Causal Inference Approaches for Dealing with Time-dependent Confounding in Longitudinal Studies, with Applications to Multiple Sclerosis Research

by

Mohammad Ehsanul Karim

B.Sc., University of Dhaka, 2004
M.S., University of Dhaka, 2005
M.Sc., The University of British Columbia, Vancouver, 2009

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

The Faculty of Graduate and Postdoctoral Studies

(Statistics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

January 2015

Abstract

Marginal structural Cox models (MSCMs) have gained popularity in analyzing longitudinal data in the presence of 'time-dependent confounding', primarily in the context of HIV/AIDS and related conditions. This thesis is motivated by issues arising in connection with dealing with time-dependent confounding while assessing the effects of beta-interferon drug exposure on disease progression in relapsing-remitting multiple sclerosis (MS) patients in the real-world clinical practice setting. In the context of this chronic, yet fluctuating disease, MSCMs were used to adjust for the time-varying confounders, such as MS relapses, as well as baseline characteristics, through the use of inverse probability weighting (IPW). Using a large cohort of $1,697$ relapsing-remitting MS patients in British Columbia, Canada ($1995-2008$), no strong association between beta-interferon exposure and the hazard of disability progression was found (hazard ratio 1.36, 95% confidence interval $0.95, 1.94$). We also investigated whether it is possible to improve the MSCM weight estimation techniques by using statistical learning methods, such as bagging, boosting and support vector machines. Statistical learning methods require fewer assumptions and have been found to estimate propensity scores with better covariate balance. As propensity scores and IPWs in MSCM are functionally related, we also studied the usefulness of statistical learning methods via a series of simulation studies. The IPWs estimated from the boosting approach were associated with less bias and better coverage compared to the IPWs estimated from the conventional logistic regression approach. Additionally, two alternative approaches, prescription time-distribution matching (PTDM) and the sequential Cox approach, proposed in the literature to deal with immortal time bias and time-dependent confounding respectively, were compared via a series of simulations. The

PTDM approach was found to be not as effective as the Cox model (with treatment considered as a time-dependent exposure) in minimizing immortal time bias. The sequential Cox approach was, however, found to be an effective method to minimize immortal time bias, but not as effective as a MSCM, in the presence of time-dependent confounding. These methods were used to re-analyze the MS dataset to show their applicability. The findings from the simulation studies were also used to guide the data analyses.

Preface

I wrote this dissertation with direction and input from Drs. P. Gustafson, J. Petkau and H. Tremlett. These studies were approved by the University of British Columbia's Clinical Research Ethics board (study number: H08-01544).

Chapter 2 is a version of the pre-copy-editing, author-produced PDF of an article accepted for publication in 'American Journal of Epidemiology' following peer review. The definitive publisher-authenticated version [Karim M. E., Gustafson P., Petkau J., Zhao Y., Shirani A., Kingwell E., Evans C., van der Kop M., Oger J., and Tremlett H. Marginal Structural Cox Models for Estimating the Association Between β-Interferon Exposure and Disease Progression in a Multiple Sclerosis Cohort. American Journal of Epidemiology, 180(2):160-171, 2014, Oxford University Press] is available online at: `http://aje.oxfordjournals.org/cgi/content/abstract/kwu125`. As part of my copyright agreement with Oxford University Press I have retained the right, after publication, to include this article in full or in part in my thesis or dissertation, provided that this is not published commercially. I was the lead investigator, responsible for concept formation, statistical analyses and interpretations of the data, as well as drafting of the manuscript. P. Gustafson, J. Petkau and H. Tremlett were supervising this project and were involved throughout the project in formation of the study concept and design and manuscript composition. H. Tremlett, P. Gustafson, E. Kingwell, J. Petkau, Y. Zhao and M. van der Kop obtained funding, and A. Shirani, C. Evans, E. Kingwell, J. Oger, and H. Tremlett provided administrative, technical, or material support. A. Shirani, E. Kingwell, M. van der Kop, J. Oger, and H. Tremlett were involved in data acquisition and P. Gustafson,

J. Petkau and Y. Zhao contributed in guiding the statistical analyses. For this manuscript, I was responsible for all of the research analysis and writing the initial draft, but all co-authors were involved in the improvement of the manuscript via a number of critical revisions.

I was the lead investigator for the projects described in Chapters 3 and 4. I was responsible for all major areas of concept formation, design of the studies and analyses, as well as the manuscript composition. P. Gustafson, J. Petkau and H. Tremlett were the supervisors on these projects and were involved throughout the project in concept formation and manuscript edits.

Table of Contents

List of Tables

List of Figures

Acknowledgements

It gives me great pleasure to express my sincere gratitude and deepest appreciation to my supervisors: Professors Paul Gustafson and John Petkau. Their mentorship, valuable suggestions regarding my research, financial support, prompt review of my writing and constant inspiration greatly helped to advance my research training and to complete this dissertation. It truly has been an honor and a privilege to work with both of them.

I would like to thank my supervisory committee member, Associate Professor Helen Tremlett (Neurology), for giving me the opportunity to work on the BeAMS (Benefits and Adverse Effects of Beta-interferon for Multiple Sclerosis) project and for including me in her family of impassioned collaborators. I am also grateful for her invaluable research feedback and careful draft revisions. I acknowledge and thank my research collaborators Drs. Yinshan Zhao, Afsaneh Shirani, Elaine Kingwell, Charity Evans, Joel Oger and Mia van der Kop for their support and patience.

Many thanks to my PhD comprehensive committee members, Professors Rollin Brant and Lang Wu, for their excellent feedback. I would like to express my sincere gratitude and appreciation to Professor Erica Moodie, external examiner from McGill University (Department of Epidemiology, Biostatistics and Occupational Health), for her valuable comments. I would like to thank Professors Hubert Wong (Health Care and Epidemiology) and Rollin Brant (Statistics) for serving as the university examiners. I would also like to thank everyone in the Department of Statistics, from faculty to staff and fellow graduate students, for making my PhD program such an enriching and pleasant experience.

I would like to acknowledge the Multiple Sclerosis (MS) Society of Canada

for the PhD Research Studentship as well as travel awards to attend conferences, the endMS Network for travel awards to attend various conferences and summer schools, the University of British Columbia (UBC) for the Ph.D. Tuition Fee Award, Graduate Student Travel Award and Faculty of Science Graduate Award and the Department of Statistics for its Graduate Teaching Assistant Award. I am also grateful for the travel grants from the Pacific Institute for the Mathematical Sciences (PIMS).

Many thanks to the MS patients for their participation in research and the BCMS neurologists who contributed to the study through patient examination and data collection (current members listed here by primary clinic): UBC MS Clinic: A. Traboulsee, MD, FRCPC (UBC Hospital MS Clinic Director and Head of the UBC MS Programs); A-L. Sayao, MD, FRCPC; V. Devonshire, MD, FRCPC; S. Hashimoto, MD, FRCPC (UBC and Victoria MS Clinics); J. Hooge, MD, FRCPC (UBC and Prince George MS Clinics); L. Kastrukoff, MD, FRCPC (UBC and Prince George MS Clinics); J. Oger, MD, FRCPC. Kelowna MS Clinic: D. Adams, MD, FRCPC; D. Craig, MD, FRCPC; S. Meckling, MD, FRCPC. Prince George MS Clinic: L. Daly, MD, FRCPC. Victoria MS Clinic: O. Hrebicek, MD, FRCPC; D. Parton, MD, FRCPC; K Pope, MD, FRCPC. We also thank P. Rieckmann, MD (Sozialstiftung Bamberg Hospital, Germany) for helpful revisions of the original CIHR grant. The views expressed in this dissertation do not necessarily reflect the views of each neurologist acknowledged.

On a personal note, I am eternally grateful to my parents who have provided me with abundance of freedom and opportunities all my life, to my sister and brother for their words of wisdom and to my wife Suborna for her love.

Vancouver, Canada — Mohammad Ehsanul Karim
January 19, 2015

Dedication

To my parents,
and to my wife, Suborna.

Chapter 1

Introduction

In most scientific research, establishing causation is the ultimate goal. Researchers usually view predictive models with a causal interpretation. Without a sense of causality in the researcher's mind, the statistical measures are merely measures of association among various variables under consideration. Simple association resulting from a poorly-designed study may sometimes be misleading or inadequate in assessing the causal relationship between variables. This is especially true in the field of epidemiology, when evaluating disease-exposure relationships from observational data. Finding the cause of a health related outcome is usually the focus. Statistical association is merely an intermediate step in the process. This led researchers to redefine various statistical and epidemiologic concepts in terms of causal mechanisms.

Multiple sclerosis (MS) is a chronic disease that affects the central nervous system, affecting an estimated 2.3 million people worldwide [1]. Beta-interferons (β-IFNs) are the most commonly prescribed immunomodulatory drugs for treating relapsing-onset MS patients. The drugs were primarily licensed or approved for use in MS based on demonstrated, but partial, effects from key short-term clinical trials [2–6]. Further, a number of side effects are associated with the use of these drugs. Since these β-IFN drugs are expensive and patients may be on the drugs for many years, long-term effectiveness of β-IFN is of great interest. However, MS is a life-long disease and appropriately following these patients for such a long time in order to assess drug effectiveness in an 'exposed' versus 'non-exposed' group of individuals is not practical, from either the ethical or cost perspective. This study has access to one of the largest population-based MS databases in the world. Utilizing this retrospective cohort of British Columbia (BC) MS

patients provides the opportunity to investigate long-term effectiveness of β-IFN under the 'real-world' clinical practice setting.

A causal interpretation of a treatment effect estimate obtained from observational data requires additional considerations and assumptions. Conventional statistical analysis tools often fail to produce unbiased estimates in the absence of randomization and the presence of time-varying confounders. In this dissertation, an MS research question motivates the improvement of analysis methodologies for the observational study data from a much broader perspective. The overarching goal of this dissertation is to assess, improve and compare the statistical tools that deal with the time-dependent confounding while estimating the causal effect of a treatment in the context of observational longitudinal drug-effectiveness studies. However, it is important to understand the assumptions behind these causal inference tools. In the following sections, we briefly illustrate the basic framework and the key assumptions that facilitate causal inference.

1.1 A Brief Overview of Causal Inference Frameworks

1.1.1 Potential Outcomes Framework

The ideas of causality date back to 1748 in the work of the philosopher Hume [7]. He defined causality in plain English as follows: "Cause is an event followed by another (effect)", and "Without the first event (cause), the second (effect) would never happen", which formed the foundation for the sufficient and necessary conditions for causality. These intuitive causal definitions (especially the second) were translated into statistical language by Neyman et al. [8] using the 'potential outcome' notion to define 'causal effect' (Neyman's framework).

Table 1.1: An illustration of defining treatment effect in terms of potential outcomes

Subject i	Covariate L	Treatment $(A = 0, 1)$	Outcome $Y_{A=1}$	Outcome $Y_{A=0}$	Causal effect $Y_{A=1} - Y_{A=0}$
1	l_1	Both†	$Y_{A=1,1}$	$Y_{A=0,1}$	$Y_{A=1,1} - Y_{A=0,1}$
2	l_1	Both	$Y_{A=1,2}$	$Y_{A=0,2}$	$Y_{A=1,2} - Y_{A=0,2}$
3	l_1	Both	$Y_{A=1,3}$	$Y_{A=0,3}$	$Y_{A=1,3} - Y_{A=0,3}$
4	l_1	Both	$Y_{A=1,4}$	$Y_{A=0,4}$	$Y_{A=1,4} - Y_{A=0,4}$
5	l_1	Both	$Y_{A=1,5}$	$Y_{A=0,5}$	$Y_{A=1,5} - Y_{A=0,5}$
Conditional summary			$\mathrm{E}(Y_{A=1}\|L = l_1)$	$\mathrm{E}(Y_{A=0}\|L = l_1)$	$\mathrm{E}(Y_{A=1} - Y_{A=0}\|L = l_1)$
6	l_2	Both	$Y_{A=1,6}$	$Y_{A=0,6}$	$Y_{A=1,6} - Y_{A=0,6}$
7	l_2	Both	$Y_{A=1,7}$	$Y_{A=0,7}$	$Y_{A=1,7} - Y_{A=0,7}$
8	l_2	Both	$Y_{A=1,8}$	$Y_{A=0,8}$	$Y_{A=1,8} - Y_{A=0,8}$
9	l_2	Both	$Y_{A=1,9}$	$Y_{A=0,9}$	$Y_{A=1,9} - Y_{A=0,9}$
10	l_2	Both	$Y_{A=1,10}$	$Y_{A=0,10}$	$Y_{A=1,10} - Y_{A=0,10}$
Conditional summary			$\mathrm{E}(Y_{A=1}\|L = l_2)$	$\mathrm{E}(Y_{A=0}\|L = l_2)$	$\mathrm{E}(Y_{A=1} - Y_{A=0}\|L = l_2)$
Marginal summary			$\mathrm{E}(Y_{A=1})$	$\mathrm{E}(Y_{A=0})$	$\mathrm{E}(Y_{A=1}) - \mathrm{E}(Y_{A=0})$

† Both treatments $A = 0$ and $A = 1$ are applied on each of the subjects.

In Neyman's framework, the causal effect is defined as the comparison of potential outcomes Y_A under treatment ($A = 1$) and no treatment ($A = 0$) conditions, i.e., $Y_{A=1}$ versus $Y_{A=0}$ for a given unit i. This comparison can be measured either in the form of a difference (additive scale) or a ratio (multiplicative scale) or some other generalized contrasts such as, simple average, median, hazard or cumulative density function of the potential outcomes. For example, the causal risk ratio can be defined as the ratio (contrast) of the means (function of potential outcome), $\mathrm{E}(Y_{A=1})$ / $\mathrm{E}(Y_{A=0})$. Deviation from the null value, zero for the difference and one for the ratio, would imply that there is a causal effect. In the hypothetical example shown in Table 1.1, each row corresponding to a given subject produces a causal effect. A conditional summary (say, average) causal effect can be calculated for each particular value of the covariate L. A marginal summary (say, average) causal effect can be calculated unconditionally.

Fisher recognized the value of randomization in estimating the treatment effect from an experiment [9, 10]. He introduced Fisher's sharp null hypothesis H_0 that under all treatment assignments ($A = 0, 1$), every unit i would produce the same outcome, i.e., $Y_{A=1,i} = Y_{A=0,i}$, for completely randomized experiments. If this hypothesis H_0 is true, then there is no treatment effect in unit i (Fisher's framework).

In evaluating causation or causal effects, randomized experiments are the best choice. Their strength stems from the principle of randomization or lack of bias towards any covariate levels, as noted by Fisher. However, experimenting with medical treatments is not always feasible. Hence researchers may have to depend on observational studies to estimate the effect of a treatment despite the fact that observational studies are more prone to various kinds of biases. Providing a causal interpretation of an estimate obtained from observational data requires additional assumptions or conditions, under which we could imagine some form of chance mechanism was involved in the process of data collection. To make causal inference using nonrandomized data in a point-treatment situation (treatment is not time-

dependent), Rubin extended the potential outcome notion (Neyman-Rubin framework) in a series of works [11–16]. If $A_i = 0, 1$ is the treatment assignment on unit i, the observed outcome Y_i for unit i can be expressed as $Y_i = A_i Y_{A=1,i} + (1 - A_i) Y_{A=0,i}$.

To estimate a causal effects for subject i, one needs the two outcomes $(Y_{A=1}, Y_{A=0})$. The obvious problem with using this hypothetical definition is that we can only observe a patient under one treatment at any given time, i.e., we can observe either $Y_{A=1}$ or $Y_{A=0}$ in Table 1.1. Crossover randomized experiments may provide a solution to this issue under some conditions, but it is not possible to conduct such experiments for irreversible health outcomes. The need to deal with unobservable quantities (i.e., missingness of half of the outcomes) is considered as the fatal flaw of the potential outcome model [17] or the fundamental problem of causal inference [18].

Even though $Y_{A=1,i} - Y_{A=0,i}$, the causal effect for a particular unit i, cannot be identified due to missing information, the average causal effect, $E(Y_{A=1}) - E(Y_{A=0})$ from the two mutually exclusive groups $A = 1$ and 0, can be estimated by $E(Y|A = 1) - E(Y|A = 0)$ if these two groups are similar in characteristics (as shown in Table 1.2 as an illustration). That is, for a binary outcome, the causal risk ratio defined by the unconditional or marginal expression $P(Y_{A=1} = 1)/P(Y_{A=0} = 1)$ can be estimated from the conditional expression or association measure $P(Y = 1|A = 1)/P(Y = 1|A = 0)$. Here the marginal ratio is estimated based on the idea that both treatments are applied to the whole population. The conditional ratio is estimated based on the idea that the treatment is applied to a part of the population and a mutually exclusive part of that population did not receive the treatment, as shown in Figure 1.1. To ensure these two groups are comparable, certain assumptions need to be made.

Table 1.2: An illustration of defining treatment effect in terms of observed outcomes

Subject i	Covariate L	Treatment $A = 0$ or 1	Outcome $Y_{A=1}$	Outcome $Y_{A=0}$	Causal effect $Y_{A=1} - Y_{A=0}$
1	l_1	1	$Y_{A=1,1}$		
2	l_1	1	$Y_{A=1,2}$		
3	l_1	0		$Y_{A=0,3}$	
4	l_1	0		$Y_{A=0,4}$	
5	l_1	0		$Y_{A=0,5}$	
Conditional summary			$\mathrm{E}(Y\|A=1, L=l_1)$	$\mathrm{E}(Y\|A=0, L=l_1)$	$\mathrm{E}(Y\|A=1, L=l_1) - \mathrm{E}(Y\|A=0, L=l_1)$
6	l_2	0		$Y_{A=0,6}$	
7	l_2	0		$Y_{A=0,7}$	
8	l_2	1	$Y_{A=1,8}$		
9	l_2	1	$Y_{A=1,9}$		
10	l_2	1	$Y_{A=1,10}$		
Conditional summary			$\mathrm{E}(Y\|A=1, L=l_2)$	$\mathrm{E}(Y\|A=0, L=l_2)$	$\mathrm{E}(Y\|A=1, L=l_2) - \mathrm{E}(Y\|A=0, L=l_2)$
Marginal summary			$\mathrm{E}(Y\|A=1)$	$\mathrm{E}(Y\|A=0)$	$\mathrm{E}(Y\|A=1) - \mathrm{E}(Y\|A=0)$

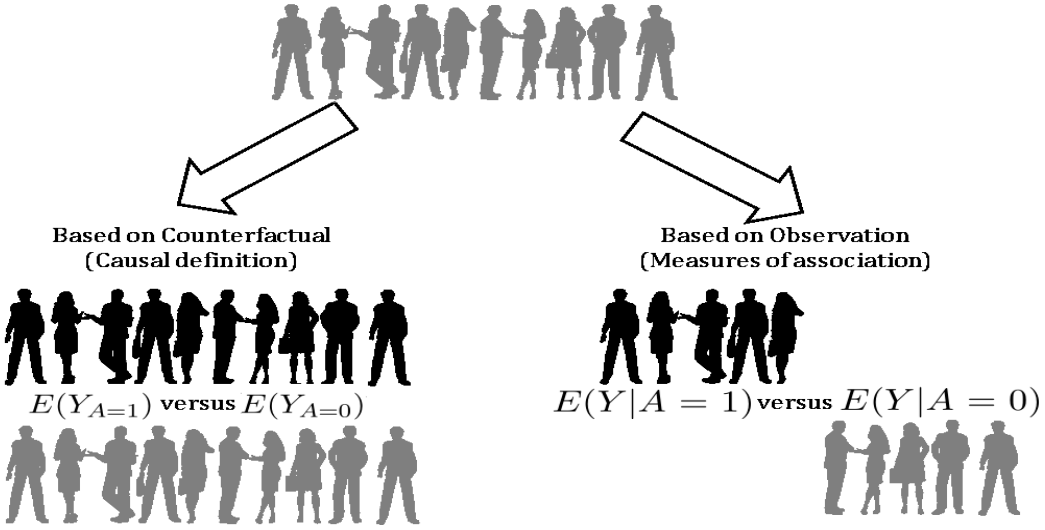


Figure 1.1: An illustration of defining treatment effect in terms of potential outcomes and observations

1.1.2 Assumptions

To be able to estimate the causal effect, the assumption of ignorability [19] or unconfoundedness [20] is required. This assumption states that $(Y_{A=1}, Y_{A=0}) \perp A$, which means the treatment assignment A and the potential outcomes are not associated. A common source of confusion would be to interpret this assumption as observed A and observed Y to be independent, which is not true if there is a treatment effect. To make a connection between this assumption and randomization, we can say that ignorability ensures the treatment assignment A and the potential outcomes $Y_{A=1}, Y_{A=0}$ are unassociated, whereas randomization assures that the treatment assignment A is unassociated with any variables, not just the joint distribution of the potential outcomes. Let us define the 'sufficient set of covariates' as a set of covariates that contains the complete information about the exposure-outcome association. Then within the levels of this set of covariates, the association measure of the exposure-outcome relationship is unconfounded. If L is a sufficient set of covariates, then $(Y_{A=1}, Y_{A=0}) \perp A|L$ is called condi-

tional ignorability within levels of L, i.e., the causal effect can be estimated within the strata matched by L or corresponding to the levels of L.

Some authors view ignorability as the combination of the assumptions of exchangeability and positivity [21, ch.3]. "Exchangeability" is denoted by $Y_a \perp A$ or $P(Y_a|A = 1) = P(Y_a|A = 0)$. If this condition is satisfied, reversal or exchange of treatment status of all the patients does not change the magnitude or direction of the treatment effect. Under this assumption, excluding the effect of treatment, sub-groups under consideration are assumed to be equivalent in all respects. Therefore, the risk in the exposed group would be the same as the risk in the unexposed group had patients in the exposed group not received the treatment or exposure. Exchangeability is also known as exogeneity [22] which is related to the concept of an exogenous variable. A variable that does not receive any causal input from any other variable in the system is called an exogenous variable. In the econometric literature, such a concept is useful in detecting confounding or deviation from causal and associational measures as well. "Positivity" is denoted by $P(A = a|L = l) > 0 \ \forall a$. This assumption requires the existence of at least one individual in each stratum of L in each of the exposure groups so that a comparable pair of subjects exists in each stratum of L in the target population, i.e., positive probability of getting assigned to each of the treatment levels in each strata.

Another assumption required to make causal inference from nonrandomized data is popularly known as the 'stable unit treatment value assumption' (SUTVA) [23]. This assumption states that the potential outcome observation Y on one unit i should be unaffected by the particular assignment of treatments $A = 0, 1$ to the other units $j \neq i$. In other words, treatment choice for one unit does not affect the outcome of any other unit. This is similar to the assumption of no interaction between units [24], but more general because SUTVA also assumes that there are no different versions of treatment, i.e., treatment does not vary in effectiveness from unit to unit. The latter part of the assumption is also known as consistency [25, 26] de-

noted by $Y_a = Y \ \forall a$.

In the literature, this consistency assumption is variously known as 'no versions of treatment' [19] or 'treatment-variation irrelevance' [27] or 'well defined interventions' [28]. That means, treatment ($A = 1$) needs to be one particular dose of a particular drug and no treatment ($A = 0$) should also mean no other treatment. For the subjects with same covariate history, the observed outcome due to this treatment $A = 1$ should be the same. As an extreme example, if it is reasonable to assume that various drugs would have the same effect (counterfactual outcome) on a given patient, then all these drugs could be listed as treatment $A = 1$ with no versions (since the outcome does not vary). In experiments, keeping the same version of the treatment is usually expected due to adherence to a precise protocol. However, in observational studies, it may be achievable if the prescription is unambiguously one particular dose of a given drug. Otherwise, if multiple treatments $A_1, A_2, \ldots, A_R$ (with possibly different effects on outcome) are being prescribed to treat similar patients, then consistency is violated due to existence of R versions of the drug. The causal effect of such a 'treatment with multiple versions' may not have any practical value. A possible remedy for such a situation is restriction, say, restricting analysis to A_i if multiple versions are separately documented in the data. Even if there is only one version of a treatment in a given observational study, the effect may vary due to noncompliance (e.g., taking pills irregularly) and thus result in a violation of the consistency assumption. The resulting average treatment effect will depend on the distribution of patients receiving various versions of the treatment in the sample. If this differs from the population distribution of patients who received various versions of the treatment in a real world setting, then the corresponding result may be biased.

The three assumptions, exchangeability, positivity and consistency, are often referred to as the 'identifiability conditions' [29]. Making a causal statement or interpretation requires that the observational study emulates a randomized experiment where all the covariates are equally distributed be-

tween the treated and the untreated groups. However, such balance between the treated and untreated groups is not usually seen in observational studies. Under these identifiability conditions, observational studies can be viewed as conditionally randomized experiments, i.e., treatment assignment can be assumed random conditional on measured covariates. Then it is possible to make causal inferences with the hope that the untestable assumptions are approximately true. Unfortunately, without subject-area knowledge or use of additional information to justify the assumptions, such inferences can not be validated.

Rosenbaum and Rubin proposed propensity score methodology based on this framework [20, 30–32]. They defined the propensity score $p = P(A = 1|L)$ as the conditional probability of receiving treatment given the measured background variables L. They also extended the assumption of conditional ignorability. The propensity score is often used for matching when there are multiple or possibly high-dimensional attributes of L.

1.1.3 Models for Longitudinal Settings

For longitudinal data structures, more sophisticated methodology needs to be adopted to account for the complexity in the data. Robins showed that, under some conditions, time-varying treatments will not have a causal interpretation even if the usual identifiability conditions hold [33, 34]. He extended the point-treatment theory further to apply in longitudinal settings where treatment may be time-varying (multiple time point treatments). Figure 1.2 (b) illustrates this extension for two time points. Here, treatment (A) assignment or choice may change in the second time point ($t = 1$) compared to that of the first time point ($t = 0$), whereas for the point-treatment situation (Figure 1.2 (a)) treatment A is assigned only once. Under this framework, potential outcomes are often denoted by the term 'counterfactuals' [35] (or 'possible worlds' [25]), while others have reservations about this nomenclature [16].

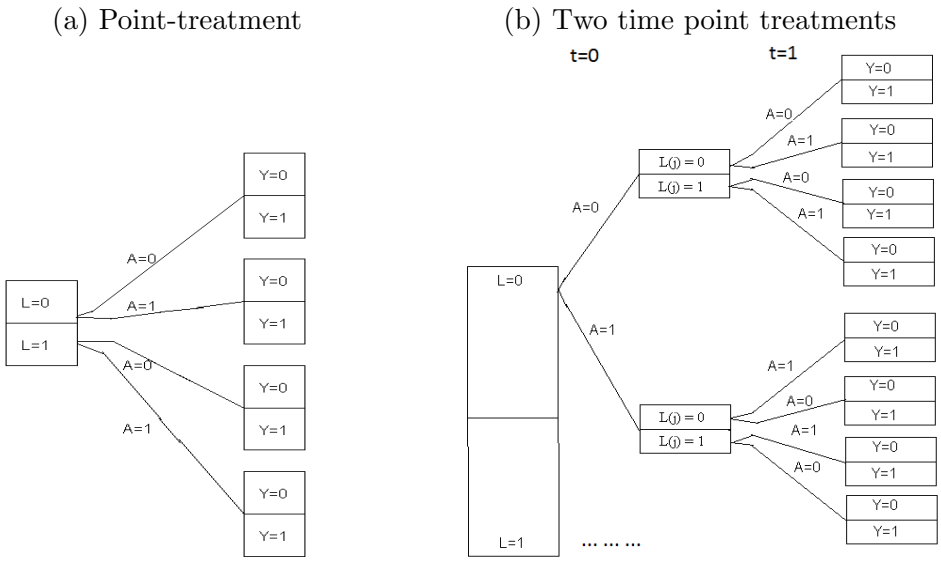


Figure 1.2: Illustration of point-treatment and two time point treatments situation

A methodology to estimate causal parameters from longitudinal counterfactual models was proposed under the so-called sequential randomization assumption (SRA) [33]. Let us define $\bar{A}(t) = (A(1), A(2), \ldots, A(t))$ as the treatment history up to time t from baseline, $\bar{a}(t)$ as the observed treatment history up to time t from baseline and $Y_{\bar{a}(t)}$ or shortly, $Y_{\bar{a}}$ as the corresponding vector of counterfactuals. Then $Y_{\bar{a}} \perp A(t)|\bar{A}(t-1)$ is known as SRA, which is an extension of the ignorability condition of the point-treatment theory: $(Y_{A=1}, Y_{A=0}) \perp A$. This assumption basically states that the joint distribution of the counterfactuals $Y_{\bar{a}}$ is independent of the current treatment given the treatment history $\bar{A}(t-1)$. Similarly, the conditional ignorability assumption $(Y_{A=1}, Y_{A=0}) \perp A|L$ of the point-treatment theory can be extended to $Y_{\bar{a}} \perp A(t)|\bar{A}(t-1), \bar{L}(t)$ where $\bar{L}(t)$ is the time-dependent covariate history. This assumption basically states that the counterfactuals $Y_{\bar{a}}$ are independent of treatment $A(t)$ at time t, conditional on the treatment and measured covariate history ($\bar{A}(t-1)$ and $\bar{L}(t)$, with the assumption that covariates of a given time t are measured before treatment assignment) up to time t.

Estimating the Causal Effect of Treatment in the Presence of Time-dependent Confounders

Models based on this SRA assumption, such as marginal structural models (MSM) [36–42], provide a way to identify the causal effect of time-dependent treatment from longitudinal data. To provide a causal interpretation of the coefficient associated with $A(t)$ on the outcome in a regression model, we need to remove any confounding effect of time-dependent covariates $\bar{L}(t)$ up to time t. One way this could happen is if $A(t)$ is an exogeneous variable (the covariate history $\bar{L}$ is not causing treatment), i.e., if $\bar{L}(t) \perp \bar{A}(t)$ or $\bar{L}(t) \perp A(t) | \bar{A}(t-1)$. In that case, the association measure will estimate the causal effect. Now, let $P\big(a(t)|\bar{a}(t-1)\big)$ denote the probability of subjects who choose treatment $A(t) = a(t)$ among the subjects with treatment history $\bar{A}(t-1) = \bar{a}(t-1)$. Similarly, let $P\big(a(t)|\bar{a}(t-1), \bar{l}(t)\big)$ denote the probability of subjects who choose treatment $A(t) = a(t)$ among the subjects with treatment history $\bar{A}(t-1) = \bar{a}(t-1)$ and covariate history $\bar{L}(t) = \bar{l}(t)$. Then

$$\omega(t) \quad = \quad \prod_{j=0}^{t} \frac{P\big(A(j)|\bar{A}(j-1), \bar{L}(j)\big)}{P\big(A(j)|\bar{A}(j-1)\big)}$$

indicates the degree to which the treatment process $A(t)$ deviates from exogeneity; "exogeneity" can be expressed as $\omega(t) \equiv 1$. For the subjects who make the predictable choices, in the event that the covariate history $\bar{L}(t)$ is a strong predictor of treatment choice $A(t)$ at time t, $\omega(t)$ will be larger, but if the covariate history does not cause or predict treatment choice $A(t)$ at time t (treatment assignment truly being an exogenous variable), then $\omega(t)$ will be 1. Suppose the exposure-outcome association is being estimated from a regression model. Then, even if $A(t)$ is exogenous, weighting the regression model by the weight $\omega^{-1}(t)$ (generally known as the inverse probability of treatment weights, discussed in §1.1.4) will provide an estimate of the causal effect. That is, an effect measure obtained from the weighted regression of

the mean of observed outcome Y on the treatment history $\bar{A}(t)$ will have a causal interpretation. Such a marginal model for the response (that averages over covariates instead of conditioning on the covariates) is popularly known as an MSM.

To explain the product operator in the formula for $\omega(t)$, let us consider the point treatment situation. Then,

$$\omega \;=\; \frac{P\big(A|L\big)}{P\big(A\big)}$$

indicates the degree to which the treatment assignment A deviates from exogeneity. Similarly, for a two-time point treatment situation ($t = 0, 1$), let $A(0)$ denote the binary treatment status at time 0 and $A(1)$ denote the binary treatment status at time 1. Also let $L(0)$ denote the binary covariate status at time 0 (baseline) and $L(1)$ denote the binary covariate status at time 1. Here $L(1)$ can possibly be affected by $A(0)$, but not the other way around since we are not dealing with retrocausality. With the convention that $A(-1) = 0$:

$$\begin{aligned}
\omega(1) &= \frac{P\big(A(1), A(0)|L(1), L(0)\big)}{P\big(A(1), A(0)\big)} \\
&= \frac{P\big(A(1)|A(0), L(1), L(0)\big)P\big(A(0)|L(1), L(0)\big)}{P\big(A(1)|A(0)\big)P\big(A(0)\big)} \\
&= \frac{P\big(A(1)|A(0), L(1), L(0)\big)}{P\big(A(1)|A(0)\big)} \times \frac{P\big(A(0)|L(0)\big)}{P\big(A(0)\big)} \\
&= \prod_{j=0}^{1} \frac{P\big(A(j)|\bar{A}(j-1), \bar{L}(j)\big)}{P\big(A(j)|\bar{A}(j-1)\big)},
\end{aligned}$$

will now indicate the degree to which the treatment assignments $\bar{A}(1) = \big(A(0), A(1)\big)$ deviate from exogeneity. This expression can be generalized to t-time points and this leads to the formula above for $\omega(t)$.

1.1.4 Inverse Probability of Treatment Weights

Weights are usually known while analyzing the data from a randomized clinical trial. While dealing with observational studies, weights need to be estimated from the observed data. To derive the weights, treatment history is assumed to be predicted by the covariate history so that an appropriate adjustment can be made. The weight $\omega^{-1}(t)$ is known as inverse probability of treatment weight (IPTW). Note that this IPTW is a generalization of the propensity score $p = P(A|L)$ and is functionally related, i.e., IPTW $= A/p + (1 - A)/(1 - p)$ [43, 44]. For the point treatment context, such weighting is equivalent to adding $\omega^{-1} - 1$ copies of corresponding subjects which will constitute a pseudo-population, where the unconfounded effect estimate can be obtained by the use of simple measures of association (say, risk ratio or risk difference). This estimate is equivalent to that of standardization methods [45, 46] where the causal risk ratio can be estimated by the standardized risk ratio for the total population,

$$\frac{P(Y_{A=1}=1)}{P(Y_{A=0}=1)} = \frac{\sum_l P(Y=1|A=1,L=l)P(L=l)}{\sum_l P(Y=1|A=0,L=l)P(L=l)}.$$

This quantity estimates the risk for 'all' the subjects in the population that are treated versus 'all' the subjects in the population that are untreated, computed from the observed quantities of Y, A and L. This is a ratio of weighted averages of the stratum L-specific risks that offers a causal interpretation (complete exposure versus complete nonexposure) under the conditional ignorability assumption. MSM generalizes the standardization methods in longitudinal settings [47–49].

MSM, therefore, treats the unobserved counterfactual potential outcomes as missing values and tries to impute stratum specific average values or, equivalently re-weights the observed values to adjust for those missing in order to rebuild the pseudo dataset (as shown in Table 1.3 as an illustration; compare to Table 1.2).

The mention of the MSM approach in the literature dates back to the 1990s, but use of this approach increased after the publication of two landmark papers in 2000 [42, 50]. These papers outlined a simple method to implement this approach using off-the-shelf software routines of logistic regression to estimate the weights. As IPTW estimation is central to the MSM approach, assessment of the assumptions of the corresponding logistic regression fits are crucial, even though rarely seen in the MSM literature [51, 52]. Alternative IPTW modelling strategies, such as statistical learning methods, require fewer assumptions and may be worth investigating.

Analysts and researchers are increasingly using the MSM approach to deal with time-dependent confounding. In observational settings, many are skeptical about weight-based estimators in general. The debate is not about the foundation of the IPTW estimators [39], but mostly about proper implementation techniques. The major criticism of this method stems from the fact that the assumptions of the MSM approach are restrictive and mostly untestable from a given dataset. There exists substantial literature about various implementation techniques [53–57].

Marginal Structural Cox Model: With event-time outcomes in the longitudinal context, censoring is usually another feature of these studies. MSM models are tailored for survival data by use of inverse probability of censoring weights (IPCW) in addition to IPTW, and these models are popularly known as marginal structural Cox models (MSCMs) [50, 58]. MSCM models will be further discussed in Chapter 2.

Table 1.3: Outcomes after stratum specific averages are imputed in the cells where the outcomes are missing

Subject i	Covariate L	Treatment† $A = 0$ or 1	Outcome $Y_{A=1}$	Outcome $Y_{A=0}$	Causal effect $Y_{A=1} - Y_{A=0}$
1	l_1	1	$Y_{A=1,1}$	$\mathrm{E}(Y\|A=0, L=l_1)$	
2	l_1	1	$Y_{A=1,2}$	$\mathrm{E}(Y\|A=0, L=l_1)$	
3	l_1	0	$\mathrm{E}(Y\|A=1, L=l_1)$	$Y_{A=0,3}$	
4	l_1	0	$\mathrm{E}(Y\|A=1, L=l_1)$	$Y_{A=0,4}$	
5	l_1	0	$\mathrm{E}(Y\|A=1, L=l_1)$	$Y_{A=0,5}$	
Conditional summary			$\mathrm{E}(Y\|A=1, L=l_1)^{\dagger\dagger}$	$\mathrm{E}(Y\|A=0, L=l_1)$	$\mathrm{E}(Y\|A=1, L=l_1) - \mathrm{E}(Y\|A=0, L=l_1)$
6	l_2	0	$\mathrm{E}(Y\|A=1, L=l_2)$	$Y_{A=0,6}$	
7	l_2	0	$\mathrm{E}(Y\|A=1, L=l_2)$	$Y_{A=0,7}$	
8	l_2	1	$Y_{A=1,8}$	$\mathrm{E}(Y\|A=0, L=l_2)$	
9	l_2	1	$Y_{A=1,9}$	$\mathrm{E}(Y\|A=0, L=l_2)$	
10	l_2	1	$Y_{A=1,10}$	$\mathrm{E}(Y\|A=0, L=l_2)$	
Conditional summary			$\mathrm{E}(Y\|A=1, L=l_2)$	$\mathrm{E}(Y\|A=0, L=l_2)$	$\mathrm{E}(Y\|A=1, L=l_2) - \mathrm{E}(Y\|A=0, L=l_2)$
Marginal summary			$\mathrm{E}(Y_{full}\|A=1)^{\dagger\dagger\dagger}$	$\mathrm{E}(Y_{full}\|A=0)$	$\mathrm{E}(Y_{full}\|A=1) - \mathrm{E}(Y_{full}\|A=0)$

‡ This simplistic illustrative example is for the point-treatment situation. However, the MSM is applicable for a more generalized longitudinal setting where treatment A values may change more than once over time.

‡† $\mathrm{E}(Y|A=1, L=.)$ values are computed from stratum L specific observed $Y_{A=.,i}$ values. If none of the values in a particular stratum $L=x$ are observed, then this method fails.

††† $\mathrm{E}(Y_{full}|A=.)$ is computed after the missing value imputation by stratum L specific averages.

1.1.5 Role of Causal Diagrams

The existence of common causes of treatment and outcome, i.e., the idea of confounding, is an important issue in any epidemiologic study since this distorts the relationship between the treatment exposure variable and the outcome variable. Even if exposure and outcome variables are not causally associated, due to relationship with this common cause, association measures may report associated exposure-outcome variables. Similarly, conditioning on common effects also may distort the exposure-outcome variable relationship and such bias is popularly known as collider bias.

For a longitudinal setting, when treatment status is a time-dependent variable and other time-dependent variables are affected by the previous treatment status, the relationship of the treatment variable with other variables can be complicated. Such complications may lead to additional bias which can be hard to detect and control. A set of graphical tools called 'directed acyclic graphs' (DAGs) or causal diagrams were developed [59–61] to define causality [62], confounder [63], selection bias [64], effect-modification [65], non-collapsibility [66] and over-adjustment [67].

These graphs are more intuitive in explaining the causal concepts compared to other methods or definitions even in complicated structures. DAGs do not deal with cyclic variables which can cause themselves, nor are suitable for assessing claims of retrocausality [68]. Using DAGs, it is possible to explain epidemiologic concepts in an intuitive way [35], an unified structural approach for detecting confounding, mediation and selection bias can be outlined [64] and a simple 6-step approach can be used [62] to identify confounding using the backdoor criterion [60].

In DAG notation, if C is a variable that is causing exposure variable E and outcome variable D, then even if E and D are not associated with each other, a statistical association measure may find these two variables to be associated, which does not imply or reflect the true causal relationship [63].

A path between two variables corresponds to statistical association and a block in the path means statistical association between these two variables is nullified. The backdoor criterion is a graphical tool to determine the existence of an unblocked non-directional path between exposure and outcome as shown in Figure 1.3. In the exposure E - outcome D association, a set of variables C fulfills this criterion, if (i) no variable in C is caused by exposure E and (ii) C blocks every path between exposure E and outcome D that is causing exposure choice E. This backdoor criterion helps researchers identify whether there is confounding present in a situation, whether such confounding can be eliminated, and what particular variables are necessary to control for the confounding.

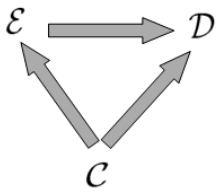
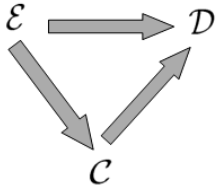
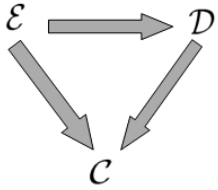
C is common cause or confounding	C is intermediate value or Mediator	C is common effect or collider
		
Unblocked non-directional path exists between E and D, i.e., E ← C → D	Unblocked directional path exists between E and D, i.e., E → C → D and E → D	Blocked non-directional path exists between E and D, i.e., E → C ← D

Figure 1.3: Relationships among exposure E, outcome variable D and a covariate C in a directed acyclic graph

When a confounder is affected by the previous exposure status, it is known as an intermediate variable. In a regression adjustment, to get a valid assessment of the exposure effect, Cox [24] suggested not to control for an intermediate variable. But this suggestion was not based on any proof or simulation. Using DAG theory, it was later shown why standard methods of estimation of treatment effects in longitudinal studies fail to produce unbiased estimates in the presence of a time-dependent risk-factor

that is also a predictor of subsequent exposure [67].

1.1.6 Time-dependent Confounders

If a confounder C is affected by the previous treatment exposure, then it is also acting as an intermediate variable between treatment E and outcome D. As C is in the causal pathway between the current treatment and future outcome, it is associated with both of them (E and D). If C is impacted by the previous treatment exposure and subsequently influences the current treatment decision, it is known as a time-dependent confounder. MSM and MSCM approaches are useful tools to deal with time-dependent confounding [64]. Throughout this thesis, we define a covariate as a "time-dependent confounder" [50, 69] if it

1. is itself affected by the previous treatment exposure and
2. predicts the future treatment decision and future outcome conditional on the past treatment exposure.

1.2 Models to Estimate the Causal Effect

1.2.1 In the Presence of Time-dependent Confounders

MSCMs are useful tools to estimate the causal effect of treatment in the presence of time-dependent confounders in the longitudinal context with event-time outcomes. This is especially true when the time-dependent confounders also act as mediators between the exposure-outcome association. However, MSCMs may require strong and untestable assumptions. Alternative methods such as structural nested models [38, 70, 71], the sequential stratification approach [72] and the sequential Cox approach [73] can be used to deal with time-dependent confounders. But those methods can be very computationally intensive compared to standard statistical tools. Among these, the sequential Cox approach is relatively new, and deserves more attention due to its simplicity.

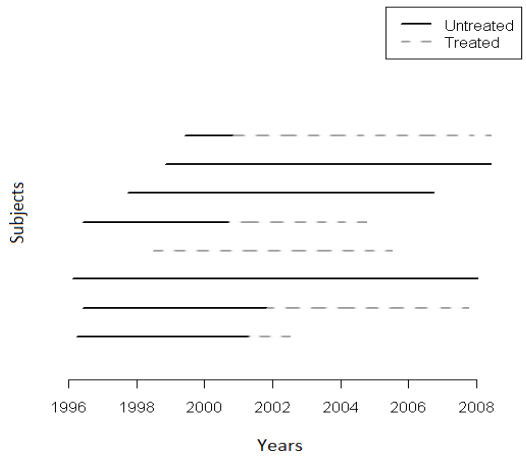


Figure 1.4: An illustration of immortal time, i.e., a delay or wait period that may exist before a subject begins to receive a treatment in an observational drug effectiveness study

1.2.2 In the Absence of a Time-dependent Confounder

A recurrent theme in this research is to find causal effects from observational data in the presence of time-dependent exposure. So far we have dealt with the complicated situation when time-dependent confounders are present. In simpler settings, when time-dependent covariates do not interact or influence future treatment, estimating the effect of time-dependent treatment is less troublesome. Under the assumptions of conditional exchangeability, consistency, correct model specification, and positivity, hazard ratios estimated from a time-dependent Cox proportional hazards model [74] will have causal interpretation [75].

1.2.3 In the Presence of Immortal Time Bias

In some observational studies, after entering into the study or reaching eligibility, there might be a wait-period before receiving treatment. A treated patient with such a wait period contributes to both treated and untreated time. The waiting time of the patients who 'survived' until treatment initia-

tion needs to be properly accounted for in the analysis. As shown in Figure 1.4, there may be patients of three kinds: (a) those who were treated from the beginning till the end of the study, (b) those who were untreated from the beginning till the end of the study, and (c) those who begin the study as an untreated patient, but after some waiting time switch onto treatment. If this waiting time is classified as exposed to treatment instead of unexposed, it offers an artificially enhanced survival advantage for the treated subjects and this phenomenon is sometimes referred to as 'immortal time bias' [76–80]. Note that the immortal time bias can occur with or without time-dependent confounding.

Time-dependent exposure modelling is one way to adjust for this immortal time bias, i.e., for survival outcomes, the time-dependent Cox proportional hazards model [74] is a suggested solution. This approach achieves the best statistical efficiency compared to the alternatives [81, ch.33]. Instead of comparing the treated versus untreated groups, this approach compares time under treatment to time not under treatment. A time-distribution matching approach was suggested [82], which offers a way to use the usual Cox model for the treatment group comparison [83]. This approach assigns new baselines to achieve balance in the follow-up time distribution of exposed and unexposed. This suggested approach is cited frequently in the recent biomedical literature [84–88]. However, it is currently unknown how well this method works in a general setting compared to a time-dependent Cox model. One of the proposed directions of this research is to assess the performance of this approach by means of simulations and theoretical calculations. We will also assess the suitability of using the sequential Cox approach instead of MSCM when a time-dependent confounder is present. We will further discuss these approaches in Chapter 4.

1.3 Organization of the Dissertation

So far we have portrayed the assumptions and the key concepts of the causal inference framework in a very general way. This framework allows us to identify and control for the time-dependent confounders. We consider three general problems related to adjustment of analyses for the possible influences of time-dependent confounders in this dissertation.

We will describe the motivating MS research problem that inspired this work in Chapter 2. MSCMs allow adjustment for time-varying confounders, as well as baseline characteristics. Most of the MSCM analysis performed in the published literature are specific to HIV/AIDS. MS is a chronic disease with features of its own. Different subjects may come to medical attention (e.g., an MS clinic) at different times. Therefore, subjects included in an observational MS study may have different baselines or cohort entry start times, different drug initiation times and may use different treatments or may switch treatments over time depending on their health conditions (e.g., relapse frequency and disease course). A carryover effect of the previous treatment may exist even after drug discontinuation. β-IFN has been found to be effective in some short-term MS clinical trials (3-5 years). The effect of this treatment on longer-term outcomes such as irreversible disability is of great interest. In this study, one of the main objectives was to assess the association between β-IFN drug exposure and disease progression in relapsing-remitting MS patients in the 'real-world' clinical practice setting. In the presence of time-varying confounders, such as MS relapses, MSCM can be a valuable tool to analyze longitudinal observational survival data. In this chapter, we set out to assess the suitability of MSCMs to analyze data from a large cohort of relapsing-remitting MS patients in BC, Canada (1995-2008).

Our data analyses and previous literature indicate that the properties of the inverse probability weights (IPWs) can influence the estimated effects from MSCM and their accuracy. Logistic regressions are generally

used to model the IPWs. Statistical learning algorithms such as bagging, support vector machines, and boosting have proved to be useful in estimating propensity scores with better covariate balance. As propensity scores and IPWs are functionally related, whether the lessons learnt from propensity scores can be translated and generalized to IPW estimation is of great interest. In Chapter 3, we will assess the performance of these proposed methods via simulated survival data that mimics a context in which both treatment status and a confounder are time-dependent. These statistical learning approaches are also applied to estimate IPWs to investigate the impact of beta-interferon treatment in delaying disability progression in the British Columbia Multiple Sclerosis (BCMS) cohort.

Prescription time-distribution matching is an approach proposed in the literature to avoid a time-dependent Cox analysis. In longitudinal survival studies, in the presence of time-dependent confounding, MSCMs are usually used to deal with such confounding. The sequential Cox approach is suggested as an alternative approach. Both the prescription time-distribution matching and the sequential Cox approaches make the interpretation of the results much more accessible to a wider audience. In Chapter 4, we assess the suitability of both of these approaches for analyzing data in the absence and presence of time-dependent confounding. These methods are also utilized to investigate the impact of beta-interferon treatment in delaying disability progression in subjects from the BCMS database. Finally, Chapter 5 briefly summarizes this research, and suggests possible directions for future research.

Chapter 2

Marginal Structural Cox Models for Estimating the Effect of Beta-interferon Exposure in Delaying Disease Progression in a Multiple Sclerosis Cohort

2.1 Introduction

Multiple sclerosis (MS) is a disease associated with damage to the myelin and nerve fibers in the brain and spinal cord. It is a life-long disease, typically manifesting in early adulthood, affecting an estimated 2 to 2.5 million people worldwide [89]. A relapsing-remitting course is the most common presenting MS phenotype; these patients can experience periods of acute worsening, known as an attack or relapse, followed by relapse-free periods with partial or full recovery. Disability may gradually worsen over time, ultimately becoming irreversible. As evident from various clinical trials, immunomodulatory drugs, such as beta-interferon (β-IFN) may reduce the risk of an MS relapse and increase the duration of relapse-free periods over the short-term [2–6]. However, their impact on longer-term outcomes such as irreversible disability is unclear.

There is a real need to determine whether the β-IFNs positively influence the MS disease course over the long-term, particularly in the 'real-world' clinical practice setting. Observational studies are the most pragmatic means of addressing this need. However, findings from recent observational studies have been contradictory with respect to the impact of β-IFN [90–92]. Possible explanations for these inconsistencies include: selection bias, informative censoring, immortal time bias, and inappropriate use of analytic tools [93, 94]. Hence the association between β-IFN and the progression of disability in clinical practice remains undetermined.

Recently researchers assessed the association of β-IFN with the time to irreversible disability outcomes among relapsing-remitting MS patients treated in the real world clinical practice setting of British Columbia, Canada, using a Cox model with time-dependent treatment exposure after adjusting for a number of important baseline confounders [92]. They were also able to compare β-IFN treated patients with two separate control cohorts - a 'historical' cohort (patients who first became β-IFN eligible prior to the approval of β-IFN in Canada in 1995) and a 'contemporary' cohort (patients who first became β-IFN eligible after the approval of β-IFN, but remained unexposed to β-IFN.). While this approach represented a considerable improvement over previous studies [95], concern remained about the potential for indication bias when the contemporary control cohort was considered [92]. Despite adjustment for a number of baseline characteristics, there were also concerns raised about the inability to adjust for subsequent (post-baseline) treatment decisions [92, 96–99]. Furthermore, since disease activity, such as relapses can drive decision-making with respect to starting or stopping β-IFN treatment [100], and might also be associated with the outcome [101], relapses could be considered a potential time-dependent confounder. Simply incorporating such confounders as covariates in a time-dependent Cox model may be inadequate to adjust for selection bias and confounding [50].

Marginal structural Cox models (MSCMs) allow estimation of the causal effects of treatment exposure on survival responses (e.g., time to disability)

in the presence of time-dependent confounding, selection bias and informative censoring [50, 58]. These models depend on model-based estimates of the inverse probability of the observed treatment and censoring status of each patient to achieve causal interpretation of the findings. Simulation studies with short-term follow-up have repeatedly shown that MSCMs are advantageous in terms of obtaining consistent estimates of the effect of time varying treatment exposures [56, 57, 102, 103]. When studying MS, a chronic disease, extended observational periods are needed, which may contribute to the construction of highly variable weights [104], and subsequently may lead to an inefficient estimate of the causal effect. Furthermore, how robust these models are when follow-up lengths differ for individual patients, as is the case in clinical practice, is largely unknown. To assess and address these practical challenges, we explored the use of different weighting approaches in MSCMs to estimate the causal effect of β-IFN on the time to irreversible disability in a cohort of relapsing-remitting MS patients from British Columbia, Canada.

2.2 Materials and Methods

2.2.1 Study Population and Measurements

This cohort study included data that were collected prospectively from MS patients who were registered at a British Columbia (BC) MS clinic and who were eligible to receive β-IFN (all preparations of β-IFN were considered as one therapeutic class). In Canada, the first β-IFN was licensed in July 1995. Therefore, patients who became eligible for β-IFN treatment for the first time between July 1995 and December 2004 were included (only the contemporary control cohort was considered). Broad eligibility criteria for receiving β-IFN treatment were adapted from the BC government's reimbursement scheme, i.e., adults ($\geq$ 18 years old) who had a diagnosis of definite MS with a relapsing-onset course and were able to walk (Expanded Disability Status Scale or EDSS $\leq$ 6.5). The first MS clinic visit at which a patient met the β-IFN eligibility criteria was considered the pa-

tient's baseline date (time = 0). The end of follow-up was December 2008. The study was approved by the University of British Columbia's Clinical Research Ethics board.

The study outcome (irreversible disability progression) was based on the EDSS [105], a standardized rating system to measure neurological impairment and disability, which ranges from 0 (indicating no disability) to 10 (death from MS). The EDSS has been widely used to describe a patient's clinical status, to quantify disability progression and to evaluate treatment response in intervention studies. Our outcome was time to reaching sustained EDSS 6. An EDSS score of 6 indicates that the patient requires intermittent or unilateral constant assistance (cane, crutch or brace) to walk about 100 meters with or without resting. Since it is possible to move back and forth along the EDSS scale, sustained EDSS 6 (i.e., confirmed after at least 150 days, with all subsequent scores being at least EDSS 6 or greater) was adopted in this study as an indicator of irreversible disability progression [92, 106, 107].

Since a patient's β-IFN exposure status might change during follow-up, this was considered as a time-dependent variable. β-IFN exposure was defined as 'any vs. none' on a monthly basis. This could be considered an improvement on the previous study design [92] in which only one treatment initiation and one termination date was considered for each treated patient. Potential confounders included: age at baseline, sex, disease duration at baseline, EDSS score at baseline and relapses.

The relapse variable was selected to be included in the model as a time-varying factor for the following reasons. Firstly, relapses may be associated with the outcome (disability progression). Studies have shown that early relapses may have a significant impact on later disability progression, even though the strength of this association may diminish with time [107]. Secondly, the β-IFNs have been shown to reduce relapse rates [2–6]; therefore, a patient's relapse status may be affected by prior β-IFN treatment. Thirdly,

the presence (or absence) of relapses might influence treatment decisions, i.e. determine whether to start or stop a β-IFN. Finally, the risk of a relapse is not constant over time; it typically decreases as the patient's disease duration and age increases [108]. Therefore, only considering those relapses that occurred prior to a patient's baseline date may be insufficient. Instead, we considered the cumulative number of relapses in the last two years (hereafter 'cumulative relapses') as a time-dependent confounder.

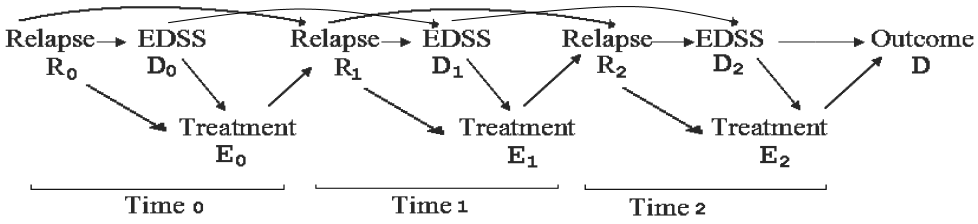


Figure 2.1: Representation of the hypothesized causal relationships in the treatment of MS with three time points $j = 0, 1, 2$.

The cumulative number of relapses could be an intermediate variable between treatment exposure and disability progression; a simplified version of this hypothesized causal relationship is outlined in Figure 2.1. In this Figure, E_j denotes the binary β-IFN exposure variable that is measured immediately after the time-dependent confounder R_j, cumulative relapse and D_j, disability progression index, i.e., EDSS score of the j-th time period. The time-dependent confounder R_j at time j is affected by prior treatment E_{j-1}. According to the causal diagram, R_0 imposes confounding for the E_0-D relationship (as relapse frequency may dictate the subsequent treatment choice and residual disability left by frequent relapses may accumulate over time leading to irreversible disability), but R_1 is an intermediate variable for the same relationship [35, 109] (as the prior β-IFN treatment may reduce relapse frequency which may allow more time to recover from residual disability left by past relapses and may contribute to slower progression of disability over time). A more detailed discussion of rationale can be found in Appendix §A.1. We also examined whether cumulative relapses were an

important predictor of subsequent treatment choices.

2.2.2 Statistical Methods

Conventional Cox model. We defined the model notations as follows: if patient i was followed from the time of β-IFN eligibility ($t = 0$) to time T_i with treatment exposure at time t described by A_{it} (1 = under treatment, 0 = not under treatment), then a_{it} was the realization of A_{it}; $\bar{a}_{it} = (a_{i1}, a_{i2}, \ldots, a_{it})$ described the observed treatment status up to time t. The patient's baseline covariates were recorded in the vector L_{i0} consisting of baseline EDSS score, disease duration, age and sex. If $\lambda_i(t|L_{i0})$ was the hazard of reaching sustained EDSS 6 at time t for patient i with baseline covariates L_{i0}, one way to model such data was with the time-dependent Cox proportional hazards model:

$$\lambda_i(t|L_{i0}) = \lambda_0(t) \exp\big(\beta_1 A_{it} + \beta_2 L_{i0}\big), \tag{2.1}$$

where $\lambda_0(t)$ was the unspecified baseline hazard, β_2 was the vector of log hazard ratios (HRs) for the baseline covariates and β_1 was the log HR of the current β-IFN status (A_{it}). Adding cumulative relapse (L_{it}) as a covariate in this model may have failed to adjust for this time-dependent confounder (discussed in detail in Appendix §A.2). Hence, the MSCM approach [42, 50] was applied instead.

Marginal Structural Cox model (MSCM). Within a counterfactual framework, in the pseudo-population, MSCMs enabled the conceptual comparison of the hazard functions for those who never received β-IFN (complete non-exposure during follow-up) with those who received β-IFN continuously (complete exposure). To accomplish this, the partial likelihood function of the Cox model (or its approximations; see Appendix §A.3) was modified such that the contribution of patient i to the risk set at time t was weighted by the inverse probability of treatment and censoring (IPTC) weight w_i to remove the possible confounding effects of both time-varying and baseline

confounders [50].

Weighting schemes. The stabilized version of the IPT weight for patient i at time t was given by:

$$sw_{it}^{T} = \prod_{j=0}^{t} \frac{pr(A_{ij} = a_{ij} | \bar{A}_{i,j-1} = \bar{a}_{i,j-1}, L_{i0} = l_{i0})}{pr(A_{ij} = a_{ij} | \bar{A}_{i,j-1} = \bar{a}_{i,j-1}, L_{i0} = l_{i0}, \bar{L}_{ij} = \bar{l}_{ij})}, \quad (2.2)$$

where $\bar{A}_{ij} = \bar{a}_{ij}$ and $\bar{L}_{ij} = \bar{l}_{ij}$ were the observed treatment history and time-varying confounder history respectively from baseline to time j. The stabilized IPT weights were inversely related to a function of the time-varying confounder cumulative relapse, since this variable appeared only in the denominator of the weights, whereas the baseline covariates were included in both the numerator and the denominator, as shown in equation (2.2). The weights sw_{it}^{T} down-weighted the person-time contributions when cumulative relapse were a strong predictor of the treatment status in the subsequent time periods, after controlling for the baseline covariates. Assuming that the denominators of the weight models were correctly specified, these weights created a pseudo-population in which cumulative relapses no longer predicted the subsequent β-IFN treatment status [41]. The β-IFN treatment effect in this pseudo-population would be the same as in the original target population [22].

Generally, when the numerator in equation (2.2) is replaced by 1, these weights become the unstabilized IPT weights, w_{it}^{T} [50]. The unstabilized weights simultaneously controls for time-varying and baseline covariates. Unlike MSCMs using stabilized versions of the weights, MSCM analyses using unstabilized versions of the weights do not need further adjustment for the baseline covariates [22]. Use of the unstabilized weights also yields consistent causal estimates but these estimates are associated with substantial variability [41].

Consistent estimation of β_1 from censored data can be achieved by in-

corporating IPC weights in the analysis [110]. Using similar logic to that leading to the IPT weights for uncensored patients, the stabilized version of the IPC weight for patient i at time t is obtained as:

$$sw_{it}^{C} = \prod_{j=0}^{t} \frac{pr(C_{ij}=0|\bar{C}_{i,j-1}=0, \bar{A}_{i,j-1}=\bar{a}_{i,j-1}, L_{i0}=l_{i0})}{pr(C_{ij}=0|\bar{C}_{i,j-1}=0, \bar{A}_{i,j-1}=\bar{a}_{i,j-1}, L_{i0}=l_{i0}, \bar{L}_{i,j-1}=\bar{l}_{i,j-1})}, \tag{2.3}$$

where C_{ij} denoted the binary censoring status taking the value of 1 if the i-th patient was censored in the j-th time and 0 otherwise and $\bar{C}_{ij} = \bar{c}_{ij}$ was the observed censoring history up to time j. The overall stabilized IPTC weights sw_{it} are obtained by multiplying sw_{it}^{T} by sw_{it}^{C} [22, 42].

Since the weights were unknown, they were estimated from the data. Logistic regression models were applied to estimate the conditional probabilities appearing in equations (2.2) - (2.3) (see Appendix §A.4).

The normalized version of the IPTC weights were calculated where each weight was normalized by the mean weight of the corresponding risk set [57]:

$$w_{it}^{(n)} = \frac{w_{it}N_t}{\sum_i Y_{it}w_{it}}, \qquad sw_{it}^{(n)} = \frac{sw_{it}N_t}{\sum_i Y_{it}sw_{it}}, \tag{2.4}$$

where Y_{it} indicated whether patient i belonged to the risk set at time t and $N_t = \sum_i^N Y_{it}$ was the total number of patients in the risk set at time t. We critically assessed the performance of all of the above mentioned weighting schemes.

To take within-subject correlation [111] into account, robust SEs are usually evaluated, which may be asymptotically conservative [42, 112]. Therefore, the 95% CIs for the causal estimate based on 500 nonparametric bootstrap samples were calculated [73, 113, 114].

IPTC weighted survival estimates. IPTC weight adjusted Kaplan-Meier survival curves did not require assumptions related to parametric survival or the Cox model. We used unstabilized IPTC weights (w or $w^{(n)}$) to adjust the survival curves. This had the added advantage of yielding marginal estimates that provided direct causal interpretations without first requiring fit of the MSCM model [115]; hence, constructing such curves served as a sensitivity analysis. However, these weights can be highly variable compared to $sw^{(n)}$ and the adjusted survival curves are prone to distortion in the presence of extreme weights. Truncation of extreme weights was applied as one ad-hoc solution to assuage the problem of extreme weights [55].

Sample code and practical guidance on implementing the weights in such direct and approximate MSCM approaches via various `R` [116] packages are included in Appendix §A.5.

2.3 Results

Of 1,697 patients included in the study, 1,297 patients were female (76%). The mean age at baseline was 39.7 years (SD = 9.7), the mean disease duration from symptom onset was 7 years (SD = 7.7) and the median EDSS score was 2 (IQR = 1).

The mean follow-up time was 4 years (IQR = 6.0 − 1.7 = 4.3), and the maximum was 12.7 years. In total there were 6,890 person-years of follow-up and 2,530 person-years of β-IFN exposure. In all, 829 patients remained untreated during follow-up. Patients at risk of reaching the outcome at the beginning of each year are shown in Figure 2.2. Overall, 138 patients reached the outcome of sustained EDSS 6. Further description of the data is provided in the Appendix §A.6.

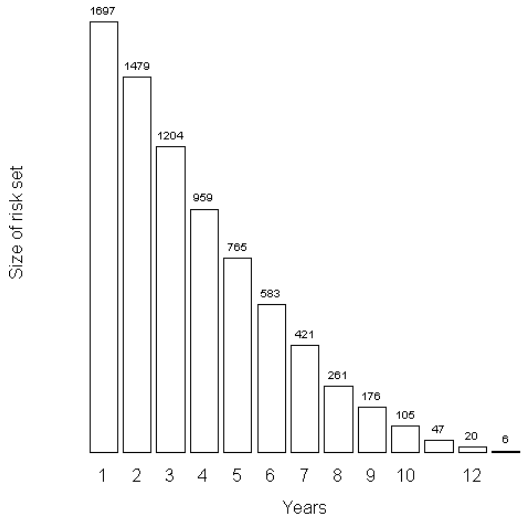


Figure 2.2: Number of patients at risk of reaching sustained EDSS 6 during the first month of each follow-up year after baseline. Failure to continue to the next risk set results from either censoring or reaching sustained EDSS 6. Analyses were performed by month, but the plot is drawn by year for simplicity.

2.3.1 Time-dependent Weights

We found the cumulative relapse variable to be a good predictor of subsequent treatment choices as evidenced by the significance in the model for the IPT weights (two-sided $P < 0.001$; see Appendix-Table A.1) and also for the IPC weights (two-sided $P = 0.03$).

The IPTC weights varied not only from patient to patient, but also by time. As the number of patients at risk decreased monotonically over time, the variation of the IPTC weights increased with follow-up time. As seen in Figure 2.3, in addition to such increasing variability, a clear upward trend over time was evident in the unstabilized weights w. The means at successive time points were much closer to one after stabilization (sw). However, an upward trend of the mean weights was still apparent as follow-up progressed. As expected, this trend was eliminated when the stabilized weights were normalized ($sw^{(n)}$). When the unstabilized weights were normalized

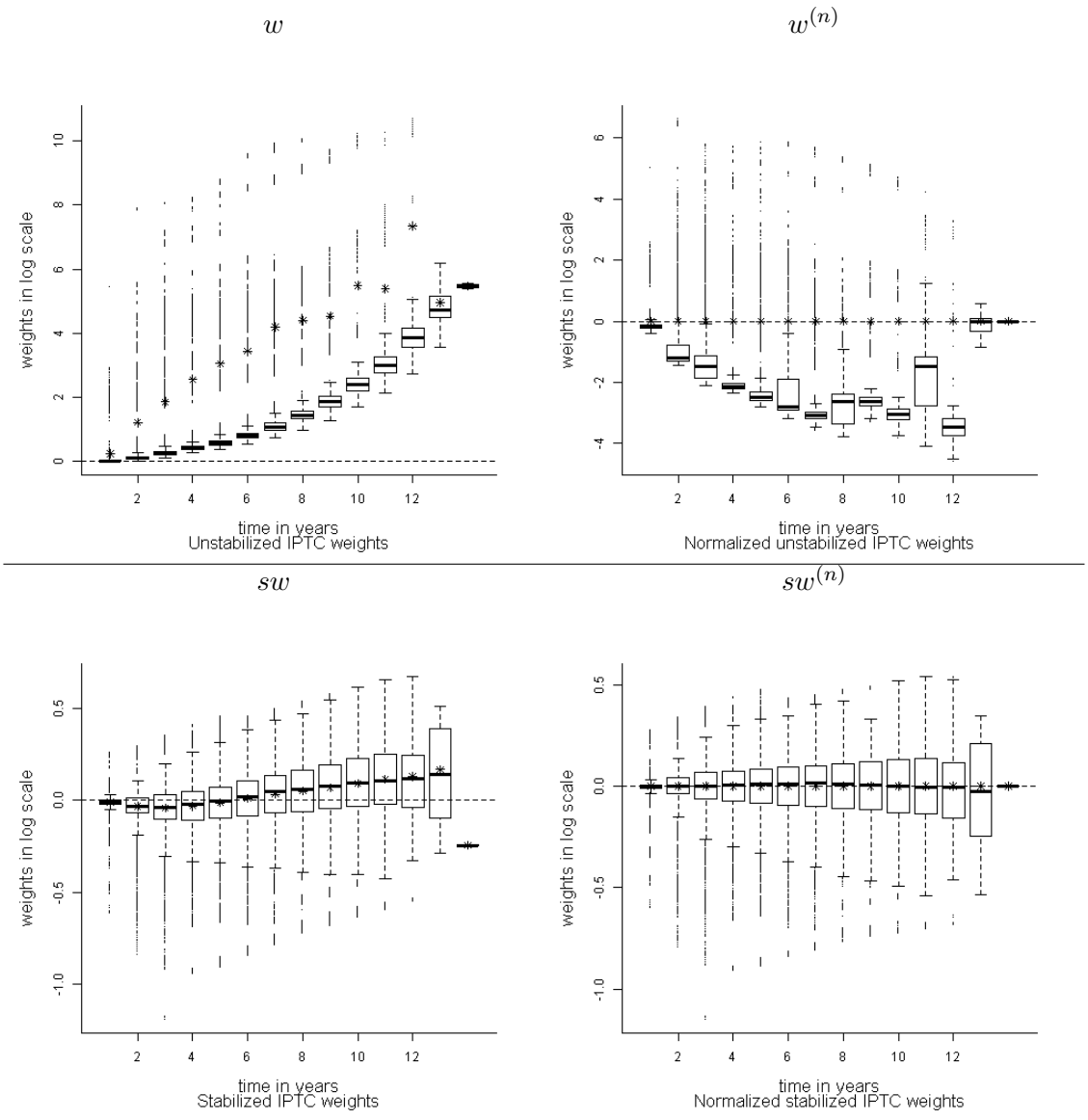


Figure 2.3: Distribution of various IPTC weighting schemes for each year of follow-up (instead of month for better visual display). The means are indicated by $*$ in each boxplot. Note that the plots do not have identical scales on the vertical axes.

($w^{(n)}$), even though the mean weight at each time point was one, the distri-

butions of the weights were highly variable and skewed.

The mean and SD of the unstabilized, unnormalized weights (w) were much larger than those of the other weights (Table 2.1), and the resulting causal effect estimate was further removed from null, with a much wider confidence interval (CI). Normalization resulted in a mean weight of one and a markedly reduced SD. Stabilization of the weights had an even greater impact on reducing the SE of the causal estimate.

A smaller range is an indication of well-behaved weights [55] that generally leads to a smaller CI for the effect estimate. In terms of this desirable property, $sw^{(n)}$ behaved better than the other schemes: these weights had a smaller range. This supported the use of $sw^{(n)}$ in this application. Also, a necessary condition for correct model specification is that the mean of the stabilized weights is one [49, 55], ideally at each time period rather than just overall. Although $sw^{(n)}$ depend on the same specifications of the treatment and censoring models as in sw, we observe that there was no tendency for the mean to deviate from one even after long follow-up (see Figure 2.3).

Table 2.1: Different versions of the IPTC weights and the corresponding causal effect of β-IFN on the hazard of reaching sustained EDSS 6 for MS patients from BC (1995-2008).

Scheme*	Stabilized	Normalized	Estimated Weights		Causal Estimates	
			Mean (log-SD)	Min-Max	HR	95 % CI
w	No	No	28.17 (6.44)	1 - 43,985.38	1.54	0.09, 26.38§
$w^{(n)}$	No	Yes	1 (2.45)	0.01 - 753.47	1.36	0.18, 10.40§
sw	Yes	No	0.99 (-2.12)	0.30 - 1.95	1.36	0.95, 1.94§
$sw^{(n)}$	Yes	Yes	1 (-2.18)	0.32 - 1.71	1.36	0.95, 1.94§#

log-SD, log of standard deviation; Min, minimum; Max, maximum; CI, confidence interval.

* The IPT numerator model included the baseline covariates EDSS, age, disease duration and sex, treatment status at previous time interval and restricted cubic spline [117] of the follow-up month number. The denominator model included the covariates considered in the numerator model and the time-dependent covariate cumulative relapses for last two years, as well as its interaction with treatment status at the previous time interval. The same model specifications were used to generate the IPC weights. With the stabilized versions of the weights, the hazard ratio model of the MSCM must include adjustment for the baseline covariates, but this is not necessary with the unstabilized versions of the weights.

§ Based on 500 nonparametric bootstrap samples with patients as sampling units.

The CI of the causal effect estimate obtained using $sw^{(n)}$ was the smallest, although equal to that obtained using sw when displayed to 2 decimal places.

2.3.2 The Causal Effect of β-IFN

Since the $sw^{(n)}$ had better properties, we relied on the corresponding MSCM estimates (see Table 2.2). The estimated HR failed to suggest a beneficial effect of the treatment, and the evidence of an association between the current beta-IFN exposure and the hazard of reaching sustained EDSS 6 was inconclusive.

Table 2.2: The marginal structural Cox model (MSCM) fit with the normalized stabilized IPTC weights $sw^{(n)}$ for time to sustained EDSS 6 to estimate the causal effect of β-IFN treatment for multiple sclerosis (MS) patients from British Columbia, Canada (1995-2008). The model was also adjusted for the baseline covariates EDSS, age, disease duration and sex.

Covariate	Estimate*	HR †	95% bootstrap CI ‡
β-IFN	0.31	1.36	0.95 - 1.94
EDSS	0.54	1.72	1.54 - 1.92§
Disease duration#	-0.19	0.83	0.66 - 1.05
Age#	0.28	1.32	1.08 - 1.62§
Sex¶	-0.22	0.80	0.55 - 1.17

HR, Hazard ratio; CI, confidence interval; EDSS, expanded disability status scale

* Estimated log HR

† HR, indicating the instantaneous risk of reaching sustained and confirmed EDSS 6

‡ Based on 500 nonparametric bootstrap samples.

§ 95% CI that does not include 1.

\# Expressed in decades.

¶ Reference level: Male

To verify the results, we also obtained the estimates from several approaches that approximate the MSCM (see Table 2.3). All the estimates from the models based on $sw^{(n)}$ were consistent. The conclusion concerning the causal effect of β-IFN on time to sustained EDSS 6 did not change with the modelling choices.

Table 2.3: Estimates of effect of β-IFN treatment on time to sustained EDSS 6 for MS patients from British Columbia, Canada (1995-2008) using different analytical approaches.

Model	Adjustment	Measures of effect	95% CI
Cox	Unweighted†	1.29§	0.91 - 1.82‡
	Weighted by $sw^{(n)}$	1.36§	0.95 - 1.94¶
Pooled logistic	Unweighted†	1.29$^{\#}$	0.91 - 1.82‡
	Weighted by $sw^{(n)}$	1.36$^{\#}$	0.96 - 1.95¶
Poisson	Weighted by $sw^{(n)}$	1.36$^{\#}$	0.96 - 1.95¶
C-log-log	Weighted by $sw^{(n)}$	1.37$^{\#}$	0.96 - 1.95¶

† Based on time-dependent β-IFN treatment exposure status and covariates measured at baseline: EDSS, age, disease duration, sex. This estimate does not have a causal interpretation; and is shown for comparison purposes.

‡ 95% CIs calculated based on robust SEs.

¶ 95% CIs obtained from 500 nonparametric bootstrap samples.

§ HR is the measure of effect obtained from a Cox model.

$^{\#}$ HR from Cox model was approximated by the odds ratio (OR) of the pooled logistic model [118, 119] (see Appendix §A.3) or, under the infrequent event assumption, by the standardized mortality ratio (SMR) from Poisson regression or by the OR from complementary log-log regression respectively. The weighted Cox [50, 57] model was approximated by weighted versions of these models. Software specifications of these analyses are reported in Appendix §A.5.

In a complementary analysis we considered longitudinal EDSS values as an additional time-varying confounder, instead of treating EDSS as a baseline covariate (see Table 2.4). Additionally, the impact of weight trimming [120] was evaluated to assess the sensitivity of the findings to the positivity assumption (see Appendix §A.7.1). The analysis was also repeated after selecting patients *via* more restricted eligibility criteria (see Appendix §A.7.2).

Further analyses were conducted to check the impact of the cumulative exposure to β-IFN over the last two years on the same outcome (see Appendix §A.7.3). We also assessed the impact of including cumulative relapses in the last year, rather than the last two years (see Appendix §A.7.4). None of these sensitivity analyses resulted in statistical evidence for an effect of treatment.

Table 2.4: Sensitivity analysis to assess the impact of EDSS as an additional time-varying confounder: The MSCM fit with the normalized stabilized IPTC weights $sw^{(n)}$ for time to sustained EDSS 6 to estimate the causal association between β-IFN treatment for patients with relapsing-onset MS, British Columbia, Canada (1995-2008)

Covariate	Estimate	HR †	95% CI ‡
β-IFN*	0.12	1.13	0.76 - 1.68
Disease duration#	-0.02	0.98	0.82 - 1.22
Age#	0.32	1.37	1.10 - 1.63 §
Sex¶	-0.36	0.70	0.47 - 1.02

HR, Hazard ratio; CI, confidence interval; EDSS, expanded disability status scale.

* The model was adjusted for cumulative relapse and EDSS as time-varying confounders and baseline covariates age, disease duration and sex. Considering EDSS as a time-varying confounder rather than a baseline covariate in the analysis does not contradict the causal diagram (Figure 2.1). All missing EDSS values were imputed via the last-value-carried-forward approach.

‡ Based on 500 nonparametric bootstrap sample estimates.

§ 95% CI that does not include 1.

Expressed in decades.

¶ Reference level: Male.

2.3.3 IPTC Weighting for Estimation of Survival Curves

We plotted IPTC weight $w^{(n)}$ adjusted Kaplan-Meier survival curves. However, the large drops in the survival plot in Figure 2.4 (b) were driven by only a few large weights. Therefore, we investigated the sensitivity of these adjusted Kaplan-Meier curves after progressively truncating $w^{(n)}$.

Table 2.5: The impact of truncation of the $w^{(n)}$ on the estimated causal effect of β-IFN on reaching sustained EDSS 6 for MS patients from British Columbia, Canada (1995-2008).

Truncation	Estimated weights		Treatment effect estimate		
percentiles[‡]	Mean (log-SD)	Min-Max	HR	SE[†]	95 % CI[†]
None	1 (2.45)	0.01 - 753.47	1.36	1.41	0.18 - 10.4
(5, 95)	0.31 (-1.24)	0.04 - 0.93	1.11	0.32	0.64 - 1.95
(10, 90)	0.3 (-1.29)	0.05 - 0.83	1.13	0.31	0.66 - 1.95
(25, 75)	0.21 (-2.2)	0.09 - 0.35	1.17	0.25	0.77 - 1.76
Median [§]	0.19 (-Inf)	0.19 - 0.19	1.29	0.23	0.91 - 1.82

log-SD, logarithmic transformation of standard deviation; Min, minimum; Max, maximum; CI, confidence interval, HR, Hazard ratio.

[†] Based on 500 nonparametric bootstrap samples.

[‡] Truncation means the extreme weights (determined by the selected percentile range) are replaced by the nearest percentile weight value.

[§] Weighting by the median of the weights gives the same estimate and CI as obtained from the simple baseline covariate adjusted Cox model (see Table 2.3).

As can be seen from Figure 2.4 (c), truncation of the 5% smallest and largest of the $w^{(n)}$ freed the curve from the excess influence of a few extreme weights (following the convention in [55]). In this application, the adjusted survival curves did not change dramatically with greater truncation (see Figure 2.4 (d)-(f)). Note that some studies do not truncate the smaller weights as truncating such weights generally does not lead to substantial changes in the effect estimates [121].

The magnitude of variability in the weights $w^{(n)}$ affected not only the adjusted survival curve, but also the CI for the causal effect obtained from the $w^{(n)}$ weighted MSCM. The CI (95% bootstrap CI $0.18 - 10.4$; see Table 2.1) was wider than that obtained with $sw^{(n)}$, even though the two causal effect estimates were the same (HR 1.36). As before, truncation of the extreme

weights was examined as another sensitivity analysis to increase the precision of the causal estimate [55]. Truncating the 5% smallest and largest of the $w^{(n)}$ had a substantial impact in this application: the CI shrunk to 0.64 - 1.95 (see Table 2.5). Table 2.5 shows that despite improving the precision of the estimate of the β-IFN treatment effects, this ad-hoc truncation approach did not alter the conclusion concerning the causal effect of β-IFN on time to sustained EDSS 6.

2.4 Discussion

By adapting an IPTC weight based MSCM approach in order to explore the impact of β-IFN on MS disability progression in the 'real-world' clinical practice setting, we did not find a significant association between β-IFN exposure and disability progression.

The possibility that cumulative number of (prior) relapses may represent a time-dependent confounder lying on the causal path of β-IFN and disability progression led us to propose this MSCM approach [122]. From the analysis, it was evident that the cumulative relapse count in the previous two years was an important factor in the weight models. This highlights the importance of controlling for this type of time-dependent confounder and justifies the additional complexity of the MSCM approach. Further advantages of using such models included the ability to adjust for potential informative censoring.

Even though an extended follow-up period is essential to adequately capture the potential effects of treatment on disease progression for chronic diseases such as MS, the duration of follow-up may vary considerably from patient to patient in observational settings. This feature of the data poses considerable challenges while applying the MSCM approach, especially when trying to obtain suitable weights. Over time, treatment exposure as well as other patient characteristics (e.g., age, disease duration, occurrence of re-

lapses) change, further contributing to the complexity of the study design. To account for these changes, the weights at a given time point need to be obtained by combining weights for each previous time period in a multiplicative manner. For patients with an extended follow-up, this may cause estimated weights for later periods to increase dramatically and the overall mean weights for these periods to deviate far from one. Also, as follow-up progresses, the decreasing number of patients 'at risk' may further contribute to high variability in the weights. Deviation from a mean of one (for the stabilized versions of the weights) at any time point is an indication of possible weight model misspecification, whereas highly variable weights may decrease the precision of the causal effect estimate [55]. Furthermore, in the presence of very large weights, near nonpositivity may result in a biased and imprecise estimate of the treatment effect [110, 123]. The large variability in follow-up periods of the MS patients prompted us to investigate the choice of appropriate weighting schemes for MSCMs.

Stabilization of the weights is generally advocated to decrease weight variation, and hence increase the precision of MSCM estimates [50]. However, the performance of these weights in the chronic disease context has not been well-studied. Here we noted that as the observation period increased, so did the upward trend of the weights. Even though the normalized weights ($sw^{(n)}$) generally possess desirable properties irrespective of the follow-up period length [57], we could find no application of these newly proposed weights to the chronic disease context in the published literature. Application of $sw^{(n)}$ completely eradicated the upward trends, in turn producing an effect estimate with slightly higher precision compared to the other weighting schemes, suggesting the potential utility of such weights in studies with longer-follow-up.

Adjusting for the time-dependent confounder 'cumulative relapses' *via* IPTC weighting ($sw^{(n)}$) moved the estimated effect of β-IFN treatment (HR 1.36) away from the null compared to the unweighted Cox model (HR 1.29). The corresponding 95% bootstrap CIs from the MSCM analyses were wider

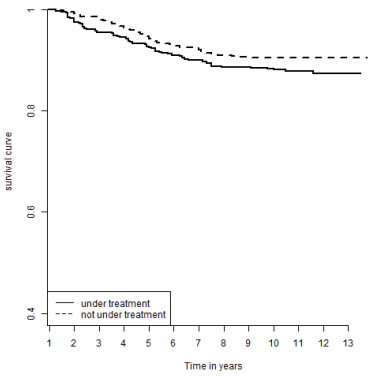
than the 95% robust CIs of the unweighted Cox model, appropriately reflecting more uncertainty as a consequence of using estimated weights. The effect estimates were consistent for the various approximations of MSCM models that we considered; none provided evidence of a significant benefit of β-IFN exposure on disease progression.

We also explored the application of other weighting schemes, such as, normalized unstabilized weights $w^{(n)}$. Using these weights, we constructed IPTC weighted adjusted survival curves. These curves serve as sensitivity analyses as their results are independent of fitting any MSCM. However, unstable survival estimates were produced as a result of a few very large weights. Moreover, as expected, use of the unstabilized weights resulted in larger SEs of the MSCM estimators than those obtained from the stabilized versions. The ad-hoc strategy of truncating extreme weights produced more stable survival curves and increased the precision of the MSCM estimate based on $w^{(n)}$. Truncation at the 5% level was enough to produce quite stable and smooth survival curves, as well as $w^{(n)}$ based MSCM estimated SEs comparable to those based on $sw^{(n)}$.

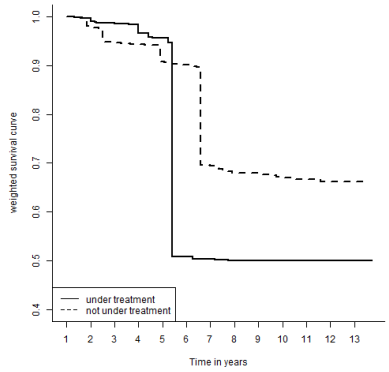
This study has limitations. In order to make a causal interpretation from the MSCM results, identifiability conditions such as positivity, consistency, conditional exchangeability and correct MSCM model specification are required [55], most of which are untestable assumptions. In addition, assuming the IPTC weight models were correctly specified, truncation of the most extreme weights might have introduced bias into the β-IFN effect estimates, reflecting the fundamental 'bias-variance trade-off' [55]. Our assessment of disease progression was based on the EDSS which has recognized limitations [124] and may not be able to tease out differences due to natural aging versus MS disability. Also, one could consider EDSS as another time-dependent confounder. Our sensitivity analysis implementing this (based on imputed missing EDSS values) substantially moved the estimated HR towards the null (HR 1.13; 95% CI $0.76, 1.68$), considerably weakening the suggestion from the main analysis of an adverse effect of treat-

ment. The near-significant point estimate (HR 1.36, 95% CI $0.95 - 1.94$) from the main results may therefore be due to residual confounding. Although we considered important confounders, residual confounding due to unmeasured covariates (both baseline and time-dependent) is still possible. Potential limitations of the observational study design to assess the association between β-IFN and disease progression are similar to those described elsewhere [92].

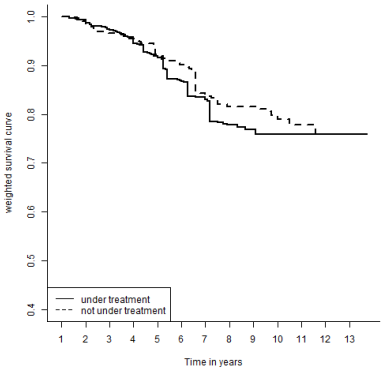
In summary, use of the Cox model alone may be inadequate to handle the challenges of analyzing longitudinal observational data. The use of such tools may partly explain the seemingly inconsistent findings regarding the effectiveness of β-IFN on disability progression in the 'real-world' MS clinical practice setting [91, 92]. Here, we carefully implemented the MSCM analysis to adjust for potential indication bias and related changes in patient characteristics which might influence the subsequent treatment decisions. Our analyses did not find any association between β-IFN exposure and the time to developing sustained EDSS 6 over the follow-up. Even though different approaches were used here, our conclusions are consistent with those of other studies [92, 125]. Furthermore, none of the sensitivity analyses in the current study changed our conclusion regarding the causal effect of β-IFN on disease progression. The consistency of the results from all of our MSCM analyses strengthen our confidence in the findings. The methods implemented here are adaptable to chronic disease settings beyond MS.



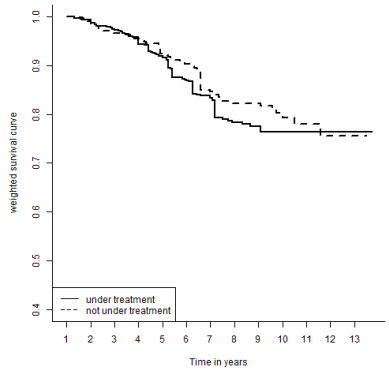
(a) Unweighted



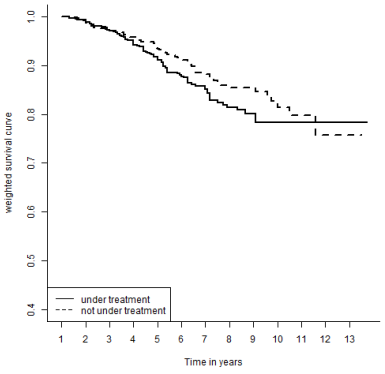
(b) Adjusted by $w^{(n)}$ (untruncated)



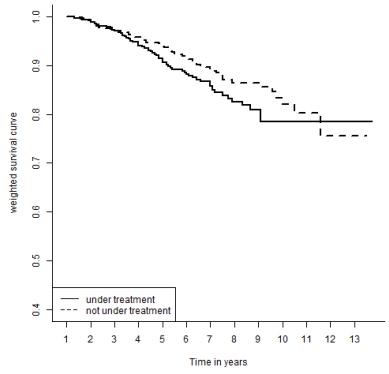
(c) Adjusted by 5% truncated $w^{(n)}$



(d) Adjusted by 10% truncated $w^{(n)}$



(e) Adjusted by 25% truncated $w^{(n)}$



(f) Adjusted by 50% truncated $w^{(n)}$

Figure 2.4: IPTC weight adjusted Kaplan-Meier-type survival curves for the effect of β-IFN on time to reaching sustained EDSS 6 for multiple sclerosis (MS) patients from British Columbia, Canada (1995-2008). The truncated weights are derived from the normalized unstabilized IPTC weights ($w^{(n)}$) so that the survival probabilities and HRs are marginal estimates with causal interpretation.

Chapter 3

The Performance of Statistical Learning Approaches to Construct Inverse Probability Weights in Marginal Structural Cox Models: A Simulation-based Comparison

3.1 Introduction

Marginal structural Cox models (MSCMs) [42, 50, 58] provide a popular approach to estimate the causal effect of time-dependent treatment from non-experimental survival data in the presence of time-dependent confounders. As discussed in Chapter 1, these models are based on the potential outcome notion of causality. In Chapter 2, we have seen that inverse probability weights (IPWs) play a key role in the MSCM approach. As with survey sampling weighting, IPWs redistribute the population by creating a pseudo-population so that the biasing effect of time-dependent confounding variables that influence the future treatment decision is removed and the association between outcome and treatment becomes unconfounded. The

validity of MSCM results based on non-experimental data depends on identifiability conditions, such as exchangeability, positivity, consistency and all models being correctly specified [55]. If these identifiability conditions hold, the resulting treatment-outcome association measures from the MSCM analysis possess a causal interpretation.

For randomized experiments, weights are usually known. Since the weights are not generally known in observational studies, we need to estimate them from the observed data. Estimation of IPWs is central to MSCM. As observed in point-treatment studies (treatment intervention occuring at a single time point in the study), MSCM estimates are highly sensitive to weight model misspecification [54]. Similar patterns are evident in longitudinal studies with moderate numbers of time periods (measurements during up to three time periods) [126, 127]. Hence, for the correct estimation of the causal parameter, the weights need to be estimated as accurately as possible. As the weights are calculated based on the product of propensity score-based estimates at each time period, longer follow-up makes weight estimates more challenging [128]. The search for techniques capable of robust estimation of IPWs is of considerable current interest [127, 129–131].

The use of logistic regression to model the exposure status is the most popular IPW estimation approach. General guidelines for IPW estimation via logistic regression are stated in the MSCM literature [55]. As estimated propensity scores (see Appendix §B.1) are subsequently utilized to create the weights for the MSCM [42, Appendix.1], findings in the propensity score literature could be valuable in the MSCM context. Propensity scores estimated from statistical learning techniques have sometimes been found to improve covariate balance compared to those estimated from logistic regression models [132, 133]. These methods provide predictions based on a relationship obtained from learning algorithms such as bootstrap aggregation (bagging), support vector machine (SVM) and boosting [134–139]. These statistical learning methods seem promising in estimating IPWs with better properties in longitudinal settings, as hypothesized by some researchers

[135, 140, 141]. However, as the implementation, mechanism and interpretation of the propensity sores and MSCM approaches differ considerably, it is not immediately clear if the lessons from propensity score literature will directly apply in the MSCM context [126].

The performance of the proposed statistical learning algorithms (bagging, SVM, boosting) have not been investigated in the context of MSCMs in the longitudinal setting. As the true weights cannot be known from observational data analysis, we need to resort to simulation to evaluate the utility of the various IPW estimation methods. Young et al. [56, 142] suggested a simulation scheme that satisfies the sufficient conditions of inverse probability weighting of the MSCM [39] and other similar methods [37, 143]. Their scheme was used and further described by subsequent simulation studies [57, 114] and later discussed elsewhere [103, 144]. Using their data generation procedure, we compare the performance of MSCMs using these proposed IPW estimation methods.

This chapter is organized as follows. In the next section, we describe the notation of the IPW estimation methods used to obtain estimates from MSCM. The next section describes the design of the simulation study, the corresponding model-specification and the metrics used for evaluating the performances of the various IPW estimation approaches. We also summarize and compare the resulting MSCM estimates. Then we present MSCM analyses on a retrospective cohort of multiple sclerosis (MS) subjects [92] (also see Chapter 2) in the next section. The chapter concludes with discussion of the results, and the strengths and weaknesses of the current study.

3.2 Marginal Structural Cox Model (MSCM)

In §2.2.2, we described the notations of MSCM in terms of time t. In this section, we will consider a fixed set of time intervals. Consider a hypothetical longitudinal study where the measurements are taken at intervals $m = 0, 1, 2, \ldots, K$. Let $t_0 = 0$ be the time of the baseline visit and L_0 be

the covariates measured at baseline. Suppose that follow-up continues to the exact failure time T. During the m-th time interval $[t_m, t_{m+1})$, binary treatment status A_m is measured immediately after recording the value of a binary covariate (L_m) in the same interval. Here, $A_m = 1$ if the subject is treated in the m-th interval and $A_m = 0$ otherwise. Similarly, $L_m = 1$ if the covariate is present for the subject in the m-th interval and $L_m = 0$ otherwise. We let $\bar{A}_m = (A_0, A_1, \ldots, A_m)$ and $\bar{L}_m = (L_0, L_1, \ldots, L_m)$ be the observed treatment history and covariate history respectively through the end of interval m and set $A_{-1} = L_{-1} = 0$. Consequently, $\bar{a}_m = (a_0, a_1, \ldots, a_m)$ and $\bar{l}_m = (l_0, l_1, \ldots, l_m)$ are the realizations of $\bar{A}_m$ and $\bar{L}_m$ respectively. After observing covariate histories until the m-th interval, we define $Y_{m+1} = I(T \leq t_{m+1})$, the indicator of failure by t_{m+1} and $\bar{Y}_{m+1} = (Y_0, Y_1, \ldots, Y_{m+1})$, the failure history through the end of interval $m+1$. By definition, subjects must be at risk at baseline, i.e., $Y_0 = 0$.

We denote a treatment regime by $\bar{a}_K = (a_0, a_1, \ldots a_m, \ldots a_K)$, a possible realization of $\bar{A}_K$. There are 2^{K+1} possible treatment regimes for a binary treatment, including $\bar{0}_K = (0, \ldots, 0)$ (never treated), $\bar{1}_K = (1, \ldots, 1)$ (always treated) and $(0, \ldots, 0, 1, \ldots, 1)$ (partly treated) etc. Let the counterfactual failure time be $T_{\bar{a}_K}$ had a subject followed a (hypothetical) regime $\bar{a}_K$. Then the counterfactual outcome history under the treatment regime is denoted by $\bar{Y}^{\bar{a}_K}_{K+1}$. Therefore, for each regime $\bar{a}_m$, we can define a MSCM as follows:

$$\lambda_{\bar{a}_m}(m) = \lambda_{\bar{0}_m}(m) \exp\big(\gamma(m, \bar{a}_m, \boldsymbol{\psi})\big), \tag{3.1}$$

where the (causal) effect is indicated by a parameter vector $\boldsymbol{\psi} = (\psi_1, \psi_2)$, γ is a known function, $\lambda_{\bar{a}_m}(m)$ and $\lambda_{\bar{0}_m}(m)$ are hazard functions for the counterfactuals $T_{\bar{a}_m}$ and $T_{\bar{0}_m}$ at time t_m. For the treatment regime $\bar{a}_m$, the causal hazard ratio is defined as $\lambda_{\bar{a}_m}(m)/\lambda_{\bar{0}_m}(m)$ comparing with $\bar{0}_m$. A causal effect is present ($\psi_1 \neq 0$) if for any $\bar{a}_m$ ($m = 0, 1, \ldots, K$), $\lambda_{\bar{a}_m}(m) \neq \lambda_{0^-_m}(m)$. The equality of the hazard functions for all $K+1$ intervals is

indicative of the absence of causal effect ($\psi_1 = 0$). We specify

$$\gamma(m, \bar{a}_m, \boldsymbol{\psi}) = \psi_1 A_m + \psi_2 L_0, \tag{3.2}$$

based on current treatment exposure [50]. See Appendix §B.2 for an extended definition of MSCM.

3.2.1 Estimation of ψ_1 from MSCM

MSCM is based on counterfactual theory. This requires creating a pseudo-population where the confounding due to the time-dependent confounder is removed from the relationship between outcome and treatment expusure. If the time-dependent confounder is a strong predictor of the treatment exposure for a patient in a given time-period, then that person-time contribution is down-weighted using IPWs. As discussed in §2.2.2, the stabilized version of the inverse probability of treatment weights is

$$sw_{im} \quad = \quad \prod_{j=0}^{m} \frac{pr(A_{ij} = a_{ij} | \bar{A}_{i,j-1} = \bar{a}_{i,j-1}, L_{i0} = l_{i0})}{pr(A_{ij} = a_{ij} | \bar{A}_{i,j-1} = \bar{a}_{i,j-1}, L_{i0} = l_{i0}, \bar{L}_{ij} = \bar{l}_{ij})};$$

see Appendix §B.3 for further details.

Simulation studies have shown that a MSCM fitted directly using the IPW weighted Cox proportional hazards model considerably reduces the variability of the estimated treatment effect [57] compared to approximate MSCM approaches, such as the IPW weighted pooled logistic regression approximation [50]. This is true even when both the direct and approximate MSCM approaches use the same weights. It was also shown that when the event rate is more frequent, the weighted pooled logistic regression approximation leads to biased estimates [56]. Therefore, we fit the MSCM directly using the Cox model with IPWs to estimate ψ_1, as was done in Chapter 2. We use the robust sandwich standard error (calculated based on residuals and weights) [145, 146] to estimate the variance of the MSCM estimators [50, 57, 104].

3.2.2 Estimation Methods of IPWs

The following methods are used to estimate the IPWs:

Logistic regression. To estimate the IPWs, treatment status at each time point is modelled with respect to the covariates associated with the treatment decision. The predicted values from this logistic regression model are the most commonly used to generate treatment weights. Logistic regression is easy to understand and interpret, but violation of its assumptions leads to invalid inference.

We use the IPWs estimated from this approach as a baseline to compare the properties of IPWs estimated from other methods.

Bagging. Bagging is a statistical learning algorithm intended to increase the power of a predictive model [147, 148]. B bootstrap samples from the original dataset are used as B training sets. Predictive classification trees are grown to make predictions for the original dataset. If the constructed trees are not pruned, predictions are generally associated with high variability but low bias due to possible over-fitting. The resulting B predictions are then aggregated and majority vote assigns a final predicted value for each treatment status (0 or 1). These predictions are generally associated with less variability and more accuracy than those obtained from using logistic regression [149, 150]. This is especially true when logistic regression provides unstable prediction, which is sometimes the case while estimating treatment weights [73].

In our simulations, we use $B = 100$ bootstrap replications. As suggested in the propensity scores literature, 10-fold cross-validation is used in order to obtain better predictions [151].

Support vector machines. SVM is a highly flexible statistical learning algorithm to find the optimal separating hyperplane (say, a straight line in 2 dimensions) that gives the best separation between the treatment

classes. The resulting separation rule puts the binary classes (treated versus untreated) are as far as possible from the hyperplane [149, 150, 152]. To facilitate finding a hyperplane that maximally separates binary classes, SVM maps (transforms) the covariates into a higher dimensional space by using the kernel function. Even with noisy data, SVMs generally perform well in classification problems.

In our simulations, we use the polynomial kernel to fit the SVM. Available software routines enable us to obtain probability estimates or predictions from the SVM fit via internal cross-validation procedures [153].

Boosting. Boosting is a general approach for improving predictions, which can be applied in many regression and classification problems [134, 149, 150, 154]. Bagging and boosting work in a similar iterative way with one exception. Bagging uses bootstrap samples whereas boosting sequentially uses a modified version of the original data in each iteration. These modifications are based on the information obtained from the previous iterations. That is, in the boosting fitting procedure, in each successive iteration $b = 1, 2, \ldots, B$, this algorithm places more weight on those treatment statuses that were misclassified in the previous iterations. Then the final predicted values of treatment status are obtained from an average (weighted by a shrinkage parameter, not a simple average as for the bagging approach) of the B predictions. When classification trees are used in the process, this method inherits their flexible properties, while being able to capture complex interactions among covariates and nonlinear effects [155].

In this chapter, $B = 1,000$ trees are used in each fit, with maximum two-way interaction ($d = 2$) of all the covariates under consideration and shrinkage parameter set to 0.01.

3.2.3 IPW schemes

Having a smaller standard deviation (SD) is desirable property for the weights [49, 55]. Even though the unstabilized weights w (see Appendix §B.3 equation (B.3)) produce consistent MSCM estimates, these weights are notorious for producing extreme weights that can ultimately lead to inefficient estimates and impractical confidence interval widths [55, 57]. Generally, unstabilized weights w are associated with high variability. The variability of the weights is an important factor to consider because more variable weights may lead to more variability of the MSCM estimates. Stabilization (see Appendix §B.3; equation (B.6)) reduces the variability of the weights, while not affecting the consistency of the estimate of the MSCM parameter [50, 156]. Even after stabilization, we may observe a few extreme weights.

Several other ad-hoc suggestions have been proposed to further reduce the variability of the weights. Although they are practically useful, they do not have much theoretical justification (see Appendix §B.5). Weight truncation [55, 104, 157, 158] reduces excess variability in the weights. Normalization (see Appendix §B.3; equation (B.8)) is a popular survey sampling technique that found its way into the MSCM literature [57]. In this study, we compare various unstabilized, stabilized and normalized versions of IPWs. We also assess their characteristics under increased levels of truncation.

3.2.4 Fitting Weight Models to Estimate IPW

For generating unstabilized, stabilized and normalized versions of the inverse probability of treatment weights using logistic regression, we used the IPW generating formulas (equations 2, 5, 7 in Appendix §B.3). The denominator model for the unstabilized and stabilized weights included the lagged value of treatment status A_{m-1}, the follow-up month index m, the time-dependent covariate L_m, its lagged vaue L_{m-1} and its interaction with lagged treatment status ($A_{m-1} \times L_m$) (for both equations 2 and 5 in Appendix §B.3). The numerator model for the stabilized weights included the lagged value of treatment status A_{m-1} and the follow-up month index m (for equation 5 in

Appendix §B.3). We get normalized unstabilized and stabilized versions of the weights by normalizing the unstabilized and stabilized weights respectively (equation 7 in Appendix §B.3). The same list of covariates are used to generate the IPW using the bagging, SVM and boosting approaches. As the second order interaction depth of the covariates ($d = 2$) is selected for the boosting approach, explicitly specifying the interaction ($A_{m-1} \times L_m$) was not necessary. The software implementation details are described in Appendix §B.4.

3.3 Design of Simulations

MSCM is a popular tool to estimate the causal effects of time-varying treatments in the presence of time-varying confounders that are affected by previous treatment exposure. To study the properties of this method, we need to be able to simulate data from a MSM with specified parameter values so that we can evaluate how well MSCM performs. Several studies have simulated data from MSCMs under different conditions [56, 57, 102, 103, 114, 142, 144, 159, 160]. We apply a data generation process due to Young et al. [56] where both treatment status and confounder values are generated based on their lagged values. We briefly describe the data generation procedure here.

For this longitudinal follow-up study, let $i = 1, 2, \ldots, n$ be the subject index. At each time interval $[t_m, t_{m+1})$, values of a single time-dependent confounder L_{im} and time-dependent treatment A_{im} are sampled from a Bernoulli distribution with probabilities p_L and p_A respectively, where p_L and p_A are defined as follows:

$$\begin{aligned} logit(p_L) &= logit\ Pr(L_m = 1 | A_{m-1}, L_{m-1}, Y_m = 0; \boldsymbol{\beta}) \\ &= \beta_0 + \beta_1 I(T_0^{-} < c) + \beta_2 A_{m-1} + \beta_3 L_{m-1}, \end{aligned} \tag{3.3}$$

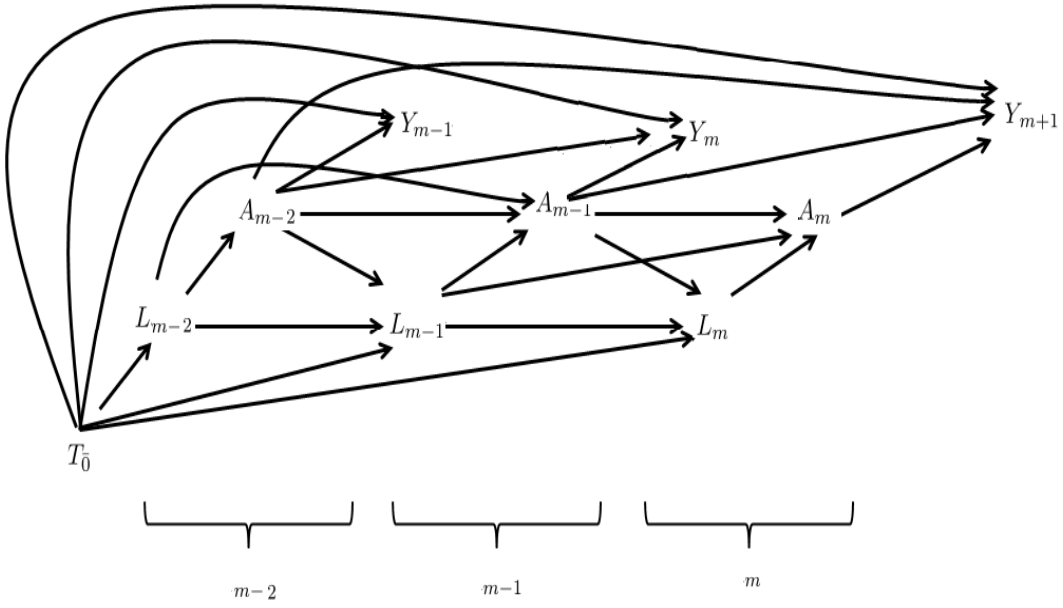


Figure 3.1: Causal diagram depicting the dependencies in the marginal structural Cox model (MSCM) data generation algorithm.

$$\begin{aligned} logit(p_A) &= logit\ Pr(A_m = 1 | L_m, A_{m-1}, L_{m-1}, Y_m = 0; \boldsymbol{\alpha}) \\ &= \alpha_0 + \alpha_1 A_{m-1} + \alpha_2 L_m + \alpha_3 L_{m-1} + \alpha_4 L_m \times A_{m-1}, \quad (3.4) \end{aligned}$$

where $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4)$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$. Here, $T_{\bar{0}}$ is the untreated counterfactual survival time and c is an arbitrary cut-point used to generate the binary variable $I(T_{\bar{0}} < c)$. The sampling distributions of L_m and A_m both depend on their previous lagged values, i.e., l_{m-1} and a_{m-1}. In particular, past treatment exposure status A_{m-1} is a predictor of L_m, which then predicts future treatment exposure A_m. The confounding in the exposure-outcome relationship arises via the following path: $Y_{m+1} \leftarrow T_{\bar{0}} \rightarrow L_m \rightarrow A_m$ (see Figure 3.1). While generating treatment status in the next interval, we also include an interaction term between past treatment status A_{m-1} and current confounder status L_m. This interaction $A_{m-1} \times L_m$ mimics the commonly occurring situation that both of these factors influence future treatment decisions.

The untreated counterfactual survival time (for the never-treated regime $\bar{0}_K \equiv \bar{0}$), $T_{i\bar{0}}$ for each person is sampled from an exponential distribution with constant hazard $\lambda_{T_{\bar{0}}}$. The counterfactual survival time under a given regime $\bar{a}_m$, $T_{i\bar{a}_m}$ is calculated from the cumulative hazard $\int_0^{m+1} \lambda_{\bar{a}_j}(j)dj \equiv \sum_{j=0}^{m} \lambda_{\bar{a}_j}(j)$. At each step of the data generation procedure, this cumulative hazard $\int_0^{m+1} \lambda_{\bar{a}_j}(j)dj$ is updated based on the new a_m value, accumulating the risk for the regime $\bar{A}_m = \bar{a}_m$. The counterfactual survival times $T_{\bar{0}}$ and $T_{\bar{A}_m}$ follow the same distribution if either $\psi_1 = 0$ or $A_m = 0$ for all m [142, 160]. Therefore, the sampled $T_{i\bar{0}}$ (with hazard $\lambda_{i\bar{0}}$ for the never-treated regime $\bar{0}$) is compared with the calculated $T_{i\bar{a}_m}$ (with cumulative hazard $\int_0^{m+1} \lambda_{\bar{a}_j}(j)dj$ for the simulated regime $\bar{a}_m$) to determine whether the subject fails in the next interval, i.e., if the sampled $T_{i\bar{0}}$ is greater than the calculated $T_{i\bar{A}_m}$, then the failure indicator $Y_{i,m+1} = 0$; otherwise $Y_{i,m+1} = 1$.

3.3.1 Simulation Specifications

In observational epidemiologic studies of drug effectiveness, confounding by indication is a common problem. This is a specific type of confounding encountered when the allocation of the treatment A is not random, and the physician's decision to assign a treatment to a particular subject is affected by factors such as the severity of disease, concurrent therapies, concomitant medical conditions L_m (say, disease activity), or combinations of these conditions (e.g., interactions). To mimic this confounding by indication in the simulation, the treatment status at each stage A_m is generated by the following factors: the previous therapy, A_{m-1}, the current and past medical conditions or symptoms, L_m and L_{m-1} respectively and the interaction $A_{m-1} \times L_m$. We assume that being treated in the previous time-period ($A_{m-1} = 1$) positively stimulates ($\alpha_1 = 1/2 = 0.5$) a subject to continue treatment in the current period ($A_m = 1$), whereas occurrence of current and past symptoms ($L_m = 1$ and $L_{m-1} = 1$ respectively) positively encourages ($\alpha_2 = 1/2 = 0.5, \alpha_3 = \log(4) = 1.39$) a subject to take the treatment

in the current period ($A_m = 1$) (in equation (3.4)). If a subject was under treatment in the previous period and the subject is currently suffering from symptoms (i.e., $A_{m-1} = 1$ and $L_m = 1$), both of these factors influence the subject positively ($\alpha_4 = \log(6/5) = 0.18$) to continue treatment in the current period ($A_m = 1$). In the absence of previous treatment ($A_{m-1} = 0$) and current or previous symptoms ($L_m = 0$ and $L_{m-1} = 0$), the subject is less likely ($\alpha_0 = \log(2/7) = -1.25$) to take the treatment in the current time-period ($A_m = 1$) and more likely to discontinue ($A_m = 0$). Therefore, the associated parameter vector in equation (3.4) is $\boldsymbol{\alpha} = (\log(2/7), 1/2, 1/2, \log(4), \log(6/5))$.

In our simulations, the time-dependent confounder, L_m, is similarly generated by the previous treatment status A_{m-1}, the lagged time-dependent confounder L_{m-1} and a binary confounder $I(T_{\bar{0}} \leq c)$ associated with the counterfactual outcome $T_{\bar{0}}$. We assume that having a survival $T_{\bar{0}}$ shorter than a cut-point c (i.e., $I(T_{\bar{0}} \leq c) = 1$) puts a subject under an increased risk ($\beta_1 = 2$) of developing a symptom ($L_m = 1$) (in equation (3.3)). The cut-point is set to $c = 30$. Being treated in the previous time interval ($A_{m-1} = 1$) reduces a subject's risk ($\beta_2 = \log(1/2) = -0.69$) of developing a new symptom ($L_m = 1$). A subject who experienced a symptom in the previous period ($L_{m-1} = 1$) is also more likely ($\beta_3 = \log(3/2) = 0.40$) to develop a new symptom in the current time-period ($L_m = 1$). In the absence of previous treatment ($A_{m-1} = 0$) and previous symptoms ($L_{m-1} = 0$), a subject is less likely ($\beta_0 = \log(3/7) = -0.85$) to develop a new symptom ($L_m = 1$). Therefore, the associated parameter vector in equation (3.3) is $\boldsymbol{\beta} = (\log(3/7), 2, \log(1/2), \log(3/2))$. The true causal effect parameter in equations (3.1) and (3.2) is set such that the treatment is less hazardous ($\psi_1 = -0.5$ in log-hazard scale) to the subjects and therefore has a beneficial effect.

Note that the same set of covariates are used to generate the treatment in the data-generation algorithm (see equation 3.4) and in the weight model fitting process, except for the follow-up index m. Using this follow-up or

visit index variable m in the weight model allows us to estimate a separate intercept for each visit (say, month) [69]. Other flexible choice of modelling (say, smoothing this index m) are also possible [50], but were not used in our weight model fitting.

To study the properties of the weight estimation procedures, we generate a large dataset with $n = 25,000$ subjects. In the simulations, we generate datasets with $n = 2,500$ subjects, each followed for up to $m = 10$ subsequent monthly visits as in previous studies [56, 57, 114]. To assess the small sample properties, these simulations are repeated for a smaller sample size ($n = 300$). $T_{i\bar{0}}$'s were sampled from an exponential distribution, with constant $\lambda_0 = 0.01$ rate of monthly events throughout the follow-up to mimic the rare disease condition. To mimic a more frequent event rate scenario, the rate is increased to constant $\lambda_0 = 0.10$ rate of monthly events throughout the follow-up. The Monte Carlo study consists of $N = 1,000$ generated datasets for each setting under consideration. The pseudocode for our simulation design is provided in Appendix §B.6.

3.3.2 Performance Metrics

We assessed the performance of the various weighting schemes by the following measures

- Bias $= \sum_{i=1}^{N}(\hat{\psi}_{1i} - \psi_1)/N$: The average difference between the true and $N = 1,000$ estimated parameters (log hazard ratio) from the MSCM model.
- SD $= \sqrt{\sum_{i=1}^{N}(\hat{\psi}_{1i} - \psi_1')^2/(N-1)}$ where $\psi_1' = \sum_{i=1}^{N}\hat{\psi}_{1i}/N$
- MSE $= \sqrt{\sum_{i=1}^{N}(\hat{\psi}_{1i} - \psi_1)^2/N}$
- Model-based SE: The average of $N = 1,000$ estimated standard errors of the estimated causal effect from the MSCM model.
- Coverage probabilities of model-based nominal 95% CIs: Proportion of the $N = 1,000$ datasets for which the true parameter is contained

in the 95% CI.

3.4 Simulation Results

3.4.1 IPW Summary

Summaries of the (untruncated) weights calculated from different approaches from one large simulated dataset with 25,000 subjects, each with up to 10 visits, are presented in Table 3.1. As expected, for each fitting approach, the mean and standard deviation are noticeably larger for the unstabilized weights w. Normalization is effective in reducing the variability of the unstabilized weights (i.e., $w^{(n)}$) but stabilization (i.e., sw) is even better. Normalization of the stabilized weights (i.e., $sw^{(n)}$) has little impact on the variability. Bagging results in a reduction in variability and SVM reduces the variability even further. With boosting, the variabilities of the unstabilized weights (w and $w^{(n)}$) increase slightly compared to those of the corresponding weights from SVM. Surprisingly, the variablities of the boosting stabilized weights (i.e., sw and $sw^{(n)}$) increase more than 5-fold compared to those from SVM.

As expected, the effect of increased levels of truncation is monotone in reducing the variability of IPWs generated from all approaches (see Appendix §B.7: Appendix Tables B.1-B.4). When the weight variability is already small, the truncation has less of an effect on variability reduction.

This data is generated under the rare disease condition ($\lambda_0 = 0.01$ in a monthly scale and the corresponding event rate is 0.010 in this dataset). The event rate becomes as frequent as 0.075 when the parameter λ_0 is increased to 0.10 in a monthly scale under the same data generating conditions.

Table 3.1: Summaries of the (untruncated) weights estimated by different methods (l = logistic, b = bagging, svm = SVM, gbm = boosting) under different weighting schemes (w = unstabilized, $w^{(n)}$ = unstabilized normalized, sw = stabilized, $sw^{(n)}$ = stabilized normalized) from the simulation study with a large $(25,000)$ number of subjects, each with up to 10 visits, under the rare event condition.

	Min.	Q1	Median	Mean	Q3	Max.	sd	p >20	p >100
$l - w$	1.21	3.96	17.82	189.70	98.09	12780.00	666.15	90.38	47.23
$l - w^{(n)}$	0.01	0.22	0.58	1.00	1.22	13.99	1.37	0.00	0.00
$l - sw$	0.33	0.79	0.94	1.00	1.19	2.54	0.34	0.00	0.00
$l - sw^{(n)}$	0.32	0.78	0.94	1.00	1.18	2.48	0.33	0.00	0.00
$b - w$	1.28	4.08	18.62	195.80	101.50	8990.00	641.09	96.09	49.35
$b - w^{(n)}$	0.01	0.26	0.66	1.00	1.26	12.56	1.31	0.00	0.00
$b - sw$	0.36	0.92	0.98	1.00	1.06	1.99	0.19	0.00	0.00
$b - sw^{(n)}$	0.35	0.92	0.98	1.00	1.06	1.95	0.19	0.00	0.00
$svm - w$	1.35	4.50	20.25	161.40	100.70	6568.00	466.04	80.87	40.52
$svm - w^{(n)}$	0.03	0.34	0.72	1.00	1.31	8.33	1.10	0.00	0.00
$svm - sw$	0.76	0.95	1.01	1.00	1.06	1.22	0.08	0.00	0.00
$svm - sw^{(n)}$	0.76	0.95	1.01	1.00	1.05	1.22	0.08	0.00	0.00
$gbm - w$	1.24	3.56	16.09	163.60	90.74	6441.00	477.65	75.26	38.22
$gbm - w^{(n)}$	0.01	0.23	0.60	1.00	1.22	8.86	1.29	0.00	0.00
$gbm - sw$	0.21	0.77	0.93	0.99	1.10	3.41	0.42	0.00	0.00
$gbm - sw^{(n)}$	0.21	0.77	0.94	1.00	1.11	3.45	0.42	0.00	0.00

3.4.2 Comparing IPW Estimation Approaches

Results of the simulation using $1,000$ datasets (each with $n = 2,500$) under the rare event condition ($\lambda_0 = 0.01$ in a monthly scale) are shown in Figures 3.2 - 3.6.

Figure 3.2 shows the bias pattern in estimating the MSCM parameter $\psi_1 = -0.5$ when the IPWs are estimated using the four different approaches. The untruncated weights generated using logistic regression and boosting successfully estimate the parameter ψ_1, while bagging introduces some bias in estimating ψ_1, and SVM yields even more bias. This description is valid for all the weighting schemes (w, $w^{(n)}$, sw and $sw^{(n)}$). Under increased levels of truncation, a clear pattern is visible: as expected, increasing the level of truncation increases the bias. SVM is clearly doing worse than logistic regression in terms of bias. Bagging is doing better than SVM, but clearly not as well as logistic regression. Boosted regression is performing as well as logistic regression and for the stabilized cases (sw, $sw^{(n)}$), it is doing slightly better. In general, the bias of all IPW estimation approaches agree at 50% truncation. As theory suggests, this indicates the bias obtained from a baseline-adjusted analysis [55].

Figure 3.3 shows the pattern of variability (SD) of the causal effect estimates from MSCM. This figure shows that the SDs of $\hat{\psi}_1$ for each set of weights under consideration are very similar for the same level of truncation. As expected, the unstabilized IPWs from all methods were associated with higher SDs. Normalizing the unstabilized IPWs reduced this variability. Stabilization was also effective in reducing variability. Normalization of the stabilized IPWs had little further effect. Figure 3.4 shows that, except for SVM, the average model-based standard errors (SE) of $\hat{\psi}_1$ were similar to the empirical SDs of $\hat{\psi}_1$.

Figure 3.5 summarizes the MSE patterns of the MSCM estimates. As the SDs from the different approaches were similar, differences in MSE were

mainly due to differences in bias. The sharp decrease in MSE at low levels of truncation in most of the curves suggests that a low level of truncation might yield better estimates (compared to no truncation) in terms of MSE.

The coverage probabilities of model-based nominal 95% CIs for ψ_1 are shown in Figure 3.6. The untruncated IPWs show good coverage when computed from the boosting or logistic regression approaches. As the bias increases and the SE decreases under increased levels of truncation, it is not surprising that the coverage probability decreases sharply as the level of truncation increases. IPWs calculated from boosting yield as good coverages as those from logistic regression. With stabilization, boosting yields slightly better coverage than logistic regression. The bagging approach does not perform well in terms of coverage probability and the performance of SVM is even worse.

3.4.3 Properties From Smaller Samples

The corresponding results from the simulation for $n = 300$ appear in Appendix Figures B.1-B.5 in Appendix §B.8.1. The bias is slightly larger with this smaller sample size, though the patterns are similar compared to the $n = 2,500$ case (Appendix Figure B.1). As expected, the SDs of $\hat{\psi}_1$ are much higher in all settings compared to the $n = 2,500$ case (Appendix Figure B.2). Except for SVM, the patterns of average SE are similar to the SDs of $\hat{\psi}_1$ (Appendix Figure B.3). In this smaller sample case, bias and variance both are larger, affecting the MSE (Appendix Figure B.4) and the patterns of the MSE curves differ from the $n = 2,500$ case. Except for the unstabilized weights in the smaller sample case, MSE increases with higher levels of truncation. For $n = 2,500$ we observe a sharp drop below 5% truncation and then an upward trend, whereas for $n = 300$ the drop continues to around 10% truncation. This suggests that the levels of truncation up to 10% might be beneficial in obtaining MSCM estimates from such smaller samples. The coverage probabilities of the model-based nominal 95% CIs are always less

than 90% for the unstabilized weights (Appendix Figure B.5). However, when IPWs are either normalized or stabilized, or both, the coverage probabilities at all low levels of truncation are at least 90%, except for the SVM approach. However, the coverage probabilities of even these weights never reach the nominal 95% level.

3.4.4 When More Events are Available

When this simulation is repeated with $n = 2,500$ but with $\lambda_0 = 0.10$ instead of $\lambda_0 = 0.01$ in a monthly scale, the level of bias is substantially reduced (Appendix Figure B.6 in Appendix §B.8.2 compared to Figure B.1). Bagging results in more bias than logistic regression and SVM is still worse. Boosting performs as well as logistic regression. The SDs of $\hat{\psi}_1$ are much lower in all settings compared to the rare event scenario (Appendix Figure B.7). The patterns of average SE are similar to the SDs of $\hat{\psi}_1$ (Appendix Figure B.8, compared with Appendix Figure B.7). In the rare event scenario for both the $n = 2,500$ and $n = 300$ cases, the average SEs are generally larger than the SDs of $\hat{\psi}_1$ when using SVM. However, this discrepancy is not as severe in the frequent event scenario. The almost constant MSEs with increased levels of truncation in this scenario suggests that truncation may not be very helpful in improving MSCM estimates in terms of MSE except for the unstabilized weights (Appendix Figure B.9). However, normalization, stabilization or both are still effective in obtaining estimates with lower MSE. The coverage probabilities of model-based nominal 95% CIs obtained from untruncated weights in this scenario are not very different than in the rare event scenario (Appendix Figure B.10). However, the coverage probabilities do not decrease nearly as quickly with higher levels of truncation. Even with 50% truncation, the coverage probabilities are close to 75% (as opposed to close to 0% in the rare event setting).

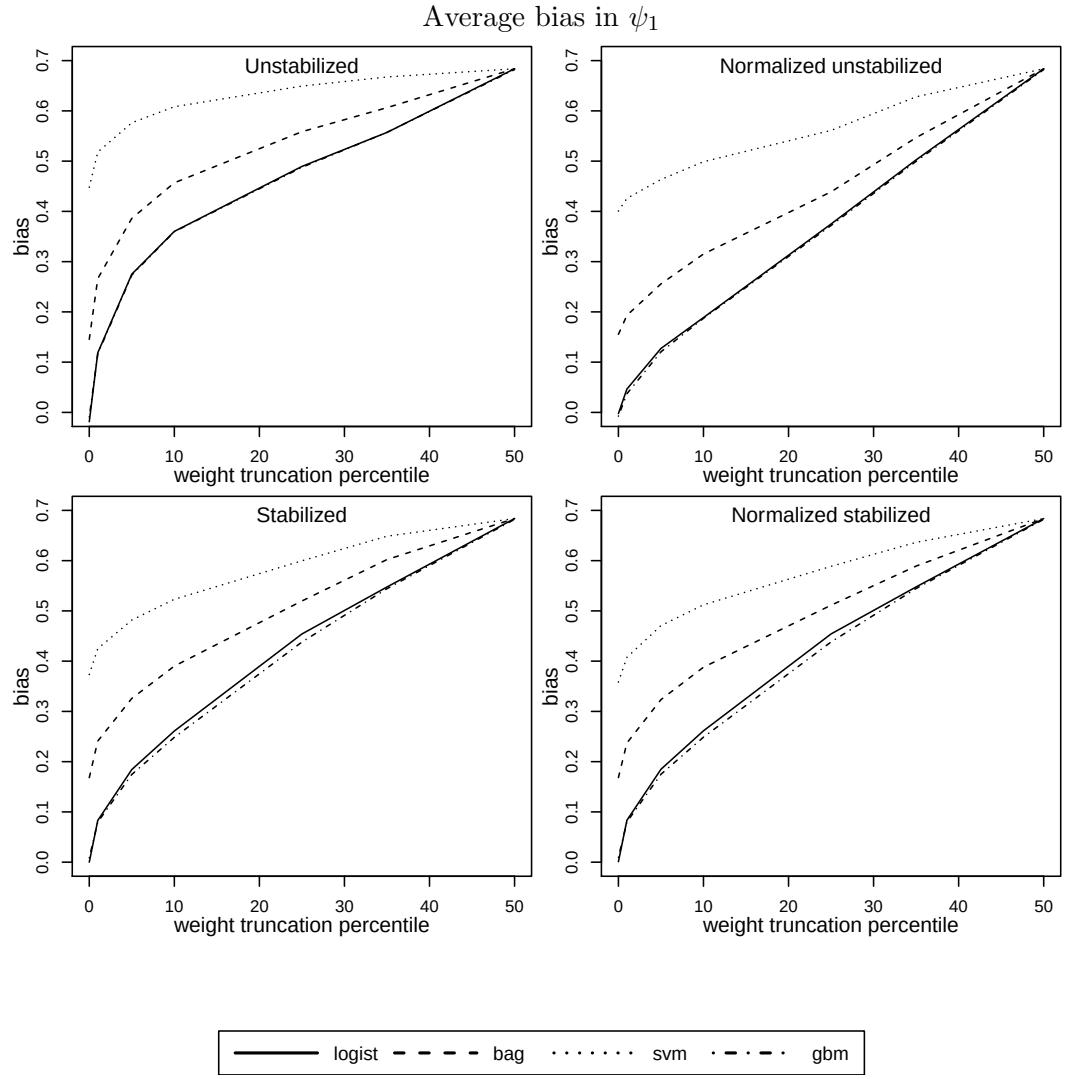


Figure 3.2: Bias of MSCM estimate $\hat{\psi}_1$ under different IPW estimation approaches when the large weights are truncated with increased levels in a simulation study of $1,000$ datasets with $2,500$ subjects observed at most 10 times.

3.4.5 Computational Time

The computational time for running the `R` process for each IPW generating approach (for estimating the unstabilized weights) using a dataset with 300

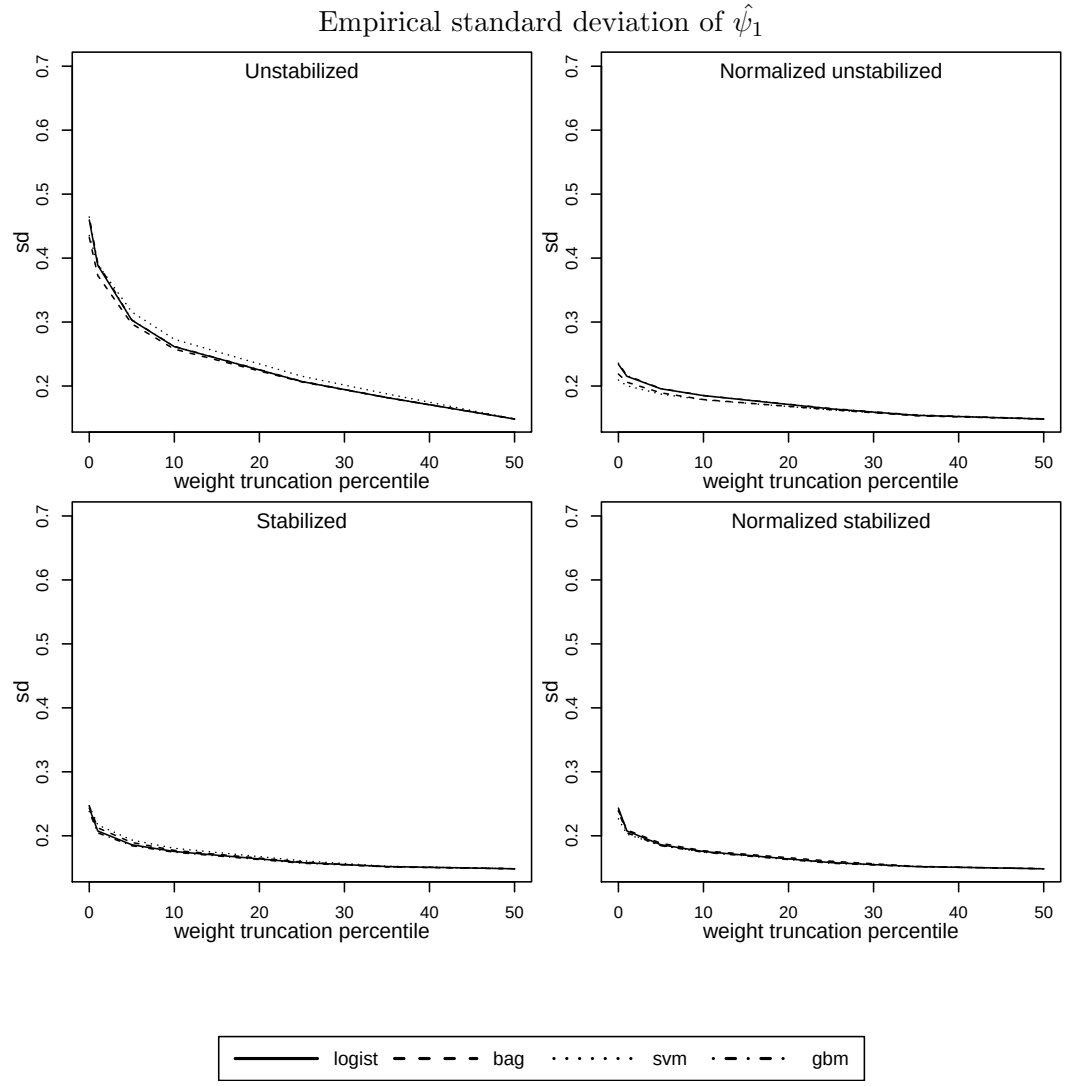


Figure 3.3: Empirical standard deviation of MSCM estimate $\hat{\psi}_1$ under different IPW estimation approaches when the large weights are truncated with increased levels in a simulation study of $1,000$ datasets with $2,500$ subjects observed at most 10 times.

subjects having up to 10 visits is reported in Table 3.2 for a Windows 7 64 bit machine.

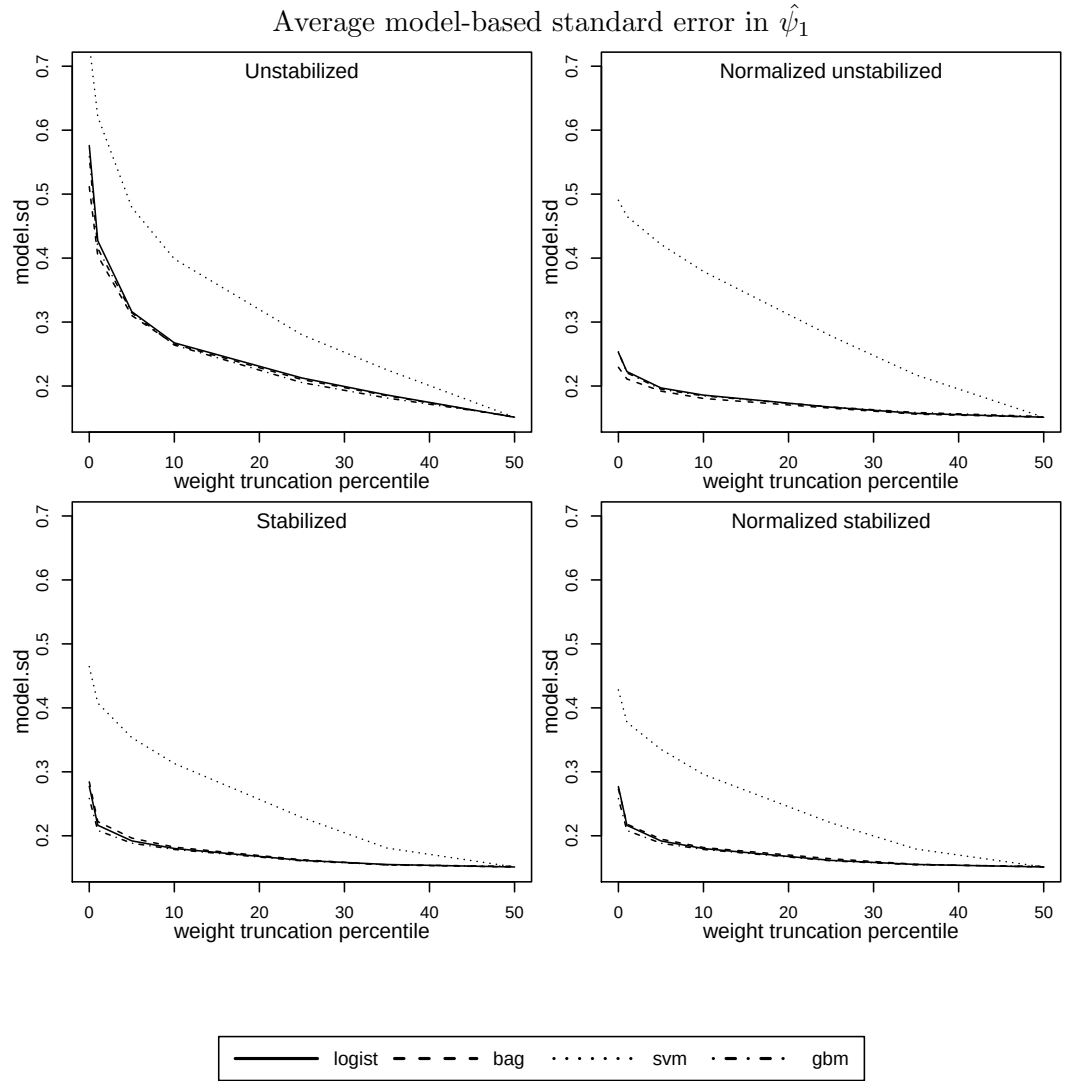


Figure 3.4: Average model-based standard error of MSCM estimate $\hat{\psi}_1$ under different IPW estimation approaches when the large weights are truncated with increased levels in a simulation study of $1,000$ datasets with $2,500$ subjects observed at most 10 times.

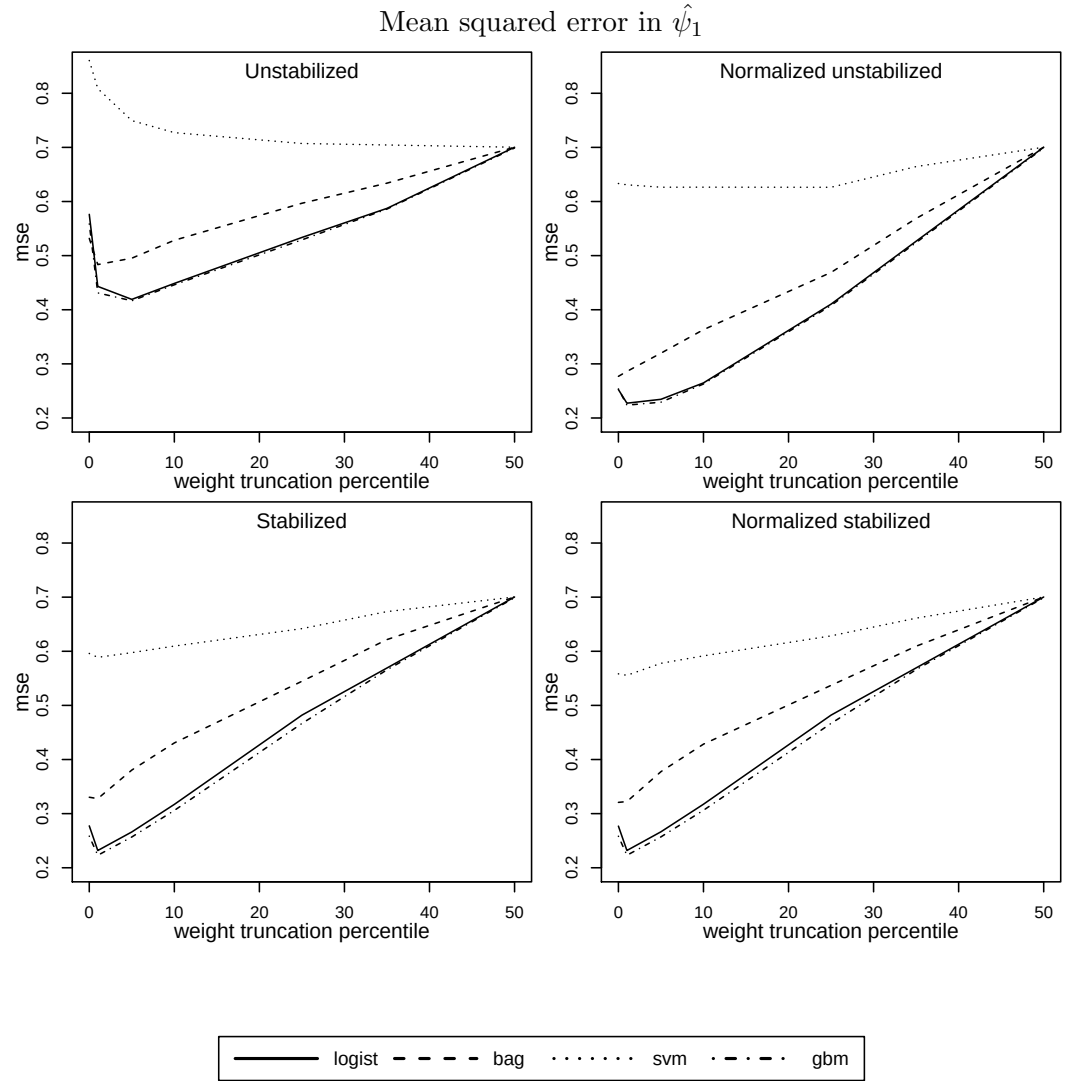


Figure 3.5: Mean squared error of MSCM estimate $\hat{\psi}_1$ under different IPW estimation approaches when the large weights are truncated with increased levels in a simulation study of $1,000$ datasets with $2,500$ subjects observed at most 10 times.

3.5 Empirical Multiple Sclerosis Application

We apply the methodologies described in this chapter in the British Columbia MS cohort (1995-2008) described in §2.2.1. The β-IFN exposure is defined

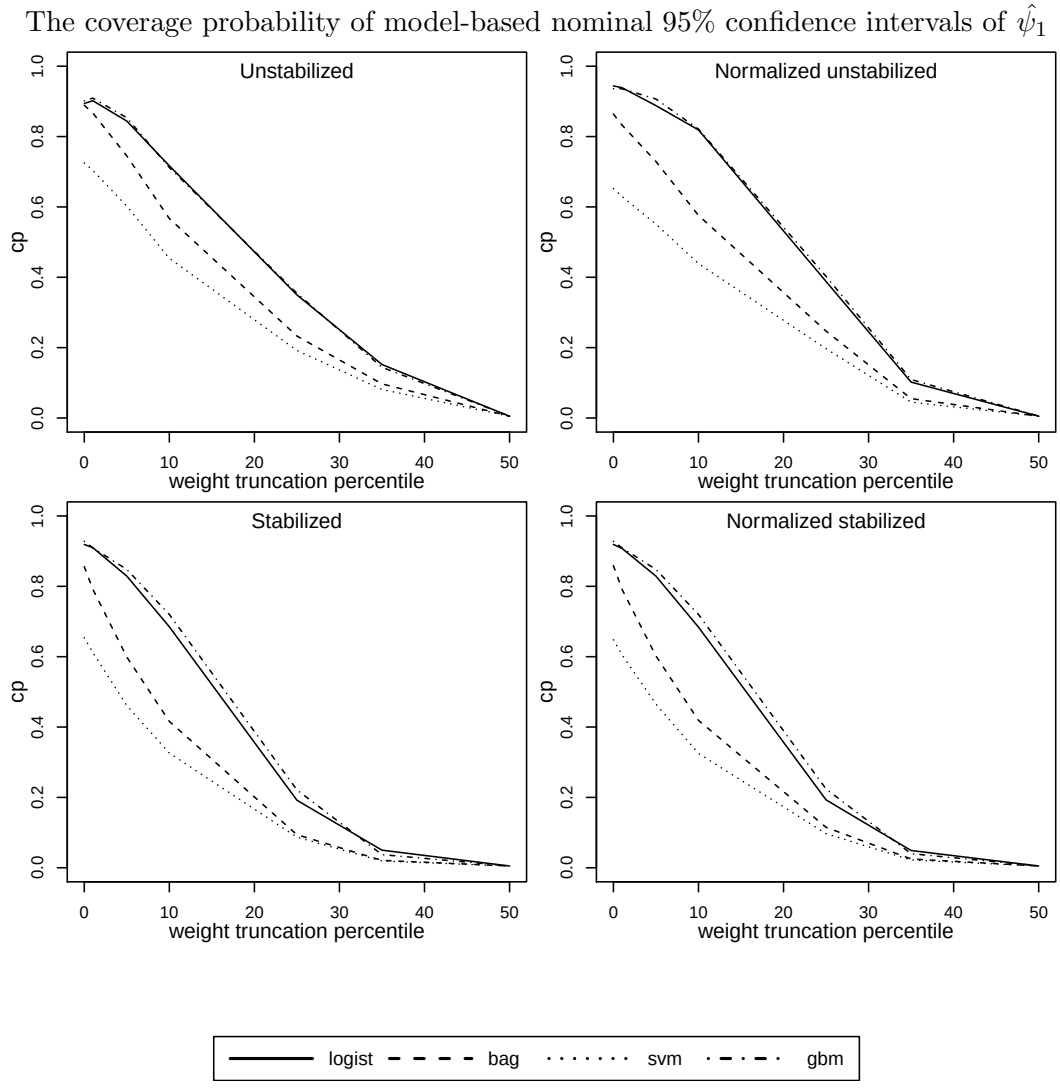


Figure 3.6: The coverage probability (cp) of model-based nominal 95% confidence intervals based on the MSCM estimate $\hat{\psi}_1$ under different IPW estimation approaches when the large weights are truncated with increased levels in a simulation study of $1,000$ datasets with $2,500$ subjects observed at most 10 times.

as a time-dependent variable A_m, measured on a monthly basis. We assess the impact of β-IFN on time to reach irreversible disability progression (sur-

vival outcome) in the real-world clinical practice setting. From the follow-up between July 1995 and December 2004, 1,697 patients are included in the study, 829 of whom never receive the β-IFN treatment. Among the 6,890 person-years of follow-up, 2,530 person-years are β-IFN exposed. Ultimately, 138 subjects reached irreversible disability, measured by sustained EDSS 6. Appendix §A.6 describes the baseline characteristics.

As discussed in in Chapter 2, MSCMs are an appropriate choice of model to adjust for the time-dependent confounder L_m cumulative relapses and baseline confounders L_0: age, sex, disease duration, and EDSS score. IPWs are estimated using the following methods: logistic regression, bagging, SVM, boosting. The resulting estimate of $\exp(\psi_1)$ and the corresponding robust standard errors are compared.

As the stabilized normalized weights $sw^{(n)}$ performed well in the simulation, we used $sw^{(n)}$ generated from different IPW estimation approaches as the MSCM weights in our analyses (equation (B.8) in Appendix §B.3). We first calculate treatment weights sw^T using the general inverse probability of treatment weight model (equation (B.6) in Appendix §B.3). In all IPW estimation methods, the numerator model included the baseline covariates L_0 (EDSS score, age, disease duration, sex), the lagged treatment status A_{m-1}, and the follow-up month index m (equation (B.7) in Appendix §B.3). The denominator model included the numerator model covariates as well as the time-dependent covariate L_m 'cumulative number of relapses for last 2 years' and its interaction with the lagged treatment status ($A_{m-1} \times L_m$) (equa-

Table 3.2: Time required to compute IPWs using various approaches

IPW estimation approach	Time (in seconds)
Logistic regression	0.02
Bagging	5.12
SVM	3.05
Boosting	50.50

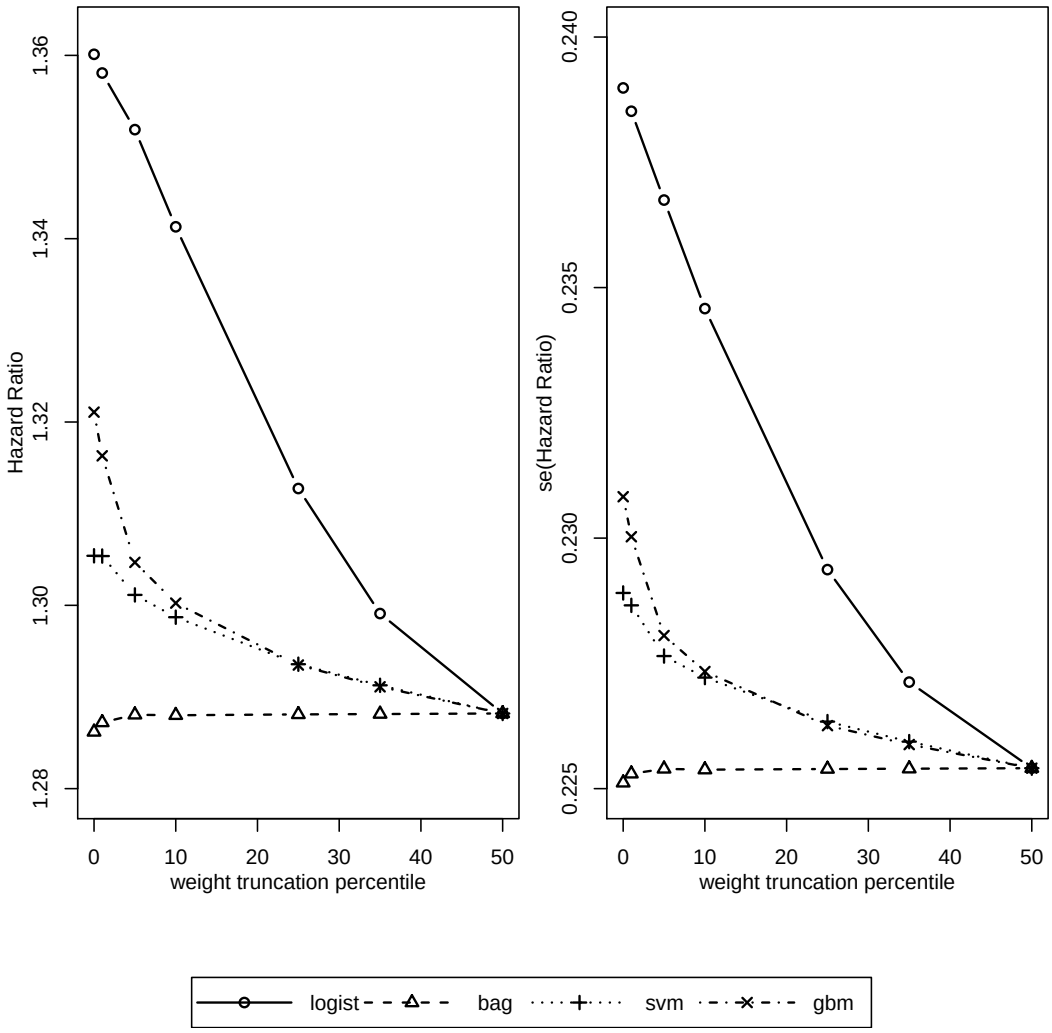


Figure 3.7: Performance of stabilized normalized weights estimated from different IPW estimation approaches for MSCM analysis in a multiple sclerosis study.

tion (B.5) in Appendix §B.3). Since this was an observational study and artificial or non-random censoring may be present, we also need to calculate censoring weights sw^C [55]. Setting censoring status as the dependent variable, the same numerator and denominator covariate specifications as in treatment weight model were used to generate the inverse probability of

censoring weights. Multiplying the treatment and censoring weights yields the IPWs $sw = (sw^T \times sw^C)$ and we normalize sw to get $sw^{(n)}$ (equation (B.8) in Appendix §B.3). The $sw^{(n)}$ weighted MSCM further adjusts for the baseline covariates (equation (B.2) in Appendix §B.2). We also assessed the impact of increased levels of weight truncation.

Figure 3.7 shows the estimated hazard ratio and corresponding robust standard error from the fitted MSCMs. IPWs generated using SVM and boosting methods show fairly similar results in terms of the estimated hazard ratio $HR = \exp(\hat{\psi}_1)$ and its robust standard errors. HR estimates based on logistic IPWs are associated with higher robust standard errors. The results from bagging do not change much under increased levels of truncation. In fitting weight models using the bagging approach, we used $B = 100$. As $B = 100$ may be an inadequate number of boostrap replicates to stabilize the misclassification error rates in the bagging approach, we repeated the analysis with $B = 1,000$. However, this did not have much impact on the MSCM estimates (data not shown). Appendix Tables B.5-B.8 in Appendix §B.9 summarize the stabilized normalized weights generated from the four approaches, the corresponding hazard ratio estimates and the confidence intervals in more detail. As shown in those tables, the IPW weights are well-behaved (mean approximately 1 and small SD) and none of the analyses yield strong evidence of an association between β-IFN exposure and time to reaching a sustained EDSS 6. Except for bagging, standard errors reduce with higher levels of truncation. For the bagging approach, estimates are close to those obtained from the baseline-adjusted analysis under all truncation levels. The performance of the IPW estimation methods to estimate ψ_1 is also provided on the log-hazards scale (Appendix Figure B.11 in Appendix §B.9). The patterns of ψ_1 in this figure are similar to that of HR in Figure 3.7. The differences in the SEs of $\hat{\psi}_1$ from the different approaches are very small.

Based on the simulation results, we know that when IPWs are estimated from the boosting approach, the corresponding MSCM estimates are better

or at least similar to that of logistic regression approach. Therefore, based on our simulation results, the use of boosting as an IPW estimation method serves as an excellent sensitivity analysis.

3.6 Discussion

The MSCM approach is built on counterfactual theory where a pseudo-population is built based on IPWs. The confounding due to the time-dependent confounder is removed from the relationship between outcome and treatment exposure in this pseudo-population. Estimation of IPWs is essential in the process.

The probability of receiving treatment given the covariates is known as the propensity score. IPWs can be thought as an extension to propensity scores when treatment is time-dependent in longitudinal studies [161]. Typically, as with propensity scores, the IPWs are estimated using logistic regression models. Assessment of the assumptions of the corresponding logistic regression fits are rarely seen in the MSCM literature. In fact, many analyses involving MSCM do not report important IPW summaries adequately [51, 52]. Alternative modelling strategies, such as statistical learning methods, have been explored in the propensity score literature. These strategies achieve the same goal of obtaining covariate balance but require fewer assumptions. As propensity scores and MSCM estimation are quite different, we investigated whether the success of these alternative methods in propensity score modelling generalizes to the MSCM context.

Various MSCM data generating algorithms are proposed in the literature. We used one of them [56] to assess the performance of the MSCMs. To make the data generating process more realistic for many disease settings, we included an interaction term between previous treatment status and current state of the confounder in deciding the next period treatment assignment. This data generating process was used to assess three settings: (i) large and (ii) small sample sizes in the rare event scenario and (iii) large

sample sizes when the event rate is more frequent. We estimated the causal effect parameter ψ_1 with a MSCM. As we know the values of the parameters generating the data, we can assess the performance of the IPW estimating methods under consideration.

We also evaluated the performance of the ad-hoc variability reduction techniques for IPWs (such as stabilization, normalization, truncation and their combinations). We found that normalized weights perform better than unnormalized weights. When applied to unstabilized weights, the improvement is noticeable in terms of bias, SD, MSE and confidence interval coverage. However, application of normalization to stabilized weights did not have much impact. This is the case even when truncation is applied. Depending on how many risk-sets are present in the study, the increased computational burden due to use of normalization might not be justified given the small gain. However, when application of unstabilized weights is desired [115, 162], normalization might be useful. A small level of truncation might also be helpful in such scenarios [128].

Among the methods under consideration, boosting estimates were associated with the least bias under the assumption of rare events. In the rare event scenario, stabilized IPWs estimated from boosting marginally outperformed the stabilized IPWs generated from logistic regression in terms of bias, MSE and coverage probability of ψ_1 estimation. However, when the event rate is more frequent, no such advantage is apparent. Compared to any other methods under consideration, MSCM estimates computed using the boosting approach were always closest to those from logistic regression, but the computational burden associated with the boosting approach is much higher.

Bagging and SVM did not perform very well in our simulation context. One reason could be that estimation of the weights generally requires estimation of probability of class membership (i.e., estimating the probability of being treated versus not). Bagging may not estimate the probability of class

membership well and may result in boundary probabilities (i.e., close to 0 or 1) more often than expected [163]. SVM approaches attempt to find the optimal dividing hyperplane [164] instead of explicitly modelling the probability of getting treatment. Such properties might make these statistical learning approaches less desirable for estimating IPWs [135].

We implemented all these IPW estimation methods in a MS dataset to show the applicability of the proposed methods in practice. We estimated the effect of β-IFN on irreversible disability progression using MSCM with stabilized normalized IPW. Except for the bagging approach, hazards ratio estimates from the different IPW adjusted analyses are fairly similar. As expected from our simulation results, HR estimates using IPWs estimated from the boosting approach were more precisely determined than those from logistic regression. Interestingly, SVM performed similarly to the boosting method. In this application, bagging failed to take into account of time-dependent confounder adequately, as the resulting estimates were similar to that obtained from the baseline-adjusted analysis (where time-dependent confounding was not controlled). This was the situation under all the increased levels of truncation considered. The results from logistic regression, SVM and boosting are consistent with the previously reported estimates [128]. None of the methods resulted in a significant effect estimate.

This study has several limitations. In this simulation, the data generation algorithm used linearity in the logit specification and also the logistic regression model was correctly specified during the simulated data analysis. For these reasons, it is not surprising that the logistic regression model performs well. However, despite these advantages favoring the logistic regression, the better performance of the boosting approach in the rare disease settings shows the utility of using this method. Statistical learning methods generally work well with high dimensional data. In this simulation, the number of covariates considered for adjustment is limited. Baseline covariates are not used in the simulation and the time-dependent confounder under consideration is a binary variable. The treatment variable under con-

sideration is binary. When we conducted the data analysis where baseline confounders were present, the results from SVM seem to be close to those from logistic regression whereas bagging did not perform as well as the simulation suggests, especially at the untruncated and lower truncation levels. Further studies are required to assess the behaviour of these methods in the presence of baseline covariates. While estimating the IPWs, better covariate balance in the propensity scores in the point-treatment studies motivated us to use the statistical learning approaches in the MSCM context. However, it is not well established in the published literature how to generalize such balancing criteria when there are many time points in a longitudinal study and when multiple time-dependent covariates are present [129]. Future studies could investigate this issue. To reduce the computational burden, we utilized robust standard errors for $\hat{\psi}_1$ [57]. However, resampling methods, such as the bootstap [165] may provide more reliable estimates of the standard error [55, 166]. The performance of these methods may be further enhanced after fine tuning of statistical learning parameters to obtain better fits. In this study, we mostly relied on the default settings offered by off-the-shelf statistical software packages which are freely available to the general researchers and epidemiologists. Future research could explore other statistical learning methods, such as neural networks and random forests [150].

Chapter 4

Comparison of Statistical Approaches Dealing with Immortal Time Bias in Drug Effectiveness Studies

4.1 Introduction

A goal of causal inference is to design an analysis plan for observational study data to emulate the conditions of a randomized clinical trial. In the absence of time-dependent covariates, this means making the subjects from different treatment groups comparable at baseline (i.e., the start of follow-up or the cohort entry time point). When dealing with survival data, survival outcome can be modelled by treatment groups, conditional on the baseline covariates, via the Cox proportional hazards model. Under the identifiability conditions (conditional exchangeability, positivity and consistency), the resulting hazard ratio estimate can be given a causal interpretation [55, 75].

In many observational drug effectiveness studies, there may be a delay or wait period before a subject begins to receive a treatment. Therefore, a subject's treatment status recorded at baseline may not be accurate for the entire duration of follow-up. Epidemiologists refer to this wait period during which the survival outcome cannot occur, in part due to the study design, as 'immortal time'. If the subject develops the event shortly after baseline, he may not get the opportunity to initiate the treatment and by design, he

is assigned to the untreated group. Failure to adjust for the change in treatment status, therefore, results in a spurious survival advantage (protective association) in favour of the treated group. This bias is popularly known as immortal (or 'immune') time bias [78, 167], time-dependent bias [168] or survival bias [82]. This bias can considerably distort the underlying hazard ratio if a large number of failures occur before the initiation of treatment or if the length of immortal time is large [169].

Although this bias was first identified in the 1970's [170], many pharmacoepidemiology studies still fail to account for this source of bias [79, 168, 171]. Recently, this issue of immortal time bias has resurfaced in pharmacoepidemiology studies while trying to implement newly popularized causal inference tools such as propensity scores (see §B.1 for a brief description). While deriving propensity scores, treatment group memberships need to be defined at baseline [172]. Some studies choose to address or at least acknowledge this immortal time bias problem by using simplified techniques such as selecting an alternative baseline unique for all subjects that makes sense clinically or excluding the subjects with wait times [173–175] or by rather complicated techniques such as risk-set matching [176–178], while others choose to ignore it completely [91, 93]. While some argue that the propensity score is an acceptable tool to deal with immortal time bias [179, 180], others are skeptical about this claim [93, 181]. To account for the immortal time bias, statisticians generally recommend adopting a proper treatment exposure definition via time-dependent analyses [78], such as use of time-dependent Cox proportional hazards models. However, as the findings from these models are expressed in terms of person-time under treatment exposure rather than in terms of treatment groups, interpretations are often not as intuitive as for a group-based comparison. Also, assumptions related to these analyses, such as treatment initiation being unrelated to the risk of subsequent failure, may be unrealistic for some situations [182]. Several other approaches have been proposed in the literature which modify the data so as to retain the treatment group-level interpretation. Prescription time-distribution matching (PTDM) [82] is one suggested approach to ad-

just for immortal time bias that is cited frequently in the recent literature [84–88] due to its simplicity.

In longitudinal studies, treatment may not be the only influential variable that may change after baseline. It is natural to have regular measurements of clinical symptoms and disease activity, and the values of these covariates may change over time. Since the predictive ability of baseline covariates may decrease over the follow-up time, considering the full covariate history of these time-dependent covariates, rather than just the baseline covariates may be preferable [183]. If these covariates do not interact with treatment exposure, a time-dependent Cox model may still be adequate to obtain an unbiased estimate of the treatment effect. However, if these covariates are affected by previous treatment (such covariates are popularly known as time-dependent confounders), the estimated hazard ratio may be biased if the time-dependent confounders are included as covariates in the time-dependent Cox model analysis [50]. In the presence of time-dependent confounding and immortal time, marginal structural Cox models (MSCM) are frequently used to estimate the causal effect of a time-dependent treatment exposure [42, 181]. MSCM is basically an extension of the propensity score methods that appropriately accommodates treatments of a time-varying nature in observational longitudinal studies with survival outcome [161] (see Chapter 2). The sequential Cox approach [73] has been proposed as an alternative to the MSCM approach. Both approaches are intended to deal with immortal time and the initiation of treatment after baseline.

Several studies have quantified the amount and direction of bias due to misclassifying or ignoring immortal time by means of simulation [184–187] or theoretically [188–190]. Via simulation of various disease contexts, it was showed repeatedly that overly optimistic estimates of treatment effects are obtained when the time-dependent nature of treatment is ignored [184, 186, 187]. In the intensive care unit context, a further simulation study showed that time-fixed analytic approaches are not generally equipped to deal with the time-dependent covariates [185]. Particularly, the landmark

method, a time-fixed analytic approach, was shown to adequately control for immortal time bias only when outcome occurs soon after initiating the treatment [188]. The nature of time-dependent bias was investigated mathematically for survival models in some studies [189, 190]. Such bias was theoretically quantified under some parametric assumptions depending on various cohort definitions [80]. To the best of our knowledge, no attempt has been made to explore the appropriateness of the PTDM or sequential Cox approaches in minimizing immortal time bias.

The focus of this chapter is to assess the performance of these proposed methods for dealing with immortal time bias. To do this, we quantify the bias due to PTDM based on the expressions derived in Suissa [80] for two other naive approaches. We also simulate survival data with time-dependent treatment exposure. Three different conditions are considered for simulation: (1) one baseline covariate present, (2) one time-dependent covariate present along with a baseline covariate and (3) one time-dependent confounder present. To assess the suitability of these methods in an application, we apply all these methods to investigate the impact of time-varying beta-interferon treatment in delaying disability progression in subjects from the British Columbia Multiple Sclerosis (MS) database (1995-2008) [92, 128].

The remainder of the chapter is organized as follows. In the next section, we describe the notation and design of the simulation study, the methods used to address immortal time bias, and the metrics used to evaluate their performances. Then we summarize the simulation and the MS data analysis results. The chapter concludes with a discussion of the results, and the implications and limitations of the current study.

4.2 Methods

4.2.1 Notation

Consider a hypothetical longitudinal study consisting of n subjects ($i = 1, 2, \ldots, n$). Let $t_0 = 0$ be the start of follow-up or the time of the baseline visit. Baseline covariates L_0 (binary or continuous) are recorded at baseline. Follow-up continues till the time of failure T or the time of censoring T^C. Regular measurements of the binary treatment status A_m (= 1 for treated and 0 otherwise), are recorded at intervals $m = 0, 1, 2, \ldots, K$. As this study is focusing mainly on the implications of immortal time, we assume that the subjects may initiate treatment at most once and continue taking the treatment thereafter till the study ends. Let treatment initiation occur at time T^A.

Let N_m be the number of failures occurring up to and including the m-th interval $[t_m, t_{m+1})$. Also, let C_m (= 1 for censoring due to dropout or artificial censoring and 0 otherwise) be the binary indicator of censoring during the m-th interval. Finally, let r_m be the risk-set consisting of subjects who are at risk of failure during the m-th interval $[t_m, t_{m+1})$. Let $\bar{a}_m = (a_0, a_1, \ldots, a_m)$ be the observed realizations of treatment history $\bar{A}_m$ up to interval m, and similarly, let $\bar{l}_m$ and $\bar{c}_m$ be the observed realizations of covariate histories $\bar{L}_m$ and censoring histories $\bar{C}_m$ up to interval m respectively. The binary indicator of failure by time t_{m+1} is defined as $Y_{m+1} = I(T \leq t_{m+1})$.

4.2.2 Analysis Approaches

In a simplified drug-effectiveness analysis, we can divide the subjects into two groups: the 'ever-treatment exposed group' consisting of the subjects who were exposed to the treatment at some point during their follow-up, and the 'never-treatment exposed group' consisting of the subjects who were never exposed to the treatment during their follow-up.

For comparison purposes, we will include two naive Cox models with time-independent treatment definitions: (1) unexposed time is misclassified as exposed time for the subjects in the ever-treatment exposed group, (2) unexposed time is excluded from the follow-up of the ever-treatment exposed group and the treatment initiation time is treated as time zero for these subjects. To address immortal time, we then apply the following approaches: (3) time-dependent Cox model with time-dependent treatment and time-dependent covariates, (4) time-dependent Cox model with time-dependent treatment and baseline values of the covariates, (5) MSCM, (6) PTDM, (7) sequential Cox approach.

Brief characteristics of the analysis approaches are shown in Table 4.1. We describe these methods in detail in the following sections using the notation defined above.

Naive Cox Model with Time-independent Treatment Definition

To demonstrate the impact of misclassifying treatment exposure by ignoring immortal time, two naive Cox analyses with time-independent treatment definitions are used to estimate the log-hazard (or log-hazard ratio). In the first approach, subjects in the ever-treatment exposed group are classified as treated for their whole duration of follow-up. This is similar to intention-to-treat principle [191] where subjects are assumed exposed to treatment immediately at the beginning of follow-up. We call this approach 'include immortal time' hereafter. Then we fit a time-invariant Cox model while adjusting for the potential baseline confounders.

In the second approach, the immortal time, i.e., time from cohort entry to the initiation of treatment, is excluded from the follow-up of the ever-treatment exposed subjects and time zero for these subjects is taken to be the time of treatment initiation T^A. However, the follow-up period for the never-treatment exposed subjects remains the same, i.e., time zero is the time of cohort entry $t_0 = 0$. We call this approach 'exclude immortal time'

hereafter. We fit a time-invariant Cox model while adjusting for the potential confounders measured at original baseline. More details about these two approaches are available in the Appendix §C.1.

Time-dependent Cox Model with Both Treatment and Covariates being Time-dependent

To avoid the difficulties related to the immortal time, statisticians frequently suggest using the time-dependent Cox model incorporating the entire treatment history [80, 82, 184, 192]. If we only consider the baseline covariates L_0, the hazard function can be expressed as the following time-dependent Cox model:

$$\lambda_T(m|L_0) = \lambda_0(m) \exp\big(\psi_1 A_m + \psi_2 L_0\big), \tag{4.1}$$

where m is the visit index, $\lambda_0(m)$ is the unspecified baseline hazard function, ψ_1 is the log HR of the time-dependent treatment status (A_m) and ψ_2 is the vector of log-hazard for the baseline covariates L_0.

To increase accuracy of the results, researchers also suggest incorporating the entire history of the relevant time-dependent covariates in the analysis [193]. In the presence of time-dependent covariates L_m (binary or continuous), equation (4.1) can be modified to:

$$\lambda_T(m|L_0, L_m) = \lambda_0(m) \exp\big(\psi_1 A_m + \psi_2 L_0 + \psi_3 L_m\big), \tag{4.2}$$

where ψ_3 is the vector of log HRs for the time-dependent covariates L_m.

Time-dependent Cox Model with Time-dependent Treatment and Baseline Covariates

We also include another time-dependent Cox analysis based on a time-dependent treatment definition. However, the history of the time-dependent covariates is restricted to the baseline values only (denoted as L_0', which ex-

cludes the post-baseline values of L_m). This is to quantify the impact of ignoring post-baseline changes in covariates. We call this the 'full cohort (base)' analysis. The hazard function is modelled as:

$$\lambda_T(m|L_0, L_0') = \lambda_0(m) \exp\left(\psi_1 A_m + \psi_2 L_0 + \psi_3' L_0'\right),$$

where ψ_3' is the vector of log HRs for the values of the time-dependent covariate at baseline, L_0'.

Marginal structural Cox model

We have already discussed this model in §2.2.2 and §3.2 so we include only a brief description here. If the time-dependent covariate L_m is influenced by past exposure, i.e., if L_m is a time-dependent confounder, playing a dual role as a confounder and an intermediate variable in the causal pathway between treatment and outcome, ψ_1 estimated from equation (4.2) may be biased [50]. Researchers need to be cautious about what covariates they include in the regression equation as covariates [194]. Instead of using L_m as a covariate in a Cox model, L_m is used to calculate the inverse probability weights (IPW) that are person-time specific measures of the degree to which L_m confounds the treatment selection process. These IPWs are then used to create the pseudo-population which will be free from the confounding effects of L_m. MSCM enables the conceptual comparison of the hazard functions for those subjects who were never exposed to treatment with those who were continuously exposed.

Stabilized IPW, sw_m, can be obtained by multiplying stabilized inverse probability of treatment weights (IPTW), sw_m^T, by stabilized inverse probability of censoring weights (IPCW), sw_m^C [42], where

$$sw_m^T = \prod_{j=0}^{m} \frac{pr(A_j = a_j|\bar{A}_{j-1} = \bar{a}_{j-1}, L_0 = l_0)}{pr(A_j = a_j|\bar{A}_{j-1} = \bar{a}_{j-1}, L_0 = l_0, \bar{L}_j = \bar{l}_j)}, \qquad (4.3)$$

and

$$sw_m^C = \prod_{j=0}^{m} \frac{pr(C_j = 0|\bar{C}_{j-1} = 0, \bar{A}_{j-1} = \bar{a}_{j-1}, L_0 = l_0)}{pr(C_j = 0|\bar{C}_{j-1} = 0, \bar{A}_{j-1} = \bar{a}_{j-1}, L_0 = l_0, \bar{L}_{j-1} = \bar{l}_{j-1})} \quad (4.4)$$

The weights sw_m are used in the time-dependent Cox model with hazard function modelled as in equation (4.1) to weight the contribution of each person-time observation so that the confounding due to L_{im} is removed. Note that IPCW is used only if non-random censoring is present. As discussed in §2.2.2, when the numerators in equations (4.3) and (4.4) are replaced by 1, these become the unstabilized IPW, w_m.

Prescription Time-Distribution Matching Approach

Although time-dependent Cox models are suitable tools for dealing with immortal time bias, these models do not offer treatment group-based interpretations as does the standard Cox model with fixed exposures. Researchers often resort to even simpler methodologies to deal with immortal time, such as the PTDM approach [82].

The essence of this approach is to redefine time zero in both the ever-treatment exposed and the never-treatment exposed groups. This is done by shifting the start of follow-up to the time of treatment initiation (the end of the immortal time period) T^A for the ever-treatment exposed subjects. The immortal time (wait) periods for the ever-treatment exposed subjects are randomly (with replacement) assigned to the never-treatment exposed subjects. The never-treatment exposed subjects who failed within their assigned wait period are excluded from further analysis. The analysis is performed based on the new time zeros, i.e., the newly defined baseline after excluding the observed or assigned immortal time from the follow-up for the ever and never-treatment exposed groups respectively. This eliminates imbalance in the excluded time distribution between the two treatment groups. Note that the random assignment of immortal time to the never-treatment exposed subjects and the subsequent exclusion of the subjects if they fail

within the assigned immortal period makes the data restructuring process random and the hazard ratio obtained from a different random assignment may be different. Further illustration and theoretical assessment of this approach is provided in Appendix §C.2.

Sequential Cox Approach

Let $[t_m, t_{m+1})$ denote the m-th interval where at least one subject initiates treatment. We want to mimic a clinical trial setting (e.g., either on treatment or off treatment during the entire duration of the follow-up) for each of these intervals where subjects initiate treatment. Based on the treatment initiation at the m-th interval, the m-th mini-trial is created as follows: only subjects who have not received any treatment before the m-th interval are considered. Among the subjects at-risk at t_m (i.e., those who have not failed or been lost to follow-up by the beginning of the m-th interval, t_m), the subjects initiating treatment during the m-th interval ($t_m < T^A \leq t_{m+1}$) are considered as the treated group, while the remaining subjects are considered as the control group. These control subjects are artificially censored at the time of later treatment initiation ($T^A > t_{m+1}$) to avoid confounding due to treatment. As these subjects are artificially censored, the analysis needs to be adjusted using IPCW. Note that if we consider 'month' as the interval unit for follow-up, there may be some intervals (i.e., months) during follow-up when no subjects initiate treatment.

In this mimicked trial, a subject is either on treatment or off treatment during the entire duration of the follow-up. Therefore, this manipulated subset of the data mimics a randomized clinical trial. A Cox proportional hazards model can be used to compare the survival experiences of these two groups. The relevant time-dependent covariate information is updated depending on the interval. In the analysis, we adjust for the baseline confounders L_0 measured at inclusion or baseline, the time-dependent covariates L_m measured at the start of m-th interval and the lagged covariates L_{m-1} consisting of the lagged value from the previous interval; this will help us

reduce bias in the estimation of the treatment effect from the m-th mini-trial data [73]. Let us denote $\tilde{L}_m = (L_0, L_{m-1}, L_m)$.

After treatment initiation, time-dependent covariate values after the m-th interval may be affected by the treatment and hence those covariate values after the m-th interval are not used in the analysis of the data for the m-th mini-trial [73]. If L_m are time-dependent confounding covariates, they are not included in $\tilde{L}_m$ as they are affected by the treatment [195, p.23].

We assume that the different mini-trials may have different baseline hazard functions but all subjects in the same mini-trial will have the same baseline hazard function. Under this assumption, a stratified Cox model is appropriate. Therefore, the hazard function for the m-th mini-trial can be written as [73]:

$$\lambda_T^m(m|L_0, \tilde{L}_m) = \lambda_{0m}(m) \exp\left(\psi_1 A_m + \psi_2' \tilde{L}_m\right) \tag{4.5}$$

where $\lambda_{0m}(m)$ is the unspecified baseline hazard function for stratum m, ψ_2' is the vector of log HRs for the time-dependent covariates $\tilde{L}_m$. This hazard function should be weighted by IPCW given in equation (4.4). Pooled logistic regression [42, 50] or Aalen's additive regression model can be used to estimate the IPCW [73, 196]. The resulting estimate will bear a causal interpretation under the assumptions of no unmeasured confounders and correct model specification for the hazard ratio and the censoring weights.

We can fit a stratified Cox model on the combined data of all mini-trials (pseudo-data), stratified by the treatment initiation time. Alternatively, a simple Cox model weighted by IPCW can be run for each of the successive mini-trials to obtain separate estimates of the treatment effect for each mini-trial, leading to the name, the sequential Cox approach. An overall estimate of the treatment effect is obtained by simply averaging the treatment effect estimates from the separate mini-trials. The overall estimate requires two additional assumptions for causal interpretation: (1) the treatment effect

is the same in all the mini-trials and (2) the treatment effect is unchanged for all covariate histories before the m-th interval, given the covariates at the m-th interval. However, if one is willing to interpret the overall effect estimate as an aggregated (averaged) effect over all the mini-trials, then the first assumption can be relaxed [73, 75].

The IPCW adjusted stratified Cox model used in the sequential Cox approach is easy to implement using standard software packages. The IPTW, the potentially unstable part of the IPW, are not used in the sequential Cox approach [73]. As the data associated with a given mini-trial can be extracted and separated quite easily from the combined mini-trial data (pseudo-data), it is also straightforward to compare the effects of early versus late treatment initiation. However, the combined mini-trial (pseudo) dataset can become large due to repeated use of the same control subjects. While inclusion of the same subject more than once may increase event rates, the SE obtained from the stratified weighted Cox analysis is invalid. Time consuming resampling methods, such as the jackknife [73] or the bootstrap [196], are required to obtain a correct SE. An illustrative data construction example is provided in Appendix §C.3 and the corresponding software implementation details are provided in Appendix §C.4.

Table 4.1: Description of the analytic methods.

Data-modify method	Method	Time-dependent Cox	Stratified	Covariate history	Weight adjusted
(1) Include IT	Cox PH	No	No	Baseline	No
(2) Exclude IT	Cox PH	No	No	Baseline	No
(3) Full cohort	Cox PH	Yes	No	Full	No
(4) Full cohort (Base)	Cox PH	Yes	No	Baseline	No
(5) MSCM	Cox PH	Yes	No	Full	Yes, IPTC
(6) PTDM	Cox PH	No	No	Baseline	No
(7) Sequential Cox	Cox PH	No	Yes	Multiple[†]	Yes, IPC

IT, Immortal time; PTDM, Prescription time distribution matching; MSCM, Marginal structural Cox model; IPT, Inverse probability of treatment; IPC, Inverse probability of censoring; IPTC, Inverse probability of treatment and censoring.

[†] For sequential Cox approach, covariate values are collected at three time points for each mini-trial: at baseline, at the period of treatment start and the lagged value at treatment start.

4.2.3 Design of Simulation

A number of schemes for simulating survival data for Cox models are available in the literature. Some generate survival times for the Cox models with time-invariant covariate by inverting cumulative hazards functions from commonly used survival distributions [197, 198]. A scheme for generating survival times for more complicated distributions such as the truncated piecewise exponential is also available [199]. These schemes have been extended to the situation with one [200–202] or more time-varying covariates [203].

To simulate survival times with or without time-dependent covariates, we adapt the permutation algorithm [204]. This algorithm simulates survival data following specified distributions of survival time conditional on any number of fixed or time-dependent covariates. In this algorithm, a permutation probability law based on the Cox model partial likelihood [83] is used as a basis for performing matching as follows. If a subject with a given set of covariates remains at risk until interval m, then the probability of that subject reaching the outcome at interval m is proportional to the subject's current hazard. This algorithm has been validated for generating survival times conditional on time-dependent treatment [205] and also when time-dependent covariates are present [183]. This algorithm has been used in several other studies dealing with generating survival data with time-dependent covariates (see for example [206–210]). A brief description of the algorithm is presented in Appendix §C.5.

A number of different simulation schemes are available in the literature to simulate survival times in the presence of a time-dependent confounder [56, 57, 102, 103, 144, 159, 160]. We adopt the data generation process of Young et al. [56] (also used in Chapter 3: see §3.3) where both treatment status and confounder are time-dependent. Data generated from this algorithm are popularly used to assess the ability of MSCMs to handle time-dependent confounders [57, 114].

4.2.4 Simulation Specifications

In our Monte Carlo study, we will generate $N = 1,000$ datasets with $n = 2,500$ subjects, each followed for up to $m = 10$ subsequent visits for each setting under consideration. For mimicking the rare disease condition, we set $\lambda_0 = 0.01$ (on a monthly scale). For mimicking a more frequent disease condition, we set $\lambda_0 = 0.10$ (on a monthly scale) and repeat the Monte Carlo study. Below we discuss the simulation schemes under consideration and the specifications that were used.

Simulation - I

We assume an exponential distribution for generating failure times T with constant $\lambda_0 = 0.01$ rate of monthly events throughout the follow-up. The exponential distribution is the simplest of all commonly used survival time distributions. However, despite its simplicity, it is often considered useful in biomedical research and is the basis for various frequently used approaches, such as, the Poisson model [186]. Therefore, we choose the exponential as the marginal distribution for generating event times. An uniform distribution $U(1, 60)$ months is assumed to generate censoring times T^C; i.e., administrative censoring is set at 5 years of follow-up. This marginal distribution of censoring time is independent of treatment exposure, as well as the failure times that were generated earlier. The accuracy of the chosen algorithm decreases with increasing rates of censoring [183] and hence we chose 60 months as the administrative censoring time-point i.e., the high upper limit of the uniform distribution. Treatment initiation time T^A is generated from an uniform distribution $U(0, 10)$ (in months). We assume treatment to be a binary variable for all subjects. This implies that the treatment has a constant impact on the hazard (multiple versions of the treatment is not acceptable); otherwise no treatment was assigned. Also, to focus on the immortal time issue, we assumed that there are no discontinuations or interruptions for those who initiate treatment. Additionally, we consider sex as a baseline confounder in these data. Subject's sex is generated based on a Bernoulli distribution where the probability of being male is 0.3. This co-

variate is also generated independent of time-dependent treatment exposure.

After generating values for the survival time T_i, the censoring time T_i^C, and the treatment and covariate matrix $X_{im} = (A_{im}, L_{i0})$ for each subject $i = 1, 2, \ldots, n$ for up to $m = 10$ time periods, the permutation algorithm [204] is used to generate survival data where treatment A_m is time-dependent but the confounder L_0 is fixed at baseline value. The effect parameters for treatment and sex on the survival outcome are set such that the treatment has a harmful effect (a log-hazard of $\psi_1 = 0.5$) and males are at a lower risk than females (a log-hazard of $\psi_2 = -0.7$). Here, the treatment having harmful effect means that a subject's survival time is shorter when she is treated compared to her survival time when she is untreated.

Simulation - II

To generate the survival times, we use exactly the same specification used in simulation - I, with the exception that we now add one time-dependent covariate, say cumulative disease activity L_m, such that higher cumulative disease activity has a higher risk (a log-hazard of $\psi_3 = \log(1.5)$). This time-dependent covariate L_m is generated based on a Bernoulli distribution with probability of disease activity increment being 0.75, accumulating the disease activity over at most $m = 10$ periods of time. As before, sex is a baseline confounder (L_0).

Simulation - III

We use the algorithm for simulating survival times in the presence of a time-dependent confounder [56]. In this simulation, counterfactual failure time $T_{i\bar{0}}$'s are sampled from an exponential distribution, with constant $\lambda_0 = 0.01$ rate of monthly events throughout the follow-up, as discussed in §3.3. The binary time-dependent confounder, L_m, is modelled by the following covariates: a binary covariate $I(T_{\bar{0}} \leq c)$, previous treatment status A_{m-1},

and the lagged variable L_{m-1}:

$$\begin{aligned} logit(p_L) &= logit\ Pr(L_m = 1|A_{m-1}, L_{m-1}, Y_m = 0; \boldsymbol{\beta}) \\ &= \beta_0 + \beta_1 I(T_{\bar{0}} < c) + \beta_2 A_{m-1} + \beta_3 L_{m-1}, \end{aligned}$$

with associated parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3) = (\log(3/7),\ 2,\ log(1/2),\ \log(3/2))$, $c = 30$ and $Y_m = I(T \leq t_m)$ (as defined in § 4.2.1).

We model treatment status at each stage A_m with the factors symptom or current medical condition L_m, past symptom L_{m-1}, and previous treatment status A_{m-1} as

$$\begin{aligned} logit(p_A) &= logit\ Pr(A_m = 1|L_m, A_{m-1}, L_{m-1}, Y_m = 0; \boldsymbol{\alpha}) \\ &= \alpha_0 + \alpha_1 L_m + \alpha_2 L_{m-1} + \alpha_3 A_{m-1}, \end{aligned}$$

with associated parameters $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2, \alpha_3) = (\log(2/7),\ 1/2,\ 1/2,\ 10)$. Current treatment status A_m is made heavily dependent on the previous treatment status A_{m-1} by setting a high parameter value ($\alpha_3 = 10$). That way, we emulate the situation where subjects switch to treatment at most once and keep on using the treatment without much interruption or discontinuation. The true causal effect parameter is set to be $\psi_1 = 0.5$. A brief description of the three simulations under consideration is provided in Table 4.2.

To study the properties of these simulated data populations, we generated datasets with $n = 25,000$ subjects, each followed for up to $m = 10$ subsequent visits, based on each simulation setting. As mentioned before, in each of the Monte Carlo studies, we generated datasets with $n = 2,500$ subjects, each followed for up to $m = 10$ subsequent visits.

Table 4.2: Three simulation settings under consideration.

	Simulation - I	Simulation - II	Simulation - III
Algorithm	Abrahamowicz et al. [204]	Abrahamowicz et al. [204]	Young et al. [56]
Time-varying treatment	Yes	Yes	Yes
Baseline covariate	Yes	Yes	No
Time-varying covariate	No	Yes	No
Time-varying confounder	No	No	Yes

4.2.5 Analytic Models Used

Simulation-I Models

In the simulation setting-I, when estimating the treatment effect, the baseline covariate L_0 is included in all the models under consideration. The Cox model is used in all these approaches. In the 'include IT (immortal time)' approach, immortal time is mislabelled as treated and in the 'exclude IT' approach, immortal time is excluded from the analysis. PTDM excludes the observed and assigned wait times. In the sequential Cox approach, stratified Cox weighted by IPCW is used, adjusting for treatment status A_m and baseline L_0. In the absence of a time-dependent covariate, stabilized IPWs are not useful. Therefore, the corresponding unstabilized IPCW model is fitted using pooled logistic regression adjusting for A_m and L_0 to predict future censoring status.

Simulation-II Models

In the simulation setting-II, when estimating the treatment effect, the baseline covariate L_0 and time-dependent covariate L_m are included in all the models under consideration. The Cox model is used in all these approaches. In the 'full cohort' and MSCM approach, all post-baseline values of L_m are used. In all the other approaches (except sequential Cox), only the baseline

values of L_m (i.e., L_0') are used. In the sequential Cox approach, the baseline covariate L_0 and three values of L_m are used: one at cohort entry, another at mini-trial entry and the lagged value before the mini-trial entry (as discussed in §4.2.2). To create the IPCWs, pooled logistic models are used. In the stabilized IPCW model (equation 4.4), the numerator model adjusts for A_m and L_0, while the denominator model adjusts for A_m, L_0 and L_m. For the MSCM, the model adjusts for only L_0 to obtain the effect of A_m. The corresponding IPTW (equation 4.3) is modelled via a pooled logistic model. For the stabilized IPTWs, the numerator model adjusts for the time index, L_0 and lagged values of A_m, while the denominator model adjusts for L_m, L_0, and lagged values of L_m and A_m to predict future treatment status.

Simulation-III Models

In the simulation setting- III, all the approaches under consideration include the time-dependent confounder L_m when estimating the treatment effect. The Cox model is used in all these approaches. In the 'full cohort' and MSCM approach, all post-baseline values of L_m are used. In all the other approaches (except sequential Cox), only values of L_m at cohort entry are used. As discussed in § 4.2.2, in the sequential Cox approach, we discard the time-dependent confounder L_m. The unstabilized IPCW model adjusts for only A_m in the absence of a time-dependent confounder. To do a sensitivity analysis, we discarded the IPCW. In another sensitivity analysis, three values of L_m are used in the sequential Cox approach: one at cohort entry, another at mini-trial entry and the lagged value before the mini-trial entry (as discussed in §4.2.2 and as in simulation-II). To create the IPCWs (equation 4.4), a pooled logistic model is used. For the stabilized IPCWs, the numerator model adjusts for A_m, while the denominator model adjusts for A_m and L_m. The IPTWs (equation 4.3) for fitting the MSCM are modelled via a pooled logistic model. The stabilized IPTW numerator model adjusts for time index and lagged values of A_m, while the denominator model adjusts for L_m and lagged values of L_m and A_m to predict future treatment status.

Among the approaches used here, the sequential Cox approach and the MSCM rely on IPWs, but the other approaches do not use any propensity scores or other weight-based approach; rather they rely only on the regression-based estimation approach. The PTDM and the included and excluded immortal time approaches are not suitable for propensity score or weight adjustment as they either use the future treatment status of the subjects to define the treatment groups or lack a baseline that is uniformly unique for all the subjects. More details about these approaches are available in the Appendices C.1 and C.2.

4.2.6 Performance Metrics

We assessed the performance of the various approaches by the following measures:

- Bias $= \sum_{i=1}^{N}(\hat{\psi}_{1i} - \psi_1)/N$: The average difference between the true and $N = 1,000$ estimated parameters (log-hazard).
- SD $= \sqrt{\sum_{i=1}^{N}(\hat{\psi}_{1i} - \psi_1')^2/(N-1)}$ where $\psi_1' = \sum_{i=1}^{N}\hat{\psi}_{1i}/N$
- Model-based SE: The average of $N = 1,000$ estimated standard errors of the estimated causal effect.
- Coverage probabilities of model-based nominal 95% CIs: Proportion of $N = 1,000$ datasets in which the true parameter is contained in the nominal 95% CI.
- Power: For a level $\alpha = 0.05$ test of $H_0 : \psi_1 = 0$, the estimated power is the proportion of nominal p-values that are less than $\alpha = 0.05$.

4.3 Application in Multiple Sclerosis

We apply the methodologies described in this chapter in the British Columbia (BC) MS cohort study (1995-2008) described in §2.2.1. The dataset was used in previous studies [92, 101] (also in Chapters 2 and 3) to estimate the effect

of β-IFN on irreversible disease progression. As before, irreversible progression of disability is measured by sustained EDSS 6 which is confirmed after at least 150 days, with all subsequent EDSS scores being 6 or greater. The treatment definition is changed in this study to allow us to demonstrate the impact of the various immortal or immune time adjustment methods. Here, once the subjects initiate β-IFN, we assume they continue taking the drug without any discontinuation until they develop the survival outcome (time to irreversible progression of disability) or become censored.

4.3.1 Analytic Models Used

Potential baseline confounders L_0 include age, sex, disease duration and EDSS score. Also, we consider the cumulative number of relapses in the last 2 years (hereafter called 'cumulative relapses') as a time-dependent confounder L_m (justified in [128], Chapter 2). All the models under consideration adjust for the baseline confounders L_0 (age, sex, disease duration and EDSS score) and the time-dependent confounder L_m (cumulative relapses) when estimating the treatment effect. The Cox model is used in all these approaches. All post-baseline values of cumulative relapses (L_m) are used only in the 'full cohort' and MSCM approaches. In all the other approaches (except sequential Cox), the value of cumulative relapses at cohort entry is used. In the sequential Cox approach, three values of cumulative relapses are used: one at cohort entry, another at mini-trial entry and the lagged value before the mini-trial entry (discussed in § 4.2.2). To create the IPCW (equation 4.4), pooled logistic models are used. The stabilized IPCW numerator model adjusts for A_m and the baseline confounders L_0, while the denominator model adjusts for A_m, L_0 and L_m. For the MSCM model, we estimated the effect of A_m after adjusting for the potential baseline confounders. The corresponding IPTWs (equation 4.3) are modelled via a pooled logistic model. The stabilized IPTW numerator model adjusts for a restricted cubic spline of the follow-up time-index, baseline confounders L_0 and lagged values of A_m to predict future treatment status. The denominator model additionally adjusts for the current and lagged values of

cumulative relapses (L_m).

4.4 Simulation Results

4.4.1 Description of the Simulated Data

To describe the data generated from the three simulation settings, we generated datasets with a larger number of subjects ($25,000$) with up to 10 subsequent visits from each simulation algorithms. For our purposes, we need to generate data such that subjects generally switch from the 'not treated' state to the 'treated' state at most once. For simulation-I and II, there are no exceptions. However, the way simulation-III is generated allows a few exceptions (19 out of $25,000$) where there are discontinuations. However, the proportion of discontinuation in the simulation-III dataset is negligible (0.00076) and we do not expect any noticeable impact in the results due to this small number of exceptions. The characteristics of the treated, untreated and partially treated groups, their failure rates and average number of visits are listed in Table 4.3. Simulation-I and II are very similar with respect to the characteristics listed here.

Table 4.3: Characteristics of three simulation settings under consideration.

Rates	Simulation-I	Simulation-II	Simulation-III
Failure	0.084	0.084	0.143
Always treated	0.051	0.051	0.261
Never treated	0.152	0.150	0.046
Partially treated	0.797	0.799	0.692
Discontinuation	-	-	0.001
Mean visits	8.949	8.943	9.367

4.4.2 Rare Event Condition

We present the results from the rare event condition ($\lambda_0 = 0.01$ in a monthly time-scale) in the three simulation settings. When a time-dependent covariate or confounder is present, PTDM is not an appropriate analysis method. This method is only appropriate for analyzing simulation setting - I. We still show the results from this analysis in the other simulation settings for comparison purposes.

Results From Simulation-I

Results from simulation-I are reported in Table 4.4. The time-dependent Cox model with treatment status (A_m) and the baseline covariate (L_0) is fitted to assess the accuracy of the survival generating permutation algorithm. The level of bias of $\hat{\psi}_1$ is negligible (0.005), the average coverage probability of model-based nominal 95% CIs is 0.946 and the corresponding power is 0.879. We consider these results as the standard for comparison purposes for this simulation setting.

When the immortal time is misclassified as exposed time, we see a substantial downward bias (-2.799). The situation improves slightly when immortal time is excluded from the analysis (bias -2.214). Applying PTDM, the bias is further reduced (-1.837), but the estimate is still substantially off the target. Also, the variability of the estimator ($SD(\hat{\psi}_1)$) for PTDM is substantially larger than for the time-dependent Cox results.

When the sequential Cox approach is applied, the amount of bias is negligible (0.007), the average coverage probability of the model-based nominal 95% CIs is 0.949; both are comparable to the time-dependent Cox results. However, the power from this approach (0.755) is slightly lower and the SD of the $\hat{\psi}_1$ (0.196) is higher than that of the time-dependent Cox approach. As there is no time-dependent covariate in this simulation, the MSCM is not fitted.

Table 4.4: Comparison of the analytical approaches to adjust for immortal time bias from simulation-I (one baseline covariate and time-dependent treatment exposure) of 1,000 datasets, each containing 2,500 subjects followed for up to 10 time-intervals.

Approach	Bias	$SD(\hat{\psi}_1)$	$se(\hat{\psi}_1)$	CP	Power
Full cohort	0.005	0.167	0.162	0.946	0.879
Included IT	-2.799	0.143	0.141	0.000	1.000
Excluded IT	-2.214	0.143	0.142	0.000	1.000
PTDM	-1.837	0.198	0.198	0.000	0.999
Sequential Cox	0.007	0.196	0.187	0.949	0.755
MSCM	-	-	-	-	-

PTDM, Prescription time distribution matching; IT, Immortal time; MSCM, Marginal structural Cox model.

In all these approaches, the empirical standard errors $SD(\hat{\psi}_1)$ (SD of the estimated parameters) are reasonably close to the average model-based standard-error ($se(\hat{\psi}_1)$). A slight discrepancy is, however, apparent with the sequential Cox approach (0.196 and 0.187 respectively, reported in the Table 4.4). Here the average standard error $se(\hat{\psi}_1) = 0.187$ is obtained from the approximate jackknife approach (see Appendix C.4). We resort to the nonparametric bootstrap method to determine more reliable estimates of the standard error. For the 1,000 datasets under consideration, the bootstrap standard error of the sequential Cox approach based on 100 nonparametric bootstrap samples is 0.189. Since the bootstrap method requires a substantial amount of computing time and estimates from both of the methods (bootstrap and approximate jackknife approach) are close, we simply report the approximate jackknife standard error estimate from now on.

Results From Simulation-II

Results from simulation-II are reported in Table 4.5. The time-dependent Cox model with treatment status (A_m), baseline covariate (L_0) and time-

dependent covariate L_m is again fitted to validate the survival generating permutation algorithm. The level of bias is negligible (0.000), the average coverage probability of the model-based nominal 95% CIs is 0.952 and the corresponding power is 0.878. These results are again considered as the standard for comparison purposes for this simulation setting.

To examine the implications of not using the post-baseline changes in the time-dependent covariate, we use only the baseline values of the time-dependent covariate, while keeping the definition of time-dependent treatment unchanged from the previous analysis. This results in some bias (-0.179). However, when we simplify the treatment definition by misclassifying the immortal time as treated time, the bias is again substantial (-2.305). Excluding the immortal time or using PTDM results in little or no improvement in terms of bias (bias -2.321 and -1.952 respectively).

When the sequential Cox approach is used in this simulation setting, we observe some bias (0.268). Even though the bias is much less than with the PTDM method, the bias is still close to that obtained from the time-dependent Cox analysis that incorporates only the baseline covariate information.

We apply MSCM with L_m handled as a time-dependent confounder, even though it is not. The corresponding bias is negligible (-0.001), the average coverage probability of the model-based nominal 95% CIs is 0.952, and the power is 0.880. These results are very similar to those for the time-dependent Cox analysis using the full cohort.

Results From Simulation-III

Results from simulation-III are reported in Table 4.6. MSCM with treatment status (A_m) is fitted to validate the survival generating permutation algorithm. The corresponding stabilized weights are generated based on relationship between treatment status (A_m) and the time-dependent con-

Table 4.5: Comparison of the analytical approaches to adjust for immortal time bias from simulation-II (one baseline covariate, one time-dependent covariate and time-dependent treatment exposure) of 1,000 datasets, each containing 2,500 subjects followed for up to 10 time-intervals.

Approach	Bias	$SD(\hat{\psi}_1)$	$se(\hat{\psi}_1)$	CP	Power
Full cohort	0.000	0.164	0.162	0.952	0.878
Full cohort (Base)	-0.179	0.189	0.189	0.842	0.394
Included IT	-2.305	0.183	0.180	0.000	1.000
Excluded IT	-2.321	0.187	0.184	0.000	1.000
PTDM	-1.952	0.233	0.233	0.000	0.999
Sequential Cox	0.268	0.190	0.185	0.696	0.978
MSCM	-0.001	0.163	0.162	0.952	0.880

PTDM, Prescription time distribution matching; IT, Immortal time; MSCM, Marginal structural Cox model.

founder L_m. The level of bias is negligible (0.029), the average coverage probability of the model-based nominal 95% CIs is 0.942, and the corresponding power is 0.734. These results are now considered as the standard for comparison purposes for this simulation setting.

The time-dependent Cox models, both using full and baseline covariate information, result in biased estimates in the presence of this time-dependent confounder (0.438 and 0.188 respectively). The average coverage probability of the model-based nominal 95% CIs for the method using the full covariate history is very low (0.251). When the immortal time is misclassified, excluded or PTDM is used to analyze data, we still see substantial bias ($-2.190, -1.917$ and -1.553 respectively).

We apply the sequential Cox approach in three different ways. First we do the analysis excluding the time-dependent confounder L_m from the analysis. The amount of bias (0.721) is lower than with PTDM, but higher than with the time-dependent Cox analysis, as was seen in [195]. The second

analysis, a sensitivity analysis that does not use IPCW in the analysis, leads to similar bias (0.720). Finally, we perform the analysis by including L_m in the stratified IPCW weighted Cox model. The bias is reduced (0.474), but comparable to the bias in the time-dependent Cox analysis with full covariate information (0.438).

Table 4.6: Comparison of the analytical approaches to adjust for immortal time bias from simulation-III (one time-dependent confounder and time-dependent treatment exposure) of $1,000$ datasets, each containing $2,500$ subjects followed for up to 10 time-intervals.

Approach	Bias	$SD(\hat{\psi}_1)$	$se(\hat{\psi}_1)$	CP	Power
Full cohort	0.438	0.168	0.169	0.251	1.000
Full cohort (Base)	0.188	0.177	0.180	0.841	0.982
Included IT	-2.190	0.199	0.198	0.000	1.000
Excluded IT	-1.917	0.194	0.193	0.000	1.000
PTDM	-1.553	0.249	0.223	0.001	0.978
Sequential Cox[#]	0.721	0.266	0.257	0.188	0.999
Sequential Cox[†]	0.720	0.266	0.256	0.185	0.999
Sequential Cox[§]	0.474	0.272	0.263	0.578	0.969
MSCM	0.029	0.201	0.205	0.942	0.734

PTDM, Prescription time distribution matching; IT, Immortal time; MSCM, Marginal structural Cox model.

[#] As described in § 4.2.2.

[†] Sequential Cox not adjusting for either the time-dependent confounder or for informative censoring.

[§] Sequential Cox adjusting for both the time-dependent confounder in the regression for estimating ψ_1 and for informative censoring via IPCW.

4.4.3 When More Events are Available

Results from the more frequent event condition are presented in the Tables C.1-C.3 in Appendix §C.6. The trends in the bias are similar compared to those in the rare event condition. In general, in all simulation settings,

the standard errors are much less than in the corresponding analyses when failure rates are rare. Bias is slightly lower in some cases. One noticeable difference is observed in simulation setting - III: in the presence of the time-dependent confounder, when the failure rate is more frequent, the time-dependent Cox and MSCM approaches yield minimal bias (0.044 and 0.000 respectively). It is not clear why this is the case. As expected, the average coverage probability of the model-based nominal 95% CIs from the time-dependent Cox approach is smaller (0.888) than that of MSCM.

4.5 Results from Multiple Sclerosis Data Analysis

Table 4.7: Summary of the estimated parameters from the relapsing-onset multiple sclerosis (MS) patients' data from British Columbia, Canada (1995-2008).

Approach	$\hat{HR}$	$se(\hat{HR})$	95% CI	Weights	
				Average (log-SD)	range
Full cohort	1.29	0.23	0.91 - 1.83		
Full cohort (Base)	1.25	0.23	0.87 - 1.79		
Included IT	1.05	0.20	0.72 - 1.52		
Excluded IT	1.53	0.30	1.05 - 2.24		
PTDM	1.26	0.24	0.86 - 1.85		
Sequential Cox	1.14	0.29	0.69 - 1.89	1.00 (-4.06)	0.63 - 2.20
MSCM	1.31	0.23	0.92 - 1.84	1.00 (-2.86)	0.37 - 1.60

PTDM, Prescription time distribution matching.
The HR for the treatment is reported. The analyses are adjusted for baseline covariates sex, EDSS score, age and disease duration, and for the time-dependent confounder 'cumulative relapses'.

To focus on the impact of immortal time in this application, we assume that the subjects remain on β-IFN treatment once they initiate the treatment, as is assumed in previous pharmacoepidemiologic studies [73, 196, 211]. Appendix §A.6 describes the baseline characteristics of the MS cohort

under consideration. As justified in our previous study [128](see Chapter 2), we consider MSCM estimates to be ideal in this context. Results are reported in Table 4.7.

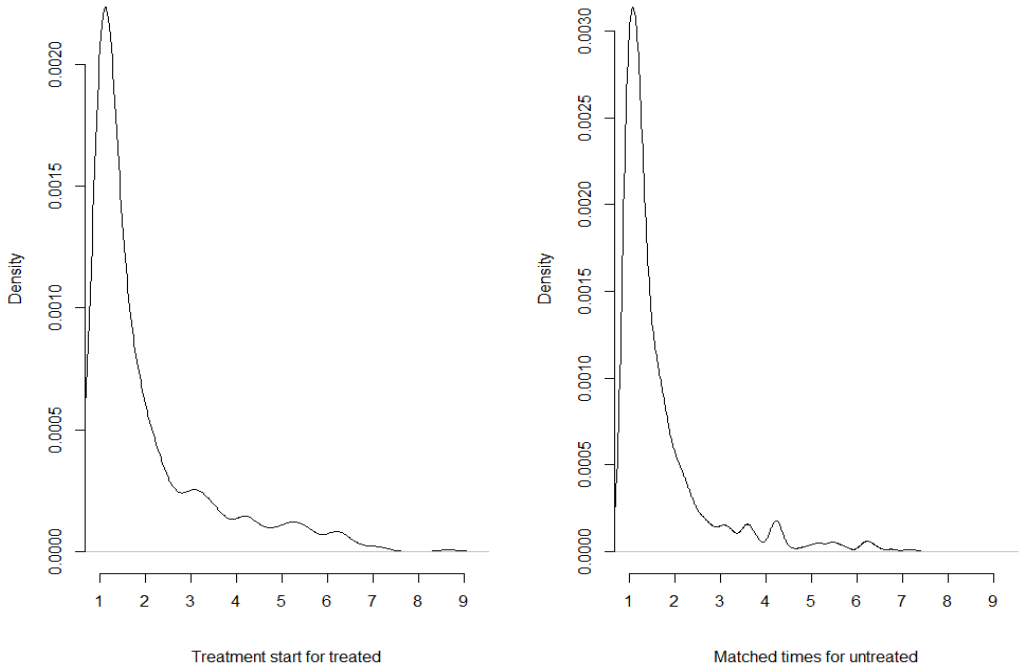


Figure 4.1: Matched wait periods (in years) from prescription time-distribution matching approach in the relapsing-onset multiple sclerosis (MS) cohort from British Columbia, Canada (1995-2008).

Wait periods (assigned for never-treatment exposed subjects and observed for ever-treatment exposed subjects) from the PTDM method are shown in Figure 4.1. As the PTDM approach produces different estimates from the same data based on random sampling of the immortal times, we estimate the HR from the MS data $1,000$ times and report the mean and SD of the estimated HR in Appendix Table C.4 in Appendix §C.7 as is

done in other studies involving random estimates [212]. The distribution of the estimated HR, depicted in Appendix Figure C.4 in Appendix §C.7.1, is moderately symmetric and the estimated HR is always above the null value of 1 but below 2.

The IPCW in the sequential Cox approach are less variable ($SD = 0.02$) than the IPW in MSCM (0.06). IPCW are estimated separately at each mini-trial data construction [73]. When they are instead estimated from the aggregated dataset [75], the HR estimate (1.11) is very close to the estimate (1.14) shown in Table 4.7 (see Appendix §C.7.2). Note that no matter how they are constructed, the IPCWs from the mini-trials are well-behaved, i.e., the averages are close to one and they have low variability (most are within the range 0.9 to 1.1 and the distributions are unimodal and symmetric; see Appendix Figures C.5 and C.6).

4.6 Discussion

Due to various practical considerations, researchers use observational survival studies to assess the impact of treatments. In such studies, in contrast to randomized clinical trials, subjects may not be exposed to the treatment at the beginning of follow-up. In longitudinal observational studies, treatment exposure in addition to other patient characteristics may change over time. Ignoring these time-varying characteristics may lead to inaccurate estimates, or possibly even to wrong conclusions being drawn. Statistical procedures, such as the time-dependent Cox model, are known to deal with time-dependent treatment and time-dependent covariate information. However, in the medical literature, it is not uncommon to see the time-independent Cox model based on only the baseline characteristics (i.e., treatment and covariate information measured at baseline) used in such circumstances, likely for the convenience of model fitting and group-based interpretation [193].

Perhaps of even greater concern, some studies employ future treatment status (who initiates treatment later in the follow-up) to classify subjects into the treatment groups [186, 189]. Comparisons between two such misclassified groups is prone to bias related to immortal time, due to the incorrect specification of risk-sets.

The Cox model assumes that treatment initiation is unrelated to the risk of subsequent failure. Such assumptions underlying the time-dependent Cox model may be untestable or difficult to assess in an epidemiological context [82]. The Cox model is sometimes considered as an oversimplified method to capture the observed process [182]. Alternative survival analysis methods, such as Poisson regression and pooled logistic regression, also suffer from the same bias when the definition of time zero for building risk-sets is not the same for all subjects [169]. Therefore, there is a need for methods that are capable of handling the time-dependent nature of longitudinal data, as well as helping us better understand the treatment-outcome mechanism so that the interpretations of the results become more appropriate.

To this end, we assess two methods that are proposed for the situation when treatment initiation occurs later than cohort entry: PTDM and sequential Cox. The appropriateness of these methods is not assessed in the literature. We design three increasingly difficult simulation settings to highlight the importance of accounting for time-dependent covariates in longitudinal studies. The first setting (simulation-I) is the simplest: only treatment initiation may be delayed and the covariate under consideration is time-fixed. In the second setting (simulation-II), we add a covariate that is time-dependent. The last setting (simulation-III) deals with the situation where a time-dependent confounder is present. PTDM and sequential Cox approaches are claimed to be appropriate analysis techniques for datasets generated from simulation settings I and III respectively. As the time-dependent Cox model is appropriate for simulation settings I and II, we use these results as the standard for comparison. For setting III, where a time-dependent confounder is present, a MSCM is the most popular and

appropriate method and hence results from this method are used as the standard for comparison in this simulation setting.

Downward bias (indicated by a negative sign in the log-hazard estimates) in the analyses ignoring immortal time ('exclude IT' approach) is consistent with previous simulation studies [184]. This indicates immortal time bias makes the treatment look more protective than it actually is. Even though the bias associated with exclusion of immortal time is generally less or equally severe than with misclassifying it ('include IT' approach), the bias is not negligible.

A widely accepted alternative to the time-dependent Cox model is PTDM [82]. Here, treatment exposure is converted into a time-independent variable so that a simple Cox model for treatment-group comparison can be applied. This conversion of time-dependent exposure into a time-independent exposure requires restructuring of the data using the PTDM approach. In this approach, new time zeros are defined after excluding the observed and assigned immortal times in the ever and never-treatment exposed groups. The excluded times (wait-periods) for the ever-treatment exposed and never-treatment exposed groups follow the same distribution. However, the baselines for all subjects are not exactly the same as in landmark analyses [213, 214]. It is not clear whether assigning random baselines will adequately address the immortal time bias. Ambiguous and inconsistent definition of the baseline time for different subjects in observational studies makes it hard to obtain an unbiased estimate of the treatment-outcome association due to entanglement of various sources of bias [215]. From the results of our simulation (simulation-I), we can see the bias is slightly less than when misclassifying or excluding immortal time. However, the bias is still substantial in this analysis (also see Appendix §C.2 for theoretical assessment), highlighting the value of setting a well-defined time zero or baseline.

The sequential Cox approach is an alternative method for estimating the treatment effect from more complex observational data settings where the

treatment is time-dependent and censoring may be non-random [73]. Especially in simulation-I, in the absence of the time-varying covariate or confounder, this method works very well in comparison to the time-dependent Cox. However, results from our simulation settings with a time-varying covariate or confounder are not as promising as claimed in the original paper [73]. MSCMs are generally more popular in dealing with time-dependent confounders. As MSCMs are extensions of time-definement Cox models, they are also used in addressing the immortal time bias [216]. Although the mechanisms and interpretations behind the sequential Cox approach and MSCM are different, both claim to have the same goal of estimating the causal effect of treatment in the presence of time-dependent confounders. However, from our simulation, we do not find the sequential Cox approach to be as effective as MSCM in the presence of a time-dependent confounder (simulation-III) or even when a time-dependent covariate which is not a time-dependent confounder (simulation-II) is present.

In simulation-III, we performed a sensitivity analysis of the sequential Cox approach without using IPC weights. This sensitivity analysis assesses the impact of artificial censoring induced in the analysis by censoring later treatment initiation cases. This analysis yielded very similar results (bias 0.720 compared to 0.721 in the original analysis). Another sensitivity analysis was performed that adjusts for the time-dependent confounder. This approach yields less bias (0.474), indicating the importance of adjusting for baseline values of the time-dependent confounder. As the time-dependent confounder values after the treatment initiation are discarded in the sequential Cox approach, it makes sense to control for the time-dependent confounder in the analysis. Even after such adjustment, the sequential Cox approach does not seem to remove the effects of time-dependent confounding, as is recently mentioned elsewhere [217]. Instead of using the full covariate history of the time-dependent covariate ($\bar{L}_m$), this approach only adjusts for a few values of the time-dependent covariates ($\tilde{L}_m = (L_0, L_{m-1}, L_m)$) as defined in the equation (4.5). This may limit the ability of this method to obtain unbiased estimates. Additionally, this approach cannot handle treat-

ment discontinuation or more than one treatment initiation [75].

On the other hand, in contrast to PTDM, the sequential Cox approach effectively removes the immortal time bias. The approach utilizes all subjects and is therefore more efficient than PTDM. It also handles different baselines properly by performing a stratified analysis. The focus is on recreating the covariate process at each treatment start using the mini-trial approach. Such focused and detailed scrutiny could yield insights about the data which may be hard to extract using a MSCM approach. Interpreting MSCM results remains a hurdle and an alternate view of the data may be helpful. Although IPTWs are avoided in the sequential Cox approach, we still need to use IPCWs. These weights are less variable and more stable than IPTW [73, 217] and appropriately handle artificial censoring caused by the censoring at later treatment start dates. However, similar to other artificial censoring correction methods, IPCW may not be an effective method in the presence of strong confounding and small sample size [218].

We apply the methods under consideration to estimate the effect of β-IFN on disease progression. The sequential Cox approach seems to have a downward bias (HR=1.14; 95% CI 0.69 - 1.89) compared to MSCM (HR=1.31; 95% CI 0.92 - 1.84). The PTDM results (HR=1.26; 95% CI 0.86 - 1.85) look closer to MSCM. However, the PTDM approach involves random assignment of wait periods for the never-treatment exposed subjects. When we repeat the analysis $1,000$ times and average the results, the estimated treatment effect (HR=1.44; 95% CI 0.97 - 2.11) looks too high with wider CI compared to that obtained from a MSCM. The random nature of the results may be an undesirable feature of this analysis. However, use of statistical methods that produce variable results is not uncommon in the epidemiologic literature [212, 219].

Similar to other simulation studies, we only investigate a handful of possible scenarios. However, the assumptions underlying the data simulation are consistent with patterns typical in epidemiologic observational survival

studies where treatment initiation may happen later for some subjects and associated covariates are measured regularly. Furthermore, our assumption of no discontinuations or interruptions in the treatment is restrictive and may not be suitable in some disease scenarios where subjects may choose different treatment strategies over the course of time. Other bias related to time-dependency, such as time-modified confounding [220], is not considered in this study. Substantial immortal time bias is induced by group-based Cox model analysis in the scenarios investigated. However, even in these extreme scenarios, methods are available that estimate the target parameter adequately with minimal bias.

One single approach may not be the most suitable to analyze all kinds of survival data. However, based on this study, we have gained considerable understanding about which approaches should be used depending on the nature of the disease mechanism. If assumptions behind the Cox model with time-dependent treatment (such as treatment assignment occurring a random times) are not reasonable in a given disease scenario, the sequential Cox approach can be used as a good alternative. However, when we need to consider post-baseline values of the time-dependent covariate to adequately model a disease process, then the sequential Cox approach is not the best alternative. In the presence of time-dependent confounders, MSCM is the best method to adjust for this type of confounding. Future research could focus on enhancing the sequential Cox approach to allow appropriate adjustments for time-dependent covariates and confounders. Future studies could also assess the impact of model misspecifications and measurement error under the same scheme used in this study.

Chapter 5

Conclusion

5.1 Summary of the Main Results

Estimating the treatment effect from drug effectiveness observational studies is challenging due to the existence of various kinds of biases. The idea behind causal inference is to design the statistical analysis of observational studies to mimic the conditions of a hypothetical randomized experiment. Such conditions will allow us to properly investigate well-formulated causal questions. This requires not only knowledge about the subject area (e.g., the condition or disease under study and associated drug exposure), but also familiarity with the statistical tools and techniques appropriate for such analyses.

Research on chronic diseases deals with multiple measurements of affected subjects over an extended follow-up period. During this time, key patient characteristics may change, including initiation or cessation of drug treatments. This means that straightforward adjustment for baseline confounders may not be adequate to answer a question about the effectiveness of a drug, where the exposure might occur months or years after 'baseline'. For example, longitudinal observational data are required to assess the impact of beta-interferon drug exposure on disease progression in relapsing-remitting multiple sclerosis (MS) patients in the 'real-world' clinical practice setting. Most commonly used causal inference tools, such as propensity scores, are not generally well-suited to deal with complex longitudinal patterns of such data, i.e., in the presence of immortal time bias and time-dependent confounding [21, ch.15]. Marginal structural Cox models (MSCMs) can be thought of as an extension of the propensity score tool that gained popular-

ity over the last decade. MSCMs provide distinct advantages over traditional approaches by allowing adjustment for time-varying confounders, such as relapses ('attacks') in our MS application, as well as baseline characteristics, through the use of inverse probability weighting (IPW). As MSCMs are extensions of the Cox model with time-dependent treatment exposure, they also allow adjustment for immortal time bias.

We assessed the suitability of MSCMs to analyze data from a large cohort of 1,697 relapsing-remitting MS patients in British Columbia, Canada (1995-2008) in Chapter 2. In the context of this observational study spanning over a decade and involving patients with a chronic, yet fluctuating disease, the recently proposed normalized stabilized weights were found to be the most appropriate choice of weights. Using these weights, no association was found between beta-interferon exposure and the hazard of disability progression (hazard ratio 1.36, 95% confidence interval $0.95, 1.94$). Additionally, findings did not change when truncated normalized unstabilized weights were used in further MSCMs and to construct IPW adjusted survival curves. Qualitatively similar conclusions from approximation approaches to the weighted Cox model (i.e., MSCM) extend confidence in the findings.

IPWs are at the heart of MSCMs. The properties of IPWs influence the estimated effects from MSCM and their accuracy. Logistic regressions are popularly used to model the IPWs. Statistical learning algorithms such as bagging, support vector machines and boosting have proved useful in generating well-balanced propensity scores. As propensity scores are used in the intermediate steps to construct IPWs, it is natural to investigate the utility of these approaches for modelling IPWs. We compared the performance of these proposed approaches in Chapter 3 using simulated survival data that mimicked a context in which both treatment status and a confounder were time-dependent. Proposed approaches are compared with respect to bias, standard error, MSE, and coverage probabilities of model-based nominal 95% confidence intervals under various weight variability reduction tech-

niques, such as normalization and increased levels of truncation. Under a rare event condition, the weights generated from boosting were found to be associated with less MSE and better coverage. Bagging and support vector machine did not perform well in this MSCM context. The study was repeated for the situation when events are more frequent and also with smaller numbers of subjects to observe the impact. In the smaller sample case, bias, variance and subsequently MSE were larger. When the event rate is more frequent, MSCM estimates computed using boosing approach were similar to those from logistic regression.

The findings from this simulation study guide an application of the MSCM to investigate the impact of beta-interferon treatment in delaying disability progression in subjects from the British Columbia Multiple Sclerosis database (1995-2008). When boosting is used to model the IPWs, MSCM estimates were similar to that obtained when IPWs are estimated from the logistic regression approach as in Chapter 2. Although the confidence interval was narrower, the conclusion remains the same (hazard ratio 1.32, 95% confidence interval $0.94, 1.86$).

In observational drug effectiveness survival studies, misclassification or exclusion of the period between cohort entry and first treatment exposure during the follow-up period may result in immortal time bias. This bias can be minimized by acknowledging a change in treatment exposure status with time-dependent analyses, such as fitting a time-dependent Cox model. Accounting for time-dependent variables in the analyses may be complex and the corresponding interpretations may not be intuitive. Furthermore, the assumptions of such an approach, such as treatment initiation being unrelated to the risk of subsequent failure, may be untestable or difficult to assess. Prescription time-distribution matching is an approach proposed in the literature to avoid the need for a time-dependent Cox analysis. In this method, the treatment initiation time distribution for the treated subjects is matched with a newly assigned baseline or time zero for untreated subjects, so that both treatment groups have a comparable time zero.

In longitudinal studies with a sequence of measurements, both treatment and the covariates under consideration may be time-dependent. Furthermore, the time-dependent covariates may be affected by the change of treatment status, i.e., time-dependent confounding may be present. MSCMs are usually used to deal with such confounding. However, these models are extensions of time-dependent Cox models and therefore fitting and interpretation of these models is also not straightforward. The sequential Cox approach is suggested as an alternative approach. This approach creates small cohorts based on each possible treatment start time and the overall treatment effect is estimated by averaging the estimated effects from all the created cohorts. Both the prescription time-distribution matching and the sequential Cox approaches break the time-dependent nature of the problem down into smaller pieces such that the findings potentially become accessible to a wider audience. In Chapter 4, we assess the suitability of both approaches for analyzing data in the absence and presence of a time-dependent confounder.

These approaches are applied to investigate the impact of beta-interferon treatment in delaying disability progression in the British Columbia Multiple Sclerosis cohort ($1995-2008$). Under the assumption that there were no treatment discontinuations, we found no convincing evidence that β-IFN reduces the hazard of disability progression with either approach (hazard ratio 1.26, 95% confidence interval $0.86, 1.85$ from the PTDM and hazard ratio 1.14, 95% confidence interval $0.69, 1.89$ from the sequential Cox approach).

5.2 Implications

Most of the MSCM analyses reported in the literature aim to model disease conditions specific to HIV/AIDS. We applied and adapted advances made in this field to better study another chronic disease, MS. In this work, we identified a time-dependent confounder using a causal diagram by incorporating subject-specific knowledge of how β-IFN treatment potentially impacts on

the disease process. We then translated the MS disease features using the MSCM framework to adjust for the time-dependent confounder.

Randomized clinical trials are not feasible ethically or practically over the long observation periods of chronic diseases such as MS. Such studies may also fail to reflect the 'real-world' clinical practice setting. Therefore, observational studies may provide invaluable information in the MS context. It is of considerable clinical importance to establish whether β-IFN has an effect in delaying long-term progression of the disease. Therefore, the aim was to estimate the effect of β-IFN treatment on the longer-term outcome of irreversible disability. Although this question was studied previously in various observational studies, results were seemingly inconsistent and contradictory. In this study, we show how the analysis should be appropriately done in the presence of a time-dependent confounder, such as MS relapses.

The present study is the first to examine the impact of beta-interferon drug exposure on disease progression in a MS cohort using a MSCM approach. This study took into account of the causal dynamics of the MS disease process over a long follow-up period and has made a notable contribution to the available evidence. The implication of this study goes far beyond the MS disease setting as it shows that normalized stabilized weights are useful for fitting MSCMs in order to study chronic disease conditions with an extended follow-up period. A large number of sensitivity analyses were carefully planned and carried out to check various assumptions, to validate the results in a restricted sub-population, and to assess the impact of covariate definitions used in the data analysis.

It has long been hypothesized that use of statistical learning techniques might improve the properties of IPWs as they were shown to improve the balance of propensity scores [135, 140, 141]. Using a simulation study and data analysis, this study is the first to investigate the utility of statistical learning methods such as bagging, SVM and boosting in creating weights. In particular, weights created using the boosting approach were shown to

improve the behaviour of IPWs and consequently provided a better estimate of the effect from MSCM with narrower confidence interval. The in-depth analysis also considered the impact of various weight variability reduction techniques, such as truncation and normalization, which reduce the variability of the IPWs.

Pharmacoepidemiological studies often suggest alternative techniques for the ease of analysis and interpretation of the results. For example, 'PTDM' was suggested as an alternative to the Cox model with time-dependent exposures to minimize immortal time bias. One set of data analyses found this method to be effective in controlling immortal time bias [82] in the sense that it provided similar results as fitting of a Cox model with time-dependent exposures.

The MSCM approach is usually used to estimate a time-dependent treatment effect in the presence of time-dependent confounding. To avoid some of the problems of fitting MSCM, such as potential instability of the IPW estimates, an alternative, the sequential Cox approach, was adopted to study drug exposure in analyzing a HIV cohort [73]. Although the estimation mechanisms are different, the target parameter, the causal effect of treatment, is the same in both approaches and both approaches also adjust for time-dependent confounding.

To date, neither of these alternative approaches were investigated in the literature in generalized settings as alternatives to the Cox model with time-dependent exposures and MSCM respectively. To the best of our knowledge, ours is the first study to investigate the generalizability of these approaches using simulation settings suitable for the contexts of these methods. Findings from our study guide the appropriate choice of analysis tool based on information such as whether time-dependent covariates are present or not and whether a covariate interacts with the treatment (exposure possibly delayed for some subjects). The simulation studies revealed that the sequential Cox approach is more useful than the PTDM approach in addressing im-

mortal time bias. On the other hand, our results indicate that this approach is not as effective as other studies have suggested in the presence of a time-dependent confounder.

5.3 Future Research

The main aim of this dissertation was to assess, validate and refine the causal inference tools to estimate the causal effect of a treatment in a realistic epidemiological context involving time-dependent confounders. We used these tools to answer an MS research question that is of great importance to MS patients: is β-IFN treatment beneficial in reducing the hazard of longer-term irreversible disability milestones. Other approaches, such as structural nested models [38, 70, 71], the sequential stratification approach [72], nonparametric g-computation approach [221–226], tree-based g-computation [227] and parametric g-computation [228–233] may also be useful in estimating treatment effects in MS in the presence of time-dependent confounders. Further research could make use of the dynamic MSCMs [234–236] and the dynamic random g-formula [237, 238] to answer questions regarding the optimal time to start β-IFN treatment. Future research could address more specific questions regarding the direct, indirect and mediated effects of β-IFN using the g-computation approach and extensions of the sequential Cox approach [196, 229].

Bibliography

[1] Evans C., Zhu F., Kingwell E., Shirani A., van der Kop M., Petkau J., Gustafson P., Zhao Y., Oger J., and Tremlett H. Association between beta-interferon exposure and hospital events in multiple sclerosis. *Pharmacoepidemiology and Drug Safety*, 2014. doi: 10.1002/pds.3667. URL `http://dx.doi.org/10.1002/pds.3667`.

[2] INFB Multiple Sclerosis Study Group. Interferon beta-1b is effective in relapsing-remitting multiple sclerosis. I. Clinical results of a multicenter, randomized, double-blind, placebo-controlled trial. *Neurology*, 43(4):655–661, 1993.

[3] INFB Multiple Sclerosis Study Group and the University of British Columbia MS/MRI Analysis Group. Interferon beta-1b in the treatment of multiple sclerosis: final outcome of the randomized controlled trial. *Neurology*, 45(7):1277–1285, 1995.

[4] Jacobs L.D., Cookfair D.L., Rudick R.A., Herndon R.M., Richert J.R., Salazar A.M., Fischer J.S., Goodkin D.E., Granger C.V., Simon J.H., et al. Intramuscular interferon beta-1a for disease progression in relapsing multiple sclerosis. *Annals of Neurology*, 39(3):285–294, 1996.

[5] Ebers, G.C. and PRISMS (Prevention of Relapses and Disability by Interferon beta-1a Subcutaneously in Multiple Sclerosis) study group. Randomised double-blind placebo-controlled study of interferon beta-1a in relapsing/remitting multiple sclerosis. *The Lancet*, 352(9139): 1498–1504, 1998.

[6] Freedman M. and the OWIMS Study Group. Evidence of interferon

beta-1a dose response in relapsing–remitting MS: the OWIMS Study. *Neurology*, 53(4):679–686, 1999.

[7] Hume D. *An enquiry concerning human understanding.* P.F. Collier & Son, 1748.

[8] Neyman J., Dabrowska D.M., and Speed T.P. On the application of probability theory to agricultural experiments. essay on principles. section 9. (translated). *Roczniki Nauk Rolniczych Tom X [in Polish, translated in] Statistical Science*, 5:465–472, 1923.

[9] Fisher R.A. *Statistical methods for research workers.* Edinburgh, 1925.

[10] Fisher R.A. et al. The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain*, 33:503–513, 1926.

[11] Rubin D.B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

[12] Rubin D.B. Assignment to treatment group on the basis of a covariate. *Journal of Educational and Behavioral statistics*, 2(1):1, 1977.

[13] Rubin D.B. Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 6(1):34–58, 1978.

[14] Rubin D.B. Estimation in parallel randomized experiments. *Journal of Educational and Behavioral Statistics*, 6(4):377, 1981.

[15] Rubin D.B. Formal mode of statistical inference for causal effects. *Journal of Statistical Planning and Inference*, 25(3):279–292, 1990.

[16] Rubin D.B. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

[17] Dawid A.P. Causal inference without counterfactuals. *Journal of the American Statistical Association*, 95(450):407–424, 2000.

[18] Holland P.W. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

[19] Rubin D.B. Statistics and causal inference: comment: which ifs have causal answers. *Journal of the American Statistical Association*, 81 (396):961–962, 1986.

[20] Rosenbaum P.R. and Rubin D.B. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1): 41–55, 1983.

[21] Hernán M.A. and Robins J.M. *Causal inference*. Chapman Hall/CRC, 2015. Forthcoming. URL: `http://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/` Last accessed: Oct-05,2014.

[22] Robins J.M. Association, causation, and marginal structural models. *Synthese*, 121(1):151–179, 1999.

[23] Rubin D.B. Randomization analysis of experimental data: The Fisher randomization test. *Journal of the American Statistical Association*, 75(371):591–593, 1980.

[24] Cox D.R. *Planning of experiments*. Wiley, 1958.

[25] Lewis D.K. *Counterfactuals*. Wiley-Blackwell, 1973.

[26] VanderWeele T.J. and Hernán M.A. Causal inference under multiple versions of treatment. *Journal of Causal Inference*, 1(1):1–20, 2013.

[27] VanderWeele T.J. and Hernán M.A. From counterfactuals to sufficient component causes and vice versa. *European Journal of Epidemiology*, 21(12):855–858, 2006.

[28] Hernán M.A. and Taubman S.L. Does obesity shorten life? the importance of well-defined interventions to answer causal questions. *International Journal of Obesity*, 32:S8–S14, 2008.

[29] Cole S.R. and Frangakis C.E. The consistency statement in causal inference: a definition or an assumption? *Epidemiology*, 20(1):3–5, 2009.

[30] Rosenbaum P.R. and Rubin D.B. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79(387):516–524, 1984.

[31] Rosenbaum P.R. and Rubin D.B. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician*, 39(1):33–38, 1985.

[32] Rubin D.B. Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine*, 127(8 Part 2):757–763, 1997.

[33] Robins J.M. 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, 1986.

[34] Robins J.M. Addendum to "a new approach to causal inference in mortality studies with a sustained exposure period - application to control of the healthy worker survivor effect". *Computers & Mathematics with Applications*, 14(9-12):923–945, 1987.

[35] Greenland S., Pearl J., and Robins J.M. Causal diagrams for epidemiologic research. *Epidemiology*, 10(1):37–48, 1999.

[36] Robins J.M. The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. *Health Service Research Methodology: a Focus on AIDS*, 113:113–159, 1989.

[37] Robins J. Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. *Biometrika*, 79(2):321–334, 1992.

[38] Robins J.M. Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics - Theory and Methods*, 23(8):2379–2412, 1994.

[39] Robins J.M. Marginal structural models. In *Proceedings of the American Statistical Association, Section on Bayesian Statistical Science*, pages 1–10. American Statistical Association, 1997.

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

[41] Robins J.M. Marginal structural models versus structural nested models as tools for causal inference. *Statistical Models in Epidemiology, the Environment and Clinical Trials*, 116:95–134, 1999.

[42] Robins J.M., Hernán M.A., and Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5): 550–560, 2000.

[43] Månsson R., Joffe M.M., Sun W., and Hennessy S. On the estimation and use of propensity scores in case-control and case-cohort studies. *American Journal of Epidemiology*, 166(3):332, 2007.

[44] Austin P.C. The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies. *Statistics in Medicine*, 29: 2137–2148, 2010.

[45] Miettinen O.S. Components of the crude risk ratio. *American Journal of Epidemiology*, 96(2):168–172, 1972.

[46] Miettinen O.S. and Cook E. Confounding: essence and detection. *American Journal of Epidemiology*, 114(4):593–603, 1981.

[47] Sato T. and Matsuyama Y. Marginal structural models as a tool for standardization. *Epidemiology*, 14(6):680–686, 2003.

[48] Newman S.C. Causal analysis of case-control data. *Epidemiologic Perspectives & Innovations*, 3(1):2–7, 2006.

[49] Hernán M.A. and Robins J.M. Estimating causal effects from epidemiological data. *Journal of Epidemiology and Community Health*, 60(7): 578–586, 2006.

[50] Hernán M.A., Brumback B., and Robins J.M. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*, 11(5):561–570, 2000.

[51] Suarez D., Borras R., and Basagana X. Differences between marginal structural models and conventional models in their exposure effect estimates: A systematic review. *Epidemiology*, 22(4):586–588, 2011.

[52] Yang S., Eaton C.B., Lu J., and Lapane K.L. Application of marginal structural models in pharmacoepidemiologic studies: a systematic review. *Pharmacoepidemiology and Drug Safety*, 23(6):560–571, 2014.

[53] van der Laan M.J. and Robins J.M. *Unified methods for censored longitudinal data and causality*. Springer Verlag, 2003.

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

[55] Cole S.R. and Hernán M.A. Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, 168(6):656–664, 2008.

[56] Young J.G., Hernán M.A., Picciotto S., and Robins J.M. Relation between three classes of structural models for the effect of a time-varying exposure on survival. *Lifetime Data Analysis*, 16(1):71–84, 2010.

[57] Xiao Y., Abrahamowicz M., and Moodie E.E.M. Accuracy of conventional and marginal structural Cox model estimators: A simulation study. *The International Journal of Biostatistics*, 6(2):1–28, 2010.

[58] Hernán M.A., Brumback B., and Robins J.M. Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *Journal of the American Statistical Association*, 96(454):440–448, 2001.

[59] Pearl J. Causal diagrams for empirical research. *Biometrika*, pages 669–688, 1995.

[60] Pearl J. *Causality: models, reasoning and inference*. Cambridge Univ Press, 2000.

[61] Spirtes P., Glymour C.N., and Scheines R. *Causation, prediction, and search*, volume 81. The MIT Press, 2000.

[62] Shrier I. and Platt R. Reducing bias through directed acyclic graphs. *BMC Medical Research Methodology*, 8(1):70, 2008.

[63] Hernández-Díaz S., Schisterman E.F., and Hernán M.A. The birth weight paradox uncovered? *American Journal of Epidemiology*, 164 (11):1115, 2006.

[64] Hernán M.A., Hernández-Díaz S., and Robins J.M. A structural approach to selection bias. *Epidemiology*, 15(5):615–625, 2004.

[65] VanderWeele T.J. and Robins J.M. Four types of effect modification: A classification based on directed acyclic graphs. *Epidemiology*, 18(5): 561, 2007.

[66] Hernán M.A., Clayton D., and Keiding N. The Simpson's paradox unraveled. *International Journal of Epidemiology*, 2011.

[67] Schisterman E.F., Cole S.R., and Platt R.W. Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology*, 20(4):488, 2009.

[68] Leibovici L. Effects of remote, retroactive intercessory prayer on outcomes in patients with bloodstream infection: randomised controlled trial. *British Medical Journal*, 323(7327):1450, 2001.

[69] Fewell Z., Hernán M.A., Wolfe F., Tilling K., Choi H., and Sterne JA. Controlling for time-dependent confounding using marginal structural models. *The Stata Journal*, 4(4):402–420, 2004.

[70] Robins J. A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *Journal of Chronic Diseases*, 40:139–161, 1987.

[71] Almirall D., Ten Have T., and Murphy S.A. Structural nested mean models for assessing time-varying effect moderation. *Biometrics*, 66 (1):131–139, 2010.

[72] Schaubel D.E., Wolfe R.A., Sima C.S., and Merion R.M. Estimating the effect of a time-dependent treatment by levels of an internal time-dependent covariate: application to the contrast between liver wait-list and posttransplant mortality. *Journal of the American Statistical Association*, 104(485):49–59, 2009.

[73] Gran J.M., Røysland K., Wolbers M., Didelez V., Sterne J.A.C., Ledergerber B., Furrer H., von Wyl V., and Aalen O.O. A sequential Cox approach for estimating the causal effect of treatment in the presence of time-dependent confounding applied to data from the Swiss HIV Cohort Study. *Statistics in Medicine*, 29(26):2757–2768, 2010.

[74] Andersen P.K. and Gill R.D. Cox's regression model for counting processes: a large sample study. *The Annals of Statistics*, 10(4):1100–1120, 1982.

[75] Lange T. and Rod N.H. Causal models. In *Handbook of survival analysis*, pages 135–151. CRC Press, 2013.

[76] Messmer B.J., Leachman R.D., Nora J.J., and Cooley D.A. Survival-times after cardiac allografts. *The Lancet*, 293(7602):954–956, 1969.

[77] Clark D.A., Stinson E.B., Griepp R.B., Schroeder J.S., Shumway N.E., and Harrison D.C. Cardiac transplantation in man. *Annals of Internal Medicine*, 75(1):15, 1971.

[78] Suissa S. Effectiveness of inhaled corticosteroids in chronic obstructive pulmonary disease: immortal time bias in observational studies. *American Journal of Respiratory and Critical Care Medicine*, 168(1): 49, 2003.

[79] Suissa S. Immortal time bias in observational studies of drug effects. *Pharmacoepidemiology and Drug Safety*, 16(3):241–249, 2007.

[80] Suissa S. Immortal time bias in pharmacoepidemiology. *American Journal of Epidemiology*, 167(4):492, 2008.

[81] Clayton D. and Hills M. *Statistical models in epidemiology*. Oxford University Press, 1993.

[82] Zhou Z., Rahme E., Abrahamowicz M., and Pilote L. Survival bias associated with time-to-treatment initiation in drug effectiveness evaluation: a comparison of methods. *American Journal of Epidemiology*, 162(10):1016–1023, 2005.

[83] Cox D.R. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 187–220, 1972.

[84] Sylvestre M.P., Huszti E., and Hanley J.A. Do OSCAR winners live longer than less successful peers? A reanalysis of the evidence. *Annals of Internal Medicine*, 145(5):361–363, 2006.

[85] Ho P.M., Fihn S.D., Wang L., Bryson C.L., Lowy E., Maynard C., Magid D.J., et al. Clopidogrel and long-term outcomes after stent implantation for acute coronary syndrome. *American Heart Journal*, 154(5):846–851, 2007.

[86] Karp I., Behlouli H., LeLorier J., and Pilote L. Statins and cancer risk. *The American Journal of Medicine*, 121(4):302–309, 2008.

[87] Ho P.M., Maddox T.M., Wang L., Fihn S.D., Jesse R.L., Peterson E.D., and Rumsfeld J.S. Risk of adverse outcomes associated with

concomitant use of clopidogrel and proton pump inhibitors following acute coronary syndrome. *Journal of American Medical Association*, 301(9):937–944, 2009.

[88] Snyder C.W., Weinberg J.A., McGwin Jr G., Melton S.M., George R.L., Reiff D.A., Cross J.M., Hubbard-Brown J., Rue III L.W., and Kerby J.D. The relationship of blood product ratio to mortality: survival benefit or survival bias? *The Journal of Trauma*, 66(2):358–362, 2009.

[89] World health organization: multiple sclerosis international federation. *Multiple Sclerosis Resources in the World.* Geneva, Switzerland: World Health Organization, 2008.

[90] Brown M.G., Kirby S., Skedgel C., Fisk J.D., Murray T.J., Bhan V., and Sketris I.S. How effective are disease-modifying drugs in delaying progression in relapsing-onset MS? *Neurology*, 69(15):1498, 2007.

[91] Trojano M., Pellegrini F., Fuiani A., Paolicelli D., Zipoli V., Zimatore G.B., Di Monte E., Portaccio E., Lepore V., Livrea P., and M.P. Amato. New natural history of interferon-β–treated relapsing multiple sclerosis. *Annals of Neurology*, 61(4):300–306, 2007.

[92] Shirani A., Zhao Y., Karim M.E., Evans C., Kingwell E., van der Kop M., Oger J., Gustafson P., Petkau J., and Tremlett H. Association between use of interferon beta and progression of disability in patients with relapsing-remitting multiple sclerosis. *Journal of American Medical Association*, 308(3):247–256, 2012.

[93] Renoux C. and Suissa S. Immortal time bias in the study of effectiveness of interferon-β in multiple sclerosis. *Annals of Neurology*, 64(1): 109–110, 2008.

[94] Koch M., Mostert J., De Keyser J., Tremlett H., and Filippini G. Interferon-beta treatment and the natural history of relapsing-remitting multiple sclerosis. *Annals of Neurology*, 63(1):125–126, 2008.

[95] Derfuss T. and Kappos L. Evaluating the potential benefit of interferon treatment in multiple sclerosis. *Journal of American Medical Association*, 308(3):290–291, 2012.

[96] Goodin D.S., Reder A.T., and Cutter G. Treatment with interferon beta for multiple sclerosis. *The Journal of the American Medical Association*, 308(16):1627–1628, 2012.

[97] Shirani A., Petkau J., and Tremlett H. Treatment with interferon beta for multiple sclerosis-reply. *Journal of American Medical Association*, 308(16):1627–1628, 2012.

[98] Greenberg B.M., Balcer L., Calabresi P.A., Cree B., Cross A., Frohman T., Gold R., Havrdova E., Hemmer B., Kieseier B.C., Lisak R., Miller M.K. A. Racke, Steinman L., Stuve O., Wiendl H., and Frohman E. Interferon beta use and disability prevention in relapsing-remitting multiple sclerosis. *Journal of American Medical Association Neurology*, 70(2):248–251, 2013.

[99] Shirani A., Zhao Y., Karim M.E., Evans C., Kingwell E., van der Kop M., Oger J., Gustafson P., Petkau J., and Tremlett H. Interferon beta and long-term disability in multiple sclerosis. *JAMA Neurology*, 70(5):651–653, 2013.

[100] Coles A. Multiple sclerosis: The bare essentials. *Neurology in Practice*, 9(2):118–126, 2009.

[101] Shirani A., Zhao Y., Karim M.E., Evans C., Kingwell E., van der Kop M., Oger J., Gustafson P., Petkau J., and Tremlett H. Investigation of heterogeneity in the association between interferon beta and disability progression in multiple sclerosis: an observational study. *European Journal of Neurology*, 21(6):835–844, 2014.

[102] Westreich D., Cole S.R., Schisterman E.F., and Platt R.W. A simulation study of finite-sample properties of marginal structural Cox proportional hazards models. *Statistics in Medicine*, 31(19):2098–2109, 2012.

[103] Havercroft W.G. and Didelez V. Simulating from marginal structural models with time-dependent confounding. *Statistics in Medicine*, 31 (30):4190–4206, 2012.

[104] Xiao Y., Moodie E.E.M., and Abrahamowicz M. Comparison of approaches to weight truncation for marginal structural Cox models. *Epidemiologic Methods*, 2(1):1–20, 2012.

[105] Kurtzke J.F. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). *Neurology*, 33(11):1444–1452, 1983.

[106] Tremlett H., Paty D., and Devonshire V. Disability progression in multiple sclerosis is slower than previously reported. *Neurology*, 66(2): 172–177, 2006.

[107] Tremlett H., Yousefi M., Devonshire V., Rieckmann P., and Zhao Y. Impact of multiple sclerosis relapses on progression diminishes with time. *Neurology*, 73(20):1616–1623, 2009.

[108] Tremlett H., Zhao Y., Joseph J., and Devonshire V. Relapses in multiple sclerosis are age-and time-dependent. *Journal of Neurology, Neurosurgery & Psychiatry*, 79(12):1368–1374, 2008.

[109] Glymour M.M. Using causal diagrams to understand common problems in social epidemiology. In Oakes J.M. and Kaufman J.S., editors, *Methods in social epidemiology*. Jossey-Bass, 2006.

[110] Robins J.M., Greenland S., and Hu F.C. Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome. *Journal of the American Statistical Association*, 94 (447):687–700, 1999.

[111] Cook N.R., Cole S.R., and Hennekens C.H. Use of a marginal structural model to determine the effect of aspirin on cardiovascular mortality in the Physicians' Health Study. *American Journal of Epidemiology*, 155(11):1045–1053, 2002.

[112] Cole S.R., Hernán M.A., Robins J.M., Anastos K., Chmiel J., Detels R., Ervin C., Feldman J., Greenblatt R., Kingsley L., Lai S., Young M., Cohen M., and Muñoz A. Effect of highly active antiretroviral therapy on time to acquired immunodeficiency syndrome or death using marginal structural models. *American Journal of Epidemiology*, 158(7):687–694, 2003.

[113] McCulloch M., Broffman M., van der Laan M., Hubbard A., Kushi L., Kramer A., Gao J., and Colford J.M. Lung cancer survival with herbal medicine and vitamins in a whole-systems approach: ten-year follow-up data analyzed with marginal structural models and propensity score methods. *Integrative Cancer Therapies*, 10(3):260–279, 2011.

[114] Ali R.A., Ali M.A., and Wei Z. On computing standard errors for marginal structural Cox models. *Lifetime Data Analysis*, 20(1):106–131, 2014.

[115] Westreich D., Cole S.R., Tien P.C., Chmiel J.S., Kingsley L., Funk M.J., Anastos K., and Jacobson L.P. Time scale and adjusted survival curves for marginal structural Cox models. *American Journal of Epidemiology*, 171(6):691–700, 2010.

[116] R Core Team. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. URL `http://www.R-project.org/`. ISBN 3-900051-07-0.

[117] Harrell F.E. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer, 2001.

[118] Thompson Jr W.A. On the treatment of grouped observations in life studies. *Biometrics*, 33(3):463–470, 1977.

[119] D'Agostino R.B., Lee M.L., Belanger A.J., Cupples L.A., Anderson K., and Kannel W.B. Relation of pooled logistic regression to time dependent Cox regression analysis: the Framingham Heart Study. *Statistics in Medicine*, 9(12):1501–1515, 1990.

[120] Platt R.W., Delaney J.A.C., and Suissa S. The positivity assumption and marginal structural models: the example of warfarin use and risk of bleeding. *European Journal of Epidemiology*, 27(2):77–83, 2012.

[121] Lee B.K., Lessler J., and Stuart E.A. Weight trimming and propensity score weighting. *PLoS one*, 6(3):e18174, 03 2011.

[122] Van der Wal W.M., Noordzij M., Dekker F.W., Boeschoten E.W., Krediet R.T., Korevaar J.C., and Geskus R.B. Comparing mortality in renal patients on hemodialysis versus peritoneal dialysis using a marginal structural model. *The International Journal of Biostatistics*, 6(1):1–19, 2010.

[123] Robins J., Orellana L., and Rotnitzky A. Estimation and extrapolation of optimal treatment and testing strategies. *Statistics in Medicine*, 27 (23):4678–4721, 2008.

[124] Willoughby E.W. and Paty D.W. Scales for rating impairment in multiple sclerosis a critique. *Neurology*, 38(11):1793–1793, 1988.

[125] Ebers G.C., Traboulsee A., Li D., Langdon D., Reder A.T., Goodin D.S., Bogumil T., Beckmann K., Wolf C., Konieczny A., and the Investigators of the 16-year Long-Term Follow-Up Study. Analysis of clinical outcomes according to original treatment groups 16 years after the pivotal ifnb-1b trial. *Journal of Neurology, Neurosurgery & Psychiatry*, 81(8):907–912, 2010.

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

[127] Imai K. and Ratkovic M. Robust estimation of inverse probability weights for marginal structural models, 2014. URL `http://imai.princeton.edu/research/MSM.html`. Technical Report, Last accessed: July 20, 2014.

[128] Karim M. E., Gustafson P., Petkau J., Zhao Y., Shirani A., Kingwell E., Evans C., van der Kop M., Oger J., and Tremlett H. Marginal Structural Cox Models for Estimating the Association Between β-Interferon Exposure and Disease Progression in a Multiple Sclerosis Cohort. *American Journal of Epidemiology*, 180(2):160–171, 2014.

[129] Gruber S., Logan R.W., Jarrín I., Monge S., and Hernán M.A. Ensemble learning of inverse probability weights for marginal structural modeling in large observational datasets. *Statistics in Medicine*, 2014. doi: 10.1002/sim.6322. URL `http://dx.doi.org/10.1002/sim.6322`.

[130] Fong C. and Imai K. Covariate balancing propensity score for general treatment regimes, 2014. URL `http://imai.princeton.edu/research/CBGPS.html`. Technical Report, Last accessed: Sept 20, 2014.

[131] Wyss R., Ellis A.R., Brookhart M.A., Girman C.J., Funk M.J., LoCasale R., and Stürmer T. The role of prediction modeling in propensity score estimation: An evaluation of logistic regression, bcart, and the covariate-balancing propensity score. *American Journal of Epidemiology*, 180(6):645–655, 2014.

[132] Lee B.K., Lessler J., and Stuart E.A. Improving propensity score weighting using machine learning. *Statistics in Medicine*, 29(3):337–346, 2010.

[133] Austin P.C. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46(3):399–424, 2011.

[134] McCaffrey D.F., Ridgeway G., and Morral A.R. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, 9(4):403–425, 2004.

[135] Westreich D., Lessler J., and Funk M.J. Propensity score estimation: machine learning and classification methods as alternatives to logistic regression. *Journal of Clinical Epidemiology*, 63(8):826, 2010.

[136] Li L., Shen C., Wu A.C., and Li X. Propensity score-based sensitivity analysis method for uncontrolled confounding. *American Journal of Epidemiology*, 174(3):345–353, 2011.

[137] Zhu Y., Ghosh D., Mukherjee B., and Mitra N. A data-adaptive strategy for inverse weighted estimation of causal effects, 2013. URL `http://works.bepress.com/debashis_ghosh/58`. Technical Report, Collection of Biostatistics Research Archive, Last accessed: June-05-2014.

[138] Keller B.S., Kim J., and Steiner P.M. Data mining alternatives to logistic regression for propensity score estimation: Neural networks and support vector machines. *Multivariate Behavioral Research*, 48 (1):164–164, 2013.

[139] Watkins S., Jonsson-Funk M., Brookhart M.A., Rosenberg S.A., O'Shea T.M., and Daniels J. An empirical comparison of tree-based methods for propensity score estimation. *Health Services Research*, 48 (5):1798–1817, 2013.

[140] Regier M.D., Moodie E.E.M., and Platt R.W. The effect of error-in-confounders on the estimation of the causal parameter when using marginal structural models and inverse probability-of-treatment weights: A simulation study. *The International Journal of Biostatistics*, 10(1):1–15, 2014.

[141] Coffman D.L. and Zhong W. Assessing mediation using marginal structural models in the presence of confounding and moderation. *Psychological Methods*, 17(4):642–664, 2012.

[142] Young J.G., Hernán M.A., Picciotto S., and Robins J.M. Simulation from structural survival models under complex time-varying data structures. In *JSM Proceedings, Section on Statistics in Epidemiology*, pages 1–6. American Statistical Association, 2008.

[143] Picciotto S, Young J, and Hernán M.A. G-estimation of structural

nested cumulative failure time models. *American Journal of Epidemiology*, 67:139, 2008.

[144] Young J.G. and Tchetgen Tchetgen E.J. Simulation from a known Cox MSM using standard parametric models for the g-formula. *Statistics in Medicine*, 33(6):1001–1014, 2014.

[145] Lin D.Y. and Wei L. The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association*, 84 (408):1074–1078, 1989.

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

[147] Breiman L. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[148] Breiman L. Arcing classifier (with discussion and a rejoinder by the author). *The Annals of Statistics*, 26(3):801–849, 1998.

[149] Kuhn Max and Johnson Kjell. *Applied predictive modeling*. Springer, 2013.

[150] James G., Witten D., Hastie T., and Tibshirani R. *An introduction to statistical learning*. Springer, 2013.

[151] Luellen J.K., Shadish W.R., and Clark M.H. Propensity scores an introduction and experimental test. *Evaluation Review*, 29(6):530–558, 2005.

[152] Fan R., Chen P., and Lin C. Working set selection using second order information for training support vector machines. *The Journal of Machine Learning Research*, 6:1889–1918, 2005.

[153] Chang C. and Lin C. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):1–27, 2011.

[154] Ridgeway G. The state of boosting. *Computing Science and Statistics*, pages 172–181, 1999.

[155] Guo S. and Fraser Mark W. *Propensity score analysis: statistical methods and applications*. Sage Publications, 2009.

[156] Robins J.M. and Hernán M.A. Estimation of the causal effects of time-varying exposures. In *Longitudinal data analysis*, pages 553–599. CRC Press, 2009.

[157] Wang Y., Petersen M.L., Bangsberg D., and van der Laan M.J. Diagnosing bias in the inverse probability of treatment weighted estimator resulting from violation of experimental treatment assignment, 2006. URL `http://biostats.bepress.com/ucbbiostat/paper211/`. Technical Report, Last accessed: July 20, 2014.

[158] Bembom O. and van der Laan M.J. Data-adaptive selection of the truncation level for inverse-probability-of-treatment-weighted estimators, 2008. URL `http://biostats.bepress.com/ucbbiostat/paper230/`. Technical Report, Last accessed: July 20, 2014.

[159] Bryan J., Yu Z., and van der Laan M.J. Analysis of longitudinal marginal structural models. *Biostatistics*, 5(3):361–380, 2004.

[160] Moodie E.E.M., Stephens D.A., and Klein M.B. A marginal structural model for multiple-outcome survival data: assessing the impact of injection drug use on several causes of death in the canadian co-infection cohort. *Statistics in Medicine*, 33(8):1409–1425, 2014.

[161] Marcus S.M., Siddique J., Ten Have T.R., Gibbons R.D., Stuart E., and Normand S.T. Balancing treatment comparisons in longitudinal studies. *Psychiatric Annals*, 38(12):805, 2008.

[162] Cole S.R. and Hernán M.A. Adjusted survival curves with inverse probability weights. *Computer Methods and Programs in Biomedicine*, 75(1):45–49, 2004.

[163] Langford J. and Zadrozny B. Estimating class membership probabilities using classifier learners. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, pages 198–205, 2005.

[164] Zhu Ji and Hastie Trevor. Kernel logistic regression and the import vector machine. In *Advances in Neural Information Processing Systems*, pages 1081–1088, 2001.

[165] Efron B. and Tibshirani R.J. *An introduction to the bootstrap*. CRC press, 1994.

[166] Brumback B.A., Hernán M.A., Haneuse S.J.P.A., and Robins J.M. Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures. *Statistics in Medicine*, 23(5): 749–767, 2004.

[167] Lash T. L and Cole S.R. Immortal person-time in studies of cancer outcomes. *Journal of Clinical Oncology*, 27(23):e55–e56, 2009.

[168] van Walraven C., Davis D., Forster A.J., and Wells G.A. Time-dependent bias was common in survival analyses published in leading clinical journals. *Journal of Clinical Epidemiology*, 57(7):672–682, 2004.

[169] Wolkewitz M., Allignol A., Harbarth S., de Angelis G., Schumacher M., and Beyersmann J. Time-dependent study entries and exposures in cohort studies can easily be sources of different and avoidable types of bias. *Journal of Clinical Epidemiology*, 65(11):1171–1180, 2012.

[170] Gail M.H. Does cardiac transplantation prolong life? A reassessment. *Annals of Internal Medicine*, 76(5):815–817, 1972.

[171] Lévesque L.E., Hanley J.A., Kezouh A., and Suissa S. Problem of immortal time bias in cohort studies: example using statins for preventing progression of diabetes. *British Medical Journal*, 340, 2010.

[172] Austin P.C. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine*, 28(25):3083–3107, 2009.

[173] Ravi B., Croxford R., Austin P.C., Lipscombe L., Bierman A.S., Harvey P.J., and Hawker G.A. The relation between total joint arthroplasty and risk for serious cardiovascular events in patients with moderate-severe osteoarthritis: propensity score matched landmark analysis. *British Medical Journal*, 347:f6187, 2013.

[174] Kiri V.A., Pride N.B., Soriano J.B., and Vestbo J. Inhaled corticosteroids in chronic obstructive pulmonary disease: results from two observational designs free of immortal time bias. *American Journal of Respiratory and Critical Care Medicine*, 172(4):460–464, 2005.

[175] Karim M.E. Can joint replacement reduce cardiovascular risk? *British Medical Journal*, 347:f6651, 2013.

[176] Li Y.P., Propert K.J., and Rosenbaum P.R. Balanced risk set matching. *Journal of the American Statistical Association*, 96(455):870–882, 2001.

[177] Lu B. Propensity score matching with time-dependent covariates. *Biometrics*, 61(3):721–728, 2005.

[178] Li Y., Schaubel D.E., and He K. Matching methods for obtaining survival functions to estimate the effect of a time-dependent treatment. *Statistics in Biosciences*, 6(1):105–126, 2014.

[179] Trojano M. and Pellegrini F. Reply. *Annals of Neurology*, 64(1):110–110, 2008.

[180] Tleyjeh I.M., Ghomrawi H.M.K., Steckelberg J.M., Montori V.M., Hoskin T.L., Enders F., Huskins W.C., Mookadam F., Wilson W.R., Zimmerman V., and Baddourb L.M. Propensity score analysis with a time-dependent intervention is an acceptable although not an optimal

analytical approach when treatment selection bias and survivor bias coexist. *Journal of Clinical Epidemiology*, 63(2):139–140, 2010.

[181] Austin P.C. and Platt R.W. Survivor treatment bias, treatment selection bias, and propensity scores in observational research. *Journal of Clinical Epidemiology*, 63(2):136–138, 2010.

[182] Kiri V.A. and MacKenzie G. Re: "immortal time bias in pharmacoepidemiology". *American Journal of Epidemiology*, 170(5):667–668, 2009.

[183] Sylvestre M.P. and Abrahamowicz M. Comparison of algorithms to generate event times conditional on time-dependent covariates. *Statistics in Medicine*, 27(14):2618–2634, 2008.

[184] Austin P.C., Mamdani M.M., Van Walraven C., and Tu J.V. Quantifying the impact of survivor treatment bias in observational studies. *Journal of Evaluation in Clinical Practice*, 12(6):601–612, 2006.

[185] Shintani A.K., Girard T.D., Arbogast P.G., Moons K.G.M., and Ely E.W. Immortal time bias in critical care research: application of time-varying cox regression for observational cohort studies. *Critical Care Medicine*, 37(11):2939, 2009.

[186] Liu J., Weinhandl E.D., Gilbertson D.T., Collins A.J., and St Peter W.L. Issues regarding 'immortal time' in the analysis of the treatment effects in observational studies. *Kidney International*, 81(4): 341–350, 2011.

[187] Ho A., Dion P.W., Yeung J.H.H., Joynt G.M., Lee A., Ng C.S.H., Chang A., So F.L., and Cheung C.W. Simulation of survivorship bias in observational studies on plasma to red blood cell ratios in massive transfusion for trauma. *British Journal of Surgery*, 99(S1):132–139, 2012.

[188] Buyse M. and Piedbois P. On the relationship between response to

treatment and survival time. *Statistics in Medicine*, 15(24):2797–2812, 1996.

[189] Beyersmann J., Gastmeier P., Wolkewitz M., and Schumacher M. An easy mathematical proof showed that time-dependent bias inevitably leads to biased effect estimation. *Journal of Clinical Epidemiology*, 61 (12):1216–1221, 2008.

[190] Beyersmann Jan, Wolkewitz Martin, and Schumacher Martin. The impact of time-dependent bias in proportional hazards modelling. *Statistics in Medicine*, 27(30):6439–6454, 2008.

[191] Gupta S.K. Intention-to-treat concept: A review. *Perspectives in Clinical Research*, 2(3):109, 2011.

[192] Wolkewitz M., Allignol A., Schumacher M., and Beyersmann J. Two pitfalls in survival analyses of time-dependent exposure: a case study in a cohort of oscar nominees. *The American Statistician*, 64(3):205–211, 2010.

[193] Leffondré K., Abrahamowicz M., and Siemiatycki J. Evaluation of cox's model and logistic regression for matched case-control data with time-dependent covariates: a simulation study. *Statistics in Medicine*, 22(24):3781–3794, 2003.

[194] Cole S.R., Platt R.W., Schisterman E.F., Chu H., Westreich D., Richardson D., and Poole C. Illustrating bias due to conditioning on a collider. *International Journal of Epidemiology*, 39(2):417–420, 2010.

[195] Gran J.M. Infectious disease modelling and causal inference, 2011. Ph.D. Thesis, University of Oslo.

[196] Røysland K., Gran J.M., Ledergerber B., Wyl V., Young J., and Aalen O.O. Analyzing direct and indirect effects of treatment using dynamic path analysis applied to data from the swiss hiv cohort study. *Statistics in Medicine*, 30(24):2947–2958, 2011.

[197] Leemis L.M. Technical note-variate generation for accelerated life and proportional hazards models. *Operations Research*, 35(6):892–894, 1987.

[198] Bender R., Augustin T., and Blettner M. Generating survival times to simulate cox proportional hazards models. *Statistics in Medicine*, 24(11):1713–1723, 2005.

[199] Zhou M. Understanding the Cox regression models with time-change covariates. *The American Statistician*, 55(2):153–155, 2001.

[200] Leemis L.M., Shih L., and Reynertson K. Variate generation for accelerated life and proportional hazards models with time dependent covariates. *Statistics & Probability Letters*, 10(4):335–339, 1990.

[201] Shih L. and Leemis L.M. Variate generation for a nonhomogeneous poisson process with time dependent covariates. *Journal of Statistical Computation and Simulation*, 44(3-4):165–186, 1993.

[202] Austin P.C. Generating survival times to simulate cox proportional hazards models with time-varying covariates. *Statistics in Medicine*, 31(29):3946–3958, 2012.

[203] Hendry D.J. Data generation for the Cox proportional hazards model with time-dependent covariates: a method for medical researchers. *Statistics in Medicine*, 33(3):436–454, 2014.

[204] Abrahamowicz M., Mackenzie T., and Esdaile J.M. Time-dependent hazard ratio: modeling and hypothesis testing with application in lupus nephritis. *Journal of the American Statistical Association*, 91 (436):1432–1439, 1996.

[205] Mackenzie T. and Abrahamowicz M. Marginal and hazard ratio specific random data generation: Applications to semi-parametric bootstrapping. *Statistics and Computing*, 12(3):245–252, 2002.

[206] Abrahamowicz M. and MacKenzie T.A. Joint estimation of time-dependent and non-linear effects of continuous covariates on survival. *Statistics in Medicine*, 26(2):392–408, 2007.

[207] Sylvestre M. and Abrahamowicz M. Flexible modeling of the cumulative effects of time-dependent exposures on the hazard. *Statistics in Medicine*, 28(27):3437–3453, 2009.

[208] Mahboubi A., Abrahamowicz M., Giorgi R., Binquet C., Bonithon-Kopp C., and Quantin C. Flexible modeling of the effects of continuous prognostic factors in relative survival. *Statistics in Medicine*, 30(12): 1351–1365, 2011.

[209] Abrahamowicz M., Beauchamp M., and Sylvestre M. Comparison of alternative models for linking drug exposure with adverse effects. *Statistics in Medicine*, 31(11-12):1014–1030, 2012.

[210] Gauvin H., Lacourt A., and Leffondré K. On the proportional hazards model for occupational and environmental case-control analyses. *BMC Medical Research Methodology*, 13(1):18, 2013.

[211] Sterne J.A.C., Hernán M.A., Ledergerber B., Tilling K., Weber R., Sendi P., Rickenbach M., Robins J.M., and Egger M. Long-term effectiveness of potent antiretroviral therapy in preventing AIDS and death: a prospective cohort study. *The Lancet*, 366(9483):378–384, 2005.

[212] Essebag V., Platt R.W., Abrahamowicz M., and Pilote L. Comparison of nested case-control and survival analysis methodologies for analysis of time-dependent exposure. *BMC Medical Research Methodology*, 5 (1):5, 2005.

[213] Dafni U. Landmark analysis at the 25-year landmark point. *Circulation: Cardiovascular Quality and Outcomes*, 4(3):363–371, 2011.

[214] Giobbie-Hurder A., Gelber R.D., and Regan M.M. Challenges of

guarantee-time bias. *Journal of Clinical Oncology*, 31(23):2963–2969, 2013.

[215] Austin P.C. and Platt R.W. Author's response: the design of observational studies-defining baseline time. *Journal of Clinical Epidemiology*, 63(2):141, 2010.

[216] Wang O., Kilpatrick R.D., Critchlow C.W., Ling X., Bradbury B.D., Gilbertson D.T., Collins A.J., Rothman K.J., and Acquavella J.F. Relationship between epoetin alfa dose and mortality: findings from a marginal structural model. *Clinical Journal of the American Society of Nephrology*, 5(2):182–188, 2010.

[217] Aalen O.O. Armitage lecture 2010: understanding treatment effects: the value of integrating longitudinal data and survival analysis. *Statistics in Medicine*, 31(18):1903–1917, 2012.

[218] Howe C.J., Cole S.R., Chmiel J.S., and Muñoz A. 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, 2011.

[219] Wolkewitz M., Beyersmann J., Gastmeier P., and Schumacher M. Efficient risk set sampling when a time-dependent exposure is present. *Methods of Information in Medicine*, 48:438–43, 2009.

[220] Platt R.W., Schisterman E.F., and Cole S.R. Time-modified confounding. *American Journal of Epidemiology*, 170(6):687–694, 2009.

[221] Robins J.M. 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):1393–1512, 1986.

[222] Diggle P., Heagerty P., Liang K., and Zeger S. Time-dependent covariates. In *Analysis of longitudinal data*, pages 245–281. Oxford University Press, 2002.

[223] Bembom O. and van der Laan M.J. Statistical methods for analyzing sequentially randomized trials. *Journal of the National Cancer Institute*, 99(21):1577–1582, 2007.

[224] Van der Wal W.M., Prins M., Lumbreras B., and Geskus R.B. A simple g-computation algorithm to quantify the causal effect of a secondary illness on the progression of a chronic disease. *Statistics in Medicine*, 28(18):2325–2337, 2009.

[225] Snowden J.M., Rose S., and Mortimer K.M. Implementation of g-computation on a simulated data set: demonstration of a causal inference technique. *American Journal of Epidemiology*, 2011.

[226] Daniel R.M., Cousens S.N., De Stavola B.L., Kenward M.G., and Sterne J.A.C. Methods for dealing with time-dependent confounding. *Statistics in Medicine*, 32(9):1584–1618, 2013.

[227] Austin P.C. Using ensemble-based methods for directly estimating causal effects: an investigation of tree-based g-computation. *Multivariate Behavioral Research*, 47(1):115–135, 2012.

[228] Taubman S.L., Robins J.M., Mittleman M.A., and Hernán M.A. Intervening on risk factors for coronary heart disease: an application of the parametric g-formula. *International Journal of Epidemiology*, 38 (6):1599–1611, 2009.

[229] Daniel R.M., De Stavola B.L., and Cousens S.N. gformula: Estimating causal effects in the presence of time-varying confounding or mediation using the g-computation formula. *The Stata Journal*, 11(4):479, 2011.

[230] Daniel R.M., De Stavola B.L., and Cousens S.N. Time-varying confounding: some practical considerations in a likelihood framework. In *Causality: statistical perspectives and applications*, pages 234–252. John Wiley & Sons, 2012.

[231] Westreich D., Cole S.R., Young J.G., Palella F., Tien P.C., Kingsley L., Gange S.J., and Hernán M.A. The parametric g-formula to

estimate the effect of highly active antiretroviral therapy on incident AIDS or death. *Statistics in Medicine*, 31(18):2000–2009, 2012.

[232] Cole S.R., Richardson D.B., Chu H., and Naimi A.I. Analysis of occupational asbestos exposure and lung cancer mortality using the g formula. *American Journal of Epidemiology*, 177(9):989–996, 2013.

[233] Garcia-Aymerich J., Varraso R., Danaei G., Camargo C.A., and Hernán M.A. Incidence of adult-onset asthma after hypothetical interventions on body mass index and physical activity: An application of the parametric g-formula. *American Journal of Epidemiology*, 179 (1):20–26, 2014.

[234] Cain L.E., Robins J.M., Lanoy E., Logan R., Costagliola D., and Hernán M.A. When to start treatment? a systematic approach to the comparison of dynamic regimes using observational data. *The International Journal of Biostatistics*, 6(2), 2010.

[235] Cain L.E., Logan R., Robins J.M., Sterne J.A., Sabin C., Bansi L., Justice A., Goulet J., van Sighem A., de Wolf F., et al. When to initiate combined antiretroviral therapy to reduce rates of mortality and AIDS in HIV-infected individuals in developed countries. *Annals of Internal Medicine*, 154(8):509–515, 2011.

[236] Ewings F.M., Ford D., Walker A.S., Carpenter J., and Copas A. Optimal CD4 Count for Initiating HIV Treatment: Impact of CD4 Observation Frequency and Grace Periods, and Performance of Dynamic Marginal Structural Models. *Epidemiology*, 25(2):194–202, 2014.

[237] Young J.G., Cain L.E., Robins J.M., O'Reilly E.J., and Hernán M.A. Comparative effectiveness of dynamic treatment regimes: an application of the parametric g-formula. *Statistics in Biosciences*, 3(1): 119–143, 2011.

[238] Schomaker M., Egger M., Ndirangu J., Phiri S., Moultrie H., Technau K., Cox V., Giddy J., Chimbetete C., and Wood R. When to

start antiretroviral therapy in children aged 2–5 years: A collaborative causal modelling analysis of cohort studies from southern africa. *PLoS Medicine*, 10(11):e1001555, 2013.

[239] Simon J.H., Jacobs L.D., Campion M., Wende K., Simonian N., Cookfair D.L., Rudick R., Herndon R., Richert J., and Salazar A. The Multiple Sclerosis Collaborative Research Group. Magnetic resonance studies of intramuscular interferon beta-1a for relapsing multiple sclerosis. *Annals of Neurology*, 43(1):79–87, 1998.

[240] Gill R.D. Understanding Cox's regression model: a martingale approach. *Journal of the American Statistical Association*, 79(386):441–447, 1984.

[241] Therneau T.M. Extending the Cox Model. Technical report, Section of Biostatistics, Mayo Clinic, Rochester, 1998. URL http://mayoresearch.mayo.edu/mayo/research/biostat/upload/58.pdf, Last accessed: June-05-2014.

[242] Cole S.R., Hudgens M.G., Tien P.C., Anastos K., Kingsley L., Chmiel J.S., and Jacobson L.P. 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, 2012.

[243] Howe C.J., Cole S.R., Mehta S.H., and Kirk G.D. Estimating the effects of multiple time-varying exposures using joint marginal structural models: alcohol consumption, injection drug use, and HIV acquisition. *Epidemiology*, 23(4):574–582, 2012.

[244] Cole S.R., Jacobson L.P., Tien P.C., Kingsley L., Chmiel J.S., and Anastos K. Using marginal structural measurement-error models to estimate the long-term effect of antiretroviral therapy on incident AIDS or death. *American Journal of Epidemiology*, 171(1):113–122, 2010.

[245] Choi H.K., Hernán M.A., Seeger J.D., Robins J.M., and Wolfe F.

Methotrexate and mortality in patients with rheumatoid arthritis: a prospective study. *The Lancet*, 359(9313):1173–1177, 2002.

[246] Horvitz D.G. and Thompson D.J. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

[247] Coffman D.L., Caldwell L.L., and Smith E.A. Introducing the at-risk average causal effect with application to HealthWise South Africa. *Prevention Science*, 13(4):437–447, 2012.

[248] Lumley T. *survey: Analysis of complex survey samples*, 2011. R package version 3.26.

[249] Therneau T. *A Package for Survival Analysis in S*, 2014. URL `http://CRAN.R-project.org/package=survival`. R package version 2.37-7, Last accessed: Sep-15,2014.

[250] Curtis L.H., Hammill B.G., Eisenstein E.L., Kramer J.M., and Anstrom K.J. Using inverse probability-weighted estimators in comparative effectiveness analyses with observational databases. *Medical Care*, 45(10):S103–S107, 2007.

[251] Cole S.R., Hernán M.A., Margolick J.B., Cohen M.H., and Robins J.M. Marginal structural models for estimating the effect of highly active antiretroviral therapy initiation on CD4 cell count. *American Journal of Epidemiology*, 162(5):471–478, 2005.

[252] Rothman K.J. and Suissa S. Exclusion of immortal person-time. *Pharmacoepidemiology and Drug Safety*, 17(10):1036–1036, 2008.

[253] Therneau T.M. *Modeling survival data: extending the Cox model*. Springer, 2000.

[254] Sylvestre M., Edens T., MacKenzie T., and Abrahamowicz M. *Package 'PermAlgo'*, 2010. URL `http://cran.r-project.org/web/packages/PermAlgo/`. Last accesses: Sep-10-2014.

Appendix A

Appendix for Chapter 2

A.1 Rationale Behind Hypothesizing that Cumulative Relapses are Lying on the Causal Path of β-IFN and Disability Progression

The exact mechanism of action of the β-IFN drugs in MS has never been fully established and is one reason why estimating the effect of these drugs in MS is not straightforward. In the absence of randomization, establishing a causal link between drug exposure and outcome requires subject-specific knowledge and careful implementation of that knowledge in the analysis. Suggesting a plausible causal path is the first step.

Relapsing-remitting patients experience relapses followed by periods of remission in which partial or complete recovery occurs. Based on the results from randomized, double-blind, placebo-controlled studies, β-IFN treatments reduced the severity and frequency of relapses [2, 4–6, 239] and hence increased the period between relapses [5].Consequently, a patient has more time to recover from the residual disability left by the past relapse. This extended period of relapse-free time due to β-IFN exposure may eventually contribute to a slower progression of disability [4, 5, 239]. However, it should be noted that while most natural history studies indicate that long-term there is minimal or no association between relapse rates and disability progression, a specific window of opportunity for relapses to contribute to disease progression may exist [107, 108].

Therefore, we hypothesized that within a short time interval the cumulative relapses are acting as an intermediate variable for the treatment and disability progression relationship, i.e., the relapse frequency is influenced by prior β-IFN treatment and a greater (lesser) relapse frequency will result in faster (slower) disability progression. Also, we assume that the cumulative relapse count in the previous time period is a confounder that may dictate the treatment choice in subsequent time periods. Furthermore, experiencing an increased number of cumulative relapses after initiating treatment will increase the probability of discontinuing treatment [100]. Hence, in this relationship, cumulative relapse is treated both as an intermediate variable and a confounder.

The causal path described above could be considered as rather simplistic. It is possible that cumulative relapse and disability progression have an unmeasured common cause (for example, low serum vitamin D levels). Should this data be available, then we would add that variable to the causal path between cumulative relapse and EDSS. Cumulative relapse would still be a time-dependent confounder and would need to be adjusted for accordingly.

A.2 Rationale Behind Using Marginal Structural Cox Model (MSCM) Instead of a Cox Model

For a longitudinal study with N patients, let $i = 1, 2, \ldots, N$ be the patient index, $t = 0, 1, \ldots, T_i$ months be the follow-up time index, A_{it} be the binary treatment status at month t (1 = treated, 0 = untreated), and L_{i0} be the baseline covariates of patient i. One possible model would express the hazard function of the time-dependent Cox model as follows:

$$\lambda_i(t|L_{i0}) = \lambda_{0t} \exp\big(\beta_1 A_{it} + \beta_2 L_{i0}\big), \tag{A.1}$$

here λ_{0t} is the unspecified baseline hazard function, β_2 is the vector of log hazard ratios (HRs) for the baseline covariates and β_1 is the log HR of the current β-IFN status (A_{it}).

Assuming no tied event times, we estimate $\beta = (\beta_1, \beta_2)$ by maximizing the partial likelihood [240]:

$$PL(\beta) = \prod_{i=1}^{N} \prod_{t=0}^{T_i} \left(\frac{Y_{it} \exp\left(\beta_1 A_{it} + \beta_2 L_{i0}\right)}{\sum_{k=1}^{N} Y_{kt} \exp\left(\beta_1 A_{kt} + \beta_2 L_{k0}\right)} \right)^{dN_{it}},$$

where Y_{it} denotes whether patient i belongs to the risk set at time t, N_{it} is the number of events in the interval $[0, t]$ and dN_{it} denotes the number of new events for patient i at month t (increment from month $t-1$, if any). This setting is more general than our case, where $N_{it} \leq 1$ and $dN_{it} = 1$ for at most 1 month.

However, ignoring the time-dependent confounder L_{it} (i.e., an intermediate variable lying in the causal pathway of the treatment and the outcome) may lead to a biased estimate of β. Simply including this variable in the Cox model as a covariate as,

$$\lambda_i(t|L_{i0}, L_{it}) = \lambda_{0t} \exp\left(\beta_1 A_{it} + \beta_2 L_{i0} + \beta_3 L_{it}\right), \tag{A.2}$$

may still produce a biased estimate if L_{it} is influenced by past exposure [50].

Inverse probability of treatment and censoring weights (IPTC; say w, sw, $w^{(n)}$, $sw^{(n)}$) are person-time specific measures of the degree to which a time-dependent variable confounds the treatment selection and censoring processes. These are used in the time-dependent Cox model to weight the contribution of each person-time observation so that confounding due to L_{it} is removed without changing the target parameter. In this way, MSCM facilitates correction for time-dependent confounding. In the MSCM, these IPTC weights are inserted in the partial likelihood function as follows [241–243]:

$$PL_w(\beta) = \prod_{i=1}^{N} \prod_{t=0}^{T_i} \left(\frac{Y_{it} \exp\left(\beta_1 A_{it} + \beta_2 L_{i0}\right)}{\sum_{k=1}^{N} Y_{kt} w_{kt} \exp\left(\beta_1 A_{kt} + \beta_2 L_{k0}\right)} \right)^{dN_{it} \times w_{it}}.$$

The gradient with respect to the parameter vector β of the log of the weighted partial likelihood $PL_w(\beta)$ yields the score function $U_w(\beta)$. Equating $U_w(\beta)$ to zero yields a set of estimating equations that can be solved using an iterative method such as the Newton-Raphson algorithm or a penalized partial likelihood approach.

A.3 Approximation of the Marginal Structural Cox Model

Let D_t be an indicator of reaching EDSS 6 for the first time between the months $t-1$ and t. The data for patients who did not reach sustained EDSS 6 and remained uncensored until follow-up month t can be modelled using the pooled logistic regression (logistic regression pooled over persons and times):

$$logit\ Pr(D_{i,t} = 1 | D_{i,t-1} = 0, A_{it}, L_{i0}) = \gamma_0(t) + \gamma_1 A_{it} + \gamma_2 L_{i0}. \qquad \text{(A.3)}$$

Here $\gamma_0(t)$ is a smooth function of the month index t, represented as a restricted cubic spline, which is often used to reduce weight variability. Just as for cubic polynomial regression, use of a restricted cubic spline forces the relationship to be smooth even on the edges [117, chapter 6]; see the `R` code in the Appendix §A.5. The log OR of the current β-IFN status in this pooled logistic regression, γ_1, is generally a good approximation of the corresponding log hazard ratio obtained from the time-dependent Cox model (β_1), provided that censoring is ignorable [244] and relatively short intervals are chosen so that the probability of outcome occurrence in each time interval is small [118, 119]. The corresponding likelihood function can be expressed as:

$$L(\gamma) \quad = \quad \prod_{i=1}^{N} \prod_{t=0}^{T_i} p_{it}^{D_{it}} (1 - p_{it})^{(1-D_{it})},$$

where $\gamma = (\gamma_0, \gamma_1, \gamma_2)$ and $logit(p_{it}) = \gamma_0(t) + \gamma_1 A_{it} + \gamma_2 L_{i0}$.

Hernán et al. [50] suggested use of weighted pooled logistic regression to approximate MSCM (IPTC weighted time-dependent Cox model) estimates of treatment effect (β_1) and others have followed this suggestion. [69, 112, 159, 211, 245]. The weighted likelihood function is then written as [244]:

$$L_w(\gamma) \quad = \quad \prod_{i=1}^{N} \prod_{t=0}^{T_i} \left(p_{it}^{D_{it}} (1 - p_{it})^{(1-D_{it})} \right)^{w_{it}}.$$

This approximate approach was suggested mainly because software available at that time was unable to handle patient-specific time-varying weights in a Cox model. It has been noted that this approximation approach is inadequate when the event is not rare [56]. Subsequently Xiao et al. [57] suggested the direct use of the Cox model weighted by IPTC weights to overcome this limitation. Through simulation, these authors also showed that direct use of the Cox model weighted by IPTC weights instead of any approximate MSCM approach [50] considerably reduced the variability of the estimated treatment effect, even when both methods use the same weights.

A.4 Weight Models Used in the Data Analysis

The stabilized IPT weights for patient i at month t are expressed as:

$$sw_{it}^{T} \quad = \quad \prod_{j=0}^{t} \frac{pr(A_{ij} = a_{ij} | \bar{A}_{i,j-1} = \bar{a}_{i,j-1}, L_{i0} = l_{i0})}{pr(A_{ij} = a_{ij} | \bar{A}_{i,j-1} = \bar{a}_{i,j-1}, L_{i0} = l_{i0}, \bar{L}_{ij} = \bar{l}_{ij})}. \quad \text{(A.4)}$$

The probability appearing in the numerator of sw^T is modeled using pooled logistic model as follows:

$$logit\ Pr(A_{ij} = 1 | \bar{A}_{i,j-1}, L_{i0}) = \alpha_0(j) + \alpha_1 A_{i,j-1} + \alpha_2 L_{i0}, \quad \text{(A.5)}$$

where treatment status at the previous time interval (A_{j-1}; $A_{-1} = 0$ for all patients), the baseline covariates (L_0; in our application, EDSS, age,

Table A.1: Estimated coefficients from the treatment model (denominator of sw_{it}^{T}) for patients with relapsing-onset multiple sclerosis (MS), British Columbia, Canada (1995-2008)

	Estimate	z-value	p-value
β-IFN$_{j-1}$	9.78	102.92	< 0.001
EDSS†	0.12	4.31	< 0.001
Age†‡	-0.07	-1.70	0.09
Disease duration†‡	-0.17	-3.17	< 0.001
Sex†	-0.07	-0.96	0.34
Cumulative relapse	0.34	7.70	< 0.001
Cumulative relapse:β-IFN$_{j-1}$	-0.55	-10.83	< 0.001

EDSS, expanded disability status scale.
* Time index is also fitted with restricted cubic spline, but the corresponding coefficients are not reported in the table.
† Baseline covariates (L_0).
‡ Expressed in decades.

disease duration, sex) and a restricted cubic spline of the follow-up month index are included as predictors. These covariates, as well as the time-varying confounder cumulative relapse (L_{ij}) and its interaction with prior treatment status are included in the denominator model:

$$\begin{aligned} logit\ Pr(A_{ij} = 1|\bar{A}_{i,j-1}, L_{i0}, \bar{L}_{ij}) &= \alpha_0(j) + \alpha_1 A_{i,j-1} + \alpha_2 L_{i0} + \\ &\quad \alpha_3 L_{ij} + \alpha_{13} A_{i,j-1} L_{ij}. \end{aligned} \tag{A.6}$$

The output of this fit is reported in Appendix-Table A.1.

The predicted value from the (denominator) model (A.6) yields the estimated probability of the patient's treatment status in that month t. Since the exposure status may vary from one time point to another, first we estimate the probability of the observed treatment status at each time point, and then obtain the probability of the observed exposure sequence of a given patient by multiplying the corresponding probabilities. The numerator of

sw_{it}^T is estimated in a similar fashion from model (A.5), where $\bar{L}_{ij}$ is not included as a predictor. Dividing the numerator model probabilities of the patient's observed treatment status a_{ij} (either 0 or 1) by the corresponding denominator model probabilities yields the estimated IPT weights sw_{it}^T that account for the confounding due to $\bar{L}_{ij}$, given the required assumptions are met.

To estimate the IPTC weights $sw_{it} = sw_{it}^T \times sw_{it}^C$, the inverse probability of censoring (IPC) weights sw_{it}^C are estimated in the same fashion. In order to produce the normalized IPTC weights $sw^{(n)}$, each weight sw is divided by its risk set's mean weight.

A.5 MSCM fitting in R

For time-dependent survival analysis, all person-time observations are pooled to make an augmented dataset. Short intervals, such as months, are chosen so that the most recently observed changes of the time-varying variables can be updated in a new row in the dataset to reflect the patient's time-varying status with respect to covariates, censoring and response. In the longitudinal analysis literature, this is referred to as the 'long' format.

Guidelines regarding IPTC weight calculations in R are available in the literature [159]. These IPTC weights can be viewed as a generalization of the Horvitz-Thompson estimator [141, 246, 247]. Recently, due to the availability of packages for the analysis of complex surveys in standard software (SAS, Stata and R), it is possible to fit the time-dependent IPTC weighted Cox model directly or via approximation, say, using the weighted pooled logistic model. In all the model choices, reliable SEs can be obtained from a reasonable number of patient-specific bootstrap samples.

- Most MSCM analyses in the literature use weighted pooled logistic regression to approximate the IPTC weighted Cox model fit. In R,

performing weighted pooled logistic regression using the `glm` function from the `base` package (with `log` link) is straightforward [159].

- Similarly, the `svyglm` function from the `survey` package can be used to implement the (weighted) pooled logistic model [247].

- With data organized in person-month format, to perform survival analysis using the weighted Cox model, we used the Andersen-Gill's counting process approach as implemented in the `svycoxph` function from the `R` package `survey` [248] with the `weights` option. Approximation via complementary-log-log and Poisson models can also be implemented using the same package. A sample code follows:

```
require(survey)
require(rms)
(weighted.design<-svydesign(id=~ID, data=long.format,
  weight=~normalized.stabilized.weight))
svycoxph(Surv(start, stop, event) ~ drug + covariate.list,
  design=weighted.design)
svyglm(event ~ drug + rcs(Time) + covariate.list,
  family=binomial(link=log),  design=weighted.design)
svyglm(event ~ drug + rcs(Time) + covariate.list,
  family=binomial(link=cloglog),  design=weighted.design)
svyglm(event ~ offset(log(stop-start))+ drug + rcs(Time) +
  covariate.list, family=poisson(),  design=weighted.design)
```

- Alternatively, the `coxph` function from the `survival` package [249] can be used to fit the weighted Cox model [57]. To handle correlated observations, the `cluster` option must be specified to identify the person-month observations from the same patient. Robust SEs are obtained by specifying the option `robust = TRUE`.

Table A.2: Characteristics of the selected cohort of patients with relapsing-onset multiple sclerosis (MS), British Columbia, Canada (1995-2008).

Characteristics	β-IFN exposed patients	β-IFN unexposed patients
Frequency	868	829
Women, n (%)	660 (76.0)	637 (76.8)
Disease duration (at baseline)	5.8[†](6.6[‡])	8.3[†](8.5[‡])
Age (at baseline)	38.1[†](9.2[‡])	41.3[†](10.0[‡])
EDSS score (at baseline)	2.0[§](0-6.5[¶])	2.0[§](0-6.5[¶])
Relapse rate / year (over the 2 years prior to baseline)[#]	0.5[§](0-1.2[#])	0.5[§](0-1.0[#])
Person-years exposed to β-IFN treatment	2,530	0
Person-years not exposed to β-IFN treatment	1,400	2,960

[†] Mean.
[‡] Standard deviation.
[§] Median.
[¶] Range.
[#] IQR.

A.6 Exclusion Criteria and Summary of Selected Cohorts

In total, 2,671 patients met the eligibility criteria to receive β-IFN treatment between July 1995 and December 2004 [92]. Of these, patients who were exposed to a non-β-IFN immunomodulatory drug, a cytotoxic immunosuppressant for MS ($n = 172$), or an MS clinical trial ($n = 21$) prior to baseline were excluded from the analysis. If the exposure occurred after baseline, data were censored at at the start of the exposure of the non--IFN treatment. Other exclusion criteria included unknown MS onset date ($n = 10$), insufficient EDSS measurements ($n = 436$), reaching of the outcome ($n = 218$) or the secondary progressive stage before the eligibility date ($n = 217$). Some patients met multiple exclusion criteria. As a result, $1,697$

patients were selected. A summary of their characteristics are reported in Table A.2.

A.7 Sensitivity Analyses

A.7.1 Sensitivity Analysis: Impact of Weight Trimming

If the weights contain extreme values, one should be concerned about the positivity assumption. The MSCM approach is built on the counterfactual framework and it is necessary to assume patients could choose treatment exposure or non-exposure at any time point. If a group of patients with similar covariate history rarely or never receive treatment, then the estimated probability of being treated would be close to zero. Conversely, if a group of patients with similar covariate history almost always or always receive treatment, then the estimated probability of being treated would be close to one. Then the corresponding fitted probability will be close to zero or one resulting in a very large or small inverse probability weight respectively. This may produce unstable estimates from the MSCM.

As a sensitivity analysis, one could restrict the analysis to the subset of patients that have probability of treatment and censoring that is reasonably removed from 0 and 1 at every time point. This procedure is known as trimming [120]. As with truncation of the weights, systematically excluding such patients may produce a biased estimate. Also, the interpretation may lack generalizability due to this restriction. However, since the patients with extreme weights are removed, a relatively stable point estimate with a smaller CI is expected.

After estimating the fitted probabilities from the weight models, if the probabilities are such that a few person-time observations are contributing too much in the pseudo-population, this may make the estimate of the causal effect unstable. In our sensitivity analysis, we removed the patients with at least one fitted value either greater than 0.95 or less than 0.05 (represented

more than 20 times in the pseudo-population). This left 1,603 patients, with 133 reaching the outcome. MSCM using $sw^{(n)}$ lead to a HR estimate of 1.33 with a 95% bootstrap CI of $0.94-1.89$. The conclusion regarding the treatment effect of β-IFN on time to sustained EDSS 6 from these results remained the same.

A.7.2 Sensitivity Analysis: Impact of More Restrictive Eligibility Criteria

Table A.3: The marginal structural Cox model (MSCM) fit with the normalized stabilized IPTC weights $sw^{(n)}$ for time to sustained EDSS 6 to estimate the causal effect of β-IFN treatment for patients with relapsing-onset multiple sclerosis (MS), British Columbia, Canada (1995-2008) selected by more restrictive eligibility criteria. The model was also adjusted for baseline covariates EDSS, age, disease duration and sex.

Covariate	Estimate*	HR †	95% CI ‡
β-IFN	0.18	1.19	0.68 - 2.11
EDSS	0.40	1.48	1.24 - 1.77§
Disease duration#	-0.14	0.87	0.55 - 1.37
Age#	0.45	1.57	1.14 - 2.18§
Sex¶	-0.33	0.72	0.38 - 1.35

HR, Hazard ratio; CI, confidence interval; EDSS, expanded disability status scale.

* Estimated log HR: negative value is indicative of a beneficial effect and positive value is indicative of a harmful effect.

† HR, indicating the instantaneous risk of reaching sustained and confirmed EDSS 6.

‡ Based on 500 nonparametric bootstrap sample estimates.

§ 95% CI that does not include 1.

Expressed in decades.

¶ Reference level: Male.

As another sensitivity analysis, a more restricted study sample was selected by defining active disease (two or more documented relapses during the two years prior to baseline) as part of the eligibility criteria, while also

including all the previous criteria. This left 747 patients in the study with 3,028 person-years of follow-up and 1,460 person-years of -IFN exposure. Only 52 of these patients reached the irreversible disease outcome.

The model fit is reported in Appendix-Table A.3. The regression coefficients and HR estimates were qualitatively similar to those reported in Table 2. The CIs from this restricted dataset were wider due to the smaller sample size. Still, the conclusion regarding the treatment effect of β-IFN on time to sustained EDSS 6 remained the same as before.

A.7.3 Sensitivity Analysis: Impact of the Cumulative Exposure to β-IFN

Table A.4: The marginal structural Cox model (MSCM) fit with the normalized stabilized IPTC weights $sw^{(n)}$ for time to sustained EDSS 6 to estimate the causal effect of cumulative exposure to β-IFN over the last two years for patients with relapsing-onset multiple sclerosis (MS), British Columbia, Canada (1995-2008). The model was also adjusted for baseline covariates EDSS, age, disease duration and sex.

Covariate	Estimate	HR †	95% CI ‡
Cumulative β-IFN*	0.53	1.70	0.64 - 4.53
EDSS	0.54	1.71	1.53 - 1.91 §
Disease duration#	-0.20	0.82	0.66 - 1.10
Age#	0.30	1.34	1.10 - 1.63 §
Sex¶	-0.23	0.79	0.55 - 1.15

HR, Hazard ratio; CI, confidence interval; EDSS, expanded disability status scale.

* Expressed as proportion of months exposed over last two years.

† HR, indicating the instantaneous risk of reaching sustained and confirmed EDSS 6.

‡ Based on 500 nonparametric bootstrap sample estimates.

§ 95% CI that does not include 1.

Expressed in decades.

¶ Reference level: Male.

We also assessed the impact of the cumulative exposure to β-IFN (proportion of months exposed) over the last two years on time to sustained EDSS 6. The model fit is reported in Appendix-Table A.4. This analysis also failed to detect a significant association between the cumulative exposure to β-IFN and the hazard of reaching sustained EDSS 6. A similar finding was observed when the cumulative exposure was restricted to the past one year only (data not shown).

A.7.4 Sensitivity Analysis: Impact of the Cumulative Number of Relapses in the Last Year

Table A.5: The marginal structural Cox model (MSCM) fit with the normalized stabilized IPTC weights $sw^{(n)}$ for time to sustained EDSS 6 to estimate the causal effect of cumulative exposure to β-IFN over the last two years for patients with relapsing-onset multiple sclerosis (MS), British Columbia, Canada (1995-2008) while considering the cumulative number of relapses in the last year as the time-varying confounder. The model was also adjusted for baseline covariates EDSS, age, disease duration and sex.

Covariate	Estimate	HR †	95% CI ‡
β-IFN*	0.31	1.36	0.96 - 1.92
EDSS	0.54	1.72	1.54 - 1.92 §
Disease duration#	-0.18	0.82	0.66 - 1.04
Age#	0.28	1.32	1.10 - 1.60 §
Sex¶	-0.22	0.80	0.55 - 1.16

HR, Hazard ratio; CI, confidence interval; EDSS, expanded disability status scale.

* Expressed as proportion of months exposed over last two years.

† HR, indicating the instantaneous risk of reaching sustained and confirmed EDSS 6.

‡ Based on 500 nonparametric bootstrap sample estimates.

§ 95% CI that does not include 1.

Expressed in decades.

¶ Reference level: Male.

We also assessed the impact of the exposure to β-IFN on time to sus-

tained EDSS 6 while considering the cumulative number of relapses in the last year (instead of the last two years) as the time-varying confounder. The model fit is reported in Appendix-Table A.5. This analysis also failed to detect a significant association between the exposure to β-IFN and the hazard of reaching sustained EDSS 6.

Appendix B

Appendix for Chapter 3

B.1 Propensity Scores

A confounder is a factor that is affects both the treatment decision and the study outcome. Propensity score techniques facilitate simultaneous adjustment for multiple confounders in observational or non-experimental settings. The propensity score p_i is defined as a subject's probability of receiving a treatment ($A_{i0} = 1$) conditional on a number of covariates ($L_{i0} = l_{i0}$) present at baseline [20]. Under the assumption of no unmeasured confounding, the treated and untreated subjects with the same propensity score will have identical distributions of baseline confounders. To balance the covariate distribution, treated and untreated subjects are selected by matching the estimated propensity scores or stratifying on the basis of the estimated propensity scores quantiles. If covariate balance is lacking in the original sample, excluding the subjects without overlapping propensity scores can restore the balance. That is why the propensity score is known as a balancing score.

Propensity scores can be used to determine the inverse probability weights (IPW) that are inversely proportional to the probability of the observed exposure status, conditional on confounders. These weights can incorporate not only the baseline covariates, but also the covariates that include post-baseline values. That is, if p_i is the propensity score, then $1/p_i$ is the weight for the exposed subject and $1/(1-p_i)$ is the weight for the unexposed subject. IPW-based estimators can be generalized to multiple exposure categories, to accommodate survival or censored data and to incorporate time-dependent exposure and covariates [42, 250]. However, the ability to deal with com-

plex problems comes at a price. The effect estimates from IPW methods can be unstable and the estimated variance needs to account for the weighted (pseudo) data. Methods exist for stabilizing the weights (see appendix B.3) and robust variance estimation methods can be used to account for weighted data [50].

B.2 Model Specification in MSCM

Based on the notation described in the text (see §3.2), in the presence of baseline covariates L_{i0}, the hazard function can be expressed as the following time-dependent Cox model:

$$\lambda_{i,\bar{A}_m}(m|L_{i0}) \quad = \quad \lambda_{\bar{0}}(m)\exp\big(\gamma(m,\bar{A}_m,\boldsymbol{\psi})\big) \tag{B.1}$$

where m is the visit index, $\lambda_{\bar{0}}(m)$ is the unspecified baseline hazard function, $\boldsymbol{\psi} = (\psi_1, \psi_2)$, ψ_1 is the log HR of the current treatment status (A_{im}), and ψ_2 is the vector of log hazard ratios (HRs) for the baseline covariates. Specifying the model for $\gamma(m, \bar{A}_m, \psi_1)$ yields:

$$\lambda_{i,\bar{A}_m}(m|L_{i0}) \quad = \quad \lambda_{\bar{0}}(m)\exp\big(\psi_1 A_{im} + \psi_2 L_{i0}\big), \tag{B.2}$$

where the impact of treatment is modelled based on only current exposure A_m (i.e., the dependence on $\bar{A}_m$ is modelled only through the current exposure A_m) [50].

In the presence of a time-dependent confounder L_{im}, we may be tempted to expand the above Cox model to:

$$\lambda_{i,\bar{A}_m}(m|L_{i0}, L_{im}) = \lambda_{\bar{0}}(m)\exp\big(\psi_1 A_{im} + \psi_2 L_{i0} + \psi_3 L_{im}\big),$$

which could still produce a biased estimate of ψ_1 if L_{im} is influenced by past exposure [50]. Nonetheless, as L_{im} is a confounder, we still need to adjust for confounding due to L_{im} somehow. IPWs are person-time specific measures of the degree to which L_{im} confounds the treatment selection pro-

cess. Therefore, in MSCM, IPWs are used in the time-dependent Cox model formulation (equation (B.2)) to weight the contribution of each person-time observation so that the confounding due to L_{im} is removed.

B.3 Model Specifications for Estimating the Weights

The unstabilized IPWs for subject i at month m are expressed as:

$$w_{im} = \prod_{j=0}^{m} \frac{1}{pr(A_{ij} = a_{ij} | \bar{A}_{i,j-1} = \bar{a}_{i,j-1}, L_{i0} = l_{i0}, \bar{L}_{ij} = \bar{l}_{ij})}, \quad \text{(B.3)}$$

As discussed in § A.4, we can estimate the probabilities in equation (B.3) by building a pooled logistic regression model for current treatment status (A_j) with the following covariates: treatment status at the previous time interval (A_{j-1}), the baseline covariates (L_0), the follow-up time index, and the time-varying confounder (L_{ij}) as follows:

$$\begin{aligned} logit\ Pr(A_{ij} = 1 | \bar{A}_{i,j-1}, L_{i0}, \bar{L}_{ij}, \boldsymbol{\alpha}) &= \alpha_0(j) + \alpha_1 A_{i,j-1} + \\ &\quad \alpha_2 L_{i0} + \alpha_3 L_{ij}, \end{aligned} \quad \text{(B.4)}$$

where $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2, \alpha_3)$.

Adding interaction terms to equation (B.4) enables us to capture the realistic scenario that the status of the confounder at time j (i.e., $L_{ij} = 0$ or 1) can potentially influence a switch onto treatment ($A_{i,j-1} = 0, A_{i,j} = 1$) differently than a switch off treatment ($A_{i,j-1} = 1, A_{i,j} = 0$), depending on the treatment status at the previous time period (i.e., $A_{i,j-1}$). In our implementation, the denominator terms are estimated from:

$$\begin{aligned} logit\ Pr(A_{ij} = 1 | \bar{A}_{i,j-1}, L_{i0}, \bar{L}_{ij}, \boldsymbol{\alpha}) &= \alpha_0(j) + \alpha_1 A_{i,j-1} + \alpha_2 L_{i0} + \\ &\quad \alpha_3 L_{ij} + \alpha_{13} A_{i,j-1} L_{ij}, \end{aligned} \quad \text{(B.5)}$$

where $\boldsymbol{\alpha}$ now includes α_{13} as well. Since we are using only the last value of the treatment history ($\bar{A}_{i,j-1} = A_{i,j-1}$) in equation (B.5), it is possible to simplify this equation by considering treatment status at the $(j-1)$-th time, i.e., whether the patient was treated ($A_{i,j-1} = 1$) or not ($A_{i,j-1} = 0$) as follows:

$$\begin{aligned} logit\ Pr(A_{ij} = 1 | A_{i,j-1} = 1, L_{i0}, L_{ij}) &= \{\alpha_0(j) + \alpha_1\} + \alpha_2 L_{i0} + \\ &\quad (\alpha_3 + \alpha_{13}) L_{ij}. \\ logit\ Pr(A_{ij} = 1 | A_{i,j-1} = 0, L_{i0}, L_{ij}) &= \alpha_0(j) + \alpha_2 L_{i0} + \alpha_3 L_{ij}. \end{aligned}$$

The predicted probabilities from equation (B.5) yield the estimated probability of the subject's treatment status at time j. Subsequently, we obtain the probability of the observed exposure sequence over m time periods of a given subject by multiplying the corresponding probabilities.

To stabilize this IPW, we use the following general formula:

$$sw_{im} = \prod_{j=0}^{m} \frac{pr(A_{ij} = a_{ij} | \bar{A}_{i,j-1} = \bar{a}_{i,j-1}, L_{i0} = l_{i0})}{pr(A_{ij} = a_{ij} | \bar{A}_{i,j-1} = \bar{a}_{i,j-1}, L_{i0} = l_{i0}, \bar{L}_{ij} = \bar{l}_{ij})}. \quad \text{(B.6)}$$

In our implementation, the numerator terms are estimated from:

$$logit\ Pr(A_{ij} = 1 | \bar{A}_{i,j-1}, L_{i0}) = \alpha_0'(j) + \alpha_1' A_{i,j-1} + \alpha_2' L_{i0}, \quad \text{(B.7)}$$

where no element of $\bar{L}_{ij}$ is included as a predictor.

Dividing the estimated numerator probabilities of the subject's observed treatment status a_{ij} (either 0 or 1) by the corresponding estimated denominator probabilities yields the estimated IPWs sw_{im} that account for the confounding due to $\bar{L}_{im}$.

The formulas for the normalized versions of the IPW are:

$$w_{im}^{(n)} = \frac{w_{im} n_m}{\sum_{i \in r_m} w_{im}}, \qquad sw_{im}^{(n)} = \frac{sw_{im} n_m}{\sum_{i \in r_m} sw_{im}}, \tag{B.8}$$

where r_m denotes the risk-set at time m, n_m denotes the total number of subjects in the risk-set and w and sw are the unstabilized and stabilized IPW weights.

B.4 Implementation of the Statistical Learning Approaches in R

To estimate the weights, the following functions can be used in R:

- We fitted the logistic regressions using the `glm` function from the `base` package (with `logit` link). Sample code is as follows:

```
# Numerator and denominator models
ww <- glm(A ~ m + L + A.lag + L.lag + A.lag*L,
          family = binomial(logit), data = dataset)
ww0 <- glm(A ~ m + A.lag, family = binomial(logit), data = dataset)
# Weight generation by exposure
dataset$wwp <- with(dataset, ifelse(A == 0,
                    1 - fitted(ww), fitted(ww)))
dataset$wwp0 <- with(dataset, ifelse(A == 0,
                    1 - fitted(ww0),fitted(ww0)))
# generating unstabilized and  stabilized weights
dataset$w <- unlist(tapply(1/dataset$wwp, dataset$id, cumprod))
dataset$sw <- unlist(tapply(dataset$wwp0/dataset$wwp,
                     dataset$id, cumprod))
```

  The `lrm` function from the `rms` package can do the same.

- We performed bootstrap aggregation or bagging using the `bagging` function from the `ipred` package. Sample code is as follows:

```
library(rpart)
library(ipred)
# Numerator and denominator models
ww <- bagging(A ~ m + L + A.lag + L.lag + A.lag*L,
              data=dataset, nbagg=100,
              control=rpart.control(xval=10))
ww0 <- bagging(A ~ m + A.lag, data=dataset, nbagg=100,
               control=rpart.control(xval=10))
# Weight generation by exposure
dataset$wwp <- with(dataset, ifelse(A == 0,
                    1 - predict(ww, type="prob"),
                    predict(ww, type="prob")))
dataset$wwp0 <- with(dataset, ifelse(A == 0,
                     1 - predict(ww0, type="prob"),
                     predict(ww0, type="prob")))
# generating unstabilized and  stabilized weights
dataset$w <- unlist(tapply(1/dataset$wwp, dataset$id, cumprod))
dataset$sw <- unlist(tapply(dataset$wwp0/dataset$wwp,
                     dataset$id, cumprod))
```

- LIBSVM is a popular implementation of the support vector machines algorithm [153]. We can make use of this implementation via the e1071 package in R. We fit SVM using svm function from this package. Sample code is as follows:

```
require(e1071)
# Numerator and denominator models
ww <- svm(as.factor(A) ~ m + as.factor(L) + as.factor(A.lag) +
          as.factor(L.lag) + as.factor(A.lag*L),
          data=dataset, probability=TRUE, kernel = "polynomial")
ww0 <- svm(as.factor(A) ~ m + as.factor(A.lag),
           data=dataset, probability=TRUE, kernel = "polynomial")
# Weight generation by exposure
newdf <- data.frame(m = dataset$m, L = dataset$L,
```

```
                    A.lag = dataset$A.lag, L.lag = dataset$L.lag)
(predw <- predict(ww, newdf, probability = TRUE))
pr.predw <- attr(predw, "prob")[,1]
newdf <- data.frame(m = dataset$m, A.lag = dataset$A.lag)
(predw0 <- predict(ww0, newdf, probability = TRUE))
pr.predw0 <- attr(predw0, "prob")[,1]
dataset$wwp <- with(dataset, ifelse(A == 0,
                    pr.predw, 1-pr.predw))
dataset$wwp0 <- with(dataset, ifelse(A == 0,
                    pr.predw0, 1-pr.predw0))
# generating unstabilized and  stabilized weights
dataset$w <- unlist(tapply(1/dataset$wwp, dataset$id, cumprod))
dataset$sw <- unlist(tapply(dataset$wwp0/dataset$wwp,
                     dataset$id, cumprod))
```

The `ksvm` function in the `kernlab` package or the `svmlight` function in the `klaR` package or the `svmpath` function in the `svmpath` package can also be used to fit SVM.

- We performed boosting using the `ps` function from the `twang` package, which utilizes the `gbm` function from the `gbm` package. Sample code is as follows:

```
require(gbm)
require(twang)
# Numerator and denominator models
ww <- ps(A ~ m + L + A.lag + L.lag, data=dataset,
         interaction.depth=2,
         stop.method="ks.mean", print.level=0,verbose=FALSE)
ww0 <- ps(A ~ m + A, data=dataset, interaction.depth=2,
          stop.method="ks.mean", print.level=0,verbose=FALSE)
# Weight generation by exposure
dataset$wwp <- with(dataset, ifelse(A == 0,
                    1-ww$ps$ks.mean.ATE, ww$ps$ks.mean.ATE))
```

```
dataset$wwp0 <- with(dataset, ifelse(A == 0,
                    1-ww0$ps$ks.mean.ATE, ww0$ps$ks.mean.ATE))
# generating unstabilized and  stabilized weights
dataset$w <- unlist(tapply(1/dataset$wwp, dataset$id, cumprod))
dataset$sw <- unlist(tapply(dataset$wwp0/dataset$wwp,
                     dataset$id, cumprod))
```

The `coxph` function from the `survival` package can be used to fit the MSCM with the `cluster` option specifying the patient identification and option `robust = TRUE` specifying estimation of robust SEs.

B.5 Post-estimation Weight Variability Reduction Techniques

Normalization. Normalization (discussed in Appendix B.3) is a relatively new proposal to change the weights in such a way that, in each risk-set, the variability is reduced while also assuring that the mean weight for each risk-set equals one [57]. Such characteristics of weights in turn contribute to reducing the sampling variability of the estimates of the causal effect. This technique can be applied on both unstabilized and stabilized weights (see equation (B.8) in Appendix B.3). The usefulness of this approach was shown in a simulation setting [57].

Truncation. Weight truncation refers to reducing weights larger than some specified value w_u to w_u and increasing weights smaller than some specified value w_l to w_l. The truncation points (w_l, w_u) are usually selected according to specified weight quantiles (say, 5% and 95%). Truncation generally reduces the variability of the causal effect estimate. When the distribution of the weights is symmetric, higher levels of truncation usually lead the effect estimate to move towards the baseline-adjusted estimate. At 50% truncation, a median weight is assigned, which leads to an estimate similar to the baseline-adjusted estimate [251]. Selecting a suitable level of truncation involves the 'variance-bias-trade-off'. Selection of this

level generally involves data-adaptive methods [104, 158]. We denoted untruncated, 1%, 5%, 25%, 35% and 50% truncated unstabilized weights as $w, w1, w5, w25, w35, w50$ respectively. This notational convention holds for all other weights as well. Many researchers use the terms trimming and truncation interchangeably [120, 121, 251], but we will maintain the definition of 'truncation' above.

B.6 Pseudocode for MSCM Data Simulation

The algorithm proposed by Young et al. [56, 142] generates data that satisfy the conditions of the following three models simultaneously: MSM, structural nested accelerated failure time model and a structural nested cumulative failure time model. The steps of this algorithm are also described elsewhere [56, 57, 114, 160]. We slightly modified the treatment generating models to include an interaction term $(A_{m-1} \times L_m)$ in the treatment generation stage to make it more realistic for many disease settings.

GET

$n \leftarrow 2500$ (large sample) or 300 (small sample);

$K \leftarrow 10$ (maximum follow-up);

$\lambda_0 \leftarrow 0.01$ (rare events) or 0.10 (frequent events);

$\beta \leftarrow [\log(3/7), 2, \log(1/2), \log(3/2)]$ (parameter vector for generating L);

$\alpha \leftarrow [\log(2/7), (1/2), (1/2), \log(4), \log(6/5)]$ (parameter vector for generating A);

$\psi_1 \leftarrow -0.5$ (true log-hazard value of the treatment effect)

COMPUTE

FOR $ID = 1$ to n

INIT: $L_{-1} \leftarrow 0$; $A_{-1} \leftarrow 0$; $Y_0 \leftarrow 0$; $H_m \leftarrow 0$; $c \leftarrow 30$

$T_{\bar{0}} \sim$ Exponential(λ_0)

FOR $m = 1$ to K

$logit\ p_L \leftarrow$ logit $Pr(L_m = 1|L_{m-1}, A_{m-1}, Y_m = 0; \beta)$

$\leftarrow \beta_0 + \beta_1 I(T_{\bar{0}} < c) + \beta_2 A_{m-1} + \beta_3 L_{m-1}$

$L_m \sim$ Bernoulli(p_L)

logit $p_A \leftarrow$ logit $Pr(A_m = 1 | L_m, L_{m-1}, A_{m-1}, Y_m = 0; \alpha)$

$\leftarrow \alpha_0 + \alpha_1 L_m + \alpha_2 A_{m-1} + \alpha_3 L_{m-1} + \alpha_4 A_{m-1} \times L_m$

$A_m \sim$ Bernoulli(p_A)

$H_m \leftarrow \int_0^{m+1} \lambda_{\bar{a}_j}(j) dj$

$\leftarrow H_m + \exp(\psi_1 A_m)$

IF $T_{\bar{0}} \geq H_m$

$Y_{m+1} \leftarrow 0$

ELSE

$Y_{m+1} \leftarrow 1$

$T_{\bar{A}_m} \leftarrow m + (T_{\bar{0}} - H_m) \times \exp(-\psi A_m)$

END IF

ENDFOR m

ENDFOR ID

PRINT

ID, m, Y_{m+1}, A_m, L_m, A_{m-1}, L_{m-1}

B.7 Describing the Characteristics of the Weights in a Simulated Population

The truncated weights estimated from various approaches are summarized in Appendix Tables B.1 - B.4. Data is generated for a very large number of subjects ($n = 25,000$), each with up to 10 visits, under the rare event condition. These tables are described in § 3.4.1.

Table B.1: Summaries of the truncated weights estimated by logistic regression (l = logistic) under different weighting schemes (w = unstabilized, $w^{(n)}$ = unstabilized normalized, sw = stabilized, $sw^{(n)}$ = stabilized normalized) from the simulation study with a large $(25,000)$ number of subjects, each with up to 10 visits, under the rare event condition.

	Min.	Q1	Median	Mean	Q3	Max.	sd	p >20	p >100
$l-w$	1.21	3.96	17.82	189.70	98.09	12780.00	666.15	0.48	0.25
$l-w1$	1.21	3.96	17.82	165.30	98.09	2728.00	425.81	0.48	0.25
$l-w5$	1.31	3.96	17.82	117.00	98.09	807.30	214.13	0.48	0.25
$l-w10$	1.47	3.96	17.82	89.15	98.09	414.70	137.36	0.48	0.25
$l-w25$	3.96	3.96	17.82	38.89	98.05	98.09	39.10	0.48	0.00
$l-w35$	6.93	6.93	17.82	24.30	44.89	44.89	17.04	0.48	0.00
$l-w50$	17.82	17.82	17.82	17.82	17.82	17.82	0.00	0.00	0.00
$l-w^{(n)}$	0.01	0.22	0.58	1.00	1.22	13.99	1.37	0.00	0.00
$l-w^{(n)}1$	0.02	0.22	0.58	0.97	1.22	6.42	1.21	0.00	0.00
$l-w^{(n)}5$	0.06	0.22	0.58	0.89	1.22	3.34	0.91	0.00	0.00
$l-w^{(n)}10$	0.10	0.22	0.58	0.82	1.22	2.38	0.75	0.00	0.00
$l-w^{(n)}25$	0.22	0.22	0.58	0.65	1.22	1.22	0.40	0.00	0.00
$l-w^{(n)}35$	0.35	0.35	0.58	0.56	0.79	0.79	0.20	0.00	0.00
$l-w^{(n)}50$	0.58	0.58	0.58	0.58	0.58	0.58	0.00	0.00	0.00
$l-sw$	0.33	0.79	0.94	1.00	1.19	2.54	0.34	0.00	0.00
$l-sw1$	0.39	0.79	0.94	1.00	1.19	2.08	0.33	0.00	0.00
$l-sw5$	0.55	0.79	0.94	1.00	1.19	1.64	0.29	0.00	0.00
$l-sw10$	0.65	0.79	0.94	0.98	1.19	1.41	0.25	0.00	0.00
$l-sw25$	0.79	0.79	0.94	0.97	1.19	1.19	0.16	0.00	0.00
$l-sw35$	0.86	0.86	0.94	0.95	1.05	1.05	0.08	0.00	0.00
$l-sw50$	0.94	0.94	0.94	0.94	0.94	0.94	0.00	0.00	0.00
$l-sw^{(n)}$	0.32	0.78	0.94	1.00	1.18	2.48	0.33	0.00	0.00
$l-sw^{(n)}1$	0.39	0.78	0.94	1.00	1.18	2.08	0.33	0.00	0.00
$l-sw^{(n)}5$	0.55	0.78	0.94	0.99	1.18	1.64	0.29	0.00	0.00
$l-sw^{(n)}10$	0.65	0.78	0.94	0.98	1.18	1.42	0.25	0.00	0.00
$l-sw^{(n)}25$	0.78	0.78	0.94	0.97	1.18	1.18	0.16	0.00	0.00
$l-sw^{(n)}35$	0.85	0.85	0.94	0.95	1.04	1.04	0.09	0.00	0.00
$l-sw^{(n)}50$	0.94	0.94	0.94	0.94	0.94	0.94	0.00	0.00	0.00

Table B.2: Summaries of the truncated weights estimated by bagging approach (b = bagging) under different weighting schemes (w = unstabilized, $w^{(n)}$ = unstabilized normalized, sw = stabilized, $sw^{(n)}$ = stabilized normalized) from the simulation study with a large $(25,000)$ number of subjects, each with up to 10 visits, under the rare event condition.

	Min.	Q1	Median	Mean	Q3	Max.	sd	p >20	p >100
$b-w$	1.28	4.08	18.62	195.80	101.50	8990.00	641.09	0.49	0.25
$b-w1$	1.29	4.08	18.62	174.00	101.50	2986.00	450.31	0.49	0.25
$b-w5$	1.30	4.08	18.62	121.70	101.50	812.10	219.95	0.49	0.25
$b-w10$	1.67	4.08	18.62	97.09	101.50	466.20	152.30	0.49	0.25
$b-w25$	4.08	4.08	18.62	40.68	101.30	101.50	40.54	0.49	0.25
$b-w35$	7.70	7.70	18.62	25.84	47.23	47.23	17.84	0.49	0.00
$b-w50$	18.62	18.62	18.62	18.62	18.62	18.62	0.00	0.00	0.00
$b-w^{(n)}$	0.01	0.26	0.66	1.00	1.26	12.56	1.31	0.00	0.00
$b-w^{(n)}1$	0.02	0.26	0.66	0.98	1.26	6.87	1.18	0.00	0.00
$b-w^{(n)}5$	0.07	0.26	0.66	0.90	1.26	3.25	0.87	0.00	0.00
$b-w^{(n)}10$	0.12	0.26	0.66	0.83	1.26	2.29	0.72	0.00	0.00
$b-w^{(n)}25$	0.26	0.26	0.66	0.69	1.26	1.26	0.40	0.00	0.00
$b-w^{(n)}35$	0.40	0.40	0.66	0.61	0.84	0.84	0.20	0.00	0.00
$b-w^{(n)}50$	0.66	0.66	0.66	0.66	0.66	0.66	0.00	0.00	0.00
$b-sw$	0.36	0.92	0.98	1.00	1.06	1.99	0.19	0.00	0.00
$b-sw1$	0.49	0.92	0.98	1.00	1.06	1.63	0.18	0.00	0.00
$b-sw5$	0.76	0.92	0.98	1.00	1.06	1.38	0.15	0.00	0.00
$b-sw10$	0.83	0.92	0.98	1.00	1.06	1.24	0.12	0.00	0.00
$b-sw25$	0.92	0.92	0.98	0.99	1.06	1.06	0.06	0.00	0.00
$b-sw35$	0.95	0.95	0.98	0.98	1.02	1.02	0.03	0.00	0.00
$b-sw50$	0.98	0.98	0.98	0.98	0.98	0.98	0.00	0.00	0.00
$b-sw^{(n)}$	0.35	0.92	0.98	1.00	1.06	1.95	0.19	0.00	0.00
$b-sw^{(n)}1$	0.48	0.92	0.98	1.00	1.06	1.62	0.18	0.00	0.00
$b-sw^{(n)}5$	0.75	0.92	0.98	1.00	1.06	1.38	0.15	0.00	0.00
$b-sw^{(n)}10$	0.82	0.92	0.98	0.99	1.06	1.24	0.12	0.00	0.00
$b-sw^{(n)}25$	0.92	0.92	0.98	0.98	1.06	1.06	0.06	0.00	0.00
$b-sw^{(n)}35$	0.94	0.94	0.98	0.98	1.01	1.01	0.03	0.00	0.00
$b-sw^{(n)}50$	0.98	0.98	0.98	0.98	0.98	0.98	0.00	0.00	0.00

Table B.3: Summaries of the truncated weights estimated by SVM approach (svm = SVM) under different weighting schemes (w = unstabilized, $w^{(n)}$ = unstabilized normalized, sw = stabilized, $sw^{(n)}$ = stabilized normalized) from the simulation study with a large $(25,000)$ number of subjects, each with up to 10 visits, under the rare event condition.

	Min.	Q1	Median	Mean	Q3	Max.	sd	p >20	p >100
$svm - w$	1.35	4.50	20.25	161.40	100.70	6568.00	466.04	0.50	0.25
$svm - w1$	1.35	4.50	20.25	149.50	100.70	2438.00	366.19	0.50	0.25
$svm - w5$	1.35	4.50	20.25	111.30	100.70	731.10	196.67	0.50	0.25
$svm - w10$	1.82	4.50	20.25	88.21	100.70	408.90	133.50	0.50	0.25
$svm - w25$	4.50	4.50	20.25	40.93	100.60	100.70	40.18	0.50	0.25
$svm - w35$	8.22	8.22	20.25	27.22	50.19	50.19	18.95	0.50	0.00
$svm - w50$	20.25	20.25	20.25	20.25	20.25	20.25	0.00	1.00	0.00
$svm - w^{(n)}$	0.03	0.34	0.72	1.00	1.31	8.33	1.10	0.00	0.00
$svm - w^{(n)}1$	0.04	0.34	0.72	0.99	1.31	5.59	1.03	0.00	0.00
$svm - w^{(n)}5$	0.11	0.34	0.72	0.93	1.31	3.13	0.81	0.00	0.00
$svm - w^{(n)}10$	0.17	0.34	0.72	0.86	1.31	2.12	0.64	0.00	0.00
$svm - w^{(n)}25$	0.34	0.34	0.72	0.76	1.31	1.31	0.40	0.00	0.00
$svm - w^{(n)}35$	0.48	0.48	0.72	0.70	0.95	0.95	0.21	0.00	0.00
$svm - w^{(n)}50$	0.72	0.72	0.72	0.72	0.72	0.72	0.00	0.00	0.00
$svm - sw$	0.76	0.95	1.01	1.00	1.06	1.22	0.08	0.00	0.00
$svm - sw1$	0.81	0.95	1.01	1.00	1.06	1.19	0.08	0.00	0.00
$svm - sw5$	0.87	0.95	1.01	1.00	1.06	1.14	0.07	0.00	0.00
$svm - sw10$	0.90	0.95	1.01	1.00	1.06	1.10	0.06	0.00	0.00
$svm - sw25$	0.95	0.95	1.01	1.01	1.06	1.06	0.04	0.00	0.00
$svm - sw35$	0.98	0.98	1.01	1.01	1.04	1.04	0.03	0.00	0.00
$svm - sw50$	1.01	1.01	1.01	1.01	1.01	1.01	0.00	0.00	0.00
$svm - sw^{(n)}$	0.76	0.95	1.01	1.00	1.05	1.22	0.08	0.00	0.00
$svm - sw^{(n)}1$	0.80	0.95	1.01	1.00	1.05	1.19	0.08	0.00	0.00
$svm - sw^{(n)}5$	0.86	0.95	1.01	1.00	1.05	1.14	0.07	0.00	0.00
$svm - sw^{(n)}10$	0.89	0.95	1.01	1.00	1.05	1.09	0.06	0.00	0.00
$svm - sw^{(n)}25$	0.95	0.95	1.01	1.00	1.05	1.05	0.04	0.00	0.00
$svm - sw^{(n)}35$	0.97	0.97	1.01	1.00	1.03	1.03	0.03	0.00	0.00
$svm - sw^{(n)}50$	1.01	1.01	1.01	1.01	1.01	1.01	0.00	0.00	0.00

Table B.4: Summaries of the truncated weights estimated by boosting approach (gbm = boosting) under different weighting schemes (w = unstabilized, $w^{(n)}$ = unstabilized normalized, sw = stabilized, $sw^{(n)}$ = stabilized normalized) from the simulation study with a large $(25,000)$ number of subjects, each with up to 10 visits, under the rare event condition.

	Min.	Q1	Median	Mean	Q3	Max.	sd	p >20	p >100
$gbm - w$	1.24	3.56	16.09	163.60	90.74	6441.00	477.65	0.46	0.23
$gbm - w1$	1.24	3.56	16.09	151.20	90.74	2415.00	376.62	0.46	0.23
$gbm - w5$	1.30	3.56	16.09	117.20	90.74	878.10	226.41	0.46	0.23
$gbm - w10$	1.54	3.56	16.09	84.72	90.74	408.30	133.56	0.46	0.23
$gbm - w25$	3.56	3.56	16.09	35.96	90.54	90.74	36.33	0.46	0.00
$gbm - w35$	6.55	6.55	16.09	22.38	41.31	41.31	15.65	0.46	0.00
$gbm - w50$	16.09	16.09	16.09	16.09	16.09	16.09	0.00	0.00	0.00
$gbm - w^{(n)}$	0.01	0.23	0.60	1.00	1.22	8.86	1.29	0.00	0.00
$gbm - w^{(n)}1$	0.02	0.23	0.60	0.99	1.22	6.75	1.23	0.00	0.00
$gbm - w^{(n)}5$	0.06	0.23	0.60	0.90	1.22	3.40	0.94	0.00	0.00
$gbm - w^{(n)}10$	0.09	0.23	0.60	0.84	1.22	2.41	0.77	0.00	0.00
$gbm - w^{(n)}25$	0.23	0.23	0.60	0.66	1.22	1.22	0.40	0.00	0.00
$gbm - w^{(n)}35$	0.35	0.35	0.60	0.57	0.81	0.81	0.21	0.00	0.00
$gbm - w^{(n)}50$	0.60	0.60	0.60	0.60	0.60	0.60	0.00	0.00	0.00
$gbm - sw$	0.21	0.77	0.93	0.99	1.10	3.41	0.42	0.00	0.00
$gbm - sw1$	0.30	0.77	0.93	0.99	1.10	2.55	0.40	0.00	0.00
$gbm - sw5$	0.45	0.77	0.93	0.98	1.10	1.82	0.34	0.00	0.00
$gbm - sw10$	0.56	0.77	0.93	0.96	1.10	1.44	0.26	0.00	0.00
$gbm - sw25$	0.77	0.77	0.93	0.93	1.10	1.10	0.13	0.00	0.00
$gbm - sw35$	0.85	0.85	0.93	0.94	1.03	1.03	0.08	0.00	0.00
$gbm - sw50$	0.93	0.93	0.93	0.93	0.93	0.93	0.00	0.00	0.00
$gbm - sw^{(n)}$	0.21	0.77	0.94	1.00	1.11	3.45	0.42	0.00	0.00
$gbm - sw^{(n)}1$	0.30	0.77	0.94	1.00	1.11	2.59	0.40	0.00	0.00
$gbm - sw^{(n)}5$	0.45	0.77	0.94	0.98	1.11	1.83	0.34	0.00	0.00
$gbm - sw^{(n)}10$	0.57	0.77	0.94	0.96	1.11	1.45	0.26	0.00	0.00
$gbm - sw^{(n)}25$	0.77	0.77	0.94	0.94	1.11	1.11	0.13	0.00	0.00
$gbm - sw^{(n)}35$	0.86	0.86	0.94	0.94	1.03	1.03	0.08	0.00	0.00
$gbm - sw^{(n)}50$	0.94	0.94	0.94	0.94	0.94	0.94	0.00	0.00	0.00

B.8 Additional Simulation Results

B.8.1 Results from Smaller Samples $n = 300$

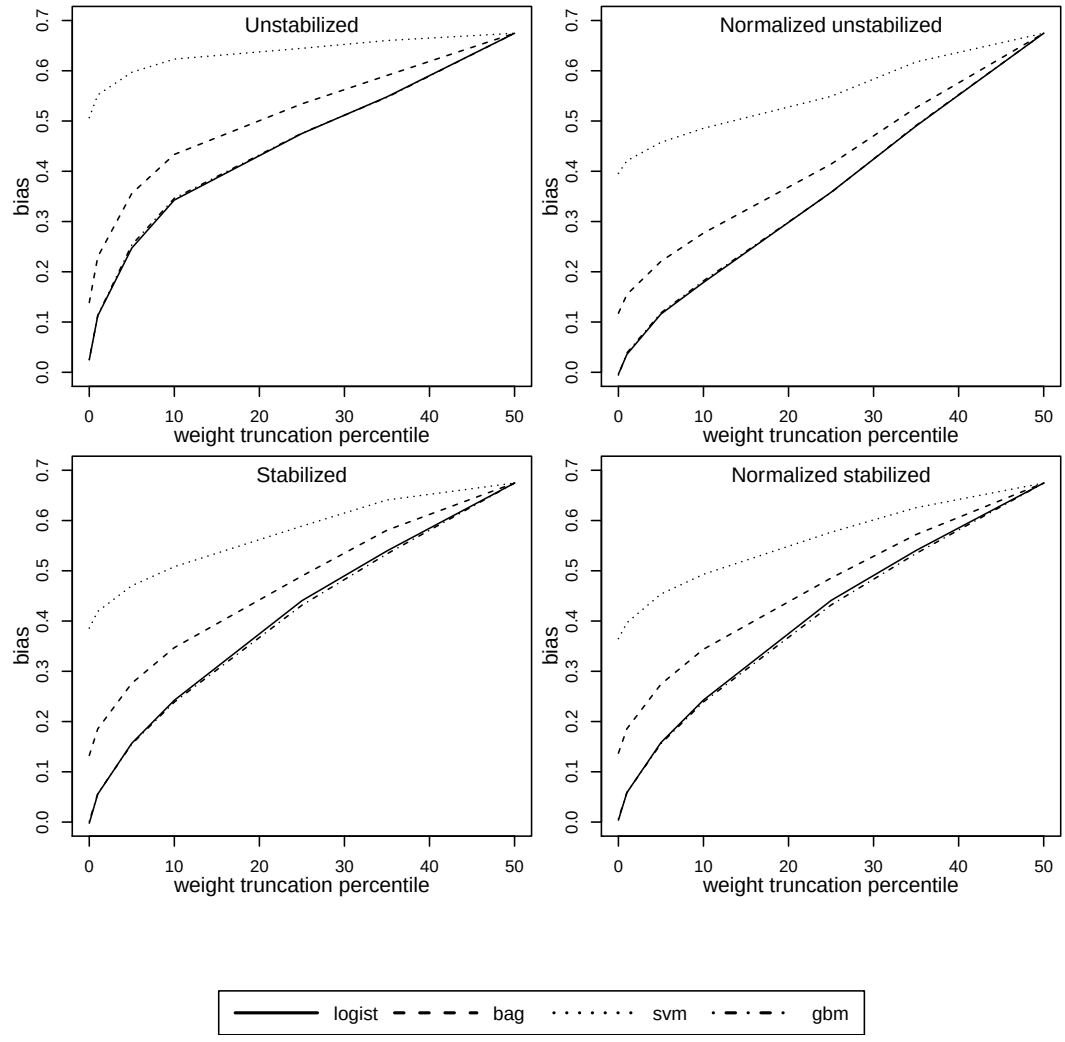


Figure B.1: Bias of MSCM estimate $\hat{\psi}_1$ under different IPW generation approaches when the large weights are progressively truncated in a simulation study of $1,000$ datasets with 300 subjects observed at most 10 times under the rare event condition.

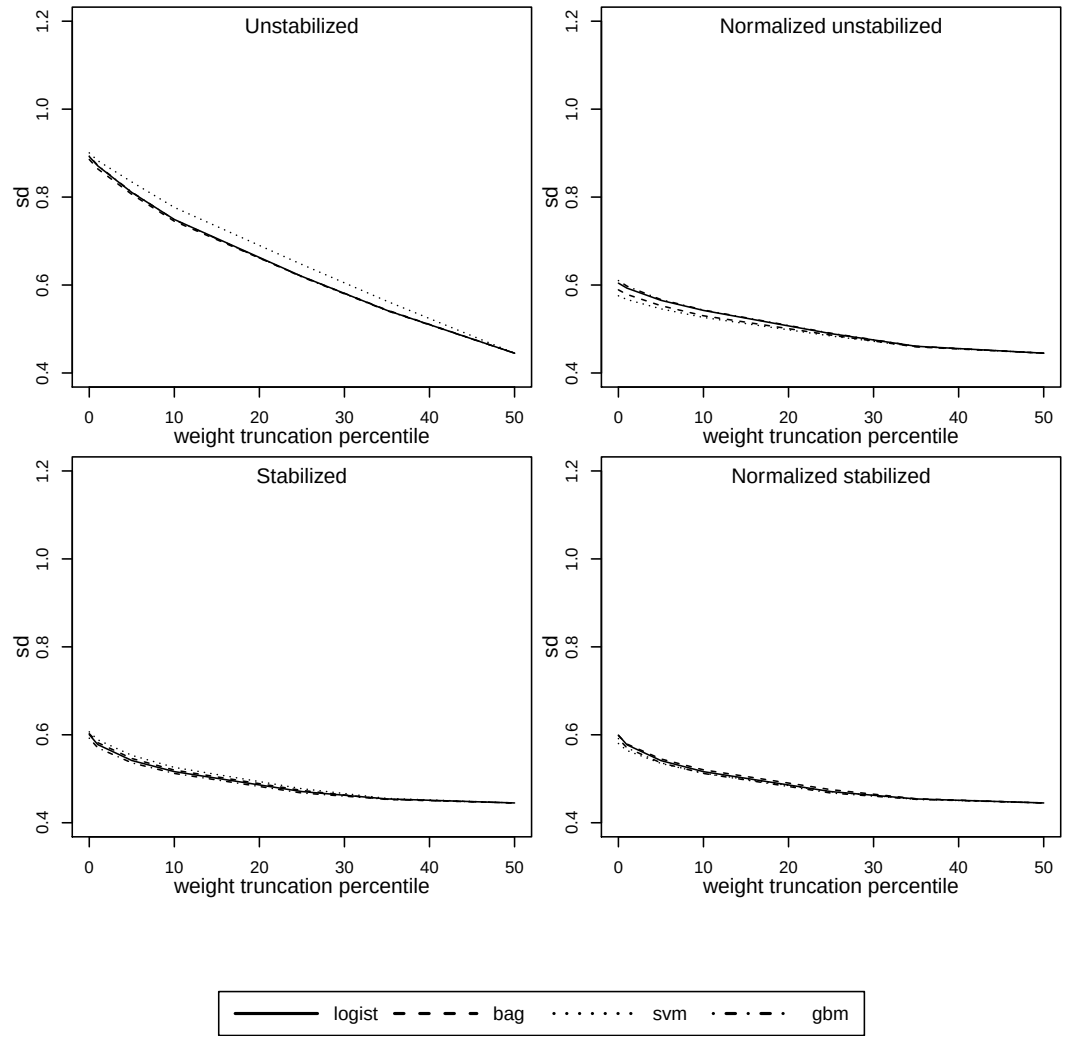


Figure B.2: Empirical standard deviation of MSCM estimate $\hat{\psi}_1$ under different IPW generation approaches when the large weights are progressively truncated in a simulation study of $1,000$ datasets with 300 subjects observed at most 10 times under the rare event condition.

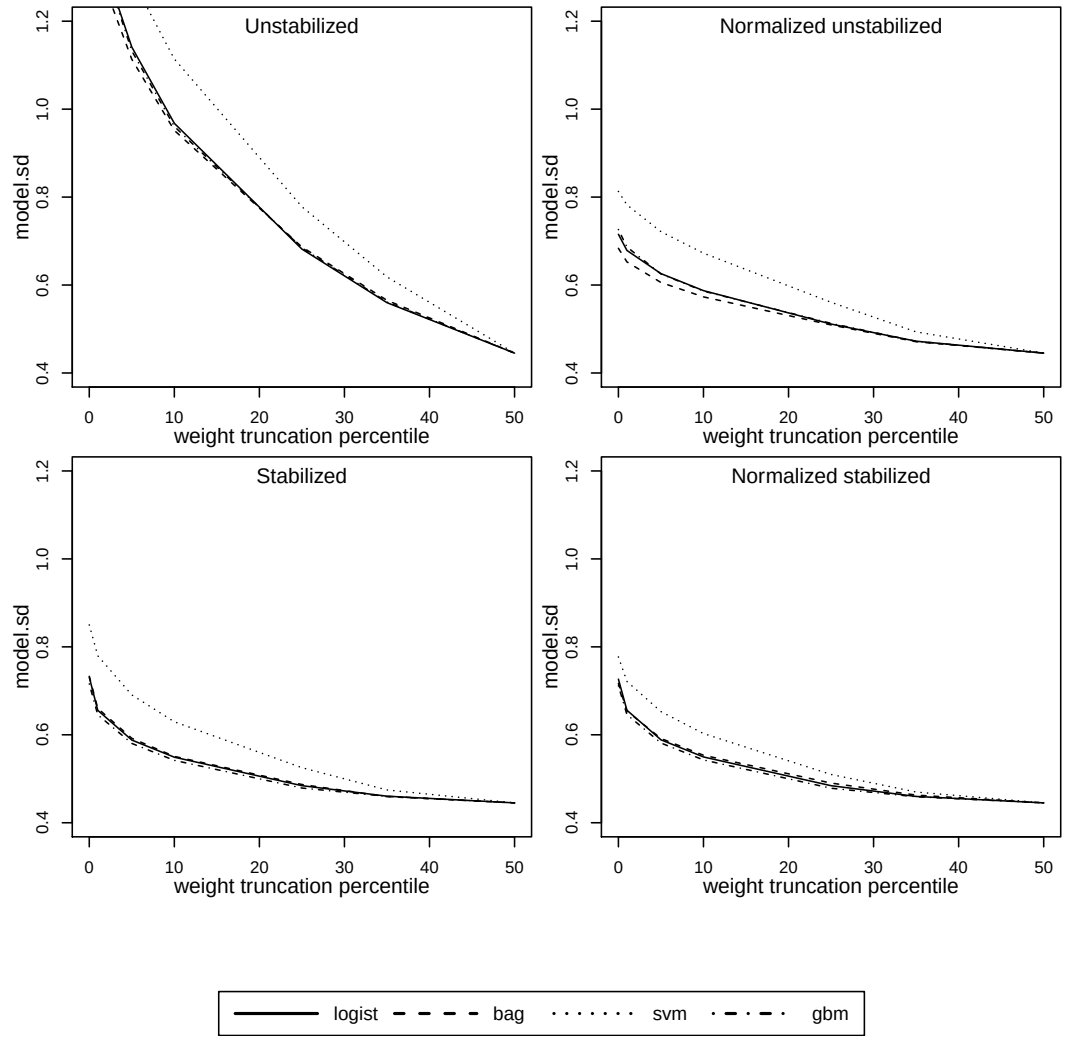


Figure B.3: Average model-based standard error of MSCM estimate $\hat{\psi}_1$ under different IPW generation approaches when the large weights are progressively truncated in a simulation study of $1,000$ datasets with 300 subjects observed at most 10 times under the rare event condition.

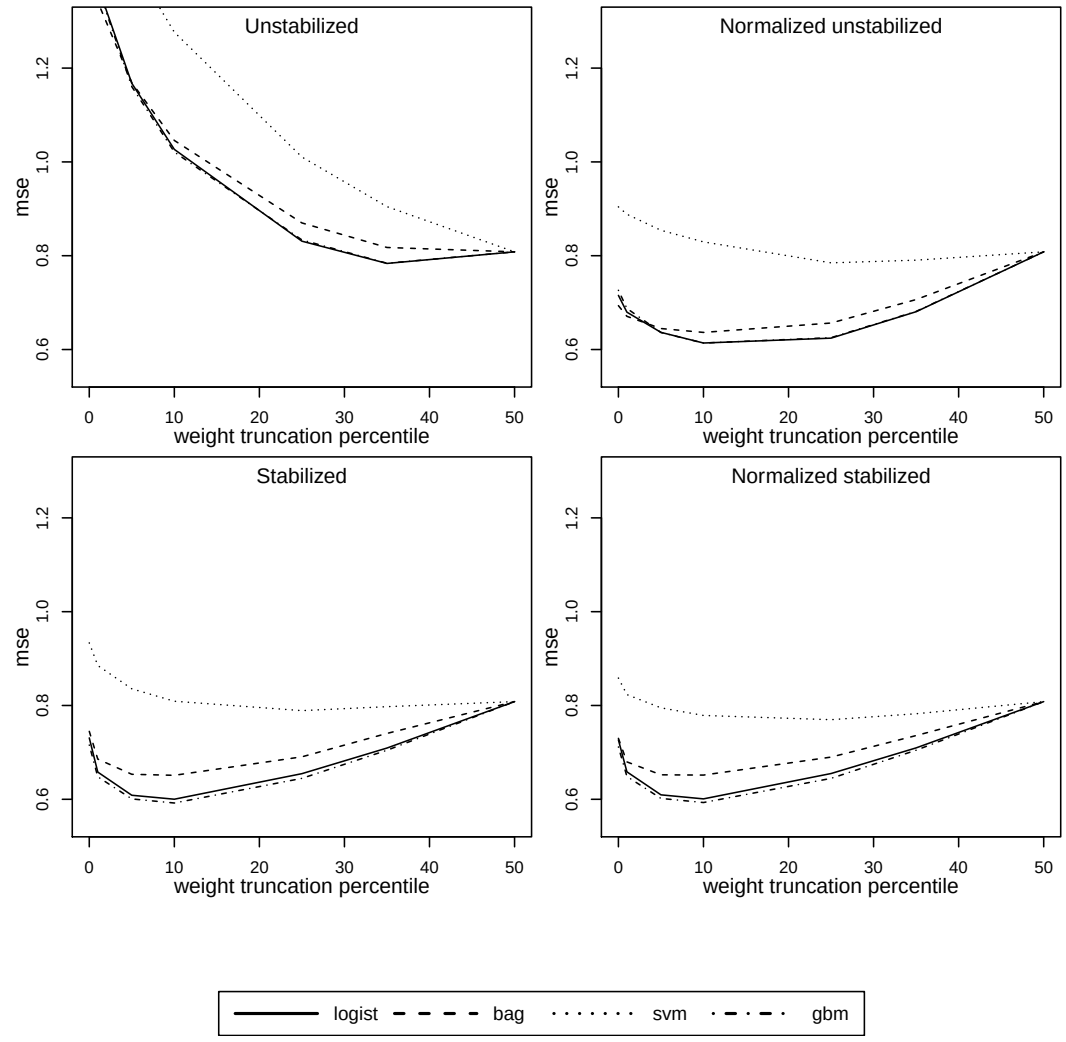


Figure B.4: Mean squared error of MSCM estimate $\hat{\psi}_1$ under different IPW generation approaches when the large weights are progressively truncated in a simulation study of $1,000$ datasets with 300 subjects observed at most 10 times under the rare event condition.

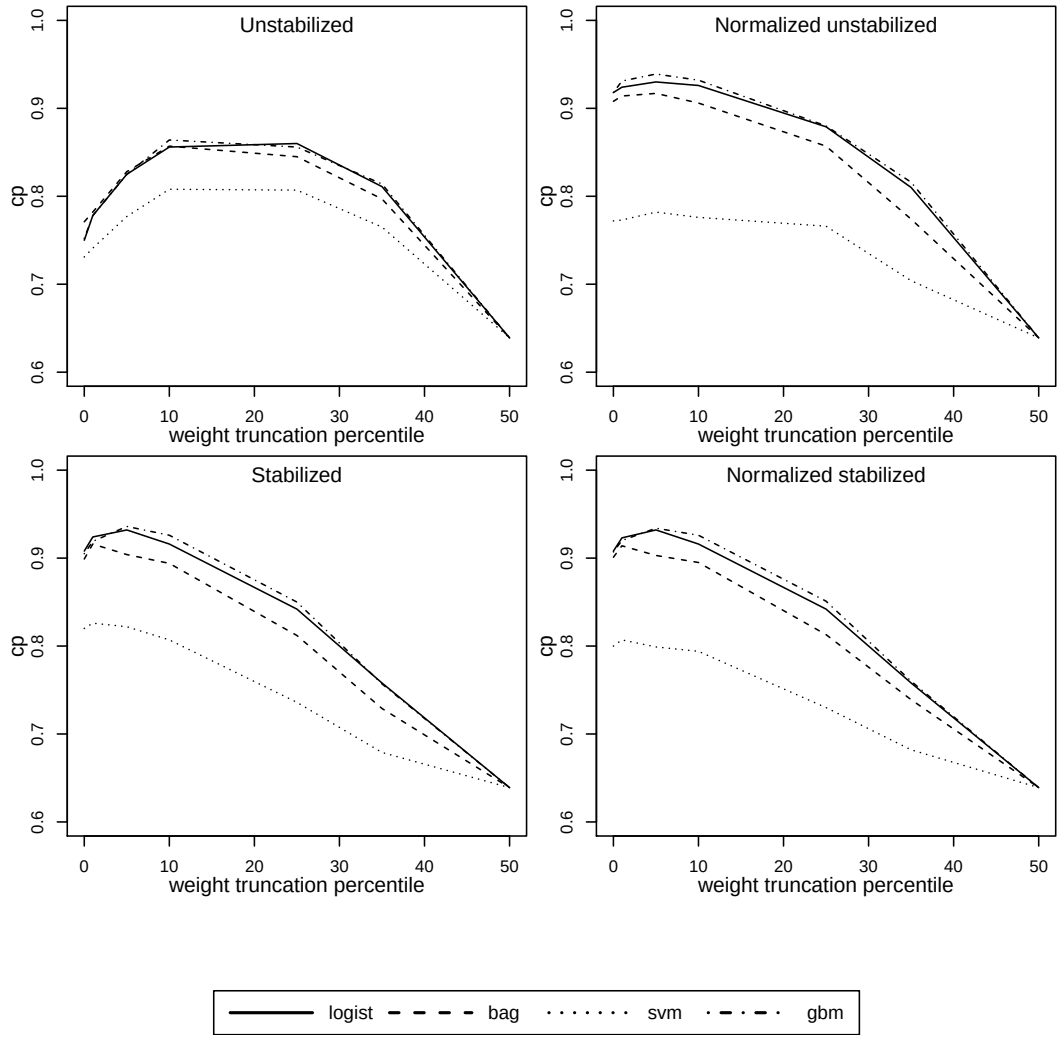


Figure B.5: The coverage probability (cp) of model-based nominal 95% confidence intervals based on the MSCM estimate $\hat{\psi}_1$ under different IPW generation approaches when the large weights are progressively truncated in a simulation study of $1,000$ datasets with 300 subjects observed at most 10 times under the rare event condition.

B.8.2 Results from the Scenario When More Events are Available for $n = 2,500$

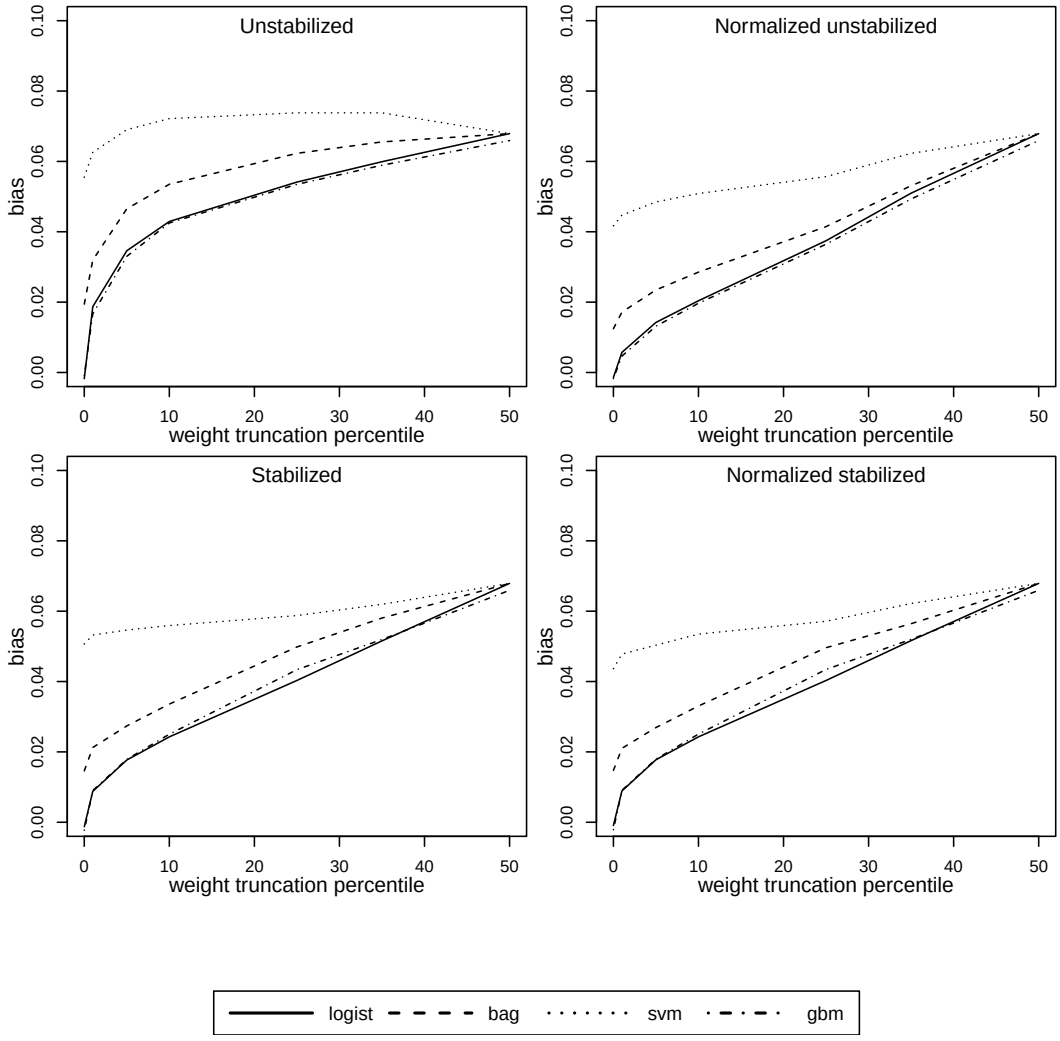


Figure B.6: Bias of MSCM estimate $\hat{\psi}_1$ under different IPW generation approaches when the large weights are progressively truncated in a simulation study of $1,000$ datasets with $2,500$ subjects observed at most 10 times when the event rate is more frequent.

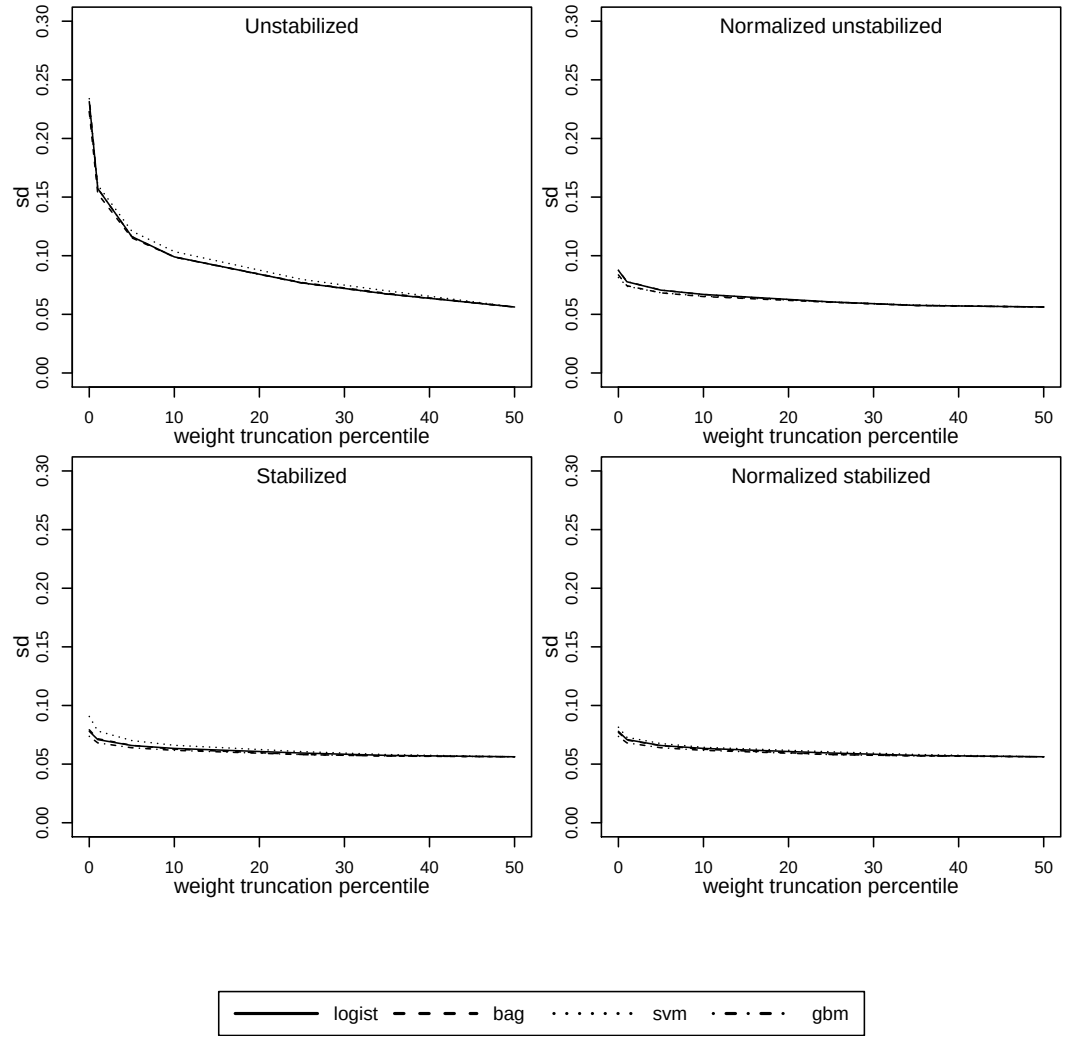


Figure B.7: Empirical standard deviation of MSCM estimate $\hat{\psi}_1$ under different IPW generation approaches when the large weights are progressively truncated in a simulation study of $1,000$ datasets with $2,500$ subjects observed at most 10 times when the event rate is more frequent.

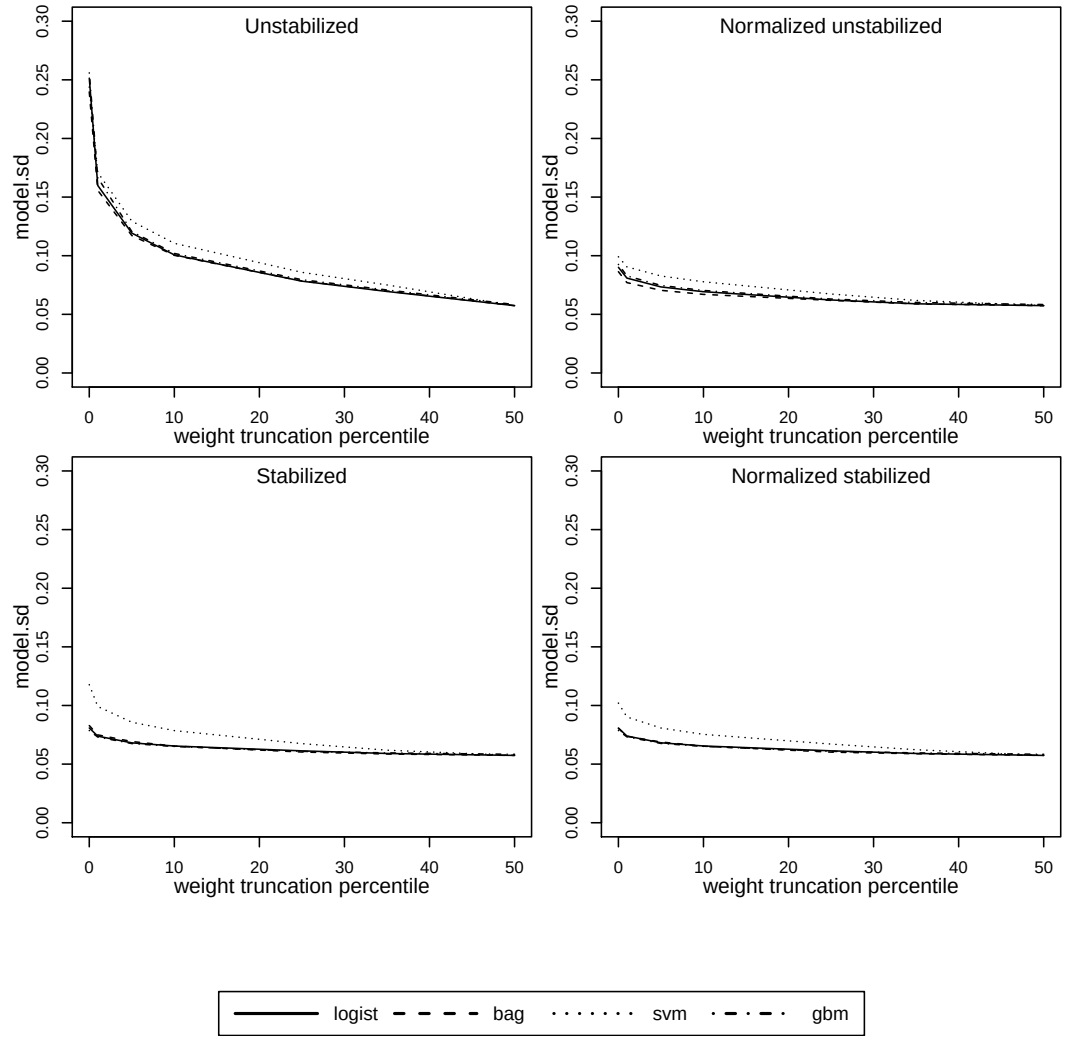


Figure B.8: Average model-based standard error of MSCM estimate $\hat{\psi}_1$ under different IPW generation approaches when the large weights are progressively truncated in a simulation study of $1,000$ datasets with $2,500$ subjects observed at most 10 times when the event rate is more frequent.

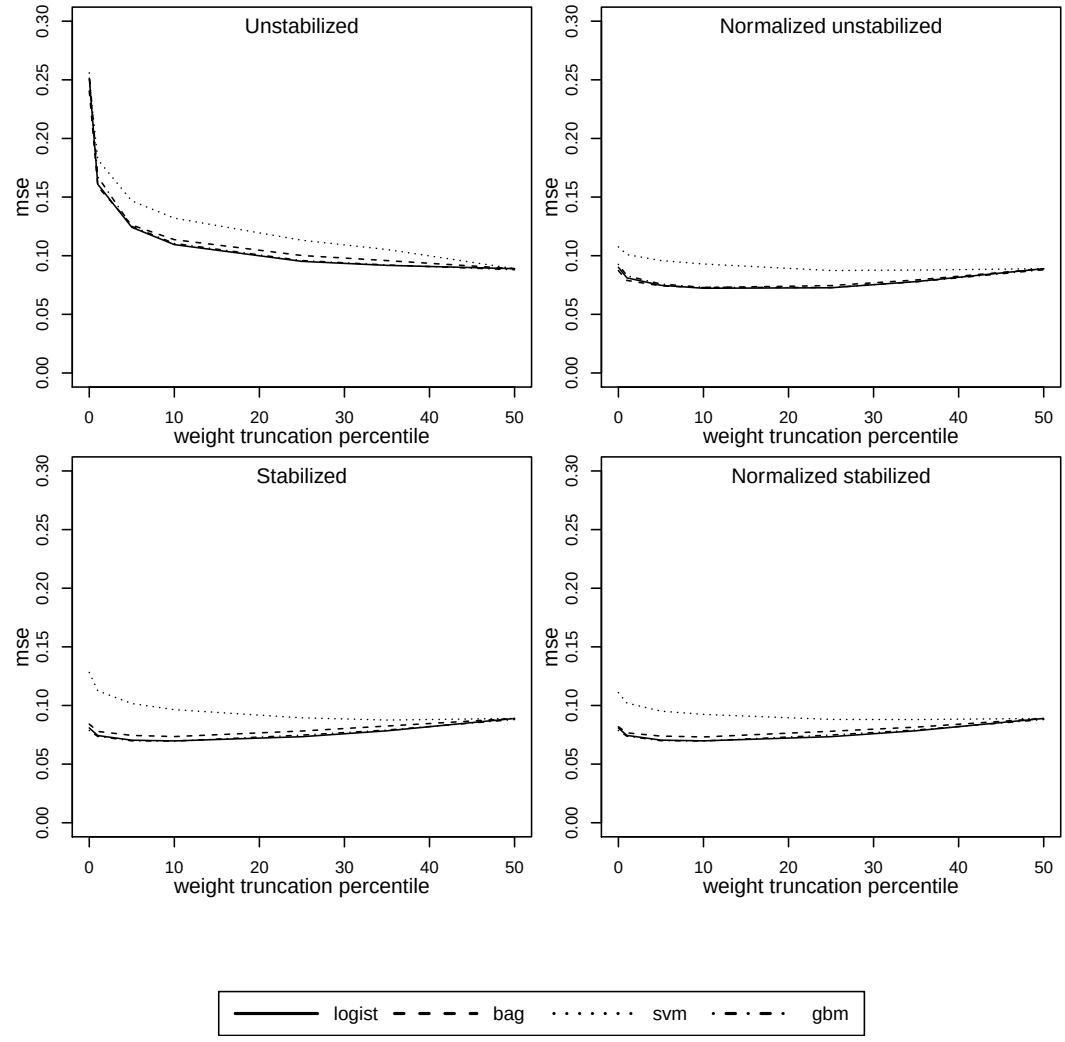


Figure B.9: Mean squared error of MSCM estimate $\hat{\psi}_1$ under different IPW generation approaches when the large weights are progressively truncated in a simulation study of $1,000$ datasets with $2,500$ subjects observed at most 10 times when the event rate is more frequent.

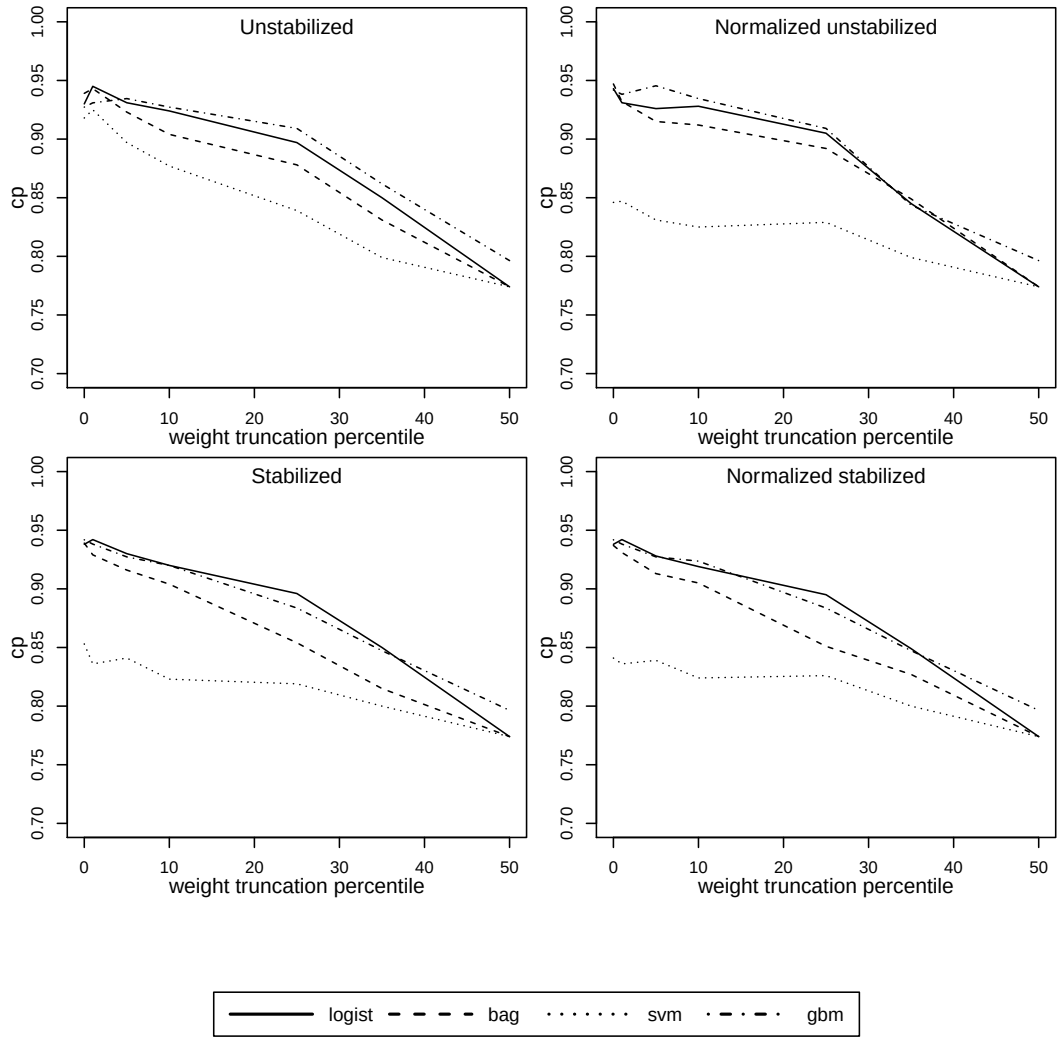


Figure B.10: The coverage probability (cp) of model-based nominal 95% confidence intervals based on the MSCM estimate under different IPW generation approaches when the large weights are progressively truncated in a simulation study of $1,000$ datasets with $2,500$ subjects observed at most 10 times when the event rate is more frequent.

B.9 Supporting Results from the Empirical MS Application

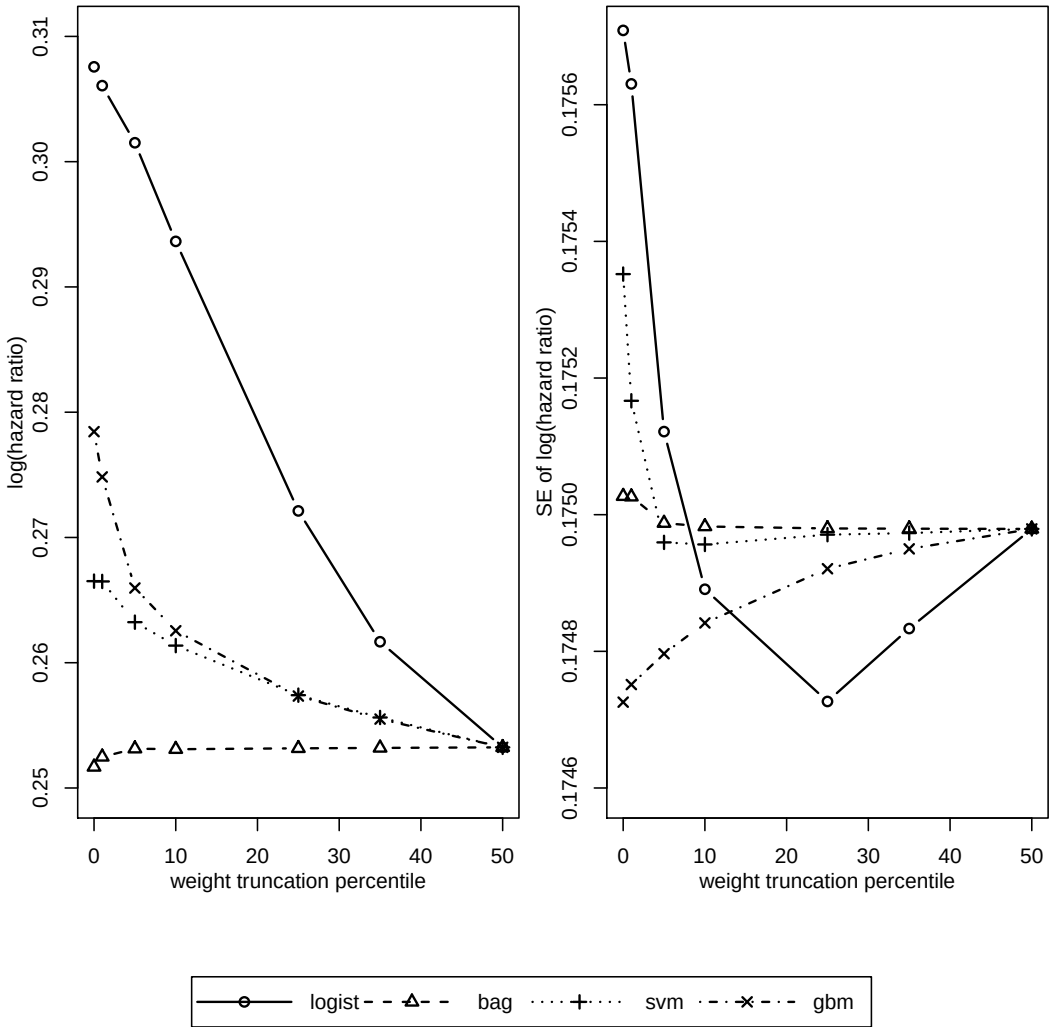


Figure B.11: Performance of stabilized normalized weights generated by different statistical learning approaches for MSCM analysis to estimate log-hazard ψ_1 in a multiple sclerosis study.

Table B.5: The impact of truncation of the $sw^{(n)}$ generated via logistic regression on the estimated causal effect of β-IFN on the hazard of reaching sustained EDSS 6 for BC MS patients (1995-2008).

Truncation percentiles	Estimated weights		Treatment effect estimate		
	Mean (log-SD)	Min-Max	HR	SE†	95% CI†
None	1.000 (-2.179)	0.317 - 1.713	1.360	0.239	0.964 - 1.919
(1, 99)	1.000 (-2.246)	0.638 - 1.275	1.358	0.239	0.963 - 1.916
(5, 95)	1.001 (-2.411)	0.812 - 1.164	1.352	0.237	0.959 - 1.905
(10, 90)	1.003 (-2.572)	0.875 - 1.125	1.341	0.235	0.952 - 1.890
(25, 75)	1.005 (-3.120)	0.950 - 1.062	1.313	0.229	0.932 - 1.849
(35, 65)	1.003 (-3.768)	0.978 - 1.030	1.299	0.227	0.922 - 1.830
Median	1.001 (-Inf)	1.001 - 1.001	1.288	0.225	0.914 - 1.815

log-SD, logarithmic transformation of standard deviation; Min, minimum; Max, maximum; CI, confidence interval; HR, Hazard ratio; SE, standard error.

† Based on robust standard error.

Table B.6: The impact of truncation of the $sw^{(n)}$ generated via bagging on the estimated causal effect of β-IFN on the hazard of reaching sustained EDSS 6 for BC MS patients (1995-2008).

Truncation percentiles	Estimated weights		Treatment effect estimate		
	Mean (log-SD)	Min-Max	HR	SE†	95% CI†
None	1.000 (-4.882)	0.967 - 1.122	1.286	0.225	0.913 - 1.813
(1, 99)	1.000 (-5.784)	0.988 - 1.020	1.287	0.225	0.913 - 1.814
(5, 95)	1.000 (-6.821)	0.997 - 1.002	1.288	0.225	0.914 - 1.815
(10, 90)	1.000 (-7.221)	0.998 - 1.001	1.288	0.225	0.914 - 1.815
(25, 75)	1.000 (-7.848)	0.999 - 1.000	1.288	0.225	0.914 - 1.815
(35, 65)	1.000 (-8.378)	0.999 - 1.000	1.288	0.225	0.914 - 1.815
Median	1.000 (-Inf)	1.000 - 1.000	1.288	0.225	0.914 - 1.815

log-SD, logarithmic transformation of standard deviation; Min, minimum; Max, maximum; CI, confidence interval; HR, Hazard ratio; SE, standard error.

† Based on robust standard error.

Table B.7: The impact of truncation of the $sw^{(n)}$ generated via SVM on the estimated causal effect of β-IFN on the hazard of reaching sustained EDSS 6 for BC MS patients (1995-2008).

Truncation percentiles	Estimated weights		Treatment effect estimate		
	Mean (log-SD)	Min-Max	HR	SE†	95% CI†
None	1.000 (-3.036)	0.420 - 1.755	1.305	0.229	0.926 - 1.841
(1, 99)	1.000 (-3.400)	0.866 - 1.102	1.305	0.229	0.926 - 1.840
(5, 95)	1.000 (-3.667)	0.937 - 1.039	1.301	0.228	0.923 - 1.833
(10, 90)	1.001 (-3.879)	0.963 - 1.030	1.299	0.227	0.922 - 1.830
(25, 75)	1.003 (-4.437)	0.988 - 1.017	1.294	0.226	0.918 - 1.823
(35, 65)	1.004 (-4.961)	0.996 - 1.012	1.291	0.226	0.916 - 1.820
Median	1.005 (-Inf)	1.005 - 1.005	1.288	0.225	0.914 - 1.815

log-SD, logarithmic transformation of standard deviation; Min, minimum; Max, maximum; CI, confidence interval; HR, Hazard ratio; SE, standard error.

† Based on robust standard error.

Table B.8: The impact of truncation of the $sw^{(n)}$ generated via boosting on the estimated causal effect of β-IFN on the hazard of reaching sustained EDSS 6 for BC MS patients (1995-2008).

Truncation percentiles	Estimated weights		Treatment effect estimate		
	Mean (log-SD)	Min-Max	HR	SE†	95% CI†
None	1.000 (-2.834)	0.348 - 1.749	1.321	0.231	0.938 - 1.861
(1, 99)	1.002 (-3.269)	0.790 - 1.108	1.316	0.230	0.935 - 1.854
(5, 95)	1.004 (-3.754)	0.945 - 1.051	1.305	0.228	0.926 - 1.838
(10, 90)	1.004 (-4.095)	0.973 - 1.032	1.300	0.227	0.923 - 1.832
(25, 75)	1.005 (-4.971)	0.997 - 1.014	1.293	0.226	0.918 - 1.822
(35, 65)	1.005 (-5.599)	1.001 - 1.010	1.291	0.226	0.916 - 1.819
Median	1.004 (-Inf)	1.004 - 1.004	1.288	0.225	0.914 - 1.815

log-SD, logarithmic transformation of standard deviation; Min, minimum; Max, maximum; CI, confidence interval; HR, Hazard ratio; SE, standard error.

† Based on robust standard error.

Appendix C

Appendix for Chapter 4

C.1 Bias Due to Incorrect Handling of Immortal Time

For simplicity, we often improperly define the treatment exposure. For example, we assume the subjects are on treatment immediately after joining a study cohort, when in reality, there may be a delay period to initiate treatment for some of the subjects. Not properly accounting for the delay period causes immortal time bias.

Let us define the notation to investigate the bias associated with immortal time. Suppose $i = 1$ indicates the ever-treatment exposed group, whereas $i = 0$ indicates the never-treatment exposed group. Further, let N_i and T_i $(i = 0, 1)$ indicate the observed number of failures and follow-up person-time in these groups.

Let $r = T_0/T_1$, the ratio of the observed person-times in the never-treatment exposed and ever-treatment exposed subjects. Denote T_{IT} as the observed immortal time, i.e., the aggregated follow-up time not under treatment in the ever-treatment exposed group, and set $f = T_{IT}/T_1$. Let N_{IT} be the number of failures observed during the immortal time. Obviously, $N_{IT} = 0$. Also, let $T_1' = T_1 - T_{IT} = (1-f) \times T_1$ denote the person-time under treatment in the ever-treatment exposed group. The total person-time not under treatment is $T_0' = T_0 + T_{IT} = r \times T_1 + f \times T_1 = T_1(r + f)$, where T_0 and T_{IT} are contributed by the never-treatment exposed and ever-treatment exposed subjects respectively. Under the assumption of constant hazard of

failure, the failure rate is calculated by the number of failures divided by the corresponding follow-up person-time. Thus, the failure rate under treatment is N_1/T_1', the failure rate not under treatment is N_0/T_0', and the failure rate ratio obtained from a time-dependent analysis is:

$$\begin{aligned} RR &= \frac{N_1/T_1'}{N_0/T_0'} \\ &= \frac{N_1/(T_1 - T_{IT})}{N_0/(T_0 + T_{IT})}. \end{aligned} \tag{C.1}$$

C.1.1 Misclassifying Immortal Time

Misclassifying the observed immortal time T_{IT} as treated time leads to the failure rate of N_1/T_1 for the ever-treatment exposed subjects, and the failure rate of N_0/T_0 for the never-treatment exposed subjects, and the failure rate ratio,

$$RR' = \frac{N_1/T_1}{N_0/T_0}.$$

Comparing RR' with the correct rate ratio RR yields [80]:

$$\begin{aligned} \frac{RR'}{RR} &= \frac{\frac{N_1/T_1}{N_0/T_0}}{\frac{N_1/(T_1-T_{IT})}{N_0/(T_0+T_{IT})}} \\ &= \frac{T_0}{T_1} \times \frac{T_1 - T_{IT}}{T_0 + T_{IT}} \\ &= r \times \frac{T_1 - f \times T_1}{T_0 + f \times T_1} \\ &= r \times \frac{T_1(1-f)}{r \times T_1 + f \times T_1} \\ &= r \times \frac{(1-f)}{(r+f)} \\ &= (1-f) \times \frac{r}{(r+f)}. \end{aligned} \tag{C.2}$$

Under the assumption of constant hazard, this approach, therefore, al-

ways underestimates the correct failure rate rato. As a lower rate ratio is indicative of less hazard or risk, this approach always overestimates (inflates) the treatment effect. Varying the r and f parameters in equation (C.2) yields the Appendix Figure C.1 (upper panel 1). We can see a larger downward bias (in RR'/RR) for increasing values of f, the fraction of the immortal person-time in the ever-treatment exposed subjects. For different ratios $r = T_0/T_1$ ($r = 0.25, 0.5, 1, 2, 4, 8$), the pattern of RR'/RR looks similar. The higher values of r yield slightly less bias (in RR'/RR).

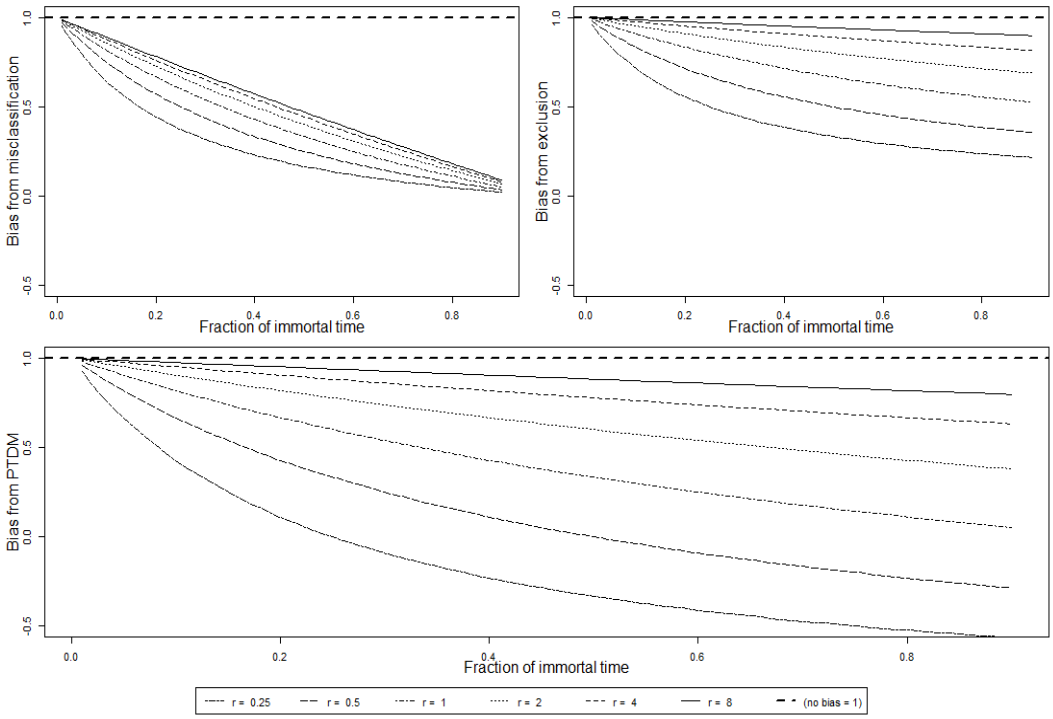


Figure C.1: Risk ratios of misclassified immortal time (RR'), excluding immortal time (RR'') and PTDM (RR'') methods compared to that of a time-dependent analysis RR in terms of various fraction of immortal time f and ratio of person-times under no treatment versus under treatment r under the assumption of constant hazard.

C.1.2 Excluding Immortal Time

Exclusion of the immortal time yields the failure rate under treatment of N_1/T_1' as N_{IT}, the number of failures during the immortal time T_{IT}, is zero. The immortal time T_{IT} is not included in the calculation of the failure rate for the untreated group, leading to failure rate N_0/T_0, and the failure rate ratio

$$RR'' = \frac{N_1/T_1'}{N_0/T_0}.$$

Comparing RR'' to the correct rate ratio RR yields [80]:

$$\begin{aligned} \frac{RR''}{RR} &= \frac{(N_1/T_1')/(N_0/T_0)}{(N_1/T_1')/(N_0/T_0')} \\ &= \frac{T_0}{T_0'} \\ &= \frac{T_0}{T_0 + T_{IT}} \\ &= \frac{r \times T_1}{r \times T_1 + f \times T_1} \\ &= \frac{r}{r+f} \end{aligned} \tag{C.3}$$

As in the previous situation, this approach, therefore, always underestimates the correct failure rate ratio, overestimating the effect of treatment. Varying the r and f parameters in equation (C.3) yields the Appendix Figure C.1 (upper panel 2). This also shows a downward bias (in RR''/RR) for increasing values of f, the fraction of the immortal person-time in the ever-treatment exposed subjects. However, the bias (in RR''/RR) is significantly reduced for the higher values of r, the ratio of the person-times in the never-treatment exposed and ever-treatment exposed subjects. If the never-treatment exposed cohort is much larger than the ever-treatment exposed cohort, the bias from this approach may be negligible, even for large fractions of immortal time f. Therefore, use of this approach may be reasonable in some settings [252].

C.2 Illustration of the Prescription Time-distribution Matching Approach

The PTDM approach can be illustrated as follows. Let us consider the time of eligibility to receive treatment as the baseline. The length of time from the first eligibility date $t_0 = 0$ to the treatment initiation T^A for the treated subjects is the wait-period or immortal time T_{IT}.

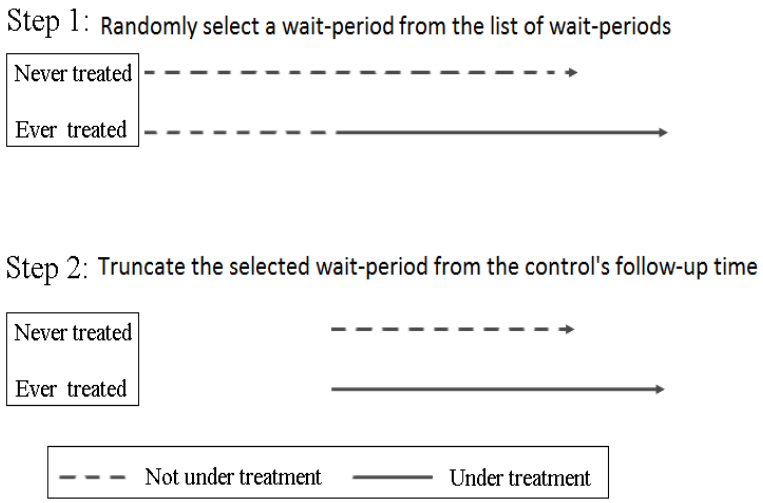


Figure C.2: An illustration of prescription time-distribution matching

To apply this method, first, the wait-periods $T_{j,IT}$ for each of the treated subjects j are listed. To achieve balance in both treatment groups, the distribution of these wait-periods $T_{j,IT}$ for the treated subjects needs to be matched with a similar part of the follow-up time for the untreated subjects. To achieve this, for each untreated subject j', a wait-period $T_{j,IT}$ is selected at random from the created list of wait-periods for the treated subjects and is assigned to the untreated subject j'. If this wait-period $T_{j,IT}$ is longer than the event time $T_{j'}$ or the censoring time $T_{j'}^C$ of this untreated subject j', the untreated subject j' gets excluded from further analysis. For

simplicity, let us first assume that both groups have the same number of subjects i.e., $N_0 = N_1$. Then this process should match the time-distribution of $T_{j,IT}$ for the treated subjects to the assigned wait-period distribution of $T_{j',IT}$ of the untreated subjects (as shown in Figure C.2). The wait-periods for the treated subjects and the matched contributions for the untreated subjects are deleted together.

The immortal time $T_{IT} \equiv \sum_{j=1}^{N_1} T_{j,IT}$ and we denote $T'_{IT} \equiv \sum_{j'=1}^{N_0} T_{j',IT}$. After excluding the observed and assigned wait-times from both groups, the unexposed time under consideration is $T''_0 = T_0 - T'_{IT}$ and the exposed time under consideration is $T'_1 = T_1 - T_{IT}$. As $T_{j,IT}$ and $T_{j',IT}$ follow the same distribution, assuming $N_1 = N_0$ we have $T_{IT} \approx T'_{IT}$, and the balance should be restored due to the elimination of the similar wait-periods from both groups. Subjects in both groups are now followed from their new baselines until reaching outcome or censoring. Under the assumption of constant hazard, the failure rate ratio is calculated as

$$RR''' = \frac{N_1/T'_1}{N_0/T''_0}.$$

Let us first derive the formula of the rate ratio RR''' for the PTDM approach under two further simplifying assumptions: $T'_{IT} = T_{IT}$ and the number of subjects discarded from the never-treatment exposed group, $N'_{IT} = 0$. We will derive the formula for more general settings later. Comparing RR''' with the correct rate ratio RR in equation (C.1) yields:

$$\begin{aligned}
\frac{RR'''}{RR} &= \frac{(N_1/T_1')/(N_0/T_0'')}{(N_1/T_1')/(N_0/T_0')} \\
&= \frac{T_0''}{T_0'} \\
&= \frac{T_0 - T_{IT}}{T_0 + T_{IT}} \\
&= \frac{r \times T_1 - f \times T_1}{r \times T_1 + f \times T_1} \\
&= \frac{r - f}{r + f}. \qquad \text{(C.4)}
\end{aligned}$$

We see that RR'''/RR can also be expressed as a function of r and f. Varying the r and f parameters yields Appendix Figure C.1 (lower panel). Unfortunately, this approach also shows a downward bias for increasing values of f, the fraction of the immortal person-time in the ever-treatment exposed subjects. As for the exclusion method, the bias (in RR'''/RR) is significantly reduced for high values of r, the ratio of the person-times in the never-treatment exposed and ever-treatment exposed subjects. However, small values of r and large values of f have much more detrimental effects on the bias compared to the misclassification and exclusion approaches.

To obtain equation (C.4), we assumed that the number of failures in assigned wait-period T_{IT}' for the untreated patients is zero; i.e., $N_{IT}' = 0$. In general, for $N_{IT}' \geq 0$, the total untreated person-time T_x of the N_{IT}' patients who had failures within the assigned wait-times is excluded and the formula becomes (set $x = T_x/T_1 \geq 0$):

$$\begin{aligned}
\frac{RR'''}{RR} &= \frac{\frac{N_1/(T_1-T_{IT})}{(N_0-N'_{IT})/(T_0-T_{IT}-T_x)}}{\frac{N_1/(T_1-T_{IT})}{N_0/(T_0+T_{IT})}} \\
&= \left(\frac{N_0}{N_0-N'_{IT}}\right)\frac{T_0-T_{IT}-T_x}{T_0+T_{IT}} \\
&= \left(\frac{N_0}{N_0-N'_{IT}}\right)\times\frac{r-f-x}{r+f} \\
&\geq \frac{r-f}{r+f}. \qquad \text{(C.5)}
\end{aligned}$$

Therefore, equation (C.4) is actually a lower bound for the bias in RR'''/RR. Also, if $N'_{IT}=0$, then $T_x=0$ as well.

Now, let us relax the assumption that $T'_{IT}=T_{IT}$. Set $T'_{IT}=q\times T_{IT}$; that is, q is the ratio of assigned and observed wait-periods. Here, $q>1$ for the setting where there are more subjects in the never-treatment exposed group than the ever-treatment exposed group, and otherwise $0<q<1$. Then, $T''_0=T_0-T'_{IT}=T_0-q\times T_{IT}$ and $T'_1=T_1-T_{IT}$, and the derivation leading to (C.5) is modified as follows:

$$\begin{aligned}
\frac{RR'''}{RR} &= \frac{\frac{N_1/(T_1-T_{IT})}{(N_0-N'_{IT})/(T_0-T'_{IT}-T_x)}}{\frac{N_1/(T_1-T_{IT})}{N_0/(T_0+T_{IT})}} \\
&= \left(\frac{N_0}{N_0-N'_{IT}}\right)\frac{T_0-T'_{IT}-T_x}{T_0+T_{IT}} \\
&= \left(\frac{N_0}{N_0-N'_{IT}}\right)\frac{T_0-q\times T_{IT}-x\times T_1}{T_0+T_{IT}} \\
&= \left(\frac{N_0}{N_0-N'_{IT}}\right)\frac{r-q\times f-x}{r+f} \qquad \text{(C.6)}
\end{aligned}$$

The equations (C.4) - (C.6) and Appendix Figure C.1 show the general pattern of bias and allow general statements about the approaches under

consideration. However, in our simulation studies, we took into account additional specific details of a more realistic epidemiological setting, such as censoring, different rates of failures, covariate under consideration, etc.

C.3 Constructing a Mini-trial in the Sequential Cox Approach

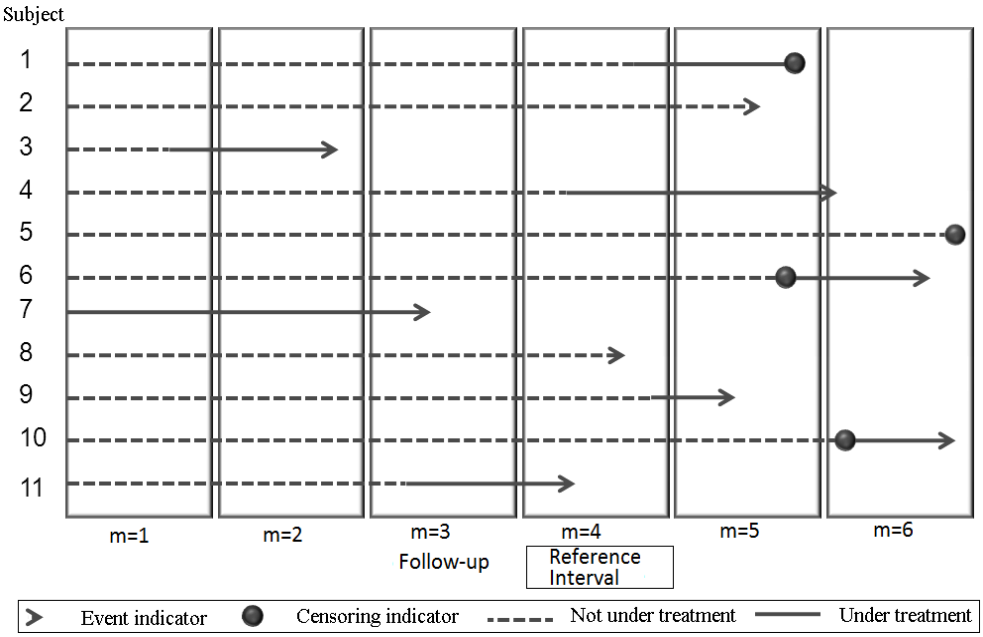


Figure C.3: An illustration of the sequential Cox approach

To illustrate the method, consider Appendix Figure C.3, where the follow–up times for 11 subjects are outlined. Patient 1 was not under treatment when she entered the study. She started taking the treatment in the $m = 4$th month and was censored during the 5th month. Similarly subject 5, who was never under treatment was censored during the 6th month. Now, suppose we want to create the mimicked trial considering the 4th month as the reference interval. We eliminate the subjects who received treatment before the 4th month, i.e., the 3rd, 7th and 11th subjects will be discarded.

Then for the subjects who started treatment after the 4th month, we censor them at the time of treatment start i.e., the 6th and 10th subjects are censored at the 5th and 6th months respectively. Then, under the assumption that treatment status remains the same for the entire month, subjects 1, 4 and 9 will be considered the treated group and subjects 2, 5, 6, 8 and 10 will be considered the control group, for the mimicked trial starting at the beginning of 4th month.

Similarly, we can identify the subjects for the treatment and control groups in the mimicked trials starting at the beginning of other months. This yields multiple mimicked RCTs, one for each of the time intervals (say, months) of treatment start. The treatment effect can be estimated separately from each mimicked trial data and then aggregated (i.e., averaged) to estimate the average treatment effect.

C.4 Implementation of the Sequential Cox Approach in R

The `coxph` function in the `survival` package [249] is used to fit both time-independent and time-dependent Cox PH models. While preparing the data for the mini-trials of the sequential Cox approach, we can code it in either long or wide form; both will produce the same result:

- In the long form, each row of the data can represent the smallest time interval to be used (such as month) and the multiple rows per subject specify the start and stop of all the intervals. Rows without any change in covariate values can be merged (one row starting at the baseline, one for starting at the m-th month and another for the lagged values) [73]. Then the counting process formulation of the `coxph` can be applied specifying the start and end time of the intervals and the corresponding event status.
- In the wide form, each subject in the m-th mini-trial will produce

only one row in the mini-trial data containing all the corresponding information at baseline, the m-th interval and the lagged data of m-th interval as separate covariates [75]. Then the standard `coxph` can be applied specifying the follow-up time and event status.

In the `coxph` function, the option `strata` is set to fit a stratified Cox model for the sequential Cox approach. Also, the options such as `cluster` and `robust = TRUE` are set to obtain the robust (sandwich) variance estimate. This is an approximate grouped jackknife variance estimate [253, p.170] when multiple observations per subject are present. To obtain bootstrap estimates [165], the `lapply` function is used on each bootstrap sample to estimate the corresponding IPCWs and subsequently the HR from a Cox PH. In this chapter we fitted pooled logistic regression using the `glm` function in the `stats` package to estimate the IPCWs. Alternatively, Aalen's additive regression can be fitted using the `aalen` function in the `timereg` package for the same purpose [75].

C.5 Survival Data Simulation via Permutation Algorithm

The algorithm has following steps:

1. For each subject $i = 1, 2, \ldots, n$, we generate the survival time T_i using a specified distribution.

2. For each subject i, we generate the censoring time T_i^C using a specified distribution.

3. We find the observed survival time $T_i^* = min(T_i, T_i^C)$ and the binary censoring indicator $C_i = I(T_i \geq T_i^C) = 1$ if censored and 0 otherwise.

4. Repeat steps 1-3 n times and sort survival status tuples (T_i^*, C_i) with respect to T_i^* in increasing order.

5. We generate n covariate matrices $X_i = (A_{im}, L_{i0}, L_{im})$ with dimensions $(m \times p)$, where the $m = 0, 1, \ldots, K$ rows correspond to the different time intervals or visits when measurements are taken and the p columns correspond to the predictor variables, including treatment (A_m), time-fixed and/or time-varying covariates (L_0 and/or L_m). For subject i, X_{im}, the m-th row of X_i, is a vector of variable values at time m.

6. According to the ordered T_i^* listed in step 3, we begin assigning the survival status tuple (T_i^*, C_i) to covariate values from X_{im} as follows. At time T_i^*, variable values (treatment and covariate) are sampled with probabilities p_{im} defined below based on the Cox model's partial likelihood:

$$p_{im} = \begin{cases} \frac{\exp(\psi X_{im})}{\sum_{j \in r_i} \exp(\psi X_{jm})}, & \text{if } C_i = 0 \\ \frac{1}{\sum_{j \in r_i} I(j \in r_i)}, & \text{if } C_i = 1, \end{cases}$$

   where ψ is the vector of log-hazards for the corresponding variables and $I(j \in r_i)$ indicates whether a subject is within a given riskset r_i for time T_i^*.

7. The subject i with the covariate values X_{im} is assigned the observed time T_i^*. The selected X_{im} is removed from further calculation.

The permutation algorithm is implemented in the `PermAlgo` package in R [254].

C.6 Additional Simulation Results

C.6.1 When More Events are Available

Table C.1: Comparison of the analytical approaches to adjust for immortal time bias from simulation-I (one baseline covariate and time-dependent treatment exposure) of $1,000$ datasets, each containing $2,500$ subjects followed for up to 10 time-intervals (frequent event case $\lambda_0 = 0.10$).

Approach	Bias	$SD(\hat{\psi}_1)$	$se(\hat{\psi}_1)$	CP	Power
Full cohort	0.000	0.061	0.060	0.951	1.000
Included IT	-2.149	0.062	0.059	0.000	1.000
Excluded IT	-1.220	0.055	0.051	0.000	1.000
PTDM	-1.284	0.073	0.070	0.000	1.000
Sequential Cox	-0.038	0.071	0.070	0.899	1.000
MSCM	-	-	-	-	-

PTDM, Prescription time distribution matching; IT, Immortal time; MSCM, Marginal structural Cox model.

Table C.2: Comparison of the analytical approaches to adjust for immortal time bias from simulation-II (one baseline covariate, one time-dependent covariate and time-dependent treatment exposure) of $1,000$ datasets, each containing $2,500$ subjects followed for up to 10 time-intervals (frequent event case).

Approach	Bias	$SD(\hat{\psi}_1)$	$se(\hat{\psi}_1)$	CP	Power
Full cohort	-0.002	0.059	0.060	0.960	1.000
Full cohort (Base)	-0.208	0.067	0.070	0.130	0.990
Included IT	-1.638	0.076	0.076	0.000	1.000
Excluded IT	-1.411	0.069	0.069	0.000	1.000
PTDM	-1.440	0.085	0.084	0.000	1.000
Sequential Cox	0.174	0.066	0.068	0.273	1.000
MSCM	-0.014	0.058	0.060	0.952	1.000

PTDM, Prescription time distribution matching; IT, Immortal time; MSCM, Marginal structural Cox model.

Table C.3: Comparison of the analytical approaches to adjust for immortal time bias from simulation-III (one time-dependent confounder and time-dependent treatment exposure) of 1,000 datasets, each containing 2,500 subjects followed for up to 10 time-intervals (frequent event case).

Approach	Bias	$SD(\hat{\psi}_1)$	$se(\hat{\psi}_1)$	CP	Power
Full cohort	0.044	0.067	0.065	0.888	1.000
Full cohort (Base)	0.007	0.068	0.066	0.942	1.000
Included IT	-2.095	0.090	0.084	0.000	1.000
Excluded IT	-1.629	0.071	0.068	0.000	1.000
PTDM	-1.575	0.090	0.074	0.000	1.000
Sequential Cox	0.202	0.099	0.099	0.464	1.000
Sequential Cox †	0.201	0.096	0.096	0.433	1.000
Sequential Cox §	0.181	0.096	0.096	0.522	1.000
MSCM	0.000	0.069	0.068	0.942	1.000

PTDM, Prescription time distribution matching; IT, Immortal time; MSCM, Marginal structural Cox model.

† Sequential Cox not adjusting for either time-dependent confounder or informative censoring.

§ Sequential Cox adjusting for both time-dependent confounder in the regression for estimating β and informative censoring via IPCW.

C.7 Additional MS Data Analysis

C.7.1 Prescription Time-distribution Matching

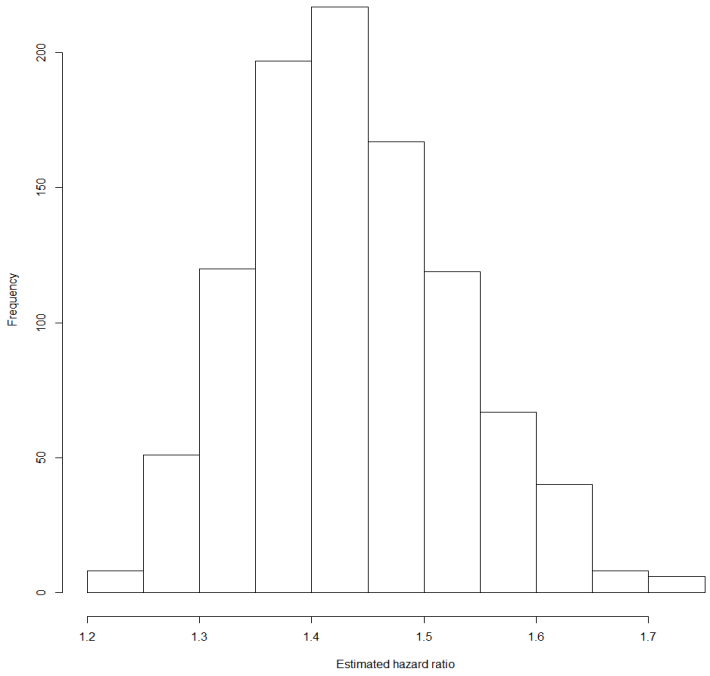


Figure C.4: Estimated hazard ratio from the PTDM method to estimate the causal effect of β-IFN on time to sustained EDSS 6 for patients with relapsing-onset multiple sclerosis (MS), British Columbia, Canada (1995-2008)

Table C.4: Mean (SD) of the estimated parameters using PTDM from the MS example with 1,000 different starting seed values.

HR	$se(\hat{HR})$	Average 95% CI
1.44 (0.09)	0.28 (0.02)	0.97 - 2.11

PTDM, Prescription time distribution matching.
The analyses are adjusted for baseline covariates: gender, EDSS score, age, disease duration and time-dependent confounder 'cumulative relapse'.

C.7.2 Sequential Cox Approach

Table C.5: Estimated hazard ratio using the sequential Cox approach to estimate the causal effect of β-IFN on time to sustained EDSS 6 for patients with relapsing-onset multiple sclerosis (MS), British Columbia, Canada (1995-2008), when IPCWs are calculated from the combined dataset of all mini-trials.

Approach	HR	$se(\hat{HR})$	95% CI	Weights	
				Average (log-SD)	range
Sequential Cox	1.11	0.29	0.66 - 1.85	1.00 (-4.15)	0.64 - 1.40

The HR for the treatment is reported. The analyses are adjusted for baseline covariates: sex, EDSS score, age, disease duration and time-dependent confounder 'cumulative relapse' measured at baseline, treatment initiation month and its lagged value.

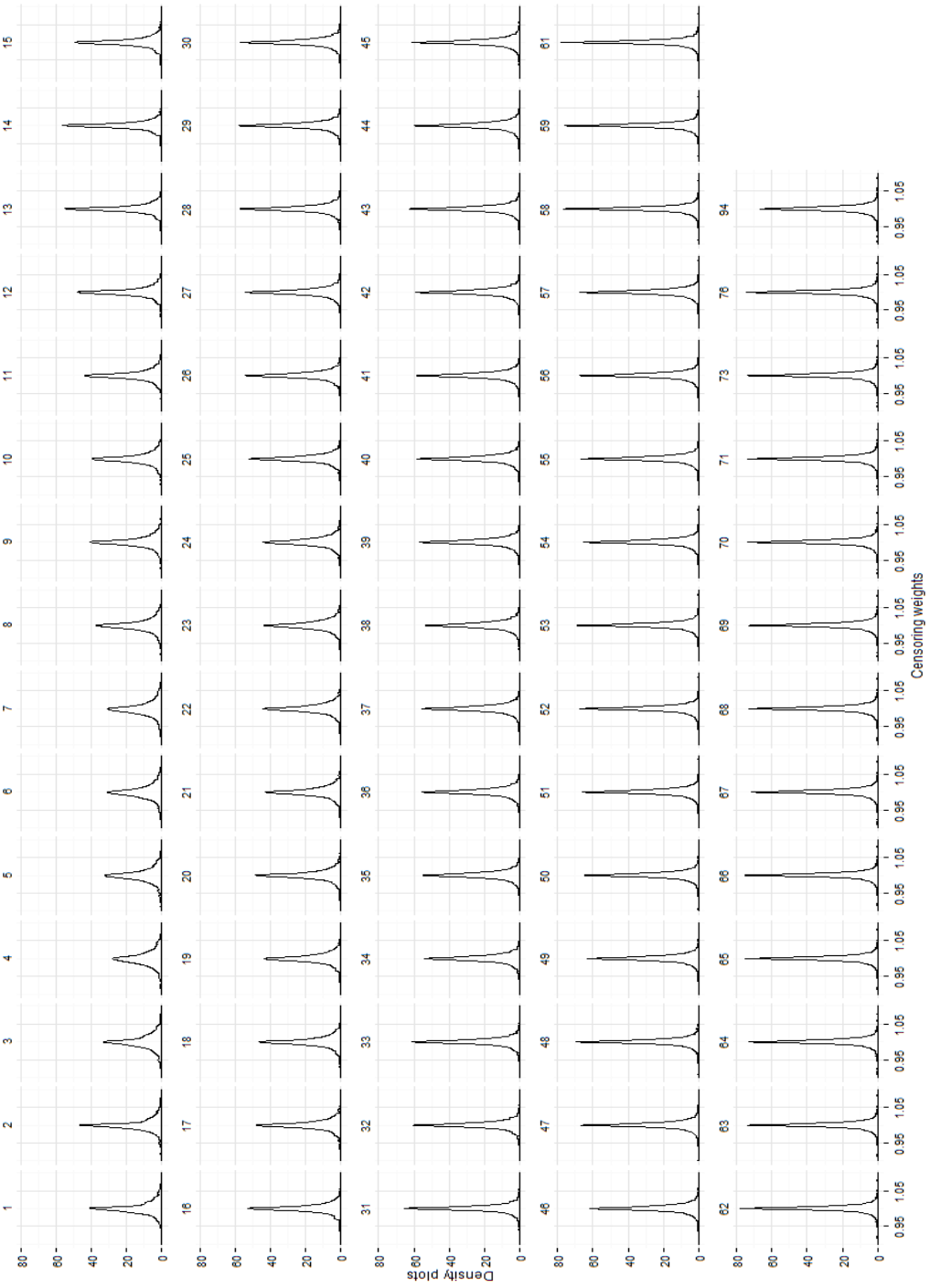


Figure C.5: Density plots of the estimated IPC weights from the MS data (estimated from each mini-trial separately) in all the reference intervals using the sequential Cox approach

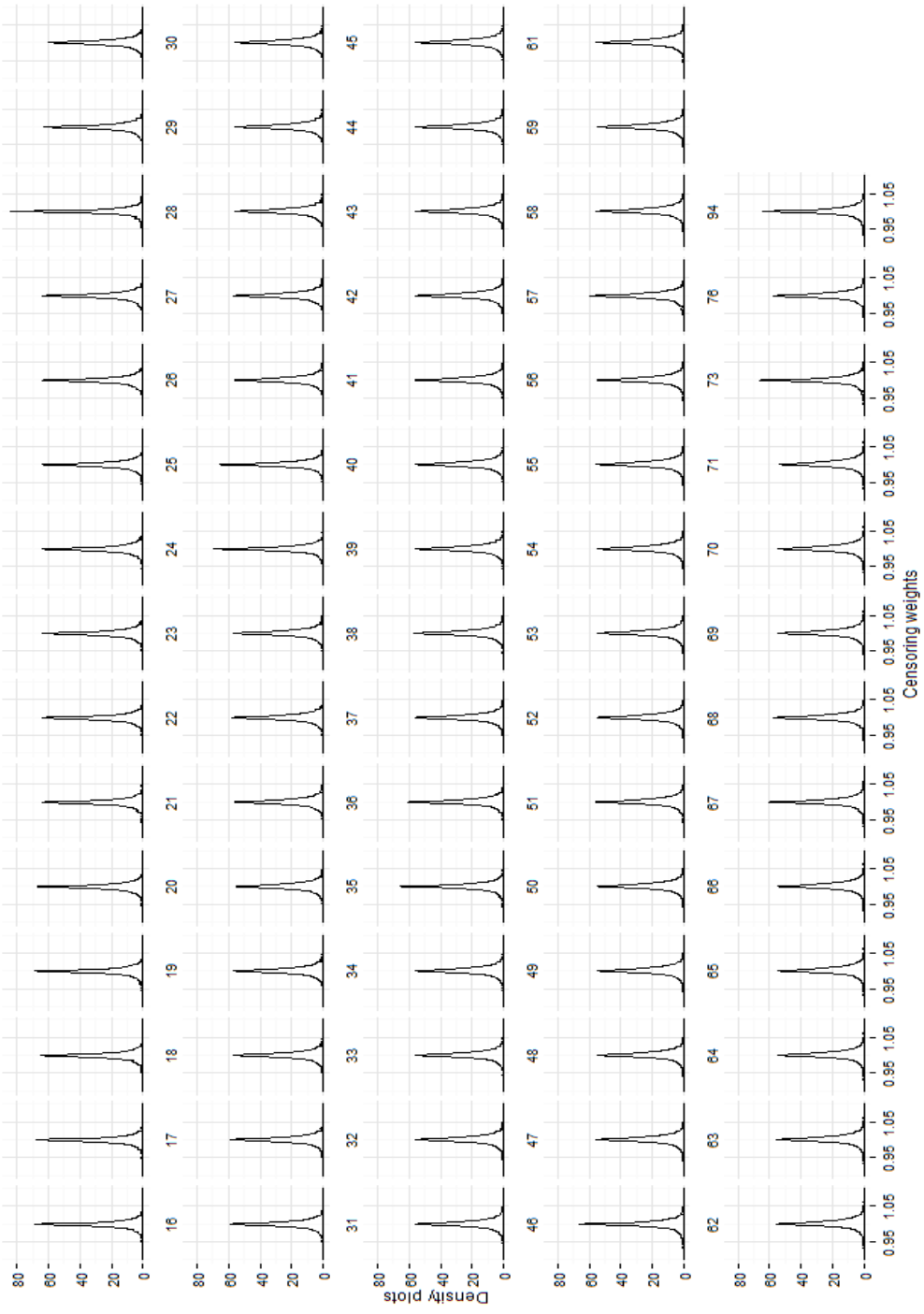


Figure C.6: Density plots of the estimated IPC weights from the MS data (estimated from the aggregated data of all mini-trials) in all the reference intervals using the sequential Cox approach