goes to 0. We start by noticing that $\{Z_{i\ell} - \bar{Z}_\ell(s;\beta_0)\}^2 \leq K^2$, where $K = \sup_{i,s}|Z_{i\ell} - \bar{Z}_\ell(s;\beta_0)| \leq \infty$. Moreover, we can write the indicator $I\{n^{-1/2}|Z_{i\ell} - \bar{Z}_\ell(s;\beta_0)| > \epsilon\} \leq I\{n^{-1/2}K > \epsilon\}$ for all i and s, implying that $I\{n^{-1/2}|Z_{i\ell} - \bar{Z}_\ell(s;\beta_0)| > \epsilon\} \to 0$ as $n \to \infty$ since $I\{n^{-1/2}K > \epsilon\} \to 0$ as $n \to \infty$. It follows that $\langle n^{-1/2}\mathcal{U}_{\ell\epsilon}(t;\beta_0), n^{-1/2}\mathcal{U}_{\ell\epsilon}(t;\beta_0)\rangle \leq \int_0^t n^{-1}\sum_{i=1}^n Y_i(s)\exp\{\beta_0^T Z_i\}K^2 I\{n^{-1/2}K > \epsilon\}\lambda_0(s)ds$, which converges in probability to 0. □

The normality of $\widehat{\beta}$ follows from a first-order Taylor expansion of the score around β_0. After expanding the score and doing algebraic manipulation to get $\sqrt{n}(\widehat{\beta} - \beta_0)$, it can be demonstrated that the product of all components on the other side of the equation converge

to a normal random variable with mean 0 and variance $\mathcal{I}^{-1}(\beta_0)$ by the Weak Law of Large Numbers, Mann-Wald, and Slutsky's Theorems. That is, $\sqrt{n}(\widehat{\beta} - \beta_0) \xrightarrow{d} N(0, \mathcal{I}^{-1}(\beta_0))$, where $\mathcal{I}(\beta_0)$ is consistently estimated by $I(\beta_0)$.

2.2.1 The Cox PH Model Under Misspecification

A primary assumption of the PH model is that the hazard function of all contrasted subpopulations remains proportional at all follow-up times. When this key assumption is violated, we can no longer rely on the well-established asymptotic theory supporting valid inference for the log-hazard ratio. A violation of the PH assumption, or non-proportional hazards (NPH), can be induced when the model is misspecified due to a missing covariate or misspecification of the functional form of a covariate that is correctly included. In addition, since inferential models should be prespecified prior to viewing the data to avoid multiple testing bias and data-driven modeling, it is not always the case that this assumption is met in practice. Still, the Cox PH model is used ubiquitously in clinical research, so it is important to consider its performance and interpretation under NPH.

Struthers and Kalbfleisch (1986) demonstrated that the Cox PH model produces estimates that are dependent on the censoring distribution when the PH assumption is violated. Specifically, under model misspecification, the estimate from the PH model, or the solution to Equation 2.1, is consistent for the solution to the following equation:

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

where $f_T(\cdot)$ denotes the density function for the event times, and $S_T(\cdot)$ and $S_C(\cdot)$ denote the survival functions for the true event and censoring times, respectively. Patterns of censoring are not typically of scientific interest, and yet this dependence implies that estimates obtained via the PH model may be study-dependent, since participant accrual and dropout patterns can differ by study. In the most concerning case, this result implies that studies with the

exact same underlying data generating mechanism will obtain different results due simply to having different participant accrual and dropout patterns. This poses serious challenges to study replicability and the interpretation of results.

2.2.2 Censoring-Robust Estimators

Many researchers have addressed the issue of censoring-time dependence by proposing estimators that reweight estimates from the PH model to a common distribution, namely, a censoring distribution that involves only administrative censoring at the study endpoint, τ. The resulting estimate can be interpreted as the average covariate effect over the study period in the absence of intermittent censoring (Xu, 2000). All reweighted estimators take on the following estimating equation form:

$$\mathcal{U}_{CR}(\beta) = \sum_{i=1}^{n} \int_0^{\infty} W_i(t) \left\{ Z_i - \frac{\sum_{j=1}^{n} Y_j(t) W_j(t) Z_j \exp\{\beta^T Z_j\}}{\sum_{j=1}^{n} Y_j(t) W_j(t) \exp\{\beta^T Z_j\}} \right\} dN_i(t) \equiv 0. \tag{2.2}$$

Xu (2000) first proposed a reweighted estimator that removes dependence on the censoring distribution when censoring is independent of covariates, where $W_i(t) = 1/\widehat{S}_C(t)$, the inverse of the left-continuous Kaplan-Meier estimate of the censoring times for all subjects combined at time t. Boyd et al. (2012) later extended this to the case where censoring may be dependent on a single binary covariate. This type of censoring, also referred to as "conditionally-independent censoring," often occurs in clinical trials where participant dropout patterns may differ by treatment arm. As a result, this requires a covariate-specific estimate of the censoring distribution for reweighting, and $W_i(t) = 1/\widehat{S}_C(t|Z_i)$ is computed using the left-continuous, Kaplan-Meier estimate of the censoring times for the group Z_i at time t. Finally, for the setting where censoring may depend on multiple, potentially continuous covariates, Nguyen and Gillen (2017) proposed a reweighted estimator that involves the use of survival trees to group individuals into censoring-similar groups for calculation of $W_i(t) = 1/\widehat{S}_C(t|\vec{Z}_i)$, where $\vec{Z}_i$ here can be a vector of multiple covariates. In each respective scenario, the solution

to Equation 2.2 is consistent for the solution to:

$$\int_0^\infty E\Bigg\{ f_T(t|Z) \times \Bigg[Z - \frac{E\{Z S_T(t|Z)\exp(\beta^T Z)\}}{E\{S_T(t|Z)\exp(\beta^T Z)\}} \Bigg] \Bigg\} dt = 0. \tag{2.3}$$

Robust Variance and Asymptotic Properties for Censoring-Robust Estimators

We focus on the censoring-robust estimator proposed by Boyd et al. (2012) for the setting of conditionally-independent censoring. The estimating equation takes the form of Equation 2.2 with $W_i(t) = 1/\widehat{S}_C(t|Z_i)$, and we will denote the resulting estimate as $\widehat{\beta}_{CIC}$, which is consistent for the solution to Equation 2.3, denoted here as β_{CIC}. When the PH assumption is violated, the inverse of the second derivative of the log-partial likelihood is no longer a consistent variance estimator. Boyd et al. (2012) derive a robust variance estimator that accounts for violation of the PH assumption but reduces to the classic variance estimator when the PH assumption holds. To start, we use the fact that Equation 2.2 can be written as:

$$\mathcal{U}_{CIC}(\beta) = \sum_{i=1}^n \delta_i W_i(t_i) Z_i - \int_0^\infty \frac{S_W^{(1)}(\beta,t)}{S_W^{(0)}(\beta,t)} dN_W(t) \tag{2.4}$$

where $S_W^{(r)}(\beta,t) = \sum_{j=1}^n Y_j(t) W_j(t) Z_j^{\otimes r} \exp\{\beta^T Z_j\}$ and $N_W(t) = \sum_{i=1}^n W_i(t) N_i(t)$. Expanding about $s_W^{(0)}(\beta,t) = \lim_{n\to\infty} S_W^{(0)}(\beta,t)$, $s_W^{(1)}(\beta,t) = \lim_{n\to\infty} S_W^{(1)}(\beta,t)$, and $\tilde{N}_W(t) = \lim_{n\to\infty} n^{-1} N_W(t)$, Equation 2.4 is asymptotically equivalent to:

$$\begin{aligned}\mathcal{U}^*_{CIC}(\beta) = \sum_{i=1}^n \mathcal{U}_i^*(\beta) = \sum_{i=1}^n \delta_i W_i(X_i) Z_i - \int_0^\infty \frac{s_W^{(1)}(\beta,t)}{s_W^{(0)}(\beta,t)} dN_W(t) - \int_0^\infty \frac{S_W^{(1)}(\beta,t)}{s_W^{(0)}(\beta,t)} d\tilde{N}_W(t) \\ + \int_0^\infty \frac{s_W^{(1)}(\beta,t) S_W^{(0)}(\beta,t)}{s_W^{(0)}(\beta,t)^2} d\tilde{N}_W(t). \end{aligned} \tag{2.5}$$

Since Equation 2.5 consists of i.i.d., mean-zero terms, $n^{-1/2}\mathcal{U}^*_{CIC}(\beta)$ is asymptotically normal with mean zero and covariance matrix $B = \lim_{n\to\infty} n^{-1} \sum_{i=1}^n \mathcal{U}_i^*(\beta)\mathcal{U}_i^*(\beta)^T$. Defining $A =$

$-\lim_{n\to\infty} n^{-1}\partial\mathcal{U}_{CIC}(\beta)/\partial\beta|_{\beta=\widehat{\beta}_{CIC}}$, it follows that the asymptotic variance of $\widehat{\beta}_{CIC}$ is given by $\lim_{n\to\infty} n^{-1}A^{-1}BA^{-1}$. Using consistent estimators of each component, it follows that the variance of $\widehat{\beta}_{CIC}$ can be estimated by:

$$V(\widehat{\beta}_{CIC}) = A^{-1}(\widehat{\beta}_{CIC})B(\widehat{\beta}_{CIC})A^{-1}(\widehat{\beta}_{CIC}) \tag{2.6}$$

where $A(\widehat{\beta}_{CIC}) = -\partial\mathcal{U}_{CIC}(\beta)/\partial\beta|_{\beta=\widehat{\beta}_{CIC}}$ and $B(\widehat{\beta}_{CIC}) = \sum_{i=1}^{n} u_i^*(\widehat{\beta}_{CIC})u_i^*(\widehat{\beta}_{CIC})^T$ with

$$u_i^*(\beta) = \delta_i W_i(X_i)\left\{Z_i - \frac{S_W^{(1)}(\beta, X_i)}{S_W^{(0)}(\beta, X_i)}\right\} - \sum_{j=1}^{n} \frac{I(X_i \geq X_j)\delta_j W_j(X_j)W_i(X_j)\exp\{\beta Z_i\}}{S_W^{(0)}(\beta, X_j)} \times \left\{Z_i - \frac{S_W^{(1)}(\beta, X_j)}{S_W^{(0)}(\beta, X_j)}\right\}.$$

Further, we have that $\widehat{\beta}_{CIC} \dot{\sim} N(\beta_{CIC}, V(\beta_{CIC}))$, where $V(\beta_{CIC}) = A^{-1}BA^{-1}$ is consistently estimated by $V(\widehat{\beta}_{CIC})$ in Equation 2.6.

Proof. Let $X_i = \min(T_i, C_i)$, $N_i(t) = I(X_i \leq t, \delta_i = 1)$, and $Y_i(t) = I(X_i \geq t)$ be the observed time, right-continuous counting process, and left-continuous at-risk process for the i^{th} individual, respectively. We also define the left-continuous weight process $W_i(t) = 1/\widehat{S}_C(t|Z_i)$, and the history process, or filtration, is defined as $\mathcal{F}_t^- = \{N_i(u), W_i(u^+), Y_i(u^+), Z_i;\ i = 1, \ldots, n;\ 0 \leq u < t\}$. We assume that (T_i, C_i, Z_i) are i.i.d. observations and that T_i and C_i are independent conditional on Z_i. Then, we define $S_W^{(r)}(\beta, t) = \sum_{j=1}^{n} Y_j(t)W_j(t)Z_j^r \exp\{\beta^T Z_j\}$ for $r = 0, 1, 2$, and $s_W^{(r)}(\beta, t) = ES_W^{(r)}(\beta, t)$. To prove this, we will appeal to Theorem 5.3 in Kalbfleisch and Prentice (2002) which implies Rebolledo's Central Limit Theorem. Theorem 5.3 requires demonstration that the conditions below hold.

There exists an open neighborhood $\mathcal{B}$ of β_{CIC} and functions $s_W^{(r)}(\beta, t)$ for $r = 0, 1, 2$, defined on $\mathcal{B} \times [0, \tau]$ which satisfy the following:

1. $\sup_{\beta\in\mathcal{B},t\in[0,\tau]} ||n^{-1}S_W^{(r)}(\beta,t) - s_W^{(r)}(\beta,t)|| \xrightarrow{p} 0$ as $n \to \infty$.

2. $s_W^{(0)}(\beta,t)$ is bounded away from 0 for $t \in [0,\tau]$.

3. For $r = 0,1,2$, $s_W^{(r)}(\beta,t)$ is a continuous function of β uniformly in $t \in [0,\tau]$, and $s_W^{(1)}(\beta,t) = \partial s_W^{(0)}(\beta,t)/\partial\beta$ and $s_W^{(2)}(\beta,t) = \partial^2 s_W^{(0)}(\beta,t)/\partial\beta^2$.

4. $\sum(\beta,\tau) = \int_0^\tau v(\beta,u)s_W^{(0)}(\beta,u)\lambda_0(u)du$ is positive definite for all $\beta \in \mathcal{B}$.

5. Z_i are bounded for all $t \in [0,\tau]$.

6. $\int_0^\tau \lambda_0(u)du < \infty$.

We start by assuming $P(Y_i(\tau) > 0) > 0$, or that there is positive probability that subject i is at risk over the inferential support interval, which requires us to have stability at τ. Since the risk set doesn't diminish, $ES_W^{(0)}(\beta,t)$ is always positive, implying Condition 2. Moreover, Condition 6 is also implied by contradiction. We will reasonably assume that Condition 5 holds. With this, the assumption of i.i.d. observations implies that $S_W^{(r)}(\beta,t)$ are comprised of independent terms and $S_W^{(r)}(\beta,t)$ converges to $s_W^{(r)}(\beta,t)$ for $r = 0,1,2$, by the Strong Law of Large Numbers. We will further assume that $P(X_i \geq t|Z_i)$ is continuous in t. Then, the fact that there exists a neighborhood $\mathcal{B}$ such that with probability one, the convergence is uniform on the set $\mathcal{B} \times [0,\tau]$ follows since $s_W^{(r)}(\beta,t)$ must be continuous in t by the Boundedness Convergence Theorem. Also, for any $t \in [0,\tau]$, the Strong Law of Large Numbers implies that $S_W^{(r)}(\beta,t) \to s_W^{(r)}(\beta,t)$ as $n \to \infty$ with probability one. We then have a sequence $S_W^{(r)}(\beta,t)$ of bounded monotonic functions (indexed by n) converging point-wise to the bounded monotonic functions $s_W^{(r)}(\beta,t)$, implying that the convergence must be uniform, or equivalently that Condition 1 holds. Condition 5 along with the Dominated Convergence Theorem imply Condition 3. We also assume that the positive definiteness from Condition 4 holds.

Now, we must establish that the estimating equation can be written as $\mathcal{U}(t) = \sum_{i=1}^n \int_0^t G_i(t)$

$dM_i(t)$ for some predictable process $G_i(t)$ and martingale $M_i(t)$, both with respect to $\mathcal{F}_t$. Define $G_i(t) = Z_i - E(Z_i)$. Since Z is assumed to be bounded, this quantity is predictable with respect to $\mathcal{F}_t$. The compensator of $W_i(t)N_i(t)$ is $H_i(t) = \int_0^\infty W_i(u)Y_i(u)\lambda_0(u)\exp(\beta^T Z)du$, giving $M_i^W(t) = W_i(t)N_i(t) - H_i(t)$ as a mean-zero martingale with respect to the filtration $\mathcal{F}_t$. We can write the estimating equation as $\mathcal{U}(\widehat{\beta}) = \sum_{i=1}^n \int_{t=0}^\infty \left[Z_i - \frac{S_W^{(1)}(\widehat{\beta},t)}{S_W^{(0)}(\widehat{\beta},t)} \right] dM_i^W(t) = 0$. Thus, the i^{th} term of $\mathcal{U}$ is a sum of a stochastic integral of a predictable process with respect to a martingale, implying $\mathcal{U}$ is a mean-zero martingale with respect to $\mathcal{F}_t$. Then, $n^{-1/2}\mathcal{U}(\beta_{CIC}, t) = \sum_{i=1}^n \left(Z_i - \frac{S_W^{(1)}(\widehat{\beta},t)}{S_W^{(0)}(\widehat{\beta},t)} \right) dM_i^W(t)$ is a martingale with predictable covariation process $\langle n^{-1/2}\mathcal{U}(\beta_{CIC})\rangle(t) = \int_0^t A(\beta_{CIC}, u)S_W^{(0)}(\beta_{CIC}, u)\lambda_0(u)du$. Generalizing this convergence result to the misspecified model is a direct application of White (1982a).

Let (X_i, Z_i) have some distribution G, and suppose there exists a unique point β_{CIC} such that $E_G(\mathcal{U}(X, Z; \beta_{CIC})) = 0$. Then under standard regularity conditions, the solution $\widehat{\beta}_{CIC}$ to the score $\mathcal{U}_{CIC}(\beta) = 0$ is asymptotically normally distributed with mean β_{CIC} and covariance matrix consistently estimated by $V(\widehat{\beta}_{CIC})$, given by Equation 2.6. □

2.3 The Nested Case-Control Design

The NCC design is included in the family of efficient sampling designs used for time-to-event data. Proposed by Thomas (1977), the NCC design is useful for maximizing information while reducing the number of subjects requiring full covariate information. This design makes use of the fact that most of the statistical information from the PH model comes from the subjects who experience events. As such, it requires covariate information from all who experience the event of interest, *cases*, and only a subset of participants who do not experience the event, *controls* (Nuño and Gillen, 2017). At each event time, the NCC sampling design draws a simple random sample (SRS) of m controls from the risk set to use in the analysis. The value of m is specified *a priori* and is typically chosen to be between one and four (Nuño and Gillen, 2017). Because m is typically small, the total sample size

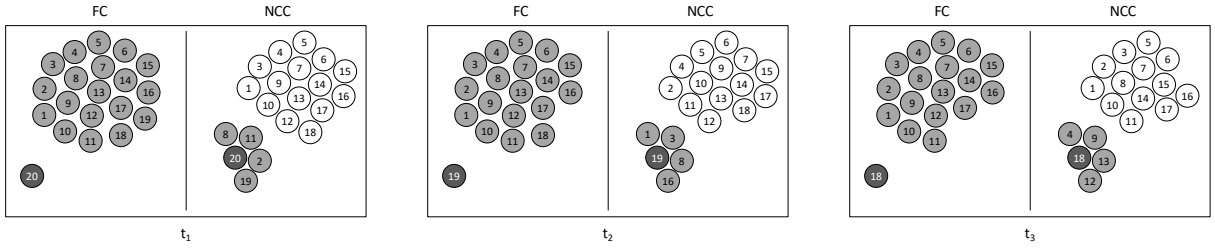


Figure 2.1: Hypothetical example of NCC sampling at three time points with cases (dark gray) and controls (light gray) used in the full cohort (left) and NCC design (right) with $m = 4$. White circles represent subjects that were not sampled in the analysis.

required under the NCC design can be considerably smaller than that of the full cohort with comparatively little loss of precision when the event is rare (Goldstein and Langholz, 1992).

Figure 2.1 shows an example of NCC design sampling compared to the full cohort at three hypothetical time points (t_1, t_2, t_3). At the first event time t_1, subject 20 experiences the event of interest. The left side of each panel represents the risk set under the full cohort, and the right side of the panel represents the data under the NCC design. Under the full cohort (left), subjects 1-19 are at risk and included in the analysis. Under the NCC design with $m = 4$ (right), only subjects 2, 8, 11, and 19 are included in the risk set for the first event along with subject 20. These four subjects were randomly sampled from the set of individuals who are at risk (not including the case). This process is repeated at all event times. Generally speaking, subjects are sampled without replacement at each event time but can be sampled for multiple risk sets under the NCC design. Subject 8 is an example of this, appearing in the sampled controls at times t_1 and t_2. Moreover, subjects that are cases can be sampled as controls at earlier event times, as seen by subject 19 who is a case at time t_2 but was sampled as a control at time t_1.

Relating this back to our estimating equation, consider taking an SRS of m controls from the risk set, R_i, at each event time X_i. Then, the resulting estimating equation is:

$$\mathcal{U}_{NCC}(\beta) = \sum_{i=1}^{n} \delta_i \left[Z_i - \frac{\sum_{j \in \tilde{R}_i} Z_j \exp\{\beta^T Z_j\}}{\sum_{j \in \tilde{R}_i} \exp\{\beta^T Z_j\}} \right] \equiv 0.$$

Note that this is similar to Equation 2.1, but the risk set, $\tilde{R}_i$ denotes the set of the m controls that were sampled from R_i at time X_i along with subject i. When the PH assumption holds, estimates obtained under the NCC design are consistent for the full cohort and asymptotically normally distributed (Goldstein and Langholz, 1992).

Current Extensions of the NCC Design

Several extensions of the NCC design have been proposed. Langholz and Thomas (1991) first explored alternate ways of using sampled controls in the analysis including a retained NCC design that used sampled controls forward in time, an augmented NCC design that used controls forward and backward in time, and an NCC design that implemented sampling without replacement. The retained NCC sampling borrows ideas from the case-cohort sampling since sampled controls are still considered an SRS at future event times. While this design remained generally unbiased in simulation studies, the authors found that it was surprisingly less efficient than the standard NCC design due to the positive sampling-induced covariances between observations in tested cohorts. The augmented NCC sampling aimed to improve efficiency of the NCC design by using sampled controls forward and backward in time. The naive use of these sampled controls in all risk sets would lead to bias from over-representation of individuals who survive for longer. To avoid this, Langholz and Thomas (1991) proposed stratified sampling by path sets defined by an individual's time-on-study to ensure a representative risk set despite using sampled controls forward and backward in time. This design resulted in consistent parameter estimation but did not significantly improve efficiency compared to the classic NCC design in the simulated settings. The NCC sampling without replacement originated from the idea of minimizing the number of controls that are sampled more than once. Similar to the augmented NCC sampling, simply sampling without replacement would lead to inconsistent estimates of the hazard ratio due to under-representation of individuals who survive for longer. As such, the same path sets are utilized and sampled with predetermined probabilities of sampling depending on their over-

all representation in the risk set. Then, individuals are sampled without replacement from the selected path set. As expected, this design resulted in more controls sampled overall, increasing precision. However, when compared to classic NCC designs using similar overall sample sizes, there was no evidence for a consistent gain in efficiency.

To improve efficiency of the NCC design, Langholz and Borgan (1995) proposed counter-matched stratified sampling when a surrogate is available. In counter-matched stratified sampling, controls are sampled from the stratum of subjects with a different covariate value than the case to increase efficiency. With a similar goal, Samuelsen (1997) proposed the use of controls forward and backward in time by including sampled controls in all risk sets for which they were at risk under the full cohort, leading to a risk set similar to that of the case-cohort design. They adjusted for any bias induced by a non-SRS using inverse probability of sampling weights (See Section 2.3.1). Building on this, Rivera and Lumley (2016a,b) combined the previous two ideas by implementing counter-matched stratified sampling with the use of sampled controls forward and backward in time.

As discussed earlier, it is important to consider the performance of estimators when model assumptions are violated. To this end, Nuño and Gillen (2020) explored the behavior of the NCC design estimators under a violation of the PH assumption. They demonstrated that the NCC design estimate is dependent on the censoring distribution as well as the number of controls m in the same way that estimates from the Cox PH model are dependent on the censoring distribution under NPH. In practice, this implies that in addition to being dependent on participant dropout and accrual patterns, estimates are also dependent on the choice of m for the study. In the setting of a single, binary covariate, they proposed an estimator that recovers the full cohort estimand by removing dependence on m through weighting of the estimating equation by a consistent estimate of the number of subjects at risk with each covariate value. For a continuous covariate with a misspecified functional form, Nuño and Gillen (2021) proposed an imputation algorithm to recover the full cohort

estimand. In both settings, the proposed estimators use information from previously sampled risk sets to maintain predictability for asymptotic arguments. Using methods analogous to those proposed by Boyd et al. (2012), they extend both estimators to be robust to changes in the censoring distribution. Both estimators also support the inclusion of adjustment covariates by using hot-deck imputation to map back to a risk set that is representative of the full cohort.

2.3.1 The Samuelsen Estimator

The work in Chapters 3 and 4 of this dissertation build upon the Samuelsen estimator. Under the originally proposed NCC design (Thomas, 1977), controls are only counted in the risk sets for which they were sampled. To improve efficiency of the NCC design, Samuelsen (1997) proposed the use of controls forward and backward in time by including all sampled controls in all risk sets for which they were at risk under the full cohort. In this case, the risk sets at each event time no longer represent an SRS of controls, nor do the terms of the resulting estimating equation have a covariance of zero conditional upon the filtration of the counting process giving rise to the data.

Samuelsen (1997) proposed an estimator that reweights each control's contribution to the estimating equation by the inverse of the probability of being included in the NCC sample. To this end, conditional on the filtration, $\mathcal{F} \equiv \{(X_i, \delta_i); i = 1, \ldots, n\}$, the probability that subject j ever experiences an event or is sampled as a control under the NCC design is:

$$p_j = \left\{ \begin{array}{ll} 1 - \prod_{X_i < X_j} \left(1 - \frac{m}{|\#R_i| - 1} \delta_i\right), & \text{if } \delta_j = 0 \\ 1, & \text{if } \delta_j = 1 \end{array} \right\}$$

where $|\#R_i| = \sum_{j=1}^{n} I(X_j \geq X_i)$ is the size of the risk set at time X_i. Furthermore, they define V_{j0} as an indicator for whether a subject j is ever sampled as a control, and $V_j = \max(V_{j0}, \delta_j)$. We note that the indicator V_{j0} only considers controls, whereas the

indicator V_j includes cases and sampled controls. The resulting estimating equation for the Samuelsen estimator is:

$$\mathcal{U}_{SAM}(\beta) = \sum_{i=1}^{n} \delta_i \left[Z_i - \frac{\sum_{j \in R_i} \frac{V_j}{p_j} Z_j \exp\{\beta^T Z_j\}}{\sum_{j \in R_i} \frac{V_j}{p_j} \exp\{\beta^T Z_j\}} \right] \equiv 0 \tag{2.7}$$

where R_i is the risk set for the full cohort, but the indicator V_j limits the contribution to participants who were included in the NCC sample.

Letting $\widehat{\beta}_{SAM}$ denote the solution to Equation 2.7, it can be shown that the variance of $\widehat{\beta}_{SAM}$ can be consistently estimated by:

$$\widehat{\Gamma} = \widehat{\Sigma}^{-1} + \widehat{\Sigma}^{-1} \Delta \widehat{\Sigma}^{-1}$$

where

$$\widehat{\Sigma} = I_{SAM}(\beta) = -\frac{\partial \mathcal{U}_{SAM}(\beta)}{\partial \beta}|_{\beta = \widehat{\beta}_{SAM}}$$

and

$$\Delta = \sum_{i=1}^{n} \left\{ V_i \frac{1 - p_i}{p_i} \mathcal{W}_i^T \mathcal{W}_i + \sum_{i \neq j} \frac{V_i V_j}{p_i p_j} \frac{(1 - p_i)(1 - p_j)\rho_{ij}}{p_i p_j} \mathcal{W}_i^T \mathcal{W}_j \right\}$$

with $\mathcal{W}_j$ denoting the score residual for subject j from the Cox PH model weighted by V_j/p_j and

$$\rho_{ij} = \prod_{k: X_k < \min(X_i, X_j)} \left\{ 1 - 2\frac{m\delta_k}{Y_k - 1} + \frac{m(m-1)\delta_k}{(Y_k - 1)(Y_k - 2)} \right\} \Big/ \left(1 - \frac{m\delta_k}{Y_k - 1} \right)^2 - 1.$$

Although there is no proof for asymptotic normality of this estimator due to a violation of the assumption of a predictable Martingale process, Samuelsen (1997) provides a heuristic argument for asymptotic normality that is supported by simulation results.

Efficiency of the Samuelsen Estimator

As mentioned previously, the Samuelsen estimator increases precision under the classic NCC design by allowing for the use of sampled controls forward and backward in time. Here, we compare the performance and efficiency of the full cohort estimator (FC), classic NCC design estimator, and the Samuelsen estimator using a simulation study. In the following simulations, $Z_1 \sim N(0.2 * Z_2, 1)$ represents a continuous predictor of interest, and $Z_2 \sim Bern(0.1)$ represents an *a priori*-specified binary adjustment covariate. To generate observations for the full cohort, $T_i \sim Exp(\lambda_i(t))$, $C_i \sim U(0.5, 3)$, $X_i = \min(C_i, T_i)$, and $\delta = I(X_i = T_i)$. We considered the data generating model: $\lambda(t) = \lambda_0 \exp\{\log(0.85)Z_1 + \log(1.4)Z_2 + \log(0.45)Z_1Z_2\}$, where $\lambda_0 = 0.06$ was chosen to yield a censoring rate of ∼90%, reflecting settings where the NCC design is often used. All simulations were run for 1000 iterations.

			$\beta_1^* = -0.349$				$\beta_2^* = 0.578$			
(89% Censoring)		Total N	$\widehat{\beta}_1$ (% Diff.)	Emp. SE	MB SE	MB Efficiency	$\widehat{\beta}_2$ (% Diff.)	Emp. SE	MB SE	MB Efficiency
FC		1000	-0.369 (0.00)	0.102	0.097	1.000	0.572 (0.00)	0.226	0.216	1.000
m = 1	NCC	208	-0.342 (-7.18)	0.151	0.148	1.527	0.460 (-19.45)	0.343	0.344	1.593
	Samuelsen	208	-0.390 (5.76)	0.150	0.140	1.446	0.610 (6.65)	0.340	0.319	1.475
m = 2	NCC	293	-0.339 (-7.99)	0.129	0.124	1.279	0.479 (-16.25)	0.271	0.284	1.311
	Samuelsen	293	-0.372 (1.00)	0.127	0.119	1.227	0.591 (3.31)	0.271	0.269	1.244
m = 3	NCC	367	-0.344 (-6.80)	0.118	0.115	1.192	0.498 (-12.81)	0.264	0.263	1.216
	Samuelsen	367	-0.372 (0.84)	0.118	0.111	1.149	0.584 (2.22)	0.264	0.250	1.157
m = 4	NCC	432	-0.348 (-5.48)	0.115	0.111	1.145	0.504 (-11.92)	0.248	0.251	1.163
	Samuelsen	432	-0.370 (0.25)	0.113	0.107	1.105	0.574 (0.33)	0.249	0.240	1.111

Table 2.1: Simulation results comparing the classic NCC design estimator and the Samuelsen estimator with an SRS for 1000 iterations. Percent difference and model based efficiency is compared to the full cohort. *Truth is based on estimates from a full cohort model with $n = 100,000$.

We implemented the classic NCC design sampling scheme which takes an SRS of controls from the risk set at each event time, and we used the unique subjects from this sample for the Samuelsen estimator. We fit a main effects model of the form: $\lambda(t) = \lambda_0(t) \exp\{\beta_1 Z_1 + \beta_2 Z_2\}$, which one might consider using in a primary analysis of the marginal effect of Z_1. We note that although this model is technically misspecified, it mimics a common analysis strategy (i.e. estimating main effects for a primary analysis before exploring potential effect modification) when effect modification may be present. Table 2.1 presents results comparing

the classic NCC design, which considers controls only at the times they are sampled and uses the Cox PH model for estimation, to the Samuelsen estimator, which uses sampled controls in all eligible risk sets. Especially for larger values of m, the Samuelsen estimator yields results similar to that of the full cohort estimator with respect to coefficient estimation. Compared to the NCC design, the Samuelsen estimator does a significantly better job at estimating both β_1 and β_2 and is more stable overall. Moreover, the efficiency of the Samuelsen estimator is better (i.e. closer to 1) than efficiency of the NCC design estimator for all choices of m in this setting.

2.3.2 Issues Under the NCC Design

In many settings, the NCC design can help maximize information while reducing data collection costs. When there is an underrepresented subpopulation, however, it can exacerbate preexisting issues caused by initial underrepresentation. As mentioned in Chapter 1, this is prevalent in AD research where Black, Hispanic, and Asian participants are less likely to participate in clinical research but are disproportionately affected by AD. Since the NCC design only samples a subset of participants from the underrepresented group into the analysis, it may be difficult to estimate subpopulation-specific effects with precision. In the most extreme case where the NCC design may not sample any of the participants from an underrepresented group, this would be impossible.

To demonstrate the lack of precision in a secondary analysis examining subpopulation-specific effects, we adopt the simulation setting from Section 2.3.1 and fit a model with an interaction between Z_1 and Z_2 of the form: $\lambda(t) = \lambda_0(t)\exp\{\beta_1 Z_1 + \beta_2 Z_2 + \gamma Z_1 Z_2\}$. Figure 2.2 compares the average estimated subpopulation-specific effects of the covariate of interest using the Samuelsen estimator with the classic NCC sampling design. For the underrepresented subpopulation ($Z_2 = 1$), the lack of precision due to fewer individuals being sampled is evident in the wider confidence intervals for estimates for that group.

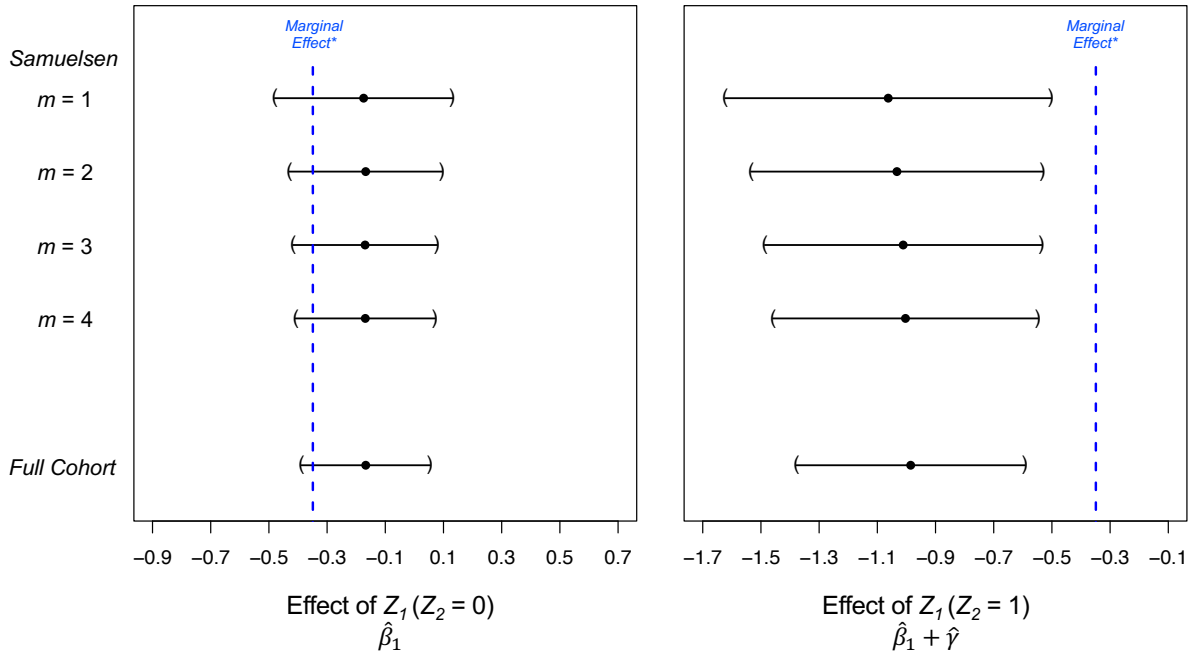


Figure 2.2: Forest plots comparing the average estimated subpopulation-specific effects in the underrepresented (right) and overrepresented (left) subpopulations under classic NCC sampling design with the Samuelsen estimator. *Marginal effect taken from the primary analysis in Section 2.3.1.

2.4 Multiple Event Survival Analysis

Chapters 4 and 5 pivot from univariate survival data to data with multiple time-to-event endpoints. Multiple event time-to-event data encompass three main types: (1) recurrent event, or sequentially occurring events of the same type (e.g., repeated hospitalizations); (2) multiple-type events, or sequential events of different types (e.g., disease progression stages); and (3) clustered or correlated events, or events occurring in related units such as family members or litter mates (Kelly and Lim, 2000). Examples of recurrent events of the same type include recurrences of seizures or hospitalizations due to a particular ailment. An example of multiple-type events is progression in AD where patients may progress to MCI and then dementia. Clustered/correlated events can include disease onset in family members or litter mates, and unlike the first two settings where event times are clustered at the individual level (i.e. multiple measurements for an individual), event times are clustered

at the group level (i.e. one measurement each from multiple members). Similarly, potential correlation between event times typically originates from multiple measurements within an individual for the first two settings, and from multiple measurements for related units in a cluster for the last type. We also note that while competing and semi-competing risks can be perceived as "multiple-type" events, the analysis methods presented in this dissertation are not immediately applicable to those settings.

Multiple event survival data are typically modeled using two approaches. Frailty models incorporate random effects analogous to those in linear or generalized linear models. Alternatively, marginal models extend univariate survival analysis techniques by accounting for within-cluster correlation through post-hoc variance adjustments. Specifically, when analyzing correlated survival data using a marginal model, the standard Cox PH model is sufficient as long as an appropriate model is chosen and the covariances are adjusted to account for correlation (Lin, 1994). Based on the characteristics of the data, one can choose a multiplicative hazards model using four key components - risk intervals, baseline hazard, risk set, and correlation adjustment. A handful of multiplicative hazards models that can be characterized by their different specifications of the four components have been proposed to handle these types of data. These include the Andersen and Gill (Andersen and Gill, 1982); Prentice, Williams, and Peterson gap-time and total-time (Prentice et al., 1981); Lee, Wei, and Amato (Lee et al., 1992); and Wei, Lin, and Weissfeld (Wei et al., 1989) models.

The choice of risk interval plays a significant role in the analysis of multiple event survival data because it determines when a subject is considered at risk for experiencing an event and can alter the interpretation of the resulting estimate. There are three main risk intervals considered in the multiple event setting: gap time, total time, and counting process. The gap time formulation considers the time between successive events, resetting the clock after each event. Due to this, it is typically used in settings of recurrent, ordered events, where an individual is not considered at risk of experiencing the k^{th} event until they have experienced

the $(k-1)^{th}$ event. In contrast, the total time formulation considers the time from a fixed origin to the occurrence of the event, and it is commonly used when analyzing events of different types or clustered events since primary interest lies in the relationship between covariates and the comparable time to an event across event types. Under the total time formulation, all units in a cluster start from time 0, and their total time until each event is measured. The counting process formulation utilizes a total time scale from a fixed origin but allows delayed entry for sequential events. That is, the units in a cluster are at risk for the same amount of time as the gap time formulation, but the risk interval for the k^{th} event starts at the time when the $(k-1)^{th}$ event occurs. Figure 2.3 demonstrates the different risk intervals for three hypothetical clusters.

Another consideration for multiple event survival data is the modeling choices for the baseline hazard. When modeling event times via the hazard function, two model specifications arise naturally depending on whether the baseline hazard varies by event type. This distinction reflects different assumptions about the event-generating process. When the occurrence of one event type does not influence the hazard of subsequent event types occurring within a cluster, a common baseline hazard model may be appropriate. An example of this occurs when analyzing the impact of a biomarker on time-to-progression to AD in family members, where each cluster is a group of related individuals. Here, one family member's progression does not impact the hazard of progression for other members conditional on being related. This type of model has also been utilized when analyzing the repeated occurrence of basal cell carcinoma, where there is no clear evidence that development of a basal cell carcinoma leads to development of another conditional on other risk factors (Pandeya, 2005). Conversely, there may also be settings where it is appropriate to allow baseline hazards to vary across event strata. Specifically, when the occurrence of an event may be related to the occurrence of another event within the same cluster, a model that assumes different baseline hazards across event types might be necessary to accurately compute the marginal hazards. Some examples of this include stroke recurrences where the risk of having a stroke increases with

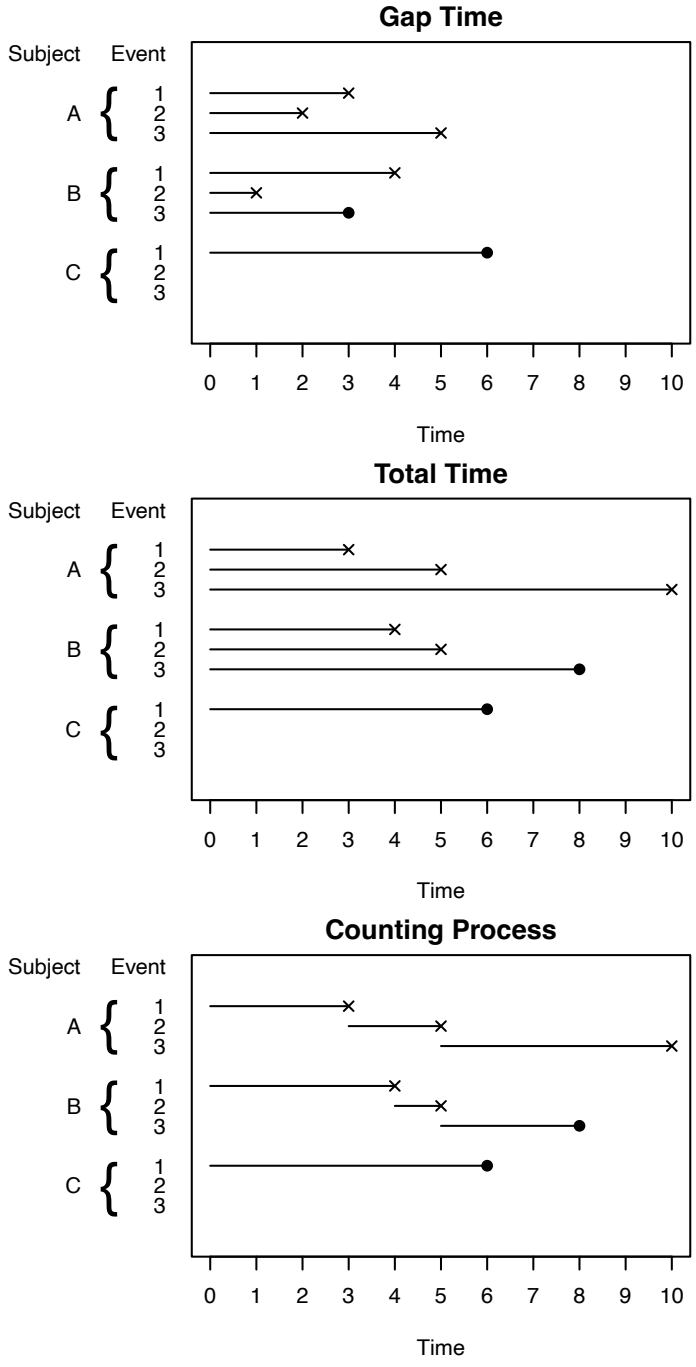


Figure 2.3: Hypothetical illustration of three different risk intervals for three subjects. Events are indicated by $\times$, and censored observations are indicated by $\bullet$.

each recurrence and progression from cognitively normal to MCI, and MCI to dementia, where the risk of progression might be different between each of the cognitive stages. Models with event-specific baseline hazards are also common in myocardial infarction, where an individual's risk of experiencing a second heart attack increases after the first (Böhm et al., 2021). As such, two types of models can be constructed based on the specification of the baseline hazard, and we define these rigorously in the next section.

In the multiple event setting, specification of the covariate effects are also important and depend on the scientific question of interest. For example, one can estimate a common covariate effect across all event strata or an event-specific covariate effect (De Jong et al., 2020). If the relationship between the events and the covariate of interest is expected to change or if interest lies in the specific covariate effects at each event type, then event-specific, or stratified effects may be more appropriate. If the covariate effect is expected to be similar across event strata, however, estimating a common effect across events can help increase precision in settings of a small sample size (Amorim and Cai, 2015). For the work in this dissertation, we will focus on categorizing models based on the specification of the baseline hazard (i.e. common vs. event-specific) in Chapter 4, and later, the choice of the type of covariate effect (i.e. common vs. event-specific) in Chapter 5.

2.4.1 Marginal Parameter Estimation

We consider the setting of right-censored, multiple time-to-event data. To maintain clarity in our presentation of these concepts, we will assume that clustering is done at the individual level, or equivalently, that multiple events are nested within individuals. The methods presented can also be applied to data clustered on the group level, however. For subject i ($i = 1, \ldots, n$) and event type k ($k = 1, ..., K_i$), let T_{ik}, C_{ik}, and $Z_{ik} = [z_{ik1}, \ldots, z_{ikp}]^T$ denote the true event time, censoring time, and $p \times 1$ covariate vector, respectively. We define $X_{ik} = \min(T_{ik}, C_{ik})$ as the observed event time and $\delta_{ik} = I(X_{ik} = T_{ik})$ as the event

indicator of event k for subject i.

We model event times via the hazard function, and as mentioned previously, two natural models arise based on the specification of the baseline hazard:

- Common baseline hazards (M1): $\lambda_{ik}(t) = \lambda_0(t)\exp\{\beta^T Z_{ik}\}$
- Event-specific baseline hazards (M2): $\lambda_{ik}(t) = \lambda_{0k}(t)\exp\{\beta^T Z_{ik}\}$

We note that these models can be easily modified to accommodate event-specific covariate effects through modification of the parameter vector β (i.e. β vs. β_k).

The parameters β for the PH model in the multiple event setting can be estimated by maximizing the partial likelihood assuming working independence, or equivalently by solving the respective estimating equations below (Lin, 1994):

$$\mathcal{U}_{M1}(\beta) = \sum_{i=1}^{n}\sum_{k=1}^{K_i}\delta_{ik}\left[Z_{ik} - \frac{\sum_{l=1}^{n}\sum_{j=1}^{K_l} Y_{lj}(X_{ik})Z_{lj}\exp\{\beta^T Z_{lj}\}}{\sum_{l=1}^{n}\sum_{j=1}^{K_l} Y_{lj}(X_{ik})\exp\{\beta^T Z_{lj}\}}\right] \equiv 0 \tag{2.8}$$

for M1, and

$$\mathcal{U}_{M2}(\beta) = \sum_{i=1}^{n}\sum_{k=1}^{K_i}\delta_{ik}\left[Z_{ik} - \frac{\sum_{l=1}^{n}\sum_{j=1}^{K_l} I(j=k)Y_{lj}(X_{ik})Z_{lj}\exp\{\beta^T Z_{lj}\}}{\sum_{l=1}^{n}\sum_{j=1}^{K_l} I(j=k)Y_{lj}(X_{ik})\exp\{\beta^T Z_{lj}\}}\right] \equiv 0 \tag{2.9}$$

for M2, where $Y_{lj}(X_{ik}) = I(X_{lj} \geq X_{ik})$ is the indicator representing whether observation lj has an observed time greater than X_{ik}. The two estimating equations, which correspond to the two types of models, differ by the sums in the compensator that represent the risk set at that event time. Under the assumption of a common baseline hazard across all event types (M1), the risk sets at each event time are similar to the case of treating all observations as independent. That is, at an event time X_{ik}, the risk set is comprised of all lj pairs that have not yet experienced the j^{th} event type or been censored, so we sum over all measurements (j) for all individuals (l). When allowing for event-specific baseline hazard functions (M2),

the risk sets at each event time are comprised only of the observations at risk with that same event type, which is captured by the indicator $I(j = k)$. Specifically, at event time X_{ik}, the risk set is comprised of all k^{th} event-type observations that have not yet experienced the event or been censored, so we sum over all individuals' k^{th} events.

2.4.2 Robust Variance and Asymptotic Properties

It can be shown that Equations 2.8 and 2.9 can be approximated by a sum of n i.i.d. random vectors, and asymptotic normality can be demonstrated using methods similar to those presented in Section 2.2.2. Due to potential correlation between an individuals' event times, standard variance estimators assuming independence are inconsistent in general. In this case, use of the sandwich or robust variance estimator can be employed (Lin, 1994). Specifically, we have that $\widehat{\beta}^{(M)} \dot{\sim} N(\beta^{(M)}, V)$ where $\beta^{(M)}$ is the true value of β for model M ($M = 1, 2$), and V is the variance. The variance of the parameters for model M, $V(\widehat{\beta}^{(M)})$, can be consistently estimated by:

$$V(\widehat{\beta}^{(M)}) = I^{-1}(\widehat{\beta}^{(M)})\ B(\widehat{\beta}^{(M)})\ I^{-1}(\widehat{\beta}^{(M)})$$

where $I(\widehat{\beta}^{(M)}) = -\frac{\partial \mathcal{U}^{(M)}(\beta)}{\partial \beta}|_{\beta=\widehat{\beta}^{(M)}}$ and

$B(\widehat{\beta}^{(M)}) = \sum_{i=1}^{n} \sum_{k=1}^{K_i} \sum_{l=1}^{K_i} W_{ik}^{(M)}(\widehat{\beta}^{(M)}) W_{il}^{(M)}(\widehat{\beta}^{(M)})^T$ with

$$W_{ik}^{(1)}(\widehat{\beta}) = \delta_{ik}\left[Z_{ik} - \frac{S^{(1)}(\widehat{\beta}, X_{ik})}{S^{(0)}(\widehat{\beta}, X_{ik})}\right] - \sum_{l=1}^{n}\sum_{j=1}^{K_l} \delta_{lj}\left[\frac{Y_{ik}(X_{lj}) \exp\{\widehat{\beta}^T Z_{ik}\}}{S^{(0)}(\widehat{\beta}, X_{lj})}\right] \times \left\{Z_{ik} - \frac{S^{(1)}(\widehat{\beta}, X_{lj})}{S^{(0)}(\widehat{\beta}, X_{lj})}\right\}$$

where $S^{(r)}(\beta, X_{ik}) = \sum_{l=1}^{n} \sum_{j=1}^{K_l} Y_{lj}(X_{ik}) Z_{lj}^{\otimes r} \exp\{\beta^T Z_{lj}\}$ for $r = 0, 1$, or

$$W_{ik}^{(2)}(\widehat{\beta}) = \delta_{ik}\left[Z_{ik} - \frac{S_k^{(1)}(\widehat{\beta}, X_{ik})}{S_k^{(0)}(\widehat{\beta}, X_{ik})}\right] - \sum_{l=1}^{n}\sum_{j=1}^{K_l} I(j=k)\delta_{lj}\left[\frac{Y_{ik}(X_{lj})\exp\{\widehat{\beta}^T Z_{ik}\}}{S_k^{(0)}(\widehat{\beta}, X_{lj})}\right] \times \left\{Z_{ik} - \frac{S_k^{(1)}(\widehat{\beta}, X_{lj})}{S_k^{(0)}(\widehat{\beta}, X_{lj})}\right\}$$

where $S_k^{(r)}(\beta, X_{ik}) = \sum_{l=1}^{n} \sum_{j=1}^{K_l} I(j=k) Y_{lj}(X_{ik}) Z_{lj}^{\otimes r} \exp\{\beta^T Z_{lj}\}$ for $r = 0, 1$.

2.4.3 NCC Design in the Multiple Event Setting

In the multiple event setting, terminal events such as death can occur. These events stop the recurrent event process and can be informative since they may be associated with the number of recurrent events experienced by an individual. In addition to marginal models that account for the potential dependency of the terminal event, parametric joint frailty models can also be used in these settings. One specific joint frailty model by Liu et al. (2004) focuses on modeling both recurrent and terminal events and has been extended by Rondeau et al. (2006) to model baseline intensity functions using splines for estimation of the intensities and absolute risks as well as for computational ease.

Using this model, Jazić et al. (2019) developed the first framework for use of the NCC design in the recurrent event setting, specifically in the presence of a terminal event. To our knowledge, this is the only current methodology that utilizes the NCC design in the recurrent event setting. For this method, the authors propose various designs that involve defining different "index events," or the cases to be used for sampling. Some examples of index events include 1) the terminal event, 2) the first instance of a recurrent event, or 3) the first of a recurrent or terminal event. Under this design, covariate information is collected for all individuals who experience the specified index event, and at each of those index event times, m individuals are sampled from those who have not yet experienced the specified index event. The NCC design cohort then consists of all index events (cases) and sampled

controls for whom recurrent and terminal event times as well as covariate information are available. To gain the most information from sampled individuals, the authors also proposed designs based on index events including 4) the second instance of the recurrent event and 5) the first of the second instance of the recurrent event or terminal event. They also use the inverse probability of being sampled into the analysis, similar to the work of Samuelsen (1997), to account for using a non-SRS.

Chapter 3

Modified Nested Case-Control Designs for Assessing Effect Modification in Underrepresented Populations

3.1 Introduction

As discussed previously, efficient sampling designs are often used to reduce data collection costs, including non-monetary costs associated with the use of existing resources, in settings where resources are limited. The NCC design helps maximize information while reducing the number of participants needed for the analysis by only requiring full covariate information from participants who experience the event of interest, *cases*, and a subset of participants who do not experience the event, *controls* (Thomas, 1977; Nuño and Gillen, 2017). To review briefly, the standard NCC sampling design requires drawing an SRS of m controls from the risk set at each event time. The value of m is specified *a priori* and is typically chosen to

be between one and four.

As we have seen in Chapter 1, the NCC design is useful in settings such as AD biomarker discovery where measurements may be burdensome or costly to obtain. Since many study participants are unwilling to undergo a lumbar puncture, CSF is a scarce resource in clinical studies (Nuño et al., 2017). Additionally, the thawing process leads not only to depletion of the sample, but also to specimen degradation over time. In this setting, use of the NCC design allows for a reduction in the number of processed CSF samples, saving precious resources.

Another challenge in clinical research is the lack of racial and ethnic representation, and its impacts are detrimental to historically underrepresented communities (Wendler et al., 2005; Chin et al., 2011). In AD research, study cohorts are predominantly comprised of highly-educated, non-Hispanic, White individuals (Raman et al., 2021), which can limit the generalizability of results. In fact, a study by Salazar et al. (2020) showed that participants from underrepresented racial and ethnic groups (e.g. Hispanic, Black, and Asian) were less likely to participate in procedures that are typical of AD clinical trials. Therefore, in addition to efficient use of limited samples, it is important that underrepresented groups are included in AD studies to allow for generalizability of results to these groups.

Two-phase sampling has been proposed to reduce data collection costs when there exists an inexpensive surrogate that is available for the entire cohort, and non-SRS methods have been implemented under this framework for settings where a biased sample may be useful. Under two-phase sampling, information that is easier to collect is used to sample a subcohort from which the expensive covariates will be measured (White, 1982b). The NCC design can be considered a type of two-phase sampling where the subjects' event time and status determines which subjects are sampled in the second phase. Although it is useful for reducing data collection costs, the NCC design can exacerbate the pre-existing issue of underrepresentation in research, as we have demonstrated in Section 2.3.2. This is because only a

subset of participants from underrepresented subpopulations are selected into the analysis, making analyzing the effect of a limited resource, such as CSF, in an underrepresented subpopulation difficult and hindering our ability to assess potential effect modification across subpopulations.

Various alternatives to an SRS in the second phase of sampling have been explored to improve efficiency while maintaining valid inference in the presence of expensive covariates (Tao et al., 2020; Robins et al., 1994; Zhou et al., 2007). Specifically, Robins et al. (1995) proposed an estimator that incorporates inverse probability of selection weighting to correct for a non-SRS two-phase sampling when the missingness of an expensive covariate is conditional on the outcome and other covariates. Under the NCC design, several biased sampling methods have also been proposed to improve efficiency. To revisit some of the designs that were mentioned in Chapter 2, Langholz and Borgan (1995) proposed counter-matched stratified sampling when a surrogate for exposure is available, and Samuelsen (1997) proposed inclusion of the NCC controls forward and backward in time, leading to a risk set similar to that of the case-cohort design (Prentice and Pyke, 1979). Rivera and Lumley (2016a,b) then extended this idea with a non-SRS, implementing counter-matched stratified sampling with the use of sampled controls forward and backward in time. In counter-matched stratified sampling, controls are sampled from the stratum of subjects with a different covariate value from the case to increase efficiency.

In this chapter, we propose an extension of the NCC sampling scheme that allows for weighted sampling of controls using observed covariate information, combining key ideas from two-phase sampling with the NCC framework. Our motivation differs from past work, as it aims to increase efficiency for assessing effect modification in a secondary analysis through pre-specified stochastic sampling of controls. In the AD biomarker discovery setting where trials often lack diversity, such an approach would increase the probability of sampling individuals from underrepresented racial and ethnic groups under the NCC design to ensure assessment

of biomarker utility across these subpopulations. Preliminary results illustrating this can be found in Section 2.3.2. We propose methods for weighted sampling based on either a binary or a continuous covariate under the NCC design. In each setting, we propose an estimator that accounts for the weighted sampling along with two robust variance estimators to ensure consistent parameter estimation and valid asymptotic inference. We use simulation studies to compare the proposed estimator to the full cohort estimator and to the standard NCC design with estimation carried out via the approach of Samuelsen (Thomas, 1977; Samuelsen, 1997). Lastly, we apply our proposed estimator to data from the NIH-funded National Alzheimer's Coordinating Center (NACC) where we consider estimating differential associations of Aβ and the risk of AD between Black and non-Black populations.

3.2 Methods

We start with a brief review of the foundational concepts presented in Chapter 2. Here, we consider the setting of right-censored time-to-event data. Let T_i, C_i, and $Z_i = [z_{i1}, \ldots, z_{ip}]^T$ denote the true event time, censoring time, and vector of covariates for subject i, ($i = 1, \ldots, n$), respectively. We define $X_i = \min(T_i, C_i)$ as the observed event time and $\delta_i = I(X_i = T_i)$ as the event indicator for subject i. Under the Cox PH model, the event times are modeled by the covariates via the hazard function:

$$\lambda_i(t) = \lambda_0(t) \exp\{\beta^T Z_i\}$$

where $\lambda_0(t)$ denotes the baseline hazard function and $\beta = [\beta_1, \ldots, \beta_p]^T$ is the vector of regression parameters representing the corresponding log-hazard ratios associated with each covariate. The parameters for the Cox PH model are estimated by maximizing the partial likelihood, or equivalently by solving the estimating equation given by:

$$\mathcal{U}_{FC}(\beta) = \sum_{i=1}^{n} \delta_i \left[Z_i - \frac{\sum_{j \in R_i} Z_j \exp\{\beta^T Z_j\}}{\sum_{j \in R_i} \exp\{\beta^T Z_j\}} \right] \equiv 0. \tag{3.1}$$

Here, R_i represents the risk set at time X_i. Specifically, when all subjects enter at time 0, R_i denotes the set of individuals that have not yet experienced the event of interest and have not been censored prior to time X_i, including the individual who experienced the current event.

Under the NCC design, only a subset of the full cohort consisting of all n individuals is used for parameter estimation. Specifically, at each event time, m controls are randomly sampled from R_i, and the resulting estimating equation is:

$$\mathcal{U}_{NCC}(\beta) = \sum_{i=1}^{n} \delta_i \left[Z_i - \frac{\sum_{j \in \tilde{R}_i} Z_j \exp\{\beta^T Z_j\}}{\sum_{j \in \tilde{R}_i} \exp\{\beta^T Z_j\}} \right] \equiv 0.$$

Note that this is similar to Equation 3.1, but the risk set, $\tilde{R}_i$, denotes the set of the m controls that were sampled from R_i at time X_i along with subject i, who experienced the event.

Under the originally proposed NCC design (Thomas, 1977), controls are only counted in the risk sets for which they were sampled. To improve efficiency of the NCC design, Samuelsen (1997) proposed the use of controls forward and backward in time by including all sampled controls in all risk sets for which they were at risk under the full cohort. Since the risk sets at each event time no longer represent an SRS of controls, the proposed estimator reweights each control's contribution to the estimating equation by the inverse of the probability of being included in the NCC sample. The probability that subject j ever experiences an event or is sampled as a control under the NCC design is:

$$p_j = \left\{ \begin{array}{ll} 1 - \prod_{X_i < X_j} \left(1 - \frac{m}{|\#R_i| - 1} \delta_i \right), & \text{if } \delta_j = 0 \\ 1, & \text{if } \delta_j = 1 \end{array} \right\}$$

where $|\#R_i| = \sum_{j=1}^{n} I(X_j \geq X_i)$, the size of the risk set at time X_i. Moreover, V_{j0} is an indicator for whether a subject j is ever sampled as a control, and $V_j = \max(V_{j0}, \delta_j)$. The

resulting estimating equation for the Samuelsen estimator is:

$$\mathcal{U}_{SAM}(\beta) = \sum_{i=1}^{n} \delta_i \left[Z_i - \frac{\sum_{j \in R_i} \frac{V_j}{p_j} Z_j \exp\{\beta^T Z_j\}}{\sum_{j \in R_i} \frac{V_j}{p_j} \exp\{\beta^T Z_j\}} \right] \equiv 0,$$

where R_i is the risk set for the full cohort, but the indicator V_j limits the contribution to participants who were included in the NCC sample.

3.2.1 Weighted Sampling on a Binary Covariate

We propose a method that allows for oversampling of select subpopulations under the NCC design. Here, we consider the case where a binary covariate is observed for all members of the full cohort and indicates that the subject is from an underrepresented subpopulation. We oversample members from the underrepresented subpopulation by first increasing the probability that controls from that group are sampled at each event time to an *a priori*-specified probability of sampling. We then correct for this oversampling by reweighting contributions to the estimating equation. We call the NCC design with weighted sampling the weighted NCC (WNCC) design.

To implement the WNCC design, let Z denote a binary covariate on which weighted sampling is implemented. Let $\pi_z(X_i)$ be the proportion of subjects at risk at X_i with $Z = z$ for $z = 0, 1$. We denote the desired probability of sampling a subject with $Z = z$ at each event time as π_z^*, then we can construct sampling weights for the two groups at time X_i as $w_z(X_i) = \pi_z^*/\pi_z(X_i)$. As the risk set sizes change, the sampling weights for each individual should change. Instead of using the traditional SRS of controls, we utilize this weighted sample of controls for the calculation of our proposed estimator.

Under this weighted sampling scheme, we can extend the idea of inclusion probabilities under a non-SRS (Rivera and Lumley, 2016b) to redefine the probability of a subject j being included in the study under the WNCC design, conditional on subject j's covariate

value Z_j as:

$$p_j^* = \left\{ \begin{array}{ll} 1 - \prod_{X_i < X_j} \left(1 - \frac{\pi_{z_j}^* m}{Y_{z_j}(X_i) - I(z_i = z_j)} \delta_i\right), & \text{if } \delta_j = 0 \\ 1, & \text{if } \delta_j = 1 \end{array} \right\}$$

where $Y_{z_j}(X_i)$ is the number of subjects at risk with $Z = z_j$ at time X_i. This is true because the probability that subject j is sampled as a control at time X_i is conditional on the value of z_j. At time X_i, the expected number of controls sampled with $Z = z_j$ is equal to $\pi_{z_j}^* m$, and the number of subjects with $Z = z_j$ who can be sampled is $Y_{z_j}(X_i) - I(z_i = z_j)$ with $I(z_i = z_j)$ being an indicator for whether subjects i and j have the same covariate value. It follows that the probability that subject j is sampled as a control at time X_i is $\frac{\pi_{z_j}^* m}{Y_{z_j}(X_i) - I(z_i = z_j)} \delta_i$.

We further define V_{j0}^* as an indicator for whether subject j is ever sampled as a control under the WNCC design, and $V_j^* = \max(V_{j0}^*, \delta_j)$ as the indicator for whether subject j is included in the analysis as a case or control under the WNCC design.

Proposition 3.1. *Denote β_0 to be the estimand corresponding to the full cohort estimator and let $\widehat{\beta}_B$ be the solution to*

$$\mathcal{U}_B(\beta) = \sum_{i=1}^{n} \delta_i \left[Z_i - \frac{\sum_{j \in R_i} \frac{V_j^*}{p_j^*} Z_j \exp\{\beta^T Z_j\}}{\sum_{j \in R_i} \frac{V_j^*}{p_j^*} \exp\{\beta^T Z_j\}} \right] \equiv 0.$$

Then $\widehat{\beta}_B \xrightarrow{p} \beta_0$.

The proof of this result is provided in Section A.1 of Appendix A. We note also that the methods from this section can easily be extended to a categorical covariate with any number of levels.

3.2.2 Variance Estimation for $\widehat{\beta}_B$

Estimating the variance of $\widehat{\beta}_B$ is done in a similar fashion to that of the Samuelsen estimator, which follows the form of a sandwich estimator accounting for correlated terms of the estimating equation. Here, the variance of $\widehat{\beta}_B$ can be estimated as:

$$\widehat{\Gamma}_{B1} = \widehat{\Sigma}_B^{-1} + \widehat{\Sigma}_B^{-1}\Delta_B\widehat{\Sigma}_B^{-1}$$

where

$$\widehat{\Sigma}_B = I_B(\widehat{\beta}_B) = -\frac{\partial \mathcal{U}_B(\beta)}{\partial \beta}|_{\beta=\widehat{\beta}_B}$$

and

$$\Delta_B = \sum_{i=1}^{n}\Bigg\{ V_i^* \frac{1-p_i^*}{p_i^*}\mathcal{W}_i^*(\widehat{\beta}_B)^T\mathcal{W}_i^*(\widehat{\beta}_B) + \sum_{i\neq j}\frac{V_i^*V_j^*}{p_i^*p_j^*}\frac{(1-p_i^*)(1-p_j^*)\rho_{ij}^*}{p_i^*p_j^*}\mathcal{W}_i^*(\widehat{\beta}_B)^T\mathcal{W}_j^*(\widehat{\beta}_B)\Bigg\}$$

with $\mathcal{W}_i^*(\widehat{\beta}_B)$ being the score residual for subject i from the Cox PH model weighted by V_i^*/p_i^*,

$$\mathcal{W}_i^*(\widehat{\beta}_B) = \delta_i\frac{V_i^*}{p_i^*}\left[Z_i - \frac{\sum_{j\in R_i}\frac{V_j^*}{p_j^*}Z_j\exp\{\widehat{\beta}_B^TZ_j\}}{\sum_{j\in R_i}\frac{V_j^*}{p_j^*}\exp\{\widehat{\beta}_B^TZ_j\}}\right] - \frac{V_i^*}{p_i^*}\sum_{j=1}^{n}\delta_j\frac{V_j^*}{p_j^*}\left[\frac{I(X_j\leq X_i)\exp\{\widehat{\beta}_B^TZ_i\}}{\sum_{k\in R_j}\frac{V_k^*}{p_k^*}\exp\{\widehat{\beta}_B^TZ_k\}}\right]\times\left\{Z_i - \frac{\sum_{k\in R_j}\frac{V_k^*}{p_k^*}Z_k\exp\{\widehat{\beta}_B^TZ_k\}}{\sum_{k\in R_j}\frac{V_k^*}{p_k^*}\exp\{\widehat{\beta}_B^TZ_k\}}\right\}$$

and

$$\rho_{ij}^{*} = \prod_{k:X_k<\min(X_i,X_j)} \Bigg\{1 - \frac{\pi_{z_i}^{*} m\delta_k}{Y_{z_i}(X_k) - I(z_k = z_i)} - \frac{\pi_{z_j}^{*} m\delta_k}{Y_{z_j}(X_k) - I(z_k = z_j)} + \frac{\pi_{z_i}^{*} m}{Y_{z_i}(X_k) - I(z_k = z_i)} \cdot \frac{\pi_{z_j}^{*}(m-1)}{Y_{z_j}(X_k) - I(z_k = z_j) - I(z_i = z_j)}\delta_k \Bigg\} \Bigg/ \left(1 - \frac{\pi_{z_i}^{*} m\delta_k}{Y_{z_i}(X_k) - I(z_k = z_i)}\right)\left(1 - \frac{\pi_{z_j}^{*} m\delta_k}{Y_{z_j}(X_k) - I(z_k = z_j)}\right) - 1.$$

As we have established in Chapter 2, a series of conditions must hold in order to demonstrate asymptotic normality of the Cox PH estimator. Similar to the Samuelsen estimator, we are unable to directly apply Rebolledo's Central Limit theorem to establish asymptotic normality due to a violation of the assumption of a predictable Martingale process in this case. As such, asymptotic normality of the WNCC estimator, along with the classic Samuelsen estimator, has yet to be formally established. Empirical results presented in Section 3.3 do, however, support the conjecture of asymptotic normality, and we provide a heuristic argument for this conjecture in Section A.2 of the Appendix.

We also propose an alternate variance estimator which takes a form similar to that proposed by Binder (1992) and uses the weights V_j^{*}/p_j^{*}, derived in this manuscript. This estimator accounts for the correlation between terms of the estimating equation via an empirical estimate of the score. That is, the variance of $\widehat{\beta}_B$ can also be estimated as:

$$\widehat{\Gamma}_{B2} = \widehat{\Sigma}_B^{-1} V_B \widehat{\Sigma}_B^{-1}$$

where

$$V_B = \sum_{i=1}^{n} \mathcal{W}_i^{*}(\widehat{\beta}_B)\mathcal{W}_i^{*}(\widehat{\beta}_B)^T.$$

3.2.3 Weighted Sampling on a Continuous Covariate

Weighted sampling can also be implemented for a continuous covariate. Suppose z is a continuous covariate which is known for the full cohort and on which weighted sampling is implemented. Using kernel density estimation, we can obtain $f_i(z)$, the estimated distribution of z for subjects with an observed time greater than X_i, at each event time X_i. We can also define $f^*(z)$ to be the desired distribution of z in the risk set at each time (e.g. a uniform density over the support of z). Using these two functions at each event time X_i, we can define a sampling weight for subject j who is at risk as $w_j^*(X_i) = f^*(z_j)/f_i(z_j)$. These sampling weights can then be used in the WNCC design.

If $f_i(z)$ is a uniformly consistent estimator of the true distribution function of z for subjects with an observed time greater than X_i, and $f_i(z_j) > 0$ for all j such that $X_j > X_i$, then we can define a consistent estimator for the probability that a subject j is included in the study under the WNCC design with weighted sampling on a continuous covariate, p_j^{**}:

$$p_j^{**} = \left\{ \begin{array}{ll} 1 - \prod_{X_i < X_j} \left(1 - \frac{w_j^*(X_i) m}{\sum_{l \in R_i} w_l^*(X_i)} \delta_i\right), & \text{if } \delta_j = 0 \\ 1, & \text{if } \delta_j = 1 \end{array} \right\}$$

Here, V_{j0}^* and V_j^* are defined as in Section 3.2.1.

Proposition 3.2. *Denote β_0 as the estimand corresponding to the full cohort estimator and let $\widehat{\beta}_C$ be the solution to*

$$\mathcal{U}_C(\beta) = \sum_{i=1}^{n} \delta_i \left[Z_i - \frac{\sum_{j \in R_i} \frac{V_j^*}{p_j^{**}} Z_j \exp\{\beta^T Z_j\}}{\sum_{j \in R_i} \frac{V_j^*}{p_j^{**}} \exp\{\beta^T Z_j\}} \right] \equiv 0.$$

Then $\widehat{\beta}_C \xrightarrow{p} \beta_0$.

The proof of this result follows closely to the proof in Section A.1 of the Appendix. We further note that a continuous covariate can also be discretized so that the method from

Section 3.2.1 would be directly applicable.

3.2.4 Variance Estimation for $\widehat{\beta}_C$

Similar to the binary covariate case, the variance of $\widehat{\beta}_C$ can be estimated as:

$$\widehat{\Gamma}_{C1} = \widehat{\Sigma}_C^{-1} + \widehat{\Sigma}_C^{-1}\Delta_C\widehat{\Sigma}_C^{-1}$$

where

$$\widehat{\Sigma}_C = I_C(\widehat{\beta}_C) = -\frac{\partial \mathcal{U}_C(\beta)}{\partial \beta}|_{\beta=\widehat{\beta}_C}$$

and

$$\Delta_C = \sum_{i=1}^{n}\Bigg\{ V_i^* \frac{1-p_i^{**}}{p_i^{**}} \mathcal{W}_i^{**}(\widehat{\beta}_C)^T \mathcal{W}_i^{**}(\widehat{\beta}_C) + \sum_{i\neq j} \frac{V_i^* V_j^*}{p_i^{**}p_j^{**}} \frac{(1-p_i^{**})(1-p_j^{**})\rho_{ij}^{**}}{p_i^{**}p_j^{**}} \mathcal{W}_i^{**}(\widehat{\beta}_C)^T \mathcal{W}_j^{**}(\widehat{\beta}_C) \Bigg\}$$

with $\mathcal{W}_i^{**}(\widehat{\beta}_C)$ being the score residual for subject i from the Cox PH model weighted by V_i^*/p_i^{**},

$$\mathcal{W}_i^{**}(\widehat{\beta}_C) = \delta_i \frac{V_i^*}{p_i^{**}} \left[Z_i - \frac{\sum_{j\in R_i} \frac{V_j^*}{p_j^{**}} Z_j \exp\{\widehat{\beta}_C^T Z_j\}}{\sum_{j\in R_i} \frac{V_j^*}{p_j^{**}} \exp\{\widehat{\beta}_C^T Z_j\}} \right] - \frac{V_i^*}{p_i^{**}} \sum_{j=1}^{n} \delta_j \frac{V_j^*}{p_j^{**}} \left[\frac{I(X_j \leq X_i)\exp\{\widehat{\beta}_C^T Z_i\}}{\sum_{k\in R_j} \frac{V_k^*}{p_k^{**}} \exp\{\widehat{\beta}_C^T Z_k\}} \right] \times \left\{ Z_i - \frac{\sum_{k\in R_j} \frac{V_k^*}{p_k^{**}} Z_k \exp\{\widehat{\beta}_C^T Z_k\}}{\sum_{k\in R_j} \frac{V_k^*}{p_k^{**}} \exp\{\widehat{\beta}_C^T Z_k\}} \right\}$$

and

$$\rho_{ij}^{**} = \prod_{k:X_k<\min(X_i,X_j)} \Bigg\{1 - \frac{w_i^*(X_k)m\delta_k}{\sum_{l\in R_k} w_l^*(X_k)} - \frac{w_j^*(X_k)m\delta_k}{\sum_{l\in R_k} w_l^*(X_k)} + \frac{0.5\cdot m(m-1)w_i^*(X_k)w_j^*(X_k)}{\sum_{l\in R_k} w_l^*(X_k)(\sum_{l\in R_k} w_l^*(X_k) - w_i^*(X_k))} + \frac{0.5\cdot m(m-1)w_i^*(X_k)w_j^*(X_k)}{\sum_{l\in R_k} w_l^*(X_k)(\sum_{l\in R_k} w_l^*(X_k) - w_j^*(X_k))}\delta_k \Bigg\} \Bigg/ \left(1 - \frac{w_i^*(X_k)m\delta_k}{\sum_{l\in R_k} w_l^*(X_k)}\right)\left(1 - \frac{w_j^*(X_k)m\delta_k}{\sum_{l\in R_k} w_l^*(X_k)}\right) - 1.$$

Details regarding the derivation of the variance estimator and a heuristic argument for approximating the limiting distribution of our estimator are provided in Section A.2 of Appendix A. An alternate variance estimator for this method takes the form proposed by Binder (1992), and would be computed using the weights V_j^{**}/p_j^{**}. That is, the variance of $\widehat{\beta}_C$ can also be estimated as:

$$\widehat{\Gamma}_{C2} = \widehat{\Sigma}_C^{-1} V_C \widehat{\Sigma}_C^{-1}$$

where

$$V_C = \sum_{i=1}^{n} \mathcal{W}_i^{**}(\widehat{\beta}_C)\mathcal{W}_i^{**}(\widehat{\beta}_C)^T.$$

3.3 Empirical Results

In this section, we present results from simulation studies illustrating the empirical performance of our estimator. In all simulations presented, Z_1 (e.g. Aβ) is a continuous predictor of interest, and Z_2 (e.g. race/ethnicity) is an *a priori*-specified binary adjustment covariate on which we performed weighted sampling. Specifically, Z_2 defines the subpopulations of interest for assessing effect modification. We note that simulations for weighted sampling on

a continuous covariate demonstrate similar results.

In Section 3.3.1, we first illustrate the utility of implementing a weighted sampling scheme under the NCC design framework. Next, in Section 3.3.2, we illustrate the performance of the proposed estimator when there is a homogeneous effect of the predictor of interest across all subpopulations defined by the binary covariate. In Section 3.3.3, we illustrate the performance of the proposed estimator in a primary analysis under effect modification. We also demonstrate the benefit of the WNCC design in a secondary analysis of the subpopulation-specific effect of the predictor of interest. We provide an analysis of the power for testing the presence of differential subpopulation-specific effects in the secondary analysis under the NCC and WNCC designs.

Throughout this section, we consider the data generating model: $\lambda(t) = \lambda_0 \exp\{\log(0.85)Z_1 + \log(1.4)Z_2 + \xi Z_1 Z_2\}$. Here, the value of ξ, the coefficient controlling the strength of effect modification, varies, and $\lambda_0 = 0.06$ was chosen to yield a censoring rate of $\sim$90% to reflect settings under which the NCC design is often used. In the full cohort, $Z_2 \sim Bern(0.2)$, $Z_1 \sim N(0.2 \cdot Z_2, 1)$, and $T_i \sim Exp(\lambda_i(t))$. Moreover, $C_i \sim U(0.5, 3)$, and $X_i = \min(C_i, T_i)$. The event indicator (δ_i) was set to 1 if the subject's observed time is the event time and 0 if the subject's observed time was the censoring time. The WNCC design was implemented to have an equal probability of sampling from each Z_2 subgroup (i.e. $P(Z_2 = 0) = P(Z_2 = 1) = 0.5$) into the risk sets at each event time. All simulations were run for 1000 iterations.

3.3.1 Weighted Sampling Performance

For these simulations, we vary the distribution of Z_2 to illustrate differences in sampled cohorts from settings where the underrepresented subpopulation comprises of 10%, 20%, and 30% of the full cohort. That is, $Z_2 \sim Bern(p)$, where $p = 0.1, 0.2$, and 0.3. Weighted sampling is then implemented such that there is an equal probability of sampling each Z_2 subgroup (i.e. $Z_2 = 0$ and $Z_2 = 1$) from the risk set at each event time. Specifically, a

weighted sample of subjects are taken such that the probability of sampling a subject with $Z_2 = 0$ is 0.5 and the probability of sampling a subject with $Z_2 = 1$ is 0.5 at each event time.

Table 3.1 displays the average number of subjects sampled into an analysis under the NCC design with an SRS and with weighted sampling. We note that although we set the probability of sampling to be equal at each event time, the proportion of individuals from each group will not always be equal due to the resampling of the same individuals at different frequencies in each subgroup. This is also the reason we see fewer unique individuals under weighted sampling compared to an SRS. From these results, it is clear that there is utility in weighted sampling for increasing the number of individuals that are sampled from the underrepresented group.

		$Z_2 \sim Bern(0.1)$		$Z_2 \sim Bern(0.2)$		$Z_2 \sim Bern(0.3)$	
		Unique N	Rare N (Prop.)	Unique N	Rare N (Prop.)	Unique N	Rare N (Prop.)
FC		1000	100 (0.10)	1000	200 (0.20)	1000	300 (0.30)
m = 1	SRS	197	24 (0.12)	208	49 (0.24)	216	75 (0.35)
	Weighted	188	51 (0.27)	204	74 (0.36)	215	93 (0.43)
m = 2	SRS	279	32 (0.11)	293	65 (0.22)	304	100 (0.33)
	Weighted	251	69 (0.28)	280	104 (0.37)	298	130 (0.44)
m = 3	SRS	351	39 (0.11)	367	79 (0.22)	380	122 (0.32)
	Weighted	304	80 (0.26)	344	126 (0.37)	370	159 (0.43)
m = 4	SRS	414	45 (0.11)	432	92 (0.21)	447	141 (0.32)
	Weighted	350	87 (0.25)	400	142 (0.35)	432	183 (0.42)

Table 3.1: Comparison of average subjects sampled in analyses under an SRS and weighted sampling for 1,000 simulations.

3.3.2 Homogeneous Effect of the Predictor of Interest

As stated previously, we evaluated the performance of the WNCC design and the proposed estimator under a homogeneous effect (i.e. when $\xi = 0$). We fit a main effects model of the form: $\lambda(t) = \lambda_0(t) \exp\{\beta_1 Z_1 + \beta_2 Z_2\}$ to evaluate the ability of the proposed estimator to

correct the bias induced by implementing weighted sampling. We compare results from our proposed estimator to results from the Samuelsen estimator under the WNCC design. Below, we report the average number of unique subjects used in each analysis, average coefficient estimates for β_1 and β_2 with percent bias from the true values, average original standard error (SE 1) estimates, average alternate standard error (SE 2) estimates, empirical standard error (Emp. SE), and the coverage probabilities for the two standard error estimates (CP 1 and CP 2).

		$\beta_1 = -0.163$						$\beta_2 = 0.336$					
$\lambda_0 = 0.06$	Unique N	$\widehat{\beta}_1$ (% Bias)	SE 1	SE 2	Emp. SE	CP 1	CP 2	$\widehat{\beta}_2$ (% Bias)	SE 1	SE 2	Emp. SE	CP 1	CP 2
FC (89% cens.)	1000	-0.164 (-1.00)	0.098	0.097	0.098	0.948	0.952	0.320 (-4.88)	0.229	0.229	0.230	0.966	0.963
Samuelsen													
m = 1	196	-0.165 (-1.59)	0.144	0.145	0.152	0.941	0.954	-0.895 (-366.10)	0.301	0.302	0.312	0.013	0.015
m = 2	270	-0.166 (-2.17)	0.121	0.122	0.125	0.950	0.952	-0.778 (-331.36)	0.266	0.267	0.271	0.008	0.008
m = 3	333	-0.165 (-1.26)	0.113	0.113	0.114	0.944	0.949	-0.692 (-305.62)	0.253	0.254	0.257	0.012	0.012
m = 4	387	-0.163 (-0.43)	0.108	0.108	0.110	0.951	0.948	-0.594 (-276.65)	0.247	0.247	0.248	0.015	0.015
Proposed													
m = 1	196	-0.168 (-3.27)	0.147	0.149	0.156	0.944	0.953	0.306 (-9.05)	0.295	0.296	0.309	0.940	0.940
m = 2	270	-0.168 (-3.41)	0.123	0.124	0.127	0.952	0.950	0.320 (-4.85)	0.260	0.261	0.264	0.954	0.954
m = 3	333	-0.167 (-2.55)	0.114	0.115	0.116	0.944	0.948	0.312 (-7.33)	0.248	0.248	0.252	0.953	0.954
m = 4	387	-0.165 (-1.54)	0.110	0.110	0.112	0.948	0.947	0.323 (-4.04)	0.242	0.242	0.242	0.950	0.952

Table 3.2: Simulation results comparing the performance of the full cohort Cox PH estimator, the Samuelsen estimator under the WNCC design, and our proposed estimator under the WNCC design for a primary analysis when there is a homogeneous effect of the predictor of interest (i.e. $\xi = 0$).

Table 3.2 outlines the results from this analysis where the model is correctly specified and the WNCC design is used. The Samuelsen estimator with no correction for the weighted sampling yielded estimates with minimal bias for β_1, the effect of Z_1, and close to nominal coverage probabilities for both variance estimators. The proposed estimator maintained similar performance with both variance estimators. For β_2, the proposed estimator significantly corrected the bias induced (compared to the Samuelsen estimator with no correction) by oversampling subjects with $Z_2 = 1$. The proposed estimator had accurate standard error estimates from both variance estimators, performing as well as the full cohort estimator for bias in the case of four controls. We note that the proposed estimator also performs well when there is a null effect for the predictor of interest, and we include these results in

Section A.3 of Appendix A.

3.3.3 Effect Modification for the Predictor of Interest

In this section, we evaluated the performance of the WNCC design and the proposed estimator when effect modification is present (i.e. when $\xi \neq 0$). In this case, $\xi = \log(0.45)$. When there is potential for effect modification, it is common first to evaluate main effects in a primary analysis, then consider interactions in a secondary analysis. Here, we started by comparing the performance of the proposed estimator to the performance of the Samuelsen estimator under the WNCC design in a primary analysis setting by fitting a main effects model of the form: $\lambda(t) = \lambda_0(t)\exp\{\beta_1 Z_1 + \beta_2 Z_2\}$. We note that the model is misspecified here; however, this is typical in practice due to the complex relationships between variables in research. Since the true values of β_1 and β_2 are determined by the censoring distribution under model misspecification (Struthers and Kalbfleisch, 1986), we took full cohort estimates from analyses with $n = 100,000$ to be the "true values."

		$\beta_1 = -0.349^*$						$\beta_2 = 0.578^*$					
$\lambda_0 = 0.06$	Unique N	$\widehat{\beta}_1$ (% Bias)	SE 1	SE 2	Emp. SE	CP 1	CP 2	$\widehat{\beta}_2$ (% Bias)	SE 1	SE 2	Emp. SE	CP 1	CP 2
FC (89% cens.)	1000	-0.368 (-5.40)	0.097	0.101	0.099	0.938	0.950	0.562 (-2.71)	0.217	0.222	0.216	0.958	0.958
Samuelsen													
m = 1	203	-0.400 (-14.51)	0.148	0.153	0.161	0.928	0.943	-0.621 (-207.40)	0.294	0.303	0.313	0.025	0.029
m = 2	279	-0.386 (-10.48)	0.124	0.130	0.130	0.938	0.950	-0.511 (-188.42)	0.257	0.266	0.259	0.010	0.013
m = 3	343	-0.385 (-10.13)	0.115	0.120	0.123	0.933	0.946	-0.423 (-173.11)	0.244	0.252	0.244	0.011	0.016
m = 4	398	-0.381 (-9.21)	0.109	0.115	0.113	0.937	0.946	-0.327 (-156.58)	0.237	0.245	0.234	0.016	0.022
Proposed													
m = 1	203	-0.388 (-11.06)	0.149	0.153	0.161	0.937	0.948	0.558 (-3.55)	0.284	0.289	0.300	0.934	0.938
m = 2	279	-0.375 (-7.38)	0.125	0.129	0.129	0.949	0.954	0.564 (-2.49)	0.249	0.254	0.252	0.948	0.954
m = 3	343	-0.376 (-7.53)	0.116	0.120	0.122	0.942	0.950	0.556 (-3.75)	0.236	0.241	0.236	0.954	0.962
m = 4	398	-0.374 (-6.93)	0.111	0.114	0.112	0.945	0.951	0.564 (-2.38)	0.230	0.235	0.228	0.958	0.963

Table 3.3: Simulation results comparing the performance of the full cohort Cox PH estimator, the Samuelsen estimator under the WNCC design, and our proposed estimator under the WNCC design for a primary analysis when effect modification is present (i.e. $\xi \neq 0$). *Truth is based on estimates from a full cohort model with $n = 100,000$.

Table 3.3 outlines the results from this analysis where effect modification is present in the data generating model but not adjusted for in the fitted model under the WNCC design. Here, we observe that the proposed estimator performs better than the Samuelsen estimator

in correcting bias when estimating both β_1 and β_2. Again, this is to be expected due to the weighted sampling that was performed on the cohort used for analyses. There is some residual bias present with the proposed estimator; however, it performs nearly as well as the full cohort with $m = 4$ controls. The alternate variance estimator (SE 2) is more conservative than the original variance estimator (SE 1) for both β_1 and β_2, and the benefits of this are evident in the nominal coverage probabilities for the proposed estimator.

To demonstrate the utility of the WNCC design and the proposed estimator in a secondary analysis (for effect modification) setting, we fit a model with an interaction between Z_1 and Z_2 of the form: $\lambda(t) = \lambda_0(t) \exp\{\beta_1 Z_1 + \beta_2 Z_2 + \gamma Z_1 Z_2\}$, using the subcohort sampled under the WNCC design as one might do in practice. We conducted the same secondary analysis using an SRS instead of weighted sampling to compare the performance of the WNCC design with the proposed estimator to the classic NCC design with the Samuelsen estimator. We generated forest plots comparing the subpopulation-specific hazard ratio estimates obtained from both analyses for $Z_2 = 0$ (i.e. $\exp\{\widehat{\beta}_1\}$) and for $Z_2 = 1$ (i.e. $\exp\{\widehat{\beta}_1 + \widehat{\gamma}\}$).

Figure 3.1 illustrates the difference in precision for estimating the effect of the predictor of interest in the underrepresented (right) and overrepresented (left) subpopulations in a secondary analysis. Comparing the WNCC sampling scheme and proposed estimator to the classic NCC design and Samuelsen estimator, we observe from the width of the 95% confidence intervals that the estimates for the underrepresented subpopulation are more precise and nearly achieve the efficiency of the full cohort in the $m = 4$ case under the WNCC design. When implementing the WNCC design, we use fewer individuals compared to the NCC design. As a result, we lose some precision when estimating the effect of the predictor of interest in the overrepresented subpopulation. In the $m = 4$ case, the width of the confidence interval under the WNCC design is approximately 6% wider for the $Z_2 = 0$ group but 10% narrower for the $Z_2 = 1$ group. Generally, the loss of precision in the overrepresented group is less of a concern given the relative widths of the confidence intervals in each Z_2 group as

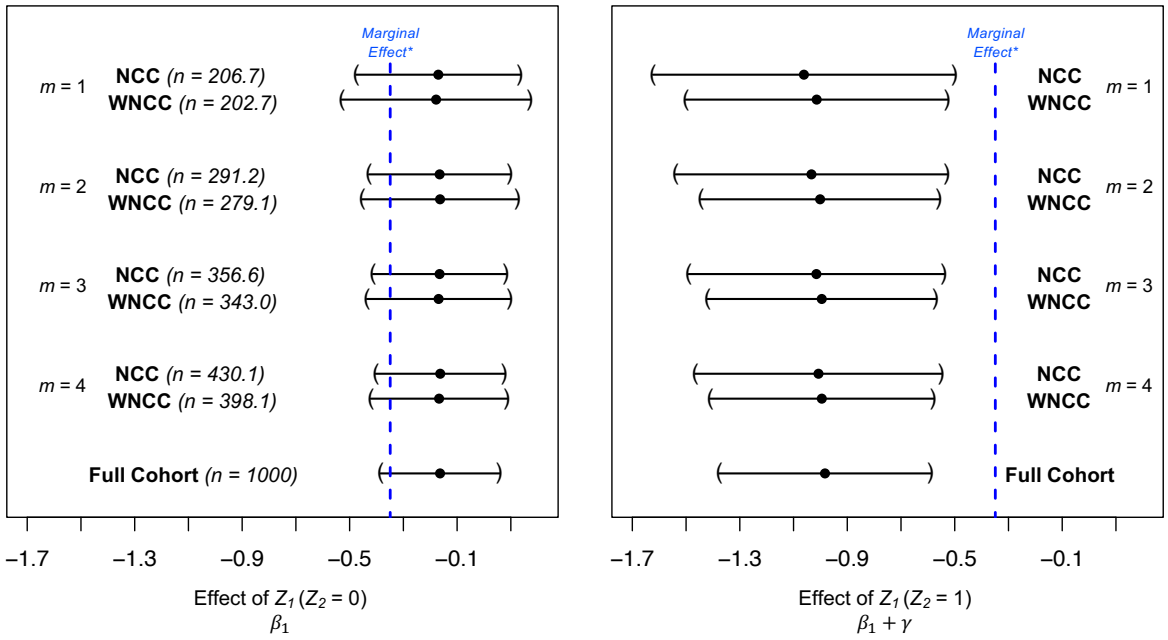


Figure 3.1: Forest plots comparing the WNCC design to the classic NCC design in a secondary analysis of subpopulation-specific effects. The right and left plots represent the estimated effect of Z_1 in the underrepresented and overrepresented subpopulations, respectively. Average total sample size (n) listed on the left side of the plot. *Marginal effect taken from the primary analysis.

well as our increased ability to study underrepresented groups which is often already limited.

We also conducted an analysis of the power for testing the hypothesis that there is a difference in the subpopulation-specific effects of Z_1, or $H_0 : \gamma = 0$, in a secondary analysis. We compared the power of this test under the WNCC design with the proposed estimator and the NCC design with the Samuelsen estimator under different values of ξ, the coefficient for the effect modification in the data generating model. We ran 1,000 iterations with $n = 1,000$ in the full cohort for each of the 10 evenly spaced values of ξ tested. For each iteration, we implemented both the WNCC and NCC design sampling schemes and used the resulting subcohorts in secondary analyses with the proposed and the Samuelsen estimators, respectively. Figure 3.2 presents the resulting power curves for each design where we observe an improvement in power for the analyses using the WNCC design compared to those using the classic NCC design. We note that the higher reported power under the WNCC and

NCC designs compared to the full cohort for smaller values of $|\xi|$ are due to stochasticity of simulations.

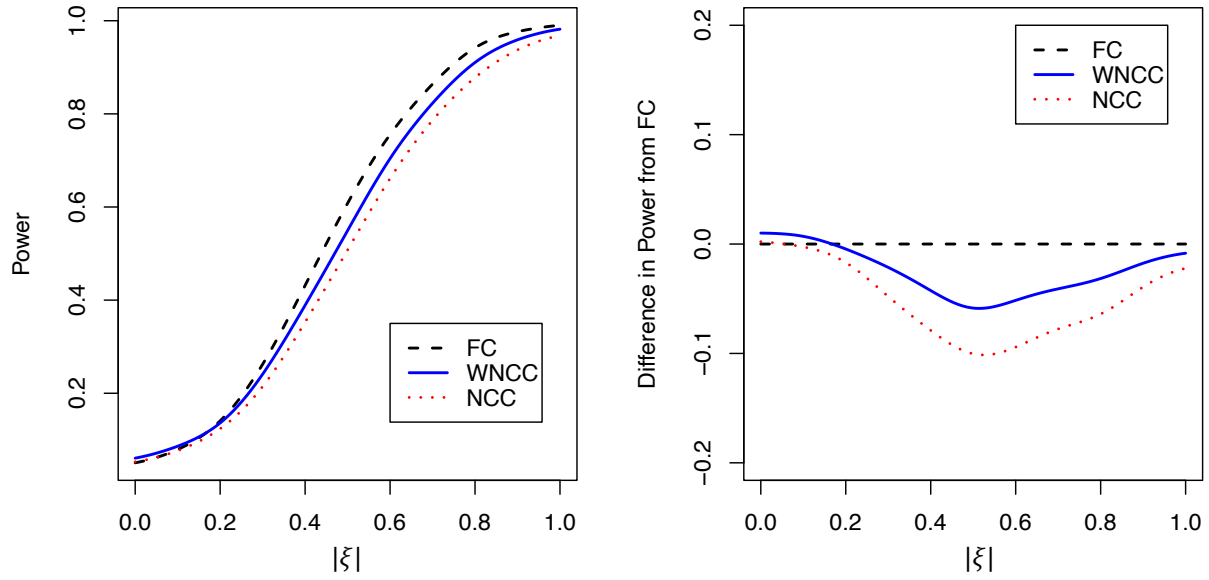


Figure 3.2: Power curves for simulations testing the hypothesis $H_0 : \gamma = 0$ comparing the WNCC design to the full cohort and NCC design analyses ($m = 4$) under multiple values of ξ, the coefficient controlling the strength of effect modification in the data generating model.

3.4 Application to NACC Data

As mentioned previously, the NCC design is especially relevant in settings where the event of interest is rare and at least one of the covariates is difficult to obtain. As such, we applied the proposed estimator to data from NACC to analyze the relationship between Aβ at baseline and time to progression to the first of MCI or dementia. We also conducted a secondary analysis of the subpopulation-specific associations between Aβ at baseline and time to progression in the Black and non-Black populations (Mayeda et al., 2016; Howell et al., 2017). We included 897 individuals, all with information regarding race, years of education, sex, age, APOE-e4 status, and with at least one Aβ measurement. These participants were all cognitively normal at the time of their first Aβ measurement. Since Aβ samples are both a rare resource and expensive to process, we utilized the NCC design as one might do in

practice. The NCC analysis was conducted as if Aβ were only processed for participants included in the NCC sample. In practice, full cohort information for a rare measure such as Aβ would not be available for all participants to conduct a full cohort analysis; however, we wanted to compare the performance of the NCC estimators to that of the full cohort.

Participants in this study were a majority non-Black (including Hispanic and non-Hispanic participants of White, Asian, Multiracial, and Native Hawaiian or Pacific Islander race), with only approximately 6.5% Black participants. Given the underrepresentation of Black individuals in this study and the fact that AD has been found to affect the Black population disproportionately, we performed weighted sampling to increase the probability of sampling Black individuals into the study under the WNCC design.

	Mean (sd) or Count (%) N = 897
Observed Time (yrs.)	7.3 (4.6)
Progressed	205 (22.9%)
Aβ (pg/mL)	684.6 (357)
Black	58 (6.5%)
Female	511 (57.0%)
≥ 1 APOE-e4 allele	326 (36.3%)
Age (yrs.)	69.4 (8.9)
Education (yrs.)	16.2 (2.6)

Table 3.4: Participant characteristics for those included in the analysis.

In the primary analysis, we adjusted for *a priori*-selected potential confounders including age, sex, APOE-e4 allele status, and years of education in addition to the predictor of interest, Aβ measurement, and race (Black vs. non-Black), the covariate used for weighted sampling. We performed weighted sampling such that the probability of sampling Black individuals and non-Black individuals at each event time was equal. We fit a Cox PH model to the full cohort, the Samuelsen estimator to a WNCC design sample, and the proposed estimator to the same WNCC sample. Table 3.5 displays results for the predictor of interest (Aβ) as well as the covariate on which weighted sampling was performed (Black vs. non-Black).

		Aβ (200 pg/mL)				Black/Non-Black			
	Unique N	Est. log HR (HR)	% Bias	SE 1 (p-value)	SE 2 (p-value)	Est.log HR (HR)	% Bias	SE 1 (p-value)	SE 2 (p-value)
FC (77% cens.)	897	-0.174 (0.840)	0.00	0.043 (<0.001)	0.042 (<0.001)	-0.291 (0.747)	0.00	0.389 (0.455)	0.413 (0.481)
Samuelsen									
m = 1	323	-0.099 (0.905)	42.83	0.055 (0.073)	0.054 (0.068)	-1.662 (0.190)	-470.85	0.445 (<0.001)	0.453 (<0.001)
m = 2	387	-0.144 (0.866)	16.93	0.045 (0.001)	0.043 (0.001)	-1.590 (0.204)	-446.19	0.432 (<0.001)	0.443 (<0.001)
m = 3	431	-0.163 (0.850)	6.21	0.046 (<0.001)	0.044 (<0.001)	-1.290 (0.275)	-343.07	0.413 (0.002)	0.421 (0.002)
m = 4	482	-0.178 (0.837)	-2.30	0.043 (<0.001)	0.042 (<0.001)	-1.046 (0.351)	-259.34	0.408 (0.010)	0.419 (0.012)
Proposed									
m = 1	323	-0.102 (0.903)	41.52	0.065 (0.120)	0.065 (0.117)	0.024 (1.025)	108.41	0.411 (0.953)	0.424 (0.954)
m = 2	387	-0.157 (0.855)	9.73	0.050 (0.002)	0.048 (0.001)	-0.351 (0.704)	-20.48	0.400 (0.381)	0.427 (0.412)
m = 3	431	-0.173 (0.841)	0.72	0.051 (0.001)	0.049 (<0.001)	-0.282 (0.754)	3.20	0.394 (0.474)	0.414 (0.496)
m = 4	482	-0.190 (0.827)	-9.58	0.047 (<0.001)	0.046 (<0.001)	-0.267 (0.766)	8.28	0.393 (0.497)	0.413 (0.518)

Table 3.5: Estimated results from NACC data for Aβ and Black vs. non-Black coefficients comparing the proposed estimator to the Samuelsen estimator with weighted sampling and full cohort analyses.

Under the full cohort, we estimate that the risk of progressing to MCI or dementia is approximately 16% lower for those with higher Aβ when comparing two subpopulations of similar age, sex, APOE-e4 status, and education but differing in baseline Aβ by 200 pg/mL. For larger values of m, our proposed estimator estimates this association with similar precision, as seen in Table 3.5. Using the WNCC sampling scheme, the number of subjects used was almost 50% less than the full cohort. Although this was not our primary goal, it highlights the utility of the efficient sampling design in reducing the potential costs associated with covariate collection. When estimating the coefficient associated with race (Black vs. non-Black), our proposed estimator corrects any potential bias induced from the use of the classic Samuelsen estimator with the weighted sampling.

In the secondary analysis, using the individuals sampled in the primary analysis under the WNCC sampling scheme, we fit a model with an interaction between Aβ and the indicator of Black race. Figure 3.3 shows the estimated hazard ratios and corresponding 95% confidence intervals (original variance estimator) for Aβ in the Black population and the non-Black population for the classic NCC design based on an SRS, the proposed WNCC design, and the full cohort. In this analysis, we observe a lower estimated hazard ratio for Aβ in the Black population compared to the full cohort marginal estimate, which was driven by the large

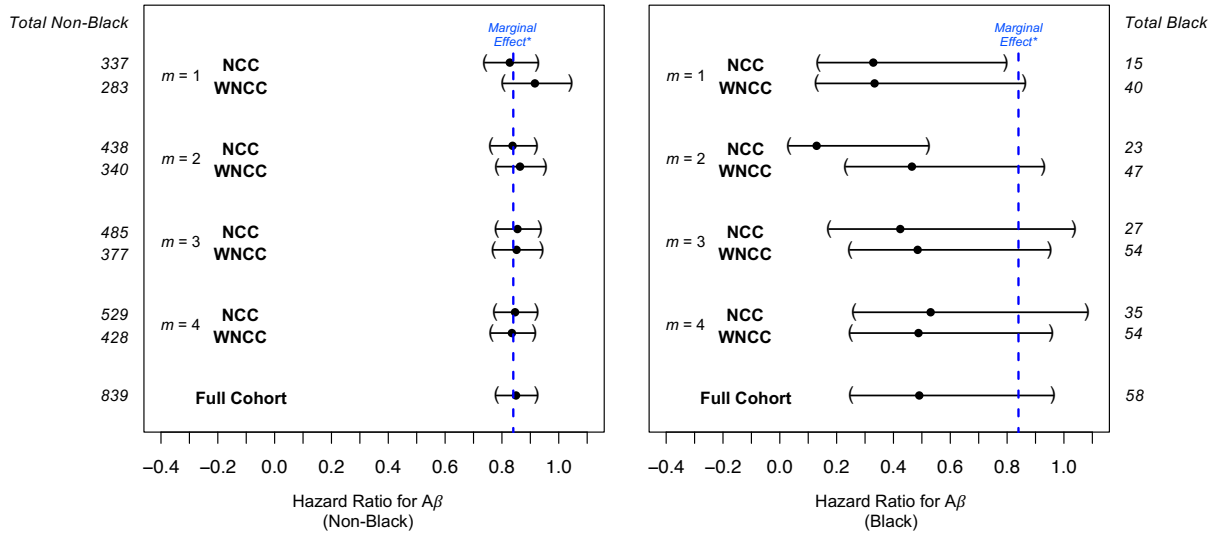


Figure 3.3: Forest plots comparing the WNCC design to the classic NCC design in a secondary analysis of subpopulation-specific hazard ratios for Aβ. The right and left plots represent the Aβ hazard ratios in the Black and non-Black subpopulations, respectively. Average total sample size (n) listed on the left side of the plot. *Marginal effect taken from the primary analysis.

proportion of non-Black participants in the full cohort analysis. Due to the oversampling of Black individuals into the WNCC sample, estimates from the proposed estimator are less biased than those provided by the classic NCC design (compared to the full cohort estimate). The WNCC estimator provides more stable estimates for a single draw compared to the classic NCC with an SRS, as seen in the $m = 2$ case. Moreover, when $m = 3$ and 4, the WNCC is nearly as efficient as the full cohort. For non-Black individuals, the estimated hazard ratios when $m = 3$ or 4 are also close to unbiased for the full cohort under both the WNCC and NCC designs. Despite undersampling non-Black individuals under the WNCC design, there is a minimal loss in efficiency (approximately 3% for the $m = 4$ case) for higher values of m.

It has been well-established that use of the NCC design can greatly reduce the number of subjects requiring full covariate information while maintaining consistency. The NCC design, however, can exacerbate underrepresentation of some subpopulations, and we pro-

	Unique N (% Black) FC = 897 (6.5)	
	SRS	Weighted Sampling
m = 1	328.1 (5.0)	316.9 (13.1)
m = 2	411.8 (5.4)	388.2 (12.6)
m = 3	472.0 (5.7)	442.9 (11.7)
m = 4	519.7 (5.8)	488.8 (10.9)

Table 3.6: Average number of unique individuals and proportion of unique individuals that are Black under an SRS and weighted sampling for 1000 iterations with NACC data.

posed weighted sampling to assuage this issue. To illustrate the utility of weighted sampling, we conducted 1000 iterations of random sampling on the NACC dataset under simple and weighted sampling schemes. The results, displayed in Table 3.6, demonstrate that under the proposed weighted sampling, the proportion of Black participants in the sample doubles for most choices of m compared to an SRS. The overall sample size is also slightly smaller under the weighted sampling scheme due to resampling of the same individuals from the Black population.

3.5 Discussion

As we've seen, the NCC design is useful in settings where the event of interest is rare and the covariate of interest is expensive or difficult to measure. Under this sampling design, however, it can be difficult to draw inference on the effect of this covariate in underrepresented subpopulations. In AD biomarker discovery, this challenge can result in disparities in therapeutic interventions and diagnoses for underrepresented ethnic and racial subpopulations. In this chapter, we proposed an extension to the Samuelsen estimator that allows for the maximization of information on the subpopulation-specific effects of an expensive covariate to preempt this issue. Our method involves weighted sampling under the NCC design with a correction using derived inclusion probabilities. In simulation studies, our method performed well for estimating effects in primary analyses while increasing power to detect effect

modification in underrepresented subpopulations. When applied to real data, our method also performed well compared to the full cohort for larger values of m while reducing the sample size by almost 50% compared to the full cohort analysis.

Under the proposed WNCC design, users are able to specify a desired distribution of sampling individuals into the analysis at each event time conditional on one of their covariate values. In our simulations for the binary covariate case, we specified an equal probability of sampling from each subpopulation defined by a binary covariate at each event time. We note, however, that the overall proportion of individuals from the underrepresented subpopulation who are included in the analysis will depend on the initial proportion of individuals from the underrepresented group in the full cohort, the hazard associated with the underrepresented group, and the initial sample size. We recommend defining similar desired probabilities for each subgroup (or an even distribution over a continuous covariate) in practice to avoid issues that may arise in the case of NPH.

Since the goal of the WNCC design is to increase representation of a select subpopulation through intentional sampling, it shares some similarities with counter-matched stratified sampling. The WNCC design is focused on increasing efficiency in the secondary analysis and accommodates sampling based a on continuous covariate, which differs from the goal of most counter-matched designs. Our proposed method also shares ideas with survey sampling methods, and hence a natural extension to consider may be the calibration of the estimated weights (Støer and Samuelsen, 2012; Rivera and Lumley, 2016a,b). Calibration has been demonstrated to increase efficiency when auxiliary variables are highly correlated with the predictor of interest (Breslow et al., 2013), and one could surely implement it in conjunction with the WNCC design. As mentioned previously, however, the main goal of our method is to increase efficiency for detecting effect modification via the inclusion of more participants from underrepresented subpopulations than an SRS, so we have not yet explored this extension to the proposed design. Moreover, in the setting of biomarker discovery, it is not guaranteed

that variables highly correlated with the biomarker measurement will be available on the full cohort.

Statistical models for biomarker discovery should be specified prior to viewing the data. As such, it is likely that assumptions for the Cox PH model will not always be completely met in practice. Struthers and Kalbfleisch (1986) demonstrated that the estimand for the Cox PH model is dependent on the censoring distribution under a misspecified model, but methods for a censoring-robust estimand, that can be interpreted as the average covariate effect over the support of the observed times (Xu, 2000), have been developed for the full cohort by Boyd et al. (2012) in the context of a randomized clinical trial and Nguyen and Gillen (2017) in the observational study setting. The extension to the NCC design by Nuño and Gillen (2020) and Nuño and Gillen (2021), mentioned in Chapter 2, should be implemented with our proposed WNCC design to obtain censoring-robust inference that will guard against potential violation of the PH assumption.

Many two-phase sampling methods are motivated by the need to maximize information with respect to a rare exposure. In this setting, we focus on maximizing information with respect to a rare resource in an underrepresented subpopulation for a single exposure; however, the NCC design and weighted sampling can be advantageous in the multiple event setting as well. One can imagine utilizing longitudinal data with multiple biomarker measurements and different stages of disease progression to estimate biomarker effects. Further, weighted sampling can be implemented in the multiple event setting to maximize information on any underrepresented subpopulations. To our knowledge, a basic framework for NCC sampling in the multiple event setting does not currently exist, so the following chapter focuses on developing this.

Chapter 4

Nested Case-Control Sampling for Marginal Estimation in Multiple Event Settings

4.1 Introduction

When studying rare events with expensive or difficult-to-obtain covariates, the NCC design provides an efficient approach to reducing data collection costs while maintaining valid statistical inference (Thomas, 1977; Nuño and Gillen, 2017). This design has proven valuable across diverse disease settings, including rare cancers, cardiology, and renal disease, by substantially lowering both financial and operational costs (Shamberger et al., 1999; Cushman, 2004; Lønning et al., 2022; Warwick et al., 2010). For example, in an analysis of risk factors for portal hypertension in children with Wilms' tumor, the NCC design was implemented to reduce the number of participants for whom radiation therapy doses were calculated and verified by an oncologist (Warwick et al., 2010), saving both time and money. Similarly, an NCC study was conducted to analyze the relationships between different genetic factors

and the risk of venous thrombosis, reducing the number of participants for which genomic DNA was analyzed (Cushman, 2004). The NCC design has been thoroughly investigated for univariate survival outcomes, and we reviewed several important extensions in Chapter 2. In Chapter 3, we proposed a weighted sampling framework motivated by current challenges in the AD biomarker discovery setting. Despite this extensive body of work and need for efficient sampling designs in diverse settings, applications of the NCC design to multiple event data remain scarce.

Recall that multiple event time-to-event data fall into three main categories: (1) recurrent event, or sequentially occurring events of the same type; (2) multiple-type events, or sequential events of different types; and (3) clustered or correlated events, or events occurring in related units such as family members or littermates (Kelly and Lim, 2000). These data are typically modeled using one of two approaches - frailty models or marginal models. In this chapter, we will focus on the marginal model approach, where the standard Cox PH model is sufficient as long as an appropriate model is chosen and the covariances are adjusted to account for potential correlation within clusters (Lin, 1994). Models are typically chosen based on the characteristics of the data using four key components - risk intervals, baseline hazard, risk set, and correlation adjustment - which are described in detail in Chapter 2. We also note that while competing and semi-competing risks can be perceived as types of multiple event data, the analysis methods presented in this chapter are not intended for immediate use in those settings.

As discussed previously, the NCC design has been extended to handle recurrent event data in the presence of a terminal event under a joint frailty model framework (Jazić et al., 2019). To our knowledge, this represents the only current NCC methodology for recurrent events. No prior work has addressed NCC sampling for marginal parameter estimation under semi-parametric models. In this chapter, we address this critical gap in the literature. To handle all types of multiple event data, we consider models based on the specification of the baseline

hazard for multiple event survival data and use these models to motivate our sampling framework. Specifically, we outline NCC sampling strategies for multiple event data that include pooling observations for sampling and stratifying sampling by event number. We demonstrate how event-specific sampling enables the fitting of marginal models that assume either common or event-specific baseline hazards (Lin, 1994), yielding consistent estimates of the full cohort estimand for both model types. We also describe scenarios in which this method yields efficiency gains compared to pooled sampling. Finally, we apply the proposed method to data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) to analyze the association between progression through Alzheimer's disease (AD) stages (i.e. cognitively normal to MCI and MCI to dementia) and p-tau, a key biomarker in AD research.

4.2 Methods

We consider right-censored, multiple time-to-event data with clustering at the individual level. Throughout this chapter, we assume multiple events are nested within individuals to maintain clarity in our presentation; however, the proposed methods can be applied to data clustered on the group level. We will also distinguish the event number (for an individual) from other uses of the word "event" by referring to the former as "event type," though this is also applicable to recurrent events of the same type. For subject i ($i = 1, \ldots, n$) and event type k ($k = 1, ..., K_i$), let T_{ik} denote the true event time, C_{ik} the censoring time, and $Z_{ik} = [z_{ik1}, \ldots, z_{ikp}]^T$ a p-dimensional covariate vector. We define the observed event total time as $X_{ik} = \min(T_{ik}, C_{ik})$ and $\delta_{ik} = I(X_{ik} = T_{ik})$ as the event indicator of event type k for subject i. We saw in Chapter 2 that different timescales can be used in the multiple event setting. Moving forward, we will consider the gap time formulation, which is often of scientific interest in the recurrent event setting; however, these methods are easily applicable to total time as well. The observed gap time can be formally defined as $G_{ik} = X_{ik} - X_{i,k-1}$ where $X_{i0} = 0$.

We start with a brief review of the notation and foundational concepts for multiple event survival data, where we again model event times via the hazard function. As outlined in Chapter 2, two model specifications arise based on whether the baseline hazard varies by event type, and this specification generally reflects different assumptions about the event-generating process. A common baseline hazard model may be chosen in settings where the occurrence of one event type does not influence the hazard of subsequent event types occurring within an individual. Conversely, a model with event-specific baseline hazards may be appropriate when the occurrence of an event type may be related to the occurrence of another event type within the same individual. The hazard function for each of these models is listed below:

- Common baseline hazards (M1): $\lambda_{ik}(t) = \lambda_0(t)\exp\{\beta^T Z_{ik}\}$
- Event-specific baseline hazards (M2): $\lambda_{ik}(t) = \lambda_{0k}(t)\exp\{\beta^T Z_{ik}\}$

The parameters β for these models can be estimated by solving the following estimating equations (Lin, 1994):

$$\mathcal{U}_{M1}(\beta) = \sum_{i=1}^{n}\sum_{k=1}^{K_i}\delta_{ik}\left[Z_{ik} - \frac{\sum_{l=1}^{n}\sum_{j=1}^{K_l} Y_{lj}(G_{ik})Z_{lj}\exp\{\beta^T Z_{lj}\}}{\sum_{l=1}^{n}\sum_{j=1}^{K_l} Y_{lj}(G_{ik})\exp\{\beta^T Z_{lj}\}}\right] \equiv 0 \tag{4.1}$$

for M1, and

$$\mathcal{U}_{M2}(\beta) = \sum_{i=1}^{n}\sum_{k=1}^{K_i}\delta_{ik}\left[Z_{ik} - \frac{\sum_{l=1}^{n}\sum_{j=1}^{K_l} I(j=k)Y_{lj}(G_{ik})Z_{lj}\exp\{\beta^T Z_{lj}\}}{\sum_{l=1}^{n}\sum_{j=1}^{K_l} I(j=k)Y_{lj}(G_{ik})\exp\{\beta^T Z_{lj}\}}\right] \equiv 0 \tag{4.2}$$

for M2, where $Y_{lj}(G_{ik}) = I(G_{lj} \geq G_{ik})$ represents whether observation lj has an observed gap time greater than G_{ik}. The variance of these parameters can be estimated via a sandwich or robust variance estimator provided in Section 2.4.2.

4.2.1 NCC Design Sampling for Multiple Event Data

We propose two NCC sampling strategies motivated by the baseline hazard specifications described above. Each sampling scheme aligns naturally with its corresponding model structure as pertains to the specification of the baseline hazard. That is, in each model setting, one might consider sampling controls from the risk sets defined under that model. For example, under the M1 model, the risk set at each event time consists of all lj pairs that have not yet experienced the j^{th} event type or been censored, regardless of the event type. Similarly, one can consider sampling from this risk set under the NCC design at each event time, adopting a **pooled** sampling approach. We term this "pooled sampling" because it ignores event type, treating all observations in the risk set as if they were independent, similar to univariate NCC sampling. In contrast, under the M2 model, the risk sets at each event time consist of observations with the same event type that have yet to experience that event or be censored. For settings where fitting an M2 model is appropriate, one can consider sampling from these event-stratified risk sets under the NCC design for an **event-specific** sampling approach. Here, the NCC design is implemented separately on each event stratum. Figure 4.1 provides a hypothetical example of viable controls under each sampling approach.

With both of these approaches, the sampling scheme coincides with the intuition behind the at-risk status under each fitted model (M1 vs. M2). As such, both of these sampling methods yield controls that are representative of the appropriate risk sets and would be valid for fitting their respective models without additional modifications under the design by Thomas (1977). In practice, however, it would be ideal to be able to fit both types of models while using the same sampling scheme. First, we present the theory for both pooled and event-specific sampling. We then demonstrate that the event-specific sampling scheme results in a minimal loss in efficiency compared to pooled sampling when fitting models with a common baseline hazard (M1) and a gain in precision when fitting models with event-specific baseline hazards (M2). Intuitively, event-specific sampling helps to maintain a proportional

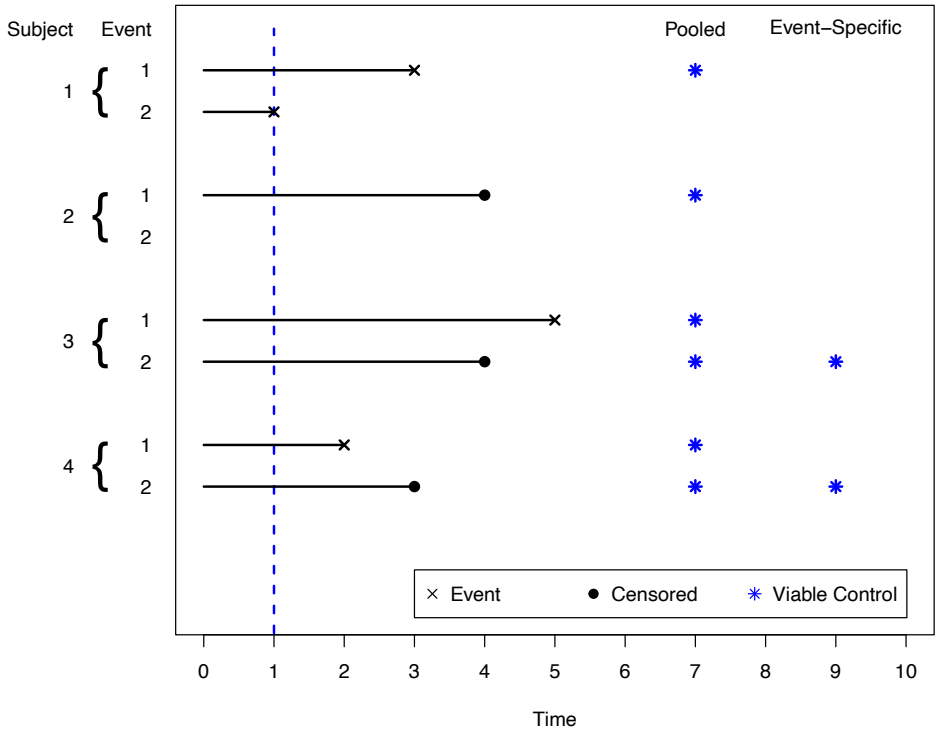


Figure 4.1: Hypothetical example of four subjects experiencing at most two events, where the length of each line represents the gap time between events. Events are indicated by $\times$, and censoring is indicated by $\bullet$. Observations considered at risk at time $t = 1$ under pooled and event-specific sampling schemes are denoted by a blue star on the right side of the plot.

number of controls to cases for each event type by design, ensuring an adequate number of controls for event-types with fewer observations when fitting M2 models and allowing for potential changes in the model being fit for the final analysis.

4.2.2 Pooled NCC Design Sampling

Similar to our work in Chapter 3 and as inspired by the work of Samuelsen (1997), we can utilize the sampled controls forward and backward in time to maximize information in the multiple event setting. To adjust for this non-SRS of controls in the analysis, we can use the inverse of the probability of being sampled. For observation lj, the probability of being

included in the analysis (i.e. as a case or sampled as a control) under pooled sampling is:

$$p_{lj}^{(P)} = \left\{ \begin{array}{ll} 1 - \prod_{(i,k):G_{ik}<G_{lj}} \left(1 - \frac{m}{|R_P(G_{ik})|-1}\delta_{ik}\right), & \text{if } \delta_{lj} = 0 \\ 1, & \text{if } \delta_{lj} = 1 \end{array} \right\}$$

where $|R_P(G_{ik})| = \sum_{l=1}^{n}\sum_{j=1}^{K_l} I(G_{lj} \geq G_{ik})$, the pooled risk set size at time G_{ik}. We can then define $V_{lj}^{(P)}$ as the indicator of whether observation lj is included in the analysis under pooled sampling. The resulting estimating equation is as follows:

$$\tilde{\mathcal{U}}_P^{(1)}(\beta) = \sum_{i=1}^{n}\sum_{k=1}^{K_i} \delta_{ik}\left[Z_{ik} - \frac{\sum_{l=1}^{n}\sum_{j=1}^{K_l} Y_{lj}(G_{ik})\frac{V_{lj}^{(P)}}{p_{lj}^{(P)}} Z_{lj}\exp\{\beta^T Z_{lj}\}}{\sum_{l=1}^{n}\sum_{j=1}^{K_l} Y_{lj}(G_{ik})\frac{V_{lj}^{(P)}}{p_{lj}^{(P)}}\exp\{\beta^T Z_{lj}\}}\right] \equiv 0 \tag{4.3}$$

for M1, and

$$\tilde{\mathcal{U}}_P^{(2)}(\beta) = \sum_{i=1}^{n}\sum_{k=1}^{K_i} \delta_{ik}\left[Z_{ik} - \frac{\sum_{l=1}^{n}\sum_{j=1}^{K_l} I(j=k)Y_{lj}(G_{ik})\frac{V_{lj}^{(P)}}{p_{lj}^{(P)}} Z_{lj}\exp\{\beta^T Z_{lj}\}}{\sum_{l=1}^{n}\sum_{j=1}^{K_l} I(j=k)Y_{lj}(G_{ik})\frac{V_{lj}^{(P)}}{p_{lj}^{(P)}}\exp\{\beta^T Z_{lj}\}}\right] \equiv 0 \tag{4.4}$$

for M2. We note again that the terms captured in the sums in the compensator depend upon the choice of model to reflect the appropriate risk sets.

Proposition 4.1. *Denote $\beta^{(1)}$ and $\beta^{(2)}$ as the estimands for the full cohort M1 and M2, respectively. Let $\widehat{\beta}_P^{(1)}$ and $\widehat{\beta}_P^{(2)}$ be the solutions to Equations 4.3 and 4.4, respectively. It follows that as $n \to \infty$, then $\widehat{\beta}_P^{(1)} \to \beta^{(1)}$. Similarly, as $n \to \infty$ and $K_i/n \to c_i$, where $0 < c_i < 1$, for $i = 1, \ldots, n$, then $\widehat{\beta}_P^{(2)} \to \beta^{(2)}$.*

The proof of this result follows closely from the proof in Section B.1 of Appendix B.

The variance of the model parameters can be estimated via a robust variance estimator, Γ_P

(Binder, 1992):

$$\widehat{\Gamma}_P = \widehat{\Sigma}_P^{-1} V_P \widehat{\Sigma}_P^{-1}$$

where

$$\widehat{\Sigma}_P = I_P(\widehat{\beta}) = -\frac{\partial \tilde{\mathcal{U}}_P^{(M)}(\beta)}{\partial \beta}|_{\beta=\widehat{\beta}_P^{(M)}} \text{ and } V_P = \sum_{i=1}^{n}\sum_{k=1}^{K_i}\sum_{l=1}^{K_i} W_{P,ik}^{(M)}(\widehat{\beta}_P^{(M)}) W_{P,il}^{(M)}(\widehat{\beta}_P^{(M)})^T$$

for $M = 1, 2$ with

$$W_{P,ik}^{(1)}(\widehat{\beta}) = \delta_{ik}\frac{V_{ik}^{(P)}}{p_{ik}^{(P)}}\left[Z_{ik} - \frac{\tilde{S}_P^{(1)}(\widehat{\beta}, G_{ik})}{\tilde{S}_P^{(0)}(\widehat{\beta}, G_{ik})}\right] - \frac{V_{ik}^{(P)}}{p_{ik}^{(P)}}\sum_{l=1}^{n}\sum_{j=1}^{K_l}\delta_{lj}\frac{V_{lj}^{(P)}}{p_{lj}^{(P)}}\left[\frac{Y_{ik}(G_{lj})\exp\{\widehat{\beta}^T Z_{ik}\}}{\tilde{S}_P^{(0)}(\widehat{\beta}, G_{lj})}\right] \times \left\{Z_{ik} - \frac{\tilde{S}_P^{(1)}(\widehat{\beta}, G_{lj})}{\tilde{S}_P^{(0)}(\widehat{\beta}, G_{lj})}\right\}$$

where $\tilde{S}_P^{(r)}(\beta, G_{ik}) = \sum_{l=1}^{n}\sum_{j=1}^{K_l} Y_{lj}(G_{ik})\frac{V_{lj}^{(P)}}{p_{lj}^{(P)}} Z_{lj}^{\otimes r}\exp\{\beta^T Z_{lj}\}$ for $r = 0, 1$, or

$$W_{P,ik}^{(2)}(\widehat{\beta}) = \delta_{ik}\frac{V_{ik}^{(P)}}{p_{ik}^{(P)}}\left[Z_{ik} - \frac{\tilde{S}_{P,k}^{(1)}(\widehat{\beta}, G_{ik})}{\tilde{S}_{P,k}^{(0)}(\widehat{\beta}, G_{ik})}\right] - \frac{V_{ik}^{(P)}}{p_{ik}^{(P)}}\sum_{l=1}^{n}\sum_{j=1}^{K_l} I(j=k)\delta_{lj}\frac{V_{lj}^{(P)}}{p_{lj}^{(P)}}\left[\frac{Y_{ik}(G_{lj})\exp\{\widehat{\beta}^T Z_{ik}\}}{\tilde{S}_{P,k}^{(0)}(\widehat{\beta}, G_{lj})}\right] \times \left\{Z_{ik} - \frac{\tilde{S}_{P,k}^{(1)}(\widehat{\beta}, G_{lj})}{\tilde{S}_{P,k}^{(0)}(\widehat{\beta}, G_{lj})}\right\}$$

where $\tilde{S}_{P,k}^{(r)}(\beta, G_{ik}) = \sum_{l=1}^{n}\sum_{j=1}^{K_l} I(j=k) Y_{lj}(G_{ik})\frac{V_{lj}^{(P)}}{p_{lj}^{(P)}} Z_{lj}^{\otimes r}\exp\{\beta^T Z_{lj}\}$ for $r = 0, 1$. Details regarding the variance estimators and limiting distribution follow closely to that of event-specific sampling presented in Section B.2 of Appendix B.

4.2.3 Event-Specific NCC Design Sampling

We continue to utilize the sampled controls forward and backward in time to maximize information under event-specific sampling. For observation lj, the probability of being included

in the analysis (i.e. as a case or sampled as a control) under event-specific sampling is:

$$p_{lj} = \left\{ \begin{array}{ll} 1 - \prod_{i:G_{ij}<G_{lj}} \left(1 - \frac{m}{|R(G_{ij})|-1}\delta_{ij}\right), & \text{if } \delta_{lj} = 0 \\ 1, & \text{if } \delta_{lj} = 1 \end{array} \right\}$$

where $|R(G_{ij})| = \sum_{l=1}^{n} I(G_{lj} \geq G_{ij})$, the event-specific risk set size at time G_{ij}. We define V_{lj} as the indicator of whether observation lj is included in the analysis under event-specific sampling, and the resulting estimating equation is as follows:

$$\tilde{\mathcal{U}}^{(1)}(\beta) = \sum_{i=1}^{n}\sum_{k=1}^{K_i} \delta_{ik}\left[Z_{ik} - \frac{\sum_{l=1}^{n}\sum_{j=1}^{K_l} Y_{lj}(G_{ik})\frac{V_{lj}}{p_{lj}} Z_{lj}\exp\{\beta^T Z_{lj}\}}{\sum_{l=1}^{n}\sum_{j=1}^{K_l} Y_{lj}(G_{ik})\frac{V_{lj}}{p_{lj}}\exp\{\beta^T Z_{lj}\}}\right] \equiv 0 \tag{4.5}$$

for M1, and

$$\tilde{\mathcal{U}}^{(2)}(\beta) = \sum_{i=1}^{n}\sum_{k=1}^{K_i} \delta_{ik}\left[Z_{ik} - \frac{\sum_{l=1}^{n}\sum_{j=1}^{K_l} I(j=k)Y_{lj}(G_{ik})\frac{V_{lj}}{p_{lj}} Z_{lj}\exp\{\beta^T Z_{lj}\}}{\sum_{l=1}^{n}\sum_{j=1}^{K_l} I(j=k)Y_{lj}(G_{ik})\frac{V_{lj}}{p_{lj}}\exp\{\beta^T Z_{lj}\}}\right] \equiv 0 \tag{4.6}$$

for M2.

Proposition 4.2. *Denote $\beta^{(1)}$ and $\beta^{(2)}$ as the estimands for the full cohort M1 and M2, respectively. Let $\widehat{\beta}^{(1)}$ and $\widehat{\beta}^{(2)}$ be the solutions to Equations 4.5 and 4.6, respectively. It follows that as $n \to \infty$, then $\widehat{\beta}^{(1)} \to \beta^{(1)}$. Similarly, as $n \to \infty$ and $K_i/n \to c_i$, where $0 < c_i < 1$, for $i = 1, \ldots, n$, then $\widehat{\beta}^{(2)} \to \beta^{(2)}$.*

The proof of this result can be found in Section B.1 of Appendix B.

We also demonstrate in Section B.2 that the variance of the model parameters can be estimated via a robust variance estimator, Γ (Binder, 1992):

$$\widehat{\Gamma} = \widehat{\Sigma}^{-1} V \widehat{\Sigma}^{-1} \tag{4.7}$$

where

$$\widehat{\Sigma} = I(\widehat{\beta}) = -\frac{\partial \tilde{\mathcal{U}}^{(M)}(\beta)}{\partial \beta}|_{\beta=\widehat{\beta}^{(M)}} \quad \text{and} \quad V = \sum_{i=1}^{n}\sum_{k=1}^{K_i}\sum_{l=1}^{K_i} W_{ik}^{(M)}(\widehat{\beta}^{(M)})W_{il}^{(M)}(\widehat{\beta}^{(M)})^T$$

for $M = 1, 2$ with

$$W_{ik}^{(1)}(\widehat{\beta}) = \delta_{ik}\frac{V_{ik}}{p_{ik}}\left[Z_{ik} - \frac{\tilde{S}^{(1)}(\widehat{\beta}, G_{ik})}{\tilde{S}^{(0)}(\widehat{\beta}, G_{ik})}\right] - \frac{V_{ik}}{p_{ik}}\sum_{l=1}^{n}\sum_{j=1}^{K_l}\delta_{lj}\frac{V_{lj}}{p_{lj}}\left[\frac{Y_{ik}(G_{lj})\exp\{\widehat{\beta}^T Z_{ik}\}}{\tilde{S}^{(0)}(\widehat{\beta}, G_{lj})}\right] \times \left\{Z_{ik} - \frac{\tilde{S}^{(1)}(\widehat{\beta}, G_{lj})}{\tilde{S}^{(0)}(\widehat{\beta}, G_{lj})}\right\}$$

where $\tilde{S}^{(r)}(\beta, G_{ik}) = \sum_{l=1}^{n}\sum_{j=1}^{K_l} Y_{lj}(G_{ik})\frac{V_{lj}}{p_{lj}} Z_{lj}^{\otimes r}\exp\{\beta^T Z_{lj}\}$ for $r = 0, 1$, or

$$W_{ik}^{(2)}(\widehat{\beta}) = \delta_{ik}\frac{V_{ik}}{p_{ik}}\left[Z_{ik} - \frac{\tilde{S}_k^{(1)}(\widehat{\beta}, G_{ik})}{\tilde{S}_k^{(0)}(\widehat{\beta}, G_{ik})}\right] - \frac{V_{ik}}{p_{ik}}\sum_{l=1}^{n}\sum_{j=1}^{K_l} I(j=k)\delta_{lj}\frac{V_{lj}}{p_{lj}}\left[\frac{Y_{ik}(G_{lj})\exp\{\widehat{\beta}^T Z_{ik}\}}{\tilde{S}_k^{(0)}(\widehat{\beta}, G_{lj})}\right] \times \left\{Z_{ik} - \frac{\tilde{S}_k^{(1)}(\widehat{\beta}, G_{lj})}{\tilde{S}_k^{(0)}(\widehat{\beta}, G_{lj})}\right\}$$

where $\tilde{S}_k^{(r)}(\beta, G_{ik}) = \sum_{l=1}^{n}\sum_{j=1}^{K_l} I(j=k) Y_{lj}(G_{ik})\frac{V_{lj}}{p_{lj}} Z_{lj}^{\otimes r}\exp\{\beta^T Z_{lj}\}$ for $r = 0, 1$. We provide details regarding the variance estimators along with a conjecture of asymptotic normality for this sampling scheme in Section B.2 of Appendix B.

4.3 Empirical Results

In this section, we present results from simulation studies demonstrating the performance of the proposed sampling and estimation methods from this chapter. We consider multiple event data with at most two recurrent events. In all simulations presented here, Z_1 is a continuous cluster-varying covariate, meaning it can change across measurements within an individual, drawn from $MVN(\vec{0}, \Sigma_{0.7})$ where $\Sigma_{0.7}$ is a 2×2 matrix with 1 on the diagonal and 0.7 on the off-diagonals. Z_2 is a cluster-invariant covariate, meaning it is constant

for all measurements from that individual, drawn from $Bern(0.5)$. We note that in these simulations, Z_1 mimics the covariate that is difficult or costly to obtain, as a slightly different approach might be taken if the expensive covariate were cluster-invariant. We provide more details regarding this approach in Section 4.5.

In all simulation settings, true event times for the k^{th} event of subject i, T_{ik}, were generated from an $Exp(\lambda_{ik})$ where λ_{ik} is taken to have either a common baseline hazard for both event types or different baseline hazards for each event type. Individual-specific censoring times C_i were generated from $U(0,1)$, and observed times X_{ik} were defined as $\min(\sum_{j=1}^{k} T_{ij}, C_i)$. Event-specific indicators δ_{ik} were 1 if $X_{ik} = \sum_{j=1}^{k} T_{ij}$ and 0 otherwise. Here, we utilize the gap time formulation and present the results from the two data generating models - M1 and M2 - below. In each scenario, we provide results from fitting the model that matches the data generating model and compare the two methods of sampling. All simulations were run for 500 iterations.

4.3.1 Common Baseline Hazard

In this section, the data generating model is $\lambda_{ik} = \lambda_0 \exp\{\log(0.7)Z_{1ik} + \log(1.4)Z_{2ik} + v_i\}$ where $v_i \sim N(0,1)$ is an individual-level random effect to induce additional correlation within subject measurements and $\lambda_0 = 0.11$ was chosen to yield an overall censoring rate of approximately 90% ($\sim$ 90% for the first event type, and $\sim$ 85% for the second event type). As stated previously, we evaluated the performance of the event-specific NCC design sampling framework and corresponding estimator from Section 4.2.3 under the setting of a common baseline hazard for both events by fitting the corresponding model of the form: $\lambda_{ik}(t) = \lambda_0(t) \exp\{\beta_1 Z_{1ik} + \beta_2 Z_{2ik}\}$. Table 4.1 demonstrates that the proposed sampling and estimation method yields coefficient estimates that perform similarly to the full cohort as well as nominal coverage probabilities for all values of m.

As mentioned previously, pooled sampling can also be implemented with methodology pre-

			$\beta_1^* = -0.339$			
	Total Meas.	Unique N	$\widehat{\beta}_1$ (% Bias)	Emp. SE	Rob. SE	Cov. Prob.
FC (90% cens.)	3288	3000	-0.334 (1.54%)	0.059	0.057	0.940
m = 1	618	541	-0.337 (0.64%)	0.084	0.082	0.944
m = 2	868	763	-0.335 (1.21%)	0.071	0.070	0.948
m = 3	1085	958	-0.333 (1.86%)	0.067	0.065	0.932
m = 4	1277	1131	-0.334 (1.58%)	0.064	0.063	0.950
			$\beta_2^* = 0.327$			
	Total Meas.	Unique N	$\widehat{\beta}_2$ (% Bias)	Emp. SE	Rob. SE	Cov. Prob.
FC (90% cens.)	3288	3000	0.323 (-1.28%)	0.117	0.118	0.956
m = 1	618	541	0.327 (0.21%)	0.163	0.162	0.952
m = 2	868	763	0.325 (-0.59%)	0.138	0.140	0.950
m = 3	1085	958	0.323 (-1.10%)	0.127	0.132	0.956
m = 4	1277	1131	0.320 (-1.93%)	0.127	0.128	0.958

Table 4.1: Simulation results comparing the performance of the full cohort PH estimator to the proposed estimator (NCC design with event-specific sampling), both assuming a common baseline hazard across event types (M1). *"True" estimates are based on a model with $n = 100,000$, and percent bias was calculated as $E\{(\widehat{\beta} - \beta^*)/|\beta^*|\} \cdot 100$.

sented in Section 4.2.2. Figure 4.2 compares average estimates and corresponding 95% confidence intervals (constructed by taking the average of all simulations) from the two sampling methods. Here, the length of the confidence intervals for event-specific sampling (green) are similar to the length of the intervals for the pooled sampling, with an increase of 1% or less on average for event-specific sampling compared to pooled sampling. Moreover, for larger values of m, both are comparable to the performance of the full cohort while using fewer overall measurements. We note that the smaller overall sample size under event-specific sampling is due to the stochastic resampling of more individuals in the smaller second event-type stratum.

4.3.2 Event-Specific Baseline Hazards

For the setting where the baseline hazard is allowed to vary by event type, we assumed a data-generating mechanism of the form: $\lambda_{ik} = \lambda_{0k} \exp\{\log(0.7)Z_{1ik} + \log(1.4)Z_{2ik} + v_i\}$. Here, $\lambda_{01} = 0.1$ and $\lambda_{02} = 0.2$ were again chosen to yield an overall censoring rate of approximately

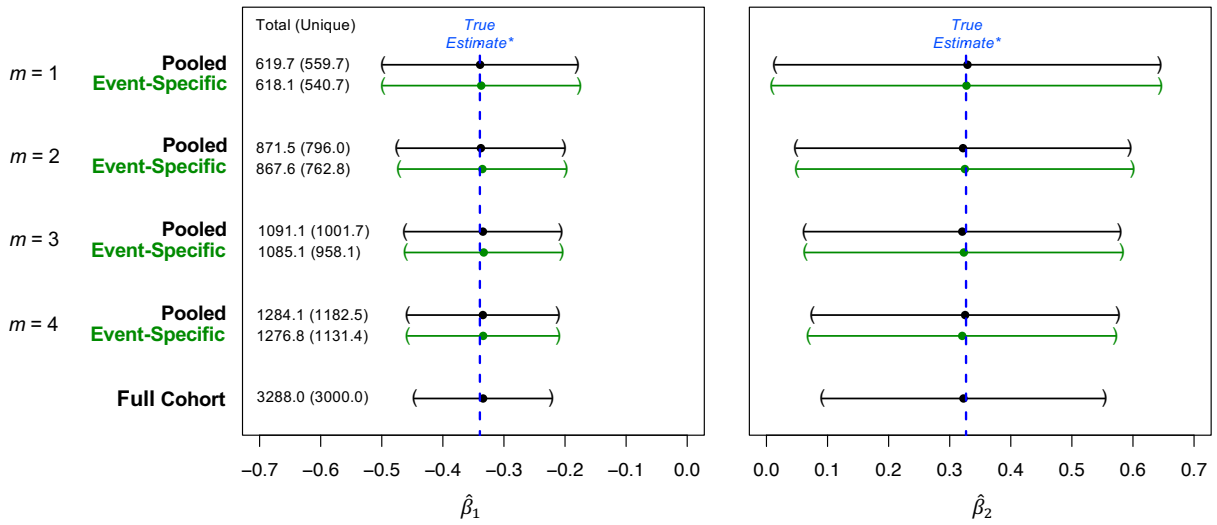


Figure 4.2: Average estimated effects and corresponding 95% confidence intervals from pooled and event-specific NCC design sampling schemes under a common baseline hazard model (M1). "Total" represents the average total number of unique observations, and "Unique" represents the average total of unique individuals sampled. *"True" estimates are based on a model with $n = 100,000$.

90% ($\sim$ 91% for the first event type, and $\sim$ 77% for the second event type) to reflect settings where the NCC design is most commonly used. We evaluated the performance of the proposed event-specific NCC design sampling framework and corresponding estimator under the setting where each event type has a different baseline hazard. We fit the corresponding model of the form: $\lambda_{ik}(t) = \lambda_{0k}(t)\exp\{\beta_1 Z_{1ik} + \beta_2 Z_{2ik}\}$. Table 4.2 demonstrates that the proposed sampling and estimation method yield estimates consistent for the full cohort estimand. The robust variance estimator also yields nominal coverage for all values of m.

We compared the average estimates and corresponding 95% confidence intervals (constructed by taking the average of all simulations) for the event-specific and pooled sampling methods in this scenario (Figure 4.3). Event-specific sampling yields 3-6% narrower confidence intervals than pooled sampling on average, attributable to more proportional control representation across event types.

			$\beta_1^* = -0.327$			
	Total Meas.	Unique N	$\widehat{\beta}_1$ (% Bias)	Emp. SE	Rob. SE	Cov. Prob.
FC (90% cens.)	3266	3000	-0.318 (2.64%)	0.057	0.055	0.948
m = 1	605	501	-0.323 (1.14%)	0.083	0.080	0.940
m = 2	845	710	-0.318 (2.58%)	0.071	0.068	0.926
m = 3	1053	896	-0.319 (2.29%)	0.065	0.063	0.948
m = 4	1236	1062	-0.319 (2.27%)	0.062	0.061	0.956
			$\beta_2^* = 0.306$			
	Total Meas.	Unique N	$\widehat{\beta}_2$ (% Bias)	Emp. SE	Rob. SE	Cov. Prob.
FC (90% cens.)	3266	3000	0.300 (-2.08%)	0.113	0.113	0.950
m = 1	605	501	0.305 (-0.49%)	0.152	0.157	0.964
m = 2	845	710	0.297 (-2.98%)	0.139	0.135	0.950
m = 3	1053	896	0.305 (-0.34%)	0.127	0.127	0.948
m = 4	1236	1062	0.300 (-1.96%)	0.128	0.123	0.938

Table 4.2: Simulation results comparing the performance of the full cohort PH estimator to the proposed estimator (NCC design with event-specific sampling), both assuming different baseline hazards across event types (M2). *"True" estimates are based on a model with $n = 100,000$, and percent bias was calculated as $E\{(\widehat{\beta} - \beta^*)/|\beta^*|\} \cdot 100$.

4.4 Application to ADNI Data

We've emphasized that the NCC design is particularly valuable for AD biomarker research, where measurements are costly to obtain and process. Recall that many biomarkers are derived from CSF which is obtained via an invasive procedure called a lumbar puncture. Collected samples are frozen and stored for use in future analyses. However, repeated freeze-thaw cycles degrade sample quality, making sample conservation essential beyond mere cost considerations.

The ADNI study collects longitudinal imaging, genetic, and biomarker data on participants across North America. Since it is of interest to preserve CSF samples, we applied our proposed NCC design and corresponding estimators to data from ADNI to estimate the association between p-tau, a key biomarker in AD, and the risk of disease progression, where we define disease progression to be either progression from cognitively normal (CN) to MCI or progression from MCI to dementia. We considered each progression to be separate events and

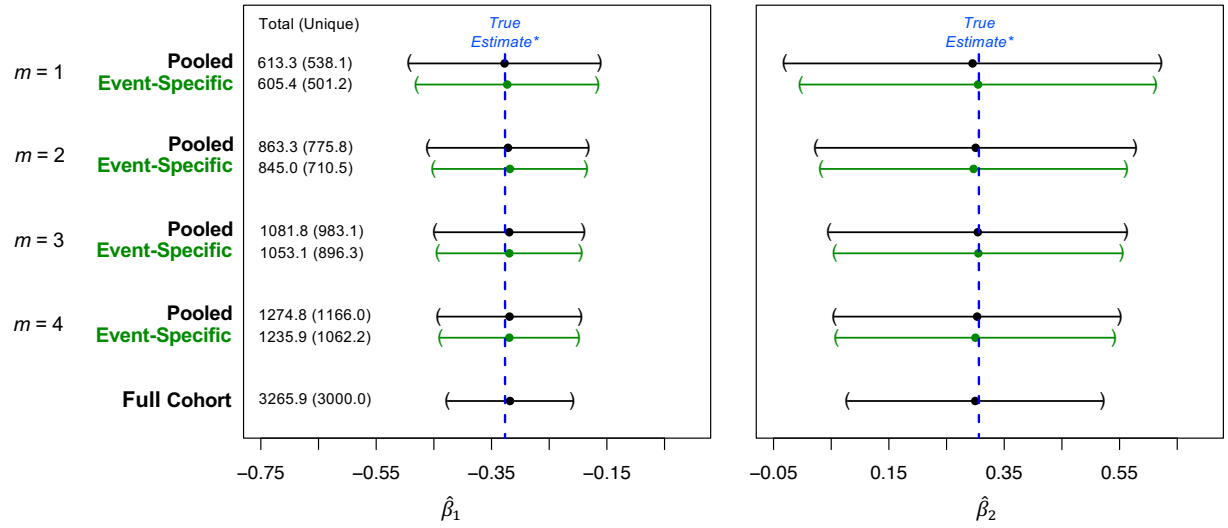


Figure 4.3: Average estimated effects and corresponding 95% confidence intervals from pooled and event-specific NCC design sampling schemes under a model with event-specific baseline hazards (M2). "Total" represents the average total number of unique observations, and "Unique" represents the average total of unique individuals sampled. *"True" estimates are based on a model with $n = 100,000$.

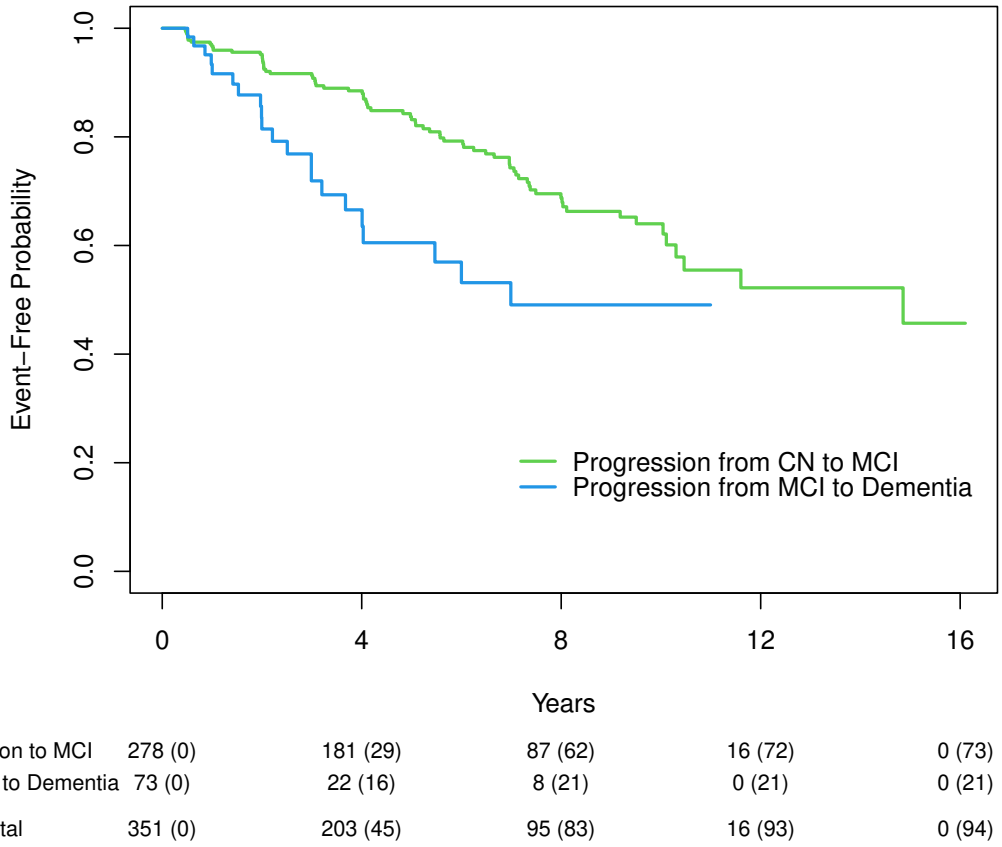
utilized longitudinal biomarker CSF p-tau measurements. We included $N = 278$ individuals who started with a diagnosis of CN and had CSF p-tau measurements and data regarding age, sex, APOE-e4 allele status, and education. In order to evaluate the performance of our estimator in this example, we also provide results from a full cohort analysis for comparison. We note, however, that this would not be possible in practice since only measurements for the sampled cohort would be available.

Table 4.3 includes participant characteristics for the full cohort, and Figure 4.4 displays event-specific Kaplan-Meier estimates. Out of $N = 278$ unique participants, 73 progressed to MCI, and of those participants, 21 progressed to dementia. Two participants who progressed to dementia were missing the timing of an MCI diagnosis, so we used the midpoint between CN and dementia diagnoses. Ideally, CSF samples would be collected at each disease milestone. Measurement timing, however, did not always align with diagnosis dates. We therefore imputed p-tau values for 63 observations using each participant's closest available measurement to the "baseline" of each event (i.e. at true baseline or at first diagnosis of

	Progression to MCI		Progression to Dementia	
	Low P-tau (< 23 pg/mL) $N = 185$	High P-tau (≥ 23 pg/mL) $N = 93$	Low P-tau (< 23 pg/mL) $N = 31$	High P-tau (≥ 23 pg/mL) $N = 42$
Progressed	40 (21.6%)	33 (35.5%)	4 (12.9%)	17 (40.5%)
P-Tau	16.9 (3.5)	32.2 (8.1)	16.4 (3.9)	33.4 (8.6)
Age	73.4 (5.6)	76.3 (6.1)	73.8 (5.5)	77.2 (5.5)
Female	94 (50.8%)	46 (49.5%)	14 (45.2%)	18 (42.9%)
1 APOE-e4 allele	36 (19.5%)	27 (29%)	4 (12.9%)	14 (33.3%)
2 APOE-e4 alleles	6 (3.2%)	2 (2.2%)	4 (12.9%)	1 (2.4%)
Education (yrs.)	16.2 (2.8)	16.4 (2.3)	16.5 (2.8)	16 (2.7)

Table 4.3: Participant characteristics for individuals stratified by event and p-tau level. Values are reported as mean (SD) for continuous variables and count (%) for discrete variables.

MCI) and adjusted for this in our analysis using the time between the "baseline" of each progression and date of the closest available p-tau measurement. To analyze the association between p-tau and disease progression, we fit a model with p-tau measurement as the predictor of interest and allowed for differing baseline hazards for each progression (M2). We utilized a gap time model formulation, meaning observed times were represented by the time since the previous progression, and focused on estimating a shared covariate effect to maximize precision. We also adjusted for age, sex, years of education, APOE-e4 allele status, and the time from the "baseline" of each progression to the date of the closest available p-tau measurement. We noticed that 47 of the imputed measurements were taken more than one year before or after the diagnosis, which could potentially bias results, so we conducted a sensitivity analysis where these observations were removed. We obtained similar results from the sensitivity analysis, so we present the results from the original imputed cohort in Table 4.4. For select covariates, we report the total number of measurements used (Total Meas.), number of unique individuals included in the analysis (Unique N), coefficient estimate (Est.) and corresponding hazard ratio (HR), standardized bias (Std. Bias), which was computed by taking the ratio of the difference from the full cohort estimate and the standard error for the full cohort estimate, standard error estimate (SE), and p-value.



Progression to MCI	278 (0)	181 (29)	87 (62)	16 (72)	0 (73)
Progression to Dementia	73 (0)	22 (16)	8 (21)	0 (21)	0 (21)
Total	351 (0)	203 (45)	95 (83)	16 (93)	0 (94)

Figure 4.4: Kaplan-Meier plot of disease progression for participants stratified by progression type (event type).

In the full cohort, we estimate that the risk of disease progression is approximately 40% higher for those with higher p-tau when comparing two subpopulations of similar age, sex, years of education, APOE-e4 status, and time from diagnosis to measurement but differing in p-tau by 10 pg/mL. This result is consistent with previous research that demonstrates an increased risk of progression associated with higher p-tau measurements (Blennow et al., 2010; Blom et al., 2009). Our proposed sampling and corresponding estimator result in some bias relative to the full cohort estimate, but generally provide results consistent with the full cohort for the predictor of interest, p-tau, as well as the adjustment covariates. Table 4.4 also shows the results for age (continuous per 10 year-increment) and APOE-e4 status, two well-researched risk factors for dementia (Hou et al., 2019; Saunders et al., 1993). Our analysis

			P-Tau (10 pg/mL)			Age (10 years)		
	Total Meas.	Unique N	Est. (HR)	Std. Bias	SE (p-value)	Est. (HR)	Std. Bias	SE (p-value)
FC (73% cens.)	351	278	0.34 (1.40)	0.00	0.09 (<0.01)	0.27 (1.31)	0.00	0.17 (0.10)
m = 1	155	123	0.34 (1.41)	0.08	0.13 (<0.01)	0.08 (1.08)	-1.16	0.25 (0.76)
m = 2	192	152	0.39 (1.48)	0.61	0.12 (<0.01)	0.40 (1.49)	0.80	0.20 (0.04)
m = 3	241	194	0.35 (1.42)	0.16	0.11 (<0.01)	0.17 (1.18)	-0.61	0.18 (0.34)
m = 4	243	194	0.30 (1.35)	-0.38	0.10 (<0.01)	0.30 (1.34)	0.16	0.17 (0.08)
			APOE-e4 - 1 Allele (vs. 0)			APOE-e4 - 2 Alleles (vs. 0)		
	Total Meas.	Unique N	Est. (HR)	Std. Bias	SE (p-value)	Est. (HR)	Std. Bias	SE (p-value)
FC (73% cens.)	351	278	0.30 (1.34)	0.00	0.23 (0.21)	1.10 (3.00)	0.00	0.35 (<0.01)
m = 1	155	123	0.46 (1.58)	0.69	0.30 (0.13)	1.29 (3.64)	0.56	0.56 (0.02)
m = 2	192	152	0.40 (1.49)	0.44	0.26 (0.12)	1.30 (3.66)	0.57	0.34 (<0.01)
m = 3	241	194	0.29 (1.33)	-0.03	0.25 (0.25)	0.86 (2.36)	-0.70	0.42 (0.04)
m = 4	243	194	0.29 (1.34)	-0.03	0.24 (0.24)	0.97 (2.62)	-0.39	0.38 (0.01)

Table 4.4: Estimated results from the application of our proposed sampling design and estimator to ADNI data, including full cohort results. Standardized bias was computed by taking the ratio of the difference from the full cohort estimate and the standard error for the full cohort.

shows that the risk of disease progression is higher for participants who are older and those with higher numbers of APOE-e4 alleles. The proposed method returns conclusions similar to the full cohort results for these variables. The estimates presented here are a result of single analyses, each with its own independent cohort for each choice of m. As such, due to the variability in sampling, it is expected that improvements in bias are not perfectly monotonic with the increase in m. Moreover, the minimal increase in total measurements from the $m = 3$ analysis to the $m = 4$ analysis can likely be attributed to this stochasticity as well as the resampling of observations in multiple risk sets. We note also that the number of individuals with two APOE-e4 alleles is small, resulting in unstable robust standard error estimates, and that the resulting hazard ratios for APOE-e4 (2) should be interpreted with this in mind. It is important to note also that if we were to perform this analysis in practice, we could reduce the number of CSF samples requiring processing by approximately 30-65%, depending on the choice of m, by using the NCC design.

4.5 Discussion

The NCC design has been thoroughly investigated in the univariate setting, but little work has been done for multiple event data. In this chapter, we explored two potential NCC sampling schemes for marginal parameter estimation in the multiple event setting - one that pools all observations in a single risk set (pooled), and another that stratifies risk sets by event type (event-specific) for sampling. The development of these sampling frameworks was motivated by models that assume a common baseline hazard for all observations (M1) and models that assume different baseline hazards for each event type (M2), respectively.

A single sampling design capable of fitting both model types offers practical advantages. We recommend using an event-specific NCC design sampling scheme in the multiple event setting for the most flexibility in fitting different types of models. This proposed sampling design maintains a balanced number of sampled controls proportional to cases across event strata and is also supported by the use of the sampled controls forward and backward in time for optimal efficiency using an extension of the Samuelsen estimator framework (Samuelsen, 1997). In simulation, our proposed estimator was consistent for the full cohort estimand, maintained valid uncertainty quantification for inference, and significantly reduced the number of participants used in the analysis. When compared to pooled sampling, the proposed event-specific sampling scheme demonstrated a gain in precision in M2 models and a minimal loss in precision when fitting M1 models despite using fewer observations.

In some settings, it may be of interest to estimate the cumulative baseline hazard function for risk prediction using the NCC sample. We suggest users implement the weighted cumulative baseline hazard function estimator proposed by Lin (2000) with event-specific sampling probabilities, p_{lj}. The data should either be stratified by event type for event-specific baseline hazards or pooled together for a common baseline hazard. In either case, event-specific sampling probability weights should be used to coincide with the way the observations were sampled.

Several considerations arise when conducting inference with multiple events. To avoid multiple testing bias, researchers should select covariates a priori, before examining the data. This includes the potential for effect modification of the selected covariates by event type. In settings where an event-specific coefficient is deemed appropriate, the proposed event-specific sampling encourages adequate representation across event types to estimate this effect modification, and we recommend fitting a model with different coefficients associated with the prespecified covariate for each event type. Moreover, the timescale should be selected exclusively with the scientific question of interest in mind. Our proposed sampling method is intended to be used for both total time and gap time formulations, but its utility is not evident for the counting process timescale due to the inconsistencies of risk set timing between subjects and event types.

A critical practical consideration involves whether expensive covariates are cluster-varying (changing with each measurement, e.g., biomarkers collected longitudinally) or cluster-invariant (constant over time, e.g., sex, race). For cluster-varying covariates, sampling observations rather than individuals makes sense, as each measurement yields new covariate information. In these settings, timing of covariate measurement for an observation may not always align with the occurrence of an event for that individual. Imputation methods can be implemented using the nearest measurement, but results should be interpreted taking this into account. For cluster-invariant covariates, covariate values are already known once an observation is sampled at the first event, so all subsequent events for that individual should be included in the analysis. From the results presented here, we advise implementing the classic NCC design with the Samuelsen estimator on the first event stratum, then including all observations from later event types with a probability of 1 to maximize the available information.

The number of events for each individual should also be considered in the analysis of multiple event data. In Section 4.4, we applied our method to the AD biomarker setting where

subjects experienced at most two events. In settings with a higher maximum number of events, some participants may have significantly more events than others, potentially biasing results when a strong covariate effect is present. To remedy this, we recommend extending the current proposed method to weight each term of the estimating equation by the inverse of the number of observations for the individual associated with that term.

Finally, as we discussed in previous chapters, one must exclude vast departures from the PH assumption for valid interpretation of results when using the PH model, and this is also true of the methods presented in this chapter. In practice, it is common for this assumption to be violated, and it has been demonstrated that the resulting estimand is dependent on the censoring distribution (Struthers and Kalbfleisch, 1986), making results difficult to replicate and interpret. Since the proposed methodology is one of the first covering the NCC design in the multiple event setting, censoring-robust estimation has yet to be developed for this case. In the next chapter, we will revisit existing censoring-robust estimators that have been developed for the full cohort (Xu, 2000; Boyd et al., 2012; Nguyen and Gillen, 2017) and discuss their performance in the multiple event setting.

Chapter 5

On the Analysis of Multiple Event Survival Data Under Non-Proportional Hazards

5.1 Introduction

The PH model relies on the strong assumption that the hazard functions of any two subpopulations remain proportional across all follow-up times. In clinical settings, covariate effects can be complex and potentially time-varying, resulting in a violation of the PH assumption, yet practitioners still often default to using the PH model. Substantial work has been done to characterize the properties of the PH model under NPH in the univariate setting, yet the implications of NPH for the analysis of multiple event data remain largely unexplored.

Recall that multiple event survival data can often be categorized as recurrent events, multiple-type events, or clustered/correlated events. Recurrent events are sequentially occurring events of the same type such as heart attacks, hospitalizations, or seizures. Multiple-type

events involve different types of events and may include progression through stages of a disease such as Alzheimer's disease, where an individual may progress from cognitively normal to having mild cognitive impairment (MCI), and from MCI to dementia. Lastly, clustered or correlated events are events occurring within related units such as an individual, family, or litter, potentially in any order. Analyzing multiple time-to-event outcomes can be advantageous for many reasons, including providing a clearer understanding of disease progression over time and increasing statistical power and precision. These data can be analyzed using methods analogous to the univariate PH model, yet the statistical implications of NPH have been relatively understudied in these settings.

In this chapter, we examine the properties of PH estimators when analyzing multiple event data under a violation of the PH assumption. We begin by reviewing existing work on robust estimation in the univariate PH model, highlighting the finding of Struthers and Kalbfleisch (1986) that, under NPH, the PH model estimate depends on the distributions of the censoring and event times. After reviewing current analysis strategies for multiple event data using PH models, we evaluate the statistical implications of the Struthers and Kalbfleisch (1986) result in the multiple event setting. Finally, through simulation studies, we demonstrate that a simple extension of univariate censoring-robust estimators is insufficient to yield interpretable and replicable estimates in the multiple event setting.

5.2 Background

5.2.1 Non-Proportional Hazards in the Univariate Setting

We start with a review of foundational concepts and begin by considering the setting of univariate, right-censored, time-to-event data, where T_i, C_i, and $Z_i = [z_{i1}, \ldots, z_{ip}]^T$ denote the true event time, censoring time, and covariate vector for subject $i = 1, \ldots, n$, respectively. We further define $X_i = \min(T_i, C_i)$ as the observed event time and $\delta_i = I(X_i = T_i)$ as the

event indicator for subject i. In this setting, practitioners are often interested in estimating the association between covariates and the time to the event of interest. One approach is the Cox PH model, which characterizes event times through the hazard function:

$$\lambda_i(t) = \lambda_0(t) \exp\{\beta^T Z_i\}$$

where $\lambda_0(t)$ denotes the baseline hazard function and $\beta = [\beta_1, \ldots, \beta_p]^T$ is the vector of regression coefficients, with each element β_j representing the log-hazard ratio for covariate j. The estimating equation for the Cox PH model can be expressed as the stochastic integral:

$$\mathcal{U}(\beta) = \sum_{i=1}^{n} \int_0^\infty \left\{ Z_i - \frac{\sum_{j=1}^n Y_j(t)\ Z_j \exp\{\beta^T Z_j\}}{\sum_{j=1}^n Y_j(t)\ \exp\{\beta^T Z_j\}} \right\} dN_i(t) \equiv 0 \tag{5.1}$$

where $Y_j(t) = I(X_j \geq t)$ is the at-risk indicator for subject j at time t, and $N_i(t)$ is the counting process representing the number of events experienced by subject i between time 0 and t. As we have discussed in Chapter 2, the model parameters are obtained by solving for β such that $\mathcal{U}(\beta) = 0$.

As noted above, a primary assumption of the PH model is that the hazard functions of any two subpopulations remain proportional at all event times, the PH assumption. Since inferential models are typically prespecified based on scientific reasoning rather than empirical model fit, it is not uncommon for the PH assumption to be violated in practice. When this occurs, estimates obtained from the Cox PH model are dependent on the distribution of the censoring times. Specifically, Struthers and Kalbfleisch (1986) demonstrated that the solution to Equation 5.1 is consistent for the solution to the following equation:

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

where $f_T(\cdot)$ denotes the density function for the event times, and $S_T(\cdot)$ and $S_C(\cdot)$ denote the survival functions for the true event and censoring times, respectively. From these results, we

see that while the patterns of censoring are not typically of scientific interest, under NPH our results may depend on the accrual and dropout patterns. Moreover, these patterns will be study-dependent, posing further challenges to replicability and interpretation of results. In the context of a clinical trial, this implies that estimated treatment effects can be influenced by participant accrual and dropout patterns when the PH assumption is violated, rendering results sensitive to factors that are not of primary scientific interest.

5.2.2 A Review of Censoring-Robust Estimators

Many researchers have addressed censoring-time dependence in the univariate setting by proposing reweighted estimating equations that target the parameter corresponding to a common reference distribution, namely, a censoring distribution involving only administrative censoring at end of study, τ. The resulting estimate can be interpreted as the average covariate effect over the study period in the absence of censoring (Xu, 2000). These reweighted estimators take the following form:

$$\mathcal{U}_{CR}(\beta) = \sum_{i=1}^{n} \int_0^{\infty} W_i(t) \left\{ Z_i - \frac{\sum_{j=1}^{n} Y_j(t) W_j(t) Z_j \exp\{\beta^T Z_j\}}{\sum_{j=1}^{n} Y_j(t) W_j(t) \exp\{\beta^T Z_j\}} \right\} dN_i(t) \equiv 0 \tag{5.2}$$

As a review, Xu (2000) first proposed a reweighted estimator that removes dependence on the censoring distribution when censoring is independent of covariates, where $W_i(t) = 1/\widehat{S}_C(t)$, that can be estimated using the left-continuous Kaplan-Meier estimate of the censoring times for all subjects combined at time t. Boyd et al. (2012) then extended this to the case where censoring may be dependent on a single, binary covariate. This type of censoring, also referred to as "conditionally-independent censoring," often occurs in clinical trials where participant dropout patterns may differ by treatment arm. As a result, this requires a covariate-specific estimate of the censoring distribution for reweighting, and $W_i(t) = 1/\widehat{S}_C(t|Z_i)$ can be computed using the left-continuous, Kaplan-Meier estimate of the censoring distribution for subjects with covariate value Z_i, evaluated at time t. Finally, in the setting where censoring

may depend on multiple, possibly continuous covariates, Nguyen and Gillen (2017) proposed the use of survival trees to group individuals into censoring-similar groups for calculation of $W_i(t) = 1/\widehat{S}_C(t|Z_i)$. In each scenario, the solution to Equation 5.2 is consistent for the solution to:

$$\int_0^\infty E\Bigg\{ f_T(t|Z) \times \Bigg[Z - \frac{E\{Z S_T(t|Z)\exp(\beta^T Z)\}}{E\{S_T(t|Z)\exp(\beta^T Z)\}} \Bigg] \Bigg\} dt = 0,$$

where the estimate is no longer dependent on the censoring distribution. In each case, the resulting estimate represents a weighted average covariate effect over the study period, where the weights are determined by the survival time distribution rather than the censoring distribution.

5.2.3 Analysis of Multiple Event Data Using the PH Model

A common approach for analyzing multiple time-to-event data is to perform marginal parameter estimation using methods analogous to the univariate Cox PH model while accounting for potential within-subject/within-cluster correlation via a post-hoc correction to variance estimates. Kelly and Lim (2000) summarize established modeling strategies, discussing various choices in defining the risk intervals, baseline hazard, and risk set, as well as approaches for accounting for within-subject correlation. An additional consideration for these models is whether one should estimate a common covariate effect or an event-specific one (De Jong et al., 2020). Four models naturally arise when considering the components of the hazard function. Formally, for cluster i ($i = 1, \ldots, n$) and event type k ($k = 1, \ldots, K_i$), we have the following models:

1. Common baseline hazard, common covariate effect: $\lambda_{ik}(t) = \lambda_0(t)\exp\{\beta^T Z_{ik}\}$
2. Event-specific baseline hazards, common covariate effect: $\lambda_{ik}(t) = \lambda_{0k}(t)\exp\{\beta^T Z_{ik}\}$
3. Common baseline hazard, event-specific covariate effect: $\lambda_{ik}(t) = \lambda_0(t)\exp\{\beta_k^T Z_{ik}\}$

4. Event-specific baseline hazards, event-specific covariate effect:

   $\lambda_{ik}(t) = \lambda_{0k}(t) \exp\{\beta_k^T Z_{ik}\}$

where $\lambda_0(t)$ and $\lambda_{0k}(t)$ are unspecified baseline hazard functions, β represents the vector of model parameters that are either shared among all event strata k or specific to event type k, and Z_{ik} is the vector of covariates for cluster i, event type k.

The parameters in model 1 can be estimated by fitting a single PH model to all of the data as if the observations were independent. Similarly, the parameters in model 3 can be estimated via a single PH model including an interaction between covariates and event type. The parameters from model 2 can be estimated by fitting a PH model that is stratified by event type, and lastly, the parameters from model 4 can be estimated by fitting separate PH models for each event type. As mentioned previously, post-hoc adjustments must be made to the variance estimators to account for within-cluster correlation for all of these models, since standard variance estimators assuming independence are generally inconsistent (Lin, 1994).

In practice, model choice is guided by the scientific rationale and relevant statistical considerations. For example, the choice between a common or event-specific baseline hazard typically reflects assumptions about the event-generating process, as we have discussed in previous chapters. A common baseline hazard assumes that all event types within a cluster occur with a common baseline rate. Conversely, event-specific baseline hazards allow the baseline rate of experiencing each event type to differ across event strata. In settings such as myocardial infarction, where an individual's risk of experiencing a second heart attack increases after the first (Böhm et al., 2021), this would be an appropriate assumption.

When deciding between a common covariate effect and an event-specific one, practitioners should consider the scientific goals of the analysis, including the hypothesized relationship between the event types and the covariates. If this relationship is expected to change or if

interest lies directly in the event-specific covariate effect, then stratified covariate effects may be more appropriate. If covariate effects are expected to be similar across all event types, however, then estimating a common covariate effect can improve precision in small-sample settings. An example of this occurs in clinical trials that test the effect of a treatment when the endpoint of interest may be recurrent or include multiple events. In this setting, primary analyses may specify a common covariate effect across events to gain power for detecting the treatment effect (Amorim and Cai, 2015).

To confirm the validity of the common treatment effect assumption, a natural diagnostic approach might include estimating the event-specific treatment effects through a completely stratified PH model, as demonstrated by Chen and Wei (1997). As in the univariate analysis of a treatment effect, the problem of censoring dependence will arise in this stratified analysis when the PH assumption does not hold. As shown above, this issue can typically be addressed using censoring-robust estimators; however, a new challenge arises with event-stratified estimates because the densities and survival functions for event times, $f_T(t|Z)$ and $S_T(t|Z)$, may vary across event strata even when the true covariate effect is identical across strata. We discuss this situation in detail in the following sections.

5.3 Multiple Event Analysis Under Non-Proportional Hazards

To extend the result of Struthers and Kalbfleisch (1986) to the multiple event setting, consider the clinical trial setting where interest lies in a single treatment effect across recurrent events. We assume that recurrent events are ordered and adopt within-individual clustering for simplicity of presentation. We note, however, that the methodology presented here is also applicable to other types of multiple event data (e.g., unordered events or events of different types) clustered at either the individual or group level.

For the remainder of this chapter, let T_{ik}, C_{ik}, and Z_i denote the event time, censoring time, and treatment indicator for the k^{th} $(k = 1, ..., K_i)$ event recurrence of subject i $(i = 1, \ldots, n)$. We assume that $T_{ik} \perp C_{ik}|Z_i$. We note that Z_i represents the treatment for subject i across multiple recurrences k but can be easily modified to represent a treatment effect that may change by event recurrence. We further define the observed event time as $X_{ik} = \min(T_{ik}, C_{ik})$ and $\delta_{ik} = I(X_{ik} = T_{ik})$ as the event indicator of event recurrence k for subject i. Suppose the baseline hazard is expected to differ by recurrence, so to gain precision for estimating the effect of the treatment Z, a model of the form $\lambda_{ik}(t) = \lambda_{0k}(t) \exp\{\beta^T Z_i\}$ (model 2) is considered. To verify that the assumption of a common β is appropriate, event-stratified models (model 4) are fit first. Stratifying the data on event k, consider estimating the stratum-specific treatment effect for the k^{th} event recurrence under the PH model, $\widehat{\beta}_{k,naive}$, using the following estimating equation:

$$\mathcal{U}^{(k)}(\beta) = \sum_{i=1}^{n_k} \int_0^\infty \left\{ Z_i - \frac{\sum_{j=1}^{n_k} Y_{jk}(t) Z_j \exp\{\beta^T Z_j\}}{\sum_{j=1}^{n_k} Y_{jk}(t) \exp\{\beta^T Z_j\}} \right\} dN_{ik}(t) \equiv 0, \tag{5.3}$$

where n_k denotes the number of subjects who are at risk for the k^{th} event recurrence.

Proposition 5.1. *The solution to Equation 5.3, $\widehat{\beta}_{k,naive}$, is consistent for the solution to:*

$$\int_0^\infty E_k \left\{ f_{T_k}(t|Z) S_{C_k}(t|Z) \times \left[Z - \frac{E_k\{Z S_{T_k}(t|Z) S_{C_k}(t|Z) \exp(\beta^T Z)\}}{E_k\{S_{T_k}(t|Z) S_{C_k(t|Z)} \exp(\beta^T Z)\}} \right] \right\} dt = 0 \tag{5.4}$$

where $E_k(\cdot)$ is the expectation taken with respect to the distribution of Z in the k^{th} stratum, $f_{T_k}(\cdot)$ is the density function, and $S_{T_k}(\cdot)$ and $S_{C_k}(\cdot)$ denote the survival functions for the true event and censoring times for the k^{th} event recurrence, respectively.

This result follows directly from the results of Struthers and Kalbfleisch (1986). Since the goal is to compare stratified treatment effects across event recurrences to assess treatment effect heterogeneity, it is crucial that the resulting estimates do not depend on factors unrelated to the scientific question of interest, such as the censoring distribution. As is apparent from

Equation 5.4, this requirement is not satisfied when the PH assumption is violated.

Implementing the censoring-robust estimator proposed by Boyd et al. (2012), we may estimate a stratum-specific treatment effect for the k^{th} event recurrence which is robust to changes in the censoring distribution, $\widehat{\beta}_{k,CR}$.

Proposition 5.2. *Denoting $\widehat{\beta}_{k,CR}$ as the estimated censoring-robust stratum-specific treatment effect for the k^{th} event recurrence, it follows that $\widehat{\beta}_{k,CR}$ is consistent for the value of β that solves:*

$$\int_0^\infty E_k\left\{ f_{T_k}(t|Z) \times \left[Z - \frac{E_k\{Z S_{T_k}(t|Z)\exp(\beta^T Z)\}}{E_k\{S_{T_k}(t|Z)\exp(\beta^T Z)\}} \right] \right\} dt = 0 \tag{5.5}$$

where $f_{T_k}(\cdot)$ and $S_{T_k}(\cdot)$ are the density and survival functions for the true event times for the k^{th} event recurrence.

Although the dependence on the censoring is removed, dependence on $f_{T_k}(\cdot)$ and $S_{T_k}(\cdot)$ still remains. In the univariate setting, this dependence does not matter. In fact, it is critical that estimates from the PH model depend on these quantities, since they provide insight into the underlying survival mechanism giving rise to the estimated log-hazard ratios. When comparing estimates from Equation 5.2 across studies in the univariate case, we are inherently assuming that each study consists of observations that are a random sample of the same study population. Equivalently, we are reasonably assuming that the underlying $f_T(\cdot)$ and therefore $S_T(\cdot)$ are the same across replications. In the multiple event case, however, $f_{T_k}(\cdot)$ and $S_{T_k}(\cdot)$ can change with each recurrence k despite being from the same population, and we will illustrate examples where this problem occurs.

5.3.1 Example: Constant Baseline Hazard with a Time-Varying Covariate Effect

In this analytic example, we consider the setting where survival times follow a distribution with a constant baseline hazard and a time-varying covariate effect. That is, $f_{T_k}(t|Z) = \lambda_k \exp\{-\lambda_k t\}$ where λ_k will be parameterized by a hazard function with an event-specific baseline hazard, or $\lambda_k = \lambda_{0k} e^{\alpha(t)^T Z}$. It follows then that we can define the event-specific survival function, $S_{T_k}(t|Z) = \exp\{-\lambda_k t\} = \exp\{-\lambda_{0k} e^{\alpha(t)^T Z} \cdot t\}$. We note that $\alpha(t)$ represents a time-varying covariate effect that is identical across all k.

Proposition 5.3. *Assuming that censoring dependence has been accounted for using appropriate censoring-robust methods, it follows from Equation 5.5 that the estimate from the PH model for the k^{th} recurrence, $\widehat{\beta}_{k,CR}$, is consistent for the value of β that solves:*

$$\int_0^\infty E_k \Bigg\{ \lambda_{0k} e^{\alpha(t)^T Z} \exp\{-\lambda_{0k} e^{\alpha(t)^T Z} \cdot t\} \times \left[Z - \frac{E_k\{Z \exp(-\lambda_{0k} e^{\alpha(t)^T Z} \cdot t) \exp(\beta^T Z)\}}{E_k\{\exp(-\lambda_{0k} e^{\alpha(t)^T Z} \cdot t) \exp(\beta^T Z)\}} \right] \Bigg\} dt = 0,$$

which is dependent on the event-specific λ_{0k} via $f_{T_k}(t|Z)$ and $S_{T_k}(t|Z)$.

In this example, it is clear that no algebraic manipulation can be done to completely remove the effect of λ_{0k} since it appears multiplicatively inside the exponential term. Concisely, the functions $f_{T_k}(t|Z)$ and $S_{T_k}(t|Z)$ are non-separable with respect to $\lambda_{0k}(t)$.

5.3.2 General Setting

In the general, non-parametric setting, we have $f_{T_k}(t|Z) = S_{T_k}(t|Z) \cdot \lambda_k(t)$ and $S_{T_k}(t|Z) = \exp[-\int_0^t \lambda_k(s) ds]$. Assuming hazard functions with event-specific baseline hazards and a time-varying effect, or $\lambda_k(t) = \lambda_{0k}(t) e^{\alpha(t)^T Z}$, we can further define $f_{T_k}(t|Z) = \exp[-\int_0^t \lambda_{0k}(s) e^{\alpha(s)^T Z} ds] \cdot \lambda_{0k}(t) e^{\alpha(t)^T Z}$, and $S_{T_k}(t|Z) = \exp[-\int_0^t \lambda_{0k}(s) \exp^{\alpha(s)^T Z} ds]$. Similar to the expo-

nential setting, $\alpha(t)$ is a time-varying covariate effect for which we are interested in estimating the average. The event-specific estimate from the PH model for the k^{th} recurrence, $\widehat{\beta}_{k,CR}$, is consistent for the value of β that solves:

$$\int_0^\infty E_k\Bigg\{\exp\Bigg[-\int_0^t \lambda_{0k}(s)e^{\alpha(s)^T Z}ds\Bigg]\cdot\lambda_{0k}(t)e^{\alpha(t)^T Z}\times \\ \Bigg[Z-\frac{E_k\{Z\exp[-\int_0^t \lambda_{0k}(s)\exp^{\alpha(s)^T Z}ds]\exp(\beta^T Z)\}}{E_k\{\exp[-\int_0^t \lambda_{0k}(s)\exp^{\alpha(s)^T Z}ds]\exp(\beta^T Z)\}}\Bigg]\Bigg\}dt=0.$$

Since the effect $\alpha(t)$ is identical across recurrence strata, the ideal estimator would produce an estimate consistent for the same quantity β across all k, or equivalently, correctly identify that the estimated average covariate effects are the same across recurrences. Due to the difference in $\lambda_{0k}(t)$ across recurrences and the non-separability of both $f_{T_k}(t|Z)$ and $S_{T_k}(t|Z)$ with respect to $\lambda_{0k}(t)$, however, $\widehat{\beta}_{k,CR}$ varies with k, resulting in an incorrect comparison of the average covariate effects across event strata. This is problematic since the baseline hazards are not typically of scientific interest and are meant to be eliminated when using this model.

Unlike the censoring distribution, which is completely a nuisance parameter in estimation, the survival distribution, although containing a nuisance parameter, is essential to estimating the treatment effect. As such, a solution such as reweighting by the inverse of an estimate of $f_{T_k}(t|Z)$ and $S_{T_k}(t|Z)$ would completely remove the signal we are trying to detect. Further, due to the complete confounding of the baseline hazards $\lambda_{0k}(t)$ and the covariate effect $\alpha(t)$ in the hazard function, disentanglement of the effect of the nuisance parameter and the effect of the covariates is not directly possible, implying that event-specific covariate effects are not truly comparable when NPH is present and baseline hazards differ across multiple events. We illustrate this concretely with empirical examples in the next section.

5.4 Numerical Illustrations

In this section, we consider the setting of a randomized clinical trial where primary interest lies in estimating a treatment effect using recurrent event data but recurrence-specific treatment effects are considered to validate the choice of estimating a common covariate effect. In the simulations presented, participants experience at most two recurrent events, and $Z_i \sim Bern(0.5)$ represents the treatment which remains constant for all events. For the k^{th} recurrence of subject i, T_{ik} is generated from a distribution with piecewise-constant hazard function where $\lambda_{ik}(t) = \lambda_{0k} \exp\{\log(0.5) \cdot I(t \leq 3)Z_i + v_i\}$ is formulated with an early-diverging hazard. We include an individual-level random effect $v_i \sim N(0, 1)$ to induce correlation within subject.

We also explore a variety of censoring settings, including settings where censoring patterns differ by recurrent event and settings similar to that explored by Boyd et al. (2012) where patterns differ by treatment arm. As such, censoring times C_{ik} were generated from a variety of "powered-uniform" distributions $[S_C(t) = 1-(t/\gamma)^r I(0 \leq t \leq \gamma)]$, denoted as $PUnif(\gamma, r)$, that differed by recurrence and/or treatment arm. More specifically, we compare results under three settings:

1. Treatment-specific censoring: $C_{i1}, C_{i2} \sim PUnif(10.5, 2.5 - Z_i)$
2. Event-specific censoring: $C_{i1} \sim PUnif(10.5, 0.5)$ and $C_{i2} \sim PUnif(10.5, 3)$
3. Treatment and event-specific censoring: $C_{i1} \sim PUnif(10.5, 0.5 + 0.5Z_i)$ and $C_{i2} \sim PUnif(10.5, 2.5 - Z_i)$

In addition to this censoring, we implement administrative censoring at $\tau = 10$ to ensure all observed times are over a common support.

Observed times X_{ik} were defined as $\min(T_{ik}, C_{ik})$, and event-specific indicators δ_{ik} were 1 if $X_{ik} = T_{ik}$ and 0 otherwise. Here, we utilize the gap-time formulation, meaning X_{ik} represents

		$\beta_1^* = -0.265$											
		First Recurrence						Second Recurrence					
	Cens.	% Cens.	Total	Est. (% Bias)	Emp. SE	Rob. SE	Cov. Prob.	%. Cens.	Total	Est. (% Bias)	Emp. SE	Rob. SE	Cov. Prob.
Cox PH	1	43.2	5000.0	-0.327 (-23.62)	0.037	0.038	0.620	41.7	2837.5	-0.326 (-23.32)	0.049	0.051	0.774
BKG	1			-0.255 (3.56)	0.042	0.043	0.946			-0.251 (5.28)	0.056	0.058	0.952
Cox PH	2	69.3	5000.0	-0.412 (-55.54)	0.051	0.051	0.172	35.6	1534.0	-0.297 (-12.29)	0.061	0.064	0.926
BKG	2			-0.257 (2.74)	0.070	0.068	0.944			-0.253 (4.51)	0.066	0.069	0.946
Cox PH	3	62.6	5000.0	-0.393 (-48.38)	0.045	0.047	0.238	43.6	1872.1	-0.327 (-23.52)	0.059	0.062	0.846
BKG	3			-0.255 (3.72)	0.060	0.060	0.934			-0.255 (3.77)	0.067	0.069	0.960

Table 5.1: Simulation results comparing the event-specific estimates from the Cox PH and BKG estimators in the setting of a common baseline hazard across event types. "True" estimates are based on a model with $n = 100,000$, and percent bias was calculated as $E\{(\widehat{\beta} - \beta_1^*)/|\beta_1^*|\} \cdot 100$.

the time since the previous recurrence. We present the results from fitting recurrence-specific Cox models of the form: $\lambda_k(t) = \lambda_{0k}(t) \exp\{\beta_k Z\}$ for $k = 1, 2$. In all simulations presented, we compare the estimates for β_1 and β_2 under the standard PH model (Cox PH) as well as the estimator proposed by Boyd et al. (2012) (hereafter BKG), which removes dependence on the censoring distribution. For the settings of common and event-specific baseline hazards, we present model estimates (Est.), empirical standard error (Emp. SE), model-based robust standard errors (Rob. SE), and corresponding coverage probability (Cov. Prob.). Since there is no "truth" in this case, we compare all estimates to the first recurrence estimate taken in the absence of censoring. As such, coverage probability and percent bias are calculated based on a large-sample ($n = 100,000$) estimate of β_1^*, the average log-hazard ratio over $[0, 10]$ in the absence of intermittent censoring for the first recurrence. All simulations were run for 500 iterations.

5.4.1 Common Baseline Hazard

We start by illustrating the baseline setting where the data-generating model is the same across all recurrences. That is, all recurrences share a common baseline hazard and the same functional form of β. The underlying data generating model is $\lambda_{ik}(t) = 0.15 \exp\{\log(0.5) \cdot I(t \leq 3)Z_i + v_i\}$. Table 5.1 presents the results for the two estimators (Cox PH and BKG).

		$\beta_1^* = -0.258$											
		First Recurrence						Second Recurrence					
	Cens.	% Cens.	Total	Est. (% Bias)	Emp. SE	Rob. SE	Cov. Prob.	%. Cens.	Total	Est. (% Bias)	Emp. SE	Rob. SE	Cov. Prob.
Cox PH	1	43.2	5000.0	-0.328 (-27.46)	0.039	0.038	0.526	20.9	2838.0	-0.407 (-57.90)	0.040	0.043	0.052
BKG	1			-0.257 (0.25)	0.043	0.043	0.952			-0.348 (-34.95)	0.046	0.047	0.485
Cox PH	2	69.3	5000.0	-0.409 (-58.71)	0.054	0.051	0.178	14.4	1535.4	-0.384 (-49.09)	0.054	0.056	0.376
BKG	2			-0.253 (1.58)	0.070	0.068	0.936			-0.350 (-36.01)	0.058	0.058	0.642
Cox PH	3	62.6	5000.0	-0.393 (-52.67)	0.049	0.047	0.180	22.8	1871.1	-0.410 (-59.36)	0.052	0.053	0.174
BKG	3			-0.253 (1.78)	0.061	0.059	0.942			-0.355 (-37.84)	0.059	0.056	0.582

Table 5.2: Simulation results comparing the event-specific estimates from the Cox PH and BKG estimators in the setting of an event-specific baseline hazard across event types. "True" estimates are based on a model with $n = 100,000$, and percent bias was calculated as $E\{(\widehat{\beta} - \beta_1^*)/|\beta_1^*|\} \cdot 100$.

As expected, the Cox PH estimator provides estimates that are dependent on the censoring distribution for both recurrences. The BKG estimator corrects for any dependence on the censoring distribution, resulting in minimal bias for estimating β_1^* and clearly targets the same quantity for both recurrences.

5.4.2 Event-Specific Baseline Hazards

Next, we consider the setting where the baseline hazard differs for each recurrence but the underlying treatment effect is the same. Specifically, the data-generating model is $\lambda_{ik}(t) = \lambda_{0k} \exp\{\log(0.5) \cdot I(t \leq 3)Z_i + v_i\}$ where $\lambda_{01} = 0.15$ and $\lambda_{02} = 0.3$ were chosen to reflect the increasing baseline hazards commonly observed in disease settings such as cancer, renal disease, and heart disease. Although an NPH effect is present in this setting, the estimated recurrence-specific treatment effects should be the same across recurrences, since the underlying relationship between the risk of experiencing the event and the treatment is identical across recurrences. Table 5.2 presents the estimated treatment effects with corresponding details for each event recurrence using the Cox PH and BKG estimators.

Similar to the common baseline hazard setting, the Cox PH estimator will change with different underlying censoring since NPH is present. Although it appears that estimates for both recurrences tend to be similar for censoring settings 2 and 3, this is not because the

estimators converge to a common estimand, but rather reflects a coincidental attenuation of the two estimates toward one another (See Figure 5.2). The BKG estimator estimates β_1^* with minimal bias and nominal coverage probability for the first recurrence, which is to be expected. For the second recurrence, the BKG estimator is clearly not targeting this same quantity and the resulting estimates represent time-averaged treatment effects with weights that differ from those for the first recurrence.

5.4.3 Across Multiple Settings

We compared the Cox PH and BKG estimators for the first and second recurrences across a range of baseline hazard ratios. Figure 5.1 presents results for the setting with no intermittent censoring, and Figure 5.2 contains results under the third censoring setting (i.e., event- and treatment-specific censoring). As seen in Figure 5.1, the Cox and BKG estimators are equivalent for the first recurrence and for the second recurrence in the absence of intermittent censoring. This is expected, as the BKG estimator is designed to target the average treatment effect in the absence of intermittent censoring. When there is no difference in the baseline hazards, all estimators agree for the first and second recurrences. As the baseline hazards separate, the differential dependence on the event-time distribution becomes evident, with the second-recurrence estimates departing increasingly from those of the first. Figure 5.2 illustrates the additional dependence on the censoring distribution. Here, both Cox estimators are biased relative to a censoring-invariant estimand. The BKG estimator removes this dependence on censoring, and when there is no difference in the baseline hazards, the two event estimates agree. As this difference increases, however, the BKG estimates diverge despite the same underlying treatment effect.

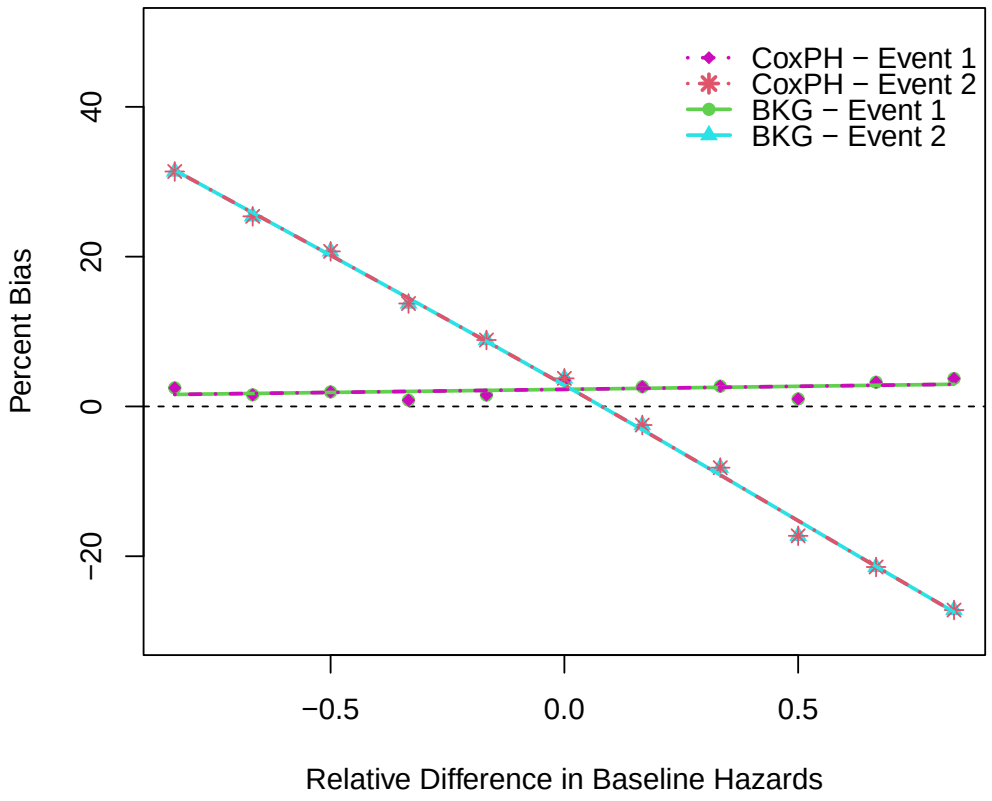


Figure 5.1: Estimated treatment effects using the Cox PH and BKG estimators across a variety of baseline hazard settings in the absence of intermittent censoring. Percent bias is calculated as $E\{(\widehat{\beta} - \beta_1^*)/|\beta_1^*|\} \cdot 100$, and relative difference in baseline hazards is $(\lambda_{01} - \lambda_{02})/\lambda_{01}$.

5.5 Discussion

It has been clearly established that under violation of model assumptions, the estimate from the Cox PH model is dependent on the distributions of the censoring and event times. Past work has addressed the dependence on the censoring distribution through development of estimators that remove this dependence and map back to a weighted average covariate effect under no intermittent censoring. Since these censoring-robust estimators have been studied only in the univariate setting, the dependence on the event-time distribution has not previously posed a challenge. The results we have presented here demonstrate that the

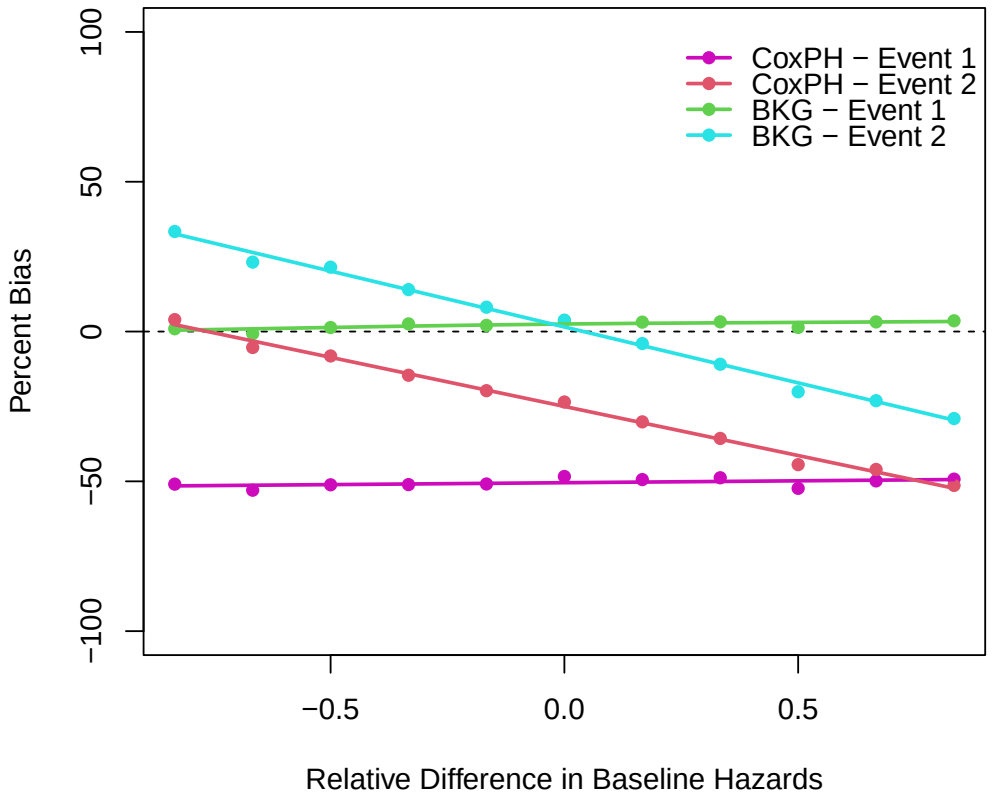


Figure 5.2: Estimated treatment effects using the Cox PH and BKG estimators across a variety of baseline hazard settings. Percent bias is calculated as $E\{(\widehat{\beta} - \beta_1^*)/|\beta_1^*|\} \cdot 100$, and relative difference in baseline hazards is $(\lambda_{01} - \lambda_{02})/\lambda_{01}$.

implications of NPH extend meaningfully to the multiple event setting.

Under NPH, standard Cox PH estimators depend on both the censoring distribution and the baseline hazard, neither of which is typically of primary scientific interest. While censoring-robust estimators resolve the first problem in the univariate setting, they are insufficient in the multiple event setting when event-specific baseline hazards differ across strata. The non-separability of the baseline hazard from the covariate effect precludes direct comparison of stratum-specific treatment effects, even after censoring adjustment. These findings suggest that practitioners should exercise caution when fitting stratified PH models to validate a

common treatment effect assumption under NPH, as apparent differences in event-specific estimates may reflect differences in the baseline hazard rather than true treatment effect heterogeneity. Our results also raise further concerns regarding the scientific interpretation of an estimated covariate effect that is pooled across multiple event times when the PH assumption is violated.

Ideally, there would be a way to remove the dependence of the estimand on the baseline hazard to create a fair comparison between event-specific treatment effects similar to the way dependence on the censoring distribution is removed. Due to the complex confounding of the baseline hazard and parameters in the survival distribution, however, we believe doing so would be difficult if not impossible. One approach might be to iteratively estimate the cumulative baseline hazard and regression coefficients, weighting the estimating equation by the inverse of the estimated baseline hazard; however, convergence of such a procedure is not guaranteed, and time-invariant estimates of a truly time-varying coefficient may ultimately be attenuated. It appears that the only way to remove this dependence on the baseline hazard function is to directly estimate the time-varying parameter function $\beta(t)$. This raises the question of whether estimating the full parameter function $\beta(t)$ is preferable to reporting a weighted average thereof. In addition to the issue of comparability across event strata, this finding raises important questions about the interpretation of the pooled treatment effect when model assumptions are violated, and whether this modeling strategy remains appropriate in such settings.

It is important to specify analyses prior to viewing the data to avoid multiple testing biases. As such, it would be likely for practitioners to plan to estimate a pooled covariate effect across recurrences and later learn that the data violate the PH assumption. In the case where NPH were to be encountered via diagnostics, we would advise against conducting the pooled analysis as it is no longer clear whether the event-specific covariate effects are empirically comparable and what the interpretation of a pooled covariate effect would be. Instead, we

would recommend reporting the stratified covariate effects along with the Kaplan-Meier plots to provide a more complete picture of the underlying behavior of the time-varying covariate effects for each event recurrence.

Chapter 6

Conclusion

Throughout this dissertation, we have emphasized the importance of tools that facilitate robust and efficient estimation for time-to-event data. Specifically, we saw the utility of the NCC design for maximizing available information while reducing data collection costs. We also saw the benefits of implementing the Samuelsen estimator to increase efficiency of the NCC design. In addition to efficiency, we highlighted the importance of using estimators that are robust to potential violations of model assumptions. Notably, censoring-robust estimators have been developed to guard against cases of NPH and allow for estimation of an interpretable and replicable estimand.

These methods have clear benefits and often improve the analysis of time-to-event data, but they are not without their limitations. As we have established, the NCC design can exacerbate pre-existing issues of underrepresentation in research by sampling a small number of individuals from that group. Additionally, there have been few examples of its use in the multiple event survival setting. Censoring-robust estimators have been thoroughly developed in the univariate setting, but there are currently no examples of their use in the multiple event setting. In fact, a unique problem arises in the multiple event setting that is not

directly accounted for by traditional censoring-robust estimators. Particularly, the estimand is dependent on the distribution of the event times, making comparisons across event strata difficult. This dissertation addressed these shortcomings to draw awareness to and improve current methodology.

In Chapter 3, we proposed a weighted NCC design that allows for oversampling of controls from underrepresented subpopulations. Motivated by the need to precisely estimate subpopulation-specific biomarker effects in AD research, this design maximizes information on the expensive-to-collect covariate while ensuring a more representative sample than the classic NCC design. In addition to implementing an efficient sampling design, our proposed estimator uses sampled controls forward and backward in time to gain the most information from them and is consistent for the estimand under the full cohort. We also developed two variance estimators that are robust to model misspecification and performed well in simulation studies.

We pivoted to multiple time-to-event endpoints in Chapter 4, where we developed an NCC design sampling framework for this setting. We presented two NCC sampling schemes that were inspired by the risk set formulations of PH models for multiple event data. Despite introducing two options, we generally recommend event-specific sampling, which stratifies sampling by event number, to pooled sampling, which treats all observations as independent. In addition to the benefits of having a primary framework for this setting, our results demonstrated that event-specific sampling often presents a gain in precision for coefficient estimates from models assuming event-specific baseline hazards and a minimal loss in precision for estimates from models that assume a common baseline hazard. We also derived corresponding variance estimators that provided nominal coverage for valid inference using models with either baseline hazard specification.

Finally, we addressed censoring-robust estimation in the multiple event setting in Chapter 5. We interrogated the impacts of all components in the quantity for which the PH estimate is

consistent, proposed by Struthers and Kalbfleisch (1986). Specifically, we highlighted that there may be settings where the estimand's dependence on the event times impacts results and that this cannot be addressed by traditional censoring-robust estimators. In the clinical trial setting where information from multiple endpoints may be used to gain precision on the treatment effect, practitioners may fit models completely stratified by event number to compare the treatment effect across these strata. Even under current censoring-robust methods, the stratified analyses are still dependent on a nuisance parameter for estimation. Specifically, we draw attention to the case where differences in the baseline hazard impact the distribution of the survival times, resulting in an estimate that changes with each event stratum despite a common underlying treatment effect. We also demonstrate that this problem cannot be easily solved using reweighting methods, leading us to question the use of the PH model under NPH in the multiple event setting.

6.1 Future Directions

Each of the proposed methods was developed as a step to achieve the overarching goal of equitable, efficient, and robust inference as motivated by challenges in AD biomarker discovery. Independently, they fill existing gaps in current statistical methodology, but when combined, they can be used to address a variety of limitations across many settings. For example, the weighted NCC design can be combined with the multiple event NCC sampling framework to assess subpopulation-specific biomarker effects in the multiple event setting. Moreover, ideas from censoring-robust estimation can be combined with these methods to ensure replicable results under NPH. Along with these direct extensions, there are a few concepts that give rise to additional future work in this area.

6.1.1 Samuelsen Power and Sample Size Calculations

In all of the aforementioned methods utilizing the NCC design, we implemented the use of sampled controls forward and backward in time to improve efficiency of the classic estimator via extensions of the Samuelsen estimator. This framework can be useful in NCC settings, as it increases efficiency by providing more information at each event time without increasing the sample size, and allows for consistent estimation of the full cohort estimand under violation of the PH assumption. Despite its benefits, the Samuelsen estimator is not as widely used as the classic NCC design, likely due to general unfamiliarity with it. On a similar note, sample size and power calculations have not yet been developed for the Samuelsen estimator, making its practical use in studies difficult for practitioners. Development of these calculations to ensure easier use of the Samuelsen framework in NCC design studies would be a key component of future work.

In Chapter 3, we proposed weighted sampling of controls in each risk set to achieve more thorough subpopulation representation for analyses. When using this method, practitioners are able to specify a desired probability of sampling controls from each subgroup; however, as observed in our examples, the distribution of sampled controls does not perfectly match the prespecified desired probabilities due to the resampling of controls at different rates for each subgroup. In future work, it would be necessary to explore this dynamic of our proposed extension in addition to the classic Samuelsen framework to ensure practitioners are able to utilize these designs with full understanding of the statistical implications. The result of this work would ideally guide practitioners' decisions about the required sample size needed to accurately assess the scientific questions of interest. This work must also be done for our proposed NCC framework for the multiple event setting presented in Chapter 4.

6.1.2 Extensions to NPH in the Multiple Event Setting

As discussed, the ideas presented in Chapter 5 were motivated by the need to estimate a common estimand across event strata in order to make a fair comparison and determine whether to estimate a combined or stratified treatment effect. In light of the impacts of the baseline hazard on the estimand via the survival distribution, there is still an ongoing discussion as to what the combined estimate really means and if it makes sense to interpret. Different arguments can be made for various settings, but it would be worthwhile to have a lengthy discussion and perhaps conduct a meta-analysis of what has been done in the past and how results may have be impacted as a result.

Moreover, the results in Chapter 5 beg the question as to whether there is a more interpretable estimand in settings where NPH may be present. Due to its explicit interpretability, one possible estimand that may be of interest is restricted mean survival time. Explicitly, restricted mean survival time is the area under the survival curve up to a specific time point, τ, or $\mu(\tau) = \int_0^\tau S(t)dt$, and it can be interpreted as the average event-free time from time 0 to τ. When comparing different groups, we can consider the difference in restricted mean survival time, or $\mu_1(\tau) - \mu_0(\tau) = \int_0^\tau \{S_1(t) - S_0(t)\}dt$. The difference in restricted mean survival time does not impose the assumption of PH, and the estimand of interest must be explicitly specified via the choice of the time τ, making it a potentially fairer way to compare across event recurrence strata.

Bibliography

Amorim, L. D. and Cai, J. (2015). Modelling recurrent events: a tutorial for analysis in epidemiology. *International Journal of Epidemiology*, 44(1):324–333.

Andersen, P. K. and Gill, R. D. (1982). Cox's Regression Model for Counting Processes: A Large Sample Study. *The Annals of Statistics*, 10(4).

Binder, D. A. (1992). Fitting Cox's proportional hazards models from survey data. *Biometrika*, 79(1):139–147.

Blennow, K., Hampel, H., Weiner, M., and Zetterberg, H. (2010). Cerebrospinal fluid and plasma biomarkers in Alzheimer disease. *Nature Reviews Neurology*, 6(3):131–144.

Blom, E. S., Giedraitis, V., Zetterberg, H., Fukumoto, H., Blennow, K., Hyman, B. T., Irizarry, M. C., Wahlund, L.-O., Lannfelt, L., and Ingelsson, M. (2009). Rapid Progression from Mild Cognitive Impairment to Alzheimer's Disease in Subjects with Elevated Levels of Tau in Cerebrospinal Fluid and the *APOE* $\varepsilon 4/\varepsilon 4$ Genotype. *Dementia and Geriatric Cognitive Disorders*, 27(5):458–464.

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

Breslow, N. E., Amorim, G., Pettinger, M. B., and Rossouw, J. (2013). Using the Whole Cohort in the Analysis of Case-Control Data: Application to the Women's Health Initiative. *Statistics in Biosciences*, 5(2):232–249.

Böhm, M., Schumacher, H., Teo, K. K., Lonn, E. M., Lauder, L., Mancia, G., Redon, J., Schmieder, R. E., Sliwa, K., Marx, N., Weber, M. A., Williams, B., Yusuf, S., Mann, J. F., and Mahfoud, F. (2021). Cardiovascular outcomes in patients at high cardiovascular risk with previous myocardial infarction or stroke. *Journal of Hypertension*, 39(8):1602–1610.

Chen, L. and Wei, L. J. (1997). Analysis of Multivariate Survival Times with Non-Proportional Hazards Models. In Bickel, P., Diggle, P., Fienberg, S., Krickeberg, K., Olkin, I., Wermuth, N., Zeger, S., Lin, D. Y., and Fleming, T. R., editors, *Proceedings of the First Seattle Symposium in Biostatistics*, volume 123, pages 23–36. Springer US, New York, NY. Series Title: Lecture Notes in Statistics.

Chin, A. L., Negash, S., and Hamilton, R. (2011). Diversity and Disparity in Dementia: The Impact of Ethnoracial Differences in Alzheimer Disease. *Alzheimer Disease & Associated Disorders*, 25(3):187–195.

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220.

Cushman, M. (2004). Estrogen Plus Progestin and Risk of Venous Thrombosis. *JAMA*, 292(13):1573.

De Jong, V. M., Moons, K. G., Riley, R. D., Tudur Smith, C., Marson, A. G., Eijkemans, M. J., and Debray, T. P. (2020). Individual participant data meta-analysis of intervention studies with time-to-event outcomes: A review of the methodology and an applied example. *Research Synthesis Methods*, 11(2):148–168.

Goldstein, L. and Langholz, B. (1992). Asymptotic Theory for Nested Case-Control Sampling in the Cox Regression Model. *The Annals of Statistics*, 20(4).

Hou, Y., Dan, X., Babbar, M., Wei, Y., Hasselbalch, S. G., Croteau, D. L., and Bohr, V. A. (2019). Ageing as a risk factor for neurodegenerative disease. *Nature Reviews Neurology*, 15(10):565–581.

Howell, J. C., Watts, K. D., Parker, M. W., Wu, J., Kollhoff, A., Wingo, T. S., Dorbin, C. D., Qiu, D., and Hu, W. T. (2017). Race modifies the relationship between cognition and Alzheimer's disease cerebrospinal fluid biomarkers. *Alzheimer's Research & Therapy*, 9(1):88.

Jazić, I., Haneuse, S., French, B., MacGrogan, G., and Rondeau, V. (2019). Design and analysis of nested case–control studies for recurrent events subject to a terminal event. *Statistics in Medicine*, 38(22):4348–4362.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. Wiley Series in Probability and Statistics. Wiley, 1 edition.

Kaplan, E. L. and Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(282):457–481.

Kelly, P. J. and Lim, L. L.-Y. (2000). Survival analysis for recurrent event data: an application to childhood infectious diseases. *Statistics in Medicine*, 19(1):13–33.

Langholz, B. and Borgan, O. R. (1995). Counter-matching: A stratified nested case-control sampling method. *Biometrika*, 82(1):69–79.

Langholz, B. and Thomas, D. C. (1991). Efficiency of Cohort Sampling Designs: Some Surprising Results. *Biometrics*, 47(4):1563.

Lee, E. W., Wei, L. J., Amato, D. A., and Leurgans, S. (1992). Cox-Type Regression Analysis for Large Numbers of Small Groups of Correlated Failure Time Observations. In Klein, J. P. and Goel, P. K., editors, *Survival Analysis: State of the Art*, pages 237–247. Springer Netherlands, Dordrecht.

Lin, D. (2000). On fitting Cox's proportional hazards models to survey data. *Biometrika*, 87(1):37–47.

Lin, D. Y. (1994). Cox regression analysis of multivariate failure time data: The marginal approach. *Statistics in Medicine*, 13(21):2233–2247.

Liu, L., Wolfe, R. A., and Huang, X. (2004). Shared Frailty Models for Recurrent Events and a Terminal Event. *Biometrics*, 60(3):747–756.

Lønning, P. E., Nikolaienko, O., Pan, K., Kurian, A. W., Eikesdal, H. P., Pettinger, M., Anderson, G. L., Prentice, R. L., Chlebowski, R. T., and Knappskog, S. (2022). Constitutional *BRCA1* Methylation and Risk of Incident Triple-Negative Breast Cancer and High-grade Serous Ovarian Cancer. *JAMA Oncology*, 8(11):1579.

Mayeda, E. R., Glymour, M., Quesenberry, C. P., and Whitmer, R. A. (2016). Inequalities in dementia incidence between six racial and ethnic groups over 14 years. *Alzheimer's & Dementia*, 12(3):216–224.

Nguyen, V. Q. and Gillen, D. L. (2017). 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.

Nuño, M. M. and Gillen, D. L. (2017). Chapter 5 - Alternative Sampling Designs for Time-to-Event Data With Applications to Biomarker Discovery in Alzheimer's Disease. In Srinivasa Rao, A. S., Pyne, S., and Rao, C., editors, *Handbook of Statistics*, volume 36, pages 105–166. Elsevier.

Nuño, M. M. and Gillen, D. L. (2020). On estimation in the nested case-control design under nonproportional hazards. *Scandinavian Journal of Statistics*, 49(1):265–286.

Nuño, M. M. and Gillen, D. L. (2021). Robust estimation in the nested case-control design under a misspecified covariate functional form. *Statistics in Medicine*, 40(2):299–311.

Nuño, M. M., Gillen, D. L., Dosanjh, K. K., Brook, J., Elashoff, D., Ringman, J. M., and Grill, J. D. (2017). Attitudes toward clinical trials across the Alzheimer's disease spectrum. *Alzheimers Res Ther*, 9(1):81.

Pandeya, N. (2005). Repeated Occurrence of Basal Cell Carcinoma of the Skin and Multifailure Survival Analysis: Follow-up Data from the Nambour Skin Cancer Prevention Trial. *American Journal of Epidemiology*, 161(8):748–754.

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

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

Raman, R., Quiroz, Y. T., Langford, O., Choi, J., Ritchie, M., Baumgartner, M., Rentz, D., Aggarwal, N. T., Aisen, P., Sperling, R., and Grill, J. D. (2021). Disparities by Race and Ethnicity Among Adults Recruited for a Preclinical Alzheimer Disease Trial. *JAMA Network Open*, 4(7):e2114364.

Rivera, C. and Lumley, T. (2016a). Using the Whole Cohort in the Analysis of Countermatched Samples. *Biometrics*, 72(2):382–391.

Rivera, C. L. and Lumley, T. (2016b). Using the entire history in the analysis of nested case cohort samples. *Statistics in Medicine*, 35(18):3213–3228.

Robins, J. M., Hsieh, F., and Newey, W. (1995). Semiparametric Efficient Estimation of a Conditional Density with Missing or Mismeasured Covariates. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(2):409–424.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of Regression Coefficients When Some Regressors are not Always Observed. *Journal of the American Statistical Association*, 89(427):846–866.

Rondeau, V., Mathoulin-Pelissier, S., Jacqmin-Gadda, H., Brouste, V., and Soubeyran, P. (2006). Joint frailty models for recurring events and death using maximum penalized likelihood estimation: application on cancer events. *Biostatistics*, 8(4):708–721.

Salazar, C. R., Hoang, D., Gillen, D. L., and Grill, J. D. (2020). Racial and ethnic differences in older adults' willingness to be contacted about Alzheimer's disease research participation. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 6(1).

Samuelsen, S. O. (1997). A Pseudolikelihood Approach to Analysis of Nested Case-Control Studies. *Biometrika*, 84(2):379–394.

Saunders, A. M., Strittmatter, W. J., Schmechel, D., St. George-Hyslop, P. H., Pericak-Vance, M. A., Joo, S. H., Rosi, B. L., Gusella, J. F., Crapper-MacLachlan, D. R., Alberts, M. J., Hulette, C., Crain, B., Goldgaber, D., and Roses, A. D. (1993). Association of apolipoprotein E allele 4 with late-onset familial and sporadic Alzheimer's disease. *Neurology*, 43(8):1467–1467.

Shamberger, R. C., Guthrie, K. A., Ritchey, M. L., Haase, G. M., Takashima, J., Beckwith, J. B., D'Angio, G. J., Green, D. M., and Breslow, N. E. (1999). Surgery-related factors and local recurrence of Wilms tumor in National Wilms Tumor Study 4. *Annals of Surgery*, 229(2):292–297.

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

Støer, N. C. and Samuelsen, S. O. (2012). Comparison of estimators in nested case–control studies with multiple outcomes. *Lifetime Data Analysis*, 18(3):261–283.

Tao, R., Zeng, D., and Lin, D.-Y. (2020). Optimal Designs of Two-Phase Studies. *Journal of the American Statistical Association*, 115(532):1946–1959.

Thomas, D. C. (1977). Addendum to a paper by FDK Liddel JC McDonald and DC Thomas. *Journal of the Royal Statistical Society. Series A (General)*, 140(4):483–485.

Van Dyck, C. H., Swanson, C. J., Aisen, P., Bateman, R. J., Chen, C., Gee, M., Kanekiyo, M., Li, D., Reyderman, L., Cohen, S., Froelich, L., Katayama, S., Sabbagh, M., Vellas, B., Watson, D., Dhadda, S., Irizarry, M., Kramer, L. D., and Iwatsubo, T. (2023). Lecanemab in Early Alzheimer's Disease. *New England Journal of Medicine*, 388(1):9–21.

Warwick, A. B., Kalapurakal, J. A., Ou, S.-S., Green, D. M., Norkool, P. A., Peterson, S. M., and Breslow, N. E. (2010). Portal hypertension in children with Wilms' tumor: a report from the National Wilms' Tumor Study Group. *International Journal of Radiation Oncology, Biology, Physics*, 77(1):210–216.

Wei, L. J., Lin, D. Y., and Weissfeld, L. (1989). Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions. *Journal of the American Statistical Association*, 84(408):1065–1073.

Wendler, D., Kington, R., Madans, J., Wye, G. V., Christ-Schmidt, H., Pratt, L. A., Brawley, O. W., Gross, C. P., and Emanuel, E. (2005). Are Racial and Ethnic Minorities Less Willing to Participate in Health Research? *PLoS Medicine*, 3(2):e19.

White, H. (1982a). Maximum Likelihood Estimation of Misspecified Models. *Econometrica*, 50(1):1.

White, J. E. (1982b). A TWO STAGE DESIGN FOR THE STUDY OF THE RELATIONSHIP BETWEEN A RARE EXPOSURE AND A RARE DISEASE1. *American Journal of Epidemiology*, 115(1):119–128.

Witbracht, M. G., Bernstein, O. M., Lin, V., Salazar, C. R., Sajjadi, S. A., Hoang, D., Cox, C. G., Gillen, D. L., and Grill, J. D. (2020). Education and Message Framing Increase Willingness to Undergo Research Lumbar Puncture: A Randomized Controlled Trial. *Frontiers in Medicine*, 7:493.

Xu, R. (2000). Estimating average regression effect under non-proportional hazards. *Biostatistics*, 1(4):423–439.

Zhou, H., Chen, J., Rissanen, T. H., Korrick, S. A., Hu, H., Salonen, J. T., and Longnecker, M. P. (2007). Outcome-Dependent Sampling: An Efficient Sampling and Inference Procedure for Studies With a Continuous Outcome. *Epidemiology*, 18(4):461–468.

Appendix A

Supplementary Material for Chapter 3

A.1 Proof of Proposition 3.1

We provide a proof of Proposition 3.1, the consistency of our estimator when conducting weighted sampling on a binary covariate, below. This proof is easily translated to Proposition 3.2 by replacing all V_j^* and p_j^* with V_j^{**} and p_j^{**}, respectively.

Proof. Let T_i, C_i, and $Z_i = [z_{i1}, \ldots, z_{ip}]^T$ denote the true event time, censoring time, and vector of covariates for subject i, $(i = 1, \ldots, n)$, respectively. We define $X_i = \min(T_i, C_i)$ as the observed time and $\delta_i = I(X_i = T_i)$ as the event indicator for subject i. Let $\mathcal{G} = \{(X_i, \delta_i, Z_i), i = 1, \ldots, n\}$, which denotes the complete cohort study information. Let $\mathcal{H} = \{(X_i, \delta_i, z_i), i = 1, \ldots, n\}$, which denotes the skeleton of the cohort and contains all information which is relevant to weighted nested case-control (WNCC) sampling. It follows then that $\mathcal{H} \subseteq \mathcal{G}$. As defined previously, V_{j0}^* is an indicator for whether subject j is ever sampled as a control under the WNCC design, and $V_j^* = \max(V_{j0}^*, \delta_j)$ is the indicator for whether subject j is included in the analysis under the WNCC design. Moreover, p_j^* (defined in Section 3.2.1) represents the probability that subject j is included in the

analysis under the WNCC design conditional on the covariate value used for weighting. Given these definitions, we have that $E\left(\frac{V_i^*}{p_i^*} \;\middle|\; \mathcal{G}\right) = E\left(\frac{V_i^*}{p_i^*} \;\middle|\; \mathcal{H}\right) = 1$. For $r = 0, 1, 2$, we define $S^{(r)}(\beta, t) = \sum_{j \in R_t} Z_j^{\otimes r} \exp\{\beta^T Z_j\}$ and $\tilde{S}^{(r)}(\beta, t) = \sum_{j \in R_t} \frac{V_j^*}{p_j^*} Z_j^{\otimes r} \exp\{\beta^T Z_j\}$. It follows that $E\{\tilde{S}^{(r)}(\beta, t)\} = E\{S^{(r)}(\beta, t)\}$. Since $\frac{1}{n}\{\tilde{S}^{(r)}(\beta, t) - S^{(r)}(\beta, t)\} = \frac{1}{n}\left\{\sum_{i=1}^n \left(\frac{V_i^*}{p_i^*} - 1\right) Y_{(i)}(t)\ Z_i^{\otimes r} \exp\{\beta^T Z_i\}\right\}$ where $Y_{(i)}(t) = I(X_i \geq t)$, then by the weak law of large numbers, $\frac{1}{n}\{\tilde{S}^{(r)}(\beta, t) - S^{(r)}(\beta, t)\} \xrightarrow{p} 0$. The proof of consistency easily follows from the work of Andersen and Gill (1982) using the fact that $\sup_t \frac{1}{n}|\tilde{S}^{(r)}(\beta, t) - S^{(r)}(\beta, t)| \xrightarrow{p} 0$ and that the pseudo-likelihood is a version of the full likelihood with $1/p_i^*$ copies of observation i. □

A.2 Derivation of Variance and Approximation of Limiting Distribution

Below, we provide details for the derivation of the variance of our estimator when conducting weighted sampling on a binary covariate. We also provide a heuristic argument for approximating the limiting distribution of this estimator. This proof is applicable to the estimator for weighted sampling on a continuous covariate by replacing V_j^*, p_j^*, $\widehat{\beta}_B$, I_B, $\mathcal{U}_B$, Σ_B, and Δ_B with V_j^{**}, p_j^{**}, $\widehat{\beta}_C$, I_C, $\mathcal{U}_C$, Σ_C, and Δ_C , respectively.

Proof. Let $\mathcal{W}_j^0$ be the score residual for subject j from the Cox PH model weighted by V_j^*/p_j^* evaluated at the true parameter value and $\mathcal{W}_j = \int \left\{ Z_j - \frac{S^{(1)}(\beta_0, t)}{S^{(0)}(\beta_0, t)} \right\} Y_{(j)}(t) \exp\{\beta_0^T Z_j\} \lambda_0(t) dt$, then it follows that $\mathcal{W}_j - \mathcal{W}_j^0 \xrightarrow{p} 0$. We can write $\frac{1}{\sqrt{n}}\left\{(\mathcal{U}_B - \mathcal{U}) - \sum_{j=1}^n \left(1 - \frac{V_j^*}{p_j^*}\right)\mathcal{W}_j\right\} = \frac{1}{\sqrt{n}} \sum_{j=1}^n \left(1 - \frac{V_j^*}{p_j^*}\right)(\mathcal{W}_j^0 - \mathcal{W}_j)$ where $\mathcal{U}$ is the score function of the classic Cox model and $\mathcal{U}_B$ is the score function of the WNCC estimator (binary). Since $\mathcal{W}_j - \mathcal{W}_j^0 \xrightarrow{p} 0$, then $\frac{1}{\sqrt{n}}\left\{(\mathcal{U}_B - \mathcal{U}) - \sum_{j=1}^n \left(1 - \frac{V_j^*}{p_j^*}\right)\mathcal{W}_j\right\} \xrightarrow{p} 0$ by Slutsky's theorem (using the previous statement). It follows that $\frac{1}{\sqrt{n}}\mathcal{U}_B$ and $\frac{1}{\sqrt{n}}\left\{\mathcal{U} + \sum_{j=1}^n \left(1 - \frac{V_j^*}{p_j^*}\right)\mathcal{W}_j\right\}$ have the same limiting distribution. We know that $E\{\mathcal{U}\} = 0$ and $E\{\sum_{j=1}^n (1 - \frac{V_j^*}{p_j^*})\mathcal{W}_j\} = 0$, and the two terms are uncorrelated.

Additionally, $Cov\{\mathcal{U}\} = \Sigma_n$ and $Cov\{\sum_{j=1}^{n}(1 - \frac{V_j^*}{p_j^*})\mathcal{W}_j\} = \Delta_n$, where $\Sigma_n = E\{I(\beta_0)\}$ and $\Delta_n = Cov\left\{\sum_{j=1}^{n}\left(1 - \frac{V_j}{p_j}\right)\mathcal{W}_j\right\}$. Then $n^{-1}\Sigma_n \to \Sigma$, for some positive definite matrix Σ and $n^{-1}I_B \to \Sigma$, and $n^{-1}\Delta_n \to \Delta$, for some positive semidefinite matrix Δ. If we assume that $n^{-1/2}\sum_{j=1}^{n}(1 - \frac{V_j^*}{p_j^*})\mathcal{W}_j$ converges in distribution to a normal distribution with expectation 0 and covariance matrix Δ, then we can show that $n^{-1/2}\mathcal{U}_B$ is asymptotically normal with expectation 0 and covariance matrix $\Sigma + \Delta$. Then by a first-order Taylor expansion of $\mathcal{U}_B$ around β_0, we have that $\sqrt{n}(\widehat{\beta}_B - \beta_0) \dot{\sim} N(0, \Sigma^{-1} + \Sigma^{-1}\Delta\Sigma^{-1})$. Estimators of the matrices Σ and Δ are given by Σ_B, the observed Fisher's information from our proposed estimator, and Δ_B, an empirical estimate of the analytic covariance of $\sum_{j=1}^{n}(1 - \frac{V_j^*}{p_j^*})\mathcal{W}_j$, respectively. □

A.3 Additional Simulation Results

Table A.1 displays results for a primary analysis under no effect of the predictor of interest. The proposed estimator performs well for estimating β_1, with nominal 95% coverage probabilities indicating a controlled Type I error. Moreover, the proposed estimator accounts for bias induced from weighted sampling when estimating β_2.

		$\beta_1 = 0.000$						$\beta_2 = 0.336$					
$\lambda_0 = 0.06$	Unique N	$\widehat{\beta}_1$ (Std. Bias)	SE 1	SE 2	Emp. SE	CP 1	CP 2	$\widehat{\beta}_2$ (% Bias)	SE 1	SE 2	Emp. SE	CP 1	CP 2
FC (89% cens.)	1000	0.000 (5.7e-3)	0.098	0.097	0.100	0.955	0.949	0.319 (-5.31)	0.227	0.227	0.229	0.949	0.951
Samuelsen													
m = 1	195	0.003 (0.031)	0.143	0.144	0.149	0.946	0.950	-0.901 (-367.72)	0.299	0.300	0.302	0.010	0.012
m = 2	269	-0.003 (-0.032)	0.121	0.121	0.129	0.937	0.935	-0.792 (-335.25)	0.264	0.264	0.267	0.006	0.006
m = 3	332	-0.003 (-0.041)	0.112	0.112	0.113	0.947	0.946	-0.691 (-305.22)	0.251	0.251	0.258	0.010	0.010
m = 4	387	0.001 (0.010)	0.108	0.108	0.111	0.944	0.945	-0.603 (-279.09)	0.245	0.245	0.250	0.014	0.016
Proposed													
m = 1	195	0.003 (0.036)	0.145	0.147	0.152	0.943	0.951	0.304 (-9.74)	0.293	0.295	0.295	0.953	0.957
m = 2	269	-0.003 (-0.031)	0.122	0.123	0.131	0.937	0.936	0.310 (-7.92)	0.258	0.259	0.263	0.946	0.946
m = 3	332	-0.004 (-0.042)	0.114	0.114	0.114	0.949	0.947	0.316 (-6.06)	0.246	0.247	0.253	0.943	0.946
m = 4	387	0.001 (0.013)	0.109	0.109	0.112	0.944	0.945	0.317 (-5.69)	0.240	0.241	0.245	0.940	0.943

Table A.1: Simulation results comparing the performance of the FC Cox PH estimator, the Samuelsen estimator under the WNCC design, and our proposed estimator under the WNCC design for a primary analysis under a null effect of the predictor of interest. We note that standardized bias was used to assess performance for estimating null effects.

Appendix B

Supplementary Material for Chapter 4

B.1 Proof of Proposition 4.2

We provide a proof of Proposition 4.2, the consistency of our proposed event-specific sampling estimator for fitting M1 and M2 below. The notation used in the following proof is applicable to M1, but we note that this proof is applicable to M2 by replacing $S^{(r)}(\beta,t)$ with $S_k^{(r)}(\beta,t) = \sum_{l=1}^{n}\sum_{j=1}^{K_l} I(j = k)Y_{lj}(t)Z_{lj}^{\otimes r}\exp\{\beta^T Z_{lj}\}$, $\tilde{S}^{(r)}(\beta,t)$ with $\tilde{S}_k^{(r)}(\beta,t) = \sum_{l=1}^{n}\sum_{j=1}^{K_l} I(j = k)\frac{V_{lj}}{p_{lj}}Y_{lj}(t)Z_{lj}^{\otimes r}\exp\{\beta^T Z_{lj}\}$ for $r = 0, 1$, $\ell^{(1)}(\beta)$ with $\ell^{(2)}(\beta)$, $\tilde{\ell}^{(1)}(\beta)$ with $\tilde{\ell}^{(2)}(\beta)$, $\widehat{\beta}^{(1)}$ with $\widehat{\beta}^{(2)}$, and $\beta^{(1)}$ with $\beta^{(2)}$. There is also an added constraint that $K_i/n \to c_i$, where $0 < c_i < 1$, for $i = 1, \ldots, n$ in addition to $n \to \infty$, for convergence.

Proof. We define T_{ik}, C_{ik}, and $Z_{ik} = [z_{ik1}, \ldots, z_{ikp}]^T$ to denote the true event time, censoring time, and vector of covariates of event type k, $(k = 1, \ldots, K_i)$, for subject i, $(i = 1, \ldots, n)$, respectively. We define $X_{ik} = \min(T_{ik}, C_{ik})$ as the observed time and $\delta_{ik} = I(X_{ik} = T_{ik})$ as the event indicator of event type k for subject i. We further define $G_{ik} = X_{ik} - X_{i,k-1}$ where $X_{i0} = 0$. Let $\mathcal{G} = \{(X_{ik}, \delta_{ik}, Z_{ik}), i = 1, \ldots, n; k = 1, \ldots, K_i\}$, which denotes the complete cohort study information. Let $\mathcal{H} = \{(X_{ik}, \delta_{ik}), i = 1, \ldots, n; k = 1, \ldots, K_i\}$, which denotes

the skeleton of the cohort and contains all information which is relevant to the event-specific sampling. It follows then that $\mathcal{H} \subseteq \mathcal{G}$. As defined in Section 4.2.3, V_{lj} is the indicator for whether observation lj was selected into the sample under event-specific sampling, and p_{lj} is the probability that observation lj is included in the analysis under event-specific sampling. It follows that $E\left(\frac{V_{lj}}{p_{lj}} \;\middle|\; \mathcal{G}\right) = E\left(\frac{V_{lj}}{p_{lj}} \;\middle|\; \mathcal{H}\right) = 1$.

We further define $S^{(r)}(\beta, t) = \sum_{l=1}^{n} \sum_{j=1}^{K_l} Y_{lj}(t) Z_{lj}^{\otimes r} \exp\{\beta^T Z_{lj}\}$ and $\tilde{S}^{(r)}(\beta, t) = \sum_{l=1}^{n} \sum_{j=1}^{K_l} \frac{V_{lj}}{p_{lj}} Y_{lj}(t) Z_{lj}^{\otimes r} \exp\{\beta^T Z_{lj}\}$ for $r = 0, 1$. Then, it follows that $E\{\tilde{S}^{(r)}(\beta, t)\} = E\{S^{(r)}(\beta, t)\}$. By the weak law of large numbers, we have that $\frac{1}{n}\{\tilde{S}^{(r)}(\beta, t) - S^{(r)}(\beta, t)\} \xrightarrow{p} 0$ as $n \to \infty$, since $\frac{1}{n}\{\tilde{S}^{(r)}(\beta, t) - S^{(r)}(\beta, t)\} = \frac{1}{n}\left\{\sum_{l=1}^{n} \sum_{j=1}^{K_l} \left(\frac{V_{lj}}{p_{lj}} - 1\right) Y_{lj}(t) Z_{lj}^{\otimes r} \exp\{\beta^T Z_{lj}\}\right\}$, and we will further assume that $\sup_t \frac{1}{n}|\tilde{S}^{(r)}(\beta, t) - S^{(r)}(\beta, t)| \xrightarrow{p} 0$. If $\ell^{(1)}(\beta)$ and $\tilde{\ell}^{(1)}(\beta)$ are the log-likelihoods for M1 in the full cohort and event-specific sample respectively, it also follows that $\frac{1}{n}\{\tilde{\ell}^{(1)}(\beta) - \ell^{(1)}(\beta)\} = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K_i} \log\left\{\frac{S^{(0)}(\beta, G_{ik})}{\tilde{S}^{(0)}(\beta, G_{ik})}\right\} \delta_{ik} \xrightarrow{p} 0$. Using the fact that $\tilde{\ell}^{(1)}(\beta)$ is a pseudo-version of $\ell^{(1)}(\beta)$ with V_{lj}/p_{lj} copies of observations lj, the proof of consistency follows from the work of Andersen and Gill (1982), stating that if $\widehat{\beta}^{(1)}$ is the unique maximizer of $\tilde{\ell}^{(1)}(\beta)$ and $\beta^{(1)}$ is the unique maximizer of $\ell^{(1)}(\beta)$, then $\widehat{\beta}^{(1)} \xrightarrow{p} \beta^{(1)}$. □

B.2 Derivation of Variance and Approximation of the Limiting Distribution

We provide details of the derivation of the variance estimator along with an argument for asymptotic normality of our proposed event-specific sampling estimator for fitting M1. These results can be extended to the M2 case by replacing $S^{(r)}(\beta, t)$ with $S_k^{(r)}(\beta, t) = \sum_{l=1}^{n} \sum_{j=1}^{K_l} I(j = k) Y_{lj}(t) Z_{lj}^{\otimes r} \exp\{\beta^T Z_{lj}\}$, $\tilde{S}^{(r)}(\beta, t)$ with $\tilde{S}_k^{(r)}(\beta, t) = \sum_{l=1}^{n} \sum_{j=1}^{K_l} I(j = k) \frac{V_{lj}}{p_{lj}} Y_{lj}(t) Z_{lj}^{\otimes r} \exp\{\beta^T Z_{lj}\}$ for $r = 0, 1$, $\tilde{U}^{(1)}(\beta)$ with $\tilde{U}^{(2)}(\beta)$, $W_{ik}^{(1)}$ with $W_{ik}^{(2)}$, and $\widehat{\beta}^{(1)}$ with $\widehat{\beta}^{(2)}$. Similarly to the previous section, there is also the added condition that $K_i/n \to c_i$, where $0 < c_i < 1$, for $i = 1, \ldots, n$ in addition to $n \to \infty$.

Proof. We start by defining $\tilde{U}^{(1)}(\beta) = \sum_{i=1}^{n}\sum_{k=1}^{K_i}\delta_{ik}\left[Z_{ik} - \frac{\tilde{S}^{(1)}(\beta,G_{ik})}{\tilde{S}^{(0)}(\beta,G_{ik})}\right]$. By doing a Taylor expansion about $\tilde{S}^{(0)}(\beta,t) = S^{(0)}(\beta,t)$ and $\tilde{S}^{(1)}(\beta,t) = S^{(1)}(\beta,t)$, we have that $\tilde{U}^{(1)}(\beta)$ is asymptotically equivalent to $U^*(\beta) \equiv \sum_{i=1}^{n}\sum_{k=1}^{K_i}\delta_{ik}\left[Z_{ik} - \frac{S^{(1)}(\beta,G_{ik})}{S^{(0)}(\beta,G_{ik})} + \frac{S^{(1)}(\beta,G_{ik})\tilde{S}^{(0)}(\beta,G_{ik})}{S^{(0)}(\beta,G_{ik})^2} - \frac{\tilde{S}^{(1)}(\beta,G_{ik})}{S^{(0)}(\beta,G_{ik})}\right]$. Here, $U^*(\beta)$ can be consistently estimated by substituting $\tilde{S}^{(0)}(\beta,t) = S^{(0)}(\beta,t)$ and $\tilde{S}^{(1)}(\beta,t) = S^{(1)}(\beta,t)$. That is, we can consistently estimate $U^*(\beta)$ by using $\sum_{i=1}^{n}\sum_{k=1}^{K} W_{ik}^{(1)}(\beta)$ where $W_{ik}^{(1)}(\beta) = \delta_{ik}\frac{V_{ik}}{p_{ik}}\left[Z_{ik} - \frac{\tilde{S}^{(1)}(\beta,G_{ik})}{\tilde{S}^{(0)}(\beta,G_{ik})}\right] - \frac{V_{ik}}{p_{ik}}\sum_{j=1}^{n}\sum_{l=1}^{K_j}\delta_{jl}\frac{V_{jl}}{p_{jl}}\left[\frac{Y_{ik}(G_{jl})\exp\{\beta^T Z_{ik}\}}{\tilde{S}^{(0)}(\beta,G_{jl})}\right] \times \left\{Z_{ik} - \frac{\tilde{S}^{(1)}(\beta,G_{jl})}{\tilde{S}^{(0)}(\beta,G_{jl})}\right\}$, the score residual for observation ik, weighted by V_{ik}/p_{ik}. It follows that we can estimate the covariance of $\tilde{U}^{(1)}(\beta)$ by $V = \sum_{i=1}^{n}\sum_{k=1}^{K_i}\sum_{l=1}^{K_i} W_{ik}^{(1)}(\widehat{\beta}^{(1)})W_{il}(\widehat{\beta}^{(1)})^T$ and then the estimated variance of our proposed estimator is $\widehat{\Gamma} = \widehat{\Sigma}^{-1}V\widehat{\Sigma}^{-1}$ where $\widehat{\Sigma} = I(\widehat{\beta}) = -\frac{\partial\tilde{U}^{(1)}(\beta)}{\partial\beta}|_{\beta=\widehat{\beta}^{(1)}}$. A conjecture of asymptotic normality follows closely from Samuelsen (1997). □