A weighted casebase framework for predicting risk in survival data

Karina Kwan

Department of Epidemiology, Biostatistics & Occupational Health

McGill University

Montréal, Québec

April, 2023

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of

Master of Science

©Karina Kwan, 2023

Abstract

Case-cohort studies are attractive for studying rare diseases where obtaining additional expensive or hard-to-access data, such as genomic sequencing, from a subset of participants is infeasible for the entire study cohort. In analyzing such studies, individual data points must be appropriately weighted to account for the biased case/control sampling. The Cox proportional hazards model is a popular semi-parametric method for analyzing survival data that provides step function risk estimates. A parametric alternative is the casebase framework, which uses finite sampling of person-moments together with logistic regression to estimate fully parametric hazard functions and smooth-in-time absolute risk functions. Unlike the Cox model, where well-tested methods exist to adjust for complex sampling designs, the casebase framework-based methods have not yet implemented weighted methods. This thesis proposes a weighted casebase framework that provides unbiased coefficient estimates and robust standard error estimates. A simulation study compares the performance of weighted Cox and casebase models. The proposed weighted analytic framework is then applied to model how cell-free DNA methylation (data obtained with the cfMeDIP-seq technology) affects risk of breast cancer in a (case-cohort) subset of individuals in the Ontario Health Study (OHS). The weighted framework performs similarly to weighted Cox models, and both are sensitive to covariate distributions and the size of the sampling fraction.

Abrégé

Les enquêtes cas-cohorte sont intéressantes pour étudier les maladies rares lorsque l'obtention de données supplémentaires coûteuses ou difficiles d'accès, telles que le séquençage génomique, auprès d'un sous-ensemble de participants n'est pas réalisable pour l'ensemble de la cohorte étudiée. Lors de l'analyse de ces études, les points de données individuels doivent être pondérés de manière appropriée pour tenir compte de l'échantillonnage biaisé des cas et des témoins. Le modèle à risques proportionnels de Cox est une méthode semiparamétrique populaire pour l'analyse des données de survie qui fournit des estimations de risque en fonction en escalier. Une alternative paramétrique est le cadre de la base de cas, qui utilise un échantillonnage fini des personnes-moments ainsi qu'une régression logistique pour estimer des fonctions de hasard entièrement paramétriques et des fonctions de risque absolu lisses par rapport au temps. Contrairement au modèle de Cox, pour lequel il existe des méthodes éprouvées permettant d'ajuster les plans d'échantillonnage complexes, les méthodes du cadre de la base de cas n'ont pas encore mis en œuvre de méthodes pondérées. Cette thèse propose un cadre de base de cas pondéré qui fournit des estimations de coefficient non biaisées et des estimations d'erreur type robustes. Une étude de simulation compare les performances des modèles pondérés de Cox et de la base de cas. Le cadre analytique pondéré proposé est ensuite appliqué pour modeler comment la méthylation de l'ADN du plasma sanguin (données obtenues avec la technologie cfMeDIP-seq) affecte le risque de cancer du sein dans un sous-ensemble d'individus (cas-cohorte) de l'Étude sur la santé Ontario (ÉSO). Le cadre pondéré fonctionne de manière

similaire aux modèles de Cox pondérés, et les deux sont sensibles aux distributions des covariables et au choix de la fraction d'échantillonnage.

Acknowledgements

I would like to thank my supervisors Prof. Celia Greenwood and Prof. David Soave for their guidance and mentorship at every step of my master's. I would also like to thank my committee member Dr. Sahir Bhatnagar for his knowledge in casebase and his feedback during the early stages.

I am grateful for my fellow classmates in the Biostatistics program as we embarked on this journey together. I am also grateful for the members of the Greenwood lab who provided me with feedback throughout my project. Thank you to Dr. Philip Awadalla, Nick Cheng, and the OICR for providing access to the OHS data.

Finally, this thesis would not have been possible without my family. Their support and endless patience kept me going.

Contribution

Karina Kwan performed all the data analysis, carried out the simulations, and wrote all chapters of this thesis. Prof. Celia Greenwood and Prof. David Soave devised the thesis topic, proposed the methods, and provided feedback on this thesis.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

5-mC 5-Methylcytosine

bootrep Bootstrap replicate

BootSE Bootstrapped standard error

BS Brier score

cb casebase

cfDNA Cell-free DNA

cfMeDIP-seq Cell-free methylated DNA immunoprecipitation sequencing

ctDNA Circulating tumor-specific cell-free DNA

coef Coefficient

Cover Coverage

COVID-19 Coronavirus disease 2019

DNA Deoxyribonucleic acid

EmpSE Empirical standard error

KM Kaplan-Meier

ModSE Model-based standard error

OICR Ontario Institute for Cancer Research

OHS Ontario Health Study

PC Principal componennts

sim Simulation

upp Upper

Var Variance

Chapter 1

Introduction

Breast cancer continues to be the most diagnosed and second-most fatal cancer among Canadian women (Brenner et al., 2022). Current screening guidelines recommend mammograms starting at different ages and at different frequencies across provinces (Canadian Partnership Against Cancer, 2018). Mammograms are used for baseline screening with additional testing should suspicious abnormalities be found. However, cancer detection is complicated by breast density. Dense breast tissues appear white on mammograms, similar to tumors, potentially leading to missed tumors. At the same time, mammograms have a high false positive rate (Canadian Partnership Against Cancer, 2020), leading to unnecessary secondary testing that can be physically and mentally taxing on patients. Overdiagnosis, where a patient is diagnosed with cancer but it would not have resulted in any symptoms or death, leads to unnecessary interventions and is also a concern (Canadian Partnership Against Cancer, 2020).

With developments in sequencing technologies, the use of genomics in providing earlier cancer detection is attractive. Improvements in existing cancer screening programs could also be made by including genomic information from individuals. Particularly, liquid biopsies, where liquid biological samples like blood are analyzed for tumor derivatives, are less invasive and could have great clinical potential (Poulet, Massias, and Taly, 2019). Unfortunately, it remains expensive to perform sequencing on all participants of

large study cohorts, such as the Ontario Health Study (Kirsh et al., 2022), and in the case of studying rare diseases like cancer, it is inefficient. The case-cohort design proposed by Prentice, 1986 provides the benefits of a cohort study but at a reduced cost by only sequencing breast cancer cases and a random subset of the non-cases. With such a design, it is important to appropriately weight cases and non-cases due to the over-representation of cases to avoid erroneous estimates and conclusions.

The Cox proportional hazards model (Cox, 1972) is a popular semi-parametric model in survival analysis. When absolute quantities are of interest, such as absolute risk, this model requires an additional estimation step and produces step function estimates. The casebase framework (Bhatnagar et al., 2022) is a fully parametric method that uses logistic regression and a finite sampling of person-moments to obtain smooth-in-time absolute risk estimates. Unlike the Cox model, the existing casebase framework does not have weighted and robust methods for biased sampling study designs. Chapter 3 of this thesis proposes a weighted casebase framework to fill this gap. Chapter 4 compares the proposed framework to weighted Cox in a simulation study and Chapter 5 illustrates the use of the proposed framework on data from the Ontario Health Study.

Chapter 2

Literature review

This chapter provides background material regarding survival analysis, the case-cohort study design, the existing casebase framework, DNA methylation, high-dimensional data, and measures of predictive model performance.

2.1 Survival analysis

Kleinbaum and Klein, 2012b describes survival analysis as data analysis concerned with time, as a variable, from the start of follow-up until an outcome of interest occurs. Survival time refers to the time period that an individual is free from an event, which may include breast cancer diagnosis, disease recurrence, and death.

Although death is an event that will occur with certainty, it may not happen during the study period. Censoring occurs when the exact survival time is unknown. In particular, right censoring occurs when the exact event time is unknown. An individual in a study may have dropped out of the study, changed physicians and been lost to follow-up, or did not have the event by the end of the study. In analysis, individuals can be coded as 0 for censored and 1 for event occurred. The risk set $R(t)$ at time t includes individuals that have not yet experienced the event up to time t.

Let T be the random variable for survival time. The survival function, the probability that an individual survives beyond time t, is defined as $S(t) = P(T > t)$ and the hazard function, the instantaneous rate of having the event in the interval $[t, t + \Delta t)$ given that an individual has survived until time t, is defined as $h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$ (Kleinbaum and Klein, 2012b). An important quantity in the medical field is the absolute risk, the probability that the event occurs in the interval $[0, t]$ (Gail, 2005). It is defined as $P(T \leq t) = \int_0^t h(u)S(u)du$ (Pfeiffer and Gail, 2017). In particular, risk charts can provide physicians a quick assessment of a patient's risk of developing breast cancer based on their age and risk factors (Woloshin, Schwartz, and Welch, 2008).

A fully non-parametric method to estimate the survival probability $S(t)$ is the Kaplan-Meier method. Ordering the k unique event times $t_{(1)} < t_{(2)} < \ldots < t_{(k)}$, the Kaplan-Meier estimate $\hat{S}(t) = \prod_{j=1}^{t} \frac{n_j - d_j}{n_j}$ where n_j is the number of individuals at risk between $(t_{(j-1)}, t_{(j)}]$, d_j is the number of events between $(t_{(j-1)}, t_{(j)}]$, and $t_{(j)} \leq t$ (Kaplan and Meier, 1958). Computing the Kaplan-Meier estimate at all event times and plotting these estimates across the event times results in a step function survival curve that provides a graphical view of how survival probabilities change over follow-up time. The Kaplan-Meier estimates can also be computed for different groups, allowing for visual comparison of the effect of different exposures on survival probabilities.

A popular semi-parametric model for survival analysis that incorporates covariates is the Cox proportional hazards model $h(t) = h_0(t) \exp(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients, $\mathbf{X} = (X_1, \ldots, X_p)$ is a $n \times p$ matrix of covariates, and $h_0(t)$ is the baseline hazard, the hazard function when $\mathbf{X} = \mathbf{0}$ (Cox, 1972). The hazard function can be separated into two components: $h_0(t)$ that only depends on time and $\exp(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta})$ that only involves the covariates.

The popularity of the model can be partly attributed to the unspecified nature of $h_0(t)$. The partial likelihood, conditioned on the failure times, does not involve $h_0(t)$ (Cox, 1972). The hazard ratio, defined as the ratio of two hazard functions, is often reported as a measure of intervention effect (Higgins, Li, and Deeks, 2022). The estimated hazard ratio

$\widehat{HR} = \frac{\hat{h}_0(t)\exp(\mathbf{X}^{*\mathrm{T}}\hat{\boldsymbol{\beta}})}{\hat{h}_0(t)\exp(\mathbf{X}^{\mathrm{T}}\hat{\boldsymbol{\beta}})} = \exp((\mathbf{X}^* - \mathbf{X})^{\mathrm{T}}\hat{\boldsymbol{\beta}})$ only requires the covariate information and the coefficient estimates $\hat{\boldsymbol{\beta}}$. The model assumes the hazard ratio is constant with respect to time such that the hazard functions of two individuals differ by a multiplicative constant, the hazard ratio (Kleinbaum and Klein, 2012d). If the assumption is not met, stratified Cox models with stratum-specific hazard functions or extended Cox models allowing for time-dependent covariates can be used (Kleinbaum and Klein, 2012e; Kleinbaum and Klein, 2012a). Estimation of the baseline hazard is not needed to estimate the hazard ratio, the hazard function, or the survival function (Kleinbaum and Klein, 2012d). However, if absolute risk estimates are of interest, the $h_0(t)$ will need to be separately estimated, such as with the Breslow estimator $\hat{h}_0(t) = \frac{d_j}{\sum_{l \in R(t)} \exp(\mathbf{X}_l^{\mathrm{T}}\boldsymbol{\beta})}$, where $R(t)$ is the risk set at time t (Breslow, 1972).

Parametric survival models may be desirable if the proportional hazards assumption is not met or to obtain smooth absolute risk estimates. Examples include the exponential model with hazard function $h(t) = \lambda$ and the Weibull model with hazard function $h(t) = \lambda p t^{p-1}$ (Kleinbaum and Klein, 2012c). The casebase framework, presented in section 2.3, allows for fitting of such parametric models.

2.2 Case-cohort study design

The case-cohort study design is a prospective study design proposed in Prentice, 1986 as an alternative to cohort and nested case-control studies. In cohort studies, individuals with a shared characteristic, such as birth in the same period or the same occupation, who have not yet developed the outcome of interest are recruited, obtain baseline measurements of exposures and covariates, and are followed over time (Barrett and Noble, 2019; Ernster, 1994). The cohort may contain individuals with different exposures measured prior to event, allowing for testing of causal relationships between the outcome of interest and the exposures (Ernster, 1994). Since measurements of exposures and other covariates are obtained for the entire cohort, the same cohort can be used to study mul-

tiple outcomes. However, obtaining data for the entire cohort can be very expensive in the context of methylation data (described in Section 2.4) and is inefficient when the outcome of interest is rare since most of the measurements are from non-cases. Nested case-control studies are more cost-effective, requiring covariate information on all cases from the cohort and selected non-cases from the risk set that are matched controls for the cases (O'Brien, Lawrence, and Keil, 2022). However, different sets of controls and cases need to be selected to study a different outcome (O'Brien, Lawrence, and Keil, 2022). In contrast to cohort and nested case-control studies, case-cohort studies only require exposure and covariate information from a random sample of the cohort, the subcohort, and all incident cases (Prentice, 1986). These studies have the benefit of being less expensive than cohort studies and the relative ease of studying multiple outcomes since the same subcohort can be reused unlike nested case-control studies. Although similar designs were proposed earlier by Kupper, McMichael, and Spirtas, 1975; O. Miettinen, 1982, Prentice, 1986 proposed a design and methods that extend beyond binary outcomes and binary covariates.

An important consideration when analyzing case-cohort data is the use of sampling weights. In such a design, cases are over-represented. In particular, when the outcome of interest is diagnosis of a rare disease, cases might make up almost half of the study population but the prevalence might be 0.1% in the general population. Figure 2.1 shows an example of a case-cohort study with a full cohort of size $N = 5,000$, a subcohort of size $n_{\text{s}} = 500$, and $n_{\text{c}} = 200$ cases. In this example, cases make up 4% of the full cohort but make up almost 30% of the case-cohort subset. Thus, care must be taken to adjust for the over-representation of the cases by including weights. Prentice, 1986 used weight at time t $w_i(t) = 1$ for cases or individuals in the subcohort and $w_i(t) = 0)$ otherwise. Barlow, 1994 used weight $w_i(t) = 1$ for cases, $w_i(t) = (N - n_{\text{c}}(t))/(n_{\text{s}} - n_{\text{c}}(t))$ for non-cases in the subcohort and $n_{\text{c}}(t)$ the number of cases at time t, and $w_i(t) = 0$ for non-cases outside the subcohort. If sampling weights are not included in the analysis, the model coefficients and their associated standard errors will be incorrect since the sampling probabilities are

related to the outcome of interest (Lumley, 2011; Therneau and Grambsch, 2000; Lavallée and Beaumont, 2015). However, including weights comes with the trade-off of increasing the variance of the coefficient estimates (Skinner and Mason, 2012).

Figure 2.1: Schema of a case-cohort study
Adapted from Kulathinal et al., 2007.

2.3 The casebase framework

The casebase framework described in Bhatnagar et al., 2022 is an alternative to the Cox proportional hazards model. The framework uses finite sampling of person-moments combined with logistic regression to provide fully parametric survival and hazard functions, allowing for smooth-in-time risk function estimation.

The concepts of person-time and person-moments underlie the methods of the framework. Person-time is defined as the length of time an individual was observed in a study while being at risk of developing the outcome of interest (Porta, 2016). The total person-time of a study, often given in the unit person-years, is the sum of follow-up times of all

individuals in a study. A person-moment is an individual's covariate profile at a particular instant in time (Hanley and O. S. Miettinen, 2009). The casebase framework uses two kinds of person-moments, termed the case series and the base series. The case series consists of all person-moments at which an event occurred. The base series are a finite sample of the infinitely many person-moments that make up the total person-time in a study (Hanley and O. S. Miettinen, 2009). Several base series sampling mechanisms are described in Hanley and O. S. Miettinen, 2009. The two-step sampling described as follows is the one used in this thesis and in Bhatnagar et al., 2022 and follows the notation of Hanley and O. S. Miettinen, 2009. In order to sample b person-moments out of the total person-time B, the first step randomly samples b individuals out of the total number of individuals in the study n using a multinomial distribution. The probability of selecting individual j is $\pi_j = t_j/B$, where t_j is the length of follow-up for individual j. Then, b moments are selected uniformly from the b individuals' follow-up time. That is, for individual j, the moment associated with their covariate information is not their length of follow-up time but selected uniformly from $\mathcal{U}(0, t_j)$. In this way, the same individual can be included in the base series more than once but with a different moment used each time and cases can be included in the base series since the time of event is not used as the moment. The size of the base series is defined relative to the size of the case series c. That is, $b = \text{ratio} \times c$, where ratio is a positive integer. Hanley and O. S. Miettinen, 2009 showed that the $b = 100c$ is a large enough base series that the variance of coefficient estimates is limited by the number of cases, and not the size of the sampling.

The derivation of the logistic regression form for the casebase model is shown in Saarela and Arjas, 2015; Saarela, 2016 and repeated here. Let $Z_i(t) \in \{0, 1\}$ be the exposure status of individual i, $\mathbf{X}_{i\cdot}$ be a $1 \times p$ vector of covariates, and $Y_i(t) \equiv \mathbf{1}_{C_i \geq t}$, where C_i is individual i's censoring time, be an indicator for individual i still being in the risk set at time t. Then let $N_i(t) \in \{0, 1, 2, \ldots\}$ be a counting process for events where $dN_i(t) = 1$ indicates an event, $R_i(t) \in \{0, 1, 2, \ldots\}$ be a non-homogeneous Poisson process where $dR_i(t) = 1$ indicates inclusion in the base series, and $Q_i(t) = N_i(t) + R_i(t)$ be a counting

process of case and base series person-moments for individual i. $\mathcal{F}_{it^-} \equiv \{N_i(u), Y_i(u) : 0 \leq u < t, Z_i(u) : 0 \leq u < t; \mathbf{X}_{i\cdot}, \mathcal{N}_{it^-} \equiv \{N_i(u) : 0 \leq u < t\}$, and $\mathcal{Z}_{it^-} \equiv \{Z_i(u) : 0 \leq u < t\}$ are the observed, observed outcome event, and observed exposure process histories, respectively. Then let $\tilde{\mathcal{F}}_{it^-}$, $\tilde{\mathcal{N}}_i(t)$, and $\tilde{\mathcal{R}}_i(t)$ be the latent versions. The intensity function $h_i(t) \equiv \lim_{\Delta t \to 0} P(\Delta\tilde{\mathcal{N}}_i(t) = 1|\mathcal{F}_{it^-})/\Delta t$ for $\tilde{\mathcal{N}}_i(t)$, which is of interest, can also be denoted as $h_i(t)dt = \mathbb{E}[d\tilde{\mathcal{N}}_i(t)|\tilde{\mathcal{F}}_{it^-}]$. Then the observed outcome process, where censoring may occur, is $\mathbb{E}[d\mathcal{N}_i(t)|\mathcal{F}_{it^-}] = Y_i(t)h_i(t)dt$. Similarly, $\tilde{\mathcal{R}}_i(t)$ also has intensity function $\rho_i(t) \equiv \lim_{\Delta t \to 0} P(\Delta\tilde{\mathcal{R}}_i(t) = 1|\mathcal{Z}_{it}; \mathbf{X}_{i\cdot})/\Delta t$.

Parameterizing the intensity function $h_i(t;\theta)$ in terms of θ, the likelihood function to be maximized is

$$L_0(\theta) = \prod_{i=1}^{n} \exp\left\{-\int_0^\tau Y_i(t)h_i(t;\theta)df\right\} \prod_{i=1}^{n} \prod_{t\in[0,\tau)} h_i(t;\tau)^{dN_i t(t)},$$

where $\prod_{t\in[0,\tau)}$ is a product integral from 0 to τ. However, conditioning on the sampled person-moments, a quasi-likelihood of the form

$$P(dN_i(t)|dQ_i(t) = 1, \mathcal{F}_{it^-}) \overset{\theta}{\propto} \left(\frac{h_i(t;\theta)^{dN_i(t)}}{h_i(t;\theta) + \rho_i(t)}\right)$$

can be obtained. This results in the partial likelihood

$$L(\theta) = \prod_{i=1}^{n} \prod_{t\in[0,\tau)} \left(\frac{h_i(t;\theta)^{dN_i(t)}}{h_i(t;\theta) + \rho_i(t)}\right)^{dQ_i(t)}.$$

Using $h_i(t;\theta)$ with a logarithmic link function, the partial likelihood has the form of a logistic likelihood with offset $\log(1/\rho_i(t))$. Using the base series sampling mechanism described earlier, the appropriate offset is $\log(B/b)$.

Unlike the Cox proportional hazards model, the casebase framework does not have weighted methods or robust standard error estimation implemented. This motivates the proposed weighted framework described in Chapter 3, with a simulation study compar-

ing the weighted framework to a weighted Cox model in Chapter 4, and an application to a real dataset in Chapter 5.

2.4 DNA methylation

DNA methylation is an epigenetic modification of the DNA molecule and in animals, mainly consists of an addition of a methyl group to cytosine at CpG dinucleotides through DNA methyltransferase (Singal and Ginder, 1999). Regions rich in CpGs, termed CpG islands, tend to be unmethylated and many are found in promoter regions (Singal and Ginder, 1999). Methylation plays an important role in development, gene regulation, and evolution. In human evolution, dinucleotide CpGs appear at a lower frequency than expected due to methylation of CpGs, which can lead to conversion of methylated cytosine to thymine (Singal and Ginder, 1999). When gene promoters are methylated, this acts as a gene silencer by preventing transcription (Singal and Ginder, 1999). Hypotheses for the mechanism underlying repressed transcription include physically blocking binding of transcription factors to recognition sites in promoters, directing binding of transcription repressors to methylated DNA, and formation of a complex that prevents transcription (Singal and Ginder, 1999).

In cancer, widespread changes in methylation occur. There is a general depletion of methylation in regions that are normally methylated, including oncogenes (Singal and Ginder, 1999). Oncogenes are mutated forms of genes that once mutated, promote tumor formation by promoting cell proliferation or preventing apoptosis (Croce, 2008). Due to the loss of methylation, cells with this mutation are then able to grow rapidly and uncontrollably. While there is an overall decrease in methylation, certain CpGs undergo hypermethylation. In particular, there is increased methylation in CpG islands of promoter regions of tumor suppressor genes (Croce, 2008). More evidence of hypermethylation is the increase in DNA methyltransferase activity, which is responsible for *de novo* DNA methylation (Singal and Ginder, 1999). However, tumor suppressor genes are only

a subset of *de novo* methylated genes. In fact, most *de novo* methylated genes are normally silenced (Klutstein et al., 2016). Such genes are normally controlled by the polycomb protein complex which acts as a transcription repressor that promotes a condensed DNA structure by binding to CpG sites (Klutstein et al., 2016). The methylation of such sites prevents activation even when the polycomb complex is not bound. The complex also promotes DNA methyltransferase activity in tumors, leading to methylation nearby (Klutstein et al., 2016). Additionally, 5-mC cytosines are frequently mutated to thymines and many such mutations occur in the tumor-suppressor gene p53 (Singal and Ginder, 1999), thereby altering methylation patterns at this tumor suppressor. Methylation thus plays an important role in tumor formation and could be useful in clinical applications.

When cells die from cell death mechanisms during the normal cell cycle, short DNA fragments from these dead cells, called cell-free DNA (cfDNA), are released into the bloodstream (Kustanovich et al., 2019). In patients with cancer, a fraction of cfDNA originates from tumor and are called circulating tumor-specific cell-free DNA (ctDNA) (Kustanovich et al., 2019). Levels of ctDNA are higher in individuals with cancer compared to individuals without, higher in individuals with advanced stages of cancer, and reflect the size and progression of the tumor (Schwarzenbach, Hoon, and Pantel, 2011; Thierry et al., 2016). It is known that methylation in these fragments contain molecular signatures from their tissues of origins and these molecular signatures are different in different tissues (Luo et al., 2021). Several studies using cfDNA have shown the ability to detect cancer in asymptomatic or early-stage individuals, supporting their use as indicators of survival CG(not really 'survival' - i.e. still alive, but perhaps indicators of time to cancer occurrence or recurrence), and detection in line with traditional testing (X. Chen et al., 2020; Fernandez-Garcia et al., 2019; Kis et al., 2017; Shen et al., 2018). Looking at DNA methylation from blood samples could be a less invasive screening tool for cancer. For example, methylation regions could be treated as covariates in statistical models to identify cancer cases in individuals not yet diagnosed and to predict the probabilities of developing or recurrence of cancer in individuals.

Many technologies exist for DNA methylation sequencing. The gold standard is whole-genome bisulfite sequencing. Bisulfite sequencing identifies methylated cytosines at base-level resolution by treating DNA with sodium bisulfite that converts unmethylated cytosines to uracil but leaves methylated cytosines as cytosines (Frommer et al., 1992). Once sequenced, unmethylated cytosines are read as thymines while methylated cytosines are read as cytosines. Although this technique provides good resolution (single-base resolution), it is costly and the chemicals used result in DNA degradation, not ideal in the context of ctDNA where the DNA are already short fragments and in relatively low abundance (Shen et al., 2018; Luo et al., 2021). A more suitable technique for analysis of cfDNA is cell-free methylated DNA immunoprecipitation sequencing (cfMeDIP-seq) which can be used for as little as 1 to 10 ng of DNA (Shen et al., 2018). Fragmented DNA, containing some methylated regions, are added to a solution containing antibodies that recognize 5-mC attached to beads (Thu et al., 2009). Only the methylated fragments will be bound to the antibodies. After washing out the antibodies, the unattached DNA fragments, and the beads, the remaining DNA fragments are amplified before being sequenced (Thu et al., 2009). Unlike bisulfite sequencing which measures methylation at almost all CpG sites, cfMeDIP-seq is sensitive to CpG density and provides relative enrichment in DNA methylation over regions of 300 bp, and not an absolute methylation level (Galardi et al., 2020; Pelizzola et al., 2008; Yong, Hsu, and P.-Y. Chen, 2016). After normalization to account for low CpG density regions, absolute and relative methylation levels can be estimated using modeling methods (Pelizzola et al., 2008).

DNA methylation

2.5 High-dimensional data

The high dimensionality of methylation data, such as the data analysed in Chapter 5, poses statistical and computational challenges. Methylation sequencing technologies can measure the methylation status of over 450,000 CpG sites (Dedeurwaerder et al., 2014),

while typically the number of samples used is on the order of tens or hundreds. Performing multiple regression with all the regions is not possible since the feature matrix $\mathbf{X}$ is too large, resulting in the singularity of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ (Johnstone and Titterington, 2009). Hypothesis testing for significance of these regions, if analyzed separately, will result in many significant results by chance due to the high number of regions, requiring procedures for multiple testing correction (Benjamini and Hochberg, 1995). Performing data analysis on such large datasets is also computationally expensive. Instead of analyzing all the methylation sites, dimension reduction summarizes the information of $\mathbf{X}$ in a lower-dimensional function of $\mathbf{X}$ without losing information relevant to the outcome $\mathbf{Y}$ (Adragni and Cook, 2009). One technique for dimension reduction is supervised principal components.

Bair et al., 2006 proposed supervised principal components, that is, principal components with a pre-filtering step of the features in $\mathbf{X}$ based on correlation with $\mathbf{Y}$. Let $\mathbf{X}$ be a $n \times p$ matrix with p features and centered with mean 0, $\mathbf{Y}$ a $n \times 1$ vector of censoring status, $\theta_1, \ldots, \theta_{\mathrm{K}}$ a $K \times 1$ vector of thresholds for univariate models, and m the number of principal components. In the original principal components, using singular value decomposition, $\mathbf{X}$ can be written as $\mathbf{UDV}^{\mathrm{T}}$ where $\mathbf{U}$ is a $n \times q$ matrix with columns $\mathbf{u_1}, \ldots, \mathbf{u_q}$ the principal components, $\mathbf{D}$ is a $q \times q$ diagonal matrix with singular values on the diagonal, and $\mathbf{V}$ is a $q \times p$ matrix. The principal components $\mathbf{u_1}, \ldots, \mathbf{u_q}$ are independent of each other and are chosen in a way to minimize the amount of variability lost when reducing $\mathbf{X}$ to the lower dimension $\mathbf{U}$ (Jolliffe and Cadima, 2016). The supervised version proceeds as follows:

1. A univariate regression or survival model is fitted on feature j, then retain the score statistic

$$\begin{aligned} s_j &= \frac{U_j(0)^2}{I_i(0)} \\ &= \frac{\left(dl_j/d\beta|_{\beta=\beta_0}\right)^2}{-dl_j^2/d\beta^2|_{\beta=\beta_0}}, \end{aligned}$$

   where l_j is the log-likelihood or partial likelihood and β is the regression or survival model coefficient. This is repeated for each of the features.

2. Singular value decomposition of the reduced feature matrix $\mathbf{X}_{\theta_i}$ is computed, where $\mathbf{X}_{\theta_i}$ contains only the features with score statistics $|s_j|$ that exceed threshold θ_i.

3. A multiple regression or survival model is fitted with the first principal component.

4. Steps 2 and 3 are repeated for all K number of thresholds ($\theta_1, \ldots, \theta_{\text{K}}$) and until m principal components are included in the model.

5. Cross-validation is performed to determine the optimal threshold and the number of principal components.

2.6 Measures of performance

The simulation study in Chapter 4 assesses the numerical properties of Cox and casebase model estimates on simulated case-cohort data following the methods described in Morris, White, and Crowther, 2019; White, 2010. The measures used are described in Table 2.1. The mean coefficient and bias measure how close $\hat{\beta}$ is to β on average. The empirical standard error (EmpSE) measures the precision of $\hat{\beta}$. The bootstrapped standard error (BootSE) is the root-mean of the variance of $\hat{\beta}_i$ over n_{bootrep} bootstrap replicates. The mean model-based (ModSE) and bootstrapped standard errors should be close to the empirical standard error such that $\mathbb{E}(\text{ModSE}^2) = \text{EmpSE}^2$ and $\mathbb{E}(\text{BootSE}^2) = \text{EmpSE}^2$ (Morris, White, and Crowther, 2019). Coverage is the probability that β is included in a confidence interval. This thesis uses a 95% confidence interval with model-based or bootstrapped standard error.

Chapter 5 illustrates the use of the weighted Cox and casebase models on a real dataset. The concordance index measures how well the model discriminates between cases and non-cases such that an individual with a smaller risk score has a longer survival time and the higher the concordance index, the better the discrimination (Harrell et al., 1982). Let $I(\cdot)$ be an indicator function, w_i be the sampling weight, T_i be the failure time, $\mathbf{x}_i$ the covariate vector, and $\hat{\boldsymbol{\beta}}$ the coefficient vector of individual i. The weighted

Table 2.1: Performance measures used in simulation study.

Performance measure	Estimate
Mean coef	$\frac{1}{n_{\text{sim}}} \sum_{i=1}^{n_{\text{sim}}} \hat{\beta}_i$
Bias	$\frac{1}{n_{\text{sim}}} \sum_{i=1}^{n_{\text{sim}}} (\hat{\beta}_i - \beta)$
EmpSE	$\sqrt{\frac{1}{n_{\text{sim}}-1} \sum_{i=1}^{n_{\text{sim}}} (\hat{\beta}_i - \bar{\beta})^2}$
Mean ModSE	$\sqrt{\frac{1}{n_{\text{sim}}} \sum_{i=1}^{n_{\text{sim}}} \widehat{\text{Var}}(\beta_i)}$
Mean BootSE	$\sqrt{\frac{1}{n_{\text{sim}}} \sum_{i=1}^{n_{\text{sim}}} \frac{1}{n_{\text{bootrep}-1}} \sum_{j=1}^{n_{\text{bootrep}}} (\hat{\beta}_{ij} - \bar{\beta}_{ij})^2}$
Coverage[a]	$\frac{1}{n_{\text{sim}}} \sum_{i=1}^{n_{\text{sim}}} 1(\hat{\beta}_{\text{low},,i} \leq \bar{\beta} \leq \hat{\beta}_{\text{upp},,i})$

Adapted from Morris, White, and Crowther, 2019; White, 2010. Mean coef: mean coefficient estimate. Mean ModSE: mean model-based standard error. Mean BootSE: mean bootstrapped standard error. EmpSE: empirical standard error.

[a]Monte Carlo standard error: $\sqrt{\frac{\widehat{\text{Cover.}} \times (1 - \widehat{\text{Cover.}})}{n_{\text{sim}}}}$

concordance index

$$\frac{\sum_{i \neq j} w_i w_j I(T_i < T_j, \mathbf{x}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}} > \mathbf{x}_j^{\mathrm{T}} \hat{\boldsymbol{\beta}}) + 0.5 \sum_{i \neq j} w_i w_j I(T_i < T_j, \mathbf{x}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}} = \mathbf{x}_j^{\mathrm{T}} \hat{\boldsymbol{\beta}})}{\sum_{i \neq j} w_i w_j I(T_i < T_j)}$$

weights the sums by the sampling weights w_i, w_j and is identical to the original concordance index if all sampling weights are equal to 1 (Soave and Lawless, 2023). Pairs of individuals are included if the survival times can be ordered (Harrell et al., 1982). That is, if both individuals experienced the event or if one experienced the event before the other is censored. Pairs with both individuals censored are not included in the concordance index.

The Brier score (Brier, 1950) is a measure of the accuracy of predicted probabilities and was initially proposed in the context of weather forecast verification. Instead of assessing whether a weather event occurs or not, a binary classification problem, the Brier score assesses the predicted probability of the weather event. Graf et al., 1999 proposed an empirical Brier score for right-censored survival data. A weighted and scaled Brier score

that takes the sampling weights of the design into account is

$$\text{BS}_\text{w}(t) = \frac{\sum_i w_i \left\{ \frac{I(C_i \geq \min(T_i, t))}{\hat{G}(\min(T_i, t))} \right\} \{I(T_i \leq t) - \hat{F}(t|\mathbf{x}_i)\}^2}{\sum_i w_i},$$

where C_i is the censoring time, $\hat{G}(\min(T_i, t))$ is the Kaplan-Meier estimate of survival probability at time $\min(T_i, t)$, and $\hat{F}(t|\mathbf{x}_i)$ is the predicted probability of event by time t given covariate vector $\mathbf{x}_i$. The Brier score above is scaled by the sum of sampling weights to ensure the computed scores are between 0 and 1.

Chapter 3

Methodology

This chapter presents the methodology underlying the proposed weighted casebase framework. Section 3.1 proposes two methods for including weights in the existing unweighted casebase framework. Section 3.2 proposes a standard error estimate that is appropriate for a biased sampling study design. The simulation studies presented in this chapter are to verify whether the estimates from the weight implementation are reasonable. The numerical properties of the model estimates under various conditions will be examined in more detail in the simulation studies in Chapter 4.

3.1 Implementation of weighted casebase framework

The casebase framework has been implemented in the R package casebase which uses the glm package to fit a logistic regression model with offset $\log(B/b)$, where B is the total follow-up time and b is the size of the base series, to account for the base series sampling. The existing framework does not permit specification of sampling weights, which prevents the use of the casebase framework in the case-cohort setting and other designs with biased sampling of individuals. The sampling fraction of the case-cohort design can be implemented by passing the inverse of the sampling fraction to the weights argument in glm, or by upsampling non-cases in the base series sampling. For the former, this

amounts to solving the weighted log-likelihood $\sum_{i=1}^{n} w_i \log(L(\beta, t, x_i))$, where w_i is the inverse of the sampling fraction and $L(\beta, t, x_i)$ is the likelihood function. For the latter, let B_i be the follow-up time of individual i. Then the total person-time of the study B can be weighted such that $B = \sum_{i=1}^{n} w_i B_i$. The offset is now $\log(\sum_{i=1}^{n} w_i B_i / b)$ and the probability of selecting individual i, the first step of person-moment sampling, is weighted such that $p_i = w_i(B_i/B)$. In other words, follow-up times from non-cases will be weighted to have greater contribution to the total person-time, and non-cases are more likely to be included in the base series.

After making this change to the casebase algorithm, a simulation study is conducted to verify the performance of the weighting methods. A full cohort of $N = 10,000$ individuals is simulated, each with a single covariate x drawn from a log-normal distribution with $\mu = 0$ and $\sigma^2 = 1$ for 1000 individuals. Survival times are generated from a Weibull-Cox distribution (Bender, Augustin, and Blettner, 2005): $T = \left(\frac{-log(U)}{\lambda \exp(\beta x)}\right)$ where $U \sim \text{Unif}(0, 1), \beta = 1.5, \lambda = 10^{-8}$ and $\nu = 4$. Censoring times are generated from an independent Weibull distribution with $\lambda = 10^{-5}$ and $\nu = 8$. The Weibull distribution parameters are chosen to obtain, on average, 200 events after censoring. Simple random sampling among the non-cases (censored) is used to select n_0 non-cases. All n_1 cases (events) are included in the case-cohort sample. Cases have sampling weights of 1 and non-cases have sampling weights of $1/$sampling fraction where the sampling fraction is $\frac{n_0}{N-n_1}$. Weighted casebase models are fitted on the case-cohort sample using either weights in casebase sampling or in glm. The ratio parameter is set to 100, 200, and 500. This process is repeated 1000 times. Model performance is assessed using bias, mean model-based standard error (ModSE), empirical standard error (EmpSE), and coverage probability of 95% confidence interval.

Tables 3.1 and 3.2 show the results of the simulation study, and the consequences of changing casebase sampling and providing weights to glm, respectively. In both Tables, that is for both implementation methods, the coefficient estimates have low bias but the model-based standard error estimates (Mean ModSE) underestimate the empirical stan-

dard errors (EmpSE), leading to poor coverage probabilities. A better standard error estimate is described in the next section.

Table 3.1: casebase models with modified casebase sampling

	Mean coef	Bias	Mean ModSE	EmpSE	Coverage (SE)
Ratio 100	1.54	0.04	0.103	0.147	0.85 (0.01)
Ratio 200	1.53	0.03	0.093	0.137	0.81 (0.01)
Ratio 500	1.52	0.02	0.083	0.124	0.83 (0.01)

$N = 10,000; n_0 = 200; n_1 = 200; p_0 = 0.02$. Mean coef: mean coefficient for log-normal covariate with $\beta = 1.5$. Mean ModSE: mean model-based standard error. EmpSE: empirical standard error. Coverage: coverage of 95% confidence interval using model-based standard error. Ratio 100, 200, 500: weighted casebase models with ratio parameter set to 100, 200, and 500, respectively.

Table 3.2: casebase models with weights in glm

	Mean coef	Bias	Mean ModSE	EmpSE	Coverage (SE)
Ratio 100	1.52	0.02	0.073	0.112	0.81 (0.01)
Ratio 200	1.52	0.02	0.071	0.109	0.80 (0.01)
Ratio 500	1.52	0.02	0.069	0.105	0.82 (0.01)

$N = 10,000; n_0 = 200; n_1 = 200; p_0 = 0.02$. Mean coef: mean coefficient for log-normal covariate with $\beta = 1.5$. Mean ModSE: mean model-based standard error. EmpSE: empirical standard error. Coverage: coverage of 95% confidence interval using model-based standard error. Ratio 100, 200, 500: weighted casebase models with ratio parameter set to 100, 200, and 500, respectively.

3.2 Adjustment to standard errors

The bootstrap proposed by Efron, 1979 can be used to obtain an empirical distribution of $\hat{\beta}$, which then allows for standard error estimation, by resampling the data. To ensure the same number of cases and non-cases are included in each bootstrap sample, stratified bootstrapping (Bickel and Freedman, 1984) is used. Cases and non-cases are treated as separate strata and individuals of each stratum are selected from simple random sampling with replacement to create a new sample for model fitting. Bootstrap resampling and model fitting is repeated 1000 times to obtain 1000 bootstrap replicates. The reported

bootstrapped standard error (BootSE) is the standard deviation of 1000 coefficient estimates.

A simulation study under the same conditions as in 3.1 is performed to assess the use of bootstrapped standard errors. Tables 3.3 and 3.4 show the results of the simulation study. In both implementation methods, the mean bootstrapped standard errors are close to the empirical standard errors, improving the coverage probabilities to 93%. With modified casebase sampling, as the size of the base series increases, both the mean bootstrapped standard errors and the empirical standard errors decrease. With weights in glm, the standard errors are nearly identical as the size of the base series changes. Additionally, the standard errors are larger with modified casebase sampling compared to including weights in glm.

Table 3.3: casebase models with modified casebase sampling and bootstrapped standard errors

	Mean coef	Bias	Mean BootSE	EmpSE	Coverage (SE)
Ratio 100	1.54	0.04	0.135	0.143	0.94 (0.01)
Ratio 200	1.53	0.03	0.124	0.135	0.94 (0.01)
Ratio 500	1.53	0.03	0.114	0.123	0.94 (0.01)

$N = 10,000; n_0 = 200; n_1 = 200; p_0 = 0.02$. Mean coef: mean coefficient for log-normal covariate with $\beta = 1.5$. Mean BootSE: mean bootstrapped standard error of 1000 bootstrap replicates. EmpSE: empirical standard error. Coverage: coverage of 95% confidence interval using bootstrapped standard error. Ratio 100, 200, 500: weighted casebase models with ratio parameter set to 100, 200, and 500, respectively.

Table 3.4: casebase models with weights in glm and bootstrapped standard errors

	Mean coef	Bias	Mean BootSE	EmpSE	Coverage (SE)
Ratio 100	1.52	0.02	0.102	0.110	0.93 (0.01)
Ratio 200	1.52	0.02	0.099	0.105	0.93 (0.01)
Ratio 500	1.52	0.02	0.096	0.102	0.94 (0.01)

$N = 10,000; n_0 = 200; n_1 = 200; p_0 = 0.02$. Mean coef: mean coefficient for log-normal covariate with $\beta = 1.5$. Mean BootSE: mean bootstrapped standard error of 1000 bootstrap replicates. EmpSE: empirical standard error. Coverage: coverage of 95% confidence interval using bootstrapped standard error. Ratio 100, 200, 500: weighted casebase models with ratio parameter set to 100, 200, and 500, respectively.

Since the mean bootstrapped standard errors in Tables 3.3 and 3.4 are close to their respective empirical standard errors, and the the coefficient estimates show very little bias, we felt comfortable proceeding with the larger simulation in Chapter 4 and the application in Chapter 5 using the proposed weighted framework. The results are similar for both weighting methods and either implementation could have been used. The implementation with weights in glm are used in Chapters 4 and 5.

3.3 Software

The proposed weighted framework using weights in glm is implemented in the R package casebaseweights (https://github.com/karinakwan/casebaseweights/), forked from the original casebase package (https://github.com/sahirbhatnagar/casebase).

Chapter 4

Simulation study

This chapter presents a simulation study to assess the numerical properties of the estimates from the proposed weighted framework in a case-cohort setting. The estimates are also compared to those from the Cox proportional hazards model in full cohort, unweighted case-cohort, and weighted case-cohort settings. The simulation study is conducted following the approach of Morris, White, and Crowther, 2019.

4.1 Data generation

We undertook a simulation study examining cohorts of sizes $N = \{$5,000, 10,000, 50,000, 100,000$\}$ to study the effect of sampling fractions and weights on model stability and on the numerical properties of the casebase parameter estimates under the weighted framework. The single covariate distributions x and Weibull parameters are as described in Table 4.1. Survival times are generated under a Cox-Weibull distribution $T = \left(\frac{-log(U)}{\lambda \exp(\beta x)}\right)^{1/\nu}$ with $U \sim$ Unif$(0, 1)$, as described in Bender, Augustin, and Blettner, 2005, so that the proportional hazards assumption is satisfied. The true coefficient for the covariate effect is fixed at $\beta = 1.5$ for all cohort sizes and covariate distributions. The generated survival times are ordered from shortest to longest and the 200 smallest survival times are defined as events such that there are exactly $n_1 = 200$ events in each simulation replicate. The

remaining $N - 200$ survival times are right-censored. To create case-cohort samples, all individuals with events (cases) are included. Of the remaining $N - 200$ censored individuals (non-cases), $n_0 = 200$ are randomly sampled without replacement. This gives sampling weights of 1 for cases and $\frac{N-200}{n_0}$ for non-cases. A thousand such replicate cohorts and case-cohort samples are created for each cohort size and covariate distribution combination.

Table 4.1: Parameters used in simulation study

Cohort size (N)	Covariate distribution (x)	Weibull scale (λ)	Weibull shape (ν)
5,000	$\mathcal{B}$(p = 0.3, n = N)	10^{-5}	6.74
10,000	$\mathcal{B}$(p = 0.3, n = N)	10^{-5}	6.75
50,000	$\mathcal{B}$(p = 0.3, n = N)	10^{-5}	6.75
100,000	$\mathcal{B}$(p = 0.3, n = N)	10^{-5}	6.75
5,000	$\mathcal{N}(\mu = 0, \sigma^2 = 1)$	10^{-5}	7.00
10,000	$\mathcal{N}(\mu = 0, \sigma^2 = 1)$	10^{-5}	7.00
50,000	$\mathcal{N}(\mu = 0, \sigma^2 = 1)$	10^{-5}	7.00
100,000	$\mathcal{N}(\mu = 0, \sigma^2 = 1)$	10^{-5}	7.00
5,000	Lognormal($\mu = 0, \sigma^2 = 1$)	10^{-5}	5.00
10,000	Lognormal($\mu = 0, \sigma^2 = 1$)	10^{-5}	5.00
50,000	Lognormal($\mu = 0, \sigma^2 = 1$)	10^{-5}	5.00
100,000	Lognormal($\mu = 0, \sigma^2 = 1$)	10^{-5}	5.00

4.2 Methods of Analysis

The simulation study compares weighted casebase models to weighted Cox models fitted on case-cohort samples. The models will also be compared to unweighted casebase and Cox models fitted on the full cohort and case-cohort samples.

Cox proportional hazards models with hazard functions of the $h(t) = h_0(t)\exp(\beta x)$ are fitted on the cohort of size N and the case-cohort sample of size $n_0 + n_1 = 400$. Unweighted Cox models fit on the entire cohort and on the case-cohort sample are listed as "Cox full" and "Cox naive", respectively, in the results. Weighted Cox regression, which uses weights in the partial likelihood, is used on the case-cohort sample to obtain robust

coefficient and standard error estimates. Weighted Cox models are listed as "Cox robust" in the results tables.

Three base series sizes are used for each of the simulation settings under the original and proposed weighted casebase frameworks to study their effect on standard error estimates. Ratios of case series to base series equal to 100, 200, and 500 are considered. Casebase models of the form $h(t) = \exp(\beta_0 + \beta_1 \log(t) + \beta_2 x)$ are fitted on the cohort of size N and the case-cohort sample of size 400. The proposed framework with stratified bootstrapping to estimate standard errors is used on the case-cohort sample. Models fitted on the entire cohort and on the case-cohort sample are listed as "Ratio full" and "Ratio naive", respectively, with the appropriate ratio parameters. Models fitted with the proposed weighted framework are listed as "Ratio robust" with the appropriate ratio parameters.

The numerical properties of the model estimates are assessed on mean coefficient estimate, bias, mean model-based or bootstrapped standard error, empirical standard error, and coverage probability as defined in Chapter 2. Bias is the mean deviation of the coefficient estimate from the true coefficient (1.5). Empirical standard error is the standard deviation of coefficient estimates. Mean model-based standard error is used for all Cox and unweighted casebase models and is the root-mean of the squared standard error of the model coefficient estimate. Mean bootstrapped standard error is used for all weighted casebase models and is the root-mean of the variance of the bootstrapped coefficient estimates.

The R package survival is used to fit all the Cox models. The R package casebase, the implementation of the original casebase framework, is used to fit the models on the full cohort and the unweighted models on the case-cohort sample. The R package casebaseweights, modified from the original package to include weights, is used to fit the weighted models on the case-cohort sample.

4.3 Results

Tables 4.2-4.5 show the results from 1000 simulation replicates with a single binomial covariate. The first four rows are models fitted on the entire cohort, the middle four are unweighted naive models fitted on the case-cohort sample, and the last four are weighted models fitted on the case-cohort sample. The models fitted on the full cohorts, for any cohort size, have unbiased coefficient estimates, model-based standard errors nearly identical to empirical standard errors, and good coverage probabilities. The naive models on the case-control samples all have biased coefficient estimates, model-based standard errors and empirical standard errors nearly identical to those of the full models (i.e. too small), and poor coverage probabilities. In contrast, all robust model coefficients have no or small bias and good coverage probabilities. Their standard errors are larger than those of the unweighted models. Model-based, bootstrapped, and empirical standard errors get smaller as the casebase ratio parameter gets larger, approaching the respective Cox standard errors.

Tables 4.6-4.9 show the results from 1000 simulation replicates with a single normal covariate. All models fitted on the full cohort have unbiased coefficient estimates, model-based standard errors nearly identical to empirical standard errors, and good coverage probabilities. The naive models have biased coefficient estimates, resulting in poor coverage probabilities. The robust models have coefficient estimates with small bias under cohort sizes $N = \{5,000, 10,000\}$ and biased coefficient estimates under cohort sizes $N = \{50,000, 100,000\}$. These robust models underestimate the empirical standard errors, leading to poor coverage probabilities. The standard errors are larger than the weighted Cox robust standard errors although the bootstrapped standard errors decrease as the ratio parameter increases. Weighted Cox and casebase models have identical bias and nearly identical empirical standard errors and coverage probabilities.

Tables 4.10-4.13 show the results from 1000 simulation replicates with a single log-normal covariate. Under cohort sizes $N = \{5,000, 10,000\}$, the full models have coeffi-

Table 4.2: Binary covariate, N = 5,000

	Mean coef	Bias	Mean ModSE or BootSE	EmpSE	Coverage (SE)
Cox full	1.51	0.01	0.149	0.148	0.96 (0.01)
Ratio 100 full	1.51	0.01	0.152	0.150	0.96 (0.01)
Ratio 200 full	1.51	0.01	0.150	0.149	0.96 (0.01)
Ratio 500 full	1.51	0.01	0.149	0.148	0.96 (0.01)
Cox naive	1.06	-0.44	0.149	0.146	0.16 (0.01)
Ratio 100 naive	1.04	-0.46	0.152	0.146	0.14 (0.01)
Ratio 200 naive	1.04	-0.46	0.151	0.145	0.13 (0.01)
Ratio 500 naive	1.04	-0.46	0.150	0.143	0.13 (0.01)
Cox robust	1.52	0.02	0.210	0.207	0.96 (0.01)
Ratio 100 robust	1.52	0.02	0.216	0.212	0.96 (0.01)
Ratio 200 robust	1.52	0.02	0.214	0.210	0.96 (0.01)
Ratio 500 robust	1.52	0.02	0.213	0.208	0.96 (0.01)

$N = 5,000; n_0 = 200; n_1 = 200; p_0 = 0.04; \text{weight}_{\text{control}} = 25$. Binary covariate with $\beta = 1.5$. Full models fitted on entire cohort of size N. Naive: unweighted models fitted on case-cohort of size $n_0 + n_1$. Robust: weighted models fitted on case-cohort of size $n_0 + n_1$. Ratio 100, 200, 500: casebase models with ratio parameter set to 100, 200, and 500, respectively. Coef: coefficient. ModSE: model-based standard error. BootSE: bootstrapped standard error. EmpSE: empirical standard error. SE: standard error. Mean ModSE for all full models, all naive models, and robust Cox model. Mean BootSE for robust casebase models.

Table 4.3: Binary covariate, N = 10,000

	Mean coef	Bias	Mean ModSE or BootSE	EmpSE	Coverage (SE)
Cox full	1.50	0.00	0.149	0.146	0.95 (0.01)
Ratio 100 full	1.50	0.00	0.152	0.148	0.96 (0.01)
Ratio 200 full	1.50	0.00	0.151	0.148	0.95 (0.01)
Ratio 500 full	1.50	0.00	0.150	0.146	0.95 (0.01)
Cox naive	1.04	-0.46	0.150	0.144	0.14 (0.01)
Ratio 100 naive	1.02	-0.48	0.153	0.146	0.12 (0.01)
Ratio 200 naive	1.02	-0.48	0.151	0.144	0.11 (0.01)
Ratio 500 naive	1.02	-0.48	0.150	0.142	0.11 (0.01)
Cox robust	1.51	0.01	0.213	0.208	0.94 (0.01)
Ratio 100 robust	1.51	0.01	0.218	0.214	0.95 (0.01)
Ratio 200 robust	1.51	0.01	0.216	0.210	0.95 (0.01)
Ratio 500 robust	1.51	0.01	0.215	0.208	0.95 (0.01)

$N = 10,000; n_0 = 200; n_1 = 200; p_0 = 0.02; \text{weight}_{\text{control}} = 50$. Binary covariate with $\beta = 1.5$. Full models fitted on entire cohort of size N. Naive: unweighted models fitted on case-cohort of size $n_0 + n_1$. Robust: weighted models fitted on case-cohort of size $n_0 + n_1$. Ratio 100, 200, 500: casebase models with ratio parameter set to 100, 200, and 500, respectively. Coef: coefficient. ModSE: model-based standard error. BootSE: bootstrapped standard error. EmpSE: empirical standard error. SE: standard error. Mean ModSE for all full models, all naive models, and robust Cox model. Mean BootSE for robust casebase models.

Table 4.4: Binary covariate, N = 50,000

	Mean coef	Bias	Mean ModSE or BootSE	EmpSE	Coverage (SE)
Cox full	1.50	0.00	0.149	0.151	0.94 (0.01)
Ratio 100 full	1.50	0.00	0.152	0.155	0.94 (0.01)
Ratio 200 full	1.50	0.00	0.151	0.153	0.95 (0.01)
Ratio 500 full	1.50	0.00	0.150	0.152	0.94 (0.01)
Cox naive	1.03	-0.47	0.150	0.147	0.12 (0.01)
Ratio 100 naive	1.00	-0.50	0.153	0.147	0.09 (0.01)
Ratio 200 naive	1.01	-0.49	0.152	0.146	0.10 (0.01)
Ratio 500 naive	1.01	-0.49	0.151	0.145	0.10 (0.01)
Cox robust	1.49	-0.01	0.215	0.212	0.96 (0.01)
Ratio 100 robust	1.49	-0.01	0.220	0.215	0.96 (0.01)
Ratio 200 robust	1.50	0.00	0.218	0.214	0.96 (0.01)
Ratio 500 robust	1.49	-0.01	0.217	0.213	0.96 (0.01)

$N = 50,000; n_0 = 200; n_1 = 200; p_0 = 0.004; \text{weight}_{\text{control}} = 250$. Binary covariate with $\beta = 1.5$. Full models fitted on entire cohort of size N. Naive: unweighted models fitted on case-cohort of size $n_0 + n_1$. Robust: weighted models fitted on case-cohort of size $n_0 + n_1$. Ratio 100, 200, 500: casebase models with ratio parameter set to 100, 200, and 500, respectively. Coef: coefficient. ModSE: model-based standard error. BootSE: bootstrapped standard error. EmpSE: empirical standard error. SE: standard error. Mean ModSE for all full models, all naive models, and robust Cox model. Mean BootSE for robust casebase models.

Table 4.5: Binary covariate, N = 100,000

	Mean coef	Bias	Mean ModSE or BootSE	EmpSE	Coverage (SE)
Cox full	1.50	0.00	0.149	0.149	0.94 (0.01)
Ratio 100 full	1.50	0.00	0.152	0.153	0.95 (0.01)
Ratio 200 full	1.50	0.00	0.151	0.149	0.95 (0.01)
Ratio 500 full	1.50	0.00	0.150	0.150	0.95 (0.01)
Cox naive	1.03	-0.47	0.150	0.143	0.12 (0.01)
Ratio 100 naive	1.00	-0.50	0.153	0.144	0.09 (0.01)
Ratio 200 naive	1.01	-0.49	0.152	0.143	0.09 (0.01)
Ratio 500 naive	1.01	-0.49	0.151	0.141	0.09 (0.01)
Cox robust	1.50	0.00	0.215	0.206	0.96 (0.01)
Ratio 100 robust	1.50	0.00	0.221	0.210	0.96 (0.01)
Ratio 200 robust	1.50	0.00	0.219	0.208	0.96 (0.01)
Ratio 500 robust	1.50	0.00	0.217	0.205	0.96 (0.01)

$N = 100,000; n_0 = 200; n_1 = 200; p_0 = 0.002; \text{weight}_{\text{control}} = 500$. Binary covariate with $\beta = 1.5$. Full models fitted on entire cohort of size N. Naive: unweighted models fitted on case-cohort of size $n_0 + n_1$. Robust: weighted models fitted on case-cohort of size $n_0 + n_1$. Ratio 100, 200, 500: casebase models with ratio parameter set to 100, 200, and 500, respectively. Coef: coefficient. ModSE: model-based standard error. BootSE: bootstrapped standard error. EmpSE: empirical standard error. SE: standard error. Mean ModSE for all full models, all naive models, and robust Cox model. Mean BootSE for robust casebase models.

Table 4.6: Normal covariate, N = 5,000

	Mean coef	Bias	Mean ModSE or BootSE	EmpSE	Coverage (SE)
Cox full	1.50	0.00	0.073	0.075	0.94 (0.01)
Ratio 100 full	1.50	0.00	0.084	0.083	0.95 (0.01)
Ratio 200 full	1.50	0.00	0.080	0.081	0.95 (0.01)
Ratio 500 full	1.50	0.00	0.076	0.078	0.94 (0.01)
Cox naive	0.84	-0.66	0.066	0.060	0.00 (0.00)
Ratio 100 naive	0.82	-0.68	0.069	0.063	0.00 (0.00)
Ratio 200 naive	0.82	-0.68	0.067	0.061	0.00 (0.00)
Ratio 500 naive	0.82	-0.68	0.066	0.060	0.00 (0.00)
Cox robust	1.53	0.03	0.137	0.162	0.86 (0.01)
Ratio 100 robust	1.53	0.03	0.146	0.164	0.88 (0.01)
Ratio 200 robust	1.53	0.03	0.144	0.164	0.88 (0.01)
Ratio 500 robust	1.53	0.03	0.142	0.163	0.87 (0.01)

$N = 5,000; n_0 = 200; n_1 = 200; p_0 = 0.04; \text{weight}_{\text{control}} = 25$. Normal covariate with $\beta = 1.5$. Full models fitted on entire cohort of size N. Naive: unweighted models fitted on case-cohort of size $n_0 + n_1$. Robust: weighted models fitted on case-cohort of size $n_0 + n_1$. Ratio 100, 200, 500: casebase models with ratio parameter set to 100, 200, and 500, respectively. Coef: coefficient. ModSE: model-based standard error. BootSE: bootstrapped standard error. EmpSE: empirical standard error. SE: standard error. Mean ModSE for all full models, all naive models, and robust Cox model. Mean BootSE for robust casebase models.

Table 4.7: Normal covariate, N = 10,000

	Mean coef	Bias	Mean ModSE or BootSE	EmpSE	Coverage (SE)
Cox full	1.50	0.00	0.072	0.071	0.95 (0.01)
Ratio 100 full	1.50	0.00	0.084	0.085	0.94 (0.01)
Ratio 200 full	1.50	0.00	0.079	0.079	0.96 (0.01)
Ratio 500 full	1.50	0.00	0.076	0.074	0.95 (0.01)
Cox naive	0.78	-0.72	0.063	0.057	0.00 (0.00)
Ratio 100 naive	0.76	-0.74	0.066	0.057	0.00 (0.00)
Ratio 200 naive	0.76	-0.74	0.064	0.056	0.00 (0.00)
Ratio 500 naive	0.76	-0.74	0.063	0.055	0.00 (0.00)
Cox robust	1.56	0.06	0.164	0.210	0.82 (0.01)
Ratio 100 robust	1.56	0.06	0.179	0.214	0.84 (0.01)
Ratio 200 robust	1.56	0.06	0.177	0.211	0.83 (0.01)
Ratio 500 robust	1.56	0.06	0.175	0.210	0.84 (0.01)

$N = 10,000; n_0 = 200; n_1 = 200; p_0 = 0.02; \text{weight}_{\text{control}} = 50$. Normal covariate with $\beta = 1.5$. Full models fitted on entire cohort of size N. Naive: unweighted models fitted on case-cohort of size $n_0 + n_1$. Robust: weighted models fitted on case-cohort of size $n_0 + n_1$. Ratio 100, 200, 500: casebase models with ratio parameter set to 100, 200, and 500, respectively. Coef: coefficient. ModSE: model-based standard error. BootSE: bootstrapped standard error. EmpSE: empirical standard error. SE: standard error. Mean ModSE for all full models, all naive models, and robust Cox model. Mean BootSE for robust casebase models.

Table 4.8: Normal covariate, N = 50,000

	Mean coef	Bias	Mean ModSE or BootSE	EmpSE	Coverage (SE)
Cox full	1.50	0.00	0.071	0.072	0.95 (0.01)
Ratio 100 full	1.50	0.00	0.085	0.084	0.95 (0.01)
Ratio 200 full	1.50	0.00	0.080	0.081	0.96 (0.01)
Ratio 500 full	1.50	0.00	0.076	0.076	0.95 (0.01)
Cox naive	0.72	-0.78	0.060	0.057	0.00 (0.00)
Ratio 100 naive	0.71	-0.79	0.063	0.057	0.00 (0.00)
Ratio 200 naive	0.70	-0.80	0.061	0.056	0.00 (0.00)
Ratio 500 naive	0.70	-0.80	0.060	0.054	0.00 (0.00)
Cox robust	1.61	0.11	0.211	0.281	0.80 (0.01)
Ratio 100 robust	1.61	0.11	0.250	0.285	0.84 (0.01)
Ratio 200 robust	1.61	0.11	0.247	0.283	0.83 (0.01)
Ratio 500 robust	1.61	0.11	0.245	0.281	0.84 (0.01)

$N = 50,000; n_0 = 200; n_1 = 200; p_0 = 0.004; \text{weight}_{\text{control}} = 250$. Normal covariate with $\beta = 1.5$. Full models fitted on entire cohort of size N. Naive: unweighted models fitted on case-cohort of size $n_0 + n_1$. Robust: weighted models fitted on case-cohort of size $n_0 + n_1$. Ratio 100, 200, 500: casebase models with ratio parameter set to 100, 200, and 500, respectively. Coef: coefficient. ModSE: model-based standard error. BootSE: bootstrapped standard error. EmpSE: empirical standard error. SE: standard error. Mean ModSE for all full models, all naive models, and robust Cox model. Mean BootSE for robust casebase models.

Table 4.9: Normal covariate, N = 100,000

	Mean coef	Bias	Mean ModSE or BootSE	EmpSE	Coverage (SE)
Cox full	1.50	0.00	0.072	0.073	0.95 (0.01)
Ratio 100 full	1.50	0.00	0.085	0.085	0.96 (0.01)
Ratio 200 full	1.50	0.00	0.080	0.080	0.96 (0.01)
Ratio 500 full	1.50	0.00	0.077	0.078	0.96 (0.01)
Cox naive	0.73	-0.77	0.061	0.059	0.00 (0.00)
Ratio 100 naive	0.71	-0.79	0.064	0.060	0.00 (0.00)
Ratio 200 naive	0.71	-0.79	0.062	0.059	0.00 (0.00)
Ratio 500 naive	0.71	-0.79	0.061	0.058	0.00 (0.00)
Cox robust	1.62	0.12	0.230	0.310	0.84 (0.01)
Ratio 100 robust	1.62	0.12	0.284	0.315	0.87 (0.01)
Ratio 200 robust	1.62	0.12	0.281	0.311	0.87 (0.01)
Ratio 500 robust	1.62	0.12	0.279	0.309	0.86 (0.01)

$N = 100,000; n_0 = 200; n_1 = 200; p_0 = 0.002; \text{weight}_{\text{control}} = 500$. Normal covariate with $\beta = 1.5$. Full models fitted on entire cohort of size N. Naive: unweighted models fitted on case-cohort of size $n_0 + n_1$. Robust: weighted models fitted on case-cohort of size $n_0 + n_1$. Ratio 100, 200, 500: casebase models with ratio parameter set to 100, 200, and 500, respectively. Coef: coefficient. ModSE: model-based standard error. BootSE: bootstrapped standard error. EmpSE: empirical standard error. SE: standard error. Mean ModSE for all full models, all naive models, and robust Cox model. Mean BootSE for robust casebase models.

cient estimates with small or no bias, nearly identical model-based and empirical standard errors, and good coverage probabilities. The robust models have coefficient estimates with small bias and good coverage probabilities. In comparison, full models on $N = \{50,000, 100,000\}$ have small bias or biased coefficient estimates. Notably, the Cox full models have poor coverage probabilities while the casebase models have larger standard errors, both model-based and empirical, and better coverage probabilities. The robust models have standard errors that underestimate the empirical standard errors and poor coverage probabilities. The Cox robust models have larger bias coefficient estimates than casebase robust models and comparable standard errors, resulting in worse coverage probabilities.

Table 4.10: Log-normal covariate, N = 5,000

	Mean coef	Bias	Mean ModSE or BootSE	EmpSE	Coverage (SE)
Cox full	1.50	0.00	0.066	0.067	0.94 (0.01)
Ratio 100 full	1.51	0.01	0.084	0.086	0.95 (0.01)
Ratio 200 full	1.51	0.01	0.078	0.077	0.95 (0.01)
Ratio 500 full	1.51	0.01	0.072	0.071	0.95 (0.01)
Cox naive	1.16	-0.34	0.068	0.070	0.01 (0.00)
Ratio 100 naive	1.18	-0.32	0.072	0.074	0.02 (0.00)
Ratio 200 naive	1.19	-0.31	0.071	0.073	0.03 (0.01)
Ratio 500 naive	1.21	-0.29	0.070	0.072	0.04 (0.01)
Cox robust	1.52	0.02	0.078	0.090	0.92 (0.01)
Ratio 100 robust	1.52	0.02	0.086	0.092	0.94 (0.01)
Ratio 200 robust	1.52	0.02	0.084	0.093	0.94 (0.01)
Ratio 500 robust	1.52	0.02	0.082	0.091	0.94 (0.01)

$N = 5,000; n_0 = 200; n_1 = 200; p_0 = 0.04;$ $\text{weight}_{\text{control}} = 25$. Log-normal covariate with $\beta = 1.5$. Full models fitted on entire cohort of size N. Naive: unweighted models fitted on case-cohort of size $n_0 + n_1$. Robust: weighted models fitted on case-cohort of size $n_0 + n_1$. Ratio 100, 200, 500: casebase models with ratio parameter set to 100, 200, and 500, respectively. Coef: coefficient. ModSE: model-based standard error. BootSE: bootstrapped standard error. EmpSE: empirical standard error. SE: standard error. Mean ModSE for all full models, all naive models, and robust Cox model. Mean BootSE for robust casebase models.

Figure 4.1 summarizes the effect of cohort size and covariate distribution among the ratio 500 robust models in terms of bias and relative difference in standard error defined as (Mean BootSE − EmpSE)/Mean BootSE. The relative difference in standard error is

Table 4.11: Log-normal covariate, N = 10,000

	Mean coef	Bias	Mean ModSE or BootSE	EmpSE	Coverage (SE)
Cox full	1.50	0.00	0.071	0.070	0.96 (0.01)
Ratio 100 full	1.52	0.02	0.107	0.110	0.96 (0.01)
Ratio 200 full	1.51	0.01	0.096	0.097	0.95 (0.01)
Ratio 500 full	1.51	0.01	0.086	0.085	0.96 (0.01)
Cox naive	1.18	-0.32	0.073	0.074	0.02 (0.00)
Ratio 100 naive	1.21	-0.29	0.078	0.078	0.06 (0.01)
Ratio 200 naive	1.22	-0.28	0.076	0.077	0.07 (0.01)
Ratio 500 naive	1.24	-0.26	0.074	0.075	0.08 (0.01)
Cox robust	1.53	0.03	0.093	0.096	0.94 (0.01)
Ratio 100 robust	1.51	0.01	0.104	0.109	0.94 (0.01)
Ratio 200 robust	1.51	0.01	0.100	0.106	0.94 (0.01)
Ratio 500 robust	1.51	0.01	0.098	0.104	0.95 (0.01)

$N = 10,000; n_0 = 200; n_1 = 200; p_0 = 0.02; \text{weight}_{\text{control}} = 50$. Log-normal covariate with $\beta = 1.5$. Full models fitted on entire cohort of size N. Naive: unweighted models fitted on case-cohort of size $n_0 + n_1$. Robust: weighted models fitted on case-cohort of size $n_0 + n_1$. Ratio 100, 200, 500: casebase models with ratio parameter set to 100, 200, and 500, respectively. Coef: coefficient. ModSE: model-based standard error. BootSE: bootstrapped standard error. EmpSE: empirical standard error. SE: standard error. Mean ModSE for all full models, all naive models, and robust Cox model. Mean BootSE for robust casebase models.

Table 4.12: Log-normal covariate, N = 50,000

	Mean coef	Bias	Mean ModSE or BootSE	EmpSE	Coverage (SE)
Cox full	1.43	-0.07	0.075	0.071	0.82 (0.01)
Ratio 100 full	1.55	0.05	0.185	0.199	0.95 (0.01)
Ratio 200 full	1.53	0.03	0.156	0.162	0.95 (0.01)
Ratio 500 full	1.52	0.02	0.130	0.135	0.94 (0.01)
Cox naive	1.20	-0.30	0.079	0.075	0.05 (0.01)
Ratio 100 naive	1.32	-0.18	0.088	0.091	0.44 (0.02)
Ratio 200 naive	1.33	-0.17	0.084	0.088	0.47 (0.02)
Ratio 500 naive	1.34	-0.16	0.082	0.085	0.48 (0.02)
Cox robust	1.43	-0.07	0.132	0.161	0.74 (0.01)
Ratio 100 robust	1.49	-0.01	0.136	0.190	0.79 (0.01)
Ratio 200 robust	1.49	-0.01	0.130	0.184	0.80 (0.01)
Ratio 500 robust	1.49	-0.01	0.124	0.173	0.80 (0.01)

$N = 50,000; n_0 = 200; n_1 = 200; p_0 = 0.004; \text{weight}_{\text{control}} = 250$. Log-normal covariate with $\beta = 1.5$. Full models fitted on entire cohort of size N. Naive: unweighted models fitted on case-cohort of size $n_0 + n_1$. Robust: weighted models fitted on case-cohort of size $n_0 + n_1$. Ratio 100, 200, 500: casebase models with ratio parameter set to 100, 200, and 500, respectively. Coef: coefficient. ModSE: model-based standard error. BootSE: bootstrapped standard error. EmpSE: empirical standard error. SE: standard error. Mean ModSE for all full models, all naive models, and robust Cox model. Mean BootSE for robust casebase models.

Table 4.13: Log-normal covariate, N = 100,000

	Mean coef	Bias	Mean ModSE or BootSE	EmpSE	Coverage (SE)
Cox full	1.42	-0.08	0.077	0.074	0.81 (0.01)
Ratio 100 full	1.59	0.09	0.239	0.258	0.96 (0.01)
Ratio 200 full	1.56	0.06	0.196	0.203	0.96 (0.01)
Ratio 500 full	1.53	0.03	0.156	0.159	0.95 (0.01)
Cox naive	1.21	-0.29	0.080	0.078	0.06 (0.01)
Ratio 100 naive	1.32	-0.18	0.088	0.090	0.45 (0.02)
Ratio 200 naive	1.34	-0.16	0.085	0.089	0.50 (0.02)
Ratio 500 naive	1.35	-0.15	0.083	0.087	0.51 (0.02)
Cox robust	1.36	-0.14	0.124	0.187	0.53 (0.02)
Ratio 100 robust	1.44	-0.06	0.127	0.203	0.66 (0.01)
Ratio 200 robust	1.44	-0.06	0.122	0.196	0.67 (0.01)
Ratio 500 robust	1.45	-0.05	0.117	0.187	0.70 (0.01)

$N = 100,000; n_0 = 200; n_1 = 200; p_0 = 0.002;$ weight$_{\text{control}} = 500$. Log-normal covariate with $\beta = 1.5$. Full models fitted on entire cohort of size N. Naive: unweighted models fitted on case-cohort of size $n_0 + n_1$. Robust: weighted models fitted on case-cohort of size $n_0 + n_1$. Ratio 100, 200, 500: casebase models with ratio parameter set to 100, 200, and 500, respectively. Coef: coefficient. ModSE: model-based standard error. BootSE: bootstrapped standard error. EmpSE: empirical standard error. SE: standard error. Mean ModSE for all full models, all naive models, and robust Cox model. Mean BootSE for robust casebase models.

used to quantify how much the mean bootstrapped standard error differs from the empirical standard error and should be close to 0. A horizontal line is drawn in red at 0. The solid black lines indicate a binomial distribution, dashed lines indicate a log-normal distribution, and dotted lines indicate a normal distribution. The choice of robust casebase model is arbitrary; the same patterns hold for all robust casebase models. The bias and relative difference in standard error are close to 0 with a binomial distribution for all cohort sizes. As cohort size increases, the bias and relative difference in standard error increase with normal and log-normal distributions. The bias is larger with a normal covariate than with a log-normal covariate. The relative difference in standard is smaller with a log-normal covariate for cohort sizes $N = \{5,000, 10,000\}$ and with a normal covariate for cohort sizes $N = \{50,000, 100,000\}$.

Various cohort sizes were used to study the effect of sampling fractions, and consequently sampling weights, on the proposed framework. The resulting sampling weights

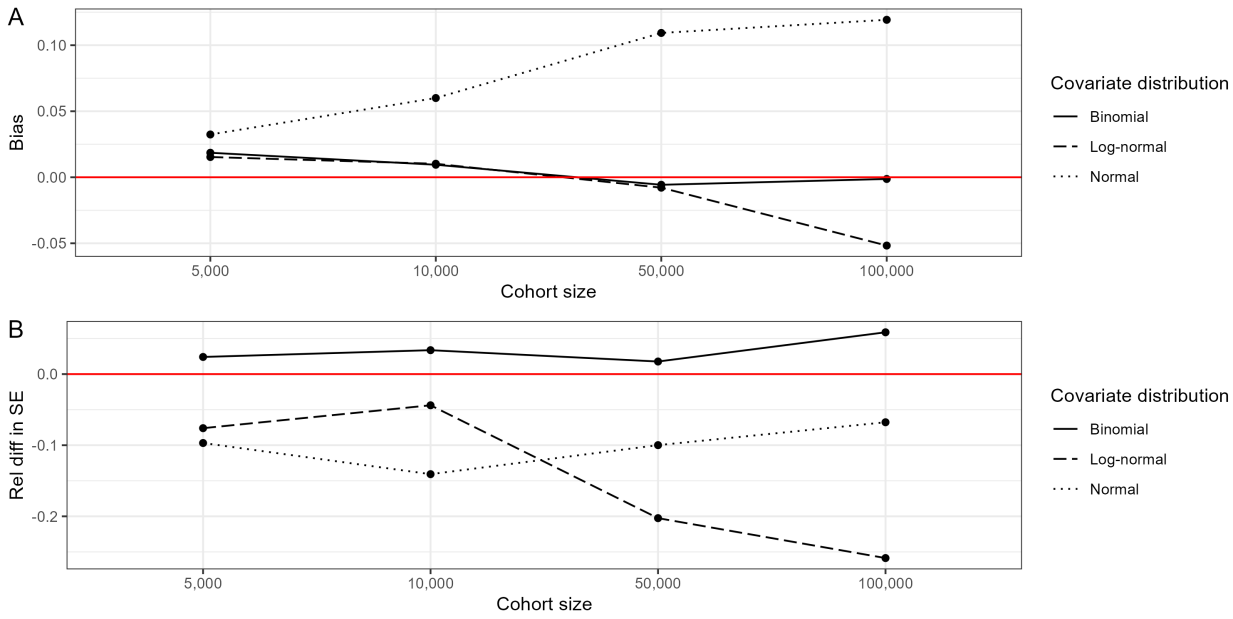

Figure 4.1: Effect of cohort size and covariate distribution

Rel diff in SE: relative difference in standard error (Mean BootSE − EmpSE)/Mean BootSE. Bias and rel diff in SE of casebase ratio 500 robust models fitted on case-cohort samples across cohort sizes N={5,000, 10,000, 50,000, 10,000} and binomial, log-normal, normal covariate distributions. Solid black line: binomial distribution; dashed line: log-normal distribution; dotted line: normal distribution.

ranged from moderate to extreme in order to study how weighted casebase fares in potentially challenging settings. The use of different covariate distributions also served as potential challenges to the framework, particularly the log-normal distribution which is skewed.

Chapter 5

A case-cohort study of breast cancer incidence using cell-free DNA methylation measurements

Following the simulation study described in Chapter 4, this chapter applies the proposed weighted casebase framework to study breast cancer incidence using cell-free DNA methylation measurements obtained from the Ontario Institute for Cancer Research (OICR). The resulting weighted casebase model is compared to a weighted Cox model using a weighted concordance and a weighted Brier score. Absolute risk curves are used to compare the resulting risk estimates to Kaplan-Meier estimates.

5.1 Ontario Health Study

The Ontario Health Study (OHS) (Kirsh et al., 2022) is an ongoing longitudinal cohort study that follows over 225,000 adults from the general Ontario population, representing 1% of the Ontario population. The study collects baseline data on lifestyle and environmental factors, and biological samples to understand their effects on participants' health. Participants were recruited between 2009 and 2017. At study entry (baseline),

all participants completed a questionnaire, which includes questions regarding socio-demographic information, self-reported health status and physical measures, family history, and lifestyle factors. Subsequent follow-up questionnaires regarding work-history and COVID-19 were sent out more recently.

A benefit to the study is the linkage to electronic health records and databases, subject to participant consent. As such, all participants have ecologic environmental measurements through linkage to the Canadian Urban Environmental Health Research Consortium. allowing for measurements of environmental exposures. Over 188,000 participants agreed to administrative health data linkage, including to the Ontario Cancer Registry. A subset of almost 40,000 participants provided blood samples, allowing researchers to perform whole genome-sequencing and genomics analyses.

5.2 Methods

5.2.1 Participant selection and methylation data

This thesis uses cfMeDIP-seq and participant data obtained from the OICR. The data collection, and sequencing and processing protocols are described in Cheng et al., 2023. A summary is provided here.

Plasma samples were obtained from participants of the OHS: 110 with breast cancer and 108 participants without a prior cancer diagnosis up to the end of the follow-up period. All participants were healthy at sample collection (baseline) with no prior history of cancer. The 110 participants with breast cancer were subsequently diagnosed with incident breast cancer during follow-up. The 108 participants were controls selected from the participants that did not develop cancer during follow-up and matched to cases on age, sex, date of plasma collection, ethnicity, smoking status, and alcohol consumption frequency. Methylated cfDNA were isolated through immunoprecipitation and sequenced in batches using the Novaseq platform with controls and cases included in the same batches (total of 8 batches). Over 9 million 300-bp regions were sequenced. Following

removal of samples with poor sequencing quality and of controls that died or had subsequent cancer diagnosis, data from 80 incident cancer patients and 70 cancer-free controls were included for a total sample size of 150 and over a million regions. Biological filtering of methylation regions was performed, removing regions from sex chromosomes and regions that are frequently methylated in blood, and retaining CpG dense and regulatory regions. This left 101,955 methylation regions to be used as predictive covariates in later analyses.

5.2.2 Dimension reduction and model building

The data are first divided into a train set, consisting of samples from batches 1 to 7, and a test set, consisting of samples from batch 8. The train set includes 124 samples, the test set 26 samples. Due to the large number of potentially uninformative regions, supervised principal components is used to reduce the dimension of the dataset. The supervised principal components method, described in Section 2.5, calculates principal components on a reduced dataset containing only methylation regions with score statistics greater than a threshold θ. In the train set, weighted Cox and weighted casebase models are fitted using supervised principal components as the predictors. Then the test set is projected onto the supervised principal components space. The models fitted on the train set are predicted in the test set using the test set supervised principal components to calculate linear predictors and probabilities of breast cancer diagnosis by time t (absolute risk). In order to determine the optimal number of principal components npc $\in \{1, \ldots, 5\}$ and the threshold $\theta \in [0, \ldots, 3]$ (tuning parameters), repeated 10-fold cross-validation is performed. The train set is divided into 10 folds with either 5 or 6 controls and either 6 or 7 cases. The sample size of each fold ranges from 11 to 13 individuals. The test set is not used in cross-validation and is only used once the final model has been fitted on the train set.

First, fold i is left out as a fold test and the remaining nine folds are used for training. Univariate Cox models, estimating the time T from sample collection to time of breast

cancer diagnosis as a function of methylation level at a single region, are fitted for each of the 101,955 methylation regions and regions with score statistics larger than θ are used to calculate principal components. Weighted Cox and weighted casebase models are fitted using j principal components and the linear predictor is calculated on fold i. A ratio of case series to base series of size 100 is used for the weighted casebase model. The linear predictor takes the form $\beta_1 \text{PC}_1 + \ldots + \beta_j \text{PC}_j$ for the weighted Cox model and the form Intercept $+ \alpha \log(\text{time}) + \beta_1 \text{PC}_1 + \ldots + \beta_j \text{PC}_j$ for the weighted casebase model. A weighted concordance is calculated on fold i using the linear predictor and the mean is taken over all 10 concordances. This is repeated for each principal component and threshold combination. The 10-fold cross-validation is repeated 100 times since a single cross-validation replicate might be affected by fold partition.

The tuning parameters with the largest mean concordance in cross-validation is used to fit the final model with the full training data set. A weighted Cox and a weighted casebase model is fitted on the entire train set with npc_{Cox} principal components and threshold θ_{Cox}, and npc_{cb} principal components and threshold θ_{cb}, respectively, then tested on the test set. A ratio of case series to base series of size 100 is used for the weighted casebase model. A weighted concordance (2.6) is used to assess the model's ability to discriminate between cases and controls in both the train and test sets. A weighted Brier score (2.6) is used to assess the accuracy of the predicted probabilities of breast cancer diagnosis. The predicted probabilities of breast cancer diagnosis during follow-up are calculated in both the train and test sets and plotted against weighted Kaplan-Meier estimates of risk to assess calibration.

5.3 Results

The tuning parameters chosen by 10-fold cross-validation is 2 principal components and a threshold of 1.89 for both the weighted Cox and the weighted casebase models, resulting in a total of 24 methylation regions. Table 5.1 shows the weighted concordances from

weighted Cox and casebase models fitted on the train set in cross-validation consisting of samples from batches 1 to 7 and predicted on the test set consisting of samples from batch 8. The Cox and casebase models have identical concordances in the train set and the casebase model has a slightly larger concordance in the test set.

Table 5.1: Weighted concordances from weighted Cox and casebase models

	Train[1]	Test
Cox	0.70	0.65
casebase	0.70	0.68

[1] Mean concordance from repeated cross-validation.

Figure 5.1 shows weighted Brier scores over follow-up time for the weighted Cox and casebase models calculated in the train set (A) and in the test set (B). The Brier scores are generally smaller for the Cox model than for the casebase model in the train set except for a few time points. In the test set, the Brier scores nearly overlap for the two models. The Brier scores in the test set are larger than those calculated in the train set, indicating again that both models are overfitted.

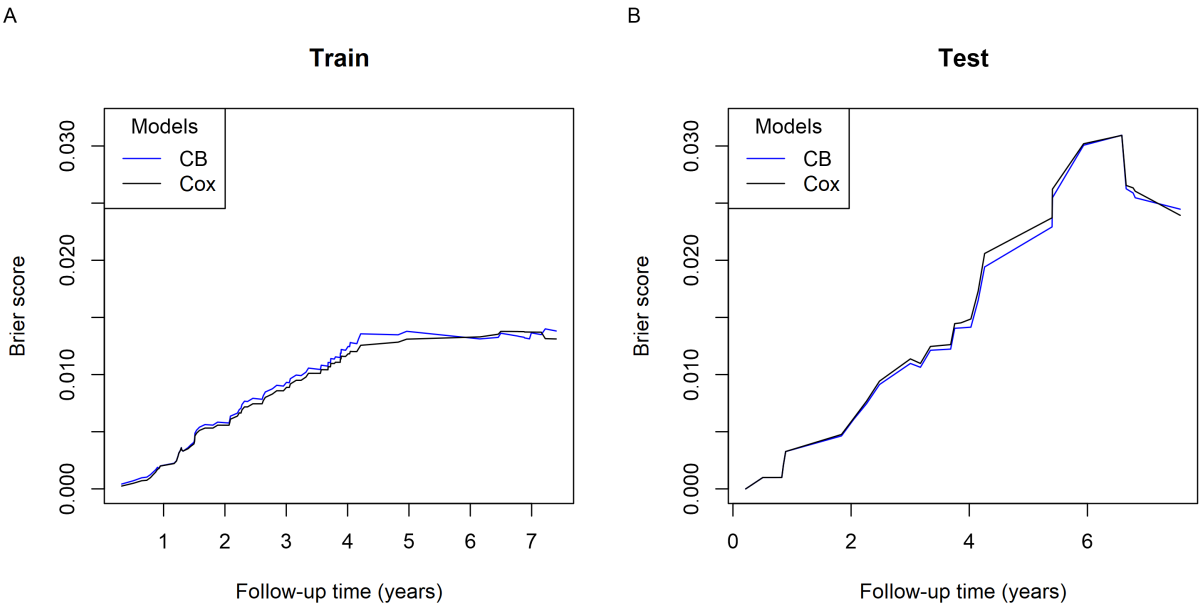

Figure 5.1: Weighted Brier score on train and test sets
(A) Train set. (B) Test set. CB = casebase.

Figure 5.2 shows the predicted probabilities of breast cancer diagnosis calculated from the casebase and Cox models, and Kaplan-Meier estimates. The train and test sets were divided into two groups based on the median of the linear predictor from the weighted Cox model, risks were estimated for the individuals of each group, and the mean probabilities for each of the two groups are shown in the figure. The fitted models overestimate risk when comparing to the Kaplan-Meier in the above median groups, especially noticeable in the test set as shown in Figure 5.2(B). Beyond 4 years of follow-up in the train set in Figure 5.2(A), the Kaplan-Meier curve in the above median group quickly jumps to a risk of 1 due to the lack of censored individuals beyond that time point. Using the median of the linear predictor from the weighted casebase model does not change Figure 5.2(A) and does not affect conclusions drawn from 5.2(B).

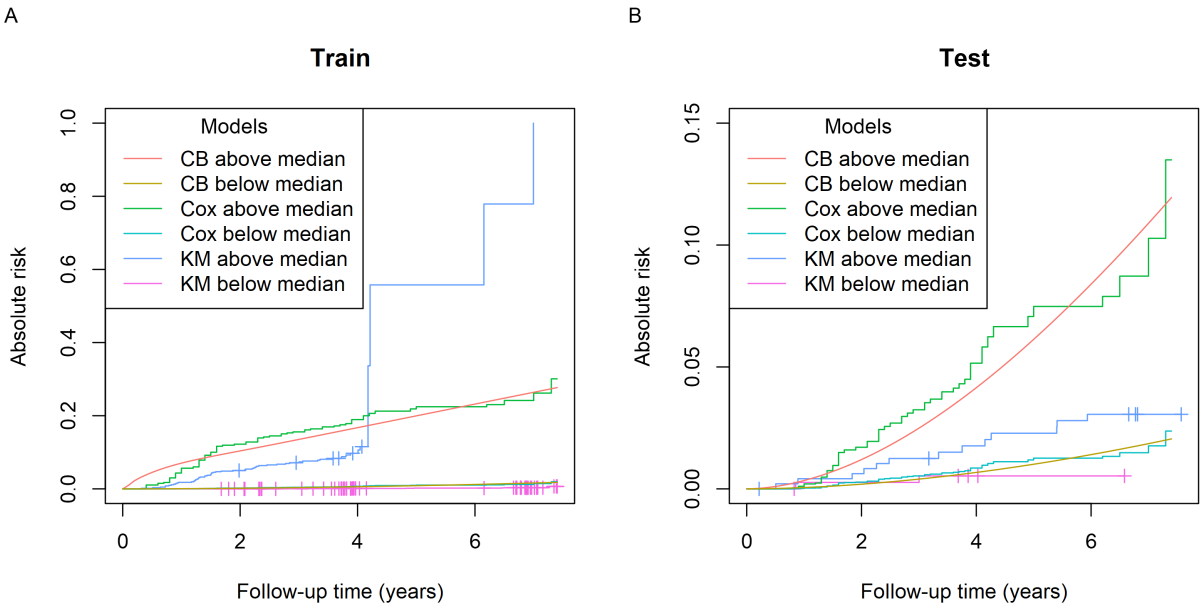

Figure 5.2: Absolute risk curves on train and test sets
(A) Train set. (B) Test set. CB = casebase. KM = Kaplan-Meier.

5.4 Ethics approval

Ethics approval of protocols to use the plasma samples from the OHS was obtained by Cheng et al., 2023 (protocol #34088). The full statement is included here:

> Patient plasma samples were obtained from the Ontario Health Study (OHS) with protocols approved by the University of Toronto Health Sciences research ethics board (protocol #34088). All participants gave written informed consent prior to participation. All samples and participant data were deidentified and assigned unique research IDs. Original OHS participant and CCO IDs are not known to anyone outside the research group. Supplementary tables do not contain any information that enables identification of the original participants.

Chapter 6

Discussion

This thesis proposed a weighted casebase framework that aimed to provide an alternative to existing weighted and robust Cox methods in the context of case-cohort data. Although the Cox proportional hazards model is popular due to the unspecified nature of the baseline hazard, it requires the proportional hazards assumption which may not always be satisfied, and for quantities like the absolute risk, will require a separate estimation of the baseline hazard. The existing casebase framework, besides being fully parametric and thus not needing a two-step estimation, can be easily implemented in software that support logistic models. The proposed framework is an extension of the existing one with the addition of weights suitable for biased sampling study designs and bootstrapped standard errors for the coefficient estimates.

Chapter 3 showed that simply supplying weights to the likelihood provides coefficient estimates with small bias but the mean standard errors of these estimates will underestimate the true standard errors. Bootstrapping provides an improved standard error estimate, with coverage probabilities closer to a nominal probability of 0.95. The simulations in this chapter showed that weights can be implemented through modification of casebase sampling or through addition of weights to glm. Both methods had similarly small bias and good coverage probabilities. However, the weights in glm method (weighting the likelihood) provide smaller mean bootstrapped and empirical standard errors.

Chapter 4 showed a simulation study with generated cohort and case-cohort data under varying covariate distributions and cohort sizes. The proposed framework provided coefficient estimates with small bias and good coverage probabilities in settings with a single binomial covariate. Although increasing the ratio parameter did not dramatically shrink the model-based or bootstrapped standard errors, consistent with Hanley and O. S. Miettinen, 2009 that a ratio of 100 is sufficient, the standard errors do tend towards the Cox estimates and could potentially achieve identical estimates at the cost of computational resources and time.

Under the single normal covariate setting, the proposed framework and the robust methods in general show poor performance as shown by the biased coefficient estimates and/or the underestimated standard errors. Particularly, control sampling probabilities of 0.004 and 0.002 provide both large bias and underestimated standard errors, possibly because the sampling weights are too large which could lead to model instability.

In contrast, the single log-normal covariate setting resulted in coefficient estimates with smaller bias than under the normal setting. Additionally, the bootstrapped standard error estimates are close to the empirical estimates under cohort sizes of $N = \{5,000, 10,000\}$ unlike the normal setting. While the coefficient estimates show small bias under cohort sizes of $N = \{50,000, 100,000\}$, the standard errors are underestimated, resulting in poor coverage probabilities. Strangely, the mean bootstrapped standard errors on the case-cohort data are smaller than the mean model-based standard errors on the full cohort data, indicating that there may have been issues in model fitting under these simulation settings. The Cox model on the full cohort has poor coverage owing to the moderate bias of the coefficient estimates, possibly due to violation of the proportional hazards assumption. It is likely that the casebase models on the full cohort would have had similar poor coverage probabilities if the associated standard errors were not so large, particularly compared to those of the Cox models. The bootstrapped standard errors of the ratio 100 and 200 robust models are smaller than the standard errors of the full models under $N = 50,000$ and similarly for all casebase robust models under $N = 100,000$.

The performance of the proposed framework is similar to that of robust Cox. The bias of the coefficient estimates are nearly identical in all the simulations with slightly smaller bias under cohort sizes of $N = \{10,000, 50,000, 100,000\}$ with a single log-normal covariate. However, robust Cox provides smaller standard errors than the proposed bootstrapped standard errors. Weighted casebase also has a longer computation time due to the base series sampling. The bootstrapped standard errors tend towards those of the robust Cox as the ratio parameter increases. Increasing the ratio parameter from 100 to 500 only slightly decreased the standard error, indicating that a ratio parameter larger than 500 will be needed to achieve smaller standard errors than those from Cox and to achieve narrower confidence intervals.

The simulation study also demonstrated the importance of including sampling weights in the analysis of case-cohort data. The naive models fitted on case-cohort data but without weights showed poor performance for both Cox and casebase. The coefficient estimates have large bias, resulting in very poor coverage probabilities. Although the mean model-based standard errors are close to their respective empirical standard errors, these standard error estimates are incorrect. Without inclusion of weights, the data are treated as if the included individuals are a representative sample of the population. This is not the case, however, since cases were purposely over-sampled. As a result, the distributions of the case-cohort sample and the full cohort will not be the same since selection of the case-cohort sample depended on event status (Pfeffermann, 1996). Generalizations to the full cohort and the general population may be incorrect and misleading.

Overall, the simulation study shows the proposed framework works under certain cohort sizes and covariate distributions. A major limitation of the framework is the bootstrapped standard error. A better standard error estimate is needed, particularly for the normal and log-normal settings. Although the stratified bootstrapping did not cause any problems under the binomial setting, it is computationally expensive, requiring 1000 bootstrap replicates for a single standard error estimate and several days of computation. An ideal standard error estimate would be model-based such that additional separate

model fitting, as in the bootstrapping method, is not necessary and will reduce computation time and memory usage.

The use of stratified bootstrapping ensured an equal number of cases and controls were included in each bootstrap replicate. With non-stratified bootstrapping, this would not necessarily be true and different bootstrap replicates would have more or fewer cases and controls. It would be of interest to investigate whether the standard errors obtained under non-stratified bootstrapping would be larger due to the increased variability in the number of cases and controls. Survey methodology often deals with weighted analysis to account for sampling design and non-response bias. There are variance estimators for designs with weighting that could be adapted for the case-cohort design and further investigation of such literature is needed (Pfeffermann, 1993; Pfeffermann, 1996; Skinner and Mason, 2012). There also exist survey resampling methods that use stratified bootstrapping and which suggest additional scaling of sampling weights may be needed for the currently implemented bootstrapped standard errors (Rao, Wu, and Yue, 1992; Rao and Wu, 1988; Kovar, Rao, and Wu, 1988; Kolenikov, 2010). A better standard error estimate, ideally robust to small control sampling probabilities and choice of covariate, could be derived from methods described in survey methodology literature.

Chapter 5 of this thesis illustrated the proposed framework on high-dimensional methylation data from the Ontario Health Study and incidence of breast cancer. The dataset was reduced from almost 110,000 methylation regions to 2 principal components calculated on 24 regions. Weighted Cox and casebase models were fitted on the train set, consisting of the first 7 batches, then tested on the test set, consisting of batch 8. Both models had a weighted concordance of 0.70 in the train test. The weighted casebase model had a weighted concordance of 0.68 in the test set, slightly better than the 0.65 of the weighted Cox model. The concordances in the train set indicate good discrimination between cases and controls but the decreases in the test set suggests model overfitting. Figure 5.1 shows Brier scores over follow-up time calculated in (A) the train set and (B) the test set. The Brier scores from both models overlap and are close to each other, with slightly smaller

scores from the Cox model. Thus, the predictive accuracy of the two models are similar with slightly better accuracy from the Cox model. The Brier scores are overall smaller in the train set than in the test set, another sign that the models overfit the train set. A similar conclusion can be drawn from the absolute risk figures.

Figure 5.2(A) shows the curves estimated on the train set. Both models overestimate the risk probability in the above median group prior to 4 years of follow-up, before the Kaplan-Meier estimates rapidly increase due to lack of censored individuals in the above median group. If there were still censored individuals in the risk set after 4 years of follow-up, the Kaplan-Meier estimates would not be so large. Nearly all controls were in the below median group, showing that the median linear predictor was able to almost perfectly separate cases from controls in the train set and another indicator of overfitting. Figure 5.2(B) shows the absolute risk curves estimated on the test set. Both models have estimates consistently larger than the Kaplan-Meier estimates but are similar to each other. Considering the incidence rate of breast cancer in Ontarian females in 2022 is projected to be 208.1 cases per 100,000 females and 383.1 cases per 100,000 females for ages 40 to 59 and ages 60 to 79, respectively (Ontario Health (Cancer Care Ontario), 2022), the models overestimate the risk probabilities, with predicted probabilities going up to 0.024. In the test set, there are no large jumps in the Kaplan-Meier estimates since there are censored individuals until the end of follow-up in the above median group. Figure 5.2 shows the benefit of using casebase to estimate the predicted probabilities compared to Cox: the estimates are continuous over follow-up time, similar to how the theoretical probabilities are continuous over time (Kleinbaum and Klein, 2012b). It is also easier for clinicians to interpret continuous estimates than step function estimates. The models did not not have good discrimination of cases and controls in the test set, based on their weighted concordances. Panel B also shows the overestimation of risk in the below groups compared to the Kaplan-Meier estimates, particularly beyond 4 years of follow-up. In short, the fitted Cox and casebase models were overfitted to the train set, resulting in poorer performance

in the test set. Additionally, both models greatly overestimated the risk of breast cancer in both the train and test sets.

The somewhat poor performance and overestimation of risk in these models indicate that improvements are necessary before conclusions and implications of the results can be drawn. First, supervised principal components only include regions with large enough association to the outcome. The resulting principal components may describe the train set well but overlook regions that would otherwise be important in the test set. Since the train set is small and the methylation data contain many zeroes, the score statistic may be sensitive enough that removing or adding an individual could be the difference between including a region or not in the principal components. While the cross-validation results showed best concordance with 2 principal components, additional models could be fitted with more than 2 principal components. Alternatively, unsupervised principal components could be used, without considering association to the outcome of the train set. Additionally, alternative splitting of the train and test sets could be used. Although the same protocols were used on all the batches and no obvious differences were observed, the two sets could be drawn from combining all the batches together. Stratified sampling could be used to ensure the age distribution, time of diagnosis since plasma collection, length of follow-up, and proportion of cases are similar in the two groups. Another consideration is the small sample size of the study. There were a total of 150 individuals, much smaller than the number of methylation regions considered. The analysis of the dataset would greatly benefit from having more samples. However, increasing the sample size is costly and challenging since it requires finding incident cases in a limited cohort size.

Chapter 7

Conclusion

This thesis is primarily concerned with the proposal of a weighted casebase framework in the context of case-cohort data. The original casebase framework, proposed as an alternative to the Cox proportional hazards model, did not have any methods for weights, unlike the Cox model. An initial simulation study in Chapter 3 showed good properties of the method. Chapter 4 showed the results of a simulation study that showed the proposed framework is comparable to weighted and robust Cox methods, and demonstrated the importance of including weights in analyses to avoid biased and incorrect estimates. The simulation study also showed the proposed framework is sensitive to control sampling probabilities and the choice of covariate distributions. This also holds true for the weighted Cox models. Chapter 5 applied weighted Cox and casebase models to predict the risk of developing breast cancer using methylated regions in pre-cancer plasma samples from the Ontario Health Study. Considerable fine-tuning of these models is necessary before making conclusions or generalizing results. However, it appears that there are signals that can be detected in an individual's blood DNA methylation prior to diagnosis by common screening methods. Early diagnosis would greatly benefit quality of life and medical care with earlier intervention.

Further work is needed to obtain a better standard error estimator that is computationally inexpensive and robust to sampling weights and covariate distributions. This

will greatly improve the proposed framework and could be an attractive alternative to weighted Cox methods.

Bibliography

Adragni, Kofi P. and R. Dennis Cook (2009). "Sufficient dimension reduction and prediction in regression". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 367.1906. Publisher: Royal Society, pp. 4385–4405.

Bair, Eric et al. (2006). "Prediction by Supervised Principal Components". In: *Journal of the American Statistical Association* 101.473. Publisher: [American Statistical Association, Taylor & Francis, Ltd.], pp. 119–137.

Barlow, W. E. (1994). "Robust variance estimation for the case-cohort design". In: *Biometrics* 50.4, pp. 1064–1072.

Barrett, David and Helen Noble (2019). "What are cohort studies?" In: *Evidence-Based Nursing* 22.4. Publisher: Royal College of Nursing Section: Research made simple, pp. 95–96.

Bender, Ralf, Thomas Augustin, and Maria Blettner (2005). "Generating survival times to simulate Cox proportional hazards models". In: *Statistics in Medicine* 24.11, pp. 1713–1723.

Benjamini, Yoav and Yosef Hochberg (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 57.1, pp. 289–300.

Bhatnagar, Sahir Rai et al. (2022). "casebase: An Alternative Framework for Survival Analysis and Comparison of Event Rates". In: *The R Journal* 14.3, pp. 59–79.

Bickel, P. J. and D. A. Freedman (1984). "Asymptotic Normality and the Bootstrap in Stratified Sampling". In: *The Annals of Statistics* 12.2. Publisher: Institute of Mathematical Statistics, pp. 470–482.

Brenner, Darren R. et al. (2022). "Projected estimates of cancer in Canada in 2022". In: *CMAJ* 194.17. Publisher: CMAJ Section: Research, E601–E607.

Breslow, N. E. (1972). "Contribution to the Discussion of the paper by D.R. Cox". In: *Journal of the Royal Statistical Society, Series B* 34, pp. 187–220.

Brier, Glenn W. (1950). "Verification of Forecasts Expressed in Terms of Probability". In: *Monthly Weather Review* 78.1. Publisher: American Meteorological Society Section: Monthly Weather Review, pp. 1–3.

Canadian Partnership Against Cancer (2018). *Breast Cancer Screening in Canada: Environmental Scan*. Toronto, ON: Canadian Partnership Against Cancer.

— (2020). *Pan-Canadian Framework for Action to Address Abnormal Call Rates in Breast Cancer Screening*. Toronto: Canadian Partnership Against Cancer.

Chen, Xingdong et al. (2020). "Non-invasive early detection of cancer four years before conventional diagnosis using a blood test". In: *Nature Communications* 11.1. Number: 1 Publisher: Nature Publishing Group, p. 3475.

Cheng, Nicholas et al. (2023). *Pre-diagnosis plasma cell-free DNA methylome profiling up to seven years prior to clinical detection reveals early signatures of breast cancer*. Pages: 2023.01.30.23285027.

Cox, D. R. (1972). "Regression Models and Life-Tables". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 34.2. Publisher: [Royal Statistical Society, Wiley], pp. 187–220.

Croce, Carlo M. (2008). "Oncogenes and Cancer". In: *New England Journal of Medicine* 358.5. Publisher: Massachusetts Medical Society, pp. 502–511.

Dedeurwaerder, Sarah et al. (2014). "A comprehensive overview of Infinium HumanMethylation450 data processing". In: *Briefings in Bioinformatics* 15.6, pp. 929–941.

Efron, B. (1979). "Bootstrap Methods: Another Look at the Jackknife". In: *The Annals of Statistics* 7.1. Publisher: Institute of Mathematical Statistics, pp. 1–26.

Ernster, V. L. (1994). "Nested Case-Control Studies". In: *Preventive Medicine* 23.5, pp. 587–590.

Fernandez-Garcia, Daniel et al. (2019). "Plasma cell-free DNA (cfDNA) as a predictive and prognostic marker in patients with metastatic breast cancer". In: *Breast Cancer Research* 21.1, p. 149.

Frommer, M et al. (1992). "A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands." In: *Proceedings of the National Academy of Sciences of the United States of America* 89.5, pp. 1827–1831.

Gail, Mitchell H. (2005). "Crude Risk". In: *Encyclopedia of Biostatistics*. John Wiley & Sons, Ltd.

Galardi, Francesca et al. (2020). "Cell-Free DNA-Methylation-Based Methods and Applications in Oncology". In: *Biomolecules* 10.12. Number: 12 Publisher: Multidisciplinary Digital Publishing Institute, p. 1677.

Graf, Erika et al. (1999). "Assessment and comparison of prognostic classification schemes for survival data". In: *Statistics in Medicine* 18.17, pp. 2529–2545.

Hanley, James A. and Olli S. Miettinen (2009). "Fitting Smooth-in-Time Prognostic Risk Functions via Logistic Regression". In: *The International Journal of Biostatistics* 5.1. Publisher: De Gruyter.

Harrell Jr, Frank E. et al. (1982). "Evaluating the Yield of Medical Tests". In: *JAMA* 247.18, pp. 2543–2546.

Higgins, Julian PT, Tianjing Li, and Jonathan J Deeks (2022). "Chapter 6: Choosing effect measures and computing estimates of effect". In: *Cochrane Handbook for Systematic Reviews of Interventions*. Version 6.3. Cochrane.

Johnstone, Iain M. and D. Michael Titterington (2009). "Statistical challenges of high-dimensional data". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 367.1906. Publisher: Royal Society, pp. 4237–4253.

Jolliffe, Ian T. and Jorge Cadima (2016). "Principal component analysis: a review and recent developments". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374.2065. Publisher: Royal Society, p. 20150202.

Kaplan, E. L. and Paul Meier (1958). "Nonparametric Estimation from Incomplete Observations". In: *Journal of the American Statistical Association* 53.282. Publisher: [American Statistical Association, Taylor & Francis, Ltd.], pp. 457–481.

Kirsh, Victoria A et al. (2022). "Cohort Profile: The Ontario Health Study (OHS)". In: *International Journal of Epidemiology*, dyac156.

Kis, Olena et al. (2017). "Circulating tumour DNA sequence analysis as an alternative to multiple myeloma bone marrow aspirates". In: *Nature Communications* 8.1, p. 15086.

Kleinbaum, David G. and Mitchel Klein (2012a). "Extension of the Cox Proportional Hazards Model for Time-Dependent Variables". In: *Survival Analysis: A Self-Learning Text*. Ed. by David G. Kleinbaum and Mitchel Klein. Statistics for Biology and Health. New York, NY: Springer, pp. 241–288.

— (2012b). "Introduction to Survival Analysis". In: *Survival Analysis: A Self-Learning Text*. Ed. by David G. Kleinbaum and Mitchel Klein. Statistics for Biology and Health. New York, NY: Springer, pp. 1–54.

— (2012c). "Parametric Survival Models". In: *Survival Analysis: A Self-Learning Text*. Ed. by David G. Kleinbaum and Mitchel Klein. Statistics for Biology and Health. New York, NY: Springer, pp. 289–361.

— (2012d). "The Cox Proportional Hazards Model and Its Characteristics". In: *Survival Analysis: A Self-Learning Text*. Ed. by David G. Kleinbaum and Mitchel Klein. Statistics for Biology and Health. New York, NY: Springer, pp. 97–159.

— (2012e). "The Stratified Cox Procedure". In: *Survival Analysis: A Self-Learning Text*. Ed. by David G. Kleinbaum and Mitchel Klein. Statistics for Biology and Health. New York, NY: Springer, pp. 201–240.

Klutstein, Michael et al. (2016). "DNA Methylation in Cancer and Aging". In: *Cancer Research* 76.12, pp. 3446–3450.

Kolenikov, Stanislav (2010). "Resampling Variance Estimation for Complex Survey Data". In: *The Stata Journal: Promoting communications on statistics and Stata* 10.2, pp. 165–199.

Kovar, J. G., J. N. K. Rao, and C. F. J. Wu (1988). "Bootstrap and Other Methods to Measure Errors in Survey Estimates". In: *The Canadian Journal of Statistics / La Revue Canadienne de Statistique* 16. Publisher: [Statistical Society of Canada, Wiley], pp. 25–45.

Kulathinal, Sangita et al. (2007). "Case-cohort design in practice – experiences from the MORGAM Project". In: *Epidemiologic perspectives & innovations : EP+I* 4, p. 15.

Kupper, L. L., A. J. McMichael, and R. Spirtas (1975). "A Hybrid Epidemiologic Study Design Useful in Estimating Relative Risk". In: *Journal of the American Statistical Association* 70.351. Publisher: [American Statistical Association, Taylor & Francis, Ltd.], pp. 524–528.

Kustanovich, Anatoli et al. (2019). "Life and death of circulating cell-free DNA". In: *Cancer Biology & Therapy* 20.8. Publisher: Taylor & Francis, pp. 1057–1067.

Lavallée, Pierre and Jean-François Beaumont (2015). "Why We Should Put Some Weight on Weights." In: *Survey Methods: Insights from the Field (SMIF)*.

Lumley, Thomas (2011). *Complex Surveys: A Guide to Analysis Using R*. Wiley. 296 pp.

Luo, Huiyan et al. (2021). "Liquid Biopsy of Methylation Biomarkers in Cell-Free DNA". In: *Trends in Molecular Medicine* 27.5, pp. 482–500.

Miettinen, Olli (1982). "Design options in epidemiologic research: An update". In: *Scandinavian Journal of Work, Environment & Health* 8. Publisher: Scandinavian Journal of Work, Environment & Health, pp. 7–14.

Morris, Tim P., Ian R. White, and Michael J. Crowther (2019). "Using simulation studies to evaluate statistical methods". In: *Statistics in Medicine* 38.11, pp. 2074–2102.

O'Brien, Katie M., Kaitlyn G. Lawrence, and Alexander P. Keil (2022). "The Case for Case–Cohort: An Applied Epidemiologist's Guide to Reframing Case–Cohort Studies to Improve Usability and Flexibility". In: *Epidemiology* 33.3, p. 354.

Ontario Health (Cancer Care Ontario) (2022). *Ontario Cancer Statistics 2022*. Toronto: Ontario Health (Cancer Care Ontario).

Pelizzola, Mattia et al. (2008). "MEDME: An experimental and analytical methodology for the estimation of DNA methylation levels based on microarray derived MeDIP-enrichment". In: *Genome Research* 18.10, pp. 1652–1659.

Pfeffermann, Danny (1993). "The Role of Sampling Weights When Modeling Survey Data". In: *International Statistical Review / Revue Internationale de Statistique* 61.2. Publisher: [Wiley, International Statistical Institute (ISI)], pp. 317–337.

— (1996). "The use of sampling weights for survey data analysis". In: *Statistical Methods in Medical Research* 5.3. Publisher: SAGE Publications Ltd STM, pp. 239–261.

Pfeiffer, Ruth M. and Mitchell H. Gail (2017). "Competing risks". In: *Absolute Risk*. Num Pages: 7. Chapman and Hall/CRC.

Porta, Miquel (2016). "Person-Time". In: *A Dictionary of Epidemiology*. Oxford University Press.

Poulet, Geoffroy, Joséphine Massias, and Valerie Taly (2019). "Liquid Biopsy: General Concepts". In: *Acta Cytologica* 63.6. Publisher: Karger Publishers, pp. 449–455.

Prentice, R. L. (1986). "A case-cohort design for epidemiologic cohort studies and disease prevention trials". In: *Biometrika* 73.1, pp. 1–11.

Rao, J. N. K. and C. F. J. Wu (1988). "Resampling Inference With Complex Survey Data". In: *Journal of the American Statistical Association* 83.401. Publisher: [American Statistical Association, Taylor & Francis, Ltd.], pp. 231–241.

Rao, J. N. K., C. F. J. Wu, and K. Yue (1992). "Some recent work on resampling methods for complex surveys". In: *Survey Methodology*.

Saarela, Olli (2016). "A case-base sampling method for estimating recurrent event intensities". In: *Lifetime Data Analysis* 22.4, pp. 589–605.

Saarela, Olli and Elja Arjas (2015). "Non-parametric Bayesian Hazard Regression for Chronic Disease Risk Assessment: Bayesian regression for risk assessment". In: *Scandinavian Journal of Statistics* 42.2, pp. 609–626.

Schwarzenbach, Heidi, Dave S. B. Hoon, and Klaus Pantel (2011). "Cell-free nucleic acids as biomarkers in cancer patients". In: *Nature Reviews Cancer* 11.6. Number: 6 Publisher: Nature Publishing Group, pp. 426–437.

Shen, Shu Yi et al. (2018). "Sensitive tumour detection and classification using plasma cell-free DNA methylomes". In: *Nature* 563.7732. Number: 7732 Publisher: Nature Publishing Group, pp. 579–583.

Singal, Rakesh and Gordon D. Ginder (1999). "DNA Methylation". In: *Blood* 93.12, pp. 4059–4070.

Skinner, Chris and Ben Mason (2012). "Weighting in the regression analysis of survey data with a cross-national application". In: *The Canadian Journal of Statistics / La Revue Canadienne de Statistique* 40.4. Publisher: [Statistical Society of Canada, Wiley], pp. 697–711.

Soave, David and Jerald F. Lawless (2023). "Regularized regression for two phase failure time studies". In: *Computational Statistics & Data Analysis* 182, p. 107703.

Therneau, Terry M. and Patricia M. Grambsch (2000). "Influence". In: *Modeling Survival Data: Extending the Cox Model*. Ed. by Terry M. Therneau and Patricia M. Grambsch. Statistics for Biology and Health. New York, NY: Springer, pp. 153–168.

Thierry, A. R. et al. (2016). "Origins, structures, and functions of circulating DNA in oncology". In: *Cancer Metastasis Reviews* 35.3, pp. 347–376.

Thu, Kelsie L. et al. (2009). "Methylated DNA Immunoprecipitation". In: *Journal of Visualized Experiments : JoVE* 23, p. 935.

White, Ian R. (2010). "Simsum: Analyses of Simulation Studies Including Monte Carlo Error". In: *The Stata Journal: Promoting communications on statistics and Stata* 10.3, pp. 369–385.

Woloshin, Steven, Lisa M. Schwartz, and H. Gilbert Welch (2008). "Risk Charts". In: *Know Your Chances: Understanding Health Statistics*. University of California Press.

Yong, Wai-Shin, Fei-Man Hsu, and Pao-Yang Chen (2016). "Profiling genome-wide DNA methylation". In: *Epigenetics & Chromatin* 9.1, p. 26.