MARGINAL STRUCTURAL COX MODELS WITH CASE-COHORT SAMPLING

Hana Lee

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

Chapel Hill
2013

Approved by:

Dr. Michael G Hudgens

Dr. Jianwen Cai

Dr. Stephen R Cole

Dr. Danyu Lin

Dr. Donglin Zeng

ABSTRACT

HANA LEE: MARGINAL STRUCTURAL COX MODELS WITH CASE-COHORT SAMPLING
(Under the direction of Drs. Dr. Michael G Hudgens and Dr. Jianwen Cai)

A common objective of biomedical cohort studies is assessing the effect of a time-varying treatment or exposure on a survival time. In the presence of time-varying confounders, marginal structural models fit using inverse probability weighting can be employed to obtain a consistent and asymptotically normal estimator of the causal effect of a time-varying treatment. This document considers estimation of parameters in the semiparametric marginal structural Cox model (MSCM) from a case-cohort study. Case-cohort sampling entails assembling covariate histories only for cases and a random subcohort, which can be cost effective, particularly in large cohort studies with low outcome rates. Following Cole et al. [2012], we consider estimating the causal hazard ratio from a MSCM by maximizing a weighted-pseudo-partial-likelihood. The estimator is shown to be consistent and asymptotically normal under certain regularity conditions. Computation of the estimator using standard survival analysis software is discussed and results from a simulation study are presented.

In the standard (associational) case-cohort Cox analysis, various methods have been proposed to improve efficiency from maximum pseudolikelihood estimators of Prentice [1986*a*] or Self and Prentice [1988]. As the presented theory of MSCM parameter estimator is developed based on Self and Prentice [1988] we briefly review those methods and discuss extension of the methods to the MSCM analysis. In addition, we proposed a new method to improve efficiency of the case-cohort MSCM analysis from a biomedical study that aims to evaluate the causal effect of treatment on a time to event. We seek to improve the efficiency by multiple imputation method which can make fuller use of covariate information that are available from full cohort. The proposed method is applied to the Multicenter AIDS Cohort Study (MACS) and the Women's Interagency HIV Study (WIHS).

ACKNOWLEDGMENTS

I would like to express my deepest appreciation to my advisor, Dr. Michael G. Hudgens who arouse my passion toward research and taught me how joyful research is. Working with him was certainly the best thing ever happened to me at Chapel Hill. I can hardly imagine my dissertation without his endless support and insightful guidance. I sincerely thank him for everything. I dare to say that he is the best advisor in the universe.

My second deepest appreciation goes to my co-advisor Dr. Jianwen Cai. She is my role model as a female biostatistician in academia. I started dreaming of working in academia while working with her. She has the substance of genius, and is full of responsibility. My dissertation is indebted to her brilliance. Someday, I also want to be a great role model for peers and students.

Without my advisors, this dissertation would not have been completed. They have guided me to the right path whenever I lost a direction. During the past two years, they were like parents holding their childs hands and teaching how to walk. Like a toddler, I had fears and pains for the first few steps. However, every step with my advisors eventually became my priceless properties. I will never forget the moments walking together.

I would like to thank my wonderful committee members, Drs. Stephen R. Cole, Danyu Lin, and Donglin Zeng as well. I cannot imagine any better committee members. Dr. Cole's brilliant idea initiated my dissertation and his support with the Multicenter AIDS Cohort and the Women's Interagency HIV studies dataset completed my dissertation. Besides his expertise in research, he is also a warmhearted person who was very supportive to my dissertation. Dr. Lin's comments and suggestions brought invaluable developments to my first paper. After my preliminary exam, Dr. Zeng provided me a brilliant idea which was further developed into my last project. I am very grateful to these committee members for spending their time regardless of their very busy schedule.

I am also thankful to my dear friends who shared both painful and joyful moments with me during my PhD years. I cannot thank enough to my family in Korea, but they were

always with me in my heart.

Last but not least, I dedicate my thesis to my love of life, Wonyul Lee.

TABLE OF CONTENTS

LIST OF TABLES . . . . . **viii**

1 Introduction **1**

2 Literature Review **4**

2.1 Cox Models 4

2.2 Causal Inference and Marginal Structural Cox Models 7

2.2.1 Causal Inference and Potential Outcomes 8

2.2.2 Marginal Structural Cox Models 11

2.3 Cox Models with Case-cohort Sampling 14

2.4 Statistical Methods to Improve Efficiency 16

3 Marginal Structural Cox Models with Case-cohort Sampling **19**

3.1 Introduction 19

3.2 Marginal Structural Cox Model Estimators 21

3.2.1 Notation, Assumptions, and Model 21

3.2.2 Inverse Probability Weights 24

3.2.3 Weighted-Psuedo-Partial-Likelihood 26

3.3 Consistency 28

3.4 Asymptotic Normality 37

3.5 Implementation and Simulation 47

3.5.1 Implementation 47

3.5.2 Simulation 50

3.6 Supplemental Material 52

4 Efficient Inference of Case-Cohort Marginal Structural Cox Models . . **62**

4.1 Introduction 62

4.2 General Methods for MSCM Case-Cohort Estimators 65

4.3 Improving Efficiency of the Estimation 68

4.3.1 Time-Varying Inverse Sampling Weights 69

4.3.2 Imputation Method 70
4.4 Results 75
4.4.1 Simulation Studies 75
4.4.2 Real Data Analysis 78
4.5 Discussion 84
5 Summary and Future Research 86
BIBLIOGRAPHY 87

LIST OF TABLES

3.1 Summary of simulation study 51

4.1 Simulation studies to compare performance of estimators 77

4.2 Simulation studies to compare performance of estimators when $\alpha = .1$ 79

4.3 Simulation studies to compare performance of estimators when $\alpha = .2$ 80

4.4 Simulation studies to compare performance of estimators when $\alpha = .3$ 81

4.5 Baseline characteristics of the full and the 50% subcohort subjects 83

4.6 Full cohort, 20% subcohort with MI, and case-cohort MSCM analyses 83

4.7 Sensitivity analysis of case-cohort and multiple imputation 84

Chapter 1

Introduction

Biomedical cohort studies are often conducted with the goal of assessing the effect of a time-varying treatment (or exposure) on a survival time. In such studies there may exist time-dependent covariates which are simultaneously (i) confounders and (ii) affected by prior treatment on the causal pathway from treatment to disease. In the presence of time-varying confounders affected by prior treatment, standard methods such as Cox regression modeling with time-varying covariates do not in general yield consistent estimators of the causal effect of treatment [Robins, 1986, 1998; Robins and Rotnitzky, 1992; Hernán, Brumback and Robins, 2001]. On the other hand, marginal structural models (MSMs) fit using inverse probability weighting can be employed to obtain consistent estimators of the causal effect of a time-varying treatment on an outcome of interest, even if there are time-varying confounders affected by prior treatment [Robins, 1999].

For example, consider the Multicenter AIDS Cohort Study (MACS), an observational study of HIV-positive homosexual men. Using data from MACS, Hernán, Brumback and Robins [2001] showed that (i) current CD4 count and *Pneumocystis carinii* pneumonia (PCP) status were independent risk factors for death and were predictive of subsequent treatment with zidovudine (AZT) and prophylaxis therapy (i.e., confounders), and (ii) prophylaxis therapy was a protective risk factor for the development of PCP subsequently. Thus, to assess the effect of AZT and prophylaxis therapy on mortality in MACS, a method is required that can appropriately account for time-varying confounders affected by prior treatment (in particular, PCP status). Applying standard (i.e., unweighted) Cox regression with time-dependent covariates to the MACS data, Hernán, Brumback and Robins [2001] reported an estimated hazard ratio of 1.85 (95% CI 1.49, 2.30) for AZT users versus

nonusers, suggesting that treatment increases the risk of death in HIV-positive homosexual men, contrary to results from randomized clinical trials. On the other hand, fitting a marginal structural Cox model (MSCM) with inverse probability weighting yielded an estimated hazard ratio for AZT of 0.67 (95% CI 0.46, 0.98), in agreement with results from randomized trials of AZT. The difference in hazard ratio estimates between the unweighted Cox regression model and the MCSM with inverse probability weighting is not surprising given the aforementioned established results about the (in)consistency of these estimators in the presence of time-varying confounders affected by prior treatment.

Recently, Cole et al. [2012] considered fitting MSCMs via inverse probability weighting in the presence of case-cohort sampling. The case-cohort study design is a cost-efficient approach to estimate treatment effects in large cohorts with low event rates, when treatment or covariate information is expensive. The design entails randomly selecting a subcohort from the entire cohort. Covariate information is then collected only from the random subcohort and from individuals that are observed to experience an event (i.e., cases), saving cost and effort relative to obtaining covariate information from the full cohort. In addition to being cost efficient, the case-cohort design enjoys other benefits. For instance, the subcohort can serve as a basis for real time covariate monitoring during the course of the study. Also, because the subcohort is chosen randomly, survival times to different diseases can be analyzed using the same subcohort [Self and Prentice, 1988].

In the presence of case-cohort sampling, Cole et al. [2012] considered estimating the causal hazard ratio of a MSCM via inverse probability weighting. Simulation studies indicated the estimator proposed by Cole et al. [2012] can perform well empirically, however no formal justification for their estimator has been developed to date. Therefore, following Cole et al. [2012], we consider estimating the causal hazard ratio of a MSCM via inverse probability weighting in case-cohort studies and establish consistency and asymptotic normality for the estimator that maximizes a weighted-pseudo-partial-likelihood (WPPL) under certain regularity conditions.

The approach utilized in this proposal entails standard counting process and martingale theory. Using this formulation readily enables practical implementation of the methods using existing survival analysis software. Framing the problem using counting processes may also be helpful in future work, e.g., in fitting MSCMs to data from nested case-control

studies or in the presence of competing risks. In the special situation that the subcohort equals the full cohort, the proposed inverse probability weighted estimator is asymptotically equivalent to the estimator in Robins [1999]. Moreover, in this case our proof gives an alternative consistency and normality proof to Robins [1999], who did not utilize the usual counting process framework.

The outline of the remainder of this document is as follows. Chapter 2 begins with an introduction to methods for survival analysis primarily focusing on Cox models. A review of case-cohort studies is next. Then we introduce MSCMs on the basis of causal inference and potential outcome framework. This Chapter concludes with a review of some statistical methods devised to improve efficiency in the standard Cox regression analysis with case-cohort sampling. In Chapter 3 the estimator of the hazard ratio of a MSCM in the presence of case-cohort sampling is introduced, and proofs of consistency of the parameter estimators under the full and the case-cohort settings are shown. Also, we establish full distributional theories of the parameter estimators under the full cohort and the case-cohort settings in the same Chapter. How to implement a MSCM using existing software such as R or SAS is described in Chapter 3.5, along with the simulation study results. Details to show asymptotic distributional theory of the case-cohort MSCM parameter estimate are provided in 3.6. In Chapter 4 we propose a new method that can improve efficiency in the case-cohort MSCM analysis. We start from a review of general methods for MSCM case-cohort estimators, including our proposed methods introduced in Chapter 3, and demonstrate why the discussed methods devised to improve efficiency in the standard case-cohort Cox regression analysis may not be applicable to the causal setting. We propose a new method which aim to utilize all subject in the estimation and show numerical study results. The proposed method is applied to a real observational HIV study data composed of two data sets, the Multicenter AIDS Cohort Study and the Women's Interagency HIV Study.

Chapter 2

Literature Review

2.1 Cox Models

Here, we assume a study consists of n unique individuals who are indexed by $i = 1, ..., n$. Let T_i denote failure (or survival) time of a subject i in a study, where $T = 0$ represents the time of initiation of follow-up and τ represents study end point. We essentially assume that the failure time is on continuous basis. Let C_i denote the time of censoring and $X_i = \min(T_i, C_i)$ denote the observed time from the subject i. $\delta(X_i) = I\{T_i < C_i\}$ is an event indicator where $I\{\cdot\}$ is an usual indicator function. In addition, let $p \times 1$ vector of $Z_i(X_i) = (Z_{1i}(X_i), ..., Z_{pi}(X_i))$ denote time-dependent covariates information collected from the subject i. Throughout we assume $(X_i, \delta_i, Z_i)(i = 1, ..., n)$ be n independent replicates of (T, δ, Z) that Z is bounded. Also, let $N_i(t)$ be a stochastic process which denote the number of failures of subject i by time t. We use the notation $dN_i(t)$ to indicate the number of events of the subject i occurred in $[t, t + dt)$ for sufficiently small dt. Since failures occur in continuous time, we only allow jumps of size 1 and no simultaneous jumps can occur in $[t, t + dt)$ for the process $N_i(t)$. Let $Y_i(t) = I\{T_i \geq t, C_i \geq t\}$ denote whether an individual is still alive and being able to be observed (to fail) at time t, having a left-continuous sample paths. This process is called "at risk" process. Then the data for the ith participant (X_i, δ_i, Z_i) can be rewritten as $\{N_i(u), Y_i(u), Z_i(u) : 0 \leq u \leq t\}$.

In biomedical studies, we are often interested in identifying/quantifying risk (or prognostic) factors related to response. Cox regression models, including Cox proportional hazards models, introduced by Cox [1972] are the most commonly used approach to explore (or adjust) for the effect of covariates that may be associated with that outcome. Let $\lambda(t|Z(t))$

denote the hazard (or risk) of being failed associated with $Z(t)$, i.e.,

$$\lambda(t|Z(t)) = \lim_{dt \to 0} \Pr(t \leq T \leq t + dt | T \geq t, Z(t))/dt.$$

Then Cox models are given by

$$\lambda(t|Z(t)) = \lambda_0(t) \exp\{\beta_0' Z(t)\} \tag{2.1}$$

where $\lambda_0(t|Z(t))$ is an unknown baseline hazard and $\beta_0 = (\beta_{0,1}, ..., \beta_{0,p})$ is a set of unknown regression parameters. $\lambda_0(\cdot)$ describing how the hazard changes over time at baseline levels of covariates, i.e., $Z(t) = 0$ for all $t \geq 0$. β_0 describes the effect of covariates on the hazard changes over time. Under this model, we can compare two hazards under different covariate levels (e.g., treated or untreated) in the logarithm scale. For instance, consider two observations i and i^* that differ in their covariate values at time t by $Z(t)$ and Z^* respectively. Then the hazard ratio for two observations is

$$\exp\{\beta_0' Z(t))\} / \exp\{\beta_0' Z^*(t)\} = \exp\{\beta_0'(Z(t) - Z^*(t))\},$$

and therefore the log of the hazard ratio $\beta_0'(Z(t) - Z^*(t))$ can be explained by the parameter β_0. When $Z(t) \equiv Z$ for all $t \geq 0$ then this models are also referred to as *the proportional hazards models* since the hazard ratio at any time t is independent of time t. Explicitly, the hazard ratio for two observations i and i^* in the above example is

$$\exp\{\beta_0' Z\} / \exp\{\beta_0' Z^*\} = \exp\{\beta_0'(Z - Z^*)\}$$

which is constant over time t.

$\lambda_0(\cdot)$ is an unknown function and parametric distributional assumptions such as uniform, exponential, weibull on $\lambda_0(\cdot)$ is available. Other than distributional assumption, a monotonic or step function assumption can also be made. However, Cox [1975] proposes a partial likelihood approach which enables to estimate the parameter of interest β_0 in (2.1) while the $\lambda_0(\cdot)$ remains unspecified. Consistent estimator for β_0 can be obtained by using the partial likelihood score function

$$U(\beta) = \sum_{i=1}^{n} \delta_i \{ Z_i(X_i) - \frac{S^{(1)}(\beta_0, X_i)}{S^{(0)}(\beta, X_i)} \} \tag{2.2}$$

where $S^{(k)}(\beta, X_i) = n^{-1} \sum_{i=1}^{n} Y_i(X_i) Z_i^{\otimes k}(X_i) \exp\{\beta' Z(X_i)\}$ for $k = 0, 1, 2$ under standard independent censoring assumption. Here, we define $a^{\otimes 0} \equiv 1$,$a^{\otimes 1} \equiv a$, and $a^{\otimes 2} = a'a$ which is defined by the $p \times p$ matrix with (i, j)th element $a_i a_j$ for $p \times 1$ vectors a. The maximum partial likelihood estimator $\hat{\beta}$, defined as the solution to the score equation $U(\hat{\beta}) = 0$, is shown to converge in distribution to Normal with mean zero and a covariance matrix which can consistently be estimated by $-\{\partial U(\beta)/\partial \beta|_{\beta=\hat{\beta}}\}^{-1}$ based on martingale formulation (Andersen and Gill [1982]). Using an integral representation, log-likelihood function corresponding to (2.2) can be written by

$$l(\beta) = \sum_{i=1}^{n} \int_0^{\tau} \beta' Z_i(u) - \log[\sum_{l=1}^{n} Y_l(u) \exp\{\beta' Z_l(u)\}] dN_i(u). \tag{2.3}$$

The theory and application of the Cox models almost always assumes an exponential form for the relative risk function on regression variables, however, other regression forms such as a linear relative risk function (e.g., $1 + \beta_0' Z$) are more natural to use in some applications. Prentice and Self [1983] addresses that a linear relative risk regression model may provide a more convenient framework for studying epidemiologic risk factor interactions than an exponential relative risk regression. Using the same counting process formulation of Andersen and Gill [1982] but with some more stability and regularity assumptions, Prentice and Self [1983] establishes asymptotic distribution theory for a class of intensity function regression models in which the usual exponential regression form is relaxed. In Prentice and Self [1983], (2.1) is extended by

$$\lambda(t|Z(t)) = \lambda_0(t) r\{\beta_0' Z(t)\} \tag{2.4}$$

and (2.3) is modified by

$$l(\beta) = \sum_{i=1}^{n} \int_0^{\tau} \beta' Z_i(u) - \log[\sum_{l=1}^{n} Y_l(u) r\{\beta' Z_l(u)\}] dN_i(u), \tag{2.5}$$

where $r\{\cdot\}$ is a generalized relative risk function, which is an arbitrary non-negative twice differentiable function assumed to be locally bounded away from zero in some neighborhood of β_0. Estimators obtained by solving $\partial l(\beta)/\partial\beta = 0$ is shown to be consistent and asymptotically normal, based on asymptotic normality of the score function along with consistency of the observed information matrix $-n\partial^2 l(\beta)/\partial\beta^2$. Some stability and regularity conditions, beyond those of Andersen and Gill [1982], are required to show the consistency of the observed information matrix.

2.2 Causal Inference and Marginal Structural Cox Models

The purpose of this chapter is to review causal inference on the basis of potential outcome framework, relevant notation and assumptions, and finally to review the MSCMs. Before we start reviewing what the causal inference is, we address that under what circumstances standard statistical methods may fail to provide causal inference.

The term *time-varying confounder* is commonly used for time-varying risk factors of an outcome of interest that also predicts the subsequent exposure (or treatment). Post-treatment variables potentially affected by treatment and also affecting the response are referred to as *intermediate variables*. Unlike randomized clinical trials, many observational studies with long-term follow-up period often incorporate time-dependent covariates which are simultaneously confounders and intermediate variables on the causal pathway from exposure to disease. As a result, it may not be possible to obtain causal interpretation from the parameter estimator obtained using standard statistical methods. It is also true when the time-to-event response is considered. Standard methods of using the ordinary Cox models (2.1) or (2.4) adjusting for the time-dependent covariates may fail to provide appropriate causal effect of the exposure on an outcome of interest.

In the presence of such time-varying confounders, marginal structural models (MSMs, Robins [1999]), or marginal structural Cox models when the failure time is of interest, are powerful tools for assessing causal effects of time-varying treatments on an outcome of interest. MSMs are used increasingly to provide semi-parametric estimates of total (Bodnar et al. [2004]), joint (Robins, Hernán and Brumback [2000]; Hernán, Brumback and Robins [2000]), and direct/indirect (VanderWeele [2009*b*]) causal effects of exposures on an outcome

in epidemiologic studies.

In following sections, we give introduction to causal inference with relevant concepts, notation, and assumptions to understand MSMs as tools to draw causal inference. Hereafter we assume an observational study wherein confounding effect exists, which interested in evaluating treatment effect on a participant's failure time such as HIV studies for example.

2.2.1 Causal Inference and Potential Outcomes

Most causal inferences are based on the idea of potential outcomes under all possible treatment assignments, introduced by Neyman [1923]. The potential outcomes, which include observed and unobserved outcomes, are sometimes called counterfactual outcomes (or simply counterfactuals) since these outcomes could have happened contrary to what we actually observed. In this framework, causal inference can be considered as a missing data problem letting the potential outcomes as missing data, especially the unobserved outcomes to be the missing outcomes.

Below we modify some of the notation introduced from earlier section to be more suitable to a hypothetical biomedical study and to the causal inference framework. Capital letters represents random variables and lower case letters represents values of the random variables or constants, the same as before. Now, let $A_i(t)$ be the treatment vector of subject i where t denotes the time since the beginning of the subject's follow-up. Let $L_i(t)$ denote a vector of assembled covariates such as CD4 counts and PCP level from subject i at time t. The subscript i will sometimes be suppressed in our notation since we assume $A_i(t)$ and $L_i(t)$ are random vectors for each subject drawn independently from a distribution common to all subjects. Let V represent baseline covariates which can be a part of $L(0)$. $A(t)$ and $L(t)$ are defined to be zero when $t < 0$. Note that a $p \times 1$ covariate vector $Z(t)$ introduced from earlier section equals to $\{A(t), L(t)\}$.

Potential (Counterfactual) Outcomes In the context of causal inference, overbars are used to represent history up to and including time t such that $\overline{A}(t) = \{A(u); 0 \leq u \leq t\}$ and $\overline{L}(t)$ is defined analogously, assuming that decisions related to treatment at t is made after obtaining the covariate information at t, i.e., $L(t)$ is temporally earlier than $A(t)$. Let $\perp\!\!\!\perp$ denote statistical independence; for example, $A \perp\!\!\!\perp B|C$ denotes A is independent of B

given C.

$\overline{a}$ represents each possible treatment plan; $\overline{a} = \{a(t) : 0 \leq t \leq \tau\}$ where τ is the study end point, same as before. Each possible value of $\overline{a}$ can be interpreted as a pre-specified treatment plan. Practical examples of $\overline{a}$ might be never treated (i.e., $a(t) = 0$ for all $t \in [0, \tau]$), treated starting at a pre-specified time t_1 (i.e., $a(t) = I[t > t_1]$), treated from baseline (i.e., $a(t) = 1$ for all $t \in [0, \tau]$), etc. Then $T_{\overline{a}}$ represents a random variable implying a subject's potential failure time had (possibly contrary to what we observe from actual study) the subject been treated with history $\overline{a}$. For example, one can use notation $T_{\overline{0}}(t)$ to represent a subject's potential failure time if he had treated from baseline, $T_{\overline{t_1}}(t)$ if he had treated since time $t_1 > 0$, and $T_{\overline{\infty}}$ if he had never treated. At each time-to-event t such as death or disease occurrence time, a set of three different failure times $(T_{\overline{0}}(t), T_{\overline{t_1}}(t), T_{\overline{\infty}}(t))$ comprises potential outcomes.

Assumptions Most causal models that are based on the idea of potential outcomes rely on the following four assumptions.

1. *Consistency* In reality, we only observe the outcome T with a subject's actual treatment history $\overline{A}$, i.e., $T = T_{\overline{a}=\overline{A}} = T_{\overline{A}}$. This identity is called the fundamental "consistency" assumption that links the potential failure times $T_{\overline{a}}$ to the observed data $(T_{\overline{A}}, \overline{A})$.

2. *No unmeasured confounders* There are no unmeasured confounders for the effect of $A(t)$ on T if, for all $\overline{a}$,

$$T_{\overline{a}} \perp\!\!\!\perp A(t) | \overline{A}(t^-), \overline{L}(t) \tag{2.6}$$

   holds (Robins [1999] and Hernán, Brumback and Robins [2000]).

3. *Positivity* We say that positivity assumption holds if $\Pr[A(t) = a | L(t) = l] > 0$ for all $a \in \{0, 1\}$ and l such that $\Pr[L(t) = l] \neq 0$.

4. *No misspecification of the model* As we always assume that the model we employ is a correct model to analyze data, a causal model to estimate the effect of treatment is assumed to be correctly specified.

Informally, consistency means that the outcome for every treated individual equals to the subject's outcome if he/she had received treatment, and the outcome for every untreated individual equals to his/her outcome had the subject remained untreated. No unmeasured confounders means that the risk of failure under the potential treatment history $\overline{a}$ among the treatment group equals to the risk under the same potential treatment history among untreated group for each $\overline{a}$. Therefore the treated and untreated groups are exchangeable as in a randomized trial. For this reason, the assumption is also called "exchangeability" or "sequential randomized assumption" in some articles since it implies that potential outcomes are exchangeable regardless of treatment history given all relevant confounder history as if in a randomized trial. However, if there exists any unmeasured confounder that predicts $A(t)$ at time t then the potential outcomes are no longer independent of treatment history. Positivity simply means that the conditional probability of receiving every value of treatment is greater than zero. No missspecification of the model assumption is the natural assumption to make any statistical inference and may be tested using sensitivity analysis.

In general, the no unmeasured confounders assumption is a crucial assumption to draw causal inference using some causal models but is not statistically testable. A more complete studies on these condition and causal inference can be traced back to Rubin [1974], Rubin [1976], Rubin [1980], Robins [1986], Greenland and Robins [1986], and Robins [1987].

Causal and Statistical Exogeneity Informally, a treatment process is referred to be a "causally exogenous" process if the conditional probability of receiving a treatment $A(t)$ given past treatment and (measured and unmeasured) prognostic factor history depends only on past history of treatment history $\overline{A}(t^-)$. Mathematical definition of causal exogeneity may vary across different articles. Definitions of presented in this proposal are adopted from Hernán, Brumback and Robins [2001]. The article defines a treatment process to be "causally exogenous" if

$$T_{\overline{a}} \perp\!\!\!\perp A(t) | \overline{A}(t^-) \tag{2.7}$$

for all treatment plans $\overline{a}$, which is equivalent to state that $T_{\overline{a}}$ is independent of $\overline{A}(t)$. Also, definition of "statistically exogenous" of treatment process is adopted from the same article

which is given by

$$\overline{L}(t) \perp\!\!\!\perp A(t) | \overline{A}(t^-). \tag{2.8}$$

This implies that conditioning on treatment history before time t, probability of receiving treatment at time t does not depend on the history of measured time-dependent prognostic factors up to t. It can be seen that (2.8) is a necessary condition for $A(t)$ to be causally exogenous, but (2.8) does not imply (2.7) due to the possibility of unmeasured confounders. Following Robins et al. [1992], Hernán, Brumback and Robins [2001] also defined that there are no unmeasured confounders for the effect of $A(t)$ on T if, for all $\overline{a}$,

$$T_{\overline{a}} \perp\!\!\!\perp A(t) | \overline{A}(t^-), \overline{L}(t) \tag{2.9}$$

holds.

Robins [1999] showed that statistical exogeneity implies causal exogeneity under the assumption of (2.9). Also it is well recognized that treatment parameters of a correctly specified association model have causal interpretation if the treatment process is causally exogenous. Therefore, causal inference can be drawn from using standard association models if condition (2.8) is true assuming that (2.9) holds.

2.2.2 Marginal Structural Cox Models

Robins [1999] introduced MSMs combined with inverse-probability-treatment-weights (IPTW) as a method to draw causal inference in the presence of confounding, which rely on the potential outcome framework. IPTW can be considered a type of inverse sampling weights to account for missing data or sampling bias problem. By weighting observations via IPTW, we can reflect back the balanced design from observational data having confounding effect (under the assumptions described from Chapter 2.2.1).

We first review IPTW and then describe MSCMs after.

Inverse-Probability-Treatment-Weighting Suppose that we can correctly model the probability of receiving treatment at time t given past treatment history and covariate history, i.e., $\Pr[A(t)|\overline{A}(t^-), \overline{L}(t)]$. Then we could measure the degree of statistical exogeneity

of the treatment process through time t by calculating a following weight at t:

$$W^T(t) = \prod_{k \le t} \frac{\Pr[A(k)|\overline{A}(k^-)]}{\Pr[A(k)|\overline{A}(k^-), \overline{L}(k)]}, \tag{2.10}$$

which is referred to as *inverse-probability-of-treatment-weights* (IPTW). Under the four assumptions of consistency, no unmeasured confounders, positivity, and no misspecification of the model to estimate the weights, we can create a hypothetical population by weighting each subject at risk at each failure time with (2.10). This hypothetical or weighted study population is known as the *pseudo-population*. Robins [1999] and Lemma A.1 of Hernán, Brumback and Robins [2001] proved that $L(t)$ no longer predicts $A(t)$ in each pseudo-population created at each failure time t (note that (2.10) equals to 1 at any time t if $L(t)$ does not predict $A(t)|\overline{A}(k^-)$, i.e., the treatment process is statistically exogenous). Then it follows that the treatment process is causally exogenous in the pseudo-population under the assumption of no unmeasured confounders. Thereby one can employ standard association models to estimate the treatment effect which can further be interpreted as a causal effect.

As the same manner, we can effectively adjust bias occurred by censoring due to loss to follow-up when time-to-event data is considered. This can be done by considering inverse-probability-of-censoring-weights (IPCW), say W^C, where

$$W^C(t) = \prod_{k \le t} \frac{\Pr[C(k) = 0|\overline{C}(k^-) = 0, \overline{A}(k^-)]}{\Pr[C(k) = 0|\overline{C}(k^-) = 0, \overline{A}(k^-), \overline{L}(k)]}, \tag{2.11}$$

under the assumptions of independent censoring and no unmeasured confounders for censoring. Here, $C(k) = 0$ means a subject remains uncensored prior to time k and $C(k) = 1$ means censored at that time.

Robins [1997] first introduced IPTW, which is called *unstablized weights*, as a tool to adjust non-ancillary treatment process in the observational study, however, it has a slightly different form than (2.10). Stablized weights has the same denominator as in (2.10) but the numerator in (2.10) is always 1 regardless of time t. Therefore it is a nondecreasing function of t since the product of probabilities in the denominator decreases over time. Robins, Hernán and Brumback [2000] suggests *stabilized weights* which is the IPTW shown in (2.10) as a substitute of the unstablized weights, and this is by far the most widely used IPTW.

Besides this, truncated (Cole and Hernán [2008]) and normalized (Xiao, Abrahamowicz and Moodie [2010]) weights are also introduced and these weights can be considered as types of stablized weighs as they all aim to adjust variability of IPTW and make it stable over time. Stablized weights are generally recommended to employ in practice as they lead to, often remarkably, more efficient estimators of causal treatment effect.

When survival data is considered, inverse-probability-weights (IPW) defined by $W(t) \equiv W^T(t) \times W^C(t)$ are the stabilized weights. For estimation of random weights $W(t)$ see Hernán, Brumback and Robins [2000], Hernán, Brumback and Robins [2001], and Cole and Hernán [2008]. Since investigators should assume that the model to estimate IPW (e.g., a logistic model), sensitivity analysis results with different model specifications will help to see validity of the correct model assumption in practice.

Marginal Structural Cox Models Marginal Structural Cox Models (MSCMs) are given by

$$\lambda_{T_{\overline{a}}}(t) = \lambda_0(t) \exp\{\beta_0' f(\overline{a}(t))\} \tag{2.12}$$

where $\lambda_{T_{\overline{a}}}(t)$ is the hazard of failure at time t if all subjects in a study population had followed treatment history $\overline{a}$ through time t, $\lambda_0(\cdot)$ is an unspecified baseline hazard function corresponding to the hazard if all subject had been untreated, and β_0 is an unknown parameter vector. If we are interested in current treatment effect of zidovudine on AIDS so that $f(\overline{a}(t))$ in (4.1) becomes $a(t)$, and $\exp(\beta_0)$ has a causal interpretation such as the ratio of the hazard of getting AIDS at any time t if all subjects had been continuously exposed to zidovudine compared with the hazard rate at t had all subjects remained unexposed. This model is a causal model for the marginal distribution of the variables $T_{\overline{a}}$ which is the potential outcomes that are generally unobserved. Hence estimation of the causal log rate ratio β_0 cannot be made directly through this model. In the absence of confounding, association implies causation thereby we can use the standard Cox regression model to obtain causal estimates. As mentioned the above, Robins [1999] showed that we can create a psuedo-population via IPW at each failure time t in which time-dependent prognostic factors no longer predict treatment history. Robins [1999] also proved that the causal relationship between treatment and hazard in the psuedo-population is the same as

in the original study population, and the estimator of treatment effect obtained by using the standard Cox regression model based on the psuedo-population converges in probability to β_0 in (4.1). Therefore the estimator of treatment effect obtained by using the ordinary time-dependent Cox model adjusting (only) for the treatment, after weighting each individual at each failure time by IPW, can have causal interpretation as it converges in probability to β in (4.1).

2.3 Cox Models with Case-cohort Sampling

The case-cohort study proposed by Prentice [1986*a*] and Self and Prentice [1988] is a cost-effective design particularly when large epidemiologic cohort studies with rare disease or infrequent event such as HIV studies are considered. This design involves random selection of a subcohort (or a stratified random sample) from the entire cohort and all participants who experience the event of interest, henceforth cases. By monitoring covariate information only for a random subcohort and for all cases we can gain cost and effort saving. The subcohort constitutes the comparison set of cases occurring at a range of failure times as well as a basis for covariate monitoring during the course of cohort follow-up (Self and Prentice [1988]).

Prentice [1986*a*] considers Cox modeling on time-to-response case-cohort data. Suppose that a random subcohort $\tilde{\mathcal{C}}$ of size $\tilde{n}$ is selected from the entire cohort $\mathcal{C}$ of size n. Then the log partial likelihood is modified by

$$l^*(\beta) = \sum_{i=1}^{n} \int_0^\tau \beta' Z_i(u) - \log[\sum_{l \in \tilde{\mathcal{C}} \cup \{i\}} Y_l(u) r\{\beta' Z_l(u)\}] dN_i(u). \tag{2.13}$$

in the presence of case-cohort sampling, which is termed a log *pseudolikelihood* by Prentice [1986*a*].

Self and Prentice [1988] provides a full range of asymptotic theory for parameter estimators of the Cox models in the presence of the case-cohort sampling using a slightly different log partial likelihood form

$$\tilde{l}(\beta) = \sum_{i=1}^{n} \int_0^\tau \beta' Z_i(u) - \log[\sum_{l \in \tilde{\mathcal{C}}} Y_l(u) r\{\beta' Z_l(u)\}] dN_i(u). \tag{2.14}$$

Estimators obtained by solving $\partial\tilde{l}(\beta)/\partial\beta = 0$ are shown to converge in probability to β_0 and asymptotically normally distributed via same techniques as in Andersen and Gill [1982] and Prentice and Self [1983], i.e., by showing asymptotic normality of the score function along with consistency of the observed information matrix. It is also shown that estimators obtained by solving $\partial l^*(\beta)/\partial\beta = 0$ and $\partial\tilde{l}(\beta)/\partial\beta = 0$ converge in probability to the same quantity, β_0, in Self and Prentice [1988].

Several other authors such as Binder [1992] and Lin and Ying [1993] expand the idea of the case-cohort design and provide estimating equations to obtain estimators in more general settings. Binder [1992] describes how to create a family of survey-related sampling plans, and provided a procedure for fitting the proportional hazards models to survey data with complex sampling designs including the case-control sampling. Estimating equation proposed in the article is an extension of the standard score function equation in (2.2), with incorporating probability of being sampled. In particular, Binder [1992] proposes a score function given by

$$U^*(\beta) = \sum_{i=1}^{n} w_i \delta_i \{Z_i(X_i) - \frac{S^{*(1)}(\beta_0, X_i)}{S^{*(0)}(\beta, X_i)}\},$$

where now the statistics were modified by

$$S^{*(r)}(\beta, X_i) = n^{-1} \sum_{i=1}^{n} w_i Y_i(X_i) Z_i^{\otimes k}(X_i) \exp\{\beta' Z(X_i)\}$$

for $k = 0, 1, 2$, and w_i is the inclusion probability for the subject i, i.e., $w_i = 1/\pi_i$ if the subject i is selected in the sample and 0 otherwise. Estimators obtained by solving $U(\beta) = 0$ are then shown to be asymptotically normally distributed. Lin and Ying [1993] provides a general solution to the problem of missing covariate data under the Cox models, considering case-cohort data as a possible example of missing covariate data. The estimating function proposed in the article is an approximation to the partial likelihood score function with full covariate measurements, which reduces to the score function of Self and Prentice [1988] in the special setting of the case-cohort designs. The approximate partial likelihood score function is given by

$$\tilde{U}(\beta) = \sum_{i=1}^{n} \delta_i \mathbf{H}_i(X_i)\{Z_i(X_i) - \frac{\tilde{S}^{(1)}(\beta_0, X_i)}{\tilde{S}^{(0)}(\beta, X_i)}\}$$

where $\mathbf{H}_i$ is a $p \times p$ diagonal matrix with indicator functions $\{H_{1i}(\cdot), ..., H_{pi}(\cdot)\}$ as the diagonal element with $H_{1i}(X_i)$ being an indicator whether $Z_{ji}(X_i)$ is available at failure time X_i, $\tilde{S}^{(k)}(\beta, X_i)$ are defined by $n^{-1}\sum_{i=1}^{n} H_{0i}(X_i)Y_i(X_i)Z_i^{\otimes k}(X_i)\exp\{\beta' Z(X_i)\}$ for $k = 0, 1, 2$ with $H_{1i}(X_i)$ being an indicator $I\{H_{ji}(X_i) = 1\}$ for all $j = 1, ..., p$. Then approximate partial likelihood estimators (APLE) are the root to the estimating equation $\tilde{U}(\beta) = 0$. The resulting parameter estimators are consistent and asymptotically normal with a covariance matrix for which a simple and consistent estimator is provided. Also, the asymptotic theory of the APLE are established on regularity conditions that are much simpler to interpret and check than those in Self and Prentice [1988].

Despite the efficiency of the sampling methods, applications of the case-cohort designs had been limited because of perceived analytic complexity, especially on the variance computation proposed from Self and Prentice [1988]. Self and Prentice [1988] variance estimator is not easy to implement as it includes computation of covariances between score contributions from pairs of different risk sets. Simple robust variance estimators are proposed by Lin and Ying [1993] and Barlow [1994] as a solution to the computational challenges in variance computation, and also practical implementation of the Cox models to the real case-cohort data is addressed by Therneau and Li [1999] and Barlow et al. [1999]. Therneau and Li [1999] describes how to obtain Self and Prentice [1988], Barlow [1994], and Lin and Ying [1993] parameter estimators along with their variance estimators using standard software packages, with SAS and S-Plus as particular examples. Barlow et al. [1999] illustrates weighting methods as model fitting techniques and provides a SAS macro that computes the weighted estimates and the robust covariance matrix.

2.4 Statistical Methods to Improve Efficiency

In the standard associational case-cohort Cox analysis, various methods have been proposed to improve efficiency from maximum pseudolikelihood estimators of Prentice [1986*a*] or Self and Prentice [1988]. In this chapter, we briefly review some of these methods who

seek to improve efficiency of the hazard ratio estimation compared to Prentice [1986*a*] and Self and Prentice [1988].

Unweighted Psuedo-Partial Likelihood Estimators As described in 2.3, Prentice [1986*a*] proposed a pseudo-likelihood approach for the hazard ratio parameter estimation in the Cox model along with heuristic procedures for parameter estimation when the case-cohort design is applied. Asymptotic distribution theory of the case-cohort maximum pseudo-likelihood estimator was developed by Self and Prentice [1988] using martingale technique and finite population convergence results. Both Prentice [1986*a*] and Self and Prentice [1988] do not accommodate case sampling or stratified sampling of controls, i.e., they considered unweighted pseudo-likelihoods.

Unweighted Psuedo-Partial Likelihood Estimators After Prentice [1986*a*] and Self and Prentice [1988], various methods have been proposed as means of improving the efficiency of the hazard ratio estimation (compared to Prentice [1986*a*] and Self and Prentice [1988]) in the standard (associational) case-cohort Cox regression analysis. Chen and Lo [1999] studied a different class of estimating equations than Prentice [1986*a*] and Self and Prentice [1988] by constructing different risk sets in the estimating equations. They proposed to utilize complete information of all cases when calculating ratio of weighted averages based on risk set information inside the estimating equations. In particular, authors proposed three different estimating equations which all use the empirical distribution of covariate Z among cases to the conditional joint distributions of (Z, X) among cases, but use different estimators of $p = \text{pr}\{\delta(X_i) = 1\}$. Chen et al. [2001] found an optimal sample reuse method via local averaging, and proposed a unified weighted estimating equation, that can be used in various sampling design, to improve efficiency.

Time-Varying Inverse-Sampling-Weights Barlow [1994] and Barlow et al. [1999] considered estimators based on weighted pseudo-likelihood estimation. At each failure time, contribution of cases and nonfailures (controls) at risk are weighted by either fixed or time-varying inverse-sampling-weights (ISW) to account for subcohort sampling.

Using All Available Covariate Data from Full Fohort Later, methods that seek to utilize some of the phase 1 covariate information were proposed. Borgan et al. [2000] considered a stratified sampling by a phase 1 variable which is a correlate of exposure, to incorporate statum-specific ISW in the estimating equation. Stratum-specific ISW can

be calculated using empirical sampling fraction within each stratum. He proposed three different estimating equations by considering different types of weights. Simulation studies suggested that the stratified estimator II with time-varying ISW, referred to as BII estimator from herein, is the most efficient among the existing estimators. Kulich and Lin [2004] established asymptotic theory for the BII type of estimators. In addition they developed a class of weighted estimators which utilize all available covariate information from the full cohort data. Proposed weighted estimators are 1) doubly weighted (DW) estimator and 2) combined doubly weighted (CDW) estimator which involve general time-varying ISW. The methods involve a modeling step for prediction of the values of each partially missing phase 2 variables, and is likely of greatest use when there are only 1 or 2 such variables. The authors suggest to use CDW estimator in practice as DW estimator is efficient only if a model to predict the phase 2 variables given all the phase 1 variables is correct. Numerical studies indicated that the CDW estimator is more efficient then other existing estimators such as Chen and Lo [1999], Borgan et al. [2000], and Chen et al. [2001]. The efficiency gain for the phase 2 covariates depends on the ability of the first-phase data to predict the true values of the partially missing variables. Later, Breslow et al. [2009*a*] and Breslow et al. [2009*b*] considered calibration or estimation of ISW by making use of phase 1 covariate information. Calibration method adjusts ISW to be as close as possible to the sampling weights subject to a certain constraint. Estimation methods uses ISW as inverse of inclusion probabilities estimated from a logistic regression model that predicts which cohort subjects are sampled at phase 2. Simulation study and real data analysis reported by Breslow et al. [2009*b*] showed that such adjustment on ISW can dramatically improve precision of the baseline hazard ratios, which are estimated for baseline covariates, i.e., a part of phase 1 variables. They also showed that the methods can improve precision for the phase 2 covariates when their values may be imputed with reasonable accuracy for the non-subcohort controls.

In Chapter 4 we demonstrate how we can extend some of the aforementioned methods might be extended to the causal setting and discuss why some of the methods might not be useful to improve efficiency in the case-cohort MSCM analysis.

Chapter 3

Marginal Structural Cox Models with Case-cohort Sampling

3.1 Introduction

Biomedical cohort studies are often conducted with the goal of assessing the effect of a time-varying treatment (or exposure) on a survival time. In such studies there may exist time-dependent covariates which are simultaneously (i) confounders and (ii) affected by prior treatment. In the presence of time-varying confounders affected by prior treatment, standard methods such as Cox regression modeling with time-varying covariates do not in general yield consistent estimators of the causal effect of treatment Robins [1986, 1998]; Robins and Rotnitzky [1992]; Hernán, Brumback and Robins [2001]. On the other hand, marginal structural models (MSM) fit using inverse probability weighting can be employed to obtain consistent estimators of the causal effect of a time-varying treatment on an outcome of interest, even if there are time-varying confounders affected by prior treatment Robins [1999].

For example, consider the Multicenter AIDS Cohort Study (MACS), an observational study of HIV-positive homosexual men. Using data from MACS, Hernán, Brumback and Robins [2001] showed that (i) current CD4 count and *Pneumocystis carinii* pneumonia (PCP) status were independent risk factors for death and were predictive of subsequent treatment with zidovudine (AZT) and prophylaxis therapy, and (ii) prophylaxis therapy was a protective risk factor for the development of PCP subsequently. Thus, to assess the effect of AZT and prophylaxis therapy on mortality in MACS, a method is required that can appropriately account for time-varying confounders affected by prior treatment (in particular, PCP status). Applying standard (i.e., unweighted) Cox regression with time-dependent covariates to the MACS data, Hernán, Brumback and Robins [2001] reported an

estimated hazard ratio of 1.85 (95% CI 1.49, 2.30) for AZT users versus nonusers, suggesting that treatment increases the risk of death in HIV-positive homosexual men, contrary to results from randomized clinical trials. On the other hand, fitting a marginal structural Cox model (MSCM) with inverse probability weighting yielded an estimated hazard ratio for AZT of 0.67 (95% CI 0.46, 0.98), in agreement with results from randomized trials of AZT. The difference in hazard ratio estimates between the unweighted Cox regression model and the MSCM with inverse probability weighting is not surprising given the aforementioned established results about the (in)consistency of the standard estimators in the presence of time-varying confounders affected by prior treatment.

Recently, Cole et al. [2012] considered fitting MSCMs via inverse probability weighting in the presence of case-cohort sampling. The case-cohort study design is a cost-efficient approach to estimate treatment effects in large cohorts with low event rates, when treatment or covariate information is expensive. The design entails randomly selecting a subcohort from the entire cohort. Covariate information is then collected only from the random subcohort and from individuals that are observed to experience an event (i.e., cases), saving cost and effort relative to obtaining covariate information from the full cohort. In addition to being cost efficient, the case-cohort design enjoys other benefits. For instance, the subcohort can serve as a basis for real time covariate monitoring during the course of the study. Also, because the subcohort is chosen randomly, survival times to different diseases can be analyzed using the same subcohort Self and Prentice [1988].

In the presence of case-cohort sampling, Cole et al. [2012] considered estimating the causal hazard ratio of a MSCM via inverse probability weighting. Simulation studies indicated the estimator proposed by Cole et al. [2012] can perform well empirically, however no formal justification for their estimator has been developed to date. Therefore, following Cole et al. [2012], we consider estimating the causal hazard ratio of a MSCM via inverse probability weighting in case-cohort studies and establish consistency and asymptotic normality for the estimator that maximizes a weighted-pseudo-partial-likelihood (WPPL) under certain regularity conditions.

The approach utilized in this paper entails standard counting process and martingale theory. This formulation readily enables practical implementation of the methods using existing survival analysis software. Framing the problem using counting processes may also

be helpful in future work, e.g., in fitting MSCMs to data from nested case-control studies or in the presence of competing risks. In the special situation that the subcohort equals the full cohort, the proposed inverse probability weighted estimator is asymptotically equivalent to the estimator in Robins [1999]. In this case our proof gives an alternative consistency and normality proof to Robins [1999], who did not utilize the usual counting process framework. We also derive a new variance estimator that arises from the counting process formulation under both full and case-cohort settings. Empirical results presented in this paper indicate that in certain scenarios the proposed variance estimator may be preferred to the so-called "robust" variance estimator Lin and Ying [1993] employed in Cole et al. [2012].

The outline of the remainder of this paper is as follows. In §3.2, estimators of the hazard ratio of a MSCM in the presence of case-cohort sampling are introduced, including the estimator proposed by Cole et al. [2012]. Consistency and asymptotic normality are established in §3.3 and §3.4, respectively. §3.5 explains how one can directly obtain the proposed inverse probability weighted estimators using standard survival analysis software, and presents a simulation study.

3.2 Marginal Structural Cox Model Estimators

3.2.1 Notation, Assumptions, and Model

Capital letters will represent random variables and lower case letters will represent values of the random variables or constants. Consider an observational cohort study where the outcome of interest is a survival time T, based on the time from study entry until some particular outcome occurs. Throughout we assume T is continuous so that there are no tied failure times between individuals. During the course of the study individuals may dropout or discontinue participation in the study, such that T is not observed but rather right censored at the last time the individual was under study. Suppose individuals may or may not elect to receive treatment at various points of time during the study. Let $A_i(t)$ indicate whether subject i is on treatment at time t. If more than one treatment is available, then $A_i(t)$ is a vector of treatment indicator variables corresponding to the joint treatment levels. In the sequel we assume $A_i(t)$ is a $p \times 1$ vector and treatment variation is irrelevant VanderWeele [2009*a*]. The subscript i will often be suppressed, when there is no ambiguity,

because we assume random vectors are drawn independently from a distribution common to all subjects. Let $L(t)$ denote a vector of covariates, such as CD4 count or PCP status, at time t. Let $L(0)$ represent baseline covariates. Overbars are used to represent history up to and including time t such that $\overline{A}(t) = \{A(u) : 0 \leq u \leq t\}$ and $\overline{L}(t)$ is defined analogously. Assume that decisions related to treatment at t are made after obtaining the covariate information at t, i.e., $L(t)$ is temporally prior to $A(t)$. For a case-cohort study, the time varying covariates $L(t)$ and treatment $A(t)$ are by design observed only for the cases and individuals in the random subcohort (while under study); $L(t)$ and $A(t)$ are missing for all other individuals. Corresponding to the subcohort, let $\tilde{\mathcal{C}}$ denote the set of indices of size $\tilde{n} \leq n$ that are randomly selected without replacement from the set $\{1, \ldots, n\}$ corresponding to the entire cohort.

Let $\overline{a}$ denote a possible (static) treatment plan, i.e., $\overline{a} = \{a(t) : 0 \leq t \leq \tau\}$ where τ is the study duration. Assume $\tau = 1$ hereafter without loss of generality. Each possible value of $\overline{a}$ can be interpreted as a prespecified treatment plan. Assuming a single treatment (i.e., $p = 1$), practical examples of $\overline{a}$ might be never treat (i.e., $a(t) = 0$ for all $t \in [0, 1]$), treat starting at a prespecified time $t_1 < 1$ (i.e., $a(t) = I\{t \geq t_1\}$ where $I\{\cdot\}$ is the usual indicator function), treat from baseline (i.e., $a(t) = 1$ for all $t \in [0, 1]$), etc. Define $T_{\overline{a}}$ to be a subject's potential failure time had (possibly contrary to what was observed in the actual study) the subject been treated according to $\overline{a}$. Let $\perp\!\!\!\perp$ denote statistical independence; e.g., $A \perp\!\!\!\perp B|C$ denotes A is independent of B given C. Assume

$$T = T_{\overline{a}} \quad \forall \overline{a} \text{ such that } a(t) = A(t) \quad \forall t \leq T, \tag{3.1}$$

$$T_{\overline{a}} \perp\!\!\!\perp A(t)|\overline{A}(t^-), \overline{L}(t) \quad \forall \overline{a}, \tag{3.2}$$

$$\text{pr}[A(t)|\overline{A}(t^-), \overline{L}(t)] > 0 \quad \forall t \in [0, 1] \text{ such that } \text{pr}[\overline{A}(t^-), \overline{L}(t)] > 0 \tag{3.3}$$

which are referred to as the *causal consistency*, *conditional exchangeability*, and *positivity* assumptions, respectively. Assumption (3.1) states that, in the absence of censoring, the observed failure time T equals the potential failure time $T_{\overline{a}}$ for all treatment plans $\overline{a}$ consistent (i.e., compatible) with the observed treatment up to time T. Assumption (3.2) states that conditional on treatment and covariate histories, treatment at time t is independent of the potential survival time under $\overline{a}$ (i.e., no unmeasured confounding). Assumption (3.3)

states that the conditional probability of receiving any particular treatment is greater than zero. Of these three assumptions, only (3.3) can be tested empirically. Sensitivity analysis may be useful in assessing the robustness of inference drawn to violations of assumption (3.2) Robins, Rotnitzky and D [1999].

Consider the MSCM

$$\lambda_{T_{\overline{a}}}(t) = \lambda_0(t) \exp\{\beta_0' f(\overline{a}(t))\}$$

where $\lambda_{T_{\overline{a}}}(t)$ is the hazard of failure at time t if all individuals in the population had followed treatment plan $\overline{a}$ through time t, $\lambda_0(t)$ is an unspecified baseline hazard function corresponding to the hazard if all individuals had been untreated through time t, $f(\overline{a}(t))$ is a specified function of treatment history up to time t, and β_0 is an unknown parameter vector. Hereafter, we consider the MSCM

$$\lambda_{T_{\overline{a}}}(t) = \lambda_0(t) r\{\beta_0' a(t)\} \tag{3.4}$$

where for notational convenience we let $r\{\cdot\} = \exp\{\cdot\}$. For example, if we are interested in the causal effect of current AZT treatment on mortality of HIV-positive homosexual men, then $r(\beta_0)$ is the ratio of the hazard of death at time t had all subjects in the population alive at time t been exposed to AZT compared to had the subjects been unexposed at time t. Note (3.4) focuses on the effect of current treatment status only; however, the results presented below are valid for any specified $f(\overline{a}(t))$.

In this paper the counting process framework is employed to study the large sample behavior of estimators of β_0. Note that all processes discussed hereafter refer to observed processes. Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a complete probability space and let $\{\mathcal{F}_t : t \in [0,1]\}$ be an increasing right-continuous family of sub σ-algebras of $\mathcal{F}$ consisting of failure times, covariates and treatment histories up to time t, and censoring histories up to time t^+ for all subjects in a cohort of size n. That is, the filtration with respect to the probability space is the same as the usual filtration, except that treatment histories are now separated from covariate histories. Let $N_i(\cdot)$ be a counting process adapted to $\mathcal{F}_t$ representing the number of failures of subject i by time t such that $dN_i(t)$ indicates the number of events of

subject i that occurred in $[t, t + dt)$ for sufficiently small dt. Because failures are assumed to occur in continuous time, we only allow jumps of size 1 and no simultaneous jumps can occur in $[t, t + dt)$. Let $C_i(t) = 0$ indicate that subject i remained uncensored prior to time t and $C_i(t) = 1$ otherwise. The treatment process $A_i(\cdot)$ and the censoring process $C_i(\cdot)$ are assumed to be piece-wise constant point processes with cadlag (right-continuous with left-hand limits) step-function sample paths. The processes $A(\cdot)$ and $C(\cdot)$ are assumed to have jumps that can occur at no more than a finite number of time points. Informally, this means that all participants follow (approximately) the same visit schedule. This assumption should be reasonable in studies with regularly scheduled follow-up visits (e.g., every six months) and good study compliance. We refer to censoring as ignorable (or noninformative) if the cause-specific hazard of being censored at t among subjects alive and uncensored does not depend on the failure times $T_{\overline{a}}$ given prior treatment/covariate history $\overline{A}(t^-)$ and $\overline{L}(t^-)$ (Hernán, Brumback and Robins [2001]). Let $Y_i(t) = I\{N_i(t) = C_i(t) = 0\}$ denote whether an individual is at-risk of being observed to fail at time t, having left-continuous sample paths, and assume $\text{pr}[Y(1) > 0] > 0$.

3.2.2 Inverse Probability Weights

Suppose that we can correctly model the probability of receiving treatment at time t given the past treatment history and covariate history. Then we can consistently estimate the following weights

$$W^T(t) = \prod_{k \le t} \frac{\text{pr}[A(k)|\overline{A}(k^-)]}{\text{pr}[A(k)|\overline{A}(k^-), \overline{L}(k)]}, \tag{3.5}$$

which will be referred to as inverse-probability-of-treatment-weights (IPTWs). Note that we can consistently estimate the numerator probabilities in (4.2) based on sample proportions because $A(\cdot)$ is assumed to have at most a finite number of jumps over the study period. Under (3.2) to (3.3), in the absence of censoring, Robins [1999] showed that a consistent estimator of the unknown parameter β_0 in (3.4) can be obtained by fitting an ordinary time-dependent Cox model with the contribution of subject i to the risk set at time t weighted by estimates of (4.2). Informally we can think of the analysis via IPTWs as reweighting the observed data set such that it has the same properties as a random sample, with respect to

the measured confounders L, from a population where $\overline{L}(t) \perp\!\!\!\perp A(t)|\overline{A}(t^-)$ holds at time t. The weighted study population is sometimes called a *pseudo-population*.

Dropout (i.e., right censoring) may introduce selection bias if dropout is associated with exposure and dropout is associated with the outcome. In the presence of such censoring, we still can obtain a consistent estimator of β_0 by fitting the ordinary Cox model but weighting a subject alive and uncensored at time t by estimates of $W^T(t) \times W^C(t)$, where

$$W^C(t) = \prod_{k \leq t} \frac{\text{pr}[C(k) = 0|\overline{C}(k^-) = 0, \overline{A}(k^-)]}{\text{pr}[C(k) = 0|\overline{C}(k^-) = 0, \overline{A}(k^-), \overline{L}(k)]}, \tag{3.6}$$

under the assumption of no unmeasured confounders for censoring, an analogous assumption to (3.3) for censoring, and assuming that we can correctly model the denominator probabilities in (3.6) Robins [1999]. Here the weighted study population can be thought of as a pseudo-population in which there is no confounding due to measured covariates or selection bias due to censoring. In §3.2.3, we will make use of the (stabilized) weights defined by $W(t) \equiv W^T(t) \times W^C(t)$ after modifying (4.2) by adding $C(k) = 0$ to the conditioning events in both the numerator and the denominator Hernán, Brumback and Robins [2000]. Hereafter $W(t)$ will be referred to as inverse-probability-weights (IPWs). Note that (4.2) and (3.6) are finite products. In addition, (3.3) ensures non-zero probabilities in the denominators of (4.2) and (3.6) and hence the IPWs at all t are bounded.

Results presented in this article are not limited to a specific form of the weights $W(t)$. The proposed methods are applicable to different inverse probability weighting analysis provided that the IPWs (or IPTWs in the absence of censoring) are bounded, such as when truncated [Cole and Hernán, 2008] and normalized [Xiao, Abrahamowicz and Moodie, 2010] weights are employed. Under the assumption of finite support of the treatment and censoring processes, unstabilized weights [Hernán, Brumback and Robins, 2001] are also bounded. However, unstabilized weights are known to be highly variable and are by design monotone increasing functions of t. Other weights such as stabilized, truncated, and normalized weights are generally recommended in practice as they lead to more efficient estimators of the causal treatment effect.

We now briefly describe estimation of the random weights $W(t)$, denoted by $\hat{W}(t)$. One may specify a pooled logistic model (treating each person-visit as an observation) to

estimate the probability in denominators of (4.2) and (3.6) at each time (for example, at each visit), then plug in the estimated probabilities to (4.2) and (3.6) Hernán, Brumback and Robins [2000, 2001]. We assume throughout that the model to estimate denominator probabilities in the IPWs is correctly specified. In practice, investigators may want to explore the sensitivity of the regression coefficients to different model specifications for estimating the weights.

3.2.3 Weighted-Psuedo-Partial-Likelihood

In this section we consider two weighted-pseudo-partial-likelihoods (WPPLs) which form the basis for obtaining consistent estimators of β_0 in the presence of case-cohort sampling. The WPPLs are formed by weighting individual contributions to the usual partial likelihoods by $W_i(t)$ assuming that $W_i(t)$ is known. In the case-cohort setting, we consider a set of individuals $\tilde{\mathcal{C}}$ of size $\tilde{n} \leq n$ that is randomly selected without replacement from the entire cohort $\{1, \ldots, n\}$.

The log-WPPL created by individual-time-specific weights at time t under the full cohort setting is given by

$$l(\beta, t; W) = \sum_{i=1}^{n} \int_0^t W_i(u)\Big[\beta' A_i(u) - \log \sum_{l=1}^{n} W_l(u) Y_l(u) r\{\beta' A_l(u)\}\Big] dN_i(u), \tag{3.7}$$

which is motivated by the weighted estimating equations proposed by Robins [1993].
The log-WPPL in the case-cohort setting is

$$\tilde{l}(\beta, t; W) = \sum_{i=1}^{n} \int_0^t W_i(u)\Big[\beta' A_i(u) - \log \sum_{l \in \tilde{\mathcal{C}}} W_l(u) Y_l(u) r\{\beta' A_l(u)\}\Big] dN_i(u). \tag{3.8}$$

Note (3.8) is slightly different from the log-WPPL proposed by Cole et al. [2012], which is

$$l^*(\beta, t; W) = \sum_{i=1}^{n} \int_0^t W_i(u)\Big[\beta' A_i(u) - \log \sum_{l \in \tilde{\mathcal{C}} \cup \{i\}} W_l(u) Y_l(u) r\{\beta' A_l(u)\}\Big] dN_i(u). \tag{3.9}$$

The log-WPPLs (3.8) and (3.9) differ only in whether a case outside the subcohort $\tilde{\mathcal{C}}$ contributes to the risk set. In the absence of weights, i.e., $W_i(u) = 1$ for all i and u, (3.8) reduces to the log-likelihood considered by Self and Prentice [1988] and (3.9) reduces to the log-likelihood considered by Prentice [1986*b*]. Estimators that maximize (3.8) or (3.9) will be shown to converge in probability to β_0.

Note that under (3.1) each (observed) counting process $N_i(\cdot)(i = 1, ..., n)$ can be uniquely decomposed into the sum of its intensity process λ_i and a local square integrable martingale M_i, i.e.,

$$N_i(t) = \int_0^t \lambda_i(u)du + M_i(t), \qquad t \in [0,1], \tag{3.10}$$

where the intensity process is given by

$$\lambda_i(t) = Y_i(t)r\{\beta_0' A_i(t)\}\lambda_0(t), \tag{3.11}$$

which embodies the same parameters as in (3.4).

Define $\hat{\beta}$, $\tilde{\beta}$, and β^* to be solutions to $\partial l(\beta, 1; \hat{W})/\partial\beta = 0$, $\partial \tilde{l}(\beta, 1; \hat{W})/\partial\beta = 0$, and $\partial l^*(\beta, 1; \hat{W})/\partial\beta = 0$, respectively. Consider the following processes

$$\begin{aligned} X(\beta, t; W) &= n^{-1}\{l(\beta, t; W) - l(\beta_0, t; W)\} \qquad (3.12)\\ &= n^{-1}\sum_{i=1}^{n}\int_0^t W_i(u)\Big[(\beta - \beta_0)' A_i(u) \\ &\qquad - \log\frac{\sum_{l=1}^n W_l(u)Y_l(u)r\{\beta' A_l(u)\}}{\sum_{l=1}^n W_l(u)Y_l(u)r\{\beta_0' A_l(u)\}}\Big]dN_i(u), \\ \tilde{X}(\beta, t; W) &= n^{-1}\{\tilde{l}(\beta, t; W) - \tilde{l}(\beta_0, t; W)\} \qquad (3.13)\\ &= n^{-1}\sum_{i\in C}\int_0^t W_i(u)\Big[(\beta - \beta_0)' A_i(u) \\ &\qquad - \log\frac{\sum_{l\in\tilde{C}} W_l(u)Y_l(u)r\{\beta' A_l(u)\}}{\sum_{l\in\tilde{C}} W_l(u)Y_l(u)r\{\beta_0' A_l(u)\}}\Big]dN_i(u) \end{aligned}$$

corresponding to (3.7) and (3.8) respectively. We will first show that $X(\beta, t; \hat{W})$ and (3.12) are asymptotically equivalent, and so are $\tilde{X}(\beta, t; \hat{W})$ and (3.13). Thus, further technical developments will be made based on (3.12) and (3.13). We then show that (3.12) and (3.13) at $t = 1$ converge in probability to functions of β which are concave with a unique

maximum β_0 under certain conditions. Using the same argument as in Andersen and Gill [1982], it follows that $\hat{\beta} \to_p \beta_0$ and $\tilde{\beta} \to_p \beta_0$. That $\beta^* \to_p \beta_0$ can be shown analogously by using $X^*(\beta, t; W) = n^{-1}\{l^*(\beta, t; W) - l^*(\beta_0, t; W)\}$. Asymptotic normality of $\hat{\beta}$ and $\tilde{\beta}$ will be shown via asymptotic normality of score statistics corresponding to (3.7) and (3.8).

3.3 Consistency

For a $p \times 1$ column vector c, let $c^{\otimes 0} = 1$, $c^{\otimes 1} = c$, and $c^{\otimes 2} = cc'$, c_i denote i-th element of c, and c_{ij} denote (i, j) element of $c^{\otimes 2}$. Norms are defined by $\|c^{\otimes 2}\| = \sup_{i,j} |c_{ij}|$, $\|c\| = \sup_i |c_i|$, and $|c| = (\sum c_i^2)^{1/2} = (c'c)^{1/2}$. Also let $r^{(0)}\{\beta' A(t)\} = r\{\beta' A(t)\}$, $r^{(1)}\{\beta' A(t)\} = A(t) r\{\beta' A(t)\}$, and $r^{(2)}\{\beta' A(t)\} = A(t)^{\otimes 2} r\{\beta' A(t)\}$.

CONDITIONS.

A (Uniform consistency of estimated weights)

$$\sup_{\substack{i \in \{1,\ldots,n\} \\ t \in [0,1]}} |\hat{W}_i(t) - W_i(t)| \equiv \mathsf{M}_{\hat{W}} \to_p 0.$$

Along with the assumption of no misspecification of the model used to estimate denominator probabilities in $W(\cdot)$, the finite number of jumps assumption on the treatment and censoring processes are sufficient for this condition to hold. From a practical point of view, having a finite number of time points when treatment status can change or when censoring might occur may be reasonable to assume in many settings. For instance, studies often have planned visits at finite discrete intervals when a patient may have treatment altered. Similarly, the censoring time for a subject is often assumed to be the last observed visit time before the subject became lost-to-follow-up.

B (Stability of weights) Individual time-specific weights $W_i(t)$ and the corresponding estimators $\hat{W}_i(t)$ are strictly positive and bounded, i.e., there exist positive real numbers M_1 and M_2 such that

$$\sup_{\substack{i \in \{1,\ldots,n\} \\ t \in [0,1]}} W_i(t) \le \mathsf{M}_1, \quad \text{and} \quad \sup_{\substack{i \in \{1,\ldots,n\} \\ t \in [0,1]}} \hat{W}_i(t) \le \mathsf{M}_2.$$

Note that $\hat{W}(\cdot)$ and $W(\cdot)$ are assumed to be predictable with respect to the filtration $\mathcal{F}_t$ because weights are determined by predictable processes: $A(\cdot)$, $L(\cdot)$, and their histories. All weights discussed in §3.2.2 satisfy the conditions A and B under any circumstances, except the unstabilized weights. Unstabilized weights satisfy conditions A and B under the assumption of finite support of $A(\cdot)$ and $C(\cdot)$.

C (Finite interval) $\int_0^1 \lambda_0(t)dt < \infty$

D (Asymptotic stability)

(i) There exists a neighborhood $\mathcal{B}_0$ of β_0 and functions $s^{(0)}, s^{(1)}$, and $s^{(2)}$ defined on $\mathcal{B}_0 \times [0,1]$ such that

$$\sup_{\substack{\beta \in \mathcal{B}_0 \\ t \in [0,1]}} \|S^{(j)}(\beta,t) - s^{(j)}(\beta,t)\| \to_p 0, \qquad j = 0,1,2$$

where $S^{(j)}(\beta,t) = n^{-1}\sum_{i=1}^n Y_i(t) r^{(j)}\{\beta' A_i(t)\}$ for $j = 0,1,2$, which are the same quantities as given in Andersen and Gill [1982] with covariates $Z_i(t)$ being replaced by the treatment process $A_i(t)$.

(ii) Let $S^{(j)}_{W_{(k)}} = n^{-1}\sum_{i=1}^n W_i(t)^k Y_i(t) r^{(j)}\{\beta' A_i(t)\}$ for $j = 0,1,2$ and $k = 1,2$. There exists a neighborhood $\mathcal{B}$ of β_0, $\mathcal{B} \subseteq \mathcal{B}_0$, and functions $s^{(j)}_{W_{(k)}}$ defined on $\mathcal{B} \times [0,1]$ such that

$$\sup_{\substack{\beta \in \mathcal{B} \\ t \in [0,1]}} \|S^{(j)}_{W_{(k)}}(\beta,t) - s^{(j)}_{W_{(k)}}(\beta,t)\| \to_p 0, \qquad j = 0,1,2; k = 1,2$$

(iii) $S^{(0)}_{W_{(1)}}(\beta_0,t)$ converges in distribution to a mean zero Gaussian random variable uniformly in t, i.e.,

$$n^{1/2}\{S^{(0)}_{W_{(1)}}(\beta_0,t) - s^{(0)}_{W_{(1)}}(\beta_0,t)\} \to_d N(0,\sigma^2(t)), \quad \text{uniformly in } t \in [0,1],$$

for some $\sigma^2(t)$.

E (Lindeberg condition) For any $\epsilon > 0$, $j = 1, ..., p$

$$n^{-1} \int_0^1 \sum_{i=1}^{n} W_i(u)^2 [A_{ij}(u) - E_{W_{(1)}}(\beta_0, u)_j]^2 Y_i(u) r\{\beta_0' A_i(u)\}$$
$$\times I\{n^{-1/2} W_i(u) | A_{ij}(u) - E_{W_{(1)}}(\beta_0, u)_j| > \epsilon\} \lambda_0(u) du \to_p 0$$

where $E = S^{(1)}/S^{(0)}$, $E_{W_{(k)}} = S^{(1)}_{W_{(k)}}/S^{(0)}_{W_{(k)}}$, and in general c_j denotes the jth component of any $p \times 1$ vector c.

If the treatment process $A(\cdot)$ is bounded (as assumed throughout this paper) and Conditions B and F are satisfied, then Condition E holds trivially.

F (Asymptotic regularity conditions) $s^{(j)}(\beta, t)$ and $s^{(j)}_{W_{(k)}}(\beta, t)$ are continuous functions of $\beta \in \mathcal{B}$ uniformly in $t \in [0, 1]$ that are bounded on $\mathcal{B} \times [0, 1]$ for $j = 0, 1, 2$ and $k = 1, 2$. For all $(\beta, t) \in \mathcal{B} \times [0, 1]$, define

$$s^{(m+1)}(\beta, t) = \frac{\partial s^{(m)}(\beta, t)}{\partial \beta}, \quad s^{(m+1)}_{W_{(k)}}(\beta, t) = \frac{\partial s^{(m)}_{W_{(k)}}(\beta, t)}{\partial \beta}$$

for $m = 0, 1$, and $e = s^{(1)}/s^{(0)}, e_{W_{(k)}} = s^{(1)}_{W_{(k)}}/s^{(0)}_{W_{(k)}}, v = s^{(2)}/s^{(0)} - e^{\otimes 2}, v_{W_{(k)}} = s^{(2)}_{W_{(k)}}/s^{(0)}_{W_{(k)}} - e^{\otimes 2}_{W_{(k)}}$, and $V_{W_{(k)}} = S^{(2)}_{W_{(k)}}/S^{(0)}_{W_{(k)}} - E^{\otimes 2}_{W_{(k)}}$ for $k = 1, 2$. Assume that $s^{(0)}$ and $s^{(0)}_{W_{(k)}}$ are bounded away from zero and the matrices

$$\Sigma = \int_0^1 v(\beta_0, t) s^{(0)}(\beta_0, t) \lambda_0(t) dt, \quad \text{and}$$
$$\Sigma_{W_{(k)}} = \int_0^1 v_{W_{(k)}}(\beta_0, t) s^{(0)}_{W_{(k)}}(\beta_0, t) \lambda_0(t) dt$$

are positive definite.

Note $e_{W_{(k)}}$ can be interpreted as the weighted average of a treatment function with the weights taking an exponential form. The positive definite condition on Σ in Andersen and Gill [1982] can easily be extended to the $\Sigma_{W_{(k)}}$ assuming $W(t)$ are bounded away from zero on $t \in [0, 1]$.

Conditions A-F are sufficient to prove consistency of $\hat{\beta}$. To prove the consistency of $\tilde{\beta}$ and β^*, the following additional condition is required to ensure asymptotic behavior of certain subcohort averages.

G (Stability of subcohort average) Define

$$\tilde{S}^{(j)}_{W_{(k)}}(\beta,t) = \tilde{n}^{-1}\sum_{i\in\tilde{\mathcal{C}}} W_i(t)^k Y_i(t) r^{(j)}\{\beta' A_i(t)\},$$

and $\tilde{E}_{W_{(k)}} = \tilde{S}^{(1)}_{W_{(k)}}/\tilde{S}^{(0)}_{W_{(k)}}$ for $j = 0,1,2$ and $k = 1,2$.

(i) (Nontrivial subcohort) $\tilde{n}n^{-1} \to_p \alpha$ for some $\alpha \in (0,1]$.

(ii) (Asymptotic normality of subcohort averages at β_0) For any $\epsilon > 0$

$$\sup_{t\in[0,1]} n^{-1}\sum_{i=1}^{n} W_i(t)^2 Y_i(t) r\{\beta_0' A_i(t)\}^2 I\{n^{-1/2} W_i(t) Y_i(t) r\{\beta_0' A_i(t)\} > \epsilon\} \to_p 0,$$
$$\sup_{t\in[0,1]} n^{-1}\sum_{i=1}^{n} W_i(t)^2 Y_i(t) \|r^{(1)}\{\beta_0' A_i(t)\}\|^2 I\{n^{-1/2} W_i(t) Y_i(t) \|r^{(1)}\{\beta_0' A_i(t)\}\| > \epsilon\}$$
$$\to_p 0,$$

and the sequences of distributions of $n^{1/2}\{\tilde{E}(\beta_0,t) - E(\beta_0,t)\}$ are tight on the product space of cadlag functions equipped with the product Skorohod topology and so are $n^{1/2}\{\tilde{E}_{W_{(1)}}(\beta_0,t) - E_{W_{(1)}}(\beta_0,t)\}$.

(iii) (Asymptotic stability and regularity of covariance function) There exists a neighborhood $\mathcal{B}$ of β_0 and functions $q^{(j)}(\beta,t,u)$ for $j = 0,1,2$, defined on $\mathcal{B}\times[0,1]^2$ such that $q^{(j)}(\beta,t,u)$ are continuous functions of $\beta \in \mathcal{B}$ uniformly in $(t,u) \in [0,1]^2$, the $q^{(j)}$ are bounded on $\mathcal{B}\times[0,1]^2$ and

$$\sup_{\substack{\beta\in\mathcal{B}\\(t,u)\in[0,1]^2}} \|Q^{(j)}(\beta,t,u) - q^{(j)}(\beta,t,u)\| \to_p 0, \quad j = 0,1,2, \quad \text{where}$$
$$Q^{(0)}(\beta,t,u) = n^{-1}\sum_{i=1}^{n} W_i(t) Y_i(t) r\{\beta_0' A_i(t)\} W_i(u) Y_i(u) r\{\beta_0' A_i(u)\},$$
$$Q^{(1)}(\beta,t,u) = n^{-1}\sum_{i=1}^{n} W_i(t) Y_i(t) r^{(1)}\{\beta_0' A_i(t)\} W_i(u) Y_i(u) r^{(1)}\{\beta_0' A_i(u)\}',$$
$$Q^{(2)}(\beta,t,u) = n^{-1}\sum_{i=1}^{n} W_i(t) Y_i(t) r\{\beta_0' A_i(t)\} W_i(u) Y_i(u) r^{(1)}\{\beta_0' A_i(u)\}.$$

Moreover, $\sup_{n\geq 1}\mathcal{E}[Q^{(j)}(\beta,t,u)]$ for $j = 0,1,2$ are bounded sequences where $\mathcal{E}$ denote expectation.

(iv) (Asymptotic stability of subcohort averages) Let $\tilde{Q}^{(j)}(\beta,t,u)$ be covariance functions based on subcohort members $i = 1,...,\tilde{n}$. Then

$$\sup_{\substack{\beta\in\mathcal{B}\\ t\in[0,1]}} \|\tilde{S}^{(0)}_{W_{(k)}}(\beta,t) - s^{(0)}_{W_{(k)}}(\beta,t)\| \to_p 0 \quad k = 1,2,$$

i.e., the subcohort average converges to the mean of the full cohort, and

$$\sup_{\substack{\beta\in\mathcal{B}\\ (t,u)\in[0,1]^2}} \|\tilde{Q}^{(j)}(\beta,t,u) - q^{(j)}(\beta,t,u)\| \to_p 0, \quad j = 0,1,2.$$

i.e., the subcohort covariance functions converge in probability to the full cohort covariance functions. In addition, $\tilde{S}^{(0)}_{W_{(1)}}(\beta_0,\cdot)$ converges in distribution to a mean zero Gaussian random variable uniformly in t, i.e.,

$$n^{1/2}\{\tilde{S}^{(0)}_{W_{(1)}}(\beta_0,t) - s^{(0)}_{W_{(1)}}(\beta_0,t)\} \to_d N(0,\tilde{\sigma}^2(t)), \quad \text{uniformly in } t \in [0,1]$$

for some $\tilde{\sigma}^2(t)$.

Condition G is the same as condition G in Self and Prentice [1988], incorporating individual-specific time-varying weights $W_i(t)(i = 1,...,n)$.

Theorem 3.3.1. *(Consistency of $\hat{\beta}$ under full cohort) Under conditions A-F, $\hat{\beta} \to_p \beta_0$.*

Proof. Consider the process $X(\beta,t;W)$ given by (3.12) and its compensator counterpart $K(\beta,t;W)$ which is given by

$$K(\beta,t;W) = n^{-1}\sum_{i=1}^{n}\int_0^t W_i(u)\Big[(\beta-\beta_0)'A_i(u) - \log\Big\{\frac{S^{(0)}_{W_{(1)}}(\beta,u)}{S^{(0)}_{W_{(1)}}(\beta_0,u)}\Big\}\Big]\lambda_i(u)du$$

where $\lambda_i(t)$ is given as in (3.11). We start by showing that

$$|\{X(\beta,t;\hat{W}) - K(\beta,t;\hat{W})\} - \{X(\beta,t;W) - K(\beta,t;W)\}| \to_p 0 \tag{3.14}$$

so that we can consider the asymptotic behavior of $X(\beta,t;W) - K(\beta,t;W)$ instead of $X(\beta,t;\hat{W}) - K(\beta,t;\hat{W})$ to prove consistency of $\hat{\beta}$. To prove (3.14), first note the term

$|\{X(\beta,t;\hat{W}) - K(\beta,t;\hat{W})\} - \{X(\beta,t;W) - K(\beta,t;W)\}|$ in (3.14) equals

$$\Big|n^{-1}\sum_{i=1}^{n}\int_0^1\Big[\hat{W}_i(u)(\beta-\beta_0)'A_i(u) - \hat{W}_i(u)\log\Big\{\frac{S^{(0)}_{\hat{W}_{(1)}}(\beta,u)}{S^{(0)}_{\hat{W}_{(1)}}(\beta_0,u)}\Big\}\Big]dM_i(u)$$
$$- n^{-1}\sum_{i=1}^{n}\int_0^1\Big[W_i(u)(\beta-\beta_0)'A_i(u) - W_i(u)\log\Big\{\frac{S^{(0)}_{W_{(1)}}(\beta,u)}{S^{(0)}_{W_{(1)}}(\beta_0,u)}\Big\}\Big]dM_i(u)\Big|.$$

Replacing $W_i(u)$ in front of $\log\{S^{(0)}_{W_{(1)}}(\beta,u)/S^{(0)}_{W_{(1)}}(\beta_0,u)\}$ with $W_i(u) - \hat{W}_i(u) + \hat{W}_i(u)$ and rearranging terms yields

$$\Big|n^{-1}\sum_{i=1}^{n}\int_0^1\{\hat{W}_i(u) - W_i(u)\}(\beta-\beta_0)'A_i(u)dM_i(u) \tag{3.15}$$
$$- n^{-1}\sum_{i=1}^{n}\int_0^1\{\hat{W}_i(u) - W_i(u)\}\log\Big\{\frac{S^{(0)}_{W_{(1)}}(\beta,u)}{S^{(0)}_{W_{(1)}}(\beta_0,u)}\Big\}dM_i(u)$$
$$- n^{-1}\sum_{i=1}^{n}\int_0^1\hat{W}_i(u)\log\Big\{\frac{S^{(0)}_{\hat{W}_{(1)}}(\beta,u)}{S^{(0)}_{\hat{W}_{(1)}}(\beta_0,u)}\Big/\frac{S^{(0)}_{W_{(1)}}(\beta,u)}{S^{(0)}_{W_{(1)}}(\beta_0,u)}\Big\}dM_i(u)\Big|.$$

Each term in (3.15) is a local square integrable martingale since $g(W_i(\cdot), A_i(\cdot))$ is predictable for any continuous function $g(\cdot)$ due to predictableness of $W_i(\cdot)$ and $A_i(\cdot)$. Because $\hat{W}_i(\cdot)$ is also bounded and predictable, the same argument can be made for $g_1(\hat{W}_i(\cdot), A_i(\cdot))$ and $g_2(W_i(\cdot), \hat{W}_i(\cdot))$ for any continuous functions $g_1(\cdot)$ and $g_2(\cdot)$. We will show that the variance process of each martingale in (3.15) converges in probability to zero, thus proving (3.14).

Let $B_1(\beta,t)$ be the variance process of the first martingale in (3.15). Then

$$B_1(\beta,t) = n^{-2}\sum_{i=1}^{n}\int_0^t\{\hat{W}_i(u) - W_i(u)\}^2(\beta-\beta_0)'A_i(u)^{\otimes 2}(\beta-\beta_0)\lambda_i(u)du$$
$$= n^{-2}\sum_{i=1}^{n}\int_0^t\{\hat{W}_i(u) - W_i(u)\}^2(\beta-\beta_0)'Y_i(u)r^{(2)}\{\beta_0'A_i(u)\}(\beta-\beta_0)\lambda_0(u)du$$
$$\le n^{-1}\int_0^t \mathsf{M}^2_{\hat{W}}(\beta-\beta_0)'\Big[n^{-1}\sum_{i=1}^{n}Y_i(u)r^{(2)}\{\beta_0'A_i(u)\}\Big](\beta-\beta_0)\lambda_0(u)du$$
$$= n^{-1}\mathsf{M}^2_{\hat{W}}\int_0^t(\beta-\beta_0)'S^{(2)}(\beta_0,u)(\beta-\beta_0)\lambda_0(u)du$$

which converges in probability to zero due to conditions A, B, D, and F. The second equality is owing to (3.11), and the inequality comes from replacing $\{\hat{W}_i(u) - W_i(u)\}^2$ by its

supremum value $\mathsf{M}^2_{\hat{W}}$. Let $B_2(\beta,t)$ be the variance process of the second martingale term in (3.15). Then

$$\begin{aligned}
B_2(\beta,t) =& n^{-2}\sum_{i=1}^{n}\int_0^t \left\{\hat{W}_i(u)-W_i(u)\right\}^2\left\{\log S^{(0)}_{W_{(1)}}(\beta,u)-\log S^{(0)}_{W_{(1)}}(\beta_0,u)\right\}^2\lambda_i(u)du\\
\le& n^{-1}\int_0^t \mathsf{M}^2_{\hat{W}}\left\{\log S^{(0)}_{W_{(1)}}(\beta,u)-\log S^{(0)}_{W_{(1)}}(\beta_0,u)\right\}^2 S^{(0)}(\beta_0,u)\lambda_0(u)du
\end{aligned}$$

which converges to zero due to conditions A, B, D, and F. Lastly, let the variance of the third martingale term in (3.15) be $B_3(\beta,t)$. Then

$$\begin{aligned}
B_3(\beta,t) =& n^{-2}\sum_{i=1}^{n}\int_0^t \hat{W}_i(u)^2\Big[\{\log S^{(0)}_{\hat{W}_{(1)}}(\beta,u)-\log S^{(0)}_{W_{(1)}}(\beta,u)\}\\
&-\{\log S^{(0)}_{\hat{W}_{(1)}}(\beta_0,u)-\log S^{(0)}_{W_{(1)}}(\beta_0,u)\}\Big]^2\lambda_i(u)du\\
=& n^{-1}\int_0^1\Big[\{\log S^{(0)}_{\hat{W}_{(1)}}(\beta,u)-\log S^{(0)}_{W_{(1)}}(\beta,u)\}\\
&-\{\log S^{(0)}_{\hat{W}_{(1)}}(\beta_0,u)-\log S^{(0)}_{W_{(1)}}(\beta_0,u)\}\Big]^2 S^{(0)}_{\hat{W}_{(2)}}(\beta_0,u)\lambda_0(u)du\\
\le& n^{-1}\int_0^1\Big[\sup_{\beta,u}|\log S^{(0)}_{\hat{W}_{(1)}}(\beta,u)-\log S^{(0)}_{W_{(1)}}(\beta,u)|^2\\
&+2\sup_{\beta,u}|\log S^{(0)}_{\hat{W}_{(1)}}(\beta,u)-\log S^{(0)}_{W_{(1)}}(\beta,u)|\sup_{u}|\log S^{(0)}_{\hat{W}_{(1)}}(\beta_0,u)-\log S^{(0)}_{W_{(1)}}(\beta_0,u)|\\
&+\sup_{u}|\log S^{(0)}_{\hat{W}_{(1)}}(\beta_0,u)-\log S^{(0)}_{W_{(1)}}(\beta_0,u)|^2\Big]S^{(0)}_{\hat{W}_{(2)}}(\beta_0,u)\lambda_0(u)du
\end{aligned}$$

which converges in probability to zero due to conditions A, B, D, F, and by the continuous mapping theorem. It follows that $X(\beta,t;\hat{W})-K(\beta,t;\hat{W})$ and $X(\beta,t;W)-K(\beta,t;W)$ in (3.14) are asymptotically equivalent processes. Thereby we proceed to describe asymptotic behavior of the process $X(\beta,t;W)-K(\beta,t;W)$. Hereafter for notation convenience we suppress W when writing $X(\beta,t;W)$ and $K(\beta,t;W)$.

Now consider $X(\beta,t)-K(\beta,t)$, which equals to

$$n^{-1}\sum_{i=1}^{n}\int_0^t W_i(u)\Big[(\beta-\beta_0)'A_i(u)-\log\Big\{\frac{S^{(0)}_{W_{(1)}}(\beta,u)}{S^{(0)}_{W_{(1)}}(\beta_0,u)}\Big\}\Big]dM_i(u),$$

which is a martingale. After some calculation, it can be shown that its variance process $B(\beta,t)$ can be simplified as

$$n^{-1}\int_0^1 \Big[(\beta-\beta_0)' S^{(2)}_{W_{(2)}}(\beta_0,u)(\beta-\beta_0) \tag{3.16}$$
$$-2(\beta-\beta_0)' S^{(1)}_{W_{(2)}}(\beta_0,u)\log\Big\{\frac{S^{(0)}_{W_{(1)}}(\beta,u)}{S^{(0)}_{W_{(1)}}(\beta_0,u)}\Big\}$$
$$+\left\{\log\Big(\frac{S^{(0)}_{W_{(1)}}(\beta,u)}{S^{(0)}_{W_{(1)}}(\beta_0,u)}\Big)\right\}^2 S^{(0)}_{W_{(2)}}(\beta_0,u)\Big]\lambda_0(u)du$$

where each term inside the integral converges in probability to a function of finite quantities $s^{(j)}_{W_{(k)}}$ on $\beta \in \mathcal{B}$ in view of conditions D and F. Therefore, (3.16) converges in probability to zero. It follows that $X(\beta,t)$ and $K(\beta,t)$ converge in probability to the same limit by the Lenglart inequality, i.e., that $\mathrm{pr}[\sup_{t,\beta}\|X(\beta,t)-K(\beta,t)>\eta\|] \le \delta/\eta^2 + \mathrm{pr}[B(\beta,1)>\delta]$ for all $\delta,\eta>0$.

Therefore, to investigate asymptotic properties of $X(\beta,1)$, consider asymptotic properties of $K(\beta,1)$ instead:

$$K(\beta,1) \to_p \int_0^1 \Big[(\beta-\beta_0)' s^{(1)}_{W_{(1)}}(\beta_0,u) - \log\Big\{\frac{s^{(0)}_{W_{(1)}}(\beta,u)}{s^{(0)}_{W_{(1)}}(\beta_0,u)}\Big\} s^{(0)}_{W_{(1)}}(\beta_0,u)\Big]\lambda_0(u)du$$

by (3.11). Let $K_l(\beta,1)$ be the limiting quantity shown in the above. Then

$$\frac{\partial K_l(\beta,1)}{\partial\beta} = \int_0^1 \Big[s^{(1)}_{W_{(1)}}(\beta_0,u) - \frac{s^{(1)}_{W_{(1)}}(\beta,u)}{s^{(0)}_{W_{(1)}}(\beta,u)} s^{(0)}_{W_{(1)}}(\beta_0,u)\Big]\lambda_0(u)du$$

which is zero at $\beta=\beta_0$. In addition, $\partial^2 K_l(\beta,1)/\partial\beta^2$ is

$$-\int_0^1 \Big[\frac{s^{(2)}_{W_{(1)}}(\beta,u)s^{(0)}_{W_{(1)}}(\beta,u) - s^{(1)}_{W_{(1)}}(\beta,u)^{\otimes 2}}{s^{(0)}_{W_{(1)}}(\beta,u)^2}\Big] s^{(0)}_{W_{(1)}}(\beta_0,u)\lambda_0(u)du$$
$$= -\int_0^1 v_{W_{(1)}}(\beta,u) s^{(0)}_{W_{(1)}}(\beta_0,u)\lambda_0(u)du$$

which equals to $-\Sigma_{W_{(1)}}$ and is negative definite when $\beta=\beta_0$ based on condition F. Therefore $K(\beta,1)$ converges to a concave function having unique maximum at β_0. This enables us to make use of Theorem II.1 in Andersen and Gill [1982] that proves in probability convergence

of $X(\beta,1)$ to the same concave function of β as does $K(\beta,1)$, with a unique maximum at $\beta = \beta_0$. Then $\hat{\beta} \to_p \beta_0$. □

Consistency of $\tilde{\beta}$ can be shown using similar arguments as in Theorem 3.3.1. It can immediately be seen that $\tilde{X}(\beta,t;\hat{W})$ is asymptotically equivalent to $\tilde{X}(\beta,t;W)$ by condition A. Therefore, we can show that $\tilde{X}(\beta,t)$ converges in probability to $K(\beta,t)$ so that the same argument as in the proof of Theorem 3.3.1 can be made. In particular, $|\tilde{X}(\beta,t) - K(\beta,t)|$ will be decomposed into two terms, $|X(\beta,t) - K(\beta,t)|$ plus a term that will converge in probability to zero.

Theorem 3.3.2. *(Consistency of $\tilde{\beta}$ under the case-cohort) Under conditions A-G, $\tilde{\beta} \to_p \beta_0$.*

Proof. First, $|\tilde{X}(\beta,t) - K(\beta,t)|$ can be rewritten as

$$
\begin{aligned}
&\Big| n^{-1} \int_0^t \sum_{i=1}^n W_i(u)(\beta - \beta_0)' A_i(u) dM_i(u) \\
&\quad - n^{-1} \int_0^t \sum_{i=1}^n W_i(u) \log\Big\{ \frac{\tilde{S}^{(0)}_{W_{(1)}}(\beta,u)}{\tilde{S}^{(0)}_{W_{(1)}}(\beta_0,u)} \Big\} dN_i(u) \\
&\quad + n^{-1} \int_0^t \sum_{i=1}^n W_i(u) \log\Big\{ \frac{S^{(0)}_{W_{(1)}}(\beta,u)}{S^{(0)}_{W_{(1)}}(\beta_0,u)} \Big\} \lambda_i(u) du \Big| \\
\le & |X(\beta,t) - K(\beta,t)| \\
&\quad + \Big| n^{-1} \int_0^t \sum_{i=1}^n W_i(u) \Big\{ \log\Big(\frac{\tilde{S}^{(0)}_{W_{(1)}}(\beta,u)}{\tilde{S}^{(0)}_{W_{(1)}}(\beta_0,u)} \Big) - \log\Big(\frac{S^{(0)}_{W_{(1)}}(\beta,u)}{S^{(0)}_{W_{(1)}}(\beta_0,u)} \Big) \Big\} dN_i(u) \Big|.
\end{aligned}
$$

We have shown that $|X(\beta,t) - K(\beta,t)| \to_p 0$. The remaining term can be decomposed as

$$
\begin{aligned}
&\Big| n^{-1} \int_0^t \Big[\sum_{i=1}^n W_i(u) \Big\{ \log\Big(\frac{\tilde{S}^{(0)}_{W_{(1)}}(\beta,u)}{\tilde{S}^{(0)}_{W_{(1)}}(\beta_0,u)} \Big) - \log\Big(\frac{S^{(0)}_{W_{(1)}}(\beta,u)}{S^{(0)}_{W_{(1)}}(\beta_0,u)} \Big) \Big\} dM_i(u) \Big] \\
&+ n^{-1} \int_0^t \sum_{i=1}^n \Big[W_i(u) \Big\{ \log\Big(\frac{\tilde{S}^{(0)}_{W_{(1)}}(\beta,u)}{\tilde{S}^{(0)}_{W_{(1)}}(\beta_0,u)} \Big) - \log\Big(\frac{S^{(0)}_{W_{(1)}}(\beta,u)}{S^{(0)}_{W_{(1)}}(\beta_0,u)} \Big) \Big\} \lambda_i(u) du \Big] \Big|.
\end{aligned}
\tag{3.17}
$$

Then the second term in (3.17) can easily be shown to converge in probability to zero in view of conditions C, D, F and G(iv). Also the martingale in (3.17) converges in probability to zero because its variance process is

$$
\begin{aligned}
&\Big| n^{-2} \int_0^t \sum_{i=1}^n W_i(u)^2 \Big[\Big\{ \log \tilde{S}^{(0)}_{W_{(1)}}(\beta, u) - \log S^{(0)}_{W_{(1)}}(\beta, u) \Big\} \\
&\quad - \Big\{ \log \tilde{S}^{(0)}_{W_{(1)}}(\beta_0, u) - \log S^{(0)}_{W_{(1)}}(\beta_0, u) \Big\} \Big]^2 \lambda_i(u) du \Big| \\
\leq &\Big| n^{-1} \int_0^t \Big[\sup_{\beta, u} |\log \tilde{S}^{(0)}_{W_{(1)}}(\beta, u) - \log S^{(0)}_{W_{(1)}}(\beta, u)| \\
&\quad + \sup_u |\log \tilde{S}^{(0)}_{W_{(1)}}(\beta_0, u) - \log S^{(0)}_{W_{(1)}}(\beta_0, u)| \Big]^2 S^{(0)}_{W_{(2)}}(\beta_0, u) \lambda_0(u) du \Big|
\end{aligned}
$$

which converges in probability to zero, again by (3.11) with conditions C, D, F and G(iv). Note that sum of supremums in the integrand (which can be taken outside the integral) converges in probability to zero by conditions D and G(iv). □

It is straightforward to show that the estimator based on (3.9) converges in probability to the same limit as $\tilde{\beta}$. An individual case's contribution to $\tilde{C}$ at its failure time (which is weighted by its IPWs) is asymptotically negligible in the sense that IPWs are bounded at all times and weighted subcohort averages are asymptotically stable (conditions B and G(iv)). This is formally stated in the following Theorem.

Theorem 3.3.3. *Under conditions A-G,* $\tilde{\beta} - \beta^* \to_p 0$.

Proof. We sketch a proof of Theorem 3.3.3. Consider the following process

$$X^*(\beta, t) = n^{-1}\{l^*(\beta, t) - l^*(\beta_0, t)\}.$$

Then $X^*(\beta, t) = n^{-1}\{\tilde{l}(\beta, t) - \tilde{l}(\beta_0, t)\} + o_p(1)$ because $n^{-1} l^*(\beta, t) = n^{-1}\tilde{l}(\beta, t) + o_p(1)$. Therefore, $X^*(\beta, t)$ and $\tilde{X}(\beta, t)$ are asymptotically equivalent processes and we can repeat the proof of Theorem 3.3.2 using $X^*(\beta, t)$ instead of $\tilde{X}(\beta, t)$. □

3.4 Asymptotic Normality

To prove asymptotic normality of $\tilde{\beta}$, we first prove asymptotic normality of the score process of the log WPPL for the full cohort setting.

Theorem 3.4.1. *(Asymptotic normality of the full cohort MSCM score statistic) Under*

conditions A-F,

$$n^{-1/2}U(\beta_0, 1) \to_d N(0, \Sigma_U)$$

where $\Sigma_U = \Sigma_{W_{(2)}} + \Delta_{W_{(1)},W_{(2)}}$ *with*

$$\Delta_{W_{(1)},W_{(2)}} = \int_0^1 \{e_{W_{(2)}}(\beta_0, u) - e_{W_{(1)}}(\beta_0, u)\}^{\otimes 2} s^{(0)}_{W_{(2)}}(\beta_0, u)\lambda_0(u)du. \tag{3.18}$$

Proof. We will refer to the score process under the full cohort setting as *the full cohort MSCM score process.* Let $U(\beta_0, t)$ be the full cohort MSCM score process at time t. Then

$$\begin{aligned} n^{-1/2}U(\beta_0, t) &= n^{-1/2}\partial l(\beta, t)/\partial\beta\Big|_{\beta=\beta_0} \\ &= n^{-1/2}\sum_{i=1}^n \int_0^t W_i(u)\Big[A_i(u) - \frac{S^{(1)}_{W_{(1)}}(\beta_0, u)}{S^{(0)}_{W_{(1)}}(\beta_0, u)}\Big]dN_i(u) \\ &= n^{-1/2}\sum_{i=1}^n \int_0^t W_i(u)\Big[A_i(u) - E_{W_{(1)}}(\beta_0, u)\Big]dM_i(u) \end{aligned} \tag{3.19}$$

The third equality follows from (3.10) and the fact that

$$n^{-1/2}\sum_{i=1}^n \int_0^t W_i(u)[A_i(u) - E_{W_{(1)}}(\beta_0, u)]\lambda_i(u)du = 0 \tag{3.20}$$

based on (3.11). Set $H_i(t) = n^{-1/2}W_i(t)[A_i(t) - E_{W_{(1)}}(\beta_0, t)]$ for $i = 1, ..., n$. This is a locally bounded predictable process. Therefore, (3.19) is a local square integrable martingale. To apply the martingale central limit theorem to the local square integrable martingale, we show that $n^{-1/2}U(\beta_0, 1) = \sum_{i=1}^n \int_0^1 H_i(t)dM_i(t)$ satisfies (i) $\int_0^1 \sum_{i=1}^n H_{ij}(t)^2 I\{|H_{ij}(t)| > \epsilon\}\lambda_i(t)dt \to_p 0$ for any $\epsilon > 0$ (the Lindeberg condition), and that (ii) variance process of (3.19) evaluated at $t = 1$ converges in probability to a finite quantity. Condition (i) is satisfied because of condition E. To see if condition (ii) is satisfied, consider variance process of $n^{-1/2}U(\beta_0, 1)$,

$$\begin{aligned} &\sum_{i=1}^n \int_0^1 H_i(u)^{\otimes 2}\lambda_i(u)du \\ &= \int_0^1 n^{-1}\sum_{i=1}^n W_i(u)^2\Big[A_i(u) - E_{W_{(1)}}(\beta_0, u)\Big]^{\otimes 2}\lambda_i(u)du \end{aligned}$$

$$
\begin{aligned}
&= \int_0^1 n^{-1} \sum_{i=1}^{n} \Big[W_i(u)^2 Y_i(u) r^{(2)}\{\beta_0' A_i(u)\} - 2W_i(u)^2 Y_i(u) r^{(1)}\{\beta_0' A_i(u)\}\{E_{W_{(1)}}(\beta_0, u)\}' \\
&\quad + W_i(u)^2 Y_i(u) r\{\beta_0' A_i(u)\} E_{W_{(1)}}(\beta_0, u)^{\otimes 2} \Big] \lambda_0(u) du \\
&= \int_0^1 \Big[S^{(2)}_{W_{(2)}}(\beta_0, u) - 2S^{(1)}_{W_{(2)}}(\beta_0, u)\{E_{W_{(1)}}(\beta_0, u)\}' + S^{(0)}_{W_{(2)}}(\beta_0, u) E_{W_{(1)}}(\beta_0, u)^{\otimes 2} \Big] \lambda_0(u) du \\
&= \int_0^1 \Big[\frac{S^{(2)}_{W_{(2)}}(\beta_0, u)}{S^{(0)}_{W_{(2)}}(\beta_0, u)} - 2\frac{S^{(1)}_{W_{(2)}}(\beta_0, u)}{S^{(0)}_{W_{(2)}}(\beta_0, u)} \Big\{ \frac{S^{(1)}_{W_{(1)}}(\beta_0, u)}{S^{(0)}_{W_{(1)}}(\beta_0, u)} \Big\}' + \Big\{ \frac{S^{(1)}_{W_{(1)}}(\beta_0, u)}{S^{(0)}_{W_{(1)}}(\beta_0, u)} \Big\}^{\otimes 2} \Big] S^{(0)}_{W_{(2)}}(\beta_0, u) \lambda_0(u) du \\
&= \int_0^1 \Big[\Big\{ \frac{S^{(2)}_{W_{(2)}}(\beta_0, u)}{S^{(0)}_{W_{(2)}}(\beta_0, u)} - \Big(\frac{S^{(1)}_{W_{(2)}}(\beta_0, u)}{S^{(0)}_{W_{(2)}}(\beta_0, u)} \Big)^{\otimes 2} \Big\} + \Big\{ \Big(\frac{S^{(1)}_{W_{(2)}}(\beta_0, u)}{S^{(0)}_{W_{(2)}}(\beta_0, u)} \Big)^{\otimes 2} \\
&\qquad - 2\frac{S^{(1)}_{W_{(2)}}(\beta_0, u)}{S^{(0)}_{W_{(2)}}(\beta_0, u)} \Big(\frac{S^{(1)}_{W_{(1)}}(\beta_0, u)}{S^{(0)}_{W_{(1)}}(\beta_0, u)} \Big)' + \Big(\frac{S^{(1)}_{W_{(1)}}(\beta_0, u)}{S^{(0)}_{W_{(1)}}(\beta_0, u)} \Big)^{\otimes 2} \Big\} \Big] S^{(0)}_{W_{(2)}}(\beta_0, u) \lambda_0(u) du \\
&= \int_0^1 \Big[V_{W_{(2)}}(\beta_0, u) + \{E_{W_{(2)}}(\beta_0, u) - E_{W_{(1)}}(\beta_0, u)\}^{\otimes 2} \Big] S^{(0)}_{W_{(2)}}(\beta_0, u) \lambda_0(u) du.
\end{aligned}
$$

Finally we can see that the variance process of $n^{-1/2}U(\beta_0, 1)$ converges in probability to

$$\Sigma_{W_{(2)}} + \Delta_{W_{(1)}, W_{(2)}} \equiv \Sigma_U \tag{3.21}$$

where $\Delta_{W_{(1)}, W_{(2)}}$ is given in (3.18). Based on conditions C and F, (3.21) is a finite quantity. Therefore, the full cohort MSCM score statistic converges in distribution to a Gaussian process with mean zero and the limiting covariance process Σ_U by the martingale central limit theorem. When $W_i(t) \equiv 1$ for all $i = 1, ..., n$ and $t \in [0, 1]$, $\Delta_{W_{(1)}, W_{(2)}}$ becomes zero and (3.21) equals to Σ which is the asymptotic variance of the score process under the full cohort. □

The score process corresponding to (3.8), which will be referred to as *case-cohort MSCM score process*, is defined by

$$
\begin{aligned}
n^{-1/2}\tilde{U}(\beta_0, t) =& n^{-1/2} \partial \tilde{l}(\beta, t)/\partial \beta \Big|_{\beta=\beta_0} \\
=& n^{-1/2} \sum_{i=1}^{n} \int_0^t W_i(u) \Big[A_i(u) - \tilde{E}_{W_{(1)}}(\beta_0, u) \Big] dN_i(u).
\end{aligned} \tag{3.22}
$$

Replacing $\tilde{E}_{W_{(1)}}(\beta_0, u)$ in (3.22) with $E_{W_{(1)}}(\beta_0, u) + \tilde{E}_{W_{(1)}}(\beta_0, u) - E_{W_{(1)}}(\beta_0, u)$, we obtain

$$
\begin{aligned}
n^{-1/2}\tilde{U}(\beta_0,t) = n^{-1/2}\sum_{i=1}^{n}\int_0^t W_i(u)\Big[A_i(u) - E_{W_{(1)}}(\beta_0,u)\Big]dM_i(u) \\
- n^{1/2}\int_0^t \Big[\tilde{E}_{W_{(1)}}(\beta_0,u) - E_{W_{(1)}}(\beta_0,u)\Big]S^{(0)}_{W_{(1)}}(\beta_0,u)\lambda_0(u)du \\
- n^{-1/2}\sum_{i=1}^{n}\int_0^t W_i(u)\Big[\tilde{E}_{W_{(1)}}(\beta_0,u) - E_{W_{(1)}}(\beta_0,u)\Big]dM_i(u).
\end{aligned} \tag{3.23}
$$

(3.23) is equivalent to

$$
\begin{aligned}
& n^{-1/2}\sum_{i=1}^{n}\int_0^t W_i(u)\Big[A_i(u) - E_{W_{(1)}}(\beta_0,u)\Big]dM_i(u) \\
& - \int_0^t D_n(u)\lambda_0(u)du \\
& - \int_0^t D_n(u)\{S^{(0)}_{W_{(1)}}(\beta_0,u)/\tilde{S}^{(0)}_{W_{(1)}}(\beta_0,u) - 1\}\lambda_0(u)du \\
& + \int_0^t n^{1/2}\{E_{W_{(1)}}(\beta_0,u) - e_{W_{(1)}}(\beta_0,u)\}\{\tilde{S}^{(0)}_{W_{(1)}}(\beta_0,u) - S^{(0)}_{W_{(1)}}(\beta_0,u)\} \\
& \qquad \times S^{(0)}_{W_{(1)}}(\beta_0,u)/\tilde{S}^{(0)}_{W_{(1)}}(\beta_0,u)\lambda_0(u)du \\
& - n^{-1/2}\sum_{i=1}^{n}\int_0^t W_i(u)\Big[\tilde{E}_{W_{(1)}}(\beta_0,u) - E_{W_{(1)}}(\beta_0,u)\Big]dM_i(u),
\end{aligned} \tag{3.24}
$$

where

$$
D_n(t) = n^{1/2}\Big[\Big\{\tilde{S}^{(1)}_{W_{(1)}}(\beta_0,t) - S^{(1)}_{W_{(1)}}(\beta_0,t)\Big\} - e_{W_{(1)}}(\beta_0,t)\Big\{\tilde{S}^{(0)}_{W_{(1)}}(\beta_0,t) - S^{(0)}_{W_{(1)}}(\beta_0,t)\Big\}\Big].
$$

The equivalence between (3.23) and (3.24) can be shown by rewriting $n^{1/2}[E_{W_{(1)}}(\beta_0,u) - e_{W_{(1)}}(\beta_0,u)\}]S^{(0)}_{W_{(1)}}(\beta_0,u)$ in the second term of (3.23) as follows:

$$
\begin{aligned}
& n^{1/2}[\tilde{E}_{W_{(1)}}(\beta_0,t) - E_{W_{(1)}}(\beta_0,t)]S^{(0)}_{W_{(1)}}(\beta_0,t) \\
=& n^{1/2}\Big[\frac{\tilde{S}^{(1)}_{W_{(1)}}(\beta_0,t)}{\tilde{S}^{(0)}_{W_{(1)}}(\beta_0,t)} - \frac{S^{(1)}_{W_{(1)}}(\beta_0,t)}{S^{(0)}_{W_{(1)}}(\beta_0,t)}\Big]S^{(0)}_{W_{(1)}}(\beta_0,t) \\
=& n^{1/2}\Big[\Big\{\frac{\tilde{S}^{(1)}_{W_{(1)}}(\beta_0,t)}{\tilde{S}^{(0)}_{W_{(1)}}(\beta_0,t)} - \frac{S^{(1)}_{W_{(1)}}(\beta_0,t)}{\tilde{S}^{(0)}_{W_{(1)}}(\beta_0,t)}\Big\} \\
& \qquad + \Big\{\frac{S^{(1)}_{W_{(1)}}(\beta_0,t)}{\tilde{S}^{(0)}_{W_{(1)}}(\beta_0,t)} - \frac{S^{(1)}_{W_{(1)}}(\beta_0,t)}{S^{(0)}_{W_{(1)}}(\beta_0,t)}\Big\}\Big]S^{(0)}_{W_{(1)}}(\beta_0,t)
\end{aligned}
$$

$$
\begin{aligned}
=& n^{1/2}\Big[\frac{1}{\tilde{S}^{(0)}_{W_{(1)}}(\beta_0,t)}\Big\{\tilde{S}^{(1)}_{W_{(1)}}(\beta_0,t) - S^{(1)}_{W_{(1)}}(\beta_0,t)\Big\} \\
&+ \frac{S^{(1)}_{W_{(1)}}(\beta_0,t)}{\tilde{S}^{(0)}_{W_{(1)}}(\beta_0,t)S^{(0)}_{W_{(1)}}(\beta_0,t)}\Big\{S^{(0)}_{W_{(1)}}(\beta_0,t) - \tilde{S}^{(0)}_{W_{(1)}}(\beta_0,t)\Big\}\Big]S^{(0)}_{W_{(1)}}(\beta_0,t) \\
=& n^{1/2}\Big[\Big\{\tilde{S}^{(1)}_{W_{(1)}}(\beta_0,t) - S^{(1)}_{W_{(1)}}(\beta_0,t)\Big\} - E_{W_{(1)}}(\beta_0,t)\Big\{\tilde{S}^{(0)}_{W_{(1)}}(\beta_0,t) - S^{(0)}_{W_{(1)}}(\beta_0,t)\Big\}\Big] \\
&\times S^{(0)}_{W_{(1)}}(\beta_0,t)/\tilde{S}^{(0)}_{W_{(1)}}(\beta_0,t) \\
=& D_n(t) + D_n(t)\{S^{(0)}_{W_{(1)}}(\beta_0,t)/\tilde{S}^{(0)}_{W_{(1)}}(\beta_0,t) - 1\} \\
&- n^{1/2}\{E_{W_{(1)}}(\beta_0,u) - e_{W_{(1)}}(\beta_0,u)\}\{\tilde{S}^{(0)}_{W_{(1)}}(\beta_0,u) - S^{(0)}_{W_{(1)}}(\beta_0,u)\}S^{(0)}_{W_{(1)}}(\beta_0,u)/\tilde{S}^{(0)}_{W_{(1)}}(\beta_0,u).
\end{aligned}
$$

Integrand of the fourth term in (3.24) can be shown to converge to in probability to zero, uniformly in t as its integrand converges to zero uniformly in t, in view of the stability conditions D and G(iv), combined with the Slutsky's theorem. The fifth term in (3.24) is a local square integrable martingale with variance process

$$
\int_0^1 \Big[\tilde{E}_{W_{(1)}}(\beta_0,u) - E_{W_{(1)}}(\beta_0,u)\Big]^{\otimes 2} S^{(0)}_{W_{(2)}}(\beta_0,u)\lambda_0(u)du
$$

which converges in probability to zero by conditions C, D, and G(iv). Therefore, if we can show that the first term in (3.24) and $D_n(u)$ converge jointly in distribution to independent Gaussian random variables then it implies that $D_n(u)$ converges in distribution to a Gaussian. This further implies that the third term in (3.24) converges in probability to zero and that the first two terms in (3.24) converge jointly in distribution to independent Gaussian random variables, which is the desired property. We start showing the joint in distribution convergence of the first term in (3.24) and $D_n(u)$ through the following Proposition taken from Self and Prentice [1988].

Proposition 3.4.1. *(Self and Prentice [1988]) Let* $\mathbf{X}_n = (X_{1n}, ..., X_{nn})$ *and* $\boldsymbol{\delta}_n = (\delta_{1n}, ..., \delta_{nn})$ *be independent random variables such that:*

(I) δ_n *is a vector of* $\tilde{n}$ *ones and* $n - \tilde{n}$ *zeros, each possible configuration of zeros and ones is equally likely and* $\tilde{n}/n \to_p \alpha \in (0,1)$.

(II) For some scalar functions of $\mathbf{X}_n$, $f_{in}(\mathbf{X}_n)$, *and for any* $\epsilon > 0$,

$$n^{-1}\sum_{i=1}^{n}[f_{in}(\mathbf{X}_n) - f_{\cdot n}(\mathbf{X}_n)]^2 I\{|f_{in}(\mathbf{X}_n) - f_{\cdot n}(\mathbf{X}_n)| > n^{1/2}\epsilon\} \to_p 0,$$

and $\mathbf{S}^2_{f_n} \to_p \sigma_f^2 > 0$, *where* $f_{\cdot n}(\mathbf{X}_n) = n^{-1}\sum_{i=1}^{n} f_{in}(\mathbf{X}_n)$ *and*

$$\mathbf{S}^2_{f_n} = n^{-1}\sum_{i=1}^{n}[f_{in}(\mathbf{X}_n) - f_{\cdot n}(\mathbf{X}_n)]^2.$$

(III) The scalar functions of $\mathbf{X}_n$, $g_n(\mathbf{X}_n)$, *converge in distribution to a Gaussian random variable with mean zero and variance* σ_g^2.

Let $h_n(\mathbf{X}_n, \boldsymbol{\delta}_n) = n^{1/2}[\tilde{n}^{-1}\sum_{i=1}^{n}\delta_{in} f_{in}(\mathbf{X}_n) - f_{\cdot n}(\mathbf{X}_n)]$, *then* $\{g_n(\mathbf{X}_n), h_n(\mathbf{X}_n, \boldsymbol{\delta}_n)\}$ *converge in distribution to a bivariate Gaussian random variable with mean zero and covariance matrix given by*

$$\begin{bmatrix} \sigma_g^2 & \mathbf{0} \\ \mathbf{0} & (1-\alpha)\alpha^{-1}\sigma_f^2 \end{bmatrix}.$$

We can show asymptotic normality of the MSCM case-cohort score statistics via Proposition 3.4.1 as shown below.

Theorem 3.4.2. *(Asymptotic normality of the case-cohort MSCM score statistic) Under conditions A-G,*

$$n^{-1/2}\tilde{U}(\beta_0, 1) \to_d N(\mathbf{0}, \Sigma_{\tilde{U}})$$

where $\Sigma_{\tilde{U}} = \Sigma_U + \Delta_\alpha$,

$$\Delta_\alpha = \int_0^1 \int_0^1 G(\beta_0, x, v)\lambda_0(x)\lambda_0(v)dxdv, \tag{3.25}$$

and $G(\beta_0, x, v)$ *is given in the proof below.*

Proof. We briefly describe some necessary steps to prove Theorem 3.4.2. Details of the proof and the calculation of the limiting covariance function are provided in the supplemental material (§ 3.6).

Our goal can be achieved by showing the first and the second term in (3.24) converge jointly to independent Gaussian random variables, so that we can claim that the limiting covariance function of the case-cohort MSCM score process is given by the sum of each of the limiting covariances. We already have in distribution convergence of the first term,

which is the full cohort MSCM score function from Theorem 4, but not that of the second term in (3.24). To show in distribution convergence of the second term to a Gaussian, we first show in distribution convergence of $D_n(\cdot)$.

Consider application of Proposition 1 to $D_n(t)$. In particular, X_{in} represents $\{W_i(u), Y_i(u), N_i(u), A_i(u); u \in [0,1]\}$, $f_{in}(\mathbf{X}_n)$ represents a linear combination of elements of $W_i(t)Y_i(t)r\{\beta_0' A_i(t)\}$ and $W_i(t)Y_i(t)r^{(1)}\{\beta_0' A_i(t)\}$. Specifically, $f_{in}(\mathbf{X}_n)$ equals to

$$\sum_{j=1}^{p} d_j [W_i(t_j)Y_i(t_j)r_j^{(1)}\{\beta_0' A_i(t_j)\} - e_{W_{(1)},j}(\beta_0, t_j)W_i(t_j)Y_i(t_j)r\{\beta_0' A_i(t_j)\}],$$

for any constants d_j, where $j = 1, ..., p$. Note that time index can vary by component index $j = 1, ..., p$. Condition (I) in Proposition 1 is satisfied by condition G(i) and the fact that the subcohort is selected by the simple random sampling without replacement. The first subcondition of condition (II) of Proposition 1 follows from the inequality used by Andersen and Gill [1982] and Self and Prentice [1988],

$$|a-b|^2 I\{|a-b| > \epsilon\} \leq 4|a|^2 I\{|a| > \epsilon/2\} + 4|b|^2 I\{|b| > \epsilon/2\}, \tag{3.26}$$

by letting $n^{-1/2} f_{in}(\mathbf{X}_n)$ be a and $n^{-1/2} f_{\cdot n}(\mathbf{X}_n)$ be b, combined with conditions D and G(ii). The second subcondition also follows from bounded property of limiting quantities implied by D, and stability and regularity property of subcohort covariance function implied by G(iii). Finally, $g_n(\mathbf{X}_n)$ represents linear combinations of elements of the full cohort MSCM score process all evaluated at a finite number of fixed time points in $[0,1]$. It can easily be seen that, for any such $g_n(\mathbf{X}_n)$, condition (III) is satisfied due to the convergence of the full cohort MSCM score process to a Gaussian process with mean zero and finite covariance function. It follows that $\{g_n(\mathbf{X}_n), h_n(\mathbf{X}_n, \delta_n)\}$ converges jointly in distribution to independent Gaussian processes equipped with aforementioned $f_{in}(\mathbf{X}_n)$ and $g_n(\mathbf{X}_n)$. Then we have joint convergence of the finite dimensional distributions of the MSCM full cohort score process and a linear combination of elements of $D_n(\cdot)$ to Gaussian distributions. It follows that $D_n(\cdot)$ converges in distribution to a multidimensional mean zero Gaussian random variable by the Cramer-Wold device. As in Self and Prentice [1988], the fact that linear functionals of the Gaussian processes are Gaussian, combined with the fact that $\lambda_0(\cdot)$

is absolutely continuous with respect to the Lebesque measure, leads to the conclusion that the second term in (3.24) converges to a Gaussian random variable. Note that the tightness condition G(ii) implies weak convergence of the process $D_n(\cdot)$ Self and Prentice [1988]. The limiting covariance function of $D_n(\cdot)$, say $G(\beta_0, x, v)$, can be shown by straightforward algebra to equal

$$G(\beta_0,x,v) = (1-\alpha)\alpha^{-1}\Big[h^{(1)}(\beta_0,x,v) - e_{W_{(1)}}(\beta_0,x)h^{(2)}(\beta_0,x,v)' \tag{3.27}$$
$$- h^{(2)}(\beta_0,v,x)e_{W_{(1)}}(\beta_0,v)' + e_{W_{(1)}}(\beta_0,x)e_{W_{(1)}}(\beta_0,v)'h^{(0)}(\beta_0,x,v)\Big]$$

under conditions D, F, G(i), G(iii), and G(iv), where $h^{(j)}(\beta,x,v)$ are given by

$$h^{(0)}(\beta,x,v) = q^{(0)}(\beta,x,v) - s^{(0)}_{W_{(1)}}(\beta,x)s^{(0)}_{W_{(1)}}(\beta,v)$$

$$h^{(1)}(\beta,x,v) = q^{(1)}(\beta,x,v) - s^{(1)}_{W_{(1)}}(\beta,x)s^{(1)}_{W_{(1)}}(\beta,v)'$$

$$h^{(2)}(\beta,x,v) = q^{(2)}(\beta,x,v) - s^{(0)}_{W_{(1)}}(\beta,x)s^{(1)}_{W_{(1)}}(\beta,v).$$

Then it can be seen that the covariance function of the limiting process for the second term in (3.24) conditional on $\mathcal{F}(1)$ is given by (3.25) Finally, it follows that the sum of first two terms in expression (3.24) converge in distribution to a Gaussian random variable with mean zero and covariance given by $\Sigma_{\tilde{U}} \equiv \Sigma_U + \Delta_\alpha$ due to independence. □

Note $S^{(j)}_{W_{(k)}}$ in condition D equals $S^{(j)}$ for all $j = 0, 1, 2$ and $k = 1, 2$ when the IPWs are equal to 1 (i.e., no weights are considered). Specifically, $\Sigma_{W_{(k)}}$ equals Σ and (3.25) equals Δ in Self and Prentice [1988], and hence, $\Sigma_{\tilde{U}} \equiv \Sigma + \Delta$ in the absence of IPWs.

Theorem 3.4.3. *(Asymptotic normality of $\tilde{\beta}$) Under conditions A-G,*

$$n^{1/2}(\tilde{\beta} - \beta_0) \to_d N(\mathbf{0}, \Sigma^{-1}_{W_{(1)}}\Sigma_{\tilde{U}}\Sigma^{-1}_{W_{(1)}})$$

where $\Sigma_{\tilde{U}}$ is given in the Theorem 3.4.2.

Proof. A Taylor expansion of the MSCM case-cohort score process around β_0 evaluated at $\tilde{\beta}$ and $t = 1$ gives

$$n^{-1/2}\tilde{U}(\beta_0,1) = \Big\{ - n^{-1}\frac{\partial^2 \tilde{l}(\dot{\beta},1)}{\partial \beta^2}\Big\} n^{1/2}(\tilde{\beta}-\beta_0) \tag{3.28}$$

for any $\dot{\beta}$ on the line segment between $\tilde{\beta}$ and β_0. It is clear that we need to show (in probability) convergence of $-n^{-1}\partial^2\tilde{l}(\dot{\beta},1)/\partial\beta^2$, for any $\dot{\beta}$ in between $\tilde{\beta}$ and β_0. First, let

$$n^{-1}\tilde{\mathcal{I}}(\beta,t) = -n^{-1}\partial^2\tilde{l}(\beta,t)/\partial\beta^2, \quad \text{and} \tag{3.29}$$

$$n^{-1}\mathcal{I}(\beta,t) = -n^{-1}\partial^2 l(\beta,t)/\partial\beta^2. \tag{3.30}$$

Here, we consider asymptotic properties of (3.30) instead of (3.29) because the two processes converge in probability to the same quantity. To see this, note

$$\begin{aligned} &\sup_{\beta,t}|n^{-1}\{\mathcal{I}(\beta,t)-\tilde{\mathcal{I}}(\beta,t)\}| \\ \leq & n^{-1}\sum_{i=1}^{n}\int_0^1 \sup_{\beta,u}|W_i(u)\{\tilde{V}_{W_{(1)}}(\beta,u)-V_{W_{(1)}}(\beta,u)\}|dN_i(u) \\ \leq & \mathsf{M}_1\int_0^1 \sup_{\beta,u}|\{\tilde{V}_{W_{(1)}}(\beta,u)-V_{W_{(1)}}(\beta,u)\}|n^{-1}\sum_{i=1}^{n}dN_i(u) \to_p 0 \end{aligned} \tag{3.31}$$

for any $(\beta,t) \in \mathcal{B}\times[0,1]$ due to conditions B, D, F, G(iv), by the continuous mapping theorem, and the fact that the total number of jumps are bounded by n. Here, $\tilde{V}_{W_{(1)}} = \tilde{S}^{(2)}_{W_{(1)}}/\tilde{S}^{(0)}_{W_{(1)}} - (\tilde{S}^{(1)}_{W_{(1)}}/\tilde{S}^{(0)}_{W_{(1)}})^{\otimes 2}$. Therefore, it is sufficient to show that $n^{-1}\mathcal{I}(\beta,1)$ converges in probability to a fixed matrix. Using (3.10), decompose $n^{-1}\mathcal{I}(\beta_0,1)$ by

$$\begin{aligned} & n^{-1}\sum_{i=1}^{n}\int_0^1 W_i(u)\Big[\frac{S^{(2)}_{W_{(1)}}(\beta_0,u)S^{(0)}_{W_{(1)}}(\beta_0,u)-\{S^{(1)}_{W_{(1)}}(\beta_0,u)\}^{\otimes 2}}{S^{(0)}_{W_{(1)}}(\beta_0,u)^2}\Big]dM_i(u) \\ & + \int_0^1\Big[\frac{S^{(2)}_{W_{(1)}}(\beta_0,u)S^{(0)}_{W_{(1)}}(\beta_0,u)-\{S^{(1)}_{W_{(1)}}(\beta_0,u)\}^{\otimes 2}}{S^{(0)}_{W_{(1)}}(\beta_0,u)^2}\Big]S^{(0)}_{W_{(1)}}(\beta_0,u)\lambda_0(u)du. \end{aligned}$$

The elements of the first term are local square integrable martingale with variance process for the (i,j) element equals

$$n^{-1}\int_0^1 \big\{V_{W_{(1)}}(\beta,u)\big\}^2_{ij} S^{(0)}_{W_{(2)}}(\beta_0,u)\lambda_0(u)du$$

which converges in probability to zero by virtue of the stability, regularity, and boundedness conditions A-F. It follows that

$$n^{-1}\mathcal{I}(\beta,1) \to_p \int_0^1 v_{W_{(1)}}(\beta_0,u)s^{(0)}_{W_{(1)}}(\beta_0,u)\lambda_0(u)du = \Sigma_{W_{(1)}} \tag{3.32}$$

for any $\beta \in \mathcal{B}$, and therefore $n^{-1}\mathcal{I}(\dot{\beta},1) \to_p \Sigma_{W_{(1)}}$ for any $\dot{\beta}$ in between $\tilde{\beta}$ and β_0. Then Theorem 3.4.2 along with (3.28) complete the proof. In particular, the covariance matrix $\Sigma^{-1}_{W_{(1)}}\Sigma_{\tilde{U}}\Sigma^{-1}_{W_{(1)}}$ has a form

$$\Sigma^{-1}_{W_{(1)}}(\Sigma_U + \Delta_\alpha)\Sigma^{-1}_{W_{(1)}} = \Sigma^{-1}_{W_{(1)}}(\Sigma_{W_{(2)}} + \Delta_{W_{(1)},W_{(2)}} + \Delta_\alpha)\Sigma^{-1}_{W_{(1)}}$$

where $\Sigma_U = \Sigma_{W_{(2)}} + \Delta_{W_{(1)},W_{(2)}}$ as in Theorem 3.4.1 and the explicit form of Δ_α is given by (3.25). □

Note (3.32) converges to Σ when $W_i(t) = 1$ for all i and t. Then $n^{1/2}(\tilde{\beta} - \beta_0)$ converges to mean zero Gaussian vector with the same variance matrix as in Self and Prentice [1988].

Based on Theorem 3.4.3, we propose a new variance estimator

$$\widehat{\text{var}}(\tilde{\beta}) = n^{-1}\hat{\Sigma}^{-1}_{W_{(1)}}(\hat{\Sigma}_{W_{(2)}} + \hat{\Delta}_{W_{(1)},W_{(2)}} + \hat{\Delta}_\alpha)\hat{\Sigma}^{-1}_{W_{(1)}}, \tag{3.33}$$

where

$$\hat{\Sigma}_{W_{(1)}} = n^{-1}\tilde{\mathcal{I}}(\tilde{\beta},1;W=\hat{W}), \tag{3.34}$$

$$\hat{\Sigma}_{W_{(2)}} = n^{-1}\tilde{\mathcal{I}}(\tilde{\beta},1;\hat{W}^2), \tag{3.35}$$

$$\hat{\Delta}_{W_{(1)},W_{(2)}} = n^{-1}\sum_{i=1}^n \int_0^1 \hat{W}_i(u)^2\Big[\tilde{E}_{\{W_{(2)}=\hat{W}^2\}}(\tilde{\beta},u) - \tilde{E}_{\{W_{(1)}=\hat{W}\}}(\tilde{\beta},u)\Big]^{\otimes 2} dN_i(u), \quad \text{and} \tag{3.36}$$

$$\hat{\Delta}_\alpha = n^{-2}\int_0^1\int_0^1 \hat{G}(\tilde{\beta},x,v)\tilde{S}^{(0)}_{W(1)}(\tilde{\beta},x)^{-1} \times \tilde{S}^{(0)}_{W(1)}(\tilde{\beta},v)^{-1}d\overline{N}_{\hat{W}}(x)d\overline{N}_{\hat{W}}(v), \tag{3.37}$$

where $\hat{W}$ or $\hat{W}^2$ means that the IPWs are replaced by $\hat{W}$ or squared values of $\hat{W}$, $\tilde{E}_{\{W_{(2)}=\hat{W}^2\}}$ and $\tilde{E}_{\{W_{(1)}=\hat{W}\}}$ denote that the IPWs in $\tilde{E}_{W_{(2)}}$ and $\tilde{E}_{W_{(1)}}$ are replaced by $\hat{W}$, $\overline{N}_{\hat{W}}(t)$ is

defined by $\sum_i \hat{W}_i(t)N_i(t)$, and $\hat{G}(\tilde{\beta},\cdot,\cdot)$ is (3.27) with $h^{(j)}(\beta_0,\cdot)$, $e_{W_{(1)}}(\beta_0,\cdot)$, and $s^{(0)}_{W_{(1)}}(\beta_0,\cdot)$ replaced by $\tilde{H}^{(j)}(\tilde{\beta},\cdot), \tilde{E}_{W_{(1)}}(\tilde{\beta},\cdot)$, and $\tilde{S}^{(0)}_{W_{(1)}}(\tilde{\beta},\cdot)$. Estimators (3.34), (3.35), and (3.36) are consistent estimators of $\Sigma_{W_{(1)}}$, $\Sigma_{W_{(2)}}$, and $\Delta_{W_{(1)},W_{(2)}}$ in view of (3.31) and (3.32) along with condition A. Estimator (3.37) is a consistent estimator of Δ_α in view of conditions A, G(ii), that $n^{-1}\overline{N}_{\hat{W}}(t)$ uniformly converges to $\int_0^t s^{(0)}_{W_{(1)}}(\beta_0,u)\lambda_0(u)du$, and that $n^{-1}\overline{N}_{\hat{W}}(1)$ is bounded in probability.

The proposed variance estimator (3.33) is different from the robust estimator proposed by Lin and Ying (LY,Lin and Ying [1993]) that is used in most MSM analyses. Both (3.33) and the LY estimator are sandwich-type estimators where the "bread" of sandwich ($\hat{\Sigma}^{-1}_{W_{(1)}}$) is the same. The difference comes from the "meat". The proposed variance estimator requires calculation of $\hat{\Delta}_\alpha$ which reflects covariance among score components induced by the subcohort sampling. The covariance matrix of the MSCM case-cohort score statistic $\Sigma_{\tilde{U}} + \Delta_\alpha$ is estimated without explicit estimation of Δ_α if the LY estimator is used. It can be seen that calculation of the LY estimator is based on (weighted) score residuals. When sample size is small, the score residuals will be correlated due to the substitution of $\tilde{\beta}$ or $\beta^\star$ for β, which might lead to underestimation of the true variance. Simulation results reported in §3.5.2 below indicate that (3.33) may be more accurate when the size of subcohort is small.

3.5 Implementation and Simulation

We have shown that we can obtain a consistent and asymptotically normally distributed estimator of treatment effect in the case-cohort setting by fitting a MSCM via inverse probability weighting. This provides theoretical justification for simulation results shown in Cole et al. [2012]. In this section we (i) describe how a MSCM can easily be fit via inverse probability weighting for either the full cohort or case-cohort setting using standard survival analysis software, such as R or SAS, and (ii) present results from a simulation study.

3.5.1 Implementation

To fit a MSCM via inverse probability weighting for a full cohort, first create a data set in which each person-visit corresponds to one row. Specifically, let each row contain a subject

identifier, visit (or date) information, treatment and time-varying confounder information at the corresponding visit/date time, and baseline covariates. Depending upon the user-defined models to estimate $W_i(t)$, the data set may be augmented by treatment/covariate histories in each row as well. For example, one might fit pooled logistic models to obtain the estimated probability of receiving treatment at time t by regressing the log-odds of receiving treatment $A(t)$ on prior treatment status (say, $A(t^-)$) alone (for the numerator in (4.2)), or with current covariate information $L(t)$ (for the denominator in (4.2)) Hernán, Brumback and Robins [2001]. Analogously, the estimated probability of being uncensored at time t can be obtained by regressing the log-odds of being uncensored ($C(t) = 0$) on current treatment status ($A(t)$) alone, or with $L(t)$. For such models flexible functional forms (e.g., splines) are often used for continuous confounders Cole and Hernán [2008]; Cole et al. [2012, 2003]. Predicted values of the denominator and numerator probabilities in (4.2) and (3.6) can then be used to calculate $\hat{W}_i(t)$ for all participants $i = 1, ..., n$ and all study visit times t. Then $\hat{W}_i(t)$ needs to be added to the data set to fit the MSCM. Finally, the data set should be prepared in the counting process type format whereby each row contains the start and stop times corresponding to the previous and current visits, along with an event status indicator for the current visit. Then standard software can be used to fit the MSCM via inverse probability weighting. For instance, using the `survival` package in R [Therneau, 2012], the following code can be used:

```
coxph(Surv(start, stop, delta) ~ trt, weight=w)
```

where `delta` is the event indicator having value 1 if an event occurred at `stop` and 0 otherwise, `trt` indicates whether an individual received treatment (assuming treatment is a scalar) over the interval `(start, stop]`, and `w` is $\hat{W}_i(t)$. The same model can be fit in SAS by using the following code:

```
proc phreg data = dataname covout;
     model (start,stop)*delta(0)=trt;
     weight w;
run;
```

Fitting a MSCM in the case-cohort setting can be accomplished with some additional data modifications. First, prepare a reduced data set including the randomly selected $\tilde{n}$

subcohort members and all cases. Second, estimate the individual-time-specific weights $W_i(t)$ based on the user-specified model as before (e.g., logistic regression), except with individuals in the subcohort that are not cases weighted by $n/\tilde{n}$ [Cole et al., 2012]. After adding the estimated individual-time-specific weights $\hat{W}_i(t)$ to each person-visit row, modify each nonsubcohort case to contribute only one line of data with start time $t_j - \epsilon$ and stop time t_j where t_j is the event time for that individual and ϵ is chosen to be very small, for instance $\epsilon = 0.0001$. This insures that nonsubcohort cases appear only in the risk set when they fail. One should make sure that the start times for nonsubcohort cases are positive, such that $t_j - \epsilon > 0$ for your choice of ϵ. This modification of the data set for the nonsubcohort cases is sufficient to obtain β^*, and the same R/SAS code as above can be employed using the modified data set. Obtaining $\tilde{\beta}$ can be accomplished with an additional data step wherein a dummy variable is coded equal to a relatively small negative value (e.g., -20) for nonsubcohort cases and 0 otherwise [Therneau and Li, 1999]. Then, $\tilde{\beta}$ can be obtained as follows in R:

```
coxph(Surv(start, stop, delta) ~ trt + offset(dummy), weight=w)
```

or in SAS:

```
proc phreg data = dataname covout;
     model (start,stop)*delta(0)=trt/offset=dummy;
     weight w;
run;
```

The `offset` term enforces a relative weight of $\exp(-20) < 10^{-8}$, assuming -20 is used for the dummy value, to the nonsubcohort cases so that they effectively do not contribute to the sum of the log (inside the integral) in (3.8). Therneau and Li [1999] suggested using -100 ($\exp(-100) < 10^{-40}$) for the dummy variable value, however, we found that sometimes the `coxph` function in R did not converge when dummy = -100; this convergence problem was observed when the event rate was very low, say 3-4%. Therefore, we recommend several dummy values be considered to ensure robustness of analysis results. The choice of dummy = -20 yielded reasonable analysis results under average event rate $\geq$5% in our simulation study.

The proposed variance estimator (3.33) requires computation of four components: $\hat{\Sigma}_{W_{(1)}}^{-1}$, $\hat{\Sigma}_{W_{(2)}}$, $\hat{\Delta}_{W_{(1)},W_{(2)}}$, and $\hat{\Delta}_{\alpha}$. The naive variance estimator obtained by fitting the Cox model with the `weight` option is the inverse of minus the second derivative of $\tilde{l}(\beta,1)$ evaluated at $\tilde{\beta}$ (i.e., $\tilde{\mathcal{I}}^{-1}(\tilde{\beta},1)$, the inverse of the observed information matrix) which is n^{-1} times $\hat{\Sigma}_{W_{(1)}}^{-1}$. Therefore, $\hat{\Sigma}_{W_{(1)}}^{-1}$ can be obtained by multiplying n times the naive variance estimate. Likewise, $\hat{\Sigma}_{W_{(2)}}$ can be obtained by multiplying n^{-1} times the inverse of the naive variance estimate obtained by fitting the Cox model with the variable `weight` equal to the square of the original weight variable. Unfortunately, it does not seem that $\hat{\Delta}_{W_{(1)},W_{(2)}}$ and $\hat{\Delta}_{\alpha}$ can be obtained as simply as $\hat{\Sigma}_{W_{(1)}}$ or $\hat{\Sigma}_{W_{(2)}}$. One can create vectors/matrices of $\tilde{S}_{\hat{W}_k}^{(j)}(\tilde{\beta},\cdot)$, and then calculate $\tilde{E}_{\hat{W}_k}(\tilde{\beta},\cdot)$, $\tilde{Q}^{(j)}(\tilde{\beta},\cdot)$, and $\tilde{H}^{(j)}(\tilde{\beta},\cdot)$ to obtain $\hat{\Delta}_{W_{(1)},W_{(2)}}$ and $\hat{\Delta}_{\alpha}$. Alternatively, one may want to apply the LY estimator in practice [Cole et al., 2012]. The LY estimator appears to perform well empirically if we have moderate subcohort size and event rate (Cole et al. [2012], §3.5.2 below), and is computationally straightforward to implement. The LY estimator associated with $\tilde{\beta}$ can be obtained by using the following R or SAS code:

```
coxph(Surv(start, stop, delta) ~ trt + offset(dummy)
                                + cluster(id), weight=w)
proc phreg data = dataname covs(aggregate) covout;
     id id;
     model (start,stop)*delta(0)=trt/offset=dummy;
     weight w;
run;
```

The LY estimator corresponding to β^* can be obtained by deleting `offset(dummy)` or `/offset=dummy`.

3.5.2 Simulation

A simulation study was conducted to examine the finite sample bias of $\tilde{\beta}$ and β^*, and performance of the proposed variance estimator (3.33) as well as the LY variance estimator. Simulations were conducted similar to Cole et al. [2012]. Briefly, potential survival times were generated according to the MSCM (3.4), and observed survival times were generated by stochastically generating time varying exposures and confounders for cohorts of size

Table 3.1: Summary of simulation study

Sub-cohort(%)	Event rate(%)	Estimator	Bias	ESE	ASE		Coverage	
					proposed	LY	proposed	LY
5	5 †	β^*	-0.12	0.49	0.55	0.42	0.97	0.91
		$\tilde{\beta}$	-0.20	0.68	0.66	0.47	0.96	0.90
	25	β^*	-0.03	0.37	0.36	0.31	0.94	0.91
		$\tilde{\beta}$	-0.04	0.44	0.37	0.35	0.92	0.91
10	5	β^*	-0.05	0.40	0.42	0.37	0.97	0.94
		$\tilde{\beta}$	-0.06	0.44	0.43	0.39	0.96	0.94
	25	β^*	-0.02	0.27	0.26	0.26	0.93	0.93
		$\tilde{\beta}$	-0.02	0.29	0.26	0.26	0.93	0.93
20	5	β^*	-0.02	0.35	0.36	0.34	0.96	0.95
		$\tilde{\beta}$	-0.02	0.36	0.36	0.35	0.96	0.95
	25	β^*	-0.01	0.21	0.20	0.20	0.94	0.94
		$\tilde{\beta}$	-0.01	0.21	0.20	0.20	0.94	0.94

Bias denotes the empirical bias of the different estimators of β_0. ASE denotes the average estimated standard errors. ESE denotes the empirical standard errors. Coverage denotes the empirical coverage of 95% Wald-type confidence intervals using either (3.33) or the LY variance estimator. † Of the 5000 estimates of β^* and $\tilde{\beta}$, one was excluded because some of the unstabilized IPWs were greater than 10^6.

$n = 1,000$ (see Cole et al. [2012] for details). While Cole et al. [2012] considered only one scenario having a 25% event rate (i.e., 25% of individuals were cases) and a 20% subcohort fraction (i.e., $\tilde{n}n^{-1} \times 100$), we considered 36 scenarios by varying both the subcohort fraction and the event rate from 5 to 30% (in increments of 5%). Censoring times were generated from uniform distributions with support chosen to achieve the desired event rate. We did not incorporate IPCWs when calculating IPWs because the censoring times were generated independent of the exposure and potential survival times. Following Cole et al. [2012], unstabilized weights were used to calculate IPWs. For each scenario 5,000 data sets were generated under the null $\beta_0 = 0$ and the alternative $\beta_0 = \log(1/2)$.

Results from the simulation study are summarized in Table 1. Only results obtained from six scenarios under the null are presented; results from other scenarios and under the alternative were similar. For all scenarios, under both the null and alternative, $\tilde{\beta}$ and β^* were nearly unbiased; that the two estimators performed similarly is not surprising in light of Theorem 3. Under the null, the proposed variance estimator was always less biased than the LY variance estimator when the subcohort fraction was only 5%, regardless of the event rate. Similarly, (3.33) was less biased regardless of the subcohort fraction when the event rate

was 5%. Both the proposed and the LY variance estimators were approximately unbiased when the subcohort fraction and event rate were both greater than 15%. Wald confidence intervals (CIs) using the LY variance estimator tended to undercover when the subcohort fraction was 5%, whereas Wald CIs using (3.33) exhibited coverage close to the nominal level for all scenarios considered. In summary, both $\tilde{\beta}$ and β^*, along with the proposed variance estimator and CIs, exhibited good finite sample properties for the scenarios considered, while performance of the LY variance estimator depended on subcohort size and event rate.

3.6 Supplemental Material

This supplementary material contains three parts: § 3.6 provides detailed steps to apply Proposition 1 to show asymptotic normality of the MSCM case-cohort score process (Theorem 3.4.2). § 3.6 justifies the application of Proposition 1 by showing that $f_{in}(\mathbf{X}_n)$ satisfies conditions in Proposition 1. § 3.6 shows detailed calculations to obtain limiting covariance function of the MSCM case-cohort score process.

Application of Proposition 1

Our goal is to show that the difference of the first two terms in (3.24), which is given by

$$\begin{aligned} & n^{-1/2}\sum_{i=1}^{n}\int_0^t W_i(u)[A_i(u) - E_{W_{(1)}}(\beta_0, u)]dM_i(u) - \int_0^t D_n(u)\lambda_0(u)du \\ =& B_n(t) - \int_0^t D_n(u)\lambda_0(u)du \\ =& B_n(t) - C_n(t) \end{aligned}$$

converges in distribution to a finite dimensional Gaussian random variable where $B_n(\cdot)$, $C_n(\cdot)$, and $D_n(\cdot)$ are defined by

$$B_n(t) = n^{-1/2}\sum_{i=1}^{n}\int_0^t W_i(u)[A_i(u) - E_{W_{(1)}}(\beta_0, u)]dM_i(u), \tag{3.38}$$

$$C_n(t) = \int_0^t D_n(u)\lambda_0(u)du, \quad \text{and} \tag{3.39}$$

$$D_n(u) = n^{1/2}\Big[\{\tilde{S}^{(1)}_{W_{(1)}}(\beta_0,u) - S^{(1)}_{W_{(1)}}(\beta_0,u)\} - e_{W_{(1)}}(\beta_0,u) \times \{\tilde{S}^{(0)}_{W_{(1)}}(\beta_0,u) - S^{(0)}_{W_{(1)}}(\beta_0,u)\}\Big]S^{(0)}_{W_{(1)}}(\beta_0,u). \tag{3.40}$$

Let $g_n(\mathbf{X}_n)$ be a linear combination of elements of the MSCM full cohort score process (B_n), i.e., for any constants c_j $(j = 1,..,p)$,

$$g_n(\mathbf{X}_n) = n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{p} c_j \int_0^t W_i(u)[A_{i,j}(u) - E_{W_{(1)},j}(\beta_0,u)]dM_i(u)$$

where the subscript j denotes the jth component of a vector. Also, let $h_n(\mathbf{X}_n,\delta_n)$ be a linear combination of elements of D_n, i.e., for any constants $d_j(j = 1,..,p)$, $f_{in}(\mathbf{X}_n)$ is given by

$$f_{in}(\mathbf{X}_n) = \sum_{j=1}^{p} d_j\Big[W_i(u_j)Y_i(u_j)r^{(1)}_j\{\beta_0' A_i(u_j)\} - e_{W(1),j}(\beta_0,u_j)W_i(u_j)Y_i(u_j)r\{\beta_0' A_i(u_j)\}\Big]. \tag{3.41}$$

Then (3.41) leads to the desired form of $h_n(\mathbf{X}_n,\delta_n)$:

$$\begin{aligned}
h_n(\mathbf{X}_n,\delta_n) =& n^{1/2}[\tilde{n}^{-1}\sum_{i=1}^{n}\delta_{in}f_{in}(\mathbf{X}_n) - f_{\cdot n}(\mathbf{X}_n)] \\
=& n^{1/2}\Big[\sum_{j=1}^{p} d_j\{\tilde{S}^{(1)}_{W_{(1)},j}(\beta_0,u_j) - e_{W_{(1)},j}(\beta_0,u_j)\tilde{S}^{(0)}_{W_{(1)}}(\beta_0,u_j)\} \\
& - \sum_{j=1}^{p} d_j\{S^{(1)}_{W_{(1)},j}(\beta_0,u_j) - e_{W_{(1)},j}(\beta_0,u_j)S^{(0)}_{W_{(1)}}(\beta_0,u_j)\}\Big] \\
=& n^{1/2}\sum_{j=1}^{p} d_j\Big[\{\tilde{S}^{(1)}_{W_{(1)},j}(\beta_0,u_j) - S^{(1)}_{W_{(1)},j}(\beta_0,u_j)\} \\
& - e_{W_{(1)},j}(\beta_0,u_j)\{\tilde{S}^{(0)}_{W_{(1)}}(\beta_0,u_j) - S^{(0)}_{W_{(1)}}(\beta_0,u_j)\}\Big]
\end{aligned}$$

which is a linear combination of elements of D_n where each jth component can be evaluated at possibly different time points u_j, i.e.,

$$h_n(\mathbf{X}_n,\delta_n) = \sum_{j=1}^{p} d_j D_{n,j}(u_j)$$

Assume that $f_{in}(\mathbf{X}_n)$ and $g_n(\mathbf{X}_n)$ satisfy conditions stated in Proposition 1, which will

be shown in the Part II later in the supplementary material. Then by varying c_j and d_j, we can show that any chosen elements of B_n and D_n jointly converge in distribution to an independent bivariate Gaussian process by application of Proposition 1. For example, consider $c_1 = d_1 = 1$ and $c_2 = ... = c_p = d_2 = ... = d_p = 0$. Then Proposition 1 states that the first element of B_n and the first element of D_n converge jointly in distribution to an independent bivariate Gaussian. In iterative fashion, we can show that jth element of B_n and kth element of D_n converge in distribution to an independent bivariate Gaussian for all combinations of $(j,k) \in [1,2,...,p] \times [1,2,...,p]$. Therefore, B_n and D_n converge in distribution to independent processes. We have shown that B_n, the MSCM full cohort score process, converges in distribution to a Gaussian process. Therefore, what we have left to show is that D_n converges in distribution to a Gaussian process (and later to show that C_n converges in distribution to a Gaussian process).

In the above arguments we have shown that, for any d_j $(j = 1,...,p)$, $\sum_{j=1}^{p} d_j D_{n,j}$ converges in distribution to a univariate Gaussian because $f_{in}(\mathbf{X}_n)$ satisfies conditions in Proposition 1 for any d_j (which, as we mentioned above, will be shown in the Part II). Therefore, it follows that D_n converges in distribution to a multidimensional mean zero Gaussian random variable by the Cramer-Wold device. As in Self and Prentice [1988], the fact that linear functionals of the Gaussian processes are Gaussian combined with the fact that $\lambda_0(\cdot)$ is absolutely continuous with respect to the Lebesque measure leads to that C_n converges to a Gaussian random variable, say C. Then it follows that $B_n - C_n$ converges to a mean zero Gaussian random variable with covariance $\Sigma_U + \Delta_\alpha$, as the limiting covariance of C_n will be shown to equal Δ_α later in the Part II.

In the next two parts, we verify that $f_{in}(\mathbf{X}_n)$ and $g_n(\mathbf{X}_n)$ satisfy conditions in Proposition 1, and show the explicit form of limiting covariance structure of C_n respectively.

Condition (II) in Proposition 1

Recall that condition (II) of Proposition 1 has the following two subconditions: For any $\epsilon > 0$,

$$n^{-1}\sum_{i=1}^{n}[f_{in}(\mathbf{X}_n) - f_{\cdot n}(\mathbf{X}_n)]^2 I_{\{|f_{in}(\mathbf{X}_n)-f_{\cdot n}(\mathbf{X}_n)|>n^{1/2}\epsilon\}} \to_p 0, \text{ and} \tag{3.42}$$

$$\mathbf{S}^2_{f_n} = n^{-1}\sum_{i=1}^{n}[f_{in}(\mathbf{X}_n) - f_{\cdot n}(\mathbf{X}_n)]^2 \to_p \sigma_f. \tag{3.43}$$

To show (3.42) based on the inequality (3.26), we need to show that for any $\epsilon > 0$,

$$n^{-1}\sum_{i=1}^{n}|f_{in}(\mathbf{X}_n)|^2 I\{|f_{in}(\mathbf{X}_n)| > n^{1/2}\epsilon/2\} \to_p 0, \quad \text{and} \tag{3.44}$$

$$n^{-1}|f_{\cdot n}(\mathbf{X}_n)|^2 I\{|f_{\cdot n}(\mathbf{X}_n)| > n^{1/2}\epsilon/2\} \to_p 0. \tag{3.45}$$

To show (3.44), recall condition G(ii): For any $\epsilon > 0$

$$\sup_t n^{-1}\sum_{i=1}^{n} W_i(t)^2 Y_i(t) r\{\beta_0' A_i(t)\}^2 \tag{3.46}$$
$$\times I\{n^{-1/2} W_i(t) Y_i(t) r\{\beta_0' A_i(t)\} > \epsilon\} \to_p 0,$$

$$\sup_t n^{-1}\sum_{i=1}^{n} W_i(t)^2 Y_i(t) \|r^{(1)}\{\beta_0' A_i(t)\}\|^2 \tag{3.47}$$
$$\times I\{n^{-1/2} W_i(t) Y_i(t) \|r^{(1)}\{\beta_0' A_i(t)\}\| > \epsilon\} \to_p 0,$$

where (3.46) implies

$$\sup_t n^{-1}\sum_{i=1}^{n} W_i(t)^2 Y_i(t) r\{\beta_0' A_i(t)\}^2 \|e_{W_{(1)}}(\beta_0, t)\|^2 \tag{3.48}$$
$$\times I\{n^{-1/2} W_i(t) Y_i(t) r\{\beta_0' A_i(t)\} \|e_{W_{(1)}}(\beta_0, t)\| > \epsilon\} \to_p 0.$$

It can be shown that (3.47) and (3.48) imply (3.44), by repeatedly applying (3.26).

Also, we can rewrite

$$f_{\cdot n}(\mathbf{X}_n) = \sum_{j=1}^{p} d_j [S^{(1)}_{W_{(1)},j}(\beta_0, t) - e_{W_{(1)},j}(\beta_0, t) S^{(0)}_{W_{(1)}}(\beta_0, t)].$$

Then (3.45) can immediately be seen by the stability property implied by condition D.

To show (3.43), note that $\mathbf{S}_{f_n}^2$ can be rewritten as

$$\begin{aligned}\mathbf{S}_{f_n}^2 &= n^{-1}\sum_{i=1}^{n}[f_{in}(\mathbf{X}_n) - f_{\cdot n}(\mathbf{X}_n)]^2 \\ &= n^{-1}\sum_{i=1}^{n} f_{in}(\mathbf{X}_n)^2 - \{f_{\cdot n}(\mathbf{X}_n)\}^2.\end{aligned} \tag{3.49}$$

For notational and calculational convenience, let

$$f_{in}(\mathbf{X}_n) = \sum_{j=1}^{p} d_j(a_j - b_j)$$

by letting

$$\begin{aligned}a_j &= W_i(u_j)Y_i(u_j)r_j^{(1)}\{\beta_0' A_i(u_j)\}, \quad \text{and} \\ b_j &= e_{W(1),j}(\beta_0, u_j)W_i(u_j)Y_i(u_j)r\{\beta_0' A_i(u_j)\}\end{aligned}$$

then calculate the form of each term in (3.49). First term in (3.49) can be written as follows:

$$\begin{aligned}&n^{-1}\sum_{i=1}^{n} f_{in}(\mathbf{X}_n)^2 = n^{-1}\sum_{i=1}^{n}\Big[\sum_{j=1}^{p} d_j(a_j - b_j)\Big]^2 \\ =&n^{-1}\sum_{i=1}^{n}\Big[\sum_{j=1}^{p} d_j^2(a_j - b_j)^2 + 2\sum_{j<k} d_j d_k(a_j - b_j)(a_k - b_k)\Big] \\ =&n^{-1}\sum_{i=1}^{n}\Big[\sum_{j=1}^{p} d_j^2\Big\{W_i(u_j)^2 Y_i(u_j) r_j^{(1)}\{\beta_0' A_i(u_j)\}^2 \\ &\quad - 2W_i(u_j)Y_i(u_j)r_j^{(1)}\{\beta_0' A_i(u_j)\}e_{W(1),j}(\beta_0, u_j)W_i(u_j)Y_i(u_j)r\{\beta_0' A_i(u_j)\} \\ &\quad + e_{W(1),j}(\beta_0, u_j)^2 W_i(u_j)^2 Y_i(u_j) r\{\beta_0' A_i(u_j)\}^2\Big\} \\ &\quad + 2\sum_{j<k} d_j d_k\Big\{W_i(u_j)Y_i(u_j)r_j^{(1)}\{\beta_0' A_i(u_j)\}W_i(u_k)Y_i(u_k)r_k^{(1)}\{\beta_0' A_i(u_k)\} \\ &\quad - W_i(u_j)Y_i(u_j)r_j^{(1)}\{\beta_0' A_i(u_j)\}e_{W(1),k}(\beta_0, u_k)W_i(u_k)Y_i(u_k)r\{\beta_0' A_i(u_k)\} \\ &\quad - e_{W(1),j}(\beta_0, u_j)W_i(u_j)Y_i(u_j)r\{\beta_0' A_i(u_j)\}W_i(u_k)Y_i(u_k)r_k^{(1)}\{\beta_0' A_i(u_k)\} \\ &\quad + e_{W(1),j}(\beta_0, u_j)W_i(u_j)Y_i(u_j)r\{\beta_0' A_i(u_j)\}e_{W(1),k}(\beta_0, u_k)W_i(u_k)Y_i(u_k) \\ &\quad \times r\{\beta_0' A_i(u_k)\}\Big\}\Big]\end{aligned}$$

Then using $Q^{(j)}(j = 0, 1, 2)$ notation defined in condition G(iii), the above equation can be

abbreviated as

$$
\begin{aligned}
n^{-1}\sum_{i=1}^{n} f_{in}(\mathbf{X}_n)^2 = \sum_{j=1}^{p} d_j^2 \Big\{ & Q^{(1)}_{(j,j)}(\beta_0, u_j, u_j) - 2e_{W(1),j}(\beta_0, u_j) Q^{(2)}_j(\beta_0, u_j, u_j) \\
& + e_{W(1),j}(\beta_0, u_j)^2 Q^{(0)}(\beta_0, u_j, u_j) \Big\} \\
+ 2\sum_{j<k} d_j d_k \Big\{ & Q^{(1)}_{(j,k)}(\beta_0, u_j, u_k) - e_{W(1),k}(\beta_0, u_k) Q^{(2)}_j(\beta_0, u_k, u_j) \\
& - e_{W(1),j}(\beta_0, u_j) Q^{(2)}_k(\beta_0, u_j, u_k) \\
& + e_{W(1),j}(\beta_0, u_j) e_{W(1),k}(\beta_0, u_k) Q^{(0)}(\beta_0, u_j, u_k) \Big\}.
\end{aligned}
$$

Now it can be seen that the above equation converges in probability to a fixed quantity in view of stability properties of $Q^{(\cdot)}$ stated in condition G(iii). The convergence of $f_{\cdot n}(\mathbf{X}_n)$ can be shown using the same manner as the above. In particular, let

$$
f_{\cdot n}(\mathbf{X}_n) = \sum_{j=1}^{p} d_j (a_j - b_j)
$$

where

$$
\begin{aligned}
a_j &= S^{(1)}_{W_{(1),j}}(\beta_0, u_j), \quad \text{and} \\
b_j &= e_{W(1),j}(\beta_0, u_j) S^{(0)}_{W_{(1)}}(\beta_0, u_j),
\end{aligned}
$$

then

$$
\begin{aligned}
\{f_{\cdot n}(\mathbf{X}_n)\}^2 = \sum_{j=1}^{p} d_j^2 \Big\{ & S^{(1)}_{W_{(1),j}}(\beta_0, u_j)^2 - 2S^{(1)}_{W_{(1),j}}(\beta_0, u_j) e_{W(1),j}(\beta_0, u_j) S^{(0)}_{W_{(1)}}(\beta_0, u_j) \\
& + e_{W(1),j}(\beta_0, u_j)^2 S^{(0)}_{W_{(1)}}(\beta_0, u_j)^2 \Big\} \\
+ 2\sum_{j<k} d_j d_k \Big\{ & S^{(1)}_{W_{(1),j}}(\beta_0, u_k) S^{(1)}_{W_{(1),k}}(\beta_0, u_k) \\
& - S^{(1)}_{W_{(1),j}}(\beta_0, u_j) e_{W(1),k}(\beta_0, u_k) S^{(0)}_{W_{(1)}}(\beta_0, u_k) \\
& - e_{W(1),j}(\beta_0, u_j) S^{(0)}_{W_{(1)}}(\beta_0, u_j) S^{(0)}_{W_{(1)}}(\beta_0, u_k) \\
& + e_{W(1),j}(\beta_0, u_j) S^{(0)}_{W_{(1)}}(\beta_0, u_j) e_{W(1),k}(\beta_0, u_k) S^{(0)}_{W_{(1)}}(\beta_0, u_k) \Big\}
\end{aligned}
$$

Then without further calculation, it can be seen that the above equation also converges to a fixed quantity by conditions D and G(iii), and therefore we prove that (3.43) holds.

Limiting Covariance function

Now we need to show the limiting covariance function of C_n. First we will show the limiting covariance function of D_n.

Let $h_n(\mathbf{X}_n, \delta_n) = D_{n,j}(u_j) + D_{n,k}(u_k)$ (i.e., let $d_j = d_k = 1$ and $d_l = 0$ for all $l \neq j$ in $\sum_{j=1}^{p} d_j D_{n,j}(u_j)$). Covariance between $D_{n,j}(u_j)$ and $D_{n,k}(u_k)$ is given by

$$\begin{aligned} &\mathrm{Cov}(D_{n,j}(u_j), D_{n,k}(u_k)) \\ &= \Big\{\mathrm{Var}(h_n(\mathbf{X}_n, \delta_n)) - \mathrm{Var}(D_{n,j}(u_j)) - \mathrm{Var}(D_{n,k}(u_k))\Big\}/2. \end{aligned} \tag{3.50}$$

Then the limiting values of (3.50) will lead to the (j,k)th components of the limiting covariance, i.e.,

$$\begin{aligned} &\lim_{n\to\infty} \mathrm{Cov}(D_{n,j}(u_j), D_{n,k}(u_k)) \\ &= \lim_{n\to\infty} \Big\{\mathrm{Var}(h_n(\mathbf{X}_n, \delta_n)) - \mathrm{Var}(D_{n,j}(u_j)) - \mathrm{Var}(D_{n,k}(u_k))\Big\}/2. \end{aligned} \tag{3.51}$$

By Proposition 1, we can obtain limiting values of $\mathrm{Var}(h_n(\mathbf{X}_n, \delta_n))$, $\mathrm{Var}(D_{n,j}(u_j))$ and $\mathrm{Var}(D_{n,k}(u_k))$ using sample covariances calculated based on corresponding $f_{in}(\mathbf{X}_n)$ equipped with condition G(iii) and G(iv). Note that condition G(iii) ensures the convergence of the finite sample covariance function to that of the limiting distribution. For notational convenience, let

$$\begin{aligned} \mathbf{F}_{in,j}(\mathbf{X}_n) = \Big[& W_i(u_j) Y_i(u_j) r_j^{(1)}\{\beta_0' A_i(u_j)\} \\ & - e_{W(1),j}(\beta_0, u_j) W_i(u_j) Y_i(u_j) r\{\beta_0' A_i(u_j)\}\Big], \quad \text{and} \\ \mathbf{F}_{\cdot n,j}(\mathbf{X}_n) = n^{-1} \sum_{i=1}^{n} \mathbf{F}_{in,j}(\mathbf{X}_n); & \qquad j = 1, ..., p. \end{aligned}$$

Now, straightforward calculation based on Proposition 1 yields that

$$
\begin{aligned}
&\lim_{n\to\infty}\Big\{\mathrm{Var}(h_n(\mathbf{X}_n,\delta_n))-\mathrm{Var}(D_{n,j}(u_j))-\mathrm{Var}(D_{n,k}(u_k))\Big\}\\
=&(1-\alpha)\alpha^{-1}\lim_{n\to\infty}n^{-1}\sum_{i=1}^{n}\Big[\mathbf{F}_{in,j}(\mathbf{X}_n)+\mathbf{F}_{in,k}(\mathbf{X}_n)-\{\mathbf{F}_{\cdot n,j}(\mathbf{X}_n)+\mathbf{F}_{\cdot n,k}(\mathbf{X}_n)\}\Big]^2\\
&-(1-\alpha)\alpha^{-1}\lim_{n\to\infty}n^{-1}\sum_{i=1}^{n}\Big[\mathbf{F}_{in,j}(\mathbf{X}_n)-\mathbf{F}_{\cdot n,j}(\mathbf{X}_n)\Big]^2\\
&-(1-\alpha)\alpha^{-1}\lim_{n\to\infty}n^{-1}\sum_{i=1}^{n}\Big[\mathbf{F}_{in,k}(\mathbf{X}_n)-\mathbf{F}_{\cdot n,k}(\mathbf{X}_n)\Big]^2\\
=&(1-\alpha)\alpha^{-1}\Big\{\lim_{n\to\infty}n^{-1}\sum_{i=1}^{n}\Big[\mathbf{F}_{in,j}(\mathbf{X}_n)+\mathbf{F}_{in,k}(\mathbf{X}_n)-\{\mathbf{F}_{\cdot n,j}(\mathbf{X}_n)+\mathbf{F}_{\cdot n,k}(\mathbf{X}_n)\}\Big]^2\\
&\qquad-\lim_{n\to\infty}n^{-1}\sum_{i=1}^{n}\Big[\mathbf{F}_{in,j}(\mathbf{X}_n)-\mathbf{F}_{\cdot n,j}(\mathbf{X}_n)\Big]^2\\
&\qquad-\lim_{n\to\infty}n^{-1}\sum_{i=1}^{n}\Big[\mathbf{F}_{in,k}(\mathbf{X}_n)-\mathbf{F}_{\cdot n,k}(\mathbf{X}_n)\Big]^2\Big\}.
\end{aligned}
$$

The whole term after $(1-\alpha)\alpha^{-1}$ can be simplified as follows:

$$
\begin{aligned}
&\lim_{n\to\infty}n^{-1}\sum_{i=1}^{n}\Big[\{\mathbf{F}_{in,j}(\mathbf{X}_n)^2+2\mathbf{F}_{in,j}(\mathbf{X}_n)\mathbf{F}_{in,k}(\mathbf{X}_n)+\mathbf{F}_{in,k}(\mathbf{X}_n)^2\}\\
&\qquad\qquad-2\{\mathbf{F}_{in,j}(\mathbf{X}_n)+\mathbf{F}_{in,k}(\mathbf{X}_n)\}\{\mathbf{F}_{\cdot n,j}(\mathbf{X}_n)+\mathbf{F}_{\cdot n,k}(\mathbf{X}_n)\}\\
&\qquad\qquad+\{\mathbf{F}_{\cdot n,j}(\mathbf{X}_n)^2+2\mathbf{F}_{\cdot n,j}(\mathbf{X}_n)\mathbf{F}_{\cdot n,k}(\mathbf{X}_n)+\mathbf{F}_{\cdot n,k}(\mathbf{X}_n)^2\}\Big]\\
&-\lim_{n\to\infty}n^{-1}\sum_{i=1}^{n}\Big[\mathbf{F}_{in,j}(\mathbf{X}_n)^2-2\mathbf{F}_{in,j}(\mathbf{X}_n)\mathbf{F}_{\cdot n,j}(\mathbf{X}_n)+\mathbf{F}_{\cdot n,j}(\mathbf{X}_n)^2\Big]\\
&-\lim_{n\to\infty}n^{-1}\sum_{i=1}^{n}\Big[\mathbf{F}_{in,k}(\mathbf{X}_n)^2-2\mathbf{F}_{in,k}(\mathbf{X}_n)\mathbf{F}_{\cdot n,k}(\mathbf{X}_n)+\mathbf{F}_{\cdot n,k}(\mathbf{X}_n)^2\Big]\\
=&\lim_{n\to\infty}n^{-1}\sum_{i=1}^{n}2\Big[\mathbf{F}_{in,j}(\mathbf{X}_n)\mathbf{F}_{in,k}(\mathbf{X}_n)-\mathbf{F}_{in,j}(\mathbf{X}_n)\mathbf{F}_{\cdot n,k}(\mathbf{X}_n)-\mathbf{F}_{in,k}(\mathbf{X}_n)\mathbf{F}_{\cdot n,j}(\mathbf{X}_n)\\
&\qquad\qquad+\mathbf{F}_{\cdot n,j}(\mathbf{X}_n)\mathbf{F}_{\cdot n,k}(\mathbf{X}_n)\Big]\\
=&\lim_{n\to\infty}2\Big[n^{-1}\sum_{i=1}^{n}\mathbf{F}_{in,j}(\mathbf{X}_n)\mathbf{F}_{in,k}(\mathbf{X}_n)-\mathbf{F}_{\cdot n,j}(\mathbf{X}_n)\mathbf{F}_{\cdot n,k}(\mathbf{X}_n)\Big],
\end{aligned}
$$

where the first term inside the bracket is given by

$$
\begin{aligned}
& n^{-1}\sum_{i=1}^{n}\mathbf{F}_{in,j}(\mathbf{X}_n)\mathbf{F}_{in,k}(\mathbf{X}_n) \\
=& n^{-1}\sum_{i=1}^{n}\Big[W_i(u_j)Y_i(u_j)r_j^{(1)}\{\beta_0' A_i(u_j)\}W_i(u_k)Y_i(u_k)r_k^{(1)}\{\beta_0' A_i(u_k)\} \\
& \quad - e_{W(1),j}(\beta_0,u_j)W_i(u_j)Y_i(u_j)r\{\beta_0' A_i(u_j)\}W_i(u_k)Y_i(u_k)r_k^{(1)}\{\beta_0' A_i(u_k)\} \\
& \quad - W_i(u_j)Y_i(u_j)r_j^{(1)}\{\beta_0' A_i(u_j)\}W_i(u_k)Y_i(u_k)r\{\beta_0' A_i(u_k)\}e_{W(1),k}(\beta_0,u_k) \\
& \quad + e_{W(1),j}(\beta_0,u_j)W_i(u_j)Y_i(u_j)r\{\beta_0' A_i(u_j)\}W_i(u_k)Y_i(u_k)r\{\beta_0' A_i(u_k)\}e_{W(1),k}(\beta_0,u_k)\Big] \\
=& Q_{(j,k)}^{(1)}(\beta_0,u_j,u_k) - e_{W(1),j}(\beta_0,u_j)Q_k^{(2)}(\beta_0,u_j,u_k) \\
& - Q_j^{(2)}(\beta_0,u_k,u_j)e_{W(1),k}(\beta_0,u_k) + e_{W(1),j}(\beta_0,u_j)Q^{(0)}(\beta_0,u_j,u_k)e_{W(1),k}(\beta_0,u_k),
\end{aligned}
$$

and the second term inside the bracket is given by

$$
\begin{aligned}
\mathbf{F}_{\cdot n,j}(\mathbf{X}_n) - \mathbf{F}_{\cdot n,k}(\mathbf{X}_n) =& S_{W_{(1),j}}^{(1)}(\beta_0,u_j)S_{W_{(1),k}}^{(1)}(\beta_0,u_k) \\
& - e_{W(1),j}(\beta_0,u_j)S_{W_{(1)}}^{(0)}(\beta_0,u_j)S_{W_{(1),k}}^{(1)}(\beta_0,u_k) \\
& - S_{W_{(1),j}}^{(1)}(\beta_0,u_j)e_{W(1),k}(\beta_0,u_k)S_{W_{(1)}}^{(0)}(\beta_0,u_k) \\
& + e_{W(1),j}(\beta_0,u_j)S_{W_{(1)}}^{(0)}(\beta_0,u_j)S_{W_{(1)}}^{(0)}(\beta_0,u_k)e_{W(1),k}(\beta_0,u_k).
\end{aligned}
$$

Then $\lim_{n\to\infty} 2\Big[n^{-1}\sum_{i=1}^{n}\mathbf{F}_{in,j}(\mathbf{X}_n)\mathbf{F}_{in,k}(\mathbf{X}_n) - \mathbf{F}_{\cdot n,j}(\mathbf{X}_n)\mathbf{F}_{\cdot n,k}(\mathbf{X}_n)\Big]$ can be rewritten as

$$
\begin{aligned}
& \lim_{n\to\infty} 2\Big[\Big\{Q_{(j,k)}^{(1)}(\beta_0,u_j,u_k) - S_{W_{(1),j}}^{(1)}(\beta_0,u_j)S_{W_{(1),k}}^{(1)}(\beta_0,u_k)\Big\} \\
& \quad - e_{W(1),j}(\beta_0,u_j)\Big\{Q_k^{(2)}(\beta_0,u_j,u_k) - S_{W_{(1)}}^{(0)}(\beta_0,u_j)S_{W_{(1),k}}^{(1)}(\beta_0,u_k)\Big\} \\
& \quad - \Big\{Q_j^{(2)}(\beta_0,u_k,u_j) - S_{W_{(1)}}^{(0)}(\beta_0,u_k)S_{W_{(1),j}}^{(1)}(\beta_0,u_j)\Big\}e_{W(1),k}(\beta_0,u_k) \\
& \quad + e_{W(1),j}(\beta_0,u_j)\Big\{Q^{(0)}(\beta_0,u_j,u_k) - S_{W_{(1)}}^{(0)}(\beta_0,u_j)S_{W_{(1)}}^{(0)}(\beta_0,u_k)\Big\}e_{W(1),k}(\beta_0,u_k)\Big] \\
=& \lim_{n\to\infty} 2\Big[H_{(j,k)}^{(1)}(\beta_0,u_j,u_k) - e_{W(1),j}(\beta_0,u_j)H_k^{(2)}(\beta_0,u_j,u_k) \\
& \quad - H_j^{(2)}(\beta_0,u_k,u_j)e_{W(1),k}(\beta_0,u_k) + e_{W(1),j}(\beta_0,u_j)H^{(0)}(\beta_0,u_j,u_k)e_{W(1),k}(\beta_0,u_k)\Big],
\end{aligned}
$$

where

$$H^{(0)}(\beta, x, v) = Q^{(0)}(\beta, x, v) - S^{(0)}_{W_{(1)}}(\beta, x) S^{(0)}_{W_{(1)}}(\beta, v)$$
$$H^{(1)}(\beta, x, v) = Q^{(1)}(\beta, x, v) - S^{(1)}_{W_{(1)}}(\beta, x) S^{(1)}_{W_{(1)}}(\beta, v)'$$
$$H^{(2)}(\beta, x, v) = Q^{(2)}(\beta, x, v) - S^{(0)}_{W_{(1)}}(\beta, x) S^{(1)}_{W_{(1)}}(\beta, v).$$

It follows that (3.51) is the (j,k)th element of $G(\beta_0, u_j, u_k)$ in view of convergence property implied by conditions D and G(iii). Then it can be seen that the limiting covariance function of D_n is given by G, and therefore we complete showing the in distribution convergence of (3.40) to a Gaussian random variable. By applying the basic properties of covariance matrix, we obtain the limiting covariance function of C_n given by Δ_α.

Chapter 4

Efficient Inference of Case-Cohort Marginal Structural Cox Models

4.1 Introduction

Marginal structural models (MSMs) are useful tools to make causal inference from longitudinal observational studies in the presence of *time-varying confounders*; time-dependent variables that predict subsequent treatment. MSMs are made upon the notion of potential outcome introduced by Neyman [1923] and Rubin [1974]. The method for obtaining MSM estimators accounts for confounding by incorporating inverse-probability-of-treatment-weights (IPTW) and enables to study causal relationship between exposures and outcome. In this paper, we focus on evaluating causal effect of treatment on time to disease occurrence in longitudinal observational studies, in the presence of confounding.

Recently, Cole et al. [2012] considered employing the case-cohort design to the MSCM analysis as a cost-efficient approach. The case-cohort study involves two-phase sampling: simple random sampling without replacement to form a full cohort from an infinite superpopulation at phase 1, and random sampling a subcohort from the full cohort as well as sampling all subjects who experience a predefined event (henceforth, cases) at phase 2. Subcohort and cases will form *case-cohort* sample. We refer variables that are observed from the full cohort to as *phase 1 variables*, and refer variables only available for the case-cohort sample to as *phase 2 variables*. We can achieve cost and effort saving by restricting collection of expensive variables based upon phase 2 subjects only. For example, high cost associated with determination of covariate information such as CD4 counts or viral load from biomarkers in HIV studies, or the cost associated with genotyping a large number of subjects in genetic studies can be avoid by employing the case-cohort design.

Prentice [1986*a*] described a pseudo-likelihood approach for the hazard ratio parameter estimation in the Cox model along with heuristic procedures for parameter estimation when the case-cohort design is applied. Asymptotic distribution theory of the case-cohort maximum pseudo-likelihood estimator was developed by Self and Prentice [1988] using martingale technique and finite population convergence results. Both Prentice [1986*a*] and Self and Prentice [1988] do not accommodate inverse weights accounting for sampling of subjects, i.e., they considered unweighted pseudo-likelihoods.

After Prentice [1986*a*] and Self and Prentice [1988], various methods have been proposed as means of improving the efficiency of the hazard ratio estimation (compared to Prentice [1986*a*] and Self and Prentice [1988]) in the standard (associational) case-cohort Cox regression analysis. Barlow [1994] and Barlow et al. [1999] considered estimators based on weighted pseudo-likelihood estimation. At each failure time, contribution of cases and nonfailures (controls) at risk are weighted by either fixed or time-varying inverse-sampling-weights (ISW) to account for subcohort sampling.

Later, methods that seek to utilize some of the phase 1 covariate information were proposed. Borgan et al. [2000] considered a stratified sampling by a phase 1 variable which is a correlate of exposure, to incorporate statum-specific ISW in the estimating equation. Stratum-specific ISW can be calculated using empirical sampling fraction within each stratum. They proposed three different estimating equations by considering different types of weights. Simulation studies suggested that the stratified estimator II with time-varying ISW, referred to as BII estimator from herein, is the most efficient among the existing estimators. Kulich and Lin [2004] established asymptotic theory for the BII type of estimators. In addition they proposed a new class of weighted estimators with general time-varying ISW; doubly weighted (DW) estimator and combined doubly weighted (CDW) estimator. The methods involve a modeling step for prediction of the values of each partially missing phase 2 variables, and is likely of greatest use when there are only 1 or 2 such variables. The authors suggest to use CDW estimator in practice as DW estimator is efficient only if a model to predict the phase 2 variables given all the phase 1 variables is correct. Briefly, the CDW estimator can be calculated through five steps:

1. Stratify by a correlate (referred to as surrogate in Kulich and Lin [2004], which is

a part of phase 1 variable) of a phase 2 variable. Stratification can be done using non-surrogate phase 1 variables as well, however, the stratification must incorporate a surrogate.

2. Develop models using subcohort controls data for prediction of the phase 2 variables and obtain estimated values of the missing phase 2 variables.

3. Evaluate time-varying ISW for all subjects.

4. Obtain BII estimator and estimate several covariance functions evaluated at the value of BII estimator to obtain a weight matrix which will affect on efficiency of CDW estimator.

5. Iteratively solve a proposed estimating equation with plugging-in BII estimator as an initial value. The proposed estimating equation involves the weight matrix mentioned above.

Numerical studies indicated that the CDW estimator is more efficient than other existing estimators such as Chen and Lo [1999], Borgan et al. [2000], and Chen et al. [2001]. The efficiency gain for the phase 2 covariates depends on the ability of the first-phase data to predict the true values of the partially missing variables. Later, Breslow et al. [2009*a*] and Breslow et al. [2009*b*] considered adjustment of ISW by calibration or estimation which making use of phase 1 covariate information. Calibration method adjusts ISW to be as close as possible to the sampling weights subject to a constraint that the cohort total of V equals to its weighted sum among sampled subjects. Estimation methods uses ISW as inverse of inclusion probabilities estimated from a logistic regression model that predicts which cohort subjects are sampled at phase 2. Simulation study and real data analysis reported by Breslow et al. [2009*b*] showed that such adjustment on ISW can dramatically improve precision of estimation for baseline covariates effects (i.e., a part of phase 1 variables that are known for all) on an outcome. They also showed that the methods can improve precision of estimation for phase 2 covariates effects when there exists a strong surrogate for the partially missing covariates.

While Barlow [1994] and Barlow et al. [1999] considered only time-varying ISW to improve efficiency, the rest of methods seek to make better use of information that are

available from all subjects by stratification using phase 1 variables that are correlated with the phase 2 variables. However, aforementioned methods showed efficiency improvement for either baseline or phase 2 hazard ratio estimation. Under a biomedical observational study in the presence of confounding, we are in slightly different situation that we have primary exposure (typically, treatment) which is available from all subjects but missing confounders information at phase 2. Primary interest lies on evaluating effect of time-varying treatment on a predefined outcome while marginal effect of phase 2 variables on the outcome is less important. Rather, the phase 2 variable information is used to account for confounding. Therefore, some of the methods described above may not be applicable (or useful) to improve efficiency of hazard ratio estimation of treatment in MSCM case-cohort analysis.

In this paper we describe how the aforementioned methods that are developed in the standard associational context can be extended to the causal setting, and discuss why some of the methods cannot be readily applicable to the causal setting. In addition, we propose a new method to improve efficiency in the MSCM case-cohort analysis, which incorporates use of all subject in the full cohort. The rest of this paper is organized as follows: In §4.2, we describe general methods to obtain consistent and asymptotically normally distributed MSCM estimators under both the full and the case-cohort settings [Lee et al., 2013]. In §4.3, we demonstrate how the general methods can be combined with some of the discussed methods in this section, and why some of the discussed methods cannot be applicable to the causal setting. A new method to improve efficiency is discussed. In §4.4, we report the results of our simulation studies, and illustrate the proposed methods with an example. We finalize this paper with discussion in §4.5.

4.2 General Methods for MSCM Case-Cohort Estimators

We assume a study comprised of n different individuals indexed by i, which aims to evaluate the effect of treatment on a time to event outcome. Capital letters will represent random variable and lower letters will represent values of variables or constants. Subject index i may be suppressed when there is no ambiguity. Let T be a failure time, C be a censoring time, and $\Delta = I(T \leq C)$. Define observed time $X = \min(T, C)$, counting process

$N(t) = I(X \le t, \Delta = 1)$, and at risk process $Y(t) = I(X \ge t)$. A subject whose failure time is observed (i.e., $\Delta_i = 1$ and $X_i = T_i$) is referred to as a case, and a censored subject (i.e., $\Delta_i = 0$ and $X_i = C_i$) is referred to as a control. Let $A(\cdot)$ be a p-vector of treatment process, let $L(\cdot)$ be a p-vector of covariate process, and let V be baseline covariates which may be a part of $L(0)$. Consider a study conducted from time 0 to time 1 where measurements are collected at c different times. Assume $L(t)$ is temporally prior to $A(t)$, i.e., decision of treatment is made after obtaining covariate information at time $t \in [0, 1]$. We use overbar notation to represent history up to and including time t; $\overline{A}(t) = \{A(u) : 0 \le u \le t\}$, $\overline{L}(t) = \{L(u) : 0 \le u \le t\}$, etc. Let $\overline{a}$ denote each possible static treatment plan, i.e., $\overline{a} = \{a(t) : 0 \le t \le 1\}$. Define $T_{\overline{a}}$ to be a subject's potential failure time had the subject been treated according to the plan $\overline{a}$, and $C(\cdot)$ to be a censoring process so that $C_i(t)$ is a censoring indicator, i.e., $C_i(t) = 0$ means that subject i is alive at time t and $C_i(t) = 1$ means that subject i is not alive at time t. Suppose that $C(t)$ is conditionally independent of $T_{\overline{a},\overline{C}(1)=0}$ given $A(t^-)$ and $L(t^-)$. Under the usual causal assumptions such that *causal consistency*, *conditional exchangeability*, and *positivity*, we can obtain the causal effect of a function of treatment using MSCMs, which are given by

$$\lambda_{T_{\overline{a}}}(t) = \lambda_0(t) \exp\{\beta_0' f(\overline{a}(t))\}$$

where $\lambda_{T_{\overline{a}}}(t)$ is the hazard of failure at time t if all individuals in the population had followed treatment plan $\overline{a}$ through time t, $\lambda_0(t)$ is an unspecified baseline hazard function corresponding to the hazard if all individuals had been untreated through time t, $f(\overline{a}(t))$ is a user-specified function of treatment history up to time t, and β_0 is an unknown relative risk parameter vector. For notational convenience, consider the following MSCMs,

$$\lambda_{T_{\overline{a}}}(t) = \lambda_0(t) r\{\beta_0' a(t)\}, \tag{4.1}$$

i.e., let us focus on the causal effect of current treatment $a(t)$, using notation $r\{\cdot\}$ instead of $\exp\{\cdot\}$. In the presence of confounding, we consider weight process $W(\cdot)$ the form of which at time t is given by

$$W^T(t) = \prod_{k \le t} \frac{\text{pr}[A(k)|\overline{A}(k^-)]}{\text{pr}[A(k)|\overline{A}(k^-), \overline{L}(k)]}.$$

Under the assumptions of conditional exchangeability and positivity, W can consistently be estimated at any time $t \in [0,1]$. In the presence of censoring, we can further account for the bias due to censoring via weighting a subject alive and uncensored at time t using estimates of $W^T(t) \times W^C(t)$, where

$$W^T(t) = \prod_{k \le t} \frac{\text{pr}[A(k)|\overline{A}(k^-), \overline{C}(k^-) = 0]}{\text{pr}[A(k)|\overline{A}(k^-), \overline{L}(k), \overline{C}(k^-) = 0}], \qquad \text{and}$$
$$W^C(t) = \prod_{k \le t} \frac{\text{pr}[C(k) = 0|\overline{C}(k^-) = 0, \overline{A}(k^-)]}{\text{pr}[C(k) = 0|\overline{C}(k^-) = 0, \overline{A}(k^-), \overline{L}(k)]}.$$

From now on, consider the following weight process

$$W(t) = W^T(t) \times W^C(t), \tag{4.2}$$

which is referred to as *inverse-probability-weights* (IPWs). By further assuming the positivity on censoring, (4.2), can consistently be estimated.

With full data, β_0 in (4.1) would be estimated by solving the following weighted estimating equation:

$$U_F(\beta) = \sum_{i=1}^{n} \int_0^1 W_i(t)[A_i(t) - E_{W_{(1)}}(t, \beta)]dN_i(t) = 0, \tag{4.3}$$

which is the weighted partial likelihood score function, where for $j = 0, 1, 2$, and $k = 1, 2$,

$$E_{W_{(1)}}(t, \beta) = S^{(1)}_{W_{(1)}}(t, \beta)/S^{(0)}_{W_{(1)}}(t, \beta),$$
$$S^{(j)}_{W_{(k)}}(\beta, t) = n^{-1} \sum_{l=1}^{n} W_l(t)^k Y_l(t) r^{(j)}\{\beta' A_l(t)\},$$
$$r^{(j)}\{\beta' A_l(t)\} = A_l(t)^{\otimes j} r\{\beta_0' A_l(t)\},$$

and $c^{\otimes j}$ are defined by $c^{\otimes 0} = 1$, $c^{\otimes 1} = c$, $c^{\otimes 2} = cc'$ for a $p \times 1$ vector c.

Consider a set of individuals $\tilde{\mathcal{C}}$ of size $\tilde{n}$ that is randomly selected without replacement from the full cohort of size $n(\ge \tilde{n})$, i.e., $\tilde{\mathcal{C}}$ is a subcohort. With the case-cohort data, we can consistently estimate β_0 in (4.1) by solving either of the following weighted estimating

equations:

$$\tilde{U}(\beta) = \sum_{i=1}^{n} \int_0^1 W_i(t)[A_i(t) - \tilde{E}_{W_{(1)}}(t,\beta)]dN_i(t) = 0, \tag{4.4}$$

$$U^{\star}(\beta) = \sum_{i=1}^{n} \int_0^1 W_i(t)[A_i(t) - E^{\star}_{W_{(1)}}(t,\beta)]dN_i(t) = 0 \tag{4.5}$$

where for $j = 0, 1, 2,$ and $k = 1, 2,$

$$\tilde{E}_{W_{(1)}}(t,\beta) = \tilde{S}^{(1)}_{W_{(1)}}(t,\beta)/\tilde{S}^{(0)}_{W_{(1)}}(t,\beta),$$
$$E^{*}_{W_{(1)}}(t,\beta) = S^{*(1)}_{W_{(1)}}(t,\beta)/S^{*(0)}_{W_{(1)}}(t,\beta)$$
$$\tilde{S}^{(j)}_{W_{(k)}}(\beta,t) = \tilde{n}^{-1} \sum_{l \in \tilde{\mathcal{C}}} W_l(t)^k Y_l(t) r^{(j)}\{\beta' A_l(t)\},$$
$$S^{*(j)}_{W_{(k)}}(\beta,t) = \tilde{n}^{-1} \sum_{l \in \tilde{\mathcal{C}} \cup \{i\}} W_l(t)^k Y_l(t) r^{(j)}\{\beta' A_l(t)\}.$$

Asymptotic theories of the full and the case-cohort MSCM estimators obtained by solving (4.3) - (4.5) have been established by Lee et al. [2013]. As Lee et al. [2013] have shown that (4.4) and (4.5) are asymptotically equivalent, we mainly focus on adjusting (4.4) to improve efficiency. Similar arguments can be made based on (4.5).

Cole et al. [2012] used (4.5) to obtain MSCM parameter estimator in simulation studies and real data analysis. Both simulation and real data analysis results demonstrated that the case-cohort MSCM parameter estimator is less efficient than the full cohort MSCM parameter estimator. This is inevitable because we only makes use of partial data due to phase 2 variables in the case-cohort analysis. In the next section, we describe how existing methods can be extended to the MSCM case-cohort analysis and why application of some of the methods would be limited in the causal setting. Then we propose a new method to improve efficiency in the MSCM case-cohort analysis.

4.3 Improving Efficiency of the Estimation

Let ξ be a binary random variable that indicates the selection of a subject into the subcohort, and α be the selection probability, i.e., $\text{pr}(\xi = 1) = \alpha$, where $\alpha > 0$. α is a known probability as $\alpha = \tilde{n}/n$, however, we can estimate sampling probability at each failure time

using observed data. For example, we can use $\hat{\alpha}(t) = \sum_{i=1}^{\tilde{n}} Y_i(t) / \sum_{i=1}^{n} Y_i(t)$, i.e., number of subcohort members who are at risk at time t divided by number of full cohort members who are at risk at time t.

4.3.1 Time-Varying Inverse Sampling Weights

Following Barlow [1994], Barlow et al. [1999], and Borgan et al. [2000], we may improve efficiency of the MSCM hazard ratio estimation in the case-cohort analysis by solving the following doubly-weighted estimating equation:

$$\tilde{U}_B(\beta) = \sum_{i=1}^{n} \int_0^1 \varrho_i(t) W_i(t) [A_i(t) - \tilde{E}_{BW_{(1)}}(t, \beta)] dN_i(t) = 0, \tag{4.6}$$

where

$$\begin{aligned} &\varrho_i(t) = \Delta_i + (1 - \Delta_i)\xi_i / \hat{\alpha}(t), \\ &\tilde{E}_{BW_{(1)}}(t, \beta) = \tilde{S}^{(1)}_{BW_{(1)}}(t, \beta) / \tilde{S}^{(0)}_{BW_{(1)}}(t, \beta), \\ &\tilde{S}^{(j)}_{BW_{(k)}}(\beta, t) = \tilde{n}^{-1} \sum_{l \in \tilde{\mathcal{C}}} \varrho_l(t) W_l(t)^k Y_l(t) r^{(j)}\{\beta' A_l(t)\}. \end{aligned} \tag{4.7}$$

(4.7) assigns weight of 1 if subject i is being a case at time t, and assigns weight of $\hat{\alpha}(t)^{-1}$ if subject i remains a subcohort control at time t. Consider new IPWs incorporating $\varrho_i(t)$,

$$W_i^{\dagger}(t) = \varrho_i(t) \times W_i(t),$$

then (4.6) becomes (4.4) where $W_i^{\dagger}(t)$ substitutes for $W_i(t)$. Therefore, we can use the variance estimator proposed by Lee et al. [2013] or the robust variance estimator by Lin and Ying [1993] as both of them are shown to perform well in Lee et al. [2013]. When subcohort size is small, the variance estimator proposed by Lee et al. [2013] is preferable.

Efficiency gain comes from using an estimated sampling probability $\hat{\alpha}$ rather than using the known true sampling probability α at a given time [Robins, Rotnitzky and Zhao, 1994]. However, in MSCM analysis, variability coming from adding ϱ to IPWs may attenuate the efficiency gain due to increase of bias. Simulation results reported by Barlow et al. [1999]

also indicated that the unweighted analysis using Prentice-type [Prentice, 1986*a*] likelihood (which corresponds to (4.5) in causal context) may be preferable due to increase of bias. Our small simulation study result in §4.4 implies that variability coming from adding another inverse probability may attenuate efficiency.

Most literature regarding efficiency improvement in the standard case-cohort analysis seek to make use of various types of ISWs. As described in §4.1, methods proposed by Borgan et al. [2000]; Kulich and Lin [2004]; Breslow et al. [2009*a*,*b*] all make use of time-varying ISWs as means of improving efficiency. Therefore, application of these method may not useful to improve efficiency in MSCM case-cohort analysis, especially in the presence of informative censoring (that IPW involves IPCW as well as IPTW). However, it will be an interesting project to compare efficiencies of different MSCM case-cohort parameters obtained based on these methods.

As previously developed methods to improve efficiency in the standard case-cohort analysis may not be advantageous in the causal setting, we propose a new method that can improve efficiency in MSCM case-cohort analysis.

4.3.2 Imputation Method

We propose to adopt the multiple imputation (MI) method to improve efficiency in MSCM case-cohort analysis. One strategy to handling missing data is substituting missing values using simple imputation and treat imputed values as if they are observed. However, single imputation does not account for the uncertainty about the predictions of missing values and estimated variances of the parameter estimates are known to be biased toward zero. MI replaces each missing value using a set of plausible values reflecting uncertainty of the imputation model which leads to a valid statistical inference. MI has been recognized as a practical and flexible method for handling missing data as it becomes widely available in most statistical packages such as R, SAS, and STATA.

Case-cohort studies can be viewed as a special type of incomplete data, where phase 2 covariates are missing at random (MAR). Therefore, we restrict our interest in imputation of missing phase 2 covariates, and do not consider missing treatment/outcome. Briefly speaking, MAR assumes that probability of missingness depends only on observed data (and is independent of unobserved part of data). This is true for the case-cohort design,

as phase 2 covariates are missing by (i) failure status which is the observed variable, and by (ii) subcohort inclusion/exclusion status which is determined by a random mechanism. In this section we describe MI with MSCM case-cohort studies and present corresponding maximum likelihood estimator. We also present asymptotic theory of MI estimator based on two papers, by Wang and Robins [1998] and Robins and Wang [2000]. Chapter 14 of Tsiatis [2006] is an excellent reference to study large sample theories for MI estimators in both frequentist and Bayesian perspectives. In what follows, we consider a Bayesian type estimator, which was referred to as *Type A* or *proper* imputation estimator in Wang and Robins [1998] and Tsiatis [2006].

Notation First, note that A and observed time X are available in practice. Therefore, we assume that L represents the phase 2 variables and that only L is missing (for non-subcohort controls) in case-cohort data. Note that L is required to calculate (4.2), but not to obtain estimator of β_0 in (4.1). Let the full cohort data be denoted by $Z = \{Z_1, .., Z_n\}$ where Z_i is assumed iid with density $f_Z(z, \beta)$. Here we slightly abuse notation and Z_i can be written as $\overline{Z}_i(1)$ to be consistent with the previous notation discussed in § 4.2. Let $R_i = 1$ if $\overline{L}_i(1)$ is observed and $R_i = 0$ otherwise, i.e., R denote the indicator of a complete data (i.e., case-cohort sample inclusion indicator) which is time-invariant. Therefore, the observed data at the end of study period $t = 1$ (i.e., case-cohort data) can be expressed as $\{R_i, Z_i; i = 1, ...n\}$. More generally, the observed data can be written as

$$\{R_i, G_{R_i}(Z_i)\}, i = 1, ..., n.$$

where G_R is a known function associated with the data coarsening variable R. Hereafter, we slightly change notation $U^F(\beta)$ given in (4.3), which denote the full cohort score function, to $U^F(z, \beta)$. We use $U^F(z, \beta)$ notation for imputed full cohort score function as well. In similar fashion, let $U\{R_i, G_{R_i}(Z_i), \beta\}$ be the observed case-cohort score function, i.e., we use notation $U\{R_i, G_{R_i}(Z_i), \beta\}$ rather than $\tilde{U}(\beta)$ or $U^*(\beta)$ in (4.4) or (4.5).

Assumptions We assume that time-varying confounders are continuous. There may be a case that one of time-varying confounders is binary or categorical. In that case, we assume that there exists continuous variable that can be mapped to the binary/categorical variable.

We assume that a (continuous) time-varying confounder is a repeated measurement. In particular, with a study of τ visits we assume that

$$L_i = (L_i(1), ..., L_i(\tau)) \sim N(D_i\psi, \Sigma_i) \tag{4.8}$$

where D_i is the design matrix for an individual i based on all available information such as baseline covariates, treatment history, event/censoring time, with time information for a fitted repeated measures model. ψ is the regression coefficients and Σ_i is the covariance matrix for L_i which includes variance-covariance parameters for the model.

MI estimator Assuming a time-varying confounder is a repeated measurement, implementation of Bayesian MI to a (longitudinal) observational case-cohort study as follows: We randomly sample each missing value m times from the conditional distribution

$$f_{Z|R_i,G_{R_i}}\{z|R_i, G_{R_i}, \psi^{(j)}\} \tag{4.9}$$

where $\psi^{(j)}$ itself is sampled from some distribution, say $f\{\psi|R_i, G_{R_i}\}$. Therefore, we first draw $\psi^{(j)}$ from $f\{\psi|R_i, G_{R_i}\}$ in the j-th imputation and then draw missing Z_i from the posterior distribution (4.9) evaluated at $\psi^{(j)}$. Simulation of (4.9) is easy if missing pattern is monotone, i.e., missing L_j for individual i implies that all subsequent variables $L_k, k \geq j$ are missing for that individual. In such setting imputation strategy could be flexible and one can implement regression, propensity [Rubin, 1987], or predictive mean matching [Heitjan and Little, 1991; Schenker and Taylor, 1996]. Monotone missing assumption is satisfied if no case-cohort subjects miss any study visits (i.e., there is no missing for the case-cohort subjects). This could be true if a study has small number of study visits. In our simulation study, we assume that a time-varying confounder is measured at two consecutive time points (or three time points if baseline is considered). We consider missing phase 2 variables for non-subcohort controls only, and therefore we could assume monotone missing pattern (e.g., if $L_i(1)$ is missing then $L_i(2)$ should be missing). If missing pattern is not monotone, the posterior distribution can be simulated using Markov chain Monte Carlo (MCMC) algorithm [Schafer, 1997]. In MACS and WHIS studies, case-cohort subjects missed some of semi-annual visits due to the long term follow-up periods. Therefore, we implement MCMC to

simulate posterior distribution. All methods and algorithms are implemented in `proc mi` procedure in SAS version 9.3.

Let Z_{ij} be covariate information of i-th individual in the j-th imputed full data where $j = 1, ..., m$, i.e., $Z_{ij} = Z_i$ if $R_i = 1$ and Z_{ij} is a sampled value from posterior distribution of the missing data if $R_i = 0$. Let $\hat{\beta}_j$ be the solution to the j-th imputed full data score equation:

$$\sum_{i=1}^{n} U^F\{Z_{ij}(\psi^{(j)}), \hat{\beta}_j\} = 0. \tag{4.10}$$

Then the MI estimator based on m imputed data sets is defined by

$$\hat{\beta}_m = m^{-1} \sum_{j=1}^{m} \hat{\beta}_j \tag{4.11}$$

and corresponding variance estimator of $\hat{\beta}_m$ proposed by Rubin [1987] is given by

$$T_m = \bar{V}_m + \{1 + m^{-1}\}B_m, \tag{4.12}$$

where $\bar{V}_m = m^{-1} \sum_{j=1}^{m} V_j$, V_j represents standard error associated with $\hat{\beta}_j$, and $B_m = (m - 1)^{-1} \sum_{j=1}^{m} (\hat{\beta}_j - \hat{\beta}_m)^2$ with $j = 1, ..., m$ being the imputation index. The term $\{1 + m^{-1}\}B_m$ is associated with between-imputation variance reflecting uncertainty due to sampling variability. MI procedure based on score equation (4.10) is referred to as *proper* impuation by Rubin [1987]. SAS `proc mianalyze` provides MI estimator (4.11) and corresponding variance estimate (4.12).

Asymptotic Distribution of MI estimator Under the assumptions of i) proper imputation, ii) correct model specification for both imputation/analysis models, and iii) with large samples (i.e., when n goes to infinity), Rubin [1987] (p.86) showed asymptotic distribution of MI estimator and corresponding variance estimator in Bayesian context; i.e., he showed asymptotic posterior distribution of $\hat{\beta}_\infty - \beta_0$ given observed data follows mean zero normal distribution, where $\hat{\beta}_\infty = \lim_{m \to \infty} \hat{\beta}_m$. In this paper, we present large sample frequentist property of the Bayesian MI estimator. Results presented in this paragraph are mainly taken from Chapter 14 of Tsiatis [2006].

Under the same assumption as in Rubin [1987], (4.11) is consistent and asymptotically normally distributed estimator of β_0 in the sense that

$$n^{1/2}(\hat{\beta}_m - \beta_0) \to_d N(0, \mathcal{T}_m),$$

where $\mathcal{T}_m$ is composed of information matrices evaluated at β_0. In particular,

$$\begin{aligned}\mathcal{T}_m =& I^F(\beta_0)^{-1} + \Big(\frac{m+1}{m}\Big) I^F(\beta_0)^{-1}\{I^F(\beta_0) - I(\beta_0)\} I^F(\beta_0)^{-1} \\ &+ \Big(\frac{m+1}{m}\Big) I^F(\beta_0)^{-1}\{I^F(\beta_0) - I(\beta_0)\}\mathrm{var}[q\{R_i, G_{R_i}(Z_i)\}]\{I^F(\beta_0) - I(\beta_0)\} I^F(\beta_0)^{-1}.\end{aligned} \tag{4.13}$$

where $I(\beta) = -E[\partial U\{R_i, G_{R_i}(Z_i), \beta\}/\partial\beta']_{\beta=\beta_0}$ and $I^F(\beta) = -E[\partial U^F(z,\beta)/\partial\beta']_{\beta=\beta_0}$ which denote information matrices based on the observed and the (imputed) full data, respectively, and $q\{R_i, G_{R_i}(Z_i)\}$ is the influence function of initial estimator of β_0 (Tsiatis [2006], p.369).

Rubin's MI variance estimator (4.12) converges in expectation to $\mathcal{T}_m$ when n goes to infinity. In addition, when m goes to infinity, (4.12) is a consistent and asymptotically unbiased estimator of $\lim_{m\to\infty} \mathcal{T}_m$ (Tsiatis [2006], p.370-371).

Caution When we refer asymptotic properties (i.e., asymptotic bias, consistency, etc) from herein, we assume that n goes to infinity but m can be finite unless otherwise stated.

Rubin's variance estimator is asymptotically unbiased (i.e., converge in *expectation* to its asymptotic variance) when MI is done using Bayesian approach. It is important to distinguish this approach with frequentist approach of MI. Frequentist approach fixes $\psi^{(j)}$ in (4.10) at maximum likelihood estimator of ψ, say $\hat{\psi}$, and sample at random from the conditional distribution $f_{Z|R_i,G_{R_i}}\{z|R_i, G_{R_i}, \hat{\psi}\}$ to obtain random quantities $Z_{ij}, j = 1, ..., m$. This approach was referred to as *improper imputation* by Rubin. Bayesian MI introduces additional variability coming from sampling $\psi^{(j)}$ from its posterior distribution at each imputation and is less efficient than frequentist MI approach. Therefore, Rubin's variance estimator will be biased when missing data is filled by improper imputation. Wang and Robins [1998] and Robins and Wang [2000] showed asymptotic distribution theories of improper imputation estimators and described estimation of asymptotic variance. Although frequentist MI is more efficient approach with finite m, no statistical packages or built-in procedures are available to obtain variance estimator of the MI estimator. In addition,

application of Wang and Robins [1998] or Robins and Wang [2000] on longitudinal study data requires calculation of score function with respect to the variance-covariance parameters associated with repeated measurements to evaluate influence function. This involves highly complex analytic form when repeated measurement is considered. As difference between frequentist and Bayesian asymptotic variance disappears as m goes to infinity we recommend implementing Bayesian MI approach with as big m as software and computing time allows. Chapter 14 (p.366- 369) of Tsiatis [2006] is a great reference to see asymptotic property of Rubin's variance estimator (4.12) and to see it compared to that of frequentist MI estimator.

Rubin [1987] stated that it is important to include all variables that are likely to be used in final analysis model (which is in this case, MSCM); leaving out some variables that are believed to be weak predictors implies that (we are certain that) those variables have no relation with the missing data. In MSCM MI analysis, we know that variables associated with time-to-event outcome (such as baseline covariates, treatment status, and confounders themselves) induce the missing data. Therefore, we include all available information that are used to model to calculate IPW and used in MSCM when imputation model is considered. As MI requires a correct imputation model specification Rubin [1987], sensitivity analysis for different model specifications or different imputation methods might be of interest to check the model assumption. Regression and MCMC methods discussed above assume multivariate normality (4.8). However, it is known that inference based on MI can be robust to departure from the assumption when missing fraction is not large [Schafer, 1997]. Therefore, departure from (4.8) assumption may be ignorable with high subcohort fraction.

4.4 Results

Below, we present simulation study and real data analysis results.

4.4.1 Simulation Studies

First, we show a small simulation study result based on doubly-weighted method using time-varying ISW, (4.6), in Table 4.1. Performance of the estimator based on (4.6) was compared to Lee et al. [2013]'s estimator which is based on estimating equation (4.4). As

$W^{\dagger}$ is still predictable, variance formula proposed in Lee et al. [2013] was employed to calculate standard error.

To compare performance of doubly-weighted estimator and Lee et al. [2013]'s estimator, we adopt the same simulation setting as in Cole et al. [2012], except censoring mechanism. Cole et al. [2012] generated censoring times according to administrative censoring mechanism, while we generated censorings from uniform distributions with support chosen to achieve the desired event rates. For details of the simulation setting, see Cole et al. [2012]. Briefly, we generated potential survival times when never exposed to treatment T_{∞}, when treated from baseline T_0, and when treated from t_1, T_{t_1}, by following Cole et al. [2012], with baseline hazard being 1 and t_1 being 0.1, for cohorts of size $n = 1,000$. Then we generated baseline treatment status, say A_0, from Bernoulli(1/3). Then two time varying confounders, say L_1 and L_2 were generated; L_1 was generated from Bernoulli with probability dependent on A_0 and T_{∞} where marginal probability equals 0.5. A second time-varying confounder was generated from standard normal distribution dependent on A_0 and T_{∞}. Note that, L_1 is a binary and L_2 is continuous variable respectively. Finally, we generated a time-varying exposure at time $t_1 = 0.1$, say A_3, from Bernoulli with probability dependent on L_1 and L_2, where marginal probability equals 0.5, for the two-thirds of subjects who were unexposed at baseline.

We set event rate to be 20% and randomly sampled 20% of subjects to form a subcohort. We generated 200 datasets under the null $\beta_0 = 0$ and the alternative $\beta_0 = \log(1/2) \approx -0.693$; β denotes the treatment effect parameter. Table 4.1 shows that the doubly-weighted estimating equation worsened the efficiency compared to (4.4). ESEs under the null and the alternative were bigger than those based on (4.4), and MSEs were bigger as well. As mentioned in §4.3.1, the simulation result may imply that variability coming from adding another inverse probability may attenuate efficiency in MSCM case-cohort analysis.

To see performance of the proposed MI estimator (4.11), we considered several different scenarios by varying subcohort fraction 10, 20, and 30%, and event rates from 10, 15, 20, and 25%. At each scenario, we generated 1,000 datasets under the null $\beta_0 = 0$ and the alternative $\beta_0 = \log(1/2)$.

In this simulation we combined simulation settings in Cole et al. [2012] and Moodie et al. [2008]. Moodie et al. [2008] compared different methods to handle missing *exposure* data

Table 4.1: Simulation studies to compare performance of estimators

Null	Bias	ASE	ESE	MSE	Cover	Power
(4.4)	-0.029	0.236	0.206	0.056	0.975	0.020
(4.6)	-0.032	0.253	0.217	0.065	0.970	0.025
Alternative						
(4.4)	-0.044	0.236	0.236	0.057	0.955	0.880
(4.6)	-0.099	0.250	0.23	0.072	0.940	0.895

Simulation studies to compare performance of estimators based on (4.6) with (4.4) when event rate and subcohort fraction equaled to 20%. Bias denotes the empirical bias of the different estimators based on (4.6) and (4.4). ASE denotes average of estimated standard errors. ESE denotes the empirical standard errors (i.e., average standard error of the estimators). MSE denotes the mean squared error calculated by $\{\text{Bias}^2 + \text{ASE}^2\}$. Cover denotes the empirical coverage of 95% Wald-type confidence intervals using Lee et al. [2013]'s variance estimator. Size/power denotes the proportion of simulated data sets where the hypothesis $\beta_0 = 0$ was rejected.

in marginal structural models, which is different missing data type than what we consider in this article. Potential survival times T_∞, T_0, and T_{t_1} were generated as described above with baseline hazard set to be 1 and t_1 set to be 0.1, for cohorts of size $n = 1,000$. Then we generated time-varying confounders and time-varying treatment by mimicking Cole et al. [2012] and Moodie et al. [2008]. A baseline covariate, say L_0, was drawn from Normal(3, 1) and a first time-varying confounder L_1 was drawn from Normal(10, 1). Then treatment status at time 0 (A_0) was generated from Bernoulli with probability dependent on L_1 where marginal probability equals 1/3. A second time-varying confounder L_2 was drawn from Normal($L_1 + \beta_0 A_0$, 1). Treatment status at time 1 (A_1) was generated from Bernoulli with probability dependent on L_2 and A_0, where marginal probability equals 0.5, for the two-thirds of subjects who were unexposed at baseline. Censoring times were from uniform distributions. For more details about the simulation setting, see Cole et al. [2012] and Moodie et al. [2008].

We assumed that A_0 and A_1 were available from all subjects in the study. To impute missing phase 2 covariates L_1 and L_2 for non-subcohort controls, we used all available information L_0, A_0, A_1, and observed time X from subcohort controls. We choose small m ($m = 5$) because we considered 15 different scenarios which required us extensive computation time. However in the real data analysis, we considered number of imputation $m = 100$. All MI analyses were done using `proc mi` and `mianalyze` procedures in SAS 9.3.

Table 4.2 - 4.4 show simulation results when subcohort sampling rates range from 10 to

30%. Numerical studies indicate that the proposed method can improve efficiency compared to Cole et al. [2012]. Empirical standard errors obtained by MI analysis are smaller than those of the case-cohort analysis and are close to those of the full cohort analysis in all scenarios. Bias based on MI are sometimes bigger than the case-cohort analysis, especially when event rate is relatively high (e.g., see bias when subcohort fraction is 10 or 20% and event rate 25%). Nonetheless, MSE values indicate that increase of bias are offset by efficiency gain. All three analysis methods exhibit correct coverage and power in all settings. Due to the efficiency gain, MI analysis is more powerful than the case-cohort and is as powerful as the full analysis.

4.4.2 Real Data Analysis

Study Cohort

We applied the proposed imputation method to a (combined) dataset comprised of HIV positive patient collected from the Multicenter AIDS Cohort Study (MACS) and the Women's Interagency HIV study(WIHS). The full cohort data analysis is consistent with that of Cole et al. [2012]. Readers who are interested in detailed information about study cohort is referred to Cole et al. [2012]. Briefly, participants in both studies were followed-up approximately every 6 months. While average years of follow-up was approximately 8 years, maximum years of follow-up was 12 years and thus maximum number of visits was 24. At each semiannual study visit, participants went through a physical examination, provided a blood sample, and completed a questionnaire about use of antiretroviral therapy, etc.

In the real data analysis, we aimed to estimate the effect of highly active antiretroviral therapy (HAART) initiation with acquired immunodeficiency syndrome (AIDS) incidence or death while adjusting for (confounding) effects of CD4 counts and HIV-1 RNA viral loads. To this end, we constructed a full cohort from the MACS and WIHS data which includes 950 HIV-1-seropositive men and women who were alive and not using antiretroviral therapies in April 1995 (because the first highly active regimen was approved on December 6, 1995). There were 211 incident AIDS or death (henceforth cases, 22%) in the full cohort. We selected a 20% random sample without replacement from the full cohort size of 950 using the same seed number as in Cole et al. [2012]. Among the 190 subcohort subjects, there were 47 cases (25%). The case-cohort consisted of 354 subjects, defined by 190 subcohort

Table 4.2: Simulation studies to compare performance of estimators when $\alpha = .1$

Null	Subcohort fraction (%)	Event rate (%)	Bias Bias	ASE	ESE	MSE	Cover	Size/ Power
Full	10	10	-0.004	0.212	0.218	0.045	0.949	0.051
Imputation			-0.003	0.213	0.219	0.045	0.952	0.058
Case-Cohort			-0.021	0.291	0.294	0.085	0.948	0.052
		15	-0.002	0.170	0.173	0.029	0.945	0.055
			-0.001	0.172	0.173	0.030	0.947	0.053
			-0.011	0.257	0.254	0.066	0.954	0.046
		20	-0.002	0.146	0.146	0.021	0.956	0.044
			0.000	0.148	0.146	0.022	0.957	0.043
			-0.006	0.242	0.239	0.059	0.952	0.048
		25	-0.002	0.130	0.129	0.017	0.961	0.039
			0.001	0.133	0.129	0.018	0.963	0.037
			-0.003	0.235	0.235	0.055	0.948	0.052
Alternative								
Full	10	10	-0.020	0.239	0.249	0.058	0.954	0.873
Imputation			-0.019	0.240	0.250	0.058	0.952	0.868
Case-Cohort			-0.024	0.307	0.319	0.095	0.946	0.640
		15	-0.017	0.185	0.189	0.035	0.945	0.977
			-0.014	0.188	0.190	0.035	0.952	0.971
			-0.013	0.266	0.269	0.071	0.953	0.749
		20	-0.014	0.156	0.159	0.024	0.948	0.993
			-0.010	0.159	0.160	0.025	0.952	0.991
			-0.006	0.249	0.250	0.062	0.959	0.799
		25	-0.013	0.137	0.140	0.019	0.949	0.999
			-0.010	0.140	0.140	0.020	0.951	0.997
			-0.004	0.241	0.240	0.058	0.950	0.819

Simulation studies to compare performance of estimators based on full cohort score equation (4.3), multiple imputation, and case-cohort score equation (4.5) when subcohort fraction equals 10%. Bias denotes the empirical bias of the different estimators of β_0. ASE denotes average estimated standard error. ESE denotes the empirical standard errors, which is defined by standard deviations of 1,000 log hazard ratio estimates. MSE denotes the mean squared error calculated by $\{\text{Bias}^2 + \text{ESE}^2\}$. Cover denotes the empirical coverage of 95% Wald-type confidence intervals using the robust variance estimator. Size/power denotes the proportion of simulated data sets where the hypothesis $\beta_0 = 0$ was rejected.

Table 4.3: Simulation studies to compare performance of estimators when $\alpha = .2$

Null	Subcohort	Event rate	Bias	ASE	ESE	MSE	Cover	Power
Full	20	10	-0.011	0.207	0.217	0.043	0.954	0.046
Imputation			-0.010	0.207	0.217	0.043	0.952	0.048
Case-Cohort			-0.020	0.246	0.269	0.061	0.916	0.084
		15	-0.013	0.171	0.174	0.029	0.948	0.052
			-0.012	0.171	0.174	0.029	0.948	0.052
			-0.020	0.214	0.224	0.046	0.928	0.072
		20	-0.012	0.146	0.142	0.021	0.960	0.040
			-0.011	0.147	0.142	0.022	0.958	0.042
			-0.008	0.195	0.193	0.038	0.956	0.044
		25	-0.010	0.130	0.127	0.017	0.964	0.036
			-0.009	0.131	0.127	0.017	0.966	0.034
			-0.015	0.185	0.188	0.034	0.942	0.058
Alternative								
Full	20	10	-0.006	0.230	0.232	0.053	0.946	0.886
Imputation			-0.006	0.231	0.232	0.053	0.948	0.878
Case-Cohort			-0.013	0.264	0.249	0.070	0.964	0.806
		15	-0.002	0.185	0.190	0.034	0.946	0.974
			-0.001	0.186	0.190	0.034	0.944	0.972
			-0.005	0.225	0.215	0.050	0.960	0.904
		20	0.012	0.155	0.156	0.024	0.946	0.988
			0.013	0.156	0.156	0.025	0.948	0.988
			0.011	0.202	0.194	0.041	0.966	0.944
		25	0.008	0.136	0.143	0.019	0.946	1
			0.010	0.137	0.143	0.019	0.950	1
			0.009	0.190	0.187	0.036	0.954	0.962

Simulation studies to compare performance of estimators based on full cohort score equation (4.3), multiple imputation, and case-cohort score equation (4.5) when subcohort fraction equals 20%.

Table 4.4: Simulation studies to compare performance of estimators when α = .3

Null	Subcohort	Event rate	Bias	ASE	ESE	MSE	Cover	Power
Full	30	10	-0.004	0.212	0.218	0.045	0.949	0.051
Imputation			-0.004	0.212	0.219	0.045	0.948	0.052
Case-Cohort			-0.005	0.235	0.243	0.055	0.947	0.053
		15	-0.002	0.170	0.173	0.029	0.945	0.055
			-0.002	0.171	0.173	0.029	0.946	0.054
			-0.002	0.197	0.197	0.039	0.942	0.058
		20	-0.002	0.146	0.146	0.021	0.956	0.044
			-0.001	0.146	0.146	0.021	0.958	0.042
			0.000	0.176	0.175	0.031	0.958	0.042
		25	-0.002	0.130	0.129	0.017	0.961	0.039
			-0.001	0.130	0.129	0.017	0.963	0.037
			0.001	0.164	0.161	0.027	0.965	0.035
Alternative								
Full	30	10	-0.020	0.239	0.249	0.058	0.954	0.873
Imputation			-0.020	0.239	0.249	0.058	0.956	0.873
Case-Cohort			-0.024	0.259	0.270	0.067	0.945	0.810
		15	-0.017	0.185	0.189	0.035	0.945	0.977
			-0.017	0.186	0.189	0.035	0.949	0.977
			-0.017	0.209	0.214	0.549	0.948	0.928
		20	-0.014	0.156	0.159	0.024	0.948	0.993
			-0.013	0.156	0.159	0.025	0.951	0.993
			-0.014	0.185	0.189	0.034	0.949	0.968
		25	-0.013	0.137	0.140	0.019	0.949	0.999
			-0.012	0.137	0.140	0.019	0.951	0.999
			-0.013	0.170	0.172	0.029	0.951	0.985

Simulation studies to compare performance of estimators based on full cohort score equation (4.3), multiple imputation, and case-cohort score equation (4.5) when subcohort fraction equals 30%.

subjects plus the 164 cases that were not selected in the subcohort. The outcome of interest was time to AIDS or death from any cause and the time-varying confounders were CD4 counts and HIV-1 RNA viral loads. Full cohort data included 9,172 person-visit records in total and case-cohort collected about 32% of the full cohort data (2,911 person-visit records), which lead to impute about 68% (6,261) of missing CD4 and viral load records along the course of study.

In the MI analysis, missing CD4 and viral load information for non-subcohort controls was imputed 100 times using `proc mi` procedure in SAS 9.3. As in simulation studies, we only used subcohort controls to build the posterior predictive distribution of missing data. We assumed that baseline CD4 or viral load information was available from all subjects, and used all available information such as treatment history (ever exposed to ART (yes/no)), gender, race, age (at study entry), and CD4 and viral load at baseline in MI analysis. Results based on m = 100 imputation results were summarized by `proc mianalyze` procedure in SAS 9.3, which makes use of Rubin's variance formula (4.12).

As a sensitivity analysis, we randomly selected subcohort 100 times by varying seed number from 1 to 100 using SAS 9.3. In this analysis, we aimed to account for sampling variability of subcohort in addition to checking against the robustness of the MI and the case-cohort analyses.

Results

Full cohort subjects characteristic is the same as in Cole et al. [2012] as we used the same dataset. The dataset consists of 61% women, 59% African American. The mean age of the full cohort participants at study entry was 39 years with standard deviation (SD) of 8, a CD4 cell count of 498 cells/mm^3 with SD of 279, and a log of HIV-1 RNA level (henceforth log of viral load) of 4.5 copies/mL with SD of 0.7 for detectable viral load values; there were 26% of missing in viral load values. The subcohort subjects had similar baseline characteristics at study entry (Table 4.5).

The full cohort analysis result is consistent with the previously reported result of Cole et al. [2012]. The inverse probability weighted hazard ratio for incident of AIDS or death was 0.41, with 95% confidence interval (CI) (0.26, 0.65). Standard error for log hazard ratio obtained by using the robust standard error was 0.23. In the MI analysis with 100

Table 4.5: Baseline characteristics of the full and the 50% subcohort subjects

Baseline Characteristic	cohort (n=950)			Subcohort (n=475)		
	%	Mean (SD)	No.	%	Mean (SD)	No.
Mean age (years)		39 (8)			39 (8)	
Female sex	61		578	58		111
African-American race	59		560	59		113
Mean CD4 cell (no. of cells/mm^3)		498 (279)			500 (260)	
Mean log_{10}viral load (no. of copies/mL)		4.5 (0.7)			3.9 (1.1)	

Baseline characteristics of the full and the 50% subcohort participants at study entry.

Table 4.6: Full cohort, 20% subcohort with MI, and case-cohort MSCM analyses

Analysis	Hazard Ratio	95% Confidence Interval (CI)	Standard Error (SE)
Full Cohort	0.41	0.26, 0.65	0.23
MI	0.48	0.30, 0.78	0.24
Case-Cohort	0.47	0.26, 0.83	0.29

Full cohort, 20% subcohort with MI, and case-cohort MSCM analyses of the causal effect of HAART initiation and incident AIDS or death among 950 men and women infected with HIV type 1 in the MACS and WIHS study, 1996-2007.

imputations, estimated hazard ratio was 0.48 with 95% CI (0.30, 0.78) and standard error of the log hazard ratio was 0.24; increase of standard error compared to the full cohort analysis was only 0.01, and width of the CI was slightly wider than that of the full cohort analysis (0.48 compared to 0.39). Estimated hazard ratio using the case-cohort analysis was 0.47, with standard error of the log hazard ratio 0.29. Compared to the full cohort analysis, increase in standard error of the case-cohort analysis was 0.07, and width of 95% CI was 0.57, which is about 1.5 times and 1.24 times wider than those of the full cohort and the MI analyses. As expected, analysis results for the MI MSCM analysis recovered much of the precision lost in the case-cohort analysis.

Table 4.7 shows sensitivity analysis of MI and the case-cohort analyses results using 100 randomly selected (20%) subcohorts. Reported estimates for case-cohort and for imputation analyses are averaged estimates. Considering full cohort result a gold standard, estimated hazard ratios based on imputation method and case-cohort analyses are slightly biased. Standard error based on the MI analysis with 100 imputations was about 1.21 times (0.29/0.24) smaller than that of the case-cohort analysis, yielding 1.27 times narrower 95% CI. Difference in standard errors of the full cohort and the MI analyses is only 0.1. Results

Table 4.7: Sensitivity analysis of case-cohort and multiple imputation

Analysis	Hazard Ratio	95% Confidence Interval (CI)	Standard Error (SE)
Full Cohort	0.41	0.26, 0.65	0.23
MI	0.50	0.32, 0.81	0.24
Case-Cohort	0.50	0.28, 0.90	0.30

Sensitivity analysis of case-cohort and multiple imputation (with 100 imputation) based on 100 random subcohorts sampled by varying seed number 1 to 100 in SAS. CI denote 95% Wald confidence interval. Standard error for multiple imputation was calculated based on Rubin's formula (1987) through MI analyze procedure in SAS.

in Table 4.7 implies that we could recover much of the precision lost from the case-cohort sampling by implementing the MI method.

4.5 Discussion

The proposed method is valid for a special type of primary exposure such that it can readily be obtained from existing study data repository. When treatment is the primary exposure, treatment assignment status or level of the treatment information given to a participant can be obtained with relatively less much cost and efforts than expensive covariate information. However, the proposed method is not suitable for studies in the presence of missing primary exposure in addition to phase 2 variables. For example, consider a study that aims to evaluate genetic variant on time to event response. One cannot readily obtain the genetic information from repository as much cost is required to validate the genetic information from the blood sample.

The proposed method aims to utilize information on all subjects in the estimating equation, and therefore it seeks to fill in missing IPWs for non-subcohort controls by imputing missing phase 2 covariates. Intuitively, this method is valid as the subcohort is selected at random from the full cohort; the phase 2 variables are missing completely at random. Therefore, estimated values of partially missing covariates based on the random sample of the full cohort should not deviate too much from the true values if (1) the sampling was truly done in random fashion, and (2) imputation model is a correct model when parametric model is used, or nonparamteric estimation is used. Simulation results indicated that parametric methods to estimate missing phase 2 variables can easily fail to improve efficiency when phase 2 variables are continuous.

The imputation method differs from the previously developed methods (in the standard associational context) which seek to utilize information available from the full cohort. In addition, we seek to make use of all subjects in the estimating estimation to improve efficiency in the case-cohort analysis. We do not require separate surrogate measurements of phase 2 variables as in Borgan et al. [2000]; Kulich and Lin [2004]; Breslow et al. [2009*a*,*b*], but time-varying confouders themselves can serve as surrogates (e.g., baseline CD4 or viral load can serve as surrogate of the following CD4 or viral load information).

The proposed estimator would be more efficient than estimators based on (4.4) or (4.5), because we use all subjects in the estimation step. Further, it could sometimes be more efficient than the full cohort estimator if imputed values are less variable than the true values (this is possible in some range of covariates). Nonetheless, bias would become bigger in such cases so MSE compared to the full cohort analysis would be larger.

Chapter 5

Summary and Future Research

In summary, we considered estimating the causal hazard ratios of MSCMs via inverse probability weighting in full cohort and the case-cohort studies. We established asymptotic theories for estimators that maximize corresponding WPPLs under certain regularity conditions, via martingale and counting process formulation. In addition we proposed new variance estimators which could be more accurate than the robust variance estimators when sample size is small. Framing the problem using standard counting process and martingale theory readily enables practical implementation of the methods using existing survival analysis software. However, implementing MSCM for the case-cohort design was shown to be not fully efficient. Therefore, we explored an imputation method that could lead to more efficient inference in the case-cohort MSCM analysis.

As we framed the problem of estimating the causal hazard ratios of MSCMs using counting processes and martingales, we may consider fitting MSCMs to data from nested case-control studies or in the presence of competing risks as next projects. Also, researchers have found that a main challenge of implementing MSMs in practice is difficulty in estimating inverse probability weights [Cole and Hernán, 2008; Howe et al., 2011; Kang and Schafer, 2007; Lefebvre, Delaney and Platt, 2008; Mortimer et al., 2005]. It has been shown that results of using MSMs via inverse-probability-weighting could be highly sensitive to model misspecification of treatment assignment model, when even number of study visits is moderate. Therefore, doubly-robust-estimation of the causal hazard ratio of MSCMs in the presence of case-cohort sampling, or combining covariate balancing propensity score method proposed by Imai and Ratkovic [2014] in the inverse-probability-weighted estimation of MSCM hazard ratio could be a topic of future work.

BIBLIOGRAPHY

Andersen, P K and R D Gill. 1982. "Cox's Regression Model for Counting Processes: A Large Sample Study." *The Annals of Statistics* 10 (4):1100–1120.

Barlow, W.E., L. Ichikawa, D. Rosner and S. Izumi. 1999. "Analysis of Case-Cohort Designs - A prospective cohort study." *Journal of Clinical Epiemiology* 52 (12):1165–1172.

Barlow, William E. 1994. "Robust Variance Estimation for the Case-Cohort Design." *Biometrics* 50:1064–1072.

Binder, D. A. 1992. "Fitting coxs proportional hazards models from survey data." *Biometrika* 79:139147.

Bodnar, L M, M Davidian, A M Siega-Riz and A A Tsiatis. 2004. "Marginal structural models for analyzing causal effects of time-dependent treatments: An application in perinatal epidemiology." *American Journal of Epidemiology* 159 (10):926–934.

Borgan, Ø, B Langholz, S O Samuelsen, L Goldstein and J Pogoda. 2000. "Exposure Stratified Case-Cohort Designs." *Lifetime Data Analysis* 6:39–58.

Breslow, N E, T Lumley, C M Ballantyne, L E Chambless and M Kulich. 2009*a*. "Improved Horvitz-Thompson estimation of model parameters from two-phase stratified samples: applications in epidemiology." *Stat Biosc* 1 (1):32–49.

Breslow, Norman E., Thomas Lumley, Christie M. Ballantyne, Lloyd E. Chambless and Michal Kulich. 2009*b*. "Using the Whole Cohort in the Analysis of Case-Cohort Data." *American Journal of Epidemiology* 169 (11):1398–1405.

Chen, K and S-H Lo. 1999. "Case-Cohort and Case-Control Analysis With Cox's Model." *Biometrika* 86:755–764.

Chen, Y C, J C Wu, T J Chen and T Wetter. 2001. "Generalized Case-Cohort Sampling." *Journal of the Royal Statistical Society, Ser. B* 63:791–809.

Cole, S R, M G Hudgens, P C Tien, K Anastos, L Kingsley, J S Chmiel and L P Jacobson. 2012. "Marginal structural models for case-cohort study designs to estimate the association of antiretroviral therapy initiation with incident AIDS or death." *American Journal of Epidemiology* 175 (5):381–390. [erratum: 175 (7), 732].

Cole, Stephen R. and Miguel A. Hernán. 2008. "Constructing Inverse Probability Weights for Marginal Structural Models." *American Journal of Epidemiology* 168 (6):656–664.

Cole, Stephen R., Miguel A. Hernán, Kathryn Anastos James M. Robins and, Joan Chmiel, Roger Detels, Carolyn Ervin, Joseph Feldman, Ruth Greenblatt, Lawrence Kingsley, Shenghan Lai, Mary Young, Mardge Cohen and Alvaro Mu noz. 2003. "Effect of Highly Active Antiretroviral Therapy on Time to Acquired Immunodeficiency Syndrome or Death using Marginal Structural Models." *American Journal of Epidemiology* 158 (7):68–694.

Cox, D. R. 1972. "Regression models and life tables (with discussion)." *Journal of the Royal statistical Society B.* 34:187–220.

Cox, D. R. 1975. "Partial Likelihood." *Biometrika* 62 (2):269–276.

Greenland, S and James M Robins. 1986. "Identifiability, exchangeability, and epidemiologic confounding." *International Journal of Epidemiology* 15 (3):412–418.

Heitjan, D F and R J A Little. 1991. "Multiple imputation for the Fatal Accident Reporting System." *Applied Statistics* 40:13–29.

Hernán, Miguel A, Babette Brumback and James M Robins. 2000. "Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men." *Epidemiology* 11 (5):561–570.

Hernán, Miguel A, Babette Brumback and James M Robins. 2001. "Marginal structural models to estimate the joint causal effect of nonrandomized treatments." *Journal of the American Statistical Association* 96 (454):440–448.

Howe, C J, S R Cole, J S Chmiel and A Mu noz. 2011. "Limitation of inverse probability-of-censoring weights in estimating survival in the presence of strong selection bias." *American Journal of Epidemiology* 173 (5):569–577.

Imai, Kosuke and Marc Ratkovic. 2014. "Covariate balancing propensity score." *Journal of Royal Statistical Society Series B* forthcoming.

Kang, J D and J L Schafer. 2007. "Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data (with discussions)." *Statistical Science* 22 (4):523–539.

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

Lee, H, M Hudgens, J Cai and S R Cole. 2013. "Marginal Structural Cox Models with Case-Cohort Sampling." *In Preparation* .

Lefebvre, G, J A C Delaney and R W Platt. 2008. "Impact of mis-specification of the treatment model on estimates from a marginal structural model." *Statistics in Medicine* 27 (18):3629–3642.

Lin, D Y and Z Ying. 1993. "Cox regression with incomplete covariate measurements." *Journal of American Statistical Association* 88 (424):1341–1349.

Moodie, Erica E M, Joseph A C Delaney, Geneviéve Lefebvre and Robert W Platt. 2008. "Missing Confounding Data in Marginal Structural Models: A Comparison of Inverse Probability Weighting and Multiple Imputation." *The International Journal of Biostatistics* 4 (1).

Mortimer, K M, R Neugebauer, M. van der Laan and I B Tager. 2005. "An application of model-fitting procedures for marginal structural models." *American Journal of Epidemiology* 162 (4):382–388.

Neyman, Jerzy Splawa. 1923. "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9." *Statistical Science* 5 (4):465–472.

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

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

Prentice, Ross L. and Steven G. Self. 1983. "Asymptotic Distribution Theory for Cox-Type Regression Models with General Relative Risk Form." *The Annals of Statistics* 11 (3):804–813.

Robins, J. M. 1993. "Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers." *Proceedings of the biopharmaceutical section, American statistical associtaion* pp. 24–33.

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

Robins, James M. 1986. "A new approach to causal inference in mortality studies with a sustained exposure period - Application to control of the healthy worker survivor effect." *Mathematical Modelling* 7 (9-12):1393–1512.

Robins, James M. 1987. "Addendum to "A new approach to causal inference in mortality studies with sustained exposure periods - Application to control of the healthy worker survivor effect"." *Computers and Mathematics with Applications* 14:923–945.

Robins, James M. 1997. "1997 proceedings of the American Statistical Association, Section on Bayesian Statistical Science." *Alexandria, VA: American Statistical Association* .

Robins, James M. 1998. "Correction for non-compliance in equivalence trials." *Statistics in Medicine* 17 (3):269–302.

Robins, James M. 1999. Marginal Structural Models versus Structural Nested Models as Tools for Causal Inference. In *Statistical Models Epidemiology: The Environment and Clinical Trials*, ed. M Elizabeth Halloran and Donald A Berry. NY: Springer-Verlag pp. 95–134.

Robins, James M. and A. Rotnitzky. 1992. Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS Epidemiology - Methodological Issues*, ed. N. P. Jewell, K. Dietz and V. T. Farewell. Boston:Birkhäuser pp. 297–331.

Robins, James M., A. Rotnitzky and Scharfstein D. 1999. Sensitivity Analysis for Selection Bias and Unmeasured Confounding in Missing Data and Causal Inference Models. In *Statistical Models in Epidemiology: The Environment and Clinical Trials*, ed. M.E. Halloran and D. Berry. Vol. 116 NY: Springer-Verlag pp. 1–92.

Robins, James M., D. Blevins, G. Ritter and M. Wulfsohn. 1992. "G-estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia on the survival of AIDS patients." *Epidemiology* 3:319–336.

Robins, James M., Miguel A. Hernán and Babette Brumback. 2000. "Marginal Structural Models and Causal Inference in Epidemiology." *Epidemiology* 11 (5):550–560.

Robins, James M and Naisyin Wang. 2000. "Inference for imputation estimators." *Biometrika* 87(1):113–124.

Rubin, D B. 1974. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology* 66:688–701.

Rubin, D B. 1976. "Inference and missing data (with discussion)." *Biometrika* 63 (3):581–592.

Rubin, D B. 1980. "Discussion of "Randomized analysis of experimental data: the Fisher randomization test" by D J Basu." *Journal of the American Statistical Association* 75:591–593.

Rubin, Donald B. 1987. *Multiple imputation for nonresponse in surveys*. New York:John Wiley & Sons.

Schafer, J L. 1997. *Analysis of Incomplete Multivariate Data*. New York:Chapman & Hall.

Schenker, N and J M G Taylor. 1996. "Partially parametric techniques for multiple imputation." *Computational Statistics and Data Analysis* 22:425–446.

Self, Steven G. and Ross L. Prentice. 1988. "Asymptotic Distribution Theory and Efficiency Results for Case-Cohort Studies." *The Annals of Statistics* 16 (1):64–81.

Therneau, Terry. 2012. *A Package for Survival Analysis in S*. R package version 2.36-12.

Therneau, Terry M and Hongzhe Li. 1999. "Computing the Cox Model for Case Cohort Designs." *Lifetime Data Analysis* 5 (2):99–112.

Tsiatis, Anastasios A. 2006. *Semiparametric Theory and Missing Data*. Springer.

VanderWeele, Tyler J. 2009*a*. "Concerning the Consistency Assumption in Causal Inference." *Epidemiology* 20 (6):880–883.

VanderWeele, Tyler J. 2009*b*. "Marginal Structural Models for the Estimation of Direct and Indirect Effects." *Epidemiology* 20 (1):18–26.

Wang, Naisyin and James M Robins. 1998. "Large-sample theory for parametric multiple imputation procedures." *Biometrika* 85(4):935–948.

Xiao, Yongling, Michal Abrahamowicz and Erica E. M. Moodie. 2010. "Accuracy of Conventional and Marginal Structural Cox Model Estimators: A Simulation Study." *The International Journal of Biostatistics* 6 (2):Article 13.