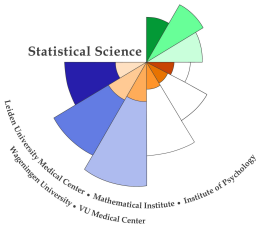
The Positivity Assumption and Marginal Structural Models in a Survival Context

Tomas D. Morley (s1722980)

Thesis advisor: Dr. Marta Fiocco

MASTER THESIS



STATISTICAL SCIENCE
FOR THE LIFE AND BEHAVIOURAL SCIENCES

Abstract

The Inverse Probability of Treatment Weighted (IPTW) estimators can be used to correctly estimate the parameters of marginal structural models (MSMs) for causal effects using observational data and a number of assumptions. In this thesis we focus on the positivity assumption which holds when there is a positive probability of receiving every level of an exposure variable for every combination of values defined by the observed confounders in the analysis. When the positivity assumption is violated, the resulting IPTW estimators may become very unstable and exhibit high variability. However, the severity with which this impacts the IPTW estimators under different conditions is not widely known or understood. In particular, to our knowledge, no existing study has investigated violations of the positivity assumption for survival analysis, or in a time dependent context more generally. This is surprising because MSMs are often applied in practice precisely because they adjust for time-dependent confounding. A novelty of this thesis is to investigate the effect of positivity violations on the performance of the IPTW-estimator in a survival context in which time dependent confounding is present.

We approach the problem in a simulation setting. One reason why the effects of positivity violations in a survival context have not been systematically studied is that existing algorithms for generating suitable data are intensive and challenging to implement. An added value of this thesis is to cast some light on this process in the hope that it will encourage other researchers to broach the subject in the future. We implement an existing algorithm in R and then extend that algorithm to incorporate violations of the positivity assumption that are propagated through time. A simulation study was carried out using the extended algorithm. We investigate how the IPTW estimators respond as strict violations of the positivity assumption become increasingly severe. As part of this study we examine the finite sample properties of the estimator and how it behaves for varied lengths of follow-up time. We also consider the case where the positivity assumption is not strictly violated but some exposure levels are rare within certain levels of the confounder. Our results indicate that even relatively benign violations of the positivity assumption can be a problem in the time-dependent context. We also find that, contrary to expectations, positivity violations are worse for studies of shorter duration. More optimistically, near violations of the positivity assumption do not appear to be serious under realistic circumstances.

Acknowledgements

I would like to extend my sincerest gratitude to Dr. M. Fiocco and Dr. C. Lancia for their support and advice while writing this thesis. In particular, I am grateful to Dr. M. Fiocco for helping me finish this thesis remotely from Switzerland and for her dedication and late night emails in these final weeks. Working alongside a demanding masters course has been far more difficult than I could have imagined. I would therefore like to thank my colleagues at Wageningen Economic Research for their understanding and patience. Special thanks go to Michiel van Dijk. I would also like to thank Dr. V. Didelez for sharing unpublished work which helped to illuminate some of the more difficult concepts encountered while writing this thesis. Finally, for their love and support I would like to thank Olivia, Ilythia, Mafalda and Milne.

Contents

Abstract **i**

Acknowledgements **ii**

1 Introduction **1**

2 Causal Inference **4**

2.1 Counterfactuals and causality 4

2.2 Confounding 5

2.3 Directed Acyclic Graphs 8

2.4 Time dependent confounding 9

2.5 Effect measures 10

2.6 Model formulation for causal effect measures 12

3 Marginal Structural Models **15**

3.1 Marginal structural models 15

3.2 Inverse Probability of Treatment Weighting 16

3.3 Assumptions 20

4 Survival Concepts **24**

4.1 Survival function 24

4.2 Hazard function 25

4.3 Proportional hazards 26

4.4 Marginal structural survival models. 27

5 Simulating from marginal structural models **29**

5.1 Monte Carlo simulations 29

5.2 Simulating from MSMs 31

5.3 Simulation algorithm 32

5.4 Overcoming non-collapsibility 33

6 Violations of positivity **43**

6.1 Positivity: definition and relevance to IPTW-estimation. 43

6.2 Types and examples of positivity violations 44

6.3 Related work on violations of the positivity assumption 46

6.4 Simulation Algorithm with Positivity Violations 47

6.5 Simulation scenarios 52

7 Simulation study **56**

7.1 Simulation set-up 56

7.2 Results 57

8 Discussion and conclusion **71**

8.1 Software 74

9 Appendix A: R code for simulation functions **76**

10 Appendix B: R code for replicating simulation study **84**

11 Appendix C: R code for replicating tables and plots in this thesis **90**

References **108**

Chapter 1

Introduction

Inverse probability of treatment weighted (IPTW) estimators for Marginal Structural Models (MSMs) are a popular class of models used to estimate causal effects. For example, the average causal effect of a binary exposure on an outcome of interest such as mortality. When experimental data is either unavailable, costly or unethical to collect, the IPTW estimator can correctly estimate causal effects using only observational data and a set of assumptions. In a survival context, or in a longitudinal setting more generally, the IPTW-estimator correctly resolves causal parameters in the presence of time-dependent confounding. However, it is often unclear when the assumptions that underlie this technique are justified or to what extent the performance of the estimator deteriorates when they are not strictly met. The IPTW-estimator is valid under the five assumptions of consistency, exchangeability, positivity, no model misspecification and no measurement error. While exchangebility has been studied in detail in the literature, the other assumptions have remained relatively neglected [1][2]. In particular, violations of the positivity assumption, which states that there is a positive probability of receiving every level of an exposure variable for every combination of values defined by the observed confounders in the analysis, remains relatively unknown. Nonetheless, positivity is of fundamental importance for estimating MSMs and is a necessary assumption for causal inference more generally.

In this thesis we explore the effect of violations of the positivity assumption on the performance of the IPTW-estimator by means of a simulation study. Since its introduction almost 20 years ago, the use of the IPTW-estimator has become widespread among epidemiologists as well as in other fields. This is in part due to the relative ease of applying the IPTW-estimator compared to other causal models such as G-computation or the the doubly robust (DR) estimator. However, a recent systematic review of twenty pharmacoepidemiologic studies, all published in 2012, indicated that the majority of these studies did not report whether the positivity assumption was checked [3]. Lack of attention toward the positivity assumption may be explained if violations of positivity are relatively benign vis-a-vis the IPTW-estimator. However it is well know that positivity violations can lead to unstable and highly variable estimates. A more pessimistic conclusion is that the relative obscurity of the positivity assumption means that it often goes undetected and unaccounted for.

This thesis offers several developments beyond what is currently available in existing studies that address the positivity assumption. First, many previous studies are consigned to a time-fixed context and do not easily generalize to the time-dependent context which is relevant for survival and longitudinal analysis. This is particularly important because positivity violations may be propagated through time as a consequence of the construction of the IPT weights. Their effects may therefore be accentuated or attenuated in a time-dependent context. In the current work we take the time-dependent context as our starting point. Second, studies which do consider positivity violations in a time-dependent context often use real data and are pedagogical in nature. As a result, it is difficult to disentangle the specific effect of positivity violations from other sources of bias which are often unavoidable when using real data. For example, model misspecification, which to some extent is present in all analysis which uses real data, can have a similar impact to positivity violations on the IPTW-estimator. A third advantage is that our approach to positivity occurs in a realistic

setting with wide applicability and one that is familiar to epidemiologists.

One reason why the effect of positivity violations on the IPTW-estimator has not been systematically studied in a longitudinal or survival context is the challenging task of simulating data under these conditions. In a nutshell, the problem is to replicate the complicated dynamics of time-dependent confounding and, specifically for survival or other non-collapsible models, to reconcile a structural or causal model with the conditional distributions typically used in Monte Carlo studies to generate data. Moreover, entertaining violations of the positivity assumption can only be achieved when the user has control over the relationship between the confounder variable and the exposure variable. This is particularly difficult in a time-dependent context because it requires control of the pathway between confounder and exposure at every point in time.

Our framework for investigating positivity violations exploits an existing algorithm for simulating data from a specific MSM in a time-dependent survival context. We extend this algorithm to allow positivity violations to occur within certain intervals of a single confounding variable. The advantage of our approach is that it can be applied to a wide range of scenarios in a systematic way and in a context which is realistic and of practical interest to analysts. Our approach is flexible enough to explore the interactions between positivity violations and the length of follow-up time in a typical survival study. As a result we can directly investigate how positivity violations are propagated through time. We also examine the finite sample properties of the IPTW-estimator under violations of the positivity assumption in varying sample sizes. Near violations of the positivity assumption, which arise when the probability of receiving a certain level of exposure is very small but not strictly zero, have similar effects to strict violations of the positivity violation. We are able to adapt our approach to investigating violations of positivity to near violations of positivity which further extends the applicability of our methods.

Our results confirm that violations of the positivity assumption can lead to a significant deterioration in the performance of the IPTW-estimator. In particular, we show that even relatively benign violations of the positivity assumption can heavily impact the performance of the estimator in terms of both the bias and variability of the estimates. When we extend the analysis to more complicated scenarios we find several interesting results. First, contrary to expectations, our results suggest that the performance of the IPTW-estimator actually improves with longer follow-up times. In contrast, an increase in the sample size of our study does not result in a decrease in bias. We also find that the IPTW-estimator actually performs quite well under conditions of near-positivity violations.

This thesis is structured as follows. In chapter 2 we introduce the counterfactual framework for causal inference, confounding and time-dependent confounding. In chapter 3 we introduced marginal structural models and the IPTW-estimator. Chapter 4 describes several core concepts in survival analysis. Chapter 5 describes the base algorithm used to generate the data used in this thesis. Chapter 6 discusses positivity violations in more detail and extends the algorithm to incorporate positivity violations. Chapter 7 presents the results of a simulation study. Chapter 8 concludes and provides a discussion and some avenues for future work.

Chapter 2

Causal Inference

Marginal structural models are a class of models for causal inference. More specifically, MSMs are models for some aspect of the distribution of counterfactual outcomes. In this chapter we introduce counterfactuals and define what is meant by a causal effect. Next, we introduce confounding, motivated by a central problem in the counterfactual framework; namely that counterfactual outcomes are never actually observed. We explain why controlling for confounders allows for correct estimation of causal effects in the time-fixed context under the three assumptions of consistency, exchangeability and positivity. We describe two important problems which emerge when moving from the time-fixed to the time-dependent context. Along the way, we also describe directed acyclic graphs and introduce the notation that will be used throughout this thesis.

2.1 Counterfactuals and causality

In the counterfactual framework [4][5][6] the causal effect of exposure A on outcome Y is defined as a contrast between the potential outcome when a subject is exposed to A and the potential outcome had that same subject remained unexposed to A. Clearly one potential outcome is counterfactual because, by definition, the same subject cannot be both exposed and unexposed to A. When A is a binary exposure we denote the potential outcome for subject i when exposed by $Y_{a=1,i}$ and $Y_{a=0,i}$ when the same subject is left unexposed. Among the exposed subjects, $Y_{a=0,i}$ is the outcome that would have occurred had the patient received $a = 0$. The causal effect for a single subject measured on a difference scale is $Y_{a=1,i} - Y_{a=0,i}$. A causal effect exists on the difference scale when $Y_{a=1,i} - Y_{a=0,i} \neq 0$.

The canonical example in the literature on estimating causal effects from MSMs is the effect of Zidovudine (AZT) on mortality among HIV positive subjects [7]. In this example, A is a binary variable equal to one if the subject received AZT and zero otherwise. The outcome Y is also a binary variable equal to one if the subject is dead and zero if the subject is alive. There are two possible ways that treatment (A) can be assigned to a single subject and two possible outcomes (Y) that can be observed. The resulting four combinations of A and Y are enumerated in table 2.1 alongside their corresponding factual and counterfactual outcomes.

Table 2.1: Enumerated potential outcomes for a single subject i

A_i	Y_i	$Y_{a=1,i}$	$Y_{a=0,i}$	$Y_{a=1,i} - Y_{a=0,i}$
1	1	1	?	?
1	0	0	?	?
0	1	?	1	?
0	0	?	0	?

The first row of table 2.1 represents the case where subject i is assigned treatment ($A = 1$) and experiences

the event $(Y = 1)$. The factual outcome $Y_{a=1}$ is observed and the counterfactual outcome in this case is $Y_{a=0}$. On the other hand, the third row of table 1 represents the case where subject i remains unexposed $(A = 0)$ and experiences the event $(Y = 1)$ in which case the factual outcome $Y_{a=0}$ is observed and $Y_{a=1}$ remains unobserved. Table 2.1 draws attention to the "fundamental problem of causal inference" [8][9]. No matter which row of data is observed, the contrast $Y_{a=1} - Y_{a=0}$ cannot be evaluated because every row of table 2.1 contains one unobserved counterfactual outcome. This has prompted several authors to cast causal problems in terms of missing data problems where the counterfactual outcomes are viewed as missing [10][11][12].

Typically interest lies in the average causal effect of treatment for a population of subjects rather than one subject. The average outcome in a population of exposed subjects is $E(Y_{a=1})$. The counterfactual outcome had the same population been left unexposed is $E(Y_{a=0})$. The average causal comparison in the counterfactual framework is then defined as

$$E(Y_{a=1}) - E(Y_{a=0}). \tag{2.1}$$

When Y is a dichotomous variable, $E(Y_a) = P(Y_a)$, and the average causal contrast can be rewritten as

$$P(Y_{a=1} = 1) - P(Y_{a=0} = 1). \tag{2.2}$$

We refer to $P(Y_a)$ as the counterfactual risk of experiencing the event $Y = 1$ under exposure $A = a$. If it was possible to observe both $P(Y_{a=1} = 1)$ and $P(Y_{a=0} = 1)$ for the same subjects, (2.2) could be immediately evaluated. But, just as in the case of a single subject, the average counterfactual outcome is not observed because the same population of subjects cannot be both exposed and unexposed. As a result, the average causal effect for a population is not immediately identifiable in the counterfactual framework.

2.2 Confounding

As counterfactual outcomes are never observed, subjects who are exposed to treatment (A=1) are typically compared to a different set of subjects who are left unexposed to treatment (A=0). Among the exposed subjects, the contrast that would ideally be made is

$$P(Y_{a=1} = 1 \mid A = 1) - P(Y_{a=0} = 1 \mid A = 1). \tag{2.3}$$

In words, the causal contrast of interest is the difference between the average outcome among the exposed when they receive treatment and the average outcome among the exposed had they, contrary to fact, remained unexposed. Because this involves the average outcome of unobserved counterfactuals $P(Y_{a=0} \mid A = 1)$, the average outcome among the exposed is typically compared to the average outcome among the unexposed

$$P(Y_{a=1} = 1 \mid A = 1) - P(Y_{a=0} = 1 \mid A = 0). \tag{2.4}$$

This carries a causal interpretation as long as $P(Y_{a=0} = 1 \mid A = 1) = P(Y_{a=0} = 1 \mid A = 0)$ or, in words, if the distribution of the counterfactual outcomes $Y_{a=0}$ is independent of the actual exposure A received. When this holds, unexposed subjects can be viewed as analogues of the exposed subjects had they, contrary to fact, not been exposed. On the other hand, if $P(Y_{a=0} = 1 \mid A = 1) \neq P(Y_{a=0} = 1 \mid A = 0)$ then (2.4), a measure of association, is confounded for (2.3), a measure of causal effect [13]. In the absence of confounding association is equal to causation. Resolving the "fundamental problem of causal inference" in the counterfactual framework requires establishing under what conditions $P(Y_{a=0} = 1 \mid A = 1)$ in (2.3) can be replaced by $P(Y_{a=0} = 1 \mid A = 0)$ as in (2.4).

One explanation for confounding is the existence of confounders, covariates which are common causes of both A and Y, which we denote by L. An example is a patient's CD4 cell count which is used by physicians to

determine whether AZT treatment should be initiated in HIV positive patients. A low CD4 count is a reason to begin AZT treatment. At the same time, CD4 is a risk factor for death and therefore associated with the outcome of interest Y. The key point is that the existence of confounders means that the distribution of the counterfactual outcomes Y_a is not independent of the actual treatment received. For example, within the group of exposed subjects, knowing that a physician initiated treatment indicates that CD4 count was low. Consequently, the risk of death is higher in the exposed group than in the unexposed group. This will be true even when no causal effect actually exists between A and Y. In this example, the distribution of the counterfactual outcomes in the exposed group $P(Y_{a=0} \mid A = 1) > P(Y_{a=0} \mid A = 0)$, the distribution of the counterfactual outcomes in the unexposed group.

2.2.1 Exchangeability

More generally, when $P(Y_a \mid A = 1) = P(Y_a \mid A = 0)$ for both $a = 0$ and $a = 1$, the *exchangeability* condition holds. Subjects are exchangeable in the sense that if the exposed and unexposed subjects were all exposed, the distribution of Y would not be different between the two groups [14]. Clearly, in the preceding example, this would not be the case. The distribution of Y in the exposed group differs from the distribution of Y in the group that is unexposed regardless of whether they actually receive their assigned exposure A. The difference arises from the different distributions of the confounder L in the exposed and unexposed groups. When the exchangeability assumption holds the distribution of the counterfactual outcomes Y_a is independent of the actual exposure received

$$Y_a \perp\!\!\!\perp A \quad \forall\, a. \tag{2.5}$$

From the missing data perspective, the exchangeability condition means that counterfactual outcomes are missing at random. When a confounder is present, the counterfactual outcomes are not missing at random because knowing that a subject comes from the exposed group ($A = 1$) predicts that subject's counterfactual risk of death under no exposure ($A = 0$) even when no causal effect of A on Y exists [15]. Causal effects like (2.3) only coincide with measures of association like (2.4) when the exchangeability condition holds.

2.2.2 Consistency

Up to this point, the expressions used to define causal effects have been cast in terms of counterfactuals. The *consistency* condition converts expressions involving counterfactuals into expressions involving ordinary conditional probabilities of measured data [16]. The consistency condition states that an individual subjects potential outcome $Y_{a,i}$, under their observed treatment $A = a$ is precisely their observed outcome Y

$$Y_{a,i} = Y_i \quad if\ A_i = a. \tag{2.6}$$

By the consistency assumption we can write

$$P(Y_{a=1} = 1 \mid A = 1) = P(Y = 1 \mid A = 1). \tag{2.7}$$

In words, (2.7) says that the expected value of the counterfactual outcome under exposure in the exposed group is exactly equal to the expected value of the observed outcome in the exposed group.

The consistency condition is violated when, for example, exposure is a variable that is not easily manipulated such as a biologic feature or when the exposure has side effects (see [17], [9] and [18] for more details and discussion). The consistency assumption makes it possible to rewrite (2.4) as (2.8)

$$P(Y_{a=1} = 1 \mid A = 1) - P(Y_{a=0} = 1 \mid A = 0) = P(Y \mid A = 1) - P(Y \mid A = 0), \tag{2.8}$$

which is an expression involving only measured data with a causal interpretation when the exchangeability condition holds and confounders are absent.

2.2.3 Conditional exchangeability

When confounders are present, a spurious marginal dependence exists between A and Y induced by L. Randomised trials avoid the statistical problems associated with confounding because A is randomly assigned to each subject ensuring that exposure is not related to any confounders. As a result, counterfactual outcomes are missing at random, confounding is absent in expectation and a naive or crude analysis which does not include the variable L carries a causal effect.

In comparison, observational studies, or studies where violations of study protocols or loss to foll up occur, often do exhibit confounding. One solution is to adjust or control for the confounding variable L. Adjustment refers to ways in which the dependence of Y on A can be adjusted to take account of the relationship of the confounder L with A and Y. Conditioning on (stratifying by) L is one method of adjustment which involves examining relations between A and Y within strata defined by the values of L [19]. Contrasts made within strata defined by the values of L will yield the correct causal effect in that strata because subjects with the same value of L are exchangeable in the absence of any other measured or unmeasured confounders. This is known as conditional exchangeability and states that the distribution of the counterfactual outcomes is independent of the exposure actually received, conditional on L

$$Y_a \perp\!\!\!\perp A|L \quad \forall\, a. \tag{2.9}$$

The intuition is that the difference in outcome between exposed and unexposed subjects cannot be due to L if all subjects in that strata have the same value of L [20]. Exposed subjects can be viewed as analogues of unexposed subjects had they been exposed, but only within the strata defined by $L = l$. Replacing (2.4) by (2.10) yields the correct causal effect within the strata defined by $L = l$

$$P(Y \mid A = 1,\ L = l) - P(Y \mid A = 0,\ L = l). \tag{2.10}$$

Knowing the value of L in (2.10) gives no more information about the distribution of the counterfactuals Y_a in that strata. If the correct causal effect is obtained in each strata $L = l$ as in (2.10), then averaging across all strata defined by L must yield the correct causal effect in the population as in (2.11).

$$\sum_l [P(Y_{a=1} \mid A = 1,\ L = l) - P(Y_{a=0} \mid A = 0,\ L = l)]P(L = l). \tag{2.11}$$

Examining relations within strata removes the effect of confounding. Exposed and unexposed within each strata are therefore exchangeable giving rise to conditional exchangeability.

2.2.4 Positivity

However, adjusting for confounders comes at a cost. First, estimating contrasts like (2.10) necessarily involves a smaller sample size because the comparison is made within a sub population of the total sample defined by $L = l$. The result is a loss of efficiency. More importantly, contrasts like (2.10) are only meaningful when there are both exposed and unexposed subjects within each strata defined by L. For example, if physicians always initiate treatment in subjects with a low CD4 cell count, then there will be no unexposed subjects with which to evaluate the second term of (2.10) at lower values of L. This condition is referred to as positivity and requires that that there be both exposed and unexposed subjects at every combination of the values of the values of the observed confounder(s) in the population under study [2]. The positivity assumption is violated when, for some confounder values, there are no exposed and unexposed subjects to be compared [1]. From the perspective of estimating $P(Y_{a=1} \mid A = 1,\ L = l)$ from (2.10) the positivity condition is equivalent to the condition that there is a positive probability of receiving exposure $A = a$ in every strata $L = l$

$$P(A = a \mid L = l) > 0 \quad \forall\, a. \tag{2.12}$$

Figure 2.1: Directed Acyclic Graphs (DAGs).

Applying the definition of conditional probability to $P(Y_{a=1} \mid A = 1,\ L = l)$ provides a slightly different perspective

$$P(Y = 1 \mid A = 1, L = l) = \frac{P(Y = 1, A = 1, L = l)}{P(A = 1, L = l)} = \frac{P(Y = 1, A = 1, L = l)}{P(A = 1 \mid L = l)P(L = l)}. \tag{2.13}$$

When $P(A = a \mid L = l)$ is non-positive, i.e. zero, the denominator of (2.13) is undefined.

Violations of positivity are more problematic than a loss of efficiency because positivity violations remain as the sample size grows whereas the efficiency of the estimate improves with the sample size. Provided that the three conditions of *consistency*, *conditional exchangeability* and *positivity* hold, (2.11) estimates the correct average causal effect even when confounder L is present.

2.3 Directed Acyclic Graphs

Causal effects, like those described in the previous two sections, can be represented using graphs. A graph consists of a finite set of vertices ν and a set of edges ϵ. The vertices of a graph correspond to a collection of random variables which follow a joint probability distribution $P(\nu)$. Edges in ϵ consist of pairs of distinct vertices and denote a certain relationship that holds between the variables [18]. The absence of an edge between two variables indicates that the variables are independent of one another and there is therefore no direct effect of the one variable on the other [9]. The direction of the causal relationship is denoted by an arrow and is acyclic because causal relationships between two variables only proceed in one direction. There are no feedback loops or mutual causation because in a causal framework a variable cannot be a cause of itself either directly or indirectly [21].

For example, figure 2.1 (left) represents a causal relationship from exposure A to outcome Y. There is a third variable L which is a common cause of both A and Y. Exposure A is assigned according to the conditional distribution $P(A \mid L = l)$. Once exposure has been assigned, the outcome Y is determined by both A and L through the conditional distribution $P(Y \mid A = a, L = l)$.

The graph encodes the spurious marginal dependence between A and Y through the "back door" path shown in figure 2.1 (left and right) as an inverted fork $A \leftarrow L \rightarrow Y$. Even when there is no causal effect from A to Y, represented in figure 2.1 (right), there is still a spurious marginal dependence between A and Y. Removing the edge between A and Y makes clear that the remaining dependence is spurious because the lack of an edge indicates that no causal effect exists. Blocking the "back door" path $A \leftarrow L \rightarrow Y$ in figure 2.1 is equivalent to estimating the causal effect of A on Y within values of L. Once we condition on L the marginal dependence between A and Y is removed. DAGs provide a useful way of recognizing causal structures like $A \leftarrow L \rightarrow Y$ which match our intuition about causal effects. A second advantage is that a DAG represent a natural way to think about simulating data because conditional independencies are explicitly encoded [22].

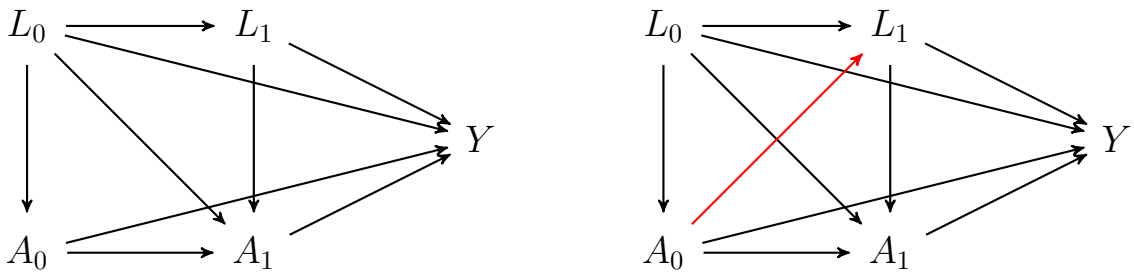


Figure 2.2: Time dependent confounding where A_0 is not a predictor of a subsequent confounder values L_1 (left). Time dependent confounding where A_0 is a predictor of subsequent confounder values L_1 (right).

2.4 Time dependent confounding

So far we have considered the time fixed context in which both exposure and confounders take on a single value. To identify a causal effect it was sufficient to block the "back door" path between exposure and outcome by conditioning on any confounders. We now broaden the setting to the time dependent case in which subjects are measured at several points in time. At each point in time a subject can receive a different exposure. Interest lies in whether exposure as a whole, the entire exposure history, has a causal effect on the survival of a patient [23].

First, we extend the time-fixed notation to include subject histories for time varying variables. Exposure and confounder histories up to visit k are represented by an overhead bar. For example, $\bar{A}_k = \{A_0, \ldots, A_k\}$ represents the vector of treatment decisions while $\bar{L}_k = \{L_0, \ldots, L_k\}$ represents the vector of measurements on the time-dependent confounder L. Variables that do not change over time such as sex, or covariates which change linearly over time like age are typically recorded at baseline ($t = 0$) and we denote the collection of baseline covariates as V_0. The outcome of interest at the end of follow-up is mortality Y which is a binary variable taking the value 1 if the subject is dead and 0 otherwise.

2.4.1 Problem 1: Spurious dependence versus indirect effects

Just as in the time-fixed case, time-dependent confounders lead to spurious associations between A and Y through a "back door" path $A_t \leftarrow L_t \rightarrow Y$. To estimate a causal effect it is necessary to block this path by conditioning on the confounding variables. Figure 2.2 (left) gives an example of this in the time dependent case for two periods ($t = 0, 1$). In the first period an exposure decision is made based on the measured confounder L_0. In the second period ($t = 1$) a new exposure decision is made based on both L_0 and L_1. Conditioning on the full confounder history $\bar{L}$ under this DAG leads to a consistent estimate of the causal effect because doing so blocks all paths between A_0, A_1 and Y except the causal path of interest.

However, the time-dependent context also admits structures like figure 2.2 (right). The addition of a causal relationship between A_0 and L_1 is represented by a red arrow. In the time dependent case it is possible for current exposure to be a determinant of future confounders which are in turn determinants of future exposure [24]. As a result the effect of A_0 on Y is mediated through L_1 on the path $A_0 \rightarrow L_1 \rightarrow Y$. Blocking this path by conditioning on $L1$ also blocks some portion of the effect of A_0 on Y and leads to a biased estimate of the full causal effect of $\bar{A}$ on Y. This presents a dilemma. On the one hand, failing to condition on L_1 admits spurious marginal dependencies between A_1 and Y and obscures the true causal effect. On the other hand, conditioning on L_1 when L_1 is an intermediate variable on the path $A_0 \rightarrow L_1 \rightarrow Y$ blocks some portion of the causal effect of A_0 on Y leading to a biased estimate of the true causal effect.

2.4.2 Problem 2: Spurious dependence versus selection bias

A second problem which emerges in the time-dependent context is selection bias, also known as collider stratification bias. Selection bias can be present in both a time-fixed and time-dependent context. However in a survival context, it becomes almost unavoidable because models for the hazard ratio (see chapter 5 section 4.3) have a built-in selection bias [25].

Selection bias is not an intuitive concept and is often explained through examples [26]. A commonly cited example is a school which admits students based on either academic ability or musical talent [18][14]. Students who have low academic ability but are admitted to the school must have been musically talented to gain admission. In other words, conditional on being admitted to the school academic ability and musical talent will be negatively correlated even if they are independent in the population of all students who applied for the school. The negative association between academic ability and musical ability is spurious and has no causal interpretation. The key point is to see that conditioning on a variable (admission to the school) which is a common effect of two other variables (musical ability and academic ability) leads to the spurious association.

Like confounding due to a common cause shown in figure 2.1, selection bias can be approached structurally and represented figuratively in a DAG [21]. In the school example, admission to the school (S) is a common effect of academic ability (A) and musical talent (M). This is represented figuratively in figure 2.3 (left). A graphical representation allows us to identify the same structure when it appears in the time-dependent context of figure 2.3 (right). Now, an unobserved variable U is a predictor of the confounder value L_t. Conditioning on L_t induces an association between U and A_0, depicted by the red arrows of figure 2.3 (right). Because U is also associated with Y in figure 2.3 (right), there is an indirect, spurious association between A_0 and Y. This is a problem for a similar reason to the indirect effect of A_0 on Y explained in section 2.4.1. Conditioning on L_1 introduces a spurious association between A_0 and Y through selection bias. Failing to condition on L_1 will lead to a biased estimate of the true causal effect when L_1 is indeed a confounder.

Of course, selection bias with a structure like figure 2.3 (right) only arises when there is an unobserved variable U. In a survival context this will be almost unvoidable because subjects vary in their susceptibility (U) to diseases like AIDS or to the negative effects of exposure like toxicity [25]. If we randomly assign exposure in a time-fixed context then the effect of U will be absent in expectation and conditional exchangeability will hold (conditional on any measured confounders). In contrast, in a time-dependent survival context, as time passes then the at risk population at any moment in time will contain only those subjects who survived up to that time. If a binary exposure A has a protective effect then subjects who receive exposure are less likely to experience the event. Knowing that a subject did not receive exposure and still survived up to, say, time $t = 10$ indicates that they are either less susceptible to the disease or perhaps unaffected by toxicity or other side effects of exposure. The key point is that although at baseline randomization ensured exposed and unexposed subjects were exchangeable, at later points in time we find that the distribution of the counterfactual outcome Y_a is no longer independent of A.

It should be clear that from the perspective of causal inference, selection bias and the indirect effect of A_0 on Y through L_1 have the same root cause. In both cases, conditioning on a confounder L induces a spurious relationship. When we condition on a confounder L under a causal structure like that of figure 2.3 we induce spurious dependencies between A and Y. However, if we fail to adjust for L when L really is a confounder then we admit marginal dependencies between A and L which are associational and not causal. Different methods are required in order to carry out causal inference in the time dependent context.

2.5 Effect measures

So far we have only considered causal effects as contrasts between potential outcomes on the difference scale. In this section we extend the discussion to causal effects measured on the ratio scale and by the odds ratio.

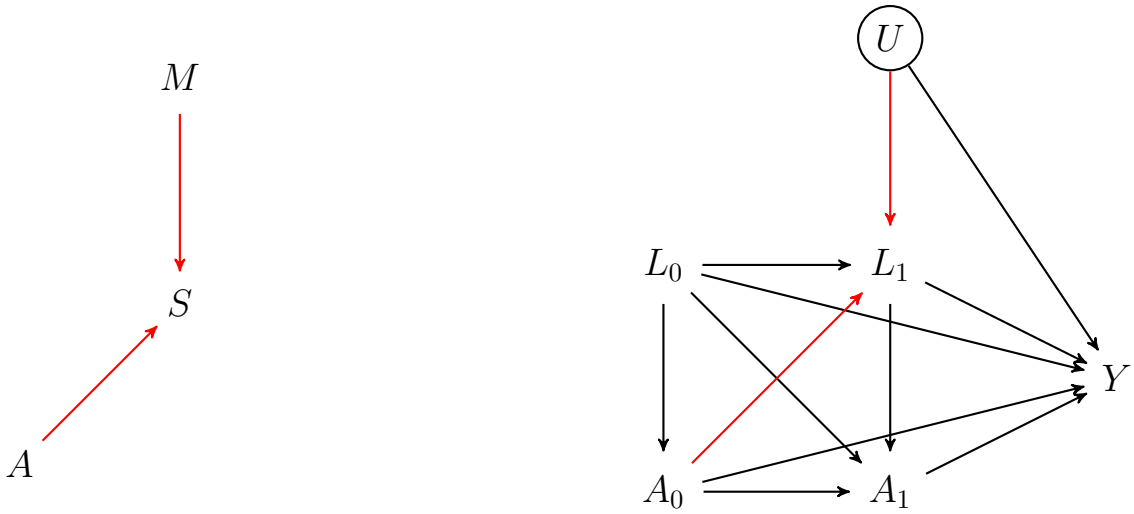


Figure 2.3: Selection bias: example of a selective school (left). Selection bias: time dependent survival context (right).

On the ratio scale, the causal contrast for a single individual is defined as the ratio between the potential outcome when subject i is exposed to A and the potential outcome when the same subjects is left unexposed to A. When no causal effect is present this ratio is equal to 1. Analogously, when Y is a binary variable, the average causal risk ratio is defined as in (2.14). A causal risk ratio equal to 1 indicates that there is no causal effect.

$$\frac{P(Y_{a=1})}{P(Y_{a=0})} \tag{2.14}$$

The causal odds ratio is defined as the ratio of the odds that the potential outcome $Y_{a=1} = 1$ when all subjects are exposed $A = 1$ versus the odds of the event that the potential outcome $Y_{a=1} = 1$ when all subjects remain unexposed $A = 0$.

$$\frac{P(Y_{a=1} = 1)}{P(Y_{a=1} = 0)} / \frac{P(Y_{a=0} = 1)}{P(Y_{a=0} = 0)} = \frac{P(Y_{a=1} = 1)P(Y_{a=0} = 0)}{P(Y_{a=1} = 0)P(Y_{a=0} = 1)}. \tag{2.15}$$

A causal odds ratio equal to 1 indicates no causal effect whereas a causal odds ration greater than 1 indicates that A and Y are associated.

Under the the conditions of conditional exchangeability, consistency and positivity, and in the absence of time dependent confounding, the causal risk ratio can be calculated by conditioning on any confounding variables L in an analogous fashion to the causal risk difference. Within the strata defined by $L = l$, for example, the correct causal risk ratio is

$$\frac{P(Y \mid A = 1, L = l)}{P(Y \mid A = 0, L = l)}, \tag{2.16}$$

and the correct causal risk ratio in the population as a whole is

$$\sum_l \left[\frac{P(Y \mid A = 1, L = l)}{P(Y \mid A = 0, L = l)}\right] P(L = l). \tag{2.17}$$

The causal odds ratio, on the other hand, is a non-collapsible effect measure. This means that the size of the odds ratio may change due to conditioning on a variable L even when L is not actually a confounder [27]. The average of the strata specific causal effects measured by the odds ratio is therefore not necessarily the same as the true marginal causal effect. This turns out to be extremely important when simulating data from a specific MSM. Further discussion is postponed to chapter 5 section 5.2.2.

2.6 Model formulation for causal effect measures

We conclude this chapter by relating the the causal risk difference, risk ratio and odds ratio to a linear parametric model formulation. Generalized linear models [28] assume that the mean of a variable X is related to a linear predictor through a link function. For the distribution of the counterfactual outcomes Y_a this can be expressed as in (2.18)

$$E(Y_a) = P(Y_a) = g^{-1}[\alpha_0 + \alpha_1 a]. \tag{2.18}$$

In words, the mean of the counterfactual distribution of Y_a is related to a linear predictor which includes exposure a through the link function g. For example, using the identity link function (2.19), allows us to capture the causal risk difference (2.20) where the mean of counterfactual distribution is linearly related to A.

$$P(Y_{a=1}) = \alpha_0 + \alpha_1 a \tag{2.19}$$

$$P(Y_{a=1}) - P(Y_{a=0}) = \alpha_0 + \alpha_1 \cdot 1 - (\alpha_0 + \alpha_1 \cdot 0) = \alpha_1. \tag{2.20}$$

Similarly, using a log link as in (2.21) allows us to express the causal risk ratio as (2.21).

$$log[P(Y_{a=1})] = \alpha_0 + \alpha_1 a \tag{2.21}$$

$$log\Big[\frac{P(Y_{a=1})}{P(Y_{a=0})}\Big] = log[P(Y_{a=1})] - log[P(Y_{a=0})] = \alpha_0 + \alpha_1 - (\alpha_0 + \alpha_1) = \alpha_1 \tag{2.22}$$

The causal risk ratio can be evaluated by (2.23)

$$\frac{P(Y_{a=1})}{P(Y_{a=0})} = e^{\alpha_1}. \tag{2.23}$$

Finally, the odds ratio can be expressed as a function of a linear regression using the logit link function as in (2.24). For notational simplicity we will write (2.24) using the logit function as in (2.25) in the remainder of this thesis.

$$log\Big[\frac{P(Y_a)}{1 - P(Y_a)}\Big] = \alpha_0 + \alpha_1 a \tag{2.24}$$

$$logit\ P(Y_a) = \alpha_0 + \alpha_1 a \tag{2.25}$$

Using (2.24) we can write the log of the causal odds ratio as (2.26)

$$log\frac{P(Y_{a=1})}{1 - P(Y_{a=1})} / \frac{P(Y_{a=0})}{1 - P(Y_{a=0})} = log\frac{P(Y_{a=1})}{1 - P(Y_{a=1})} - log\frac{P(Y_{a=0})}{1 - P(Y_{a=0})} = \alpha_0 + \alpha_1 1 - (\alpha_0 + \alpha_1 0) = \alpha_1, \tag{2.26}$$

and the causal odds ratio is therefore

$$\frac{P(Y_{a=1})}{1-P(Y_{a=1})}/\frac{P(Y_{a=0})}{1-P(Y_{a=0})} = e^{\alpha_1}. \tag{2.27}$$

Chapter 3

Marginal Structural Models

Marginal structural models (MSMs) are a class of models which can be used to estimate causal effects in the presence of time-dependent confounding. Chapter 1 section 2.4 described how conditioning on confounder history $\bar{L}$ in a time-dependent context introduces bias by blocking the path $\bar{A} \rightarrow \bar{L} \rightarrow Y$ and potentially opening the path $\bar{A} \rightarrow \bar{L} \leftarrow U \rightarrow Y$. As a result, under these circumstances, conditioning on a confounder will not resolve the correct causal effect, while failing to condition on a confounder will lead to bias through the path $\bar{A} \leftarrow \bar{L} \rightarrow Y$. Marginal structural models can be used to estimate causal effects under these conditions using only observed data and a set of assumptions.

In this chapter we begin by defining exactly what is meant by a marginal structural model. Next we describe the rationale behind the inverse probability of treatment weighting estimator which can be used to estimate a marginal structural model for causal effects in a time dependent context. Estimation of MSMs requires specifying two models. A weight model is specified for the probability of receiving exposure conditional on any covariates. A structural model is also specified relating the outcome of interest Y to exposure A. The weight model is used to construct subject or subject-time specific weights which are then applied in a weighted estimation of the structural model. Marginal structural models are only valid when the three conditions of exchangeability, consistency and positivity introduced in chapter 1, are met. We connect each of these assumptions to the estimation of marginal structural models and introduce the assumptions of no measurement error and no model misspecification which are also required for unbiased estimation of marginal structural models.

3.1 Marginal structural models

A marginal structural model is a model for some aspect of the distribution of counterfactual outcomes. Chapter 1 section 2.6 explained how a causal effect measured by the risk difference, risk ratio or odds ratio, could be expressed by a model formulation. A MSM for the odds ratio as a measure of causal effect, for example, can be formulated in terms of counterfactual outcomes as in (3.1).

$$logit\ P(Y_a = 1) = \alpha_0 + \alpha_1 a \tag{3.1}$$

Equation (3.1) models the relationship between the mean of the counterfactual outcomes Y_a and exposure A in a linear logistic fashion. The resulting causal effect is the odds ratio e^{α_1}. Importantly, equation (3.1) is marginal with respect to any confounders meaning that equation (3.1) does not involve conditioning on any confounders. The term structural, which migrated from the econometrics and social sciences literature, is sometimes used in place of causal to describe models for causal effects. Marginal structural models are models for counterfactual outcomes and therefore they are causal models. They are also marginal over confounders, earning the name marginal structural model or MSM.

In contrast, an associational model like (3.2) may carry a causal interpretation in the time-fixed case under the assumptions of consistency, conditional exchangeability and positivity (see chapter 1 section 2.2 and section 2.6). The coefficient β_1 in (3.2) will only equal the causal parameter α_1 in (3.1) when no confounding is present in which case the associational parameter β_1 coincides with the causal parameter α_1. In fact, due to non-collapsibility, it may be the case that $\alpha_1 \neq \beta_1$, even in the absence of confounding. We again postpone further discussion on non-collapsibility to chapter 5 section 5.2.2. Model (3.2) is not marginal because it involves conditioning on L.

$$logit\ P(Y = 1 \mid A = a, L = l) = \beta_0 + \beta_1 a + \beta_2 l \tag{3.2}$$

In the time dependent context the distinction between a MSM and an associational or regression model like (3.2) becomes clearer. For example, in a survival setting, a marginal structural model for the hazard function could be specified as a linear logistic function of the period specific exposures up to the current exposure a_t.

$$logit\ P(Y_{\bar{a}_t} = 1) = \alpha + \beta_1 a_0 + \cdots + \beta_k a_t \tag{3.3}$$

In the time-dependent context with a confounder L the regression formulation of (3.3) is (3.4). The regression formulation will not carry a causal effect even when the consistency, exchangeability and positivity conditions are met. This is because the act of conditioning on L blocks the path $\bar{A} \rightarrow \bar{L} \rightarrow Y$ and potentially opens the path $\bar{A} \rightarrow \bar{L} \leftarrow U \rightarrow Y$ (see section 2.4). Failing to condition on L, on the other hand, leaves open the common cause path $\bar{A} \leftarrow \bar{L} \rightarrow Y$.

$$logit\ P(Y = 1 \mid \bar{A}_t = \bar{a}_t, \bar{L}_t = \bar{l}_t) = \alpha + \beta_1 a_0 + \cdots + \beta_k a_t + \delta_1 l_0 + \cdots + \delta_k l_t \tag{3.4}$$

Estimation of a MSM for aspects of the distribution of counterfactual outcomes requires methods which do not involve conditioning on $\bar{L}$. We now describe a technique which can be used to estimate MSMs in a time dependent context.

3.2 Inverse Probability of Treatment Weighting

MSMs can be estimated in the presence of time-dependent confounding by applying a technique called inverse probability of treatment weighting (IPTW). We begin this section by describing the rationale and intuition behind why IPTW estimation can be used to estimate causal models in a time-dependent context. In particular, we explain why the IPTW-estimator avoids conditioning on a confounder L thereby avoiding the problems described in chapter 1 section 2.4. This is followed by a description of how to construct both the unstabilized and stabilized IPT weights.

3.2.1 Rationale behind IPTW

In chapter 1 section 2.2.3, when a confounder L was present, exposed and unexposed subjects were conditionally exchangeable within the strata defined by values of L provided there are no other measured or unmeasured confounders. By conditioning on (stratifying by) a particular value of $L = l$, the unexposed subjects within that strata could be viewed as analogues of exposed subjects, had the exposed subjects, contrary to fact, not been exposed. The correct causal effect could be evaluated by contrasting the outcomes in the exposed and unexposed subjects within each strata and then averaging strata specific causal effects across all strata. However, we saw in section 2.4 that this does not work in a time-dependent context.

Instead, suppose that within the strata defined by $L = l$ we knew in advance that the probability a subject is exposed is $p(A = 1 \mid L = l) = \frac{1}{3}$ and the probability that a subject is unexposed is $p(A = 0 \mid L = l) = \frac{2}{3}$. If there are twelve subjects in the strata defined by $L = l$ we would expect the exposed and unexposed subjects to appear in the proportions $\frac{1}{3}$ and $\frac{2}{3}$ respectively. In other words, we would expect to see four exposed

subjects and eight unexposed subjects. This situation is represented in table 3.1 for the strata defined by $L = l$. In table 3.1, half the exposed subjects experience the event $Y_{a=1} = 1$ and six out of eight unexposed subjects experience the event $Y_{a=0} = 1$.

Table 3.1: Exposed and unexposed subjects within the strata $L = l$

Subject	$Y_{a=1}$	$Y_{a=0}$
1	1	?
2	1	?
3	0	?
4	0	?
5	?	1
6	?	1
7	?	1
8	?	1
9	?	1
10	?	1
11	?	0
12	?	0

Within the strata $L = l$ we can view the exposed subjects as analogues of the unexposed subjects had they been exposed because of conditional exchangeability. Half the exposed subjects experienced the event $Y = 1$. If the eight unexposed subjects had been exposed, then we would expect the proportion of subjects who experienced the event to also be 1/2, the same proportion as was factually the case in the exposed subjects. Equation (3.5) carries exactly this logic.

$$P(Y_{a=0} = 1 \mid A = 1, L = l) = P(Y_{a=1} = 1 \mid A = 1, L = l) = P(Y = 1 \mid A = 1, L = l) = \frac{1}{2} \tag{3.5}$$

The first inequality in (3.5) holds by the conditional exchangeability condition introduced in chapter 1 section 2.2.1 and the second equality holds by the consistency assumption introduced in chapter 1 section 2.2.2. Although we do not know the counterfactual outcomes $Y_{i,a=1}$ for any particular unexposed subject in table 3.1, we would expect half the unexposed subjects to experience the event $(Y = 1)$ and half not to experience the event $Y = 0$, if they had been exposed.

Combined with the subjects in table 3.1 who actually did receive treatment we would expect the vector $Y_{a=1}$ to contain six subjects who experience the event and six who do not. The key point is to recognize that weighting the four subjects who were exposed by the inverse probability of appearing in the strata $L = l$ achieves exactly this result. The inverse of the probability of exposure in strata $L = l$ is

$$w_i = \frac{1}{pr(A = 1 \mid L = l)} = \frac{1}{1/3} = 3. \tag{3.6}$$

Weighting-up each exposed subject by w_i creates three replicates of each exposed subject and $4 \times 3 = 12$ replicates in total. Similarly weighting-up the unexposed subjects by the probability of remaining unexposed in the strata $L = l$ creates 1.5 copies of each unexposed subject and twelve unexposed subjects in total. In the new population the probability of exposure is 0.5 because there are exactly twelve exposed subjects and twelve unexposed subjects. In other words, L is no longer a predictor of whether or not a subject receives exposure in the new population and A is independent of L.

$$L \perp\!\!\!\perp A$$

We know from chapter 1 section 2.2, that the causal effect of A on Y within $L = l$ is correct under the assumptions of exchangeability, consistency and positivity. Crucially, the causal effect in the new population

in strata $L = l$ is the same as the causal effect in the original population in strata $L = l$ because the same proportion of subjects experience the event in the new population as in the original population.

Repeating the weighting process in every strata creates a pseudo-population in which A is independent of L in every strata. Subjects are marginally exchangeable because, under conditional exchangeability, the new population contains equal numbers of exposed and unexposed subjects who are the same persons under a different exposure level [15].

3.2.2 IPTW in the time-dependent context

The logic behind the IPTW estimator extends to the time-dependent context. Exposure and confounder histories evolve over time. Subjects with the same confounder history $\bar{L}_t$ and exposure history prior to the current treatment $\bar{A}_{t-1}$ are exchangeable. The intuition is that the difference in outcome between subjects who receive exposure at time t and those who remain unexposed at time t cannot be due to $(\bar{A}_{t-1}, \bar{L}_t)$ among subjects with the same values of $(\bar{A}_{t-1}, \bar{L}_t)$. Formally

$$Y_{\bar{a}_t} \perp\!\!\!\perp A_t \mid \bar{A}_{t-1}, \bar{L}_t \tag{3.7}$$

Weighting each subject by the inverse probability of receiving their own exposure conditional on their exposure and confounder trajectory $pr(A_t = 1 \mid \bar{A}_{t-1} = \bar{a}_{i,t-1}, \bar{L}_t = \bar{l}_{i,t})$ creates a pseudo-population in an analogous fashion to the time-fixed case. In the new population L_t no longer predicts A_t and subjects are exchangeable at every point in time provided that there are no unmeasured confounders. This is known as sequential exchangeability.

The most important point is that within the pseudo-population created by weighting by the IPTW the correct causal effect can be evaluated without the need to condition on L. Conditioning on L blocked intermediate effects as described in chapter 1 section 2.4. In the pseudo-population, estimation of the causal effect is performed unconditionally and the intermediate effect of $\bar{A}$ on Y is not blocked by conditioning. As a result the spurious dependencies introduced by blocking some portion of the effect of $\bar{A}$ on Y through $\bar{L}$ on the path $\bar{A} \rightarrow \bar{L} \rightarrow Y$ or via selection bias on the path $\bar{A} \rightarrow \bar{L} \leftarrow U \rightarrow Y$ do not arise.

Robins et al (2000) give the example of a follow-up study of the effect of AZT on whether or not HIV is detectable in the blood of patients. Exposure A_t is AZT at time t and the outcome Y is a dichotomous variable equal to 1 if HIV is detected in the blood of the subjects at the end of follow-up. Follow-up occurs for T time periods. A MSM for the causal effect of A on Y in this context is

$$logit\ pr(Y_{\bar{a}} = 1) = \beta_0 + \beta_1 cum(\bar{a}), \tag{3.8}$$

which models the counterfactual outcome at the end of follow up as a linear logistic function of the cumulative AZT exposure to the end of follow up. The function $cum(\cdot)$ is the cumulative sum and $cum(\bar{a})$ is the cumulative sum of exposure history. In this case, the probability that a subject receives their full exposure trajectory is the product of the probability that they receive their exposure in each (discrete) time period. Each subject receives the terminal weights

$$w_i = \frac{1}{\prod_{t=0}^{T-1} pr_t(A_{i,t} \mid \bar{A}_{i,t-1}, \bar{L}_{i,t})}. \tag{3.9}$$

Where $\bar{a}_T$ represents the patients complete exposure trajectory up to time interval $T-1$. Applying weighted logistic regression to the MSM (3.8) using the weights w_i will estimate the correct causal parameter β_1 under five assumptions which we describe in section 3.3.

3.2.3 Construction of IPTW weights

Up to this point we have assumed that the probability of receiving exposure within the strata defined by L or $(\bar{A}_{t-1}, \bar{L}_t)$ is known. Subjects are weighted by the inverse of the probability of receiving their exposure or exposure trajectory. In reality, these probabilities are unknown and need to be estimated from the data. In this section we describe precisely how to estimate these probabilities in order to construct the weights in a time-dependent context.

Estimating the subject specific weight probabilities in both (3.9) means estimating the probability that a subject receives their exposure within each time period t. In other words, we need a subject-time specific model for the probability $pr(A_t \mid \bar{A}_{t-1}, \bar{L}_t)$ which can then be used to fit a probability per subject and per time period. As an example, (3.10) models the probability that a subject is exposed at time t as a linear logistic function of the current time period, previous exposure and current values of the confounder L.

$$logit\ pr(A_t = 1 \mid \bar{A}_{t-1} = \bar{a}_{t-1}, \bar{L}_t = \bar{l}_t) = \alpha_0 + \alpha_1 t + \alpha_2 a_{t-1} + \alpha_3 l_t \tag{3.10}$$

It is worth mentioning the difference between a linear logit model for the counterfactual outcomes $Y_{\bar{a}}$ as in (3.8) and a linear logit model for the weight probabilities as in (3.10). In the former case, interest lies in the causal effect of A and Y measured by the odds ratio. The relationship between the logit formulation and the odds ratio is described in more detail in chapter 2 section 2.6. In the latter case of (3.10), interest lies in fitting the actual response probability $pr(A_t = a_{i,t} \mid \bar{A}_{t-1} = \bar{a}_{i,t-1}, \bar{L}_t = \bar{l}_{i,t})$ in order to construct the weights. The use of a logistic model, as opposed to modelling $pr(A_t = 1 \mid \bar{A}_{t-1} = \bar{a}_{t-1}, \bar{L}_t = \bar{l}_t)$ directly as a linear function of exposure and confounder history, is to ensure that the response probabilities lie on the interval 0 to 1.

After estimating the parameters $(\alpha_0, \alpha_1, \alpha_2, \alpha_3)$ in (3.10), the probability that a subject i is exposed conditional on their own exposure and confounder history can be evaluated as in (3.11).

$$\hat{pr}(A_t = 1 \mid \bar{A}_{t-1} = \bar{a}_{i,t-1}, \bar{L}_t = \bar{l}_{i,t}) = \frac{1}{1 + e^{-(\hat{\alpha}_0 + \hat{\alpha}_1 t + \hat{\alpha}_2 a_{i,t-1} + \hat{\alpha}_3 l_{i,t})}}. \tag{3.11}$$

The weights are referred to as inverse probability of treatment weights because each subject is weighted by the estimated probability that they receive their own exposure and confounder history [7]. Estimating and predicting the weights per subject only depends on observed data. Equation (3.11) fits the weights for the exposed subjects ($A_t = 1$). For the unexposed subjects the correct estimated probability of remaining unexposed ($A_t = 0$) is therefore

$$\begin{aligned} \hat{pr}(A_t = 0 \mid \bar{A}_{t-1} = \bar{a}_{i,t-1}, \bar{L}_t = \bar{l}_{i,t}) &= 1 - \hat{pr}(A_t = 1 \mid \bar{A}_{t-1} = \bar{a}_{i,t-1}, \bar{L}_t = \bar{l}_{i,t}) \\ &= 1 - \frac{1}{1 + e^{-(\hat{\alpha}_0 + \hat{\alpha}_1 t + \hat{\alpha}_2 a_{i,t-1} + \hat{\alpha}_3 l_{i,t})}}. \end{aligned} \tag{3.12}$$

The predicted weights are then used in place of the probabilities in (3.9).

3.2.4 Stabilized weights

The denominator in the subject-specific weights can be extremely variable if the probability that a subject receives their treatment is very strongly associated with L [7]. For example, if the probability that a subject receives their treatment within the strata $L = l_0$ is 0.05 then that subject will contribute twenty copies of themselves to the analysis. Very large inverse probability weights can result in a few subjects dominating the analysis leading to an increase in variance.

Stabilized weights (3.13), in contrast to the unstabilized weights (3.9), are advisable because they lead to more efficient estimation. The stabilized weight equivalent of (3.9), for example is

$$sw_i = \frac{\prod_{t=0}^{T-1} p_t(A_{i,t} \mid \bar{A}_{i,t-1})}{\prod_{t=0}^{T-1} p_t(A_{i,t} \mid \bar{A}_{i,t-1}, \bar{L}_{i,t})}. \tag{3.13}$$

The numerator for the weights is estimated in the same way as the denominator. The important difference is that the numerator of (3.13) does not condition on the confounder L. Consequently, an interpretation of the stabilized weights is that they quantify the degree to which exposure is statistically non-exogenous [29]. The stabilized weights for a subject will only be equal to one when the numerator and denominator are equal. In other words, when the probability of receiving current exposure does not depend on L in which case, there is no confounding for that subject.

As a concrete example, model (3.10) could be adapted for the numerator of (3.13) by removing the terms in L

$$logit\ pr(A_t = 1 \mid \bar{A}_{t-1} = \bar{a}_{t-1}) = \alpha_0 + \alpha_1 t + \alpha_2 a_{t-1}. \tag{3.14}$$

The subject specific response probabilities from (3.14) can then be used to predict the response probabilities in the numerator of (3.13). It can be shown that the mean of the stabilized weights for all subjects is equal to one [15].

3.3 Assumptions

The IPTW-estimator can be used to estimate a MSM under the assumptions of consistency, exchangeability, positivity, no measurement error and no model misspecification. The first three of these assumptions were described in relation to causal inference and confounding in chapter 1 section 2.2. In this section we explain how these assumptions relate to the estimation of marginal structural models via the IPTW-estimator.

3.3.1 Consistency

The consistency condition states that an individual subjects potential outcome $Y_{a,i}$, under their observed treatment A is precisely their observed outcome Y

$$Y_{a,i} = Y_i \quad if \quad A_i = a. \tag{3.15}$$

The consistency assumption is necessary to make inferences about Y_a using observational data because it connects the observational data to the potential outcomes. In the context of IPTW estimation, the consistency assumption is used to justify why we can treat the observed data in table 3.1 as a potential outcome. In an experimental or observational setting we only ever observe Y_i and A_i. The consistency assumption justifies why we can write

$$P(Y = 1 \mid A = 1) = P(Y_{a=1} \mid A = 1). \tag{3.16}$$

3.3.2 Exchangeability

Exchangeability is the assumption that the distribution of the counterfactual outcomes Y_a is independent of exposure.

$$Y_a \perp\!\!\!\perp A \quad \forall\ a$$

Conditional exchangeability in chapter 1 section 2.2.3 held when, within levels, of a confounder variable the difference between exposed and unexposed subjects were not due to common causes.

$$Y_a \perp\!\!\!\perp A|L \quad \forall\, a$$

That is, within levels defined by $L = l$ subjects are exchangeable. In the time-dependent setting this was extended to a subject's complete trajectory.

Violations of exchangeability occur when the distribution of confounding variable differs between the exposed and unexposed groups. Unmeasured confounders would therefore violate the exchangeability condition. For that reason, exchangeability implies the *no unmeasured confounding* assumption. In practice, information on a large number of confounders can be gathered and expert opinion used to justify control for all relevant confounders. The exchangeability assumption is untestable [30][31].

3.3.3 Positivity

The positivity assumption states that there must be a positive (i.e. non-zero) probability that a subject is exposed within every strata defined by the exposure and confounder histories. In IPTW-estimation each subject is weighted by the inverse of their probability of experiencing their own exposure, for example in (3.13) if any of the period specific probability of receiving exposure and confounder history are zero then the weights are undefined.

In a similar way to the unstabilized weights, violations, or near-violations of the positivity weights, lead to very low estimated probabilities that a subject receives their own exposure and confounder trajectory. This results in an increase both the bias and the variance of the causal effects estimate. Violations of the positivity can be one reason why the mean of the stabilized weights deviate from one. We postpone further discussion of the positivity assumption to chapter 6.

3.3.4 No model misspecification

Estimating MSMs with a continuous confounder involves specifying a model for the weight probabilities and a structural model relating exposure to outcome. Both models require correct specification to obtain unbiased estimates. The structural model requires positing a relationship between exposure and outcome. For example, this relationship may be best captures through a linear relationship, threshold dose-response or a model accounting for long and short term effects of exposure [9].

Importantly, incorrect specification of the weight model for the weights can lead to large weights for some subjects. Similar to the positivity assumption, weight probability model misspecification is one possible explanation for observed deviations of the mean of the stabilized weights from one. A misspecified-specified model can lead to larger weights increasing bias and variance. Parametric models are unlikely to be perfectly specified but should provide a good approximation of the true model. Several studies have examined the effect of model misspecification in the estimation of MSMs including [9]. In the simulation setting of this thesis the true MSM and model for the weight probabilities is known and therefore model misspecification is not a problem

3.3.5 No measurement error

Measurement error can affect the outcome, exposure or confounder and other covariates used to estimate MSMs. This can arise due to faulty equipment, poor recall by survey respondents or simply carelessness and rounding. In each case the observed variable X^* differs from the true underlying variable X. In general this will result in bias but the extent of that bias depends on the process through which the error is introduced and whether with error is recorded. In this thesis we employ a simulation algorithm to generate data from a known marginal structural model and the simulated variables have no measurement error. Analogous to the

case of no model misspecification, this makes it possible to study the properties of MSMs in the absence of measurement error. More detail on the effect of measurement error in a causal context can be found in [32].

Chapter 4

Survival Concepts

This chapter reviews several important concepts in survival analysis which are relevant to this thesis. Survival analysis is the study of the distribution of life times. We introduce the three key quantities used in survival analysis; the probability density or mass function, the survival function and the hazard function and explain how they relate to one another. Just as in Chapter 2, confounding can arise in survival models requiring some form of adjustment which is often performed by the Cox proportional hazards model for the hazard function. The issue of estimating causal effects in the presence of time-dependent confounding extends straightforwardly to survival analysis which is inherently time-dependent in nature. In particular, the hazard function, which involves conditioning on survival up to time t, means that selection bias is almost automatically an issue even in randomized trials. Marginal structural survival models are one solution to dealing with the issue of time-dependent confounding. The basic concepts reviewed in this chapter are a condensed version of those presented in Klein and Moeschberger (2003) [33].

4.1 Survival function

The survival time or lifetime T is the time interval between a well defined start point or time origin t_0 and a well defined end point t_e. For example the time between an individuals birth (t_0) and death (t_e), or the time at which a subject is exposed to a drug (t_0) and the time at which they die from a disease (t_e).

The survival function evaluated at time t is the probability that an individual survives beyond time t which is equivalent to the probability that $T > t$. The survival function is expressed by (4.1).

$$S(t) = P(T > t) = 1 - F(t). \tag{4.1}$$

The second equality in (4.1) holds because $P(T > t) = 1 - P(T < t) = 1 - F(t)$ where $F(\cdot)$ is the cumulative distribution function. At any point in time, the survival function tells us the probability that the event occurs at some point subsequent to the current time. In a sense the survival function evaluated at t "leans in" to subsequent time periods because it monitors the probability that the event occurs beyond t. In the continuous case the survival function is the integral of the probability density function $f(t)$ between the current time t and ∞. The survival function at t_0 is $S(t_0) = 1$ and $S(\infty) = 0$ confirming that, over a long enough time period the probability of survival is zero.

$$S(t) = P(T > t) = \int_t^{\infty} f(x)dx \tag{4.2}$$

Consequently the probability density function can be expressed in terms of the derivative of the survival function.

$$f(t) = -\frac{dS(t)}{dt}, \tag{4.3}$$

In the continuous time context, the survival function is a continuous and monotonically non-increasing function. In many real world applications data is gathered, and events occur, in discrete time steps and intervals. For example, relapse among cancer patients may only be detected when the patient visits a hospital in which case time cannot be treated as continuous. The discrete equivalent of (4.2) for the survival function is given by (4.4)

$$S(t) = P(T > t) = \sum_{t_j > t} p(t), \tag{4.4}$$

where $p(t) = p(T = t)$ is the probability mass function for the probability that the event is experienced at time t. In words, (4.4) expresses the probability that the event takes place beyond time t as the sum of the discrete period specific probabilities of the event occurring. The difference between the survival function evaluated at two adjacent points in time is equal to the probability mass function.

$$p(t_j) = S(t_{j-1}) - S(t_j) \tag{4.5}$$

4.2 Hazard function

The hazard function (also known as the hazard rate) expresses *the approximate probability of an individual of age t experiencing the event in the next instant.* In other words, the hazard function evaluated at every time period t_j monitors the risk of experiencing the event in the subsequent time period t_{j+1}. This is analogous to the way the survival function "leans in" to all subsequent time periods, with the difference that the hazard function monitors the risk of experiencing the event across the subsequent time period rather than all subsequent time periods. In the continuous case the hazard function is given by (4.6).

$$\lambda_t = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t \mid T \geq t)}{\Delta t} \tag{4.6}$$

The conditional statement in the numerator of (4.6) suggests a close link between the hazard function and the survival function because the condition that $T \geq t$ indicates that (4.6) monitors the risk of the outcome between t and $t + \Delta t$ among those who reach t event free. The probability that the event has not occurred by time x is equivalent to saying that the event will occur at some later date which is exactly what the survival function monitors. In the continuous case the relationship between the hazard, survival and probability density functions is given by (4.7).

$$\lambda_t = \frac{f(t)}{S(t)} = \frac{-dln(S(t))}{dt} \tag{4.7}$$

In the discrete case the relationship between the survival, hazard and probability mass function can be expressed more clearly using basic probability calculus. The discrete equivalent of (4.6) is given by (4.8)

$$\lambda_{t_j} = P(T = t_j \mid T \geq t_j) = \frac{P(T = t_j, T \geq t_j)}{P(T \geq t_j)} = \frac{P(T = t_j)}{P(T \geq t_j)} = \frac{p(t_j)}{S(t_{j-1})}, \tag{4.8}$$

where the first equality is the definition of the hazard function in the discrete case, the second equality holds by the definition of conditional probability and the third because the events $T = t_j$ and $T \geq t_j$ only coincide when $T = t_j$. The fourth inequality holds because $P(T \geq t) = S(t_{j-1})$ and in the discrete context the probability $P(T \geq t_j)$ is equivalent to the survival function at time t_{j-1} "leaning-in" to the subsequent

periods. For example, in a study that begins on Monday and ends on Sunday, the $P(T \geq Tuesday) = S(Monday)$.

Using (4.5), the last term of (4.8) can be rewritten to yield an expression for the hazard rate solely in terms of the survival function at t_j and t_{j-1}.

$$\lambda_{t_j} = 1 - \frac{S(t_j)}{S(t_{j-1})} \tag{4.9}$$

The survival function can be written as the product of conditional survival probabilities.

$$S(t) = \prod_{t \leq t_j} \frac{S(t_j)}{S(t_{j-1})} = \prod_{t \leq t_j} 1 - \lambda_{t_j} \tag{4.10}$$

The second equality in (4.10) follows from (4.9) and shows that the survival function is determined by the period specific hazard functions.

4.3 Proportional hazards

In a typical research setting interest often lies in how the survival function differs between subjects who have been exposed to some kind of treatment ($A = 1$) and those who have been left unexposed ($A = 0$). One approach is to construct survival functions for each group separately and contrast the survival functions between the two groups. This contrast will have a causal interpretation when the unexposed group can be viewed as analogous to the exposed group had they remained unexposed. In other words, when subjects in each group are exchangeable.

When confounding due to a confounder is present, adjusting for the effect of the confounder can lead to correct estimation of the causal effect of the exposure on the outcome. The Cox proportional hazards model for the hazard function allows for adjustment of confounders [34]. The central assumption in this model is that the effect of covariates is multiplicative with respect to a baseline hazard function. The basic model is

$$\lambda_t[Z] = \lambda_t^* c(\beta^{tr} Z). \tag{4.11}$$

Where $c(\cdot)$ is a function and $\beta^{tr} Z$ is a linear predictor of the covariates Z. We use λ_t^* to denote the baseline hazard at time t. The covariates Z in (4.11) act on the baseline hazard function λ_t^* in a multiplicative fashion. The Cox proportional hazard model can be extended to include time dependent variables as in (4.12).

$$\lambda_t[Z_t] = \lambda_t^* c(\beta^t Z_t) \tag{4.12}$$

Like the odds ratio, the hazard ratio (HR) is an effect measure. For a binary variable A the HR is defined as the ratio of the hazards in the exposed versus the unexposed groups. For example, if c is the exponential function with linear predictor $\alpha + \beta a$, then the HR between those exposed and those unexposed to A is given by (4.13).

$$\frac{\lambda_t[A=1]}{\lambda_t[A=0]} = \frac{\lambda_t^* exp(\beta_0 + \beta_1)}{\lambda_t^* exp(\beta_0)} = exp(\beta_1) \tag{4.13}$$

The assumption of proportional hazards is clear in (4.13) because the HR between exposed and unexposed subjects does not depend on elapsed time although the individual hazard functions may differ at each time point [35].

In discrete time the hazard function is approximately equal to the odds provided that the hazard is small (less than 0.1 at each point in time) [36][37]. As a result, the HR can be approximated by the odds ratio. We know from chapter 2 section 2.6 that we can express the odds ratio in a model formulation using the logit link function. As the HR is approximated by the odds ratio, it can be approximated by estimating a pooled regression of all subject-time observations. For example, the model in (4.14) corresponds to a HR of $exp(\beta_2)$.

$$logit\ \lambda_t[A_t = a_t] = \beta_0 + \beta_1 t + \beta_2 a_t \tag{4.14}$$

We saw in section 4.2 equation (4.10) that the survival function is completely determined by the hazard function. Estimating a model like (4.14) by logistic regression gives us the hazard function at each point of time. This can then be used to evaluate the survival function as in (4.10).

4.4 Marginal structural survival models.

We end this chapter by describing marginal structural survival models. Chapters 2 and 3 explained why MSMs are necessary in a time-dependent context. Essentially, in a time-dependent context conditioning on a confounder L_t introduces bias through a spurious dependency between Y and A but failing to condition on L permits confounding due to L being a common cause of A and Y. In a survival context the in-built selection bias of the HR is also of concern [25].

Interest now lies in estimating structural survival functions. We denote the structural hazard function by $\lambda_t^{\bar{a}}$ and the structural survival function by $S_{\bar{a}}(t)$. A model like (4.15) is a marginal structural model for the hazard function. It is marginal over any confounders and β_2 represents the causal effect of A on survival Y.

$$logit\ \lambda_t^{\bar{a}} = \beta_0 + \beta_1 t + \beta_2 a_t \tag{4.15}$$

Conceptually estimation of (4.15) by the IPTW-estimator is no different than what has already been described in chapter 3 section 3.3.3. What is different is that (4.15) is a subject-time specific model, as opposed to the subject specific models for longitudinal data that we have considered so far. This has consequences for the weights which are also subject-time specific. The relevant unstabilized weights in a survival setting are the subject-time specific weights in the product of the denominator of (4.16).

$$w_{i,t} = \frac{1}{\prod_{\tau=0}^{t} p_\tau(A_{\tau,i} \mid \bar{A}_{\tau-1,i}, \bar{L}_{\tau,i})} \tag{4.16}$$

The corresponding stabilized weights are given by,

$$sw_{i,t} = \frac{\prod_{\tau=0}^{t} p_\tau(A_{\tau,i} \mid \bar{A}_{\tau-1,i})}{\prod_{\tau=0}^{t} p_\tau(A_{\tau,i} \mid \bar{A}_{\tau-1,i}, \bar{L}_{\tau,i})}. \tag{4.17}$$

Estimation of (4.15) involves estimating a pooled logistic model for the hazard function weighted using the stabilized weights (4.17). As explained in section 4.2, the hazard function completely determines the survival function. Constructing the structural survival curve, therefore, involves first estimating $\lambda_t^{\bar{a}_t}$ and then evaluating (4.18) at all points in time.

$$S_{\bar{a}}(t) = \prod_{\tau=0}^{t} 1 - \lambda_t^{\bar{a}} \tag{4.18}$$

Chapter 5

Simulating from marginal structural models

In this chapter we start by describing in general terms the logic behind Monte Carlo simulations. Next, we consider how simulating data from a MSM deviates from the typical set-up of a Monte Carlo simulation. Simulating data from a specific MSM is a very challenging task because many of the effect measures used in a time-fixed and a time-dependent context are non-collapsible. This includes the odds ratio and the hazard ratio. In this chapter we explain non-collapsibility and why it creates a problem. Fundamentally, the problem is that there is a mismatch between the conditional probabilities that are typically used to induce confounding in simulated data, and a specific MSM which is, by definition, marginal over any confounders. We describe the algorithm of Bryan et al (2004) and Havercroft and Didelez (2012) which provides a solution to this problem. An added value of this thesis is to describe this solution in detail and provide some intuition for why it works.

5.1 Monte Carlo simulations

In statistical research, interest often lies in the estimation of a population parameter θ. When only one sample X_1 from the population is available, statistical methods are applied to obtain an estimate $\hat{\theta}_1$ of the population parameter based on that sample. If a second sample X_2 drawn from the same population was available, applying the same statistical method would result in a second estimate $\hat{\theta}_2$, and so on for more samples. In reality, only one sample is typically available and we rely on the sampling distribution of the $\hat{\theta}_i$ to draw statistical inferences about the population parameter θ.

Monte Carlo simulations flip this process on its head. We start with a known truth θ and simulate data according to that truth. The aim of a simulation study is to assess statistical methods in relation to that known truth. Statistical research is a process through which we discover information about an unknown truth. Simulation studies, on the other hand, are processes through which we re-discover, or attempt to re-discover, information about a known truth in order to assess the appropriateness of a certain statistical method under the scenarios considered.

For example, (5.1) relates the conditional mean of Y to the binary variable A in a linear logistic fashion. In a simulation context the parameters $\theta = (\alpha, \beta)$ are the known truth. The binary variable A could be generated by a random number generator such as the *rbernoulli* function in the *R statistical language* and Y is evaluated from θ and A. The ability to generate random numbers is fundamental to Monte Carlo simulation. For a sample of size n, simulating data according to (5.1) would yield the pair (Y_n, A_n) on which the performance of a statistical method can be tested.

$$logit(P(Y = 1 \mid A = a)) = \alpha + \beta a \tag{5.1}$$

In a single simulation a finite sample of size n is drawn to generate data (Y_n, X_n). Applying a statistical method, such as logistic regression, to (Y_n, X_n) to recover θ gives an indication of the performance of logistic regression by comparing the θ obtained from the simulation with the true values used to generate the data.

In a real sample of data drawn from a medical trial, say, there will be sample variability. Drawing a different sample from the same population will yield a different estimate than the first sample because of that sample variability. The same is true in data drawn from one simulation and generated by a random number generator. For that reason, a simulation trial typically includes more than one simulation, say $B = 100$ or $B = 1000$ simulations. The results from each trial are combined in order to remove the effect of sample variability. For example by obtaining the average $\bar{\beta}$ from (5.1) across all simulations and comparing this to the true value.

The steps in a simulation study can be summarised as follows [38].

- Step 1: Specify the artificial population
- Step 2: Simulate a sample of data according to the artificial population of size n
- Step 3: Calculate the parameter of interest from the simulated data
- Step 4: Repeat steps 2 and 3 B times, where B is the number of trials
- Step 5: Evaluate the performance of the method based on some summary of B trials.

5.1.1 Performance measures in simulation studies

In this section we describe measures used to assess the performance of statistical methods in simulation studies [39]. Typically, statistical methods in simulation studies are assessed in terms of both bias and variability. Methods which result in a small bias but large variability have limited applicability in the real world because of the uncertainty associated with their true value. On the other hand, methods with large bias an low variability are potentially misleading. For these reasons, a set of performance measures which reflect the bias variance trade-off are preferred. In this thesis we use the bias, standard error and root mean squared error (MSE) as measures of the bias and variance of the IPTW-estimator when violations of the positivity assumption are present. In this section we use β from (5.1) as an example.

In each trial B simulations take place. Combined across all simulations gives the average estimate of interest

$$\bar{\hat{\beta}} = \frac{1}{B}\sum_{i=1}^{B}\hat{\beta}_i \tag{5.2}$$

And the empirical SE defined by (5.3)

$$SE(\hat{\beta}) = \sqrt{1/(B-1)\sum_{i=1}^{B}(\hat{\beta}_i - \bar{\hat{\beta}})^2} \tag{5.3}$$

The bias, defined in (5.4), evaluates the performance of a statistical method on its ability to resolve the correct parameter β from which the data was simulated. The closer the bias to zero the less biased the statistical method.

$$bias(\hat{\beta}) = \bar{\hat{\beta}} - \beta \tag{5.4}$$

The mean square error evaluates the accuracy of the statistical method in one measure which combines both the bias and variability. The MSE is always positive and smaller the MSE the accurate the method.

$$MSE(\hat{\beta}) = (\bar{\hat{\beta}} - \beta)^2 + (SE(\hat{\beta}))^2 \tag{5.5}$$

5.2 Simulating from MSMs

In section 5.1 we described the general set-up for Monte Carlo simulation. Simulating from MSMs in a time-dependent survival context is more complicated for two reasons. First, in a time-dependent setting we need data which exhibits time-dependent confounding as described in section 2.4. Because we consider a survival setting, this should include selection bias because effect measures like the hazard ratio have an in-built selection bias. More challenging, MSMs are marginal over any confounders. This poses a problem when considering MSMs for non-collapsible effect measures like the hazard ratio or the odds ratio. The crucial point is that there is a mismatch between the MSM from which we wish to simulate data, and the conditional probabilities that are used to introduce confounding through an association between Y and L. In the remainder of this chapter we first describe non-collapsibility in more detail and then explain how it relates to simulating data which exhibits time-dependent confounding from a specific MSM. We then present an existing algorithm which can be used to generate data with the structure we need. Few studies have investigated violations of the IPTW-estimator assumptions in a time-dependent context. Part of the reason is the challenge of implementing and understanding. An added value of this thesis is the extra detail and intuition provided to understand this process.

5.2.1 Noncollapsibility of survival effect measures

Chapter 2 section 2.2 described how confounding arises when a confounder L is a common cause of both A and Y. As a result, the marginal, or unadjusted, effect of A on Y is different from the strata specific effects. On the other hand, when L is unrelated to A the effect of A on Y is the same across all strata defined by L. In that case, marginalizing across L can still lead to the correct causal effect.

Non-collapsibility is a feature of certain effect measures whereby conditioning on a covariate L which is related to Y changes the size of the effect measure, even when A and L are not related [27]. At first glance this seems paradoxical [40]. For example, if the effect of a drug A on outcome Y is equal across male and female subjects, we would expect the effect in the total population to equal the strata specific effects because the total population is the aggregate of all males and females. However, when the effect is measured by the odds ratio this may not be the case because the odds ratio is a non-collapsible even when no confounding is present. The same is true of the hazard ratio.

Collapsibility, on the other hand, occurs when collapsing (marginalizing) over a covariate yields the same effect estimate under no confounding. Collapsibility is a useful property because correct analysis can be performed in the marginal population reducing the dimensionality and computational effort which arises when many subgroups need to be taken account of in the analysis [22]. When confounding is present but the effect measure is non-collapsible, the degree to which non-collapsibility distorts the true causal effects can be evaluated using the IPTW estimator because the IPTW-estimator estimates causal effects without conditioning on confounders [41].

Chapter 2 section 2.6 explained that the causal odds ratio could be expressed in a model formulation by relating the odds to a linear predictor on a logistic scale. Non-collapsibility can also be explained by reference to model formulation [13]. For example, if we are interested in the causal effect of A on Y as measured by the odds ratio we can specify a MSM such as (5.6). The aim in the simulation study is to simulate data which obeys a MSM such as (5.6) and still exhibits confounding. This applies in both the time-fixed context and in the time-dependent context.

$$logit\ P(Y_a) = \gamma_0 + \gamma_1 a \tag{5.6}$$

One method is to first generate a confounder L and then assign A according to $P(A \mid L = l)$ and finally assign Y according to $P(Y \mid A = a, L = l)$. An example of a model for $P(Y \mid A = a, L = l)$ is

$$logit\ P(Y \mid A = a, L = l) = \gamma_0^* + \gamma_1^* a + \gamma_2^* l. \tag{5.7}$$

The problem is that marginalizing (5.7) over L will not result in the MSM (5.6) because (5.7) is non-collapsible. Simulating data that obeys (5.7) and then applying the IPTW-estimator to estimate the parameters of (5.6) will not return the parameters (γ_0^*, γ_1^*) even when all IPTW assumptions are met. The implication for this thesis is that assessing the performance of the IPTW-estimator under violations of the positivity assumption will not be possible because we are unable to disentangle the effect of non-collapsibility from the effect of positivity violations on the true parameter used to generate the data.

5.3 Simulation algorithm

In this section we describe the algorithm that we will implement in order to study the effect of positivity violations on the performance of the IPTW-estimator. Several algorithms have been developed for simulating data from a MSM [42][43][14]. However few of these tackle the problem of non-collapsibility and are therefore unsuitable for many effect measures in a survival setting or indeed for non-collapsible effect measures like the odds ratio. Second, the observational structure in many existing algorithms does not exhibit selection bias which is also important in a survival context due to the in built selection bias of the hazard ratio. The algorithm developed first in Bryan et al (2004) and subsequently extended upon in Havercroft and Didelez (2012) does include these important elements. First, the algorithm simulates from a known survival MSM with the characteristics of observational data outlined in section 2.4 and section 5.2. In other words, the algorithm overcomes the issue of non-collapsibility whilst maintaining a confounder structure. Second, the pathway $L \rightarrow A$ in the algorithm is under the control of the user, a point on which the algorithm of Bryan (2004) and Havercroft and Didelez (2012) depart. It is therefore possible to compare the parameters of a known MSM with the parameters recovered from data in which positivity violations have been introduced by the user and these violations are propagated through time. In this section we describe the most important aspects of this algorithm and why it is suitable for use in this thesis.

Figure 5.1 shows the DAG from which data is simulated in Havercroft and Didelez (2012). The solution to non-collapsibility in Bryan (2004) and Havercroft, Didelez (2012) is to introduce a fourth variable U as in figure 5.1. No direct link between L and Y exists but U is a common cause of L and Y. Factorising the joint density represented graphically by the DAG in figure 5.1 yields (5.8).

$$\begin{aligned} P(Y, A, L, U) &= P(Y \mid A, L, U)P(A, L, U) \\ &= P(Y \mid A, L, U)P(A \mid L, U)P(L, U) \\ &= P(Y \mid A, L, U)P(A \mid L, U)P(L \mid U)P(U) \end{aligned} \tag{5.8}$$

The independencies encoded in the DAG of figure 5.1 (right) imply (5.9) and (5.10)

$$P(Y \mid A, L, U) = P(Y \mid A, U) \tag{5.9}$$

$$P(A \mid L, U) = P(A \mid L), \tag{5.10}$$

so that (5.8) can be re-written as (5.11).

$$P(Y, A, L, U) = P(Y \mid A, U)P(A \mid L)P(L \mid U)P(U) \tag{5.11}$$

MSMs are models for causal effects as opposed to associational effects. Replacing Y in the left hand side of (5.9) with Y_a, the counterfactual outcome under exposure A, can be achieved by replacing the conditional distributions of A with $\mathbb{I}_{A=a}$, the indicator function which is equal to 1 if $A = a$ and zero otherwise. The intuition here is to think of making an intervention to set $A = a$. For example, when A has two levels

$A = \{0, 1\}$, setting $A = 1$ means that the $P(A = 1 \mid L) = 1$. Or, in other words, $A = 1$ regardless of the value of L. This can be expressed compactly for both levels of A using the indicator function in (5.12).

$$\mathbb{I}_a = \begin{cases} 1 & \text{if } A = a \\ 0 & \text{if } A \neq a \end{cases} \tag{5.12}$$

Setting $A = a$ in the remaining conditional distributions of (5.11) yields (5.13).

$$P(Y_a, L, U) = P(Y \mid A = a, U)P(L \mid U)P(U)\mathbb{I}_{A=a} \tag{5.13}$$

Averaging (5.13) over all values of either L or U yields an expression for the mean of the counterfactual distribution $P(Y_a)$ in terms of A and L or U, shown in (5.14).

$$\begin{aligned} P(Y_a) &= \sum_{l,u} P(Y \mid A = a, U)P(L \mid U)P(U) \\ &= \sum_{l} P(Y \mid A = a, L)P(L) \\ &= \sum_{u} P(Y \mid A = a, U)P(U) \end{aligned} \tag{5.14}$$

Both the second and third equalities of (5.14) hold because $P(Y_a, L, U)$ can be factorised in more than one way, as shown in (5.15).

$$P(Y_a, L, U) = P(Y_a \mid L, U)P(L \mid U)P(U) = P(Y_a \mid L, U)P(U \mid L)P(L), \tag{5.15}$$

This suggests two possibilities for simulating from a given MSM. We could specify a distribution for L, $P(L)$ and use this to generate data according to $P(Y \mid A = a, L)P(L)$ or we could specify a distribution for U, $P(U)$ and use this to generate data according to $P(Y \mid A = a, U)P(U)$. The problem is that to obtain the marginal structural model on the left hand side of (5.14) we need to marginalize (collapse) over either L or U. However, survival effect measures, like the hazard ratio or its discrete time equivalent the odds ratio, are typically non-collapsible. If we specify a model for $P(Y \mid A = a, L)$ such as (5.16), then when we come to average over L we will get α_1^* rather than the causal effect we specified which was α_1

$$logit\ P(Y \mid A = a, L) = \alpha_0 + \alpha_1 a + \alpha_2 l \tag{5.16}$$

$$logit\ P(Y \mid A = a) = \alpha_0^* + \alpha_1^* a \tag{5.17}$$

The same would be true if we use U instead of L. The issue is that there is a mismatch between the marginal model which we want to simulate from, and the conditional model we need to satisfy (5.14).

5.4 Overcoming non-collapsibility

One solution to simulating from a MSM for a non-collapsible effect measure was proposed in Bryan et al (2004) and implemented in both Bryan et al (2004) and Havercroft and Didelez (2012). Implementing and understanding the full algorithm is challenging because the observational structure of time-dependent data is very complex and models for non-collapsible effect measures require an approach that does not rely on assigning the outcome Y as a function of the confounder L. We begin by describing the inverse probability

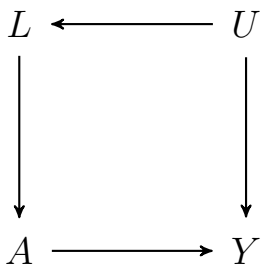


Figure 5.1: Directed Acyclic Graphs (DAGs).

transform method which is crucial for understanding the algorithm. Next, we consider a time-fixed version of the algorithm which is simpler than, but carries much of the intuition behind, the full algorithm for generating time-dependent data. Although alluded to, the time-fixed case is not covered in depth in either of the papers we follow. Finally we consider the time-dependent survival case of the algorithm. We use this thesis as an opportunity to provide an extensive discussion of how the algorithm works which will be of use to guide future researchers.

5.4.1 Inverse probability integral transform method

The inverse probability integral transform method for sampling random numbers relies on the fact that if X is a continuous random variable with CDF $F_X(x)$, then

$$U = F_X(x) \sim Uniform(0, 1). \tag{5.18}$$

Consequently, a continuous or discrete random variable from **any** distribution can be generated by first generating a random uniform variable U and then transforming it according to (5.19). The details underlying the inverse probability transform method are widely known and can be found in many standard references [44].

$$x = F_X^{-1}(u) \tag{5.19}$$

Figure 5.2 gives two examples to illustrate the inverse probability transform method. Figure 5.2 (left) shows the case for a discrete Bernoulli random variable. In the discrete case, a value of x_i is assigned when

$$F_X(x_{i-1}) < U \leq F_X(x_i). \tag{5.20}$$

Which in the case of a Bernoulli random variable in figure 5.2 (right) is equivalent to the statement

$$X = \begin{cases} 1 & \text{if } F_X(x=0) < U \leq F_X(x=1) \\ 0 & \text{if } U \leq F_X(x=0) \end{cases} \tag{5.21}$$

In figure 5.2 (right) the case of a continuous variable with an exponential distribution is shown. In the case of an exponential variable a value x can be drawn by

$$x = F_X^{-1}(u) = -\frac{1}{\lambda} log(1-u), \tag{5.22}$$

Where the second equality follows from the inverse of the CDF of an exponential distribution. If figure 5.2 (right) represents a distribution of survival times then the process for generating survival times works as

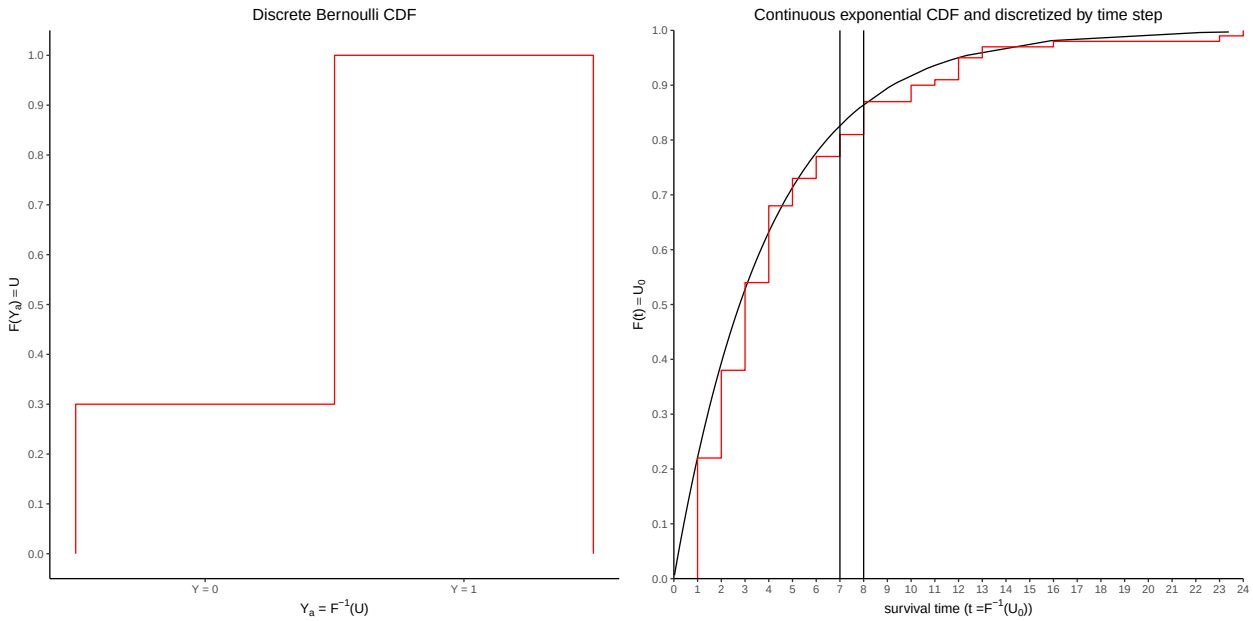


Figure 5.2: Examples of the inverse probability integral transform method. Discrete Bernoulli (left) and continuous and discrete exponential (right).

follows. First randomly generate a value u from a uniform distribution on the interval zero to one. Next, transform that value using (5.22) to obtain x, the subjects survival time. Figure 5.2 (right) also indicates how a discretized version of a continuous random variable can be drawn. For example, if survival times are distributed according to an exponential distribution but subjects are only measured at discrete intervals. Figure 5.2 (right) shows a discrete time step function analogue to the CDF of the exponential distribution. Drawing values from the discrete CDF works by (5.20). First a uniform random variable is drawn. Next, we check which discrete time interval u falls in using (5.20) and assign a survival time of x_i. Provided that we know the CDF represented in figure 5.2 (right), we can draw survival times for any subject. Repeated many times, this process results in a sample of discrete survival times from the chosen distribution.

5.4.2 The time-fixed case

Now we describe a time-fixed or point treatment version of the algorithm for non-collapsible effect measures like the odds ratio. In this thesis we do not actually carry out any simulations in a time-fixed context nor is the time-fixed case explicitly explained in either Bryan et al (2004) or Havercroft and Didelez (2012). The reason for including it here is that the time-fixed context is useful for highlighting several important aspects of the algorithm before moving on to the more complicated time-dependent survival context.

The most natural procedure would be to generate a confounder L and then assign exposure according to $P(A \mid L)$. Finally Y is assigned according to $P(Y \mid A, L)$. This structure exhibits confounding because L is a common cause of both A and Y. Unfortunately, we know from section 5.4.1 that the odds ratio is non-collapsible. We can specify parameters $(\gamma_0^*, \gamma_1^*, \gamma_2^*)$ and assign Y using (5.23). However, due to non-collapsibility of the odds ratio, the simulated data will not correspond to the MSM in (5.24), as is our aim. Under non-collapsibility $(\gamma_0^*, \gamma_1^*) \neq (\gamma_0, \gamma_1)$. Therefore, applying the IPTW-estimator to data generated by (5.23), will not resolve the parameters (γ_0, γ_1) even when all the IPTW assumptions are met.

$$logit\ P(Y\ |A = a, L) = \gamma_0 * + \gamma_1 * a + \gamma_2 * l \tag{5.23}$$

$$logit\ P(Y_a) = \gamma_0 + \gamma_1 a \tag{5.24}$$

Instead, we take a different approach. We start by drawing a random uniform variable on the interval between zero and one as in (5.25).

$$U \sim unif(0,1) \tag{5.25}$$

Next, a confounder L is assigned by using the inverse probability transform method according to the CDF of a suitable distribution. A suitable distribution for the confounder in the algorithm of Havercroft and Didelez (2012) is the gamma distribution with shape parameter $k = 3$ and scale parameter $\theta = 154$. This distribution is suitable because it produces data which resembles the CD4 count of subjects in a realistic HIV cohort study [45]. Obtaining a value of L can then be achieved by (5.26) in the same way as the example of an exponential distribution in section 5.4.1 equation (5.22).

$$l = F_L^{-1}(u) = \frac{1}{\Gamma(k)\theta^k}\gamma\Big(k, \frac{u}{\theta}\Big) \tag{5.26}$$

In (5.26), $\Gamma()$ is the gamma function and $\gamma()$ is the lower incomplete gamma function. Next, a value of A is assigned according to $P(A \mid L)$ by (5.27) and (5.28).

$$logit(P(A \mid L)) = \theta_0 + \theta_1 L \tag{5.27}$$

$$A \sim Bernoulli(p = logit^{-1}(\theta_0 + \theta_1 L)) \tag{5.28}$$

In words, the conditional probability of receiving exposure $P(A \mid L)$ depends on L in a linear logistic fashion. Assignment of A is achieved by drawing A from a Bernoulli distribution with parameter p equal to the conditional probability $P(A \mid L)$. The result is that A depends on L.

For L to be a confounder it needs to be a common cause of A and Y. In other words, we need to introduce some dependence between L and Y without resorting to a model for the conditional probability of Y given L such as (5.23). The solution is to apply the inverse probability transform method for a second time to generate Y. Crucially, we use the same value of U to generate L and Y. As a result L and Y are associated through their common cause U. We now describe how this works in practice. We draw $A = a$ as described in (5.27) and (5.28). Next we use the value of A drawn in (5.28) and the parameters of the chosen MSM (γ_0, γ_1) to evaluate $P(Y_a)$ from our chosen MSM (5.24). In a time-fixed setting we can think of Y as a Bernoulli random variable with a CDF like 5.2 (left). Suppose that after assigning A we find $P(Y_a = 1) = 0.7$. The CDF of Y can be specified by $F(0) = 0.3$ and $F(1) = 1$. Using this CDF and the inverse transform method, we can assign a value of Y according to (5.29) in an analogous manner to (5.21) in section 5.4.1.

$$Y = \begin{cases} 1 & \text{if } F_{Y_a}(0) < U \leq F_{Y_a}(1) \\ 0 & \text{if } U \leq F_{Y_a}(0) \end{cases} = \begin{cases} 1 & \text{if } 0.3 < U \leq 1 \\ 0 & \text{if } U \leq 0.3 \end{cases} \tag{5.29}$$

The important point is that Y and L are associated through their common cause U, but non-collapsibility does not arise because that association does not involve conditioning on L. Moreover, conditioning on L is sufficient to block the pathway from U to A in the DAG in figure 5.1. This is important when it comes to specifying the weight model because, as we described in section 3.2.1, IPTW estimators are only valid when subjects within strata defined by L are exchangeable. This would not be the case if the distribution of U differed between the exposed and unexposed subjects within the strata defined by L. The double use of the inverse probability transform method is the kernel of the algorithm and in the following section we describe how it extends to the time-dependent survival context.

5.4.3 The time-dependent survival case

In the survival context the solution to non-collapsibility is the same. Our aim is to simulate survival times which obey a specific MSM for the hazard function. Models for survival effect measures, like the hazard ratio or its discrete time equivalent the odds ratio, are non-collapsible. The solution proposed in Bryan et al (2004) is to exploit the relationship between the CDF, survival function and hazard function. First, we specify a chosen MSM for the hazard function. For example,(5.30) which relates the probability of experiencing the event in the next time interval to exposure at time t and the current time period.

$$logit\ \lambda_t^{\bar{a}} = \gamma_0 + \gamma_1 a_t + \gamma_2 t \tag{5.30}$$

Next, we generate a random uniform variable for each subject as in (5.31).

$$U_0 \sim Uniform(0,1) \tag{5.31}$$

This value is then mapped to the MSM in (5.30) through (5.32).

$$U_0 = F_{\bar{a}}(t) = 1 - S_{\bar{a}}(t) = 1 - \prod_{\tau=0}^{t}(1-\lambda_\tau^{\bar{a}}) \tag{5.32}$$

The first equality in (5.32) follows from the inverse probability transform method. The second equality follows from the relationship between the CDF and the survival function. The final equality follows because the survival function is completely determined by the hazard function, as explained in chapter 4 section 4.2.

We are considering a discrete time event similar to the step function in figure 5.2 (right). We assign a survival time in an analogous manner to (5.20) in section 5.4.1.

$$F_{\bar{a}}(t-1) < U_0 \leq F_{\bar{a}}(t) \tag{5.33}$$

Using (5.32) we can rewrite (5.33) to get (5.34).

$$1 - \prod_{\tau=0}^{t-1}(1-\lambda_\tau^{\bar{a}}) < U_0 \leq 1 - \prod_{\tau=0}^{t}(1-\lambda_\tau^{\bar{a}}) \tag{5.34}$$

Procedurally, as soon as $1-\prod_{\tau=0}^{t}(1-\lambda_\tau^{\bar{a}}) \geq U_0$ the subject experiences the event and we obtain their survival time. In this way we generate survival times which obey a marginal structural model. As in the time fixed case the same value of U_0 is used to generate an initial value of the confounder L_0 according to (5.35). The survival time and the confounder L are therefore related through their common ancestor U_0.

$$l_0 = F_{L_0}^{-1}(u) = \frac{1}{\Gamma(k)\theta^k}\gamma\Big(k, \frac{u_0}{\theta}\Big) \tag{5.35}$$

Creating a dependency between the survival time and L in this way avoids specifying a conditional model for the hazard function λ_t. We have focused on the mechanics and details of the method which are somewhat scarce in the papers of Bryan et al (2004) and Havercroft and Didelez (2012). A full proof that this procedure does simulate survival times that obey a specific MSM can be found in Havercroft and Didelez (2012) Appendix B. Here we simply link what has been discussed so far to their result that the survival times obtained from (5.34) do obey the marginal structural model for the hazard function $\lambda_t^{\bar{a}}$. The event that a subject survives the first t time periods and fails in the interval $t+1$ is the event $\{Y_{t+1} = 1, \bar{Y}_t = 0\}$. This event is equivalent to (5.36).

$$\{Y_{t+1}, \bar{Y}_t\} \Leftrightarrow \Big\{1 - \prod_{\tau=0}^{t-1}(1-\lambda_\tau^{\bar{a}}) < U_0 \leq 1 - \prod_{\tau=0}^{t}(1-\lambda_\tau^{\bar{a}})\Big\} \tag{5.36}$$

The probability that this event occurs is given by (5.37).

$$\begin{aligned} & P\Big(U_0 \leq 1 - \prod_{\tau=0}^{t}(1-\lambda_\tau^{\bar{a}})\Big) - P\Big(U_0 \leq 1 - \prod_{\tau=0}^{t-1}(1-\lambda_\tau^{\bar{a}})\Big) \\ & = 1 - \prod_{\tau=0}^{t}(1-\lambda_\tau^{\bar{a}}) - \Big(1 - \prod_{\tau=0}^{t-1}(1-\lambda_\tau^{\bar{a}})\Big) \\ & = \lambda_t^{\bar{a}} \prod_{\tau=0}^{t-1} 1 - \lambda_\tau^{\bar{a}} \end{aligned} \tag{5.37}$$

The final line of (5.37) is the product of the hazard function $\lambda_t^{\bar{a}}$ evaluated at time t and the survival function $S(t-1) = \prod_{\tau=0}^{t-1} 1 - \lambda_\tau^{\bar{a}}$. In other words, the last line of (5.37) is the probability of the event that the subject survives the first $t-1$ intervals ($S_{\bar{a}}(t-1) = P(T > t-1)$) and experiences the event in the t th interval conditional on surviving up to the t th interval $\lambda_t^{\bar{a}} = P(T = t \mid T \geq t)$. What (5.37) shows is that survival times generated in the manner described have the correct probability distribution.

The remaining conditional dependencies in figure 5.3 are relatively straightforward and are explained alongside the full algorithm presented in section 5.4.4. Fundamentally what distinguishes this algorithm from competing algorithms is that it avoids a mismatch between the MSM that we wish to draw data from, and the conditional probabilities which are normally used to simulate data. This is what we have described as the kernel of the algorithm and the solution is the double use of the inverse probability method to generate survival times and confounder values.

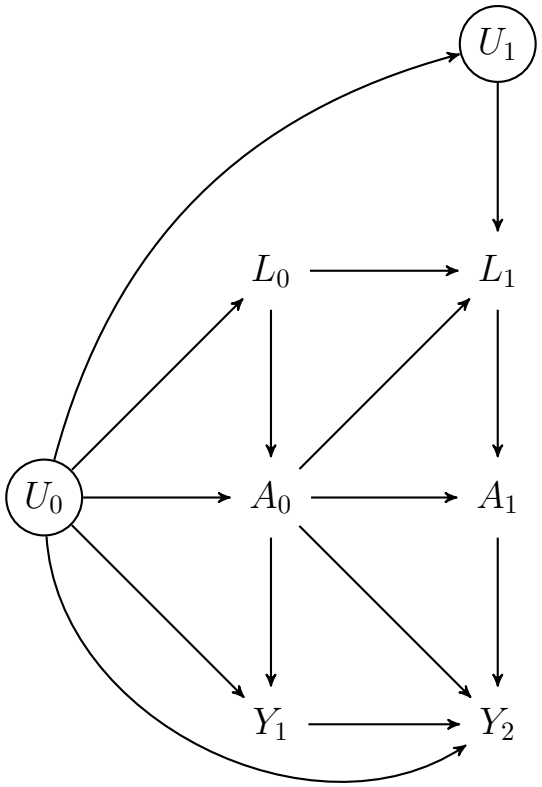


Figure 5.3: DAG for time-dependent data generating base algorithm from Havercroft and Didelez (2012)

5.4.4 Base Algorithm

To conclude this chapter we present the full algorithm from Havercroft and Didelez (2012) Appendix A and walk through its main points. Lines 3 - 14 in Algorithm 1 represent the baseline values for a single subject. A random uniform variable is generated and the inverse probability transform method is used to generate a baseline confounder value L_0 from a gamma distribution with shape parameter $k = 3$ and scale parameter $\theta = 154$. A little random noise is added by ϵ_0 so that L_0 is not completely determined by U_0.

It is assumed on line 6 that no subject was exposed before the start of follow up. Exposure is assigned according to (5.38) and (5.39). This is essentially the same as (5.27) and (5.28) in the time-fixed case of section 5.4.2. The difference is that exposure at time t also depends on time ($t = 0$ at baseline on line 7). Exposure depends on L_t and, because the idea is to mimic an HIV cohort study, a value of 500 is subtracted from L_t where 500 represents a conservative lower threshold for a healthy CD4 count [45]. The parameter $\theta_2 < 0$ so When L_t falls below 500 the subject is more likely to be assigned exposure than non exposure.

$$logit(P(A_t \mid \bar{L}_t)) = \theta_0 + \theta_1 t + \theta_2(L_t - 500) \tag{5.38}$$

$$A_t \sim Bern(logit^{-1}(\theta_0 + \theta_1 t + \theta_2(L_t - 500)) \tag{5.39}$$

The baseline hazard is fitted on line 10 according to (5.40) (where $t = 0$ at baseline) and if this hazard is greater than or equal to U_0 the subject fails for exactly the reason explained in section 5.4.3 (in the baseline case $1 - \prod_{\tau=0}^{t=0}(1 - \lambda_{\tau,i}) = \lambda_{\tau,i}$).

$$\lambda_t^{\bar{a}_t} = logit^{-1}(\gamma_0 + \gamma_1((1-a_t)t + a_t t^*) + \gamma_3 a_t(t-t^*)) \tag{5.40}$$

Lines 15-35 essentially repeat the baseline case at subsequent time points while the subject is still alive. A new value of U_t is generated for each time interval on line 8. The role of U_t after baseline should not be confused with that of U_0 which is used for the kernel of the algorithm to create dependence between L and Y. Subsequent values of U_t do create selection bias, however, because, shown on lines 23 and 24, L_t and U is associated with Y at each time period through its common ancestor U_0. Conditioning on L creates selection bias by inducing a spurious relationship between U_t and A and hence between A_t and Y. The IPTW-estimator corrects for selection bias but we need it here because as soon as we violate the positivity assumption the IPTW-estimator is no longer valid. We need a realistic time-dependent survival setting and because of the in-built selection bias in the hazard ratio we only achieve this if selection bias is present in the data generating algorithm.

Line 19 controls the flow of the algorithm so that changes to exposure and confounder values are only updated at measuring times every k time intervals. The idea is that the state of the subject (alive or dead) is measured at every point in time. The exposure and confounder values between measuring times are assumed to be the same (lines 20-21). This brings us to the last point about the algorithm. At measuring times L_t is a function of previous exposure (line 24) and is a predictor of subsequent exposure (line 26). In other words the issue of time-dependent confounding explained in section 2.4.1 is present in the algorithm. One a subject starts exposure lines 25-28, they stay on exposure for the remainder of the study until death or end

of follow-up.

Algorithm 1: Base algorithm from Havercroft and Didelez (2012) Appendix A

Result: Simulation algorithm

initialization;
for *i in 1, . . . , n* **do**
  $U_{0,i} \sim U[0,1]$
  $\epsilon_{0,i} \sim N(0,20)$
  $L_{0,i} \leftarrow F^{-1}_{\Gamma(3,154)}(U_{i,0}) + \epsilon_{0,i}$
  $A_{-1,i} \leftarrow 0$
  $A_{0,i} \leftarrow Bern(logit^{-1}(\theta_0 + \theta_2(L_{0,i} - 500)))$
  if $A_{0,i} = 1$ **then**
    $T^* \leftarrow 0$;
  $\lambda_{0,i} \leftarrow logit^{-1}(\gamma_0 + \gamma_2 A_{0,i})$
  if $\lambda_{0,i} \geq U_{0,i}$ **then**
    $Y_{1,i} \leftarrow 0$
  else
    $Y_{1,i} \leftarrow 1$
  for *t in 1, . . . , T* **do**
    while $Y_{t,i} = 0$ **do**
      $\Delta_{t,i} \sim N(0, 0.05)$
      $U_{t,i} \leftarrow min\{1, max\{0, U_{t-1,i} + \Delta_{t,i}\}\}$
      if $t \neq 0 (mod\ k)$ **then**
        $L_{t,i} \leftarrow L_{t-1,i}$
        $A_{t,i} \leftarrow A_{t-1,i}$
      else
        $\epsilon_{t,i} \sim N(100(U_{t,i} - 2), 50)$
        $L_{t,i} \leftarrow max\{0, L_{t-1,i} + 150 A_{t-k,i}(1 - A_{t-k-1,i}) + \epsilon_{t,i}\}$
        if $A_{t-1,i} = 0$ **then**
          $A_{t,i} \sim Bernoulli(logit^{-1}(\theta_0 + \theta_1 t + \theta_2(L_{t,i} - 500)))$
        else
          $A_{t,i} \leftarrow 1$
        if $A_{t,i} = 1$ *and* $A_{t-k,i} = 0$ **then**
          $T^* \leftarrow t$
      $\lambda_{t,i} \leftarrow logit^{-1}(\gamma_0 + \gamma_1[(1 - A_{t,i})t + A_{t,i}T^*] + \gamma_2 A_{t,i} + \gamma_3 A_{t,i}(t - T^*))$
      if $1 - \prod_{\tau=0}^{t}(1 - \lambda_{\tau,i}) \geq U_{0,i}$ **then**
        $Y_{t+1,i} = 1$
      else
        $Y_{t+1,i} = 0$

Chapter 6

Violations of positivity

The IPTW estimator for causal effects is valid under the five assumptions of exchangeability, consistency, positivity, no model misspecification and no measurement error. The positivity assumption was introduced in chapter 1 section 2.2.4 as one of the three necessary conditions underlying causal inference. The relevance of the positivity assumption to IPTW estimation of MSMs was explained in chapter 3 section 3.3.3. MSMs are models for causal effects in a study population. It must be possible to estimate the average causal effect in every subset of the population defined by any confounders in the analysis. This will only be possible when the positivity assumption is met [46]. When positivity violations are present they can lead to bias and high variability in the resulting IPTW-estimators. Despite its importance, the positivity assumption has received relatively little attention in the literature [2][1]. To our knowledge no systematic simulation study exists which explores the effect of positivity violations in a realistic longitudinal or survival setting. This is surprising because a major advantage of MSMs is that they can be used to adjust for time-dependent confounding.

In this chapter we begin by explaining the positivity assumption in more detail. We distinguish between structural and random violations of the positivity assumption and give examples of both. We also describe near violations of the positivity assumption which have similar impacts on the IPTW-estimators. We then explain how we extend the base algorithm introduced in chapter 5 to allow positivity violations. Finally, we outline the four scenarios which will be investigated in a simulation study in chapter 7.

6.1 Positivity: definition and relevance to IPTW-estimation.

Positivity is the condition that there is a non-zero or positive probability that a subject is exposed to each level of a treatment (A) within each strata defined by the confounders (L) in the analysis. Along with consistency and conditional exchangeability (see chapter 2 section 2.2 for more details), the positivity condition is one of the three fundamental conditions behind causal inference. Equation (6.1) expresses the positivity condition formally

$$P(A \mid L) > 0. \tag{6.1}$$

Since values of A and L are not specified in (6.1), this condition holds for all values of A and L. In the counterfactual framework the positivity assumption has an intuitive meaning. Causal contrasts are made between the average outcome in a population when they are exposed to treatment with the average outcome in the same population had they remained unexposed. This contrast is only meaningful when the same subjects can be both exposed and unexposed to treatment

The IPTW-estimator for estimating the parameters of a MSM avoids directly conditioning on L in the structural model in order to estimate causal effects in the presence of time-dependent confounding. However,

the positivity condition is still necessary otherwise the IPT weights, shown for a time fixed case in (6.2), are undefined. The denominator of (6.2) is the probability that a subject receives their own exposure conditional on their own confounder value. When this probability is very small or equal to zero the weights will become very large or will be undefined.

$$w_i = \frac{1}{p(A = a_i \mid L = l_i)}. \tag{6.2}$$

In this thesis we focus on estimating MSMs using the IPTW estimator in a survival context. The weights in (6.2) can be extended for a survival context to (6.3) (see chapter 4 section 4.4 for more details). The denominator of (6.3) is the cumulative product of the conditional probability that each subject receives their own exposure conditional on their exposure and confounder histories.

$$w_{i,t} = \frac{1}{\prod_{\tau=0}^{t} p_\tau(A_{i,\tau} \mid \bar{A}_{i,\tau-1}, \bar{L}_{i,\tau})} \tag{6.3}$$

The corresponding condition to (6.1) in the survival context is that the conditional probability in the denominator of (6.3) is positive for every subject i in every time period.

$$p_\tau(A_{i,\tau} \mid \bar{A}_{i,\tau-1}, \bar{L}_{i,\tau}) > 0. \tag{6.4}$$

6.2 Types and examples of positivity violations

Violations of the positivity assumption can be structural or random. Structural violations relate closely to the idea of counterfactuals and causal models because they involve situations in which a subgroup of the population cannot possibly receive one or more levels of the exposure variable. Structural violations are deterministic in the sense that knowing which subgroup a subject belongs to tells us exactly which exposure that subject will (or will not) receive. Random violations, on the other hand, arise when, by chance, no subjects are exposed or unexposed. This is particularly the case when data is sparse or there are many confounding variables that need to be adjusted for in the analysis. In this section we consider both types of violation and discuss specific examples in which they can arise.

6.2.1 Structural violations

Structural violations of the positivity assumption occur when a subject cannot possibly be exposed, or if an individual is always exposed within some levels of the strata defined by the confounding variables. These violations are structural or deterministic because exposure is fully determined by some levels of the confounder [1]. Structural violations of the positivity assumption are often called structural zeroes because when the denominator in (6.2) or (6.3) is zero or close to zero the weights are undefined or become very large.

Many examples of structural zeroes arise in epidemiology. A common example is the healthy worker effect in occupational epidemiology. The effect of a chemical (A) on health (Y) is confounded by being at work (L), a proxy for health status [6] [7]. If exposure to the chemical can only occur at work then within the subgroup defined by $L =$ "*not at work*", the probability of being exposed is zero. In other words, people who are not at work cannot possibly be exposed to the chemical. Similar examples of structural violations occur in many other studies. For example, structural zeroes arise in the context of rates of preterm birth and racial segregation [47], or in the context of fetal position and perinatal outcomes [48]. Studies of pulmonary function (Y) and smoking (A) where age is a confounder are likely to find structural zeroes within younger age groups because in many countries young children are prevented from smoking [49]. This is equally true for the behaviour of elderly people. In one study of serious health problems (L), physical activity (A) and

mortality (Y), structural zeroes arise when no elderly people with serious health problems undertake any physical activity [50].

What these examples have in common is that the counterfactual outcomes in some sub groups of the population are not well defined. In order to make the causal contrasts described in chapter 2 section 2.1 and section 2.5 it must be possible, at least in theory, to estimate the mean of the counterfactual distribution under every level of exposure. As an example, the counterfactual outcome of mortality among elderly patients who undertake physical activity is not well defined in the sub group with serious health problems because it is difficult to conceive of serious ill elderly people undertaking any physical activity. If the counterfactual is undefined then so is the resulting causal contrast.

A final example in which structural zeroes may arise is in studies where protocols or guidelines are in place which recommend or prohibit certain levels of the exposure based on indications or contraindications for treatment. This is the case when, for example, physicians base their decision to commence exposure to AZT based on CD4 cell count which is below a certain threshold. When physicians comply with the clinical protocols they may inadvertently be positivity non-compliant because the resulting observational data from a trial will contain no unexposed subjects below a threshold CD4 cell count as specified in the guidelines or protocol. This becomes more important when the guidelines are very strict.

6.2.2 Random violations

Random violations of positivity, also called random zeroes, arise when, by chance, no individuals or all individuals, are exposed or unexposed within a certain strata or subgroup defined by the confounders. Unlike structural zeroes, there is no structural or deterministic reason why no subjects received certain levels of exposure within that level of the confounder. As a result, the counterfactual causal contrasts remain well-defined but the positivity assumption is practically violated. Random zeroes are more likely to occur when there are many confounding variables each with many levels or the data is sparse. Intuitively, when there are only two or three individuals with each strata defined by the confounders (L), the chances that no one is exposed to a certain level of the treatment (A) is higher than when many individuals fall within that strata. In a time-dependent context the conditional probability (6.4) is taken over the exposure history. In other words, (6.4) models the probability of receiving exposure based on previous exposure history, current and previous confounder history and survival to the present time period. The number of subjects who fall within this group may be relatively small so we would expect that in a time-dependent context random zeroes arise more readily.

6.2.3 Effect of positivity violations on the IPTW-estimator

Whether they are structural or random, positivity violations cause the denominator of the weights in (6.3) to become very small and the weights to become very large or infinite. When a subject has a very large weight they contribute many copies of themselves to the weighted estimator. As a result, a small number of subjects can have a strong influence on the analysis which potentially causes bias. At the same time, the variability of the weights will also increase leading to a loss of precision in the estimated parameters of the resulting MSM.

In the time-dependent survival context, the denominator of the weights in (6.3) is a product of the probability that a subject receives their own exposure and confounder history at each point in time. As we move through time, this product decreases and we would expect the presence of a very small probability in any time period to be propagated through time. Put differently, we expect the bias and loss of precision due to positivity violations to increase with the length of the study as measured by the number of time points.

6.2.4 Near-Violations of Positivity

The effects of positivity violations described in section 6.2.3 stem from the small estimated probabilities in the denominator of (6.2). Similar results would occur when the probability that a subject receives a certain

level of exposure is not strictly speaking zero, but nevertheless is very low. Under these circumstances we can expect that the bias and variance of the IPTW-estimator to also be very large and driven by a small number of individuals.

Near-violations of positivity are relevant because in trials which involve protocols or guidelines, it can often be the case that, within a subgroup defined by certain strata, it is rare to see certain exposures applied. For example, when CD4 count falls below a certain level many, but possibly not all, physicians will use this as an indication to initiate treatment. The result in this setting is that there is a very small but non-zero probability that a subject receives a certain level of the exposure and we would expect the effect of near positivity violations on the IPTW-estimator to be similar to strict violations.

Studying near violations of the positivity assumption is important because in many applied papers, only strict violations are actually checked. It is not clear in practice how close we can move to positivity violations before the bias or RMSE is so large that the results are no longer reliable.

6.3 Related work on violations of the positivity assumption

Several published studies have already investigated violations of the positivity assumption with respect to the IPTW-estimator or other causal estimators. One group of studies approaches the problem in a pedagogical manner using real data as an example of how the positivity assumption should be diagnosed and approached in an applied setting. A second group of studies is more systematic and relies on simulation studies which are similar to the current work. Here, we briefly identify the most important of these previous studies and explain how they relate to the current work.

An early example is Mortimer et al (2005) [51] who estimate a marginal structural model using the IPTW-estimator for time-dependent data in a study of pulmonary function in asthmatic children. Their results emphasised the need to check for the positivity assumption explicitly and illustrated how incorrect inference could occur when that assumption is violated. In Bembom and van der Laan (2007), the IPTW-estimator as well as the G-computation, Double Robust (DR) and Structural nested mean models (SNMM) are studied under violations of positivity. The setting is longitudinal and studies mortality among elderly subjects. An important conclusion is that violations of the positivity assumption are more severe for the IPTW-estimator than for the remaining causal estimators. However, all four estimators are affected by violations because the remaining estimators must rely fully on model assumptions which are not testable from the data. In a later paper, Cole and Hernan (2008) demonstrated the effect of violations of both the positivity assumption and the no model misspecification assumption in the context of an HIV cohort study. Their results point to an important trade-off between the exchangeability assumption and the positivity assumption. By categorizing a continuous confounder by increasingly fine levels, Cole and Hernan (2008) show how random zeroes can arise. This message is particularly important when the number of confounders is large relative to the sample size.They conclude their paper by pointing to the need for future work which formally studies these trade-offs.

So far, each of the aforementioned studies has used real data. For understanding the impact of positivity violations this has several limitations. First, disentangling the effect of positivity violations from other sources of bias is typically not possible when using real data. For example, important confounders may be missed or unmeasured. In general, it will be difficult to ascertain that the positivity assumption is violated and the remaining assumptions underlying the IPTW-estimator are met. Model misspecification, like positivity violations, leads to large and unstable estimates of the weights which means that in a real data setting diagnosis of positivity violations will be difficult as some degree of model misspecification is virtually unavoidable. A second point is that real data does not permit us to design scenarios that might be of interest. For example, the performance of the estimator at varying sample sizes. We now turn to several simulation studies which have been conducted on the effect of positivity violations on the IPTW-estimator.

An early example of a paper which formally studies positivity violations in a simulation context is Neugebauer and van der Laan (2005) [52]. In their paper, the IPTW-estimator as well as the G-computation and Doubly Robust (DR) estimators are compared when violations of the positivity assumption is present. Their results suggest that the IPTW-estimator performs less well when positivity violations are present than the remaining

two estimators. This comparison takes place in a time-fixed or point treatment setting which means the results have limited generalizability with regard to the complexities that arise in a longitudinal or survival context. The MSM from which they simulate data is linear, which means that non-collapsibility and the associated problems outlined in chapter 5 section 5.2.2 is not an issue. However, this also means that their algorithm cannot be used to study common effect measures like the odds ratio, nor can it be easily generalized to a survival setting.

Wang et al (2006) [53] and Petersen et al (2012) [54] use a similar approach for assessing the bias introduced due to positivity violations in the IPTW-estimator. In both papers data is generated from a linear MSM in a time fixed context. The parametric bootstrap is used to draw samples from the generated data to which both the G-computation and IPTW-estimators are applied. Because the G-computation estimator is known to be consistent under violations of positivity - as long as model assumptions are fully met - the difference in the two estimates averaged over the bootstrapped samples is due to non-positivity bias. Positivity violations are random zeroes, introduced by controlling the probabilities in the weight model. Structural zeroes are not considered in either Wang (2006) or Petersen (2012). Both papers indicate that positivity violations are a significant source of bias when using the IPTW-estimator. Neither paper reports results in a time-dependent setting. Similar to Neugebauer and van der Laan (2005), the MSM is linear and cannot be generalized to the case of logistic regression or a survival setting.

Naimi et al (2011) [55] is probably the study which most closely relates to the current work. They simulate data to mimic the healthy worker effect in occupational epidemiology. Each subject is either at work (L = 1) or not at work (L=0). When the subject is not at work they are assigned no exposure (A = 0) deterministically. When the subject is at work they are assigned levels of exposure according to a Bernoulli random variable. The context is time-dependent, albeit for only two periods. The authors reflect in the discussion of their results that further work is needed to establish whether or not bias due to positivity violations is accentuated or attenuated in longer follow-up times: a novelty of this thesis. Selection bias is also present in their algorithm and introduced in a similar way to the algorithm of Havercroft and Didelez (2012) through a latent health variable.

6.4 Simulation Algorithm with Positivity Violations

In this section we describe two extensions made to the algorithm of Havercroft and Didelez (2012) which make it possible to introduce positivity violations. The first extension introduces positivity violations below a certain threshold of the confounding variable L. Below this threshold all subjects are exposed and there is a structural zero probability of non-exposure. The idea is that confounder values below the threshold are a strict indication for exposure.

The second extension to the algorithm allows for near positivity violations. To study the effect of near violations of the positivity assumption we need some control over how these violations are introduced into the simulated data. In previous work on near violations of the positivity assumption, a model for $P(A \mid L)$ is specified in such a way that the probability of receiving a certain level of exposure is low within some values of the confounders. The algorithm typically provides control over the intensity of the violations. This approach is not suitable for our purposes because we aim to investigate the scenario that positivity violations occur for some, but not all, subjects in a deterministic way. Roughly, the idea is that some subjects are positivity compliant in that they have a positive probability of receiving each level of the exposure within every strata defined by the confounder. On the other hand, some subjects are positivity non-compliant at a threshold value, meaning that below this threshold they always receive a specific level of the exposure. To add some realism to this scenario we can think of positivity non-compliant subjects as being treated by physicians who strictly adhere to clinical trial protocols or guidelines. The end result should be a sample of subjects, some of whom are positivity compliant and some are not. The question is for what threshold and for what proportion of positivity non-compliant subjects is the IPTW-estimator impaired.

6.4.1 Thresholding and Positivity Violations

We now explain exactly how positivity violations are introduced into the algorithm by means of a threshold τ below which no subject is exposed. There is a slight abuse of notation here. We used τ in, for example, the weights of (6.3) to represent time periods. From this point on τ represents a threshold values of the confounder variable L. We could have used a different symbol, but τ seemed like an appropriate symbol for a threshold.

First a value of the confounder at time t, L_t, is assigned. If L_t is below a threshold τ the subject begins treatment and remains on treatment for the remainder of the study. If, on the other hand, L_t is above the threshold τ the subject has a positive probability of receiving each level of the exposure. Put slightly differently, below the threshold exposure is assigned deterministically and above the threshold exposure is assigned stochastically. Procedurally these steps are summarised in algorithm 2.

Algorithm 2: Threshold positivity violations

Result: Positivity violations within strata of L

1 initialization (τ, L_t);
2 **if** $L_t < \tau$ **then**
3 $\quad A_{t,i} \leftarrow 1$
4 **else**
5 $\quad A_{t,i} \leftarrow Bern(logit^{-1}(\theta_0 + \theta_1 t + \theta_2(L_{t,i} - 500)))$
6

The steps in algorithm 2 allow us to vary the threshold below which positivity violations occur and quantify how extreme these violations are in their effect on the IPTW-estimator. This is an extension on the work of Naimi et al (2011) because it combines the idea that positivity violations are deterministic within some discrete interval $[0, \tau]$ with control over the size of that interval and the presence of a continuous, rather than binary, confounder.

A crucial point about the base algorithm is that we have control over the pathway $L_t \rightarrow A_t$ at all points in time. Combined with algorithm 2, positivity violations are introduced at whichever point in time the confounder L_t falls below a certain threshold. For example, if an HIV patient enters the study with a baseline CD4 count of 700 cells/μl and, after 10 time points this deteriorates below the threshold to, say, 100 cells/μl, positivity violations will follow after the 10th time point. In this way positivity violations are both introduced and propagated through time but the advent of each violation will vary by subject in the study. Algorithm 3 embeds algorithm 2 in the base algorithm and the changes are highlighted in pink. Lines 7-10 of algorithm 2 deal with the initial or baseline values. Lines 29-32 deal with all subsequent time points and,

in the wider structure of the generated data, allow the positivity violations to be propagated through time.

Algorithm 3: Base algorithm and threshold positivity violations

Result: Simulation algorithm

initialization τ;
for *i in 1, . . . , n* **do**
  $U_{0,i} \sim U[0,1]$
  $\epsilon_{0,i} \sim N(0,20)$
  $L_{0,i} \leftarrow F^{-1}_{\Gamma(3,154)}(U_{i,0}) + \epsilon_{0,i}$
  $A_{-1,i} \leftarrow 0$
  if $L_0 < \tau$ **then**
    $A_{0,i} \leftarrow 1$
  else
    $A_{0,i} \leftarrow Bern(logit^{-1}(\theta_0 + \theta_2(L_{0,i} - 500)))$
  if $A_{0,i} = 1$ **then**
    $T^* \leftarrow 0$;
  $\lambda_{0,i} \leftarrow logit^{-1}(\gamma_0 + \gamma_2 A_{0,i})$
  if $\lambda_{0,i} \geq U_{0,i}$ **then**
    $Y_{1,i} \leftarrow 0$
  else
    $Y_{1,i} \leftarrow 1$
  for *t in 1, . . . , T* **do**
    while $Y_{t,i} = 0$ **do**
      $\Delta_{t,i} \sim N(0, 0.05)$
      $U_{t,i} \leftarrow min\{1, max\{0, U_{t-1,i} + \Delta_{t,i}\}\}$
      if $t \neq 0 (mod\ k)$ **then**
        $L_{t,i} \leftarrow L_{t-1,i}$
        $A_{t,i} \leftarrow A_{t-1,i}$
      else
        $\epsilon_{t,i} \sim N(100(U_{t,i} - 2), 50)$
        $L_{t,i} \leftarrow max\{0, L_{t-1,i} + 150 A_{t-k,i}(1 - A_{t-k-1,i}) + \epsilon_{t,i}\}$
        if $A_{t-1,i} = 0$ **then**
          if $L_t < \tau$ **then**
            $A_{t,i} \leftarrow 1$
          else
            $A_{t,i} \leftarrow Bern(logit^{-1}(\theta_0 + \theta_1 t + \theta_2(L_{t,i} - 500)))$
        else
          $A_{t,i} \leftarrow 1$
        if $A_{t,i} = 1$ *and* $A_{t-k,i} = 0$ **then**
          $T^* \leftarrow t$
      $\lambda_{t,i} \leftarrow logit^{-1}(\gamma_0 + \gamma_1[(1 - A_{t,i})t + A_{t,i}T^*] + \gamma_2 A_{t,i} + \gamma_3 A_{t,i}(t - T^*))$
      if $1 - \prod_{\tau=0}^{t}(1 - \lambda_{\tau,i}) \geq U_{0,i}$ **then**
        $Y_{t+1,i} = 1$
      else
        $Y_{t+1,i} = 0$

6.4.2 Proportion of positivity compliant doctors.

The second extension we make to the base algorithm allows near violations of the positivity assumption to occur. To do so, we introduce a second parameter π which represents the proportion of positivity compliant subjects. The idea is that some subjects are positivity compliant in the sense that there is a positive probability that they will receive every level of a binary exposure variable (A). Subjects who are not positivity compliant can be thought of as following some kind of rule such as "always receive treatment when CD4 cell count falls below 350 cells/μl". Where study protocols exist, these subjects strictly adhere to the protocol.

We generate data in which only a proportion π of subjects are positivity compliant and the remaining proportion $1-\pi$ are positivity non-compliant. First we generate a random uniform variable P (called P to avoid clashing with U in the original algorithm) per subject on the interval from zero to one. If $\pi \leq P$ the subject is assigned to the positivity non compliant group. The positivity non-compliant subjects are deterministically assigned exposure at the point when their confounder value L_t first falls below a threshold value τ. Above this threshold they have a positive probability of receiving every level of the exposure. Positivity compliant subjects, on the other hand, have a positive probability of receiving every level of the exposure for every value of the confounder. The steps are summarised in algorithm 4.

Algorithm 4: Proportion of positivity compliant subjects

Result: Positivity violations within strata of L for positivity non-compliant subjects

initialization (τ, π, L_t);
$P \sim U[0,1]$
if $\pi \leq P$ **then**
  if $L_t < \tau$ **then**
    $A_{t,i} \leftarrow 1$
  else
    $A_{t,i} \leftarrow Bern(logit^{-1}(\theta_0 + \theta_1 t + \theta_2(L_{t,i} - 500)))$
else
  $A_{t,i} \leftarrow Bern(logit^{-1}(\theta_0 + \theta_1 t + \theta_2(L_{t,i} - 500)))$

The result of algorithm 4 is to create a sample of data in which a proportion of subjects π who are positivity compliant are mixed with a proportion of subjects $1-\pi$ who are positivity non-compliant. As the proportion of compliant subjects decreases we would expect to see more and more extreme violations of the positivity assumption in the strata below the threshold τ. Our near violations of positivity include both deterministic violations of the positivity assumption in the interval $[0, \tau]$ with subjects who do not violate the assumption. In contrast, earlier work has focused on near violations of the positivity assumption which occur in certain strata defined by the confounders by chance. Our hope is that this set-up will go some way towards the realistic setting in which clinical guidelines or protocols are followed, but perhaps not strictly. The strictness of the study can be emulated by varying the parameter π. This will be of particular use when certain confounders are very strong indications or contraindications for some levels of the exposure.

It should be clear from algorithms 2 and 4 that our approach to creating threshold positivity violations is nested within our approach to introducing near violations of positivity. Setting $\pi = 0$ in algorithm 4 would achieve the same result as algorithm 2. Algorithm 5 embeds algorithm 4 in the base algorithm. The changes to the base algorithm are highlighted in pink. Just as in algorithm 3, we retain control over the pathway $L_t \rightarrow A_t$ which means that near violations of the positivity assumption are propagated across time. Our study, therefore, takes place in a realistic time-dependent survival context in which near violations of the

positivity assumption occur.

Algorithm 5: Base algorithm and proportion of positivity compliant subjects

Result: Simulation algorithm

initialization (τ, π);
for *i in 1, . . . , n* **do**
  $U_{0,i} \sim U[0,1]$
  $\epsilon_{0,i} \sim N(0, 20)$
  $L_{0,i} \leftarrow F^{-1}_{\Gamma(3,154)}(U_{i,0}) + \epsilon_{0,i}$
  $A_{-1,i} \leftarrow 0$
  $P \sim U[0,1]$
  if $\pi \leq P$ **then**
    if $L_0 < \tau$ **then**
      $A_{0,i} \leftarrow 1$
    else
      $A_{0,i} \leftarrow Bern(logit^{-1}(\theta_0 + \theta_2(L_{0,i} - 500)))$
  else
    $A_{0,i} \leftarrow Bern(logit^{-1}(\theta_0 + \theta_2(L_{0,i} - 500)))$
  if $A_{0,i} = 1$ **then**
    $T^* \leftarrow 0$;
  $\lambda_{0,i} \leftarrow logit^{-1}(\gamma_0 + \gamma_2 A_{0,i})$
  if $\lambda_{0,i} \geq U_{0,i}$ **then**
    $Y_{1,i} \leftarrow 0$
  else
    $Y_{1,i} \leftarrow 1$
  for *t in 1, . . . , T* **do**
    while $Y_{t,i} = 0$ **do**
      $\Delta_{t,i} \sim N(0, 0.05)$
      $U_{t,i} \leftarrow min\{1, max\{0, U_{t-1,i} + \Delta_{t,i}\}\}$
      if $t \neq 0 (mod\ k)$ **then**
        $L_{t,i} \leftarrow L_{t-1,i}$
        $A_{t,i} \leftarrow A_{t-1,i}$
      else
        $\epsilon_{t,i} \sim N(100(U_{t,i} - 2), 50)$
        $L_{t,i} \leftarrow max\{0, L_{t-1,i} + 150 A_{t-k,i}(1 - A_{t-k-1,i}) + \epsilon_{t,i}\}$
        if $A_{t-1,i} = 0$ **then**
          if $\pi \leq P$ **then**
            if $L_t < \tau$ **then**
              $A_{t,i} \leftarrow 1$
            else
              $A_{t,i} \leftarrow Bern(logit^{-1}(\theta_0 + \theta_1 t + \theta_2(L_{t,i} - 500)))$
          else
            $A_{t,i} \leftarrow Bern(logit^{-1}(\theta_0 + \theta_1 t + \theta_2(L_{t,i} - 500)))$
        else
          $A_{t,i} \leftarrow 1$
        if $A_{t,i} = 1$ *and* $A_{t-k,i} = 0$ **then**
          $T^* \leftarrow t$
      $\lambda_{t,i} \leftarrow logit^{-1}(\gamma_0 + \gamma_1[(1 - A_{t,i})t + A_{t,i}T^*] + \gamma_2 A_{t,i} + \gamma_3 A_{t,i}(t - T^*))$
      if $1 - \prod_{\tau=0}^{t}(1 - \lambda_{\tau,i}) \geq U_{0,i}$ **then**
        $Y_{t+1,i} = 1$
      else
        $Y_{t+1,i} = 0$

6.5 Simulation scenarios

Algorithms 3 and 5 make it possible to simulate data in a time-dependent survival context when the positivity violation is either violated or the data exhibits near violations of positivity. In this section we outline four scenarios for investigating positivity violations. As described in section 6.4.1 and section 6.4.2, we can vary several parameters to introduce more or less extreme violations of positivity. We can control τ, the threshold below which no subject is exposed to treatment, and we can control π, the proportion of positivity compliant doctors to create near violations of positivity. By virtue of the design of the original algorithm, we also have control of n, the number of subjects and T the number of follow-up time points. Combining all four parameters allows for a range of possible scenarios in which positivity is violated or nearly violated. Here, we describe four possible scenarios and explain why they are of interest and our expectations of the results. Full results of each scenario are presented in chapter 7.

6.5.1 Scenario 1: Positivity violations thresholds

In the first scenario we investigate how the performance of the IPTW-estimator varies with the threshold τ below which no subject is exposed to treatment. The probability of remaining unexposed in the strata defined by $L \leq \tau$ is, in theory, zero as in (6.6). In practice, we use a parametric model for (6.5). The parametric model will smoothe over the structural zeroes and the fitted probabilities will be very close to, but not actually, equal to zero.

$$P(A = 0 \mid L \leq \tau) = 0 \tag{6.5}$$

We expect that both the bias and variance of the IPTW estimates of the parameters of the MSM will increase as the threshold increases. The reason is twofold. First, the wider the threshold the more likely we are to observe extreme weights. To see why, consider the a study of exposure A on a health outcome among people of different age groups. If no-one in the study at age 7 is exposed, it might be possible to lump the age 7 and 8 year olds into one coarser group. The probability of exposure in the coarser group will be balanced by the parametric model which will smoothe over the positivity violations among 7 year olds by borrowing information from the 8 year olds. However, if no-one between the ages 7 and 20 are exposed then the same smoothing effect of the parametric model will lead to a very small probability of exposure at age 7. Thus, as we extend the size of the interval $[0, \tau]$ in which positivity violations occur we expect to see more extreme weights. Subjects with large weights contribute many more copies of themselves to the analysis which leads to bias and high variance in the estimate.

A second reason why we expect the performance of the IPTW-estimator to deteriorate as τ increases, is that by casting a larger net we capture more subjects who, as a result of the positivity violations will have large weights. Thus we have more subjects with high weights in specific strata of the confounders. We therefore expect the IPTW-estimator to be affected more at larger values of τ.

6.5.2 Scenario 2: Number of subjects

The base algorithm also allows us to control the sample size of the data generated in each simulation. This means that we can combine control of the positivity violations by means of the threshold τ with control of the sample size n to investigate the finite sample properties of the IPTW-estimator under positivity violations.

Studying the finite-sample properties of the IPTW-estimator is important because if the performance of the estimator under violations of the positivity assumption, increases with the sample size, this may recommend its use to researchers only in cases where the sample size is sufficiently large to mitigate the negative effects of positivity violations. Moreover, it is often assumed by analysts that problems with statistical techniques generally improve with larger sample sizes. This matches the intuitive idea that more data and therefore more information can overcome problems in statistical techniques.

However, there area reasons to believe that the IPTW-estimator would be impaired under positivity violations at any sample size. First, the effect of positivity violations, as pointed out in section 2.2.4, remain even as the sample size increases. Positivity violations mean that the correct causal contrasts cannot be evaluated within certain subgroups defined by the confounders. Second, evidence on the effect of model misspecification, suggests that the IPTW-estimator performed better in smaller samples when the weight model was misspecified when measured by the RMSE [56]. Model misspecification, like positivity violations, leads to large and highly variable weights and remains a problem even in large samples.

Another consideration is the role of random zeroes which are more likely to arise when the data is sparse. At smaller sample size of, say, $n = 100$ we might expect random zeroes to arise which, from a practical point of view, have the same effect on the IPTW-estimator. This is an important point because our simulation study attempts to isolate the bias due to positivity violations which we deliberately introduce. Random violations of the positivity which occur outside of our control are more likely to arise in smaller sample sizes. Our study only includes one confounder variable. In more realistic settings with more confounders the bias due to random zeroes is more likely. Nonetheless, we take care to distinguish between random violations of the positivity assumption which may occur in small sample sizes with the deterministic violations we introduce in algorithms 3 and 5.

6.5.3 Scenario 3: Positivity Violations Thresholds By Length of Follow-up Time

In the third scenario we combine control over the threshold τ with control over the length of follow up time T. Specifically, we consider a set of follow up times and compare the effect of increasing the threshold τ within each of these follow-up times. Interest lies in whether positivity violations are accentuated or attenuated over a larger number of points in time. To our knowledge, no previous study has investigated the relationship between positivity violations and follow-up time in a longitudinal or survival context.

Section 6.2.3 explained one reason why we might expect the bias and variability of the IPTW estimates to increase as the length of follow-up increases. The denominator of (6.3) is a cumulative product of the conditional probability of exposure given confounder and exposure histories at every point in time. We expect that a positivity violation will result in a small fitted probability at the point in time where the violation occurs. In subsequent time periods the cumulative product of the time specific probabilities will lead the denominator of (6.3) to rapidly shrink and the corresponding weights will become very large. In other words, our expectation is that a longer follow-up time will accentuate the impact of positivity violations.

However, we again have to be careful interpreting the results because there is a greater chance of random zeroes in longer follow-up times than shorter follow-up times. This occurs because the probability of exposure is conditional on previous exposure and confounder histories. The positivity assumption states that there should be a positive probability of exposure within each strata defined by these histories. What this means is that we need enough subjects who have received the same exposure and similar confounder histories in order to calculate the weights. For the same reason that more confounders would lead to a greater chance of random zeroes, so too would a longer follow-up time lead to more random zeroes because only a small number of subjects are likely to have the same or similar confounder and exposure histories. We would therefore expect the bias and variance of the IPTW-estimator to increase with time. Our focus is largely on the effect of the structural positivity violations so we take care that random zeroes may also drive our results.

6.5.4 Scenario 4: Near positivity violations and positivity compliant doctors.

In the final scenario we relax the strict violations of positivity used in scenarios one, two and three. Instead, we apply algorithm 5 in order to introduce near violations of positivity below a certain threshold τ for a proportion $1-\pi$ of subjects. The idea is to investigate how the performance of the IPTW-estimator depends on both τ and π. For example, when study guidelines contain strict rules for exposure we might find that these are followed by 80% of subjects. In this case 80% of subjects are positivity non-compliant in the sub group of L to which the study guidelines pertain. In algorithm 5 this sub group is defined by the parameter

τ. The question is whether the remaining 20% of positivity compliant subjects are enough to guarantee the positivity assumption overall.

In terms of expectations, near violations of the positivity assumption are expected to have less of an impact on the IPTW-estimator that strict violations. Strict violations of the positivity assumption will lead to very large weights because the estimated conditional probability of receiving certain levels of the exposure will be low or almost zero. Near violations will not be as extreme because they do involve some exposed and non-exposed subjects at each level of the confounder. Nonetheless, the estimated conditional probabilities in the denominator of (6.4) may still be small enough to lead to inflated weights. The consequence is greater bias and variance in the IPTW estimates of the parameters of the MSM.

Chapter 7

Simulation study

In this chapter we present the results of investigating the four scenarios described in chapter 6 section 6.5. We begin this chapter by describing the simulation set-up including the parameters of the MSM from which we generate data. We then present the results of the Monte Carlo simulation study used to assess the performance of the IPTW-estimator under varying degrees of positivity violations and in different contexts.

7.1 Simulation set-up

In this thesis we follow the work of Bryan et al (2004) and Havercroft and Didelez (2012) who both use a very similar simulation set-up to generate their data. We use the same parameters as those found in Havercroft and Didelez (2012) which have been calibrated to generate data which is similar to the Swiss HIV Cohort study [45].

Each simulated data set consists of n subjects for whom a binary outcome variable Y, a binary exposure variable A and a continuous confounder L are generated per time period. Subjects are followed across T time periods with monitoring times at each discrete time point as in (7.1). A monitoring time means that the current state of the subject is recorded at each of these points in time. At each point in time the state of the subject can be either alive ($Y_t = 0$) or dead ($Y_t = 1$). After experiencing the event a subject is no longer followed. There is no censoring in any of our simulations, except for those subjects still alive at the end of the study who are right-censored at the end of follow-up. The algorithm has been deliberately calibrated to ensure that only a small number of subjects remain alive after $T = 40$ time points.

$$\{t_0 = 0, t_1 = 1, \ldots, t_{T-1} = T - 1\}. \tag{7.1}$$

Measuring times occur every five intervals at which point the confounder L is measured and exposure is assigned based on the conditional distribution of exposure A given L. The first measuring time is t_0 followed by t_5, t_{10}, and so on at intervals of five. This matches the set-up of Bryan et al (2004) and Havercroft et al (2012). Each subject is assigned an exposure level at a measuring time by (7.2) and (7.3). The parameters $(\theta_0, \theta_1, \theta_2)$ are given by (7.4). The conditional probability that a subject receives their own exposure is exactly the denominator of the IPTW weights. Model (7.2), therefore, is the weight model in this case, as well as the model for assigning exposure. When a time point is not a measuring time the subject receives the same value of both the confounder and exposure as in the previous time period. Once treatment is initiated, the subject remains on treatment until either failure or the end of follow-up.

$$logit(P(A_t \mid \bar{L}_t)) = \theta_0 + \theta_1 t + \theta_2 (L_t - 500) \tag{7.2}$$

$$A_t \sim Bern(logit^{-1}(\theta_0 + \theta_1 t + \theta_2 (L_t - 500)) \tag{7.3}$$

$$(\theta_0, \theta_1, \theta_2) = (-0.405, 0.0205, -0.00405). \tag{7.4}$$

The same MSM for the hazard function is used in both Bryan et al (2004) and Havercroft and Didelez (2012) and is shown in (7.5). The parameters $(\gamma_0, \gamma_1, \gamma_2, \gamma_3)$ are given in (7.6). The MSM (7.5) is a logistic model for the odds ratio. The parameters of this model can be viewed as discrete time approximations to the HR, as described in chapter 4.

$$\lambda_t^{\bar{a}_t} = logit^{-1}(\gamma_0 + \gamma_1((1 - a_t)t + \gamma_2 a_t t^*) + \gamma_3 a_t(t - t^*)) \tag{7.5}$$

$$(\gamma_0, \gamma_1, \gamma_2, \gamma_3) = (-3, 0.05, -1.5, 0.1). \tag{7.6}$$

With the parameters specified in (7.4) and (7.6) data can be generated from algorithms 3 and 5 for a given sample of size n and a follow-up time T. We will vary these two parameters in our results along with the threshold below which positivity is introduced into the model (τ) and the proportion of positivity compliant subjects (π). Specific values of these parameters will be specified for each scenario considered.

Each simulation trial consists of $B = 100$ simulations. A larger number of simulations per trial is desirable to lessen the impact of sample variability on the combined results of each trial. However, the algorithm is quite time intensive and $B = 100$ simulations per trial obtained reliable results in a reasonable amount of time.

7.2 Results

We now present the results of the simulation exercise to investigate the performance of the IPTW-estimator for estimating MSMs in a time dependent context under violations of the positivity assumption. The extended algorithm, which was described in chapter 6 section 6.3, is used to generate data in order to investigate the four scenarios outlined in chapter 6 section 6.4. The extended algorithm makes it possible to vary two parameters which are used to introduce positivity violations into each generated dataset. Specifically, we control a threshold level τ of the confounder L below which no subject receives treatment, and the proportion of positivity compliant subjects π. We also vary the number of subjects n per dataset to investigate how positivity violations influence the finite sample properties of the IPTW-estimator. Finally, we also vary the length of follow up time T in order to investigate how the effects of positivity violations are propagated through time.

For each simulation we use several criteria to assess the performance of the IPTW-estimator under varying degrees of positivity violation. Positivity violations are know to affect both the bias and the variance of the IPTW estimator and it is therefore necessary to use criteria which will evaluate both of these features [9]. For each simulation we estimate the bias, standard deviation and root mean squared error (RMSE). Each of these measures is described in more detail in chapter 5 section 5.1.2.

7.2.1 Scenario 1: Positivity Violations Thresholds

In the first scenario we investigate the effect of positivity violations on the performance of the IPTW-estimator. Violations of the positivity assumption are introduced by varying the parameter τ, the threshold below which subjects are always exposed to treatment. In other words, there is a structural zero probability that a subject will remain unexposed when their confounder value L falls in the interval $[0, \tau]$.

Table 7.1 presents parameter estimates from a Monte Carlo simulation with $B = 100$ repetitions of $n = 1000$ subjects and thresholds of $\tau = \{0, 100, 200, 350\}$. The true parameter estimates used to generate the simulated data are also shown in the first column of table 7.1 for reference. We include a threshold of $\tau = 0$ (no positivity violations) to demonstrate that the IPTW-estimator does resolve the correct parameter in the

absence of positivity violations. As expected, the parameter estimates are biased when positivity violations are present below the threshold τ. Moreover, this mainly affects γ_0, the intercept, and γ_2, the parameter most directly related to the effect of exposure. The bias for these two parameters becomes noticeably larger as the threshold τ increases. Indeed at higher levels of τ ($\tau = \{300, 350\}$) the sign of γ_2 flips from being negative to positive. In other words, the effect of exposure flips from being preventative to actually increasing the chance of experiencing the event in the next period. In contrast parameters (γ_1, γ_3) remain relatively consistent across each scenario. An important point is that as τ increases the intercept γ_0 and γ_2 are moving in opposite directions. This will have consequences for assessing the effect of positivity violations on the estimates of the structural hazard and survival curves. The strength of the bias in each of these parameters will determine how closely the structural hazard and survival curves match the true curves. A second point is that the standard deviation of all four parameters increases with the threshold τ. This reflects the fact that positivity violations, and the resulting large weights, lead to greater variability in the IPTW estimates. In scenario 1, and in subsequent scenarios, we focus attention on γ_2 because it is the parameter most directly associated with exposure.

Table 7.1: Pooled logistic regression results for data simulated with positivity violations at increasing threshold levels τ.

	True	$\tau = 0$	$\tau = 100$	$\tau = 200$	$\tau = 300$	$\tau = 350$
γ_0	-3	-3.014 (0.094)	-3.131 (0.106)	-3.841 (0.144)	-5.848 (0.578)	-6.524 (0.815)
γ_1	0.05	0.051 (0.011)	0.052 (0.010)	0.070 (0.017)	0.058 (0.043)	0.053 (0.066)
γ_2	-1.5	-1.525 (0.156)	-1.382 (0.120)	-0.647 (0.177)	1.155 (0.567)	1.554 (0.796)
γ_3	0.1	0.102 (0.014)	0.100 (0.009)	0.090 (0.013)	0.093 (0.013)	0.094 (0.022)

In figure 7.2 we extend the analysis to studying the bias, standard deviation and RMSE of γ_2 under many thresholds separated by step sizes of 25. As the threshold increases the bias of the IPTW estimate of γ_2 increases sharply. The standard deviation of the Monte Carlo estimate of γ_2, on the other hand, remains relatively constant until a threshold of 175, at which point it increases considerably. Our results suggest that the bias introduced by structural violations of positivity is of greater concern than the variability. The RMSE shows the combined effect of bias and variability on the IPTW estimate of γ_2.

Finally, in figure 7.3 we plot the structural hazard and survival curves under thresholds $\tau = \{100, 200, 300, 350\}$. For comparison the true hazard function and survival function are also plotted and correspond to the case where $\tau = 0$. The dotted lines reflect the hazard and survival under exposure while the single solid line in each plot represents the same functions under no exposure. In order to make a meaningful comparison each line represents the situation where exposure is always or never received in every period of time. What we see in the structural hazard functions is that the probability of experiencing the event in the next period is actually lower at higher thresholds compared to lower thresholds. Correspondingly, the probability of survival at higher thresholds is greater than for lower thresholds. This result is counter intuitive because on the one hand greater violations of the positivity assumption lead to a smaller or even positive effect of exposure on outcome but at the same time greater thresholds also appear to suggest a higher chance of survival at every time point. The reason for this is that the bias in γ_0 is stronger than the bias in γ_2 and works in the opposite direction. As a result, the hazard function is actually lower even though the preventative effect of exposure is decreasing.

7.2.2 Scenario 2: Number of subjects

In the second scenario we extend the analysis to consider the finite sample properties of the IPTW-estimator under violations of the positivity assumption. Just as in scenario 1, violations of the positivity assumption are introduced by varying the threshold τ so that subjects with a confounder value L below τ always receive exposure. We perform the same simulations as in scenario 1 individually for sample sizes of $n = \{100, 200, 300, 400, 500, 1000\}$.

Table 7.2 contains the results of these Monte Carlo trials for each sample size at several thresholds τ. We only focus on the parameter γ_2 because our interest is primarily in the parameter most closely related to the effect of exposure. As expected, the effect of the positivity violations within each sample size increases with the threshold τ. Across sample sizes we see some evidence that the bias of the IPTW-estimate of γ_2 is higher for smaller sample sizes. However, this is not markedly different except for the largest threshold of $\tau = 350$.

Table 7.2: Pooled logistic regression results for γ_2 from data simulated with positivity violations at different sample sizes (n) and threshold levels (τ).

n	τ	$\hat{\gamma}_2$	s.d.	bias	RMSE
100	0	-1.644	0.418	-0.144	0.442
100	100	-1.400	0.365	0.100	0.378
100	200	-0.687	0.588	0.813	1.004
100	350	6.033	5.995	7.533	9.628
200	0	-1.592	0.255	-0.092	0.272
200	100	-1.439	0.283	0.061	0.289
200	200	-0.697	0.328	0.803	0.868
200	350	4.106	4.855	5.606	7.416
300	0	-1.575	0.229	-0.075	0.241
300	100	-1.442	0.201	0.058	0.210
300	200	-0.741	0.299	0.759	0.816
300	350	2.892	3.463	4.392	5.593
400	0	-1.546	0.199	-0.046	0.205
400	100	-1.402	0.195	0.098	0.219
400	200	-0.719	0.260	0.781	0.823
400	350	2.304	2.684	3.804	4.655
500	0	-1.552	0.204	-0.052	0.211
500	100	-1.355	0.189	0.145	0.239
500	200	-0.712	0.209	0.788	0.815
500	350	2.058	1.989	3.558	4.077
1000	0	-1.533	0.145	-0.033	0.149
1000	100	-1.376	0.107	0.124	0.164
1000	200	-0.683	0.192	0.817	0.839
1000	350	1.425	0.675	2.925	3.002

Figure 7.3 presents the bias, standard deviation and RMSE of γ_2 for all six sample sizes at increasing thresholds separated by a step size of 25. An interesting result is that the bias of the IPTW-estimate is more or less the same across all sample sizes until a threshold of around 250. After this point the bias in the IPTW estimate decreases with the sample size. This result is interesting because it demonstrates that the bias due to positivity violations does not improve with sample size, at least for smaller values of τ. This matches our expectations because positivity violations are fundamentally linked to counterfactual outcomes and therefore causal effects. What figure 7.3 tells us is that regardless of the sample size the IPTW-estimator cannot resolve the correct causal effect as measured by γ_2 when the positivity assumption is violated. It is not immediately clear why the divergence in bias in figure 7.3 occurs around a threshold of 250. On the

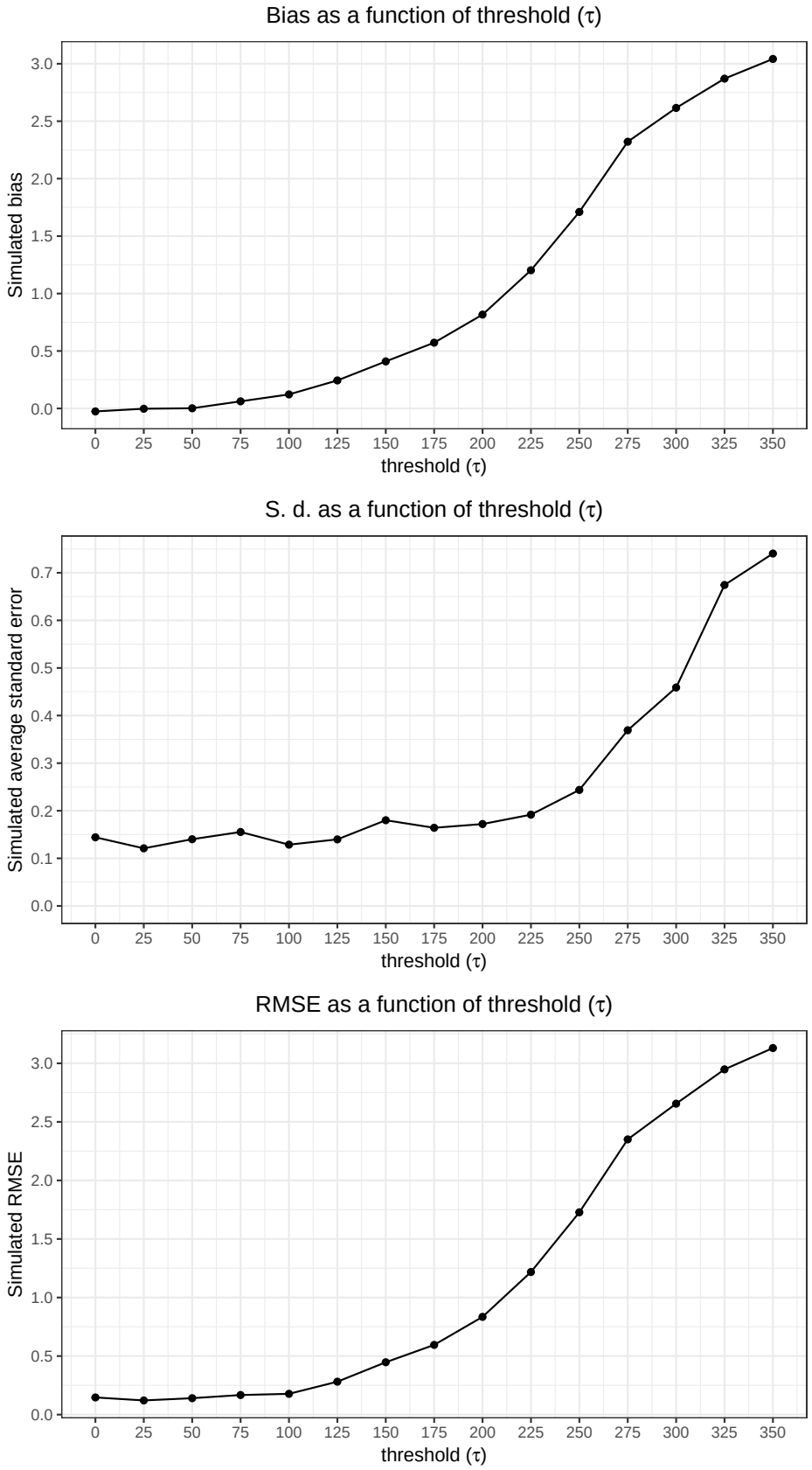


Figure 7.1: Monte Carlo bias, standard deviation (s.d.) and root mean squared error (RMSE) for γ_2 at various thresholds (τ) and $T = 40, k = 5$. $B = 100$ Monte Carlo repetitions per threshold.

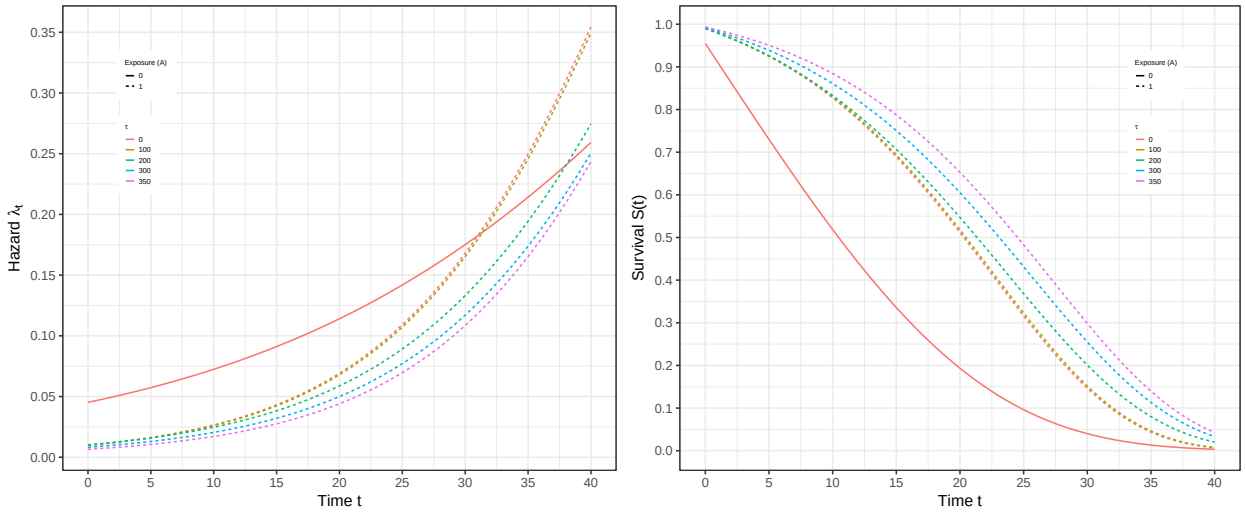


Figure 7.2: Structural hazard and survival curves under different thresholds (τ) and $T = 40, k = 5$.

other hand the variability of the estimate as measured by the standard deviation is larger for smaller sample sizes and this is also reflected in the RMSE.

The structural hazard and survival functions are plotted in figure 7.4 for sample sizes of $n = \{200, 1000\}$ for several values of the threshold τ. The set-up of figure 7.4 is the same as in scenario 1. In general the survival curves are more tightly clustered together for the smaller sample size $n = 200$ than $n = 1000$. From the point of view of estimating survival as an outcome of interest, the IPTW estimator actually performs better in the smaller sample size than the larger sample size. Partly this is again due to the fact that the intercept γ_0 and γ_2 work in different directions when the positivity assumption is violated. As a result, from the point of view of the hazard and survival functions it appears that the IPTW estimator performs better in small sample sizes even though at higher thresholds the causal effect of exposure as measured by the per period HR will be more biased than for larger sample sizes.

7.2.3 Scenario 3: Positivity Violations By Length of Follow-up Time

In scenario 3 we consider how the impact of positivity violations on the IPTW-estimator varies with the length of the follow-up time. Violations of the positivity assumption are introduced by varying the thresholds τ. Subjects with confounder values below this threshold are always treated. We simulate data separately for three follow-up times $T = \{20, 30, 40\}$. For each follow-up time we generate $B = 100$ Monte Carlo datasets per threshold τ. We consider values of τ between zero and 350 at step sizes of 25.

We focus our attention on γ_2 and the results of our simulations are shown in table 7.3. Each row of table 7.3 represents a simulation trial defined by the values of T and τ shown. For each combination of T the IPTW estimator performs correctly when $\tau = 0$. As τ increases the performance of the IPTW-estimator rapidly decreases. Table 7.3 shows greater bias for shorter time periods compared to longer time periods. For example, at a threshold of $\tau = 200$ the bias for a follow up time of $T = 20$ is 1.100. For the same threshold and a follow-up time of $T = 40$ the bias is only 0.947. The standard deviation of the estimate is also lower for shorter time periods compared to longer time periods.

Figure 7.5 plots the bias, standard deviation and RMSE of γ_2 for all combinations of T and τ. We see that, across all scenarios we consider, the bias and standard deviation are larger for smaller follow-up times when compared to longer ones. The difference in bias is not particularly pronounced but it is consistent across all the thresholds we consider. This runs counter to our expectations because, as discussed in section 6.5.3, we expect the cumulative product in the denominator to increase rapidly whenever the positivity assumption

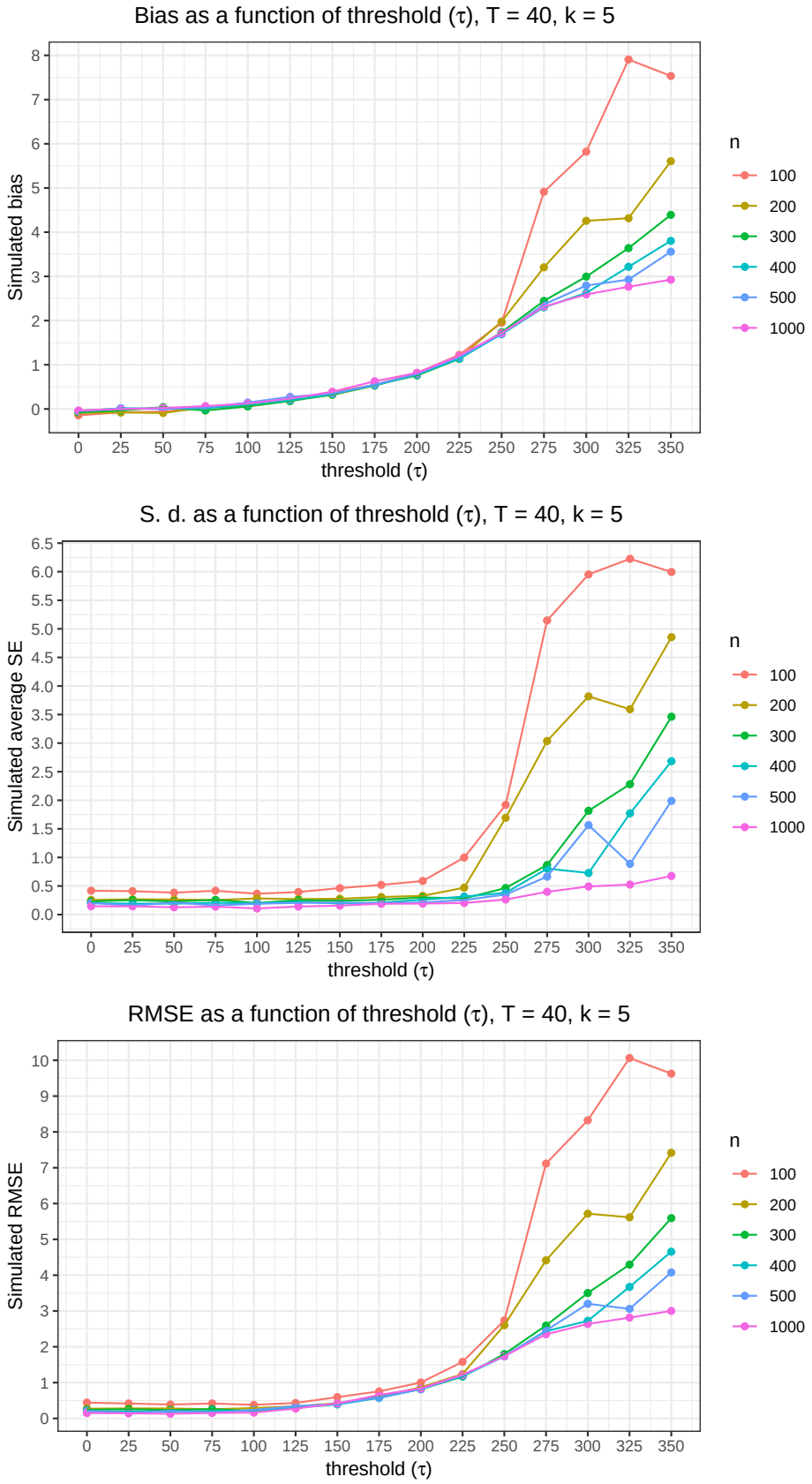


Figure 7.3: Monte Carlo bias, standard deviation (s.d.) and root mean squared error (RMSE) for γ_2 at various thresholds (τ) and sample sizes (n), $T = 40, k = 5$. $B = 100$ Monte Carlo repetitions per threshold.

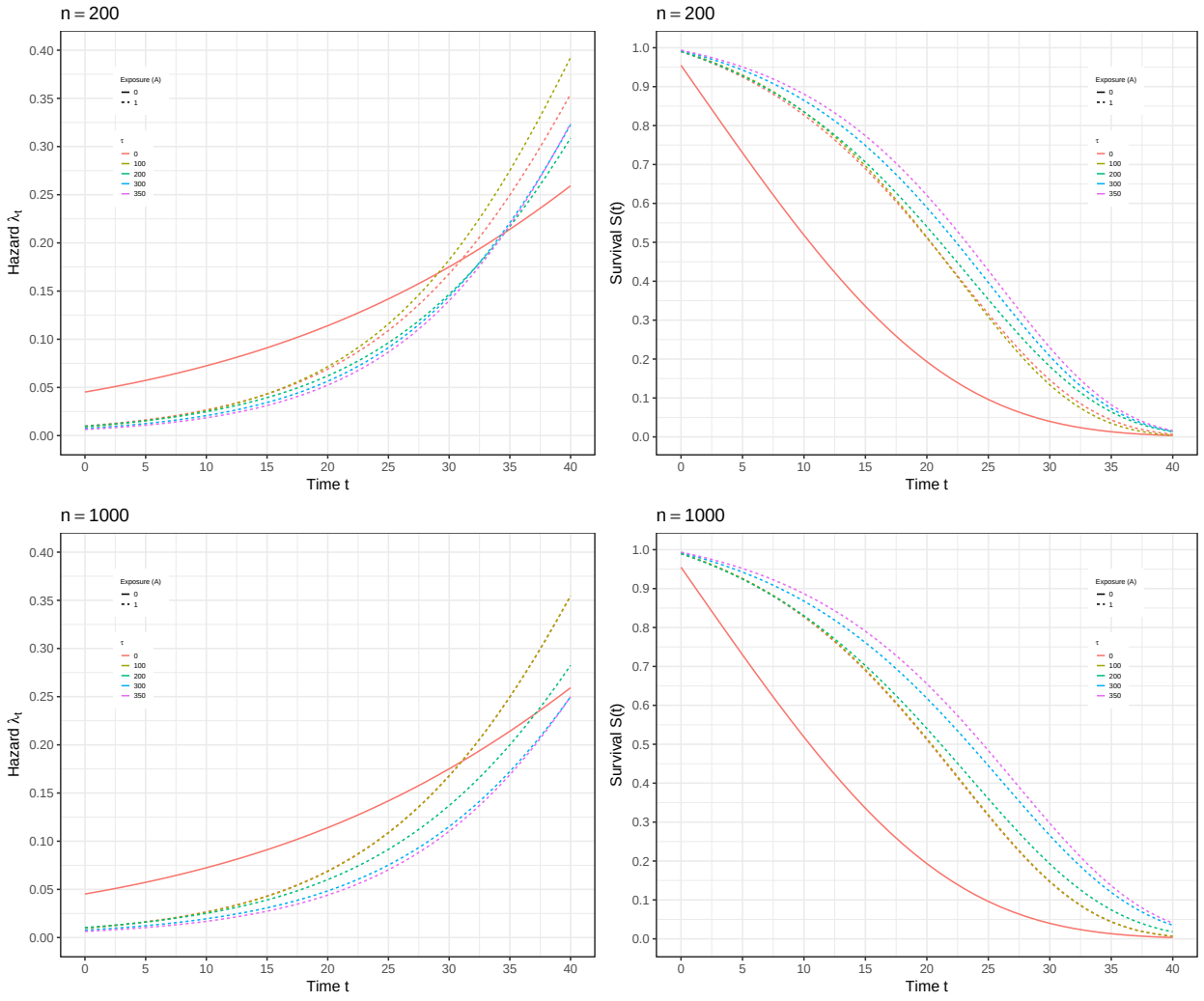


Figure 7.4: Structural hazard and survival curves under different thresholds (τ) for sample sizes of $n = \{200, 1000\}$ and $T = 20, k = 5$.

is violated. Longer time periods should, in theory, lead to larger weights and consequently greater bias and variability in the IPTW estimates.

However, in a survival context we need to consider the at risk population at each point in time. Our algorithm generates data which exhibits time dependent confounding. Low values of the confounder L_t are a predictor of subsequent exposure and of death. Subjects with a low CD4 count are therefore at greater risk of experiencing the event. Subjects with a low value of L_t also fall into the interval $[0, \tau]$ in which positivity violations are introduced. It follows that in studies with shorter follow-up times a higher proportion of subjects are likely to have confounder values in the interval $[0, \tau]$ whereas in longer follow-up times these subjects will either die, or, will have a higher value of L_t precisely because they have been exposed to treatment. Positivity violations impact the IPTW-estimator through the IPTW weights. In shorter studies we expect that a higher proportion of subjects have confounder values L_t that fall into the interval $[0, \tau]$. The impact of these weights is commensurately larger.

Table 7.3: Pooled logistic regression results for γ_2 from data simulated with positivity violations at different threshold levels at varying follow up times.

T	τ	$\hat{\gamma}_2$	s.d.	bias	RMSE
20	0	-1.489	0.124	0.011	0.125
20	100	-1.238	0.109	0.262	0.284
20	200	-0.400	0.146	1.100	1.109
20	300	1.553	0.606	3.053	3.112
20	350	4.108	5.003	5.608	7.515
30	0	-1.507	0.112	-0.007	0.112
30	100	-1.283	0.088	0.217	0.234
30	200	-0.489	0.133	1.011	1.020
30	300	1.509	0.522	3.009	3.054
30	350	2.146	0.798	3.646	3.732
40	0	-1.510	0.109	-0.010	0.109
40	100	-1.342	0.125	0.158	0.202
40	200	-0.553	0.136	0.947	0.956
40	300	1.367	0.427	2.867	2.899
40	350	1.700	0.694	3.200	3.275

7.2.4 Scenario 4: Near positivity violations and positivity compliant doctors.

The final simulation extends the analysis to the situation in which a certain proportion of doctors π are positivity compliant while the remaining proportion $1 - \pi$ are positivity non-compliant below a certain threshold τ. this situation is more appropriately viewed as near-violations of the positivity assumption because there are still some exposed and unexposed subjects at each level of the confounder L. The IPTW estimator may still be unstable because the IPTW estimates when there are only a small number of exposed subjects within a strata are likely to be very large.

Table 7.4 presents results for several values of π at thresholds of $\tau = \{100, 200, 350\}$. As expected, the lower the proportion of positivity compliant doctors π, the greater the bias in the results. Several interesting results emerge. First, near-violations of positivity lead to far less bias than the full violations considered up to this point. Still, if 50% of doctors are positivity non-compliant at a threshold of 200, substantial bias can be introduced ($\hat{\gamma}_2 = -1.261$ versus $\gamma_2 = -1.5$).

Second, the bias in the estimate of γ_2 was greater for a threshold of 200 than 350 almost as soon as the proportion of positivity compliant doctors exceeded zero. This is shown more clearly in figure 7.5 and runs counter to expectations because earlier simulations strongly suggest that higher thresholds induced greater bias. When near-violations of positivity are present, it is not necessarily the case that higher levels of

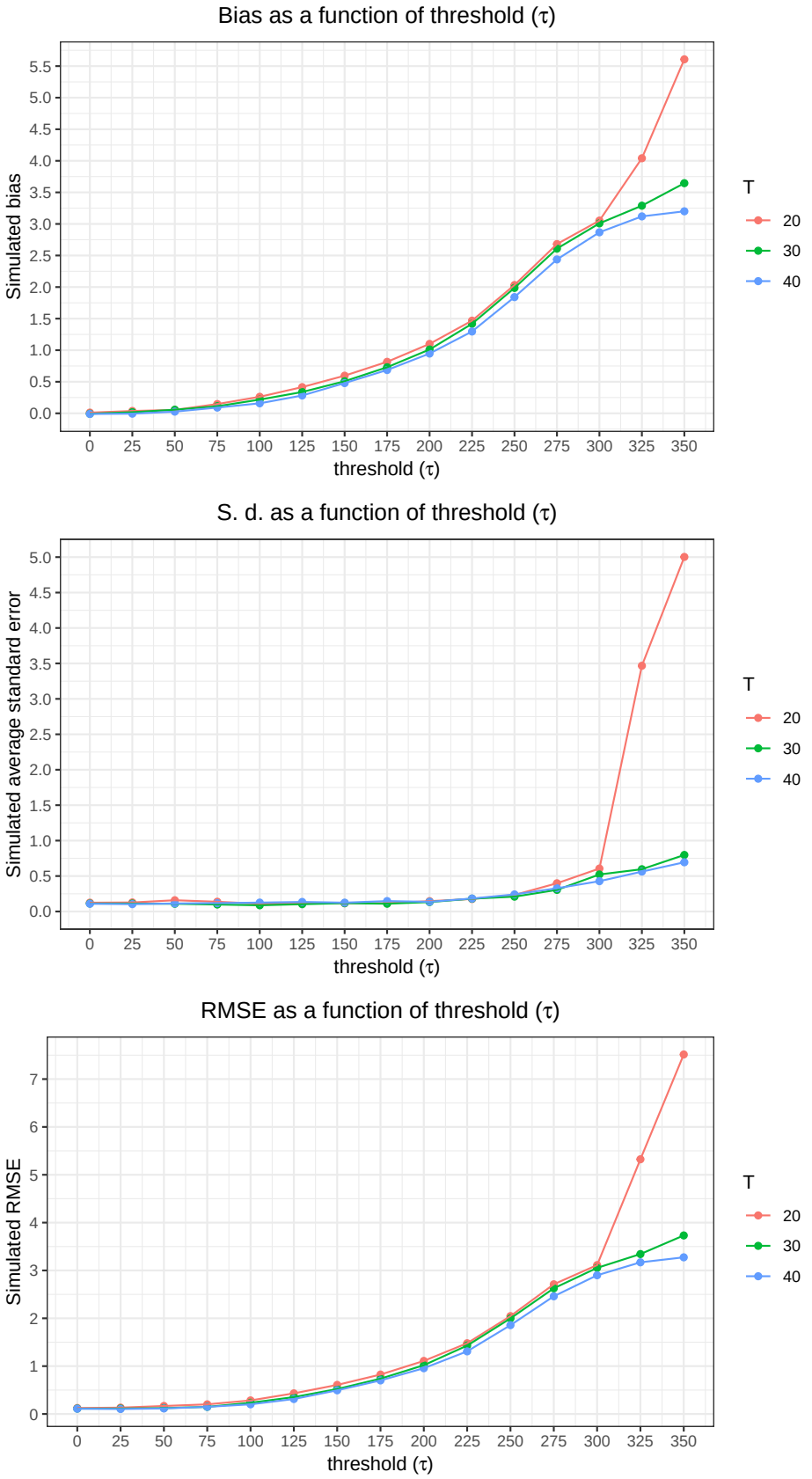


Figure 7.5: Monte Carlo bias, standard deviation (s.d.) and root mean squared error (RMSE) for γ_2 at various thresholds (τ) and follow up times (T). Visits at every five intervals $k = 5$. $B = 100$ Monte Carlo repetitions per threshold.

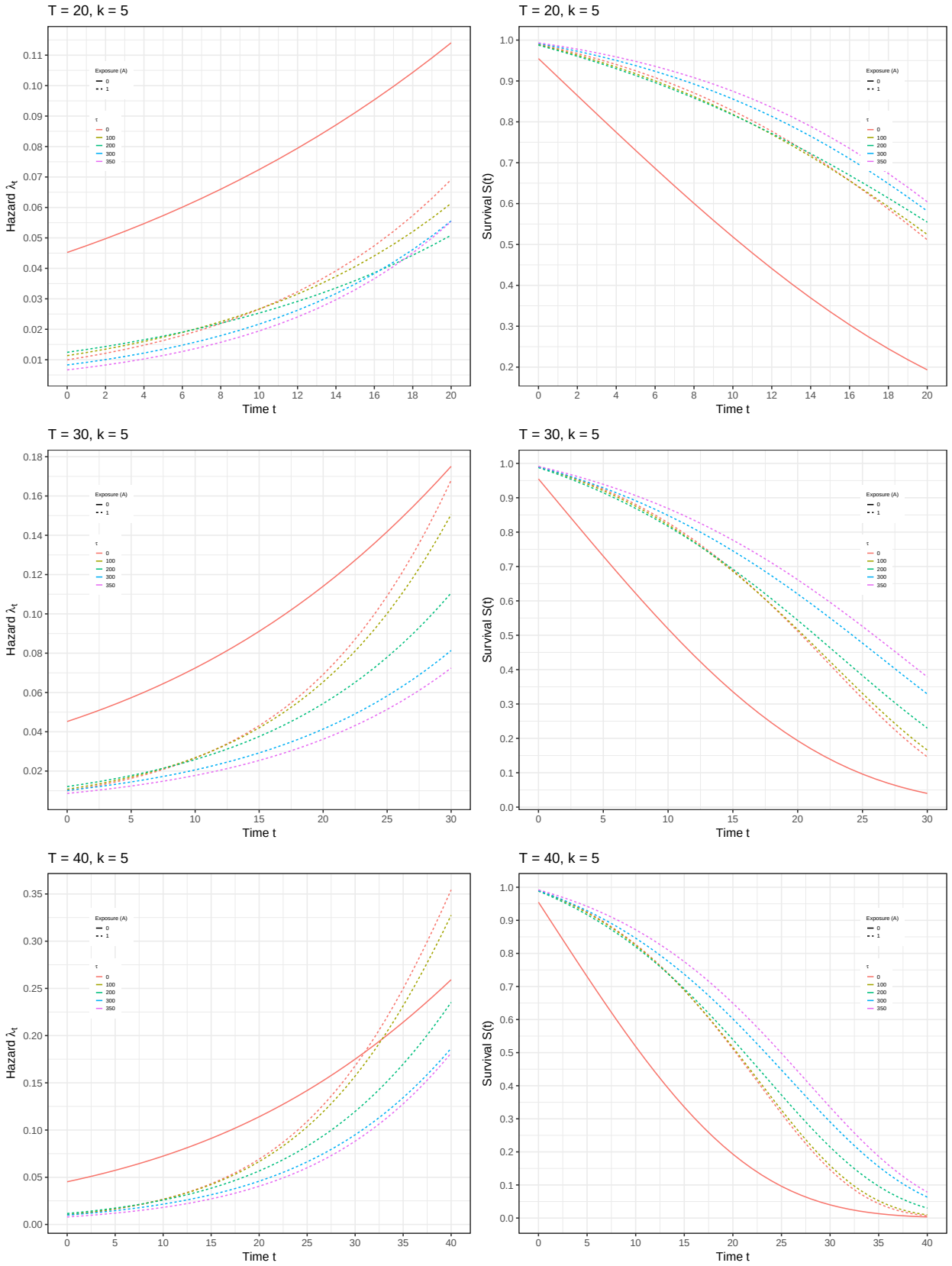


Figure 7.6: Structural hazard and survival curves under different thresholds (τ) and $T = \{20, 30, 40\}, k = 5$.

thresholds lead to higher bias or root MSE. All scenarios eventually converged to a small bias when $\pi = 1$, when no positivity violations are present.

Table 7.4: Pooled logistic regression results for γ_2 from data simulated with positivity non-compliant subjects at different threshold levels and sample sizes.

τ	π	$\hat{\gamma}_2$	s.d.	bias	RMSE
100	1	-1.541	0.129	-0.041	0.135
100	0.9	-1.555	0.150	-0.055	0.160
100	0.5	-1.454	0.134	0.046	0.142
200	1	-1.504	0.136	-0.004	0.136
200	0.9	-1.479	0.155	0.021	0.156
200	0.5	-1.261	0.160	0.239	0.287
300	1	-1.526	0.140	-0.026	0.142
300	0.9	-1.500	0.148	0.000	0.148
300	0.5	-1.306	0.176	0.194	0.263
350	1	-1.553	0.165	-0.053	0.174
350	0.9	-1.538	0.164	-0.038	0.169
350	0.5	-1.545	0.158	-0.045	0.165

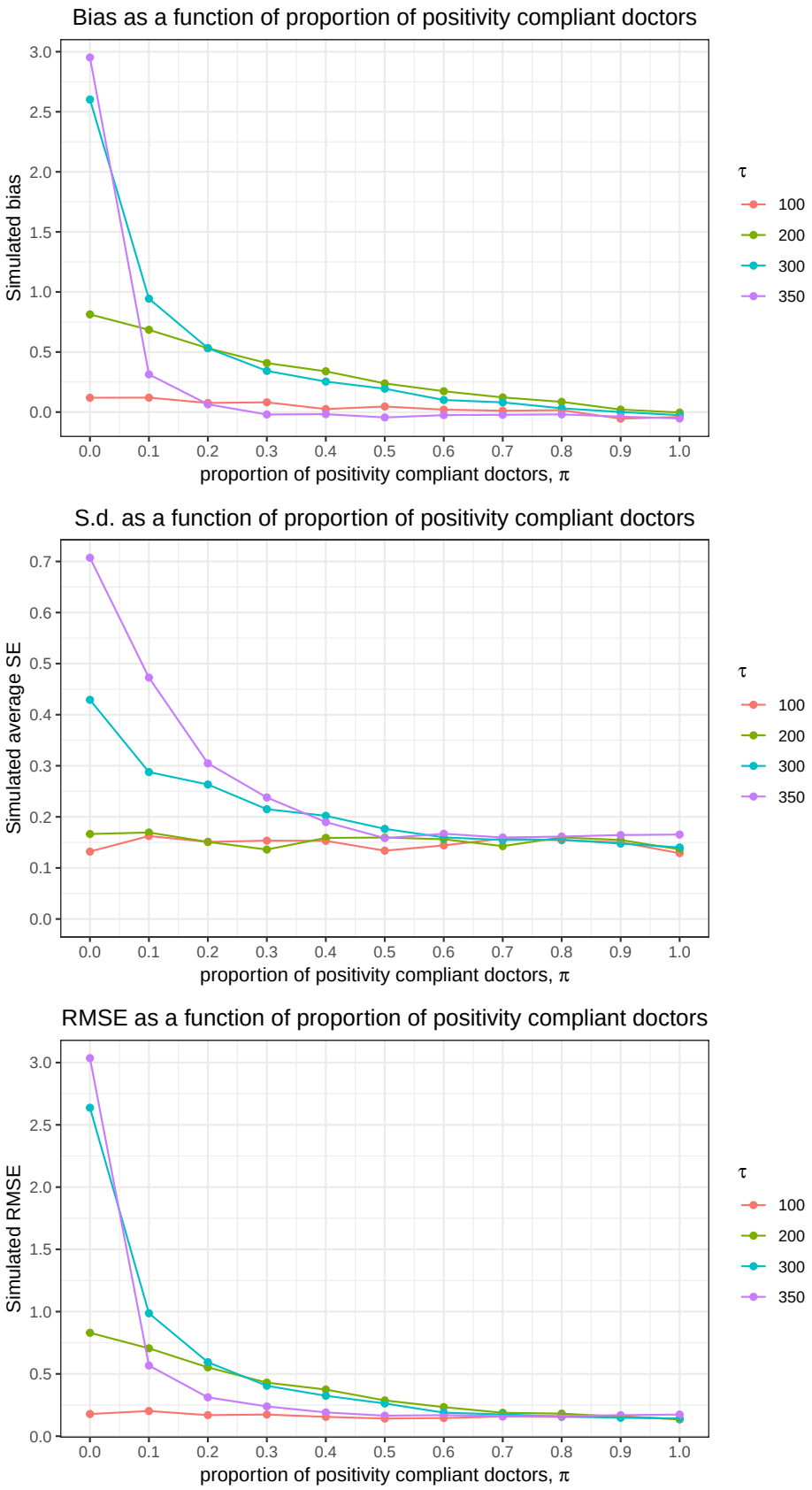


Figure 7.7: Monte Carlo bias, standard deviation (s.d.) and root mean squared error (RMSE) for γ_2 at various thresholds (τ) and proportions of positivity compliant subjects π, $T = 40, k = 5$. $B = 100$ Monte Carlo repetitions per threshold

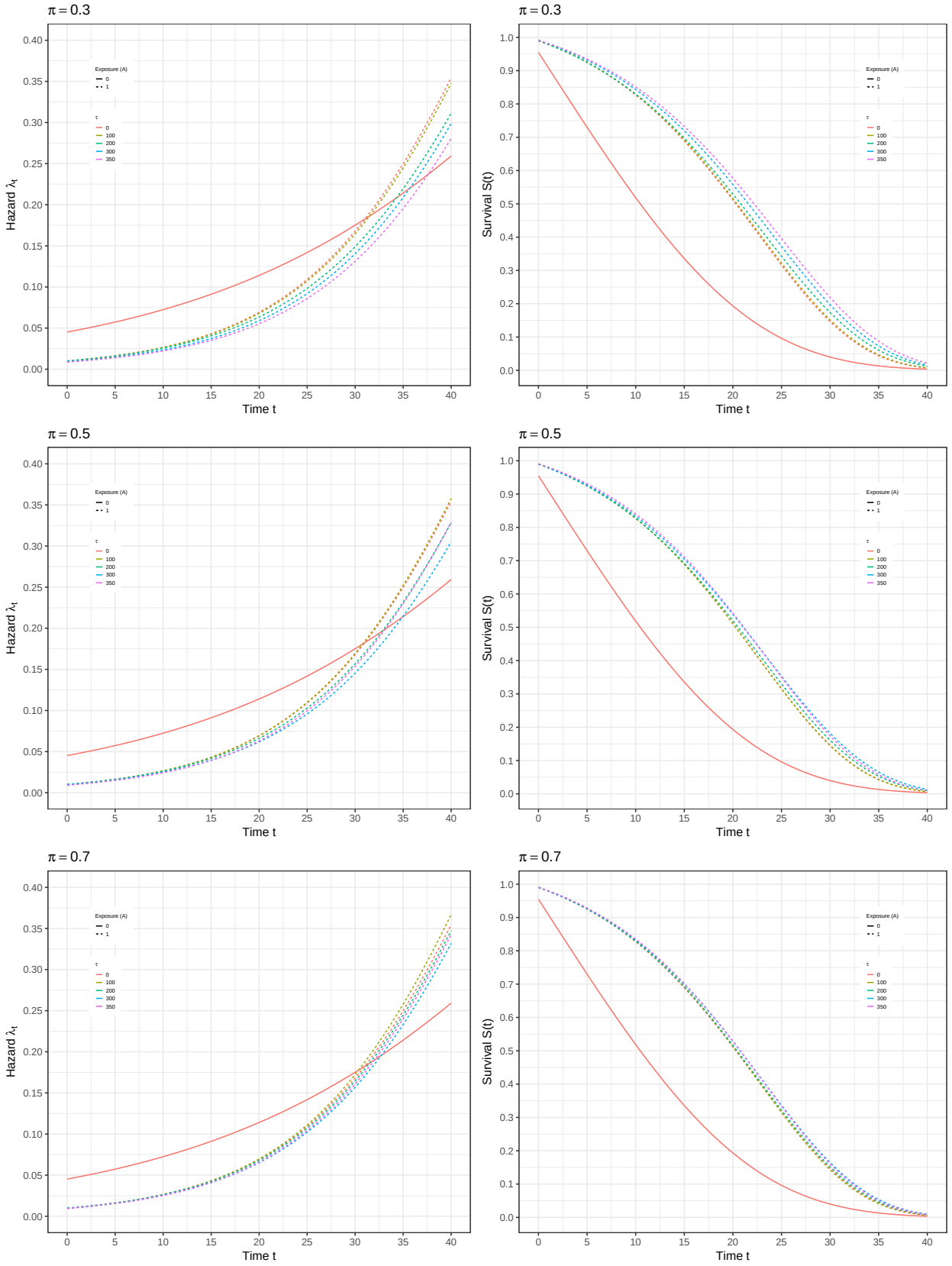


Figure 7.8: Structural hazard and survival curves under different thresholds (τ) for sample sizes of $n = \{200, 1000\}$ and $T = 20, k = 5$.

Chapter 8

Discussion and conclusion

This thesis has focused on violations of the positivity assumption in a survival context in which the complex issues surrounding time-dependent confounding arise. This work is relevant for several reasons. First, while it is well known that identification of causal effects is only valid under the positivity condition, the extent to which violations of the positivity assumption impact the IPTW-estimator is less well understood. Second, to our knowledge, no existing study has considered the effect of positivity violations on the IPTW-estimator in a realistic longitudinal or survival context. MSMs are an important class of models precisely because they allow for adjustment due to time dependent confounding. In spite of this, we know of only one study which considers a two period setting while most existing studies of the positivity assumption take place in a time-fixed or point treatment setting. This leaves open questions about how the IPTW estimator will perform in a realistic survival context under violations of the positivity assumption.

We extend an existing algorithm for simulating data from a specific MSM in which time-dependent confounding and selection bias are present. Our extensions to this algorithm make it possible to introduce positivity violations which are then propagated through time in a manner which is consistent with time-dependent data. Next we consider four specific scenarios which investigate how the performance of the IPTW estimator varies under different types of positivity violation. First we study the relationship between the bias and the size of the interval in which positivity violations arise. Next we check to see if the same behaviour is found at all sample sizes. We then consider more directly how positivity violations are propagated through time by varying the length of follow-up. Finally we move away from strict violations of the positivity assumption to consider near violations introduced when only a proportion of subjects experience positivity violations within some interval.

Our results suggest that even minor violations of the positivity assumption can have large implication for the performance of the IPTW-estimator when estimating survival models. More concretely, even when the interval within which the positivity violations occur is small, the bias of the IPTW-estimator can still be very large. On the other hand the variability of the estimates remained relatively consistent at increasing sizes of the interval in which positivity violations occur. Eventually, however, the variability of the estimates also increased substantially. Our results suggest that bias, as opposed to variability is the more dominant consideration when the positivity assumption is violated. We found little evidence that the bias due to positivity violations was affected by sample size. This matched our expectations because positivity violations undermine a fundamental condition of causal inference. Increasing the sample size will not reverse the effect of the violations. Contrary to expectations our results suggest that studies with shorter follow up times had a slightly higher bias than longer follow-up times, although the difference was not marked. Slightly more optimistically we find that when the proportion of positivity compliant subjects is high the bias is relatively minor. Overall our results suggests that, in a time-dependent survival context, analysts must take care to check for positivity violations which can occur at any point in time.

This work has limitations. We have focused on the case where positivity violations occur within a single interval of the confounder variable which is bounded below by zero. The idea was to mimic a situation where

medical guidelines or protocols are strictly enforced and exposure begins when values of the confounder fall below a certain threshold. Our work can be extended straightforwardly to accommodate more varied intervals in which positivity violations occur. Another limitation is that we consider only one continuous confounder. An interesting extension would be to consider two or more confounders in which the decision to initiate treatment is strictly enforced but based on more than one confounder. For example, when CD4 count is below 300 cells/mm^3 and viral load is greater than 10000 copies/ml. It is not immediately clear how the IPTW estimator would behave in this setting. This may also be of practical use because diagnosis of positivity violations is already difficult in a time-dependent setting because the violations can occur at any point in time. When more than one confounder is present it may be of practical use to analysts to have a methodological approach for judging the likely effects of the violations. However, we found that the base algorithm was fairly fragile in the sense that even minor changes to the original parameters could lead to unexpected results. For example, we experimented with different lengths of the interval between measuring times. We found that smaller intervals led to biased IPTW estimations of the true parameters even when all IPTW assumptions were met. Extending this algorithm to new contexts would require very careful adaptation. In spite of these limitations our results do provide a first step toward broadening the understanding of positivity violations when time dependent confounding is present and the outcome of interest is survival.

A second extension would be to use the work in this thesis to investigate the trade-off between confounding bias and positivity violations. In some cases it may be possible to extrapolate regions of nonpositivty by categorizing a continuous confounder, or re-categorizing a categorical confounder, so that the interval in which positivity violations occur is subsumed by the new categories. For example, if violations of the positivity assumption are present in the interval $[0, 100]$ then defining a categorical confounder with a level which covers the interval $[0, 200]$ would avoid the positivity violation by extrapolation. However, doing so comes at the cost of increasing confounding bias whenever categorization results in residual confounding. In our approach investigating the trade-off would work as follows. First data would be generated with violations of the positivity assumption within a predefined interval. Next, we categorize the confounder L to cover a region larger than that interval. Finally we evaluate the performance of of the IPTW-estimator and judge whether the loss or gain in performance justifies the decision to categorize the confounder in the first place. To our knowledge, no methodological approach for evaluating the trade-off between confounding bias and positivity bias has been considered in the literature (although the need for one has been identified [46][2]).

We have restricted our attention to the IPTW estimator but several other techniques for estimating causal effects from causal data also exist. Future work might consider comparing the IPTW-estimator with the doubly robust estimator or the G-computation under positivity violations. Finally, we can use our approach to weigh the methods that have been suggested for dealing with the extreme weights caused by positivity violations. For example, we can generate data in which the positivity assumption is violated and then compare the estimates after truncating the weights or removing observations with extreme weights altogether. As the effects of these techniques are fairly ad-hoc and not well understood it is valuable to have an approach through which they can be more formally studied.

8.1 Software

All simulations and analysis carried out in this thesis use the R programming language. The code written to produce the results reported in this thesis are attached as appendices. Appendix A contains the functions used to generate the data. Appendix B contains the code to reproduce the results reported in chapter 7 including all Monte Carlo experiments. Appendix C contains code to create the tables and figures in this thesis.

We use a random number generator in order to generate the data used in this thesis. This random number generator is based on a Mersenne-Twister and all functions to generate random numbers are part of the base R programming language.

Chapter 9

Appendix A: R code for simulation functions

```r
# define inverse logit (expit) function
expit <- function(x) return(exp(x)/(1 + exp(x)))

# Simulaton function Havercroft Didelez (2012) algorithm.
sim <- function(K, k, gam, theta, id){

  # define lists for holding A, L, U and Y
  A <- numeric(K) # treatment
  A_1 <- numeric(K) # lagged treatment
  L <- numeric(K) # CD4 count
  U <- numeric(K) # general health
  Y <- numeric(K) # outcome
  hazard <- numeric(K)
  surv <- numeric(K)

  # set time of starting treatment to zero initially
  Ts = 0

  # set the initial values of U, U[0], to a
  # randomly generated value from a uniform
  # distribution a measure of general health
  U[1] <- runif(1)
  A_1[1] = 0
  Y[1] <- 0
  L[1] = max(0, qgamma(U[1], shape=3, scale=154) + rnorm(1, 0, 20))

  # set A[1]
  A[1] = rbinom(n = 1, size = 1, prob = expit(theta[1] + theta[3] * (L[1] - 500)))

  # initial value of hazard
  hazard[1] = expit(gam[1] + gam[3] * A[1])
  surv[1] = 1 - hazard[1]

  #
```

```
if(surv[1] <= (1 - U[1])){ # check this
  Y[1] = 1
  Tsurv = 0
}

# loop over all time periods.
for (t in 2:K) {
  if (Y[t-1] == 1) {
    break
  }
  A_1[t] = A[t-1]
  U[t] = min(1, max(0, U[t-1] + rnorm(1,0,0.05)))
  if ((t-1) %% k == 0) {
    L[t] <- max(L[t-1] + 150*A[t-k]*(1 - A_1[t-k]) + rnorm(1, 100*(U[t] - 2), 50), 0)
    A[t] <- (1 - A[t-1])*rbinom(n = 1, size = 1, prob = expit(theta[1] + theta[2] * (t-1) + theta[3] =
  }
  else {
    L[t] <- L[t-1]
    A[t] <- A[t-1]
  }

  # if treatment began between the previous and current treatment set Ts <- t-1
  if (A[t-1] == 0 && A[t] == 1) {
    Ts <- t-1
  }
  hazard[t] <- 1/(1 + exp(-(gam[1] + ((1 - A[t])*(t-1) + (A[t])*Ts)*gam[2] + gam[3]*(A[t]) + (A[t])*(
  surv[t] = surv[t-1] * (1 - hazard[t])

  if (surv[t] <= (1 - U[1])) {
    Y[t] <- 1
  }
  else {
    Y[t] <- 0
  }
}

# if they died in the last period then we want the last period time,
# otherwise we want the current time period.
cut <- ifelse(Y[t-1] == 1, t-1, t)

Y <- Y[1:cut]
A <- A[1:cut]
A_1 <- A_1[1:cut]
L <- L[1:cut]
hazard <- hazard[1:cut]
surv <- surv[1:cut]

# visit is up to the last time point, i.e when the patient dies, or end of follow-up
visit <- seq(from = 0, to = cut-1)

# make variables for regression
d1 = (1-A) * visit + A * Ts
d3 = A * (visit - Ts)
```

```
  L500 <- L - 500

  # put all variables in one dataframe
  df <- data.frame(id, visit, Y, A, A_1, L, L500, d1, d3, hazard, surv)

  # return dataframe
  df
}

# simulate code with positivity violation below a threshold
sim_pos1 <- function(K, k, gam, theta, id, lower){

  # define lists for holding A, L, U and Y
  A <- numeric(K) # treatment
  A_1 <- numeric(K) # lagged treatment
  L <- numeric(K) # CD4 count
  U <- numeric(K) #
  Y <- numeric(K) # outcome
  hazard <- numeric(K)
  surv <- numeric(K)

  # set time of starting treatment to zero initially
  Ts = 0

  # set the initial values of U, U[0], to a
  # randomly generated value from a uniform
  # distribution a measure of general health
  U[1] <- runif(1)
  A_1[1] = 0
  L[1] = max(0, qgamma(U[1], shape=3, scale=154) + rnorm(1, 0, 20))

  # set A[1]
  if (L[1] < lower){
    A[1] <- 1
  } else {
    A[1] = rbinom(n = 1, size = 1, prob = expit(theta[1] + theta[3] * (L[1] - 500)))
  }

  # initial value of lambda
  hazard[1] = expit(gam[1] + gam[3] * A[1])
  surv[1] = 1 - hazard[1] # check this

  #
  if(surv[1] <= (1 - U[1])){ # check this
    Y[1] = 1
  }

  # loop over all time periods.
  for (t in 2:K) {
    if (Y[t-1] == 1) {
      break
    }
    A_1[t] = A[t-1]
```

```
  U[t] = min(1, max(0, U[t-1] + rnorm(1,0,0.05)))
  if ((t-1) %% k == 0) {
    L[t] <- max(L[t-1] + 150*A[t-k]*(1 - A_1[t-k]) + rnorm(1, 100*(U[t] - 2), 50), 0)
    if (L[t] < lower){
      A[t] <- 1
    } else {
      A[t] <- (1 - A[t-1])*rbinom(n = 1, size = 1, prob = expit(theta[1] + theta[2] * (t-1) + theta[3]
    }
  }
  else {
    L[t] <- L[t-1]
    A[t] <- A[t-1]
  }

  # if treatment began between the previous and current treatment set Ts <- t-1
  if (A[t-1] == 0 && A[t] == 1) {
    Ts <- t-1
  }
  hazard[t] <- 1/(1 + exp(-(gam[1] + ((1 - A[t])*(t-1) + (A[t])*Ts)*gam[2] + gam[3]*(A[t]) + (A[t])*(
  surv[t] = surv[t-1] * (1 - hazard[t])

  if (surv[t] <= (1 - U[1])) {
    Y[t] <- 1
  }
  else {
    Y[t] <- 0
  }
}

# if they died in the last period then we want the last period time,
# otherwise we want the current time period.
cut <- ifelse(Y[t-1] == 1, t-1, t)

Y <- Y[1:cut]
A <- A[1:cut]
A_1 <- A_1[1:cut]
L <- L[1:cut]
hazard <- hazard[1:cut]
surv <- surv[1:cut]

# visit is up to the last time point, i.e when the patient dies, or end of follow-up
visit <- seq(from = 0, to = cut-1)

# make variables for regression
d1 = (1-A) * visit + A * Ts
d3 = A * (visit - Ts)
L500 <- L - 500

# put all variables in one dataframe
df <- data.frame(id, visit, Y, A, A_1, L, L500, d1, d3, hazard, surv)

# return dataframe
df
```

```
}

#
# simulate code with positivity violation below a threshold
sim_pos2 <- function(K, k, gam, theta, id, lower, prop){

  # define lists for holding A, L, U and Y
  A <- numeric(K) # treatment
  A_1 <- numeric(K) # lagged treatment
  L <- numeric(K) # CD4 count
  U <- numeric(K) #
  Y <- numeric(K) # outcome
  hazard <- numeric(K)
  surv <- numeric(K)

  # set time of starting treatment to zero initially
  Ts = 0

  # set the initial values of U, U[0], to a
  # randomly generated value from a uniform
  # distribution a measure of general health
  U[1] <- runif(1)
  A_1[1] = 0
  L[1] = max(0, qgamma(U[1], shape=3, scale=154) + rnorm(1, 0, 20))

  # set A[1] - decision is based on whether the doctor is positivity compliant for this patient
  uni <- runif(1)
  if(prop > uni){
    A[1] <- rbinom(n = 1, size = 1, prob = expit(theta[1] + theta[3] * (L[1] - 500)))
  } else {
    if(L[1] < lower){
      A[1] <- 1
    } else {
      A[1] <- rbinom(n = 1, size = 1, prob = expit(theta[1] + theta[3] * (L[1] - 500)))
    }
  }

  # initial value of lambda
  hazard[1] = expit(gam[1] + gam[3] * A[1])
  surv[1] = 1 - hazard[1] # check this

  #
  if(surv[1] <= (1 - U[1])){ # check this
    Y[1] = 1
  }

  # loop over all time periods.
  for (t in 2:K) {
    if (Y[t-1] == 1) {
      break
    }
    A_1[t] = A[t-1]
    U[t] = min(1, max(0, U[t-1] + rnorm(1,0,0.05)))
```

```
  if ((t-1) %% k == 0) {
    L[t] <- max(L[t-1] + 150*A[t-k]*(1 - A_1[t-k]) + rnorm(1, 100*(U[t] - 2), 50), 0)
    if(prop > uni){
      A[t] <- (1 - A[t-1])*rbinom(n = 1, size = 1, prob = expit(theta[1] + theta[2] * (t-1) + theta[3]
    } else {
      if(L[t] < lower){
        A[t] <- 1
      } else {
        A[t] <- (1 - A[t-1])*rbinom(n = 1, size = 1, prob = expit(theta[1] + theta[2] * (t-1) + theta
      }
    }
  }
  else {
    L[t] <- L[t-1]
    A[t] <- A[t-1]
  }

  # if treatment began between the previous and current treatment set Ts <- t-1
  if (A[t-1] == 0 && A[t] == 1) {
    Ts <- t-1
  }
  hazard[t] <- 1/(1 + exp(-(gam[1] + ((1 - A[t])*(t-1) + (A[t])*Ts)*gam[2] + gam[3]*(A[t]) + (A[t])*(
  surv[t] = surv[t-1] * (1 - hazard[t])

  if (surv[t] <= (1 - U[1])) {
    Y[t] <- 1
  }
  else {
    Y[t] <- 0
  }
}

# if they died in the last period then we want the last period time,
# otherwise we want the current time period.
cut <- ifelse(Y[t-1] == 1, t-1, t)

Y <- Y[1:cut]
A <- A[1:cut]
A_1 <- A_1[1:cut]
L <- L[1:cut]
hazard <- hazard[1:cut]
surv <- surv[1:cut]

# visit is up to the last time point, i.e when the patient dies, or end of follow-up
visit <- seq(from = 0, to = cut-1)

# make variables for regression
d1 = (1-A) * visit + A * Ts
d3 = A * (visit - Ts)
L500 <- L - 500

# put all variables in one dataframe
df <- data.frame(id, visit, Y, A, A_1, L, L500, d1, d3, hazard, surv)
```

```
  # return dataframe
  df
}

# get IPTW weights
get_weights <- function(data, k) {
  # data: data simulated from one of three above functions
  # k: time interval between which exposure and confounder can be updated.

  # model for denominator (includes L)
  den <- glm(A ~ visit + L500, family=binomial, subset=(visit %% k == 0 & A_1==0), data=data)
  # model for numerator (no L)
  num <- glm(A ~ visit, family=binomial, subset=(visit %% k == 0 & A_1==0), data=data)
  treat.prob.den <- rep(1, dim(data)[1])
  treat.prob.num <- rep(1, dim(data)[1])
  treat.prob.den[data$visit %% k == 0 & data$A_1==0]  <- ifelse(data$A[data$visit %% k == 0 & data$A_1==
  treat.prob.num[data$visit %% k == 0 & data$A_1==0]  <- ifelse(data$A[data$visit %% k == 0 & data$A_1==
  treat.weights <- unlist(tapply(treat.prob.num/treat.prob.den, data$id, cumprod))
  data <- cbind(data, estprob=treat.prob.den, ip.weights=treat.weights)
  return(data)
}

# simulate data for n subjects
sim_n <- function(n, K, k, gam, theta, lower){
  get_weights(bind_rows(lapply(1:n, function(i) sim(K = K, k = k, gam = gam, theta = theta, id = i))), l
}

# simulate data for n subjects - posit violations 1
sim_n_pos1 <- function(n, K, k, gam, theta, lower){
  get_weights(bind_rows(lapply(1:n, function(i) sim_pos1(K = K, k = k, gam = gam, theta = theta, id = i
}

# simulate data for n subjects - posit violations 2
sim_n_pos2 <- function(n, K, k, gam, theta, lower, prop){
  get_weights(bind_rows(lapply(1:n, function(i) sim_pos2(K = K, k = k, gam = gam, theta = theta, id = i
}

# monte carlo function for B repetitions - base algorithm
mc <- function(B, n, K, k, gam, theta){
  m = matrix(ncol = length(gam), nrow = B)
  for(rep in 1:B){
    df <- sim_n(n, K, k, gam, theta)
    m[rep, ] <- coef(glm(Y ~ d1 + A + d3, family=quasibinomial, data=df, weights=ip.weights))
  }
  m
}

# monte carlo function for B repetitions - strict violations
mc_pos1 <- function(B, n, K, k, gam, theta, lower){
  m = matrix(ncol = length(gam), nrow = B)
  for(rep in 1:B){
    df <- sim_n_pos1(n, K, k, gam, theta, lower = lower)
    m[rep, ] <- coef(glm(Y ~ d1 + A + d3, family=quasibinomial, data=df, weights=ip.weights))
```

```
  }
  m
}

# monte carlo function for B repetitions - near violations
mc_pos2 <- function(B, n, K, k, gam, theta, lower, prop){
  m = matrix(ncol = length(gam), nrow = B)
  for(rep in 1:B){
    df <- sim_n_pos2(n, K, k, gam, theta, lower = lower, prop = prop)
    m[rep, ] <- coef(glm(Y ~ d1 + A + d3, family=quasibinomial, data=df, weights=ip.weights))
  }
  m
}
```

Chapter 10

Appendix B: R code for replicating simulation study

```
# clear workspace
rm(list=ls())

# get packages
library(dplyr)
library(ggplot2)

# source in functions
source("C:/Users/Tom/OneDrive/Documents/leiden/thesis/thesis_sim_functions.R")

# ============================================================================
# model parameters for simulation
theta=c(-0.405,0.0205,-0.00405)
gam=c(-3,0.05,-1.5,0.1)

# ============================================================================
# Experiment 1: effect of increasing threshold on bias.

# MC regression results with a few fixed violations
df0 = mc_pos1(B = 100, n = 1000, K = 40, k = 5, gam=gam, theta = theta, lower = 0)
df100 = mc_pos1(B = 100, n = 1000, K = 40, k = 5, gam=gam, theta = theta, lower = 100)
df200 = mc_pos1(B = 100, n = 1000, K = 40, k = 5, gam=gam, theta = theta, lower = 200)
df300 = mc_pos1(B = 100, n = 1000, K = 40, k = 5, gam=gam, theta = theta, lower = 300)
df350 = mc_pos1(B = 100, n = 1000, K = 40, k = 5, gam=gam, theta = theta, lower = 350)

# combine results into one dataframe and save for later.
df <- as.data.frame(cbind(true = gam,
            paste0(sprintf("%.3f", round(colMeans(df0),3)), " (", sprintf("%.3f", round(apply(df0, 2, s
            paste0(sprintf("%.3f", round(colMeans(df100),3)), " (", sprintf("%.3f", round(apply(df100, 1
            paste0(sprintf("%.3f", round(colMeans(df200),3)), " (",sprintf("%.3f", round(apply(df200, 2
            paste0(sprintf("%.3f", round(colMeans(df300),3)), " (",sprintf("%.3f", round(apply(df300, 2
            paste0(sprintf("%.3f", round(colMeans(df350),3)), " (",sprintf("%.3f", round(apply(df350, 2

# write to file for use later
```

```r
write.csv(df,
          file = "C:/Users/Tom/OneDrive/Documents/leiden/thesis/exp1_reg_mc.csv",
          row.names = FALSE)

# ------------------------------------------------------------------------------
# loop over thresholds to compare bias, sd and rMSE

#specify thresholds at which positivty violations will be introduced.
threshold <- seq(0, 350, by = 25)

# matrix for storing the mean of each parameter (which we need for constructig survival
# and hazard curves) and the sd of gamma 2 which is the parameter most closely
# associated with exposure A
tm <- as.data.frame(matrix(nrow = length(threshold), ncol = 5))
names(tm) <- c("pmean_gam0", "pmean_gam1", "pmean_gam2", "pmean_gam3", "psd_gam2")

# loop through thresholds and record mc results per threshold.
for(i in 1:length(threshold)){

  # get iteration
  print(threshold[i])

  # monte carlo results with B reps
  tmp <- mc_pos1(B = 100, n = 1000, K = 40, k = 5, gam=gam, theta = theta, lower = threshold[i])

  # record the mean of each parameter and the sd of gamma 2
  tm[i, c("pmean_gam0", "pmean_gam1", "pmean_gam2", "pmean_gam3")] <- colMeans(tmp)
  tm[i, "psd_gam2"] <- sd(tmp[, 3])
}

# calculate bias, and root MSE
tm$bias <- tm$pmean_gam2 - (-1.50)
tm$sim_mse <- tm$bias^2 + tm$psd_gam2^2
tm$sim_rmse <- sqrt(tm$sim_mse)
tm$threshold <- threshold

# write to file for use later
write.csv(tm,
          file = "C:/Users/Tom/OneDrive/Documents/leiden/thesis/exp1.csv",
          row.names = FALSE)

# plot of the bias against the threshold
ggplot(data = tm, aes(x = threshold, y = psd_gam2)) +
  geom_point() +
  geom_line() +
  theme_bw() +
  labs(title = "Bias as a function of threshold, K = 40, k = 5") +
  theme(plot.title = element_text(hjust = 0.5))

# ==============================================================================
# Experiment 2: Effect of thresholds at different time frames K = 20, 30, 40, 50
# one full run for a single K takes ~ 30 mins -> 1.5 hours in total
threshold <- seq(0, 350, by = 25)
```

```r
K <- 20
#gam20 = c(-1, 0.05, -3, 0.25)

tm20 <- as.data.frame(matrix(nrow = length(threshold), ncol = 8))
names(tm20) <- c("pmean_gam0", "pmean_gam1", "pmean_gam2", "pmean_gam3",
                 "psd_gam0", "psd_gam1",  "psd_gam2",  "psd_gam3")
for(i in 1:length(threshold)){
  print(threshold[i])
  tmp <- mc_pos1(B = 100, n = 1000, K = K, k = 5, gam=gam, theta = theta, lower = threshold[i])
  # record the mean of each parameter and the sd of gamma 2
  tm20[i, c("pmean_gam0", "pmean_gam1", "pmean_gam2", "pmean_gam3")] <- colMeans(tmp)
  tm20[i, c("psd_gam0", "psd_gam1",  "psd_gam2",  "psd_gam3")] <- apply(tmp, 2, sd)
}

# combine in dataframe
tm20$bias <- tm20$pmean_gam2 - (gam[3])
tm20$pbias <- (tm20$pmean_gam2 - (gam[3]))/(gam[3]) * 100
tm20$apbias <- abs(tm20$pbias)
tm20$sim_mse <- tm20$bias^2 + tm20$psd_gam2^2
tm20$sim_rmse <- sqrt(tm20$sim_mse)
tm20$threshold <- threshold
tm20$K <- 20

# K = 30
threshold <- seq(0, 350, by = 25)
K <- 30
#gam30=c(-2,0.05,-2.25,0.15)

tm30 <- as.data.frame(matrix(nrow = length(threshold), ncol = 8))
names(tm30) <- c("pmean_gam0", "pmean_gam1", "pmean_gam2", "pmean_gam3",
                 "psd_gam0", "psd_gam1",  "psd_gam2",  "psd_gam3")
for(i in 1:length(threshold)){
  print(threshold[i])
  tmp <- mc_pos1(B = 100, n = 1000, K = K, k = 5, gam=gam, theta = theta, lower = threshold[i])
  # record the mean of each parameter and the sd of gamma 2
  tm30[i, c("pmean_gam0", "pmean_gam1", "pmean_gam2", "pmean_gam3")] <- colMeans(tmp)
  tm30[i, c("psd_gam0", "psd_gam1",  "psd_gam2",  "psd_gam3")] <- apply(tmp, 2, sd)
}

# combine in dataframe
tm30$bias <- tm30$pmean_gam2 - (gam[3])
tm30$pbias <- (tm30$pmean_gam2 - (gam[3]))/(gam[3]) * 100
tm30$apbias <- abs(tm30$pbias)
tm30$sim_mse <- tm30$bias^2 + tm30$psd_gam2^2
tm30$sim_rmse <- sqrt(tm30$sim_mse)
tm30$threshold <- threshold
tm30$K <- 30

# K = 40
threshold <- seq(0, 350, by = 25)
K <- 40
#gam40=c(-3,0.05,-1.5,0.1)
```

```
tm40 <- as.data.frame(matrix(nrow = length(threshold), ncol = 8))
names(tm40) <- c("pmean_gam0", "pmean_gam1", "pmean_gam2", "pmean_gam3",
                 "psd_gam0", "psd_gam1",  "psd_gam2",  "psd_gam3")
for(i in 1:length(threshold)){
  print(threshold[i])
  tmp <- mc_pos1(B = 100, n = 1000, K = K, k = 5, gam=gam, theta = theta, lower = threshold[i])
  # record the mean of each parameter and the sd of gamma 2
  tm40[i, c("pmean_gam0", "pmean_gam1", "pmean_gam2", "pmean_gam3")] <- colMeans(tmp)
  tm40[i, c("psd_gam0", "psd_gam1",  "psd_gam2",  "psd_gam3")] <- apply(tmp, 2, sd)
}

# combine in dataframe
tm40$bias <- tm40$pmean_gam2 - (gam[3])
tm40$pbias <- (tm40$pmean_gam2 - (gam[3]))/(gam[3]) * 100
tm40$apbias <- abs(tm40$pbias)
tm40$sim_mse <- tm40$bias^2 + tm40$psd_gam2^2
tm40$sim_rmse <- sqrt(tm40$sim_mse)
tm40$threshold <- threshold
tm40$K <- 40

# combine into one dataframe for plotting
df <- rbind(tm20, tm30, tm40)

# write to file for use later
write.csv(df,
          file = "C:/Users/Tom/OneDrive/Documents/leiden/thesis/exp2.csv",
          row.names = FALSE)

# plot of the rMSE against the threshold
ggplot(data = df, aes(x = threshold, y = psd_gam2,
                      color = factor(K))) +
  geom_point() +
  geom_line() +
  theme_bw() +
  guides(fill=guide_legend(title="Follow-up time")) +
  labs(title = "Bias as a function of threshold, K = 40, k = 1") +
  theme(plot.title = element_text(hjust = 0.5))

# ============================================================================
# Experiment 3, varying the sample size to see finite sample properties.
# how does positivity affect the results in smaller samples comparing
# a sample size of 200 with one of 1000
threshold <- seq(0, 350, by = 25)
ns <- c(100, 200, 300, 400, 500, 1000)
L <- list()

for(n in ns){
  tm <- as.data.frame(matrix(nrow = length(threshold), ncol = 5))
  names(tm) <- c("pmean_gam0", "pmean_gam1", "pmean_gam2", "pmean_gam3", "psd_gam2")
  for(i in 1:length(threshold)){
    print(threshold[i])
    tmp <- mc_pos1(B = 100, n = n, K = 40, k = 5, gam=gam, theta = theta, lower = threshold[i])
    # record the mean of each parameter and the sd of gamma 2
```

```
    tm[i, c("pmean_gam0", "pmean_gam1", "pmean_gam2", "pmean_gam3")] <- colMeans(tmp)
    tm[i, "psd_gam2"] <- sd(tmp[, 3])
  }
  L[[as.character(n)]] <- cbind(n, threshold, tm)
}

# combine in dataframe
df <- as.data.frame(do.call(rbind, L))
df$bias <- df$pmean_gam2 - (-1.50)
df$sim_mse <- df$bias^2 + df$psd_gam2^2
df$sim_rmse <- sqrt(df$sim_mse)

# write to file for use later
write.csv(df,
          file = "C:/Users/Tom/OneDrive/Documents/leiden/thesis/exp3.csv",
          row.names = FALSE)

# plot of the rMSE against the threshold
ggplot(data = df, aes(x = threshold, y = sim_rmse,
                      color = factor(n))) +
  geom_point() +
  geom_line() +
  theme_bw() +
  guides(fill=guide_legend(title="Follow-up time")) +
  labs(title = "Bias as a function of threshold, K = 40, k = 5") +
  theme(plot.title = element_text(hjust = 0.5))


# ============================================================================
# Experiment 4: Effect of different proportions of positivity compliant doctors.
# for thresholds 100, 200, 300, 400, 500 all shown on same curve.

# ----------------------------------------------------------------------------
# loop over thresholds to compare bias, sd and rMSE
props <- seq(1, 0, by = -0.1)
thresholds <- c(100, 200, 300, 350)
L <- list()

# loop through thresholds and record mc results per proportion.
for(threshold in thresholds){
  tm <- as.data.frame(matrix(nrow = length(threshold), ncol = 5))
  names(tm) <- c("pmean_gam0", "pmean_gam1", "pmean_gam2", "pmean_gam3", "psd_gam2")
  for(i in 1:length(props)){
    print(props[i])
    tmp <- mc_pos2(B = 100, n = 1000, K = 40, k = 5, gam=gam, theta = theta, lower = threshold, prop = p
    # record the mean of each parameter and the sd of gamma 2
    tm[i, c("pmean_gam0", "pmean_gam1", "pmean_gam2", "pmean_gam3")] <- colMeans(tmp)
    tm[i, "psd_gam2"] <- sd(tmp[, 3])
  }
  L[[as.character(threshold)]] <- cbind(threshold, props, tm)
}

# combine in dataframe
```

```r
df <- as.data.frame(do.call(rbind, L))
df$bias <- df$pmean_gam2 - (-1.50)
df$sim_mse <- df$bias^2 + df$psd_gam2^2
df$sim_rmse <- sqrt(df$sim_mse)

# write to file for use later
write.csv(df,
          file = "C:/Users/Tom/OneDrive/Documents/leiden/thesis/exp4.csv",
          row.names = FALSE)

# plot of the bias against the prop
ggplot(data = df, aes(x = props, y = bias,
                      color = factor(threshold))) +
  geom_point() +
  geom_line() +
  theme_bw() +
  labs(title = " as a function of threshold, K = 40, k = 5") +
  theme(plot.title = element_text(hjust = 0.5))
```

Chapter 11

Appendix C: R code for replicating tables and plots in this thesis

```
# get packages
library(pacman)
p_load(char=c("rprojroot", "pander", "stargazer", "ggplot2", "latex2exp",
              "grid", "gridExtra", "dplyr"), install=TRUE)

# pander options
panderOptions('table.split.table', 300)
A <- c("1", "1", "0", "0")
Y <- c("1", "0", "1", "0")
Ya0 <- c("?", "?", "1", "0")
Ya1 <- c("1", "0", "?", "?")
Ya1_Ya0 <- rep("?", 4)
df <- as.data.frame(cbind(A, Y, Ya1, Ya0, Ya1_Ya0))
names(df) <- c("$A_i$", "$Y_i$", "$Y_{a=1, i}$", "$Y_{a=0, i}$", "$Y_{a=1, i} - Y_{a=0, i}$")
pander(df, caption = "Enumerated potential outcomes for a single subject $i$")
id <- 1:12
Ya1 <- c("1", "1", "0", "0", rep("?", 8))
Ya0 <- c(rep("?", 4), rep("1", 6), "0", "0")
df <- as.data.frame(cbind(id, Ya1, Ya0))
names(df) <- c("Subject", "$Y_{a=1}$", "$Y_{a=0}$")
pander(df, caption = "Exposed and unexposed subjects within the strata $L=l$")

# ------------------------------------------------------
# Discrete Bernoulli case
# probability
p <- 0.7
df <- data.frame(x = c(0, 0, 1, 1, 2, 2),
                 y = c(0, 1 - p, 1 - p, 1, 1, 0))

#
gg_bern <- ggplot(data = df, aes(x = x, y = y)) +
  geom_path(color = "red") +
  scale_x_continuous(breaks = c(0.5, 1.5),
                     limits = c(0, 2),
```

```
                     label = c("Y = 0", "Y = 1")) +
  scale_y_continuous(breaks = scales::pretty_breaks(n = 10)) +
  theme_classic() +
  labs(title = "Discrete Bernoulli CDF",
       x = TeX("Y_a = F^{-1}(U)"),
       y = TeX("$F(Y_a) = U$")) +
  theme(plot.title = element_text(hjust = 0.5))

set.seed(13878)
x <- rexp(100, 0.25)
y <- pexp(x, rate = 0.25)
P <- ecdf(x)
yd <- P(seq(0, ceiling(max(x)), 1))
t <- 0:24
df <- data.frame(x, y)
df2 <- data.frame(t, yd)
gg_exp <- ggplot(data = df, aes(x = x, y = y)) +
  geom_line() +
  geom_step(data = df2, aes(x = t, y = yd), color = "red") +
  #stat_ecdf(geom = "step", n = 25, color = "red") +
  theme_classic() +
  scale_x_continuous(expand = c(0, 0),
                     breaks = scales::pretty_breaks(n = 20)) +
  scale_y_continuous(expand = c(0, 0),
                     breaks = scales::pretty_breaks(n = 10)) +
  geom_vline(xintercept = c(7, 8)) +
  labs(title = "Continuous exponential CDF and discretized by time step",
       x = TeX("survival time (t =$F^{-1}(U_0)$)"),
       y = TeX("$F(t) = U_0$"))  +
  theme(plot.title = element_text(hjust = 0.5))

grid.arrange(gg_bern, gg_exp, ncol = 2)
exp1_reg_mc <- read.csv(file = "exp1_reg_mc.csv")
param <- c("$\\gamma_0$", "$\\gamma_1$", "$\\gamma_2$", "$\\gamma_3$")
exp1_reg_mc <- cbind(param, exp1_reg_mc)
names(exp1_reg_mc) <- c("", "True", "$\\tau = 0$", "$\\tau = 100$", "$\\tau = 200$", "$\\tau = 300$", "$
pander(exp1_reg_mc, caption = "Pooled logistic regression results for data simulated with positivity vio
exp1 <- read.csv(file = "C:/Users/Tom/OneDrive/Documents/leiden/thesis/exp1.csv")
# bias
gg_bias <- ggplot(data = exp1, aes(x = threshold, y = bias)) +
  geom_point() +
  geom_line() +
  scale_x_continuous(breaks = seq(from = 0, to = 350, by = 25)) +
  scale_y_continuous(breaks = scales::pretty_breaks(n = 10)) +
  theme_bw() +
  labs(title = TeX("Bias as a function of threshold ($\\tau$)"),
       x = TeX("threshold ($\\tau$)"),
       y = "Simulated bias") +
  theme(plot.title = element_text(hjust = 0.5))

# sd
gg_sd <- ggplot(data = exp1, aes(x = threshold, y = psd_gam2)) +
  geom_point() +
```

```
  geom_line() +
  scale_x_continuous(breaks = seq(from = 0, to = 350, by = 25)) +
  scale_y_continuous(limits = c(0, NA), breaks = scales::pretty_breaks(n = 10)) +
  theme_bw() +
  labs(title = TeX("S.d. as a function of threshold ($\\tau$)"),
       x = TeX("threshold ($\\tau$)"),
       y = "Simulated average standard error") +
  theme(plot.title = element_text(hjust = 0.5))

# rMSE
gg_rMSE <- ggplot(data = exp1, aes(x = threshold, y = sim_rmse)) +
  geom_point() +
  geom_line() +
  scale_x_continuous(breaks = seq(from = 0, to = 350, by = 25)) +
  scale_y_continuous(breaks = scales::pretty_breaks(n = 10)) +
  theme_bw() +
  labs(title = TeX("RMSE as a function of threshold ($\\tau$)"),
       x = TeX("threshold ($\\tau$)"),
       y = "Simulated RMSE") +
  theme(plot.title = element_text(hjust = 0.5))

# combine all three into one plot
grid.arrange(gg_bias, gg_sd, gg_rMSE, nrow = 3,
             top = "")
# expit or logit^-1 function
expit <- function(x) return(exp(x)/(1 + exp(x)))

# function for evaluating the MSM
msm <- function(gam, t, ts, a) gam[1] + gam[2] * ((1 - a) * (t-1) + a * ts) + gam[3] * a + gam[4] * a *

get_dat <- function(gam, ts, a, K, thresh){
  t <- ts:K
  lp <- sapply(t, function(t) msm(gam, t = t, ts = ts, a = a))
  hazard <- expit(lp)
  surv <- cumprod(1 - hazard)
  data.frame(t = t, threshold = thresh, a = a, lp = lp, hazard = hazard, surv = surv)
}

# no positivity for exposed and non-exposed
t0_a1 <- get_dat(gam = c(-3.00, 0.05, -1.50, 0.10), ts = 0, a = 1, K = 40, thresh = 0)
t0_a0 <- get_dat(gam = c(-3.00, 0.05, -1.50, 0.10), ts = 0, a = 0, K = 40, thresh = 0)

# positivity violations at thresholds of 100, 200, 300 and 350
gam100 <- as.numeric(exp1[exp1$threshold==100, 1:4])
t100_a1 <- get_dat(gam = gam100, ts = 0, a = 1, K = 40, thresh = 100)
gam200 <- as.numeric(exp1[exp1$threshold==200, 1:4])
t200_a1 <- get_dat(gam = gam200, ts = 0, a = 1, K = 40, thresh = 200)
gam300 <- as.numeric(exp1[exp1$threshold==300, 1:4])
t300_a1 <- get_dat(gam = gam300, ts = 0, a = 1, K = 40, thresh = 300)
gam350 <- as.numeric(exp1[exp1$threshold==350, 1:4])
t350_a1 <- get_dat(gam = gam350, ts = 0, a = 1, K = 40, thresh = 350)

# combine into one data frame for plotting
```

```
df <- rbind(t0_a0, t0_a1, t100_a1, t200_a1, t300_a1, t350_a1)

# Hazard
gg_haz <- ggplot(data = df, aes(x = t, y = hazard,
                                color = factor(threshold),
                                linetype = factor(a))) +
  geom_line() +
  scale_x_continuous(breaks = scales::pretty_breaks(n = 10)) +
  scale_y_continuous(breaks = scales::pretty_breaks(n = 10)) +
  guides(color=guide_legend(
                keywidth=0.1,
                keyheight=0.1,
                default.unit="inch"),
         linetype=guide_legend(
                keywidth=0.1,
                keyheight=0.1,
                default.unit="inch")) +
  theme_bw() +
  labs(title = TeX("K = 40, k = 5"),
       x = TeX("Time t"),
       y = TeX("Hazard $\\lambda_t$"),
       color = TeX("$\\tau$"),
       linetype = "Exposure (A)") +
  theme(plot.title = element_blank(),
        legend.position = c(0.15, 0.75),
        legend.text = element_text(size = 5),
        legend.title = element_text(size = 5))

# Survival
gg_surv <- ggplot(data = df, aes(x = t, y = surv,
                                color = factor(threshold),
                                linetype = factor(a))) +
  geom_line() +
  scale_x_continuous(breaks = scales::pretty_breaks(n = 10)) +
  scale_y_continuous(breaks = scales::pretty_breaks(n = 10)) +
  guides(color=guide_legend(
                keywidth=0.1,
                keyheight=0.1,
                default.unit="inch"),
         linetype=guide_legend(
                keywidth=0.1,
                keyheight=0.1,
                default.unit="inch")) +
  theme_bw() +
  labs(title = TeX("K = 40, k = 5"),
       x = TeX("Time t"),
       y = TeX("Survival S(t)"),
       color = TeX("$\\tau$"),
       linetype = "Exposure (A)") +
  theme(plot.title = element_blank(),
        legend.position = c(0.85, 0.75),
        legend.text = element_text(size = 5),
        legend.title = element_text(size = 5))
```

```
# combine into one plot
grid.arrange(gg_haz, gg_surv, nrow = 1)
exp3 <- read.csv(file = "exp3.csv")
tmp <- dplyr::filter(exp3, threshold %in% c(0, 100, 200, 350))
tmp <- dplyr::select(tmp, n, threshold, pmean_gam2, psd_gam2, bias, sim_rmse)

# round to 3 decimal places, keep zeroes
tmp$pmean_gam2 <- sprintf("%.3f", round(tmp$pmean_gam2,3))
tmp$psd_gam2 <- sprintf("%.3f", round(tmp$psd_gam2,3))
tmp$bias <- sprintf("%.3f", round(tmp$bias,3))
tmp$sim_rmse <- sprintf("%.3f", round(tmp$sim_rmse,3))

# table of longer and shorter follow up times effects on bias at different thresholds
names(tmp)[-c(1)] <- c("$\\tau$", "$\\hat\\gamma_2$", "s.d.", "bias", "RMSE")
pander(tmp, caption = "Pooled logistic regression results for $\\gamma_2$ from data simulated with posi
# bias
gg_bias <- ggplot(data = exp3, aes(x = threshold, y = bias,
                                   color = factor(n))) +
  geom_point() +
  geom_line() +
  scale_x_continuous(breaks = seq(from = 0, to = 350, by = 25)) +
  scale_y_continuous(breaks = scales::pretty_breaks(n = 10)) +
  theme_bw() +
  labs(title = TeX("Bias as a function of threshold ($\\tau$), T = 40, k = 5"),
       x = TeX("threshold ($\\tau$)"),
       y = "Simulated bias",
       color = "n") +
  theme(plot.title = element_text(hjust = 0.5))

# sd
gg_sd <- ggplot(data = exp3, aes(x = threshold, y = psd_gam2,
                                 color = factor(n))) +
  geom_point() +
  geom_line() +
  scale_x_continuous(breaks = seq(from = 0, to = 350, by = 25)) +
  scale_y_continuous(limits = c(0, NA), breaks = scales::pretty_breaks(n = 10)) +
  theme_bw() +
  labs(title = TeX("S.d. as a function of threshold ($\\tau$), T = 40, k = 5"),
       x = TeX("threshold ($\\tau$)"),
       y = "Simulated average SE",
       color = "n") +
  theme(plot.title = element_text(hjust = 0.5))

# rMSE
gg_rMSE <- ggplot(data = exp3, aes(x = threshold, y = sim_rmse,
                                   color = factor(n))) +
  geom_point() +
  geom_line() +
  scale_x_continuous(breaks = seq(from = 0, to = 350, by = 25)) +
  scale_y_continuous(breaks = scales::pretty_breaks(n = 10)) +
  theme_bw() +
  labs(title = TeX("RMSE as a function of threshold ($\\tau$), T = 40, k = 5"),
       x = TeX("threshold ($\\tau$)"),
```

```
       y = "Simulated RMSE",
       color = "n") +
  theme(plot.title = element_text(hjust = 0.5))

# combine all three into one plot
grid.arrange(gg_bias, gg_sd, gg_rMSE, nrow = 3,
             top = "")

# no positivity for exposed and non-exposed at K = 20
t0_a1 <- get_dat(gam = c(-3.00, 0.05, -1.50, 0.10), ts = 0, a = 1, K = 40, thresh = 0)
t0_a0 <- get_dat(gam = c(-3.00, 0.05, -1.50, 0.10), ts = 0, a = 0, K = 40, thresh = 0)

# positivity violations at thresholds of 100, 200, 300 and 350 for n = 200
gam100_n200 <- as.numeric(exp3[exp3$threshold==100 & exp3$n==200, c("pmean_gam0", "pmean_gam1", "pmean_
t100_a1_n200 <- get_dat(gam = gam100_n200, ts = 0, a = 1, K = 40, thresh = 100)
gam200_n200 <- as.numeric(exp3[exp3$threshold==200 & exp3$n==200, c("pmean_gam0", "pmean_gam1", "pmean_
t200_a1_n200 <- get_dat(gam = gam200_n200, ts = 0, a = 1, K = 40, thresh = 200)
gam300_n200 <- as.numeric(exp3[exp3$threshold==300 & exp3$n==200, c("pmean_gam0", "pmean_gam1", "pmean_
t300_a1_n200 <- get_dat(gam = gam300_n200, ts = 0, a = 1, K = 40, thresh = 300)
gam350_n200 <- as.numeric(exp3[exp3$threshold==350 & exp3$n==200, c("pmean_gam0", "pmean_gam1", "pmean_
t350_a1_n200 <- get_dat(gam = gam350_n200, ts = 0, a = 1, K = 40, thresh = 350)

# combine into one data frame for plotting
df_n200 <- rbind(t0_a1, t0_a0, t100_a1_n200, t200_a1_n200, t300_a1_n200, t350_a1_n200)

# positivity violations at thresholds of 100, 200, 300 and 350 for n = 1000
gam100_n1000 <- as.numeric(exp3[exp3$threshold==100 & exp3$n==1000, c("pmean_gam0", "pmean_gam1", "pmea
t100_a1_n1000 <- get_dat(gam = gam100_n1000, ts = 0, a = 1, K = 40, thresh = 100)
gam200_n1000 <- as.numeric(exp3[exp3$threshold==200 & exp3$n==1000, c("pmean_gam0", "pmean_gam1", "pmea
t200_a1_n1000 <- get_dat(gam = gam200_n1000, ts = 0, a = 1, K = 40, thresh = 200)
gam300_n1000 <- as.numeric(exp3[exp3$threshold==300 & exp3$n==1000, c("pmean_gam0", "pmean_gam1", "pmea
t300_a1_n1000 <- get_dat(gam = gam300_n1000, ts = 0, a = 1, K = 40, thresh = 300)
gam350_n1000 <- as.numeric(exp3[exp3$threshold==350 & exp3$n==1000, c("pmean_gam0", "pmean_gam1", "pmea
t350_a1_n1000 <- get_dat(gam = gam350_n1000, ts = 0, a = 1, K = 40, thresh = 350)

# combine into one data frame for plotting
df_n1000 <- rbind(t0_a1, t0_a0, t100_a1_n1000, t200_a1_n1000, t300_a1_n1000, t350_a1_n1000)

# Hazard n = 200
gg_haz_n200 <- ggplot(data = df_n200, aes(x = t, y = hazard,
                                color = factor(threshold),
                                linetype = factor(a))) +
  geom_line() +
  scale_x_continuous(breaks = scales::pretty_breaks(n = 10)) +
  scale_y_continuous(limits = c(0, 0.4), breaks = scales::pretty_breaks(n = 10)) +
  guides(color=guide_legend(
                keywidth=0.1,
                keyheight=0.1,
                default.unit="inch"),
         linetype=guide_legend(
                keywidth=0.1,
                keyheight=0.1,
                default.unit="inch")) +
```

```
  theme_bw() +
  labs(title = TeX("$n = 200$"),
       x = TeX("Time t"),
       y = TeX("Hazard $\\lambda_t$"),
       color = TeX("$\\tau$"),
       linetype = "Exposure (A)") +
  theme(legend.position = c(0.15, 0.75),
        legend.text = element_text(size = 5),
        legend.title = element_text(size = 5))

# Hazard n = 1000
gg_haz_n1000 <- ggplot(data = df_n1000, aes(x = t, y = hazard,
                                color = factor(threshold),
                                linetype = factor(a))) +
  geom_line() +
  scale_x_continuous(breaks = scales::pretty_breaks(n = 10)) +
  scale_y_continuous(limits = c(0, 0.4), breaks = scales::pretty_breaks(n = 10)) +
  guides(color=guide_legend(
                keywidth=0.1,
                keyheight=0.1,
                default.unit="inch"),
         linetype=guide_legend(
                keywidth=0.1,
                keyheight=0.1,
                default.unit="inch")) +
  theme_bw() +
  labs(title = TeX("$n = 1000$"),
       x = TeX("Time t"),
       y = TeX("Hazard $\\lambda_t$"),
       color = TeX("$\\tau$"),
       linetype = "Exposure (A)") +
  theme(legend.position = c(0.15, 0.75),
        legend.text = element_text(size = 5),
        legend.title = element_text(size = 5))

# Survival n=200
gg_surv_n200 <- ggplot(data = df_n200, aes(x = t, y = surv,
                                color = factor(threshold),
                                linetype = factor(a))) +
  geom_line() +
  scale_x_continuous(breaks = scales::pretty_breaks(n = 10)) +
  scale_y_continuous(breaks = scales::pretty_breaks(n = 10)) +
  guides(color=guide_legend(
                keywidth=0.1,
                keyheight=0.1,
                default.unit="inch"),
         linetype=guide_legend(
                keywidth=0.1,
                keyheight=0.1,
                default.unit="inch")) +
  theme_bw() +
  labs(title = TeX("$n = 200$"),
       x = TeX("Time t"),
```

```
       y = TeX("Survival S(t)"),
       color = TeX("$\\tau$"),
       linetype = "Exposure (A)") +
  theme(legend.position = c(0.85, 0.75),
        legend.text = element_text(size = 5),
        legend.title = element_text(size = 5))

# Survival n=1000
gg_surv_n1000 <- ggplot(data = df_n1000, aes(x = t, y = surv,
                                color = factor(threshold),
                                linetype = factor(a))) +
  geom_line() +
  scale_x_continuous(breaks = scales::pretty_breaks(n = 10)) +
  scale_y_continuous(breaks = scales::pretty_breaks(n = 10)) +
  guides(color=guide_legend(
                keywidth=0.1,
                keyheight=0.1,
                default.unit="inch"),
         linetype=guide_legend(
                keywidth=0.1,
                keyheight=0.1,
                default.unit="inch")) +
  theme_bw() +
  labs(title = TeX("$n = 1000$"),
       x = TeX("Time t"),
       y = TeX("Survival S(t)"),
       color = TeX("$\\tau$"),
       linetype = "Exposure (A)") +
  theme(legend.position = c(0.85, 0.75),
        legend.text = element_text(size = 5),
        legend.title = element_text(size = 5))

# combine into one plot
grid.arrange(gg_haz_n200, gg_surv_n200,
             gg_haz_n1000, gg_surv_n1000,
             nrow = 2)
exp2 <- read.csv(file = "exp2.csv")
tmp <- dplyr::filter(exp2, threshold %in% c(0, 100, 200, 300, 350))
tmp <- dplyr::select(tmp, K, threshold, pmean_gam2, psd_gam2, bias, sim_rmse)

# round to 3 decimal places, keep zeroes
tmp$pmean_gam2 <- sprintf("%.3f", round(tmp$pmean_gam2,3))
tmp$psd_gam2 <- sprintf("%.3f", round(tmp$psd_gam2,3))
tmp$bias <- sprintf("%.3f", round(tmp$bias,3))
tmp$sim_rmse <- sprintf("%.3f", round(tmp$sim_rmse,3))

# table of longer and shorter follow up times effects on bias at different thresholds
names(tmp) <- c("T", "$\\tau$", "$\\hat\\gamma_2$", "s.d.", "bias", "RMSE")
pander(tmp, caption = "Pooled logistic regression results for $\\gamma_2$ from data simulated with posi
# bias
gg_bias <- ggplot(data = exp2, aes(x = threshold, y = bias,
                                   color = factor(K))) +
  geom_point() +
```

```
  geom_line() +
  scale_x_continuous(breaks = seq(from = 0, to = 350, by = 25)) +
  scale_y_continuous(breaks = scales::pretty_breaks(n = 10)) +
  theme_bw() +
  labs(title = TeX("Bias as a function of threshold ($\\tau$)"),
       x = TeX("threshold ($\\tau$)"),
       y = "Simulated bias",
       color = "T") +
  theme(plot.title = element_text(hjust = 0.5))

# sd
gg_sd <- ggplot(data = exp2, aes(x = threshold, y = psd_gam2,
                                 color = factor(K))) +
  geom_point() +
  geom_line() +
  scale_x_continuous(breaks = seq(from = 0, to = 350, by = 25)) +
  scale_y_continuous(limits = c(0, NA), breaks = scales::pretty_breaks(n = 10)) +
  theme_bw() +
  labs(title = TeX("S.d. as a function of threshold ($\\tau$)"),
       x = TeX("threshold ($\\tau$)"),
       y = "Simulated average standard error",
       color = "T") +
  theme(plot.title = element_text(hjust = 0.5))

# rMSE
gg_rMSE <- ggplot(data = exp2, aes(x = threshold, y = sim_rmse,
                                   color = factor(K))) +
  geom_point() +
  geom_line() +
  scale_x_continuous(breaks = seq(from = 0, to = 350, by = 25)) +
  scale_y_continuous(breaks = scales::pretty_breaks(n = 10)) +
  theme_bw() +
  labs(title = TeX("RMSE as a function of threshold ($\\tau$)"),
       x = TeX("threshold ($\\tau$)"),
       y = "Simulated RMSE",
       color = "T") +
  theme(plot.title = element_text(hjust = 0.5))

# combine all three into one plot
grid.arrange(gg_bias, gg_sd, gg_rMSE, nrow = 3,
             top = "")

# no positivity for exposed and non-exposed at K = 20
#gam20 = c(-1, 0.05, -3, 0.25)
gam=c(-3,0.05,-1.5,0.1)
t0_a1_K20 <- get_dat(gam = gam, ts = 0, a = 1, K = 20, thresh = 0)
t0_a0_K20 <- get_dat(gam = gam, ts = 0, a = 0, K = 20, thresh = 0)

# positivity violations at thresholds of 100, 200, 300 and 350
gam100_K20 <- as.numeric(exp2[exp2$threshold==100 & exp2$K==20, c("pmean_gam0", "pmean_gam1", "pmean_gam
t100_a1_K20 <- get_dat(gam = gam100_K20, ts = 0, a = 1, K = 20, thresh = 100)
gam200_K20 <- as.numeric(exp2[exp2$threshold==200 & exp2$K==20, c("pmean_gam0", "pmean_gam1", "pmean_gam
t200_a1_K20 <- get_dat(gam = gam200_K20, ts = 0, a = 1, K = 20, thresh = 200)
```

```r
gam300_K20 <- as.numeric(exp2[exp2$threshold==300 & exp2$K==20, c("pmean_gam0", "pmean_gam1", "pmean_gam
t300_a1_K20 <- get_dat(gam = gam300_K20, ts = 0, a = 1, K = 20, thresh = 300)
gam350_K20 <- as.numeric(exp2[exp2$threshold==350 & exp2$K==20, c("pmean_gam0", "pmean_gam1", "pmean_gam
t350_a1_K20 <- get_dat(gam = gam350_K20, ts = 0, a = 1, K = 20, thresh = 350)

# combine into one data frame for plotting
df_K20 <- rbind(t0_a1_K20, t0_a0_K20, t100_a1_K20, t200_a1_K20, t300_a1_K20, t350_a1_K20)

# no positivity for exposed and non-exposed at K = 30
t0_a1_K30 <- get_dat(gam = gam, ts = 0, a = 1, K = 30, thresh = 0)
t0_a0_K30 <- get_dat(gam = gam, ts = 0, a = 0, K = 30, thresh = 0)

# positivity violations at thresholds of 100, 200, 300 and 350
gam100_K30 <- as.numeric(exp2[exp2$threshold==100 & exp2$K==30, c("pmean_gam0", "pmean_gam1", "pmean_gam
t100_a1_K30 <- get_dat(gam = gam100_K30, ts = 0, a = 1, K = 30, thresh = 100)
gam300_K30 <- as.numeric(exp2[exp2$threshold==200 & exp2$K==30, c("pmean_gam0", "pmean_gam1", "pmean_gam
t200_a1_K30 <- get_dat(gam = gam300_K30, ts = 0, a = 1, K = 30, thresh = 200)
gam300_K30 <- as.numeric(exp2[exp2$threshold==300 & exp2$K==30, c("pmean_gam0", "pmean_gam1", "pmean_gam
t300_a1_K30 <- get_dat(gam = gam300_K30, ts = 0, a = 1, K = 30, thresh = 300)
gam350_K30 <- as.numeric(exp2[exp2$threshold==350 & exp2$K==30, c("pmean_gam0", "pmean_gam1", "pmean_gam
t350_a1_K30 <- get_dat(gam = gam350_K30, ts = 0, a = 1, K = 30, thresh = 350)

# combine into one data frame for plotting
df_K30 <- rbind(t0_a1_K30, t0_a0_K30, t100_a1_K30, t200_a1_K30, t300_a1_K30, t350_a1_K30)

# no positivity for exposed and non-exposed at K = 40
t0_a1_K40 <- get_dat(gam = gam, ts = 0, a = 1, K = 40, thresh = 0)
t0_a0_K40 <- get_dat(gam = gam, ts = 0, a = 0, K = 40, thresh = 0)

# positivity violations at thresholds of 100, 200, 300 and 350
gam100_K40 <- as.numeric(exp2[exp2$threshold==100 & exp2$K==40, c("pmean_gam0", "pmean_gam1", "pmean_gam
t100_a1_K40 <- get_dat(gam = gam100_K40, ts = 0, a = 1, K = 40, thresh = 100)
gam400_K40 <- as.numeric(exp2[exp2$threshold==200 & exp2$K==40, c("pmean_gam0", "pmean_gam1", "pmean_gam
t200_a1_K40 <- get_dat(gam = gam400_K40, ts = 0, a = 1, K = 40, thresh = 200)
gam300_K40 <- as.numeric(exp2[exp2$threshold==300 & exp2$K==40, c("pmean_gam0", "pmean_gam1", "pmean_gam
t300_a1_K40 <- get_dat(gam = gam300_K40, ts = 0, a = 1, K = 40, thresh = 300)
gam350_K40 <- as.numeric(exp2[exp2$threshold==350 & exp2$K==40, c("pmean_gam0", "pmean_gam1", "pmean_gam
t350_a1_K40 <- get_dat(gam = gam350_K40, ts = 0, a = 1, K = 40, thresh = 350)

# combine into one data frame for plotting
df_K40 <- rbind(t0_a1_K40, t0_a0_K40, t100_a1_K40, t200_a1_K40, t300_a1_K40, t350_a1_K40)


# Hazard K = 20
gg_haz_K20 <- ggplot(data = df_K20, aes(x = t, y = hazard,
                                color = factor(threshold),
                                linetype = factor(a))) +
  geom_line() +
  scale_x_continuous(breaks = scales::pretty_breaks(n = 10)) +
  scale_y_continuous(breaks = scales::pretty_breaks(n = 10)) +
  guides(color=guide_legend(
                keywidth=0.1,
                keyheight=0.1,
```

```
                default.unit="inch"),
         linetype=guide_legend(
                keywidth=0.1,
                keyheight=0.1,
                default.unit="inch")) +
  theme_bw() +
  labs(title = TeX("T = 20, k = 5"),
       x = TeX("Time t"),
       y = TeX("Hazard $\\lambda_t$"),
       color = TeX("$\\tau$"),
       linetype = "Exposure (A)") +
  theme(legend.position = c(0.15, 0.75),
        legend.text = element_text(size = 5),
        legend.title = element_text(size = 5))

# Hazard K = 30
gg_haz_K30 <- ggplot(data = df_K30, aes(x = t, y = hazard,
                                color = factor(threshold),
                                linetype = factor(a))) +
  geom_line() +
  scale_x_continuous(breaks = scales::pretty_breaks(n = 10)) +
  scale_y_continuous(breaks = scales::pretty_breaks(n = 10)) +
  guides(color=guide_legend(
                keywidth=0.1,
                keyheight=0.1,
                default.unit="inch"),
         linetype=guide_legend(
                keywidth=0.1,
                keyheight=0.1,
                default.unit="inch")) +
  theme_bw() +
  labs(title = TeX("T = 30, k = 5"),
       x = TeX("Time t"),
       y = TeX("Hazard $\\lambda_t$"),
       color = TeX("$\\tau$"),
       linetype = "Exposure (A)") +
  theme(legend.position = c(0.15, 0.75),
        legend.text = element_text(size = 5),
        legend.title = element_text(size = 5))

# Hazard K = 40
gg_haz_K40 <- ggplot(data = df_K40, aes(x = t, y = hazard,
                                color = factor(threshold),
                                linetype = factor(a))) +
  geom_line() +
  scale_x_continuous(breaks = scales::pretty_breaks(n = 10)) +
  scale_y_continuous(breaks = scales::pretty_breaks(n = 10)) +
  guides(color=guide_legend(
                keywidth=0.1,
                keyheight=0.1,
                default.unit="inch"),
         linetype=guide_legend(
                keywidth=0.1,
```

```
                keyheight=0.1,
                default.unit="inch")) +
  theme_bw() +
  labs(title = TeX("T = 40, k = 5"),
       x = TeX("Time t"),
       y = TeX("Hazard $\\lambda_t$"),
       color = TeX("$\\tau$"),
       linetype = "Exposure (A)") +
  theme(legend.position = c(0.15, 0.75),
        legend.text = element_text(size = 5),
        legend.title = element_text(size = 5))


# Survival K = 20
gg_surv_K20 <- ggplot(data = df_K20, aes(x = t, y = surv,
                                color = factor(threshold),
                                linetype = factor(a))) +
  geom_line() +
  scale_x_continuous(breaks = scales::pretty_breaks(n = 10)) +
  scale_y_continuous(breaks = scales::pretty_breaks(n = 10)) +
  guides(color=guide_legend(
                keywidth=0.1,
                keyheight=0.1,
                default.unit="inch"),
         linetype=guide_legend(
                keywidth=0.1,
                keyheight=0.1,
                default.unit="inch")) +
  theme_bw() +
  labs(title = TeX("T = 20, k = 5"),
       x = TeX("Time t"),
       y = TeX("Survival S(t)"),
       color = TeX("$\\tau$"),
       linetype = "Exposure (A)") +
  theme(legend.position = c(0.85, 0.75),
        legend.text = element_text(size = 5),
        legend.title = element_text(size = 5))

# Survival K = 30
gg_surv_K30 <- ggplot(data = df_K30, aes(x = t, y = surv,
                                color = factor(threshold),
                                linetype = factor(a))) +
  geom_line() +
  scale_x_continuous(breaks = scales::pretty_breaks(n = 10)) +
  scale_y_continuous(breaks = scales::pretty_breaks(n = 10)) +
  guides(color=guide_legend(
                keywidth=0.1,
                keyheight=0.1,
                default.unit="inch"),
         linetype=guide_legend(
                keywidth=0.1,
                keyheight=0.1,
                default.unit="inch")) +
```

```
  theme_bw() +
  labs(title = TeX("T = 30, k = 5"),
       x = TeX("Time t"),
       y = TeX("Survival S(t)"),
       color = TeX("$\\tau$"),
       linetype = "Exposure (A)") +
  theme(legend.position = c(0.85, 0.75),
        legend.text = element_text(size = 5),
        legend.title = element_text(size = 5))

# Survival K = 40
gg_surv_K40 <- ggplot(data = df_K40, aes(x = t, y = surv,
                                color = factor(threshold),
                                linetype = factor(a))) +
  geom_line() +
  scale_x_continuous(breaks = scales::pretty_breaks(n = 10)) +
  scale_y_continuous(breaks = scales::pretty_breaks(n = 10)) +
  guides(color=guide_legend(
                keywidth=0.1,
                keyheight=0.1,
                default.unit="inch"),
         linetype=guide_legend(
                keywidth=0.1,
                keyheight=0.1,
                default.unit="inch")) +
  theme_bw() +
  labs(title = TeX("T = 40, k = 5"),
       x = TeX("Time t"),
       y = TeX("Survival S(t)"),
       color = TeX("$\\tau$"),
       linetype = "Exposure (A)") +
  theme(legend.position = c(0.85, 0.75),
        legend.text = element_text(size = 5),
        legend.title = element_text(size = 5))

# combine into one plot
grid.arrange(gg_haz_K20, gg_surv_K20,
             gg_haz_K30, gg_surv_K30,
             gg_haz_K40, gg_surv_K40,nrow = 3)
exp4 <- read.csv(file = "exp4.csv")
tmp <- dplyr::filter(exp4, props %in% c(0.25, 0.5, 0.75, 0.9, 1))
tmp <- dplyr::select(tmp, threshold, props, pmean_gam2, psd_gam2, bias, sim_rmse)

# round to 3 decimal places, keep zeroes
tmp$pmean_gam2 <- sprintf("%.3f", round(tmp$pmean_gam2,3))
tmp$psd_gam2 <- sprintf("%.3f", round(tmp$psd_gam2,3))
tmp$bias <- sprintf("%.3f", round(tmp$bias,3))
tmp$sim_rmse <- sprintf("%.3f", round(tmp$sim_rmse,3))

# table of longer and shorter follow up times effects on bias at different thresholds
names(tmp) <- c("$\\tau$", "$\\pi$", "$\\hat\\gamma_2$", "s.d.", "bias", "RMSE")
pander(tmp, caption = "Pooled logistic regression results for $\\gamma_2$ from data simulated with posi
# bias
```

```r
gg_bias <- ggplot(data = exp4, aes(x = props, y = bias,
                                   color = factor(threshold))) +
  geom_point() +
  geom_line() +
  scale_x_continuous(breaks = seq(from = 0, to = 1, by = 0.1)) +
  scale_y_continuous(breaks = scales::pretty_breaks(n = 10)) +
  theme_bw() +
  labs(title = "Bias as a function of proportion of positivity compliant doctors",
       x = TeX("proportion of positivity compliant doctors, $\\pi$"),
       y = "Simulated bias",
       color = expression(tau)) +
  theme(plot.title = element_text(hjust = 0.5))

# sd
gg_sd <- ggplot(data = exp4, aes(x = props, y = psd_gam2,
                                   color = factor(threshold))) +
  geom_point() +
  geom_line() +
  scale_x_continuous(breaks = seq(from = 0, to = 1, by = 0.1)) +
  scale_y_continuous(limits = c(0, NA), breaks = scales::pretty_breaks(n = 10)) +
  theme_bw() +
  labs(title = "S.d. as a function of proportion of positivity compliant doctors",
       x = TeX("proportion of positivity compliant doctors, $\\pi$"),
       y = "Simulated average SE",
       color = expression(tau)) +
  theme(plot.title = element_text(hjust = 0.5))

# rMSE
gg_rMSE <- ggplot(data = exp4, aes(x = props, y = sim_rmse,
                                   color = factor(threshold))) +
  geom_point() +
  geom_line() +
  scale_x_continuous(breaks = seq(from = 0, to = 1, by = 0.1)) +
  scale_y_continuous(breaks = scales::pretty_breaks(n = 10)) +
  theme_bw() +
  labs(title = "RMSE as a function of proportion of positivity compliant doctors",
       x = TeX("proportion of positivity compliant doctors, $\\pi$"),
       y = "Simulated RMSE",
       color = expression(tau)) +
  theme(plot.title = element_text(hjust = 0.5))

# combine all three into one plot
grid.arrange(gg_bias, gg_sd, gg_rMSE, nrow = 3,
             top = "")

# no positivity for exposed and non-exposed at K = 20
t0_a1 <- get_dat(gam = c(-3.00, 0.05, -1.50, 0.10), ts = 0, a = 1, K = 40, thresh = 0)
t0_a0 <- get_dat(gam = c(-3.00, 0.05, -1.50, 0.10), ts = 0, a = 0, K = 40, thresh = 0)

# positivity violations at thresholds of 100, 200, 300 and 350 for pi = 0.3
gam100_pi03 <- as.numeric(exp4[exp4$threshold==100 & exp4$props==0.3, c("pmean_gam0", "pmean_gam1", "pm
t100_a1_pi03 <- get_dat(gam = gam100_pi03, ts = 0, a = 1, K = 40, thresh = 100)
gam200_pi03 <- as.numeric(exp4[exp4$threshold==200 & exp4$props==0.3, c("pmean_gam0", "pmean_gam1", "pm
```

```
t200_a1_pi03 <- get_dat(gam = gam200_pi03, ts = 0, a = 1, K = 40, thresh = 200)
gam300_pi03 <- as.numeric(exp4[exp4$threshold==300 & exp4$props==0.3, c("pmean_gam0", "pmean_gam1", "pme
t300_a1_pi03 <- get_dat(gam = gam300_pi03, ts = 0, a = 1, K = 40, thresh = 300)
gam350_pi03 <- as.numeric(exp4[exp4$threshold==350 & exp4$props==0.3, c("pmean_gam0", "pmean_gam1", "pme
t350_a1_pi03 <- get_dat(gam = gam350_pi03, ts = 0, a = 1, K = 40, thresh = 350)

# combine into one data frame for plotting
df_pi03 <- rbind(t0_a1, t0_a0, t100_a1_pi03, t200_a1_pi03, t300_a1_pi03, t350_a1_pi03)

# positivity violations at thresholds of 100, 200, 300 and 350 for pi = 0.5
gam100_pi05 <- as.numeric(exp4[exp4$threshold==100 & exp4$props==0.5, c("pmean_gam0", "pmean_gam1", "pme
t100_a1_pi05 <- get_dat(gam = gam100_pi05, ts = 0, a = 1, K = 40, thresh = 100)
gam200_pi05 <- as.numeric(exp4[exp4$threshold==200 & exp4$props==0.5, c("pmean_gam0", "pmean_gam1", "pme
t200_a1_pi05 <- get_dat(gam = gam200_pi05, ts = 0, a = 1, K = 40, thresh = 200)
gam300_pi05 <- as.numeric(exp4[exp4$threshold==300 & exp4$props==0.5, c("pmean_gam0", "pmean_gam1", "pme
t300_a1_pi05 <- get_dat(gam = gam300_pi05, ts = 0, a = 1, K = 40, thresh = 300)
gam350_pi05 <- as.numeric(exp4[exp4$threshold==350 & exp4$props==0.5, c("pmean_gam0", "pmean_gam1", "pme
t350_a1_pi05 <- get_dat(gam = gam350_pi05, ts = 0, a = 1, K = 40, thresh = 350)

# combine into one data frame for plotting
df_pi05 <- rbind(t0_a1, t0_a0, t100_a1_pi05, t200_a1_pi05, t300_a1_pi05, t350_a1_pi05)

# positivity violations at thresholds of 100, 200, 300 and 350 for pi = 0.7
gam100_pi07 <- as.numeric(exp4[exp4$threshold==100 & exp4$props==0.7, c("pmean_gam0", "pmean_gam1", "pme
t100_a1_pi07 <- get_dat(gam = gam100_pi07, ts = 0, a = 1, K = 40, thresh = 100)
gam200_pi07 <- as.numeric(exp4[exp4$threshold==200 & exp4$props==0.7, c("pmean_gam0", "pmean_gam1", "pme
t200_a1_pi07 <- get_dat(gam = gam200_pi07, ts = 0, a = 1, K = 40, thresh = 200)
gam300_pi07 <- as.numeric(exp4[exp4$threshold==300 & exp4$props==0.7, c("pmean_gam0", "pmean_gam1", "pme
t300_a1_pi07 <- get_dat(gam = gam300_pi07, ts = 0, a = 1, K = 40, thresh = 300)
gam350_pi07 <- as.numeric(exp4[exp4$threshold==350 & exp4$props==0.7, c("pmean_gam0", "pmean_gam1", "pme
t350_a1_pi07 <- get_dat(gam = gam350_pi07, ts = 0, a = 1, K = 40, thresh = 350)

# combine into one data frame for plotting
df_pi07 <- rbind(t0_a1, t0_a0, t100_a1_pi07, t200_a1_pi07, t300_a1_pi07, t350_a1_pi07)

# Hazard pi = 0.3
gg_haz_pi03 <- ggplot(data = df_pi03, aes(x = t, y = hazard,
                                color = factor(threshold),
                                linetype = factor(a))) +
  geom_line() +
  scale_x_continuous(breaks = scales::pretty_breaks(n = 10)) +
  scale_y_continuous(limits = c(0, 0.4), breaks = scales::pretty_breaks(n = 10)) +
  guides(color=guide_legend(
                keywidth=0.1,
                keyheight=0.1,
                default.unit="inch"),
         linetype=guide_legend(
                keywidth=0.1,
                keyheight=0.1,
                default.unit="inch")) +
  theme_bw() +
  labs(title = TeX("$\\pi = 0.3$"),
       x = TeX("Time t"),
```

```
       y = TeX("Hazard $\\lambda_t$"),
       color = TeX("$\\tau$"),
       linetype = "Exposure (A)") +
  theme(legend.position = c(0.15, 0.75),
        legend.text = element_text(size = 5),
        legend.title = element_text(size = 5))

# Hazard pi = 0.5
gg_haz_pi05 <- ggplot(data = df_pi05, aes(x = t, y = hazard,
                                color = factor(threshold),
                                linetype = factor(a))) +
  geom_line() +
  scale_x_continuous(breaks = scales::pretty_breaks(n = 10)) +
  scale_y_continuous(limits = c(0, 0.4), breaks = scales::pretty_breaks(n = 10)) +
  guides(color=guide_legend(
                keywidth=0.1,
                keyheight=0.1,
                default.unit="inch"),
         linetype=guide_legend(
                keywidth=0.1,
                keyheight=0.1,
                default.unit="inch")) +
  theme_bw() +
  labs(title = TeX("$\\pi = 0.5$"),
       x = TeX("Time t"),
       y = TeX("Hazard $\\lambda_t$"),
       color = TeX("$\\tau$"),
       linetype = "Exposure (A)") +
  theme(legend.position = c(0.15, 0.75),
        legend.text = element_text(size = 5),
        legend.title = element_text(size = 5))

# Hazard pi = 0.7
gg_haz_pi07 <- ggplot(data = df_pi07, aes(x = t, y = hazard,
                                color = factor(threshold),
                                linetype = factor(a))) +
  geom_line() +
  scale_x_continuous(breaks = scales::pretty_breaks(n = 10)) +
  scale_y_continuous(limits = c(0, 0.4), breaks = scales::pretty_breaks(n = 10)) +
  guides(color=guide_legend(
                keywidth=0.1,
                keyheight=0.1,
                default.unit="inch"),
         linetype=guide_legend(
                keywidth=0.1,
                keyheight=0.1,
                default.unit="inch")) +
  theme_bw() +
  labs(title = TeX("$\\pi = 0.7$"),
       x = TeX("Time t"),
       y = TeX("Hazard $\\lambda_t$"),
       color = TeX("$\\tau$"),
       linetype = "Exposure (A)") +
```

```
  theme(legend.position = c(0.15, 0.75),
        legend.text = element_text(size = 5),
        legend.title = element_text(size = 5))

# Survival pi = 0.3
gg_surv_pi03 <- ggplot(data = df_pi03, aes(x = t, y = surv,
                                color = factor(threshold),
                                linetype = factor(a))) +
  geom_line() +
  scale_x_continuous(breaks = scales::pretty_breaks(n = 10)) +
  scale_y_continuous(breaks = scales::pretty_breaks(n = 10)) +
  guides(color=guide_legend(
                keywidth=0.1,
                keyheight=0.1,
                default.unit="inch"),
         linetype=guide_legend(
                keywidth=0.1,
                keyheight=0.1,
                default.unit="inch")) +
  theme_bw() +
  labs(title = TeX("$\\pi = 0.3$"),
       x = TeX("Time t"),
       y = TeX("Survival S(t)"),
       color = TeX("$\\tau$"),
       linetype = "Exposure (A)") +
  theme(legend.position = c(0.85, 0.75),
        legend.text = element_text(size = 5),
        legend.title = element_text(size = 5))

# Survival pi = 0.5
gg_surv_pi05 <- ggplot(data = df_pi05, aes(x = t, y = surv,
                                color = factor(threshold),
                                linetype = factor(a))) +
  geom_line() +
  scale_x_continuous(breaks = scales::pretty_breaks(n = 10)) +
  scale_y_continuous(breaks = scales::pretty_breaks(n = 10)) +
  guides(color=guide_legend(
                keywidth=0.1,
                keyheight=0.1,
                default.unit="inch"),
         linetype=guide_legend(
                keywidth=0.1,
                keyheight=0.1,
                default.unit="inch")) +
  theme_bw() +
  labs(title = TeX("$\\pi = 0.5$"),
       x = TeX("Time t"),
       y = TeX("Survival S(t)"),
       color = TeX("$\\tau$"),
       linetype = "Exposure (A)") +
  theme(legend.position = c(0.85, 0.75),
        legend.text = element_text(size = 5),
        legend.title = element_text(size = 5))
```

```
# Survival pi = 0.7
gg_surv_pi07 <- ggplot(data = df_pi07, aes(x = t, y = surv,
                                color = factor(threshold),
                                linetype = factor(a))) +
  geom_line() +
  scale_x_continuous(breaks = scales::pretty_breaks(n = 10)) +
  scale_y_continuous(breaks = scales::pretty_breaks(n = 10)) +
  guides(color=guide_legend(
                keywidth=0.1,
                keyheight=0.1,
                default.unit="inch"),
         linetype=guide_legend(
                keywidth=0.1,
                keyheight=0.1,
                default.unit="inch")) +
  theme_bw() +
  labs(title = TeX("$\\pi = 0.7$"),
       x = TeX("Time t"),
       y = TeX("Survival S(t)"),
       color = TeX("$\\tau$"),
       linetype = "Exposure (A)") +
  theme(legend.position = c(0.85, 0.75),
        legend.text = element_text(size = 5),
        legend.title = element_text(size = 5))

# combine into one plot
grid.arrange(gg_haz_pi03, gg_surv_pi03,
             gg_haz_pi05, gg_surv_pi05,
             gg_haz_pi07, gg_surv_pi07,
             nrow = 3)
## NA
## NA
##
```

References

1. Hernán MA. Beyond exchangeability: The other conditions for causal inference in medical research. 2012.

2. Westreich D, Cole SR. Invited commentary: positivity in practice. American journal of epidemiology. 2010;171:674–7.

3. Yang S, Eaton CB, Lu J, Lapane KL. Application of marginal structural models in pharmacoepidemiologic studies: A systematic review. 2014.

4. Neyman J. On the application of probability theory to agricultural experiments: principles (in Polish with German summary). Roczniki Nauk Rolniczch. 1923;10:21–51. doi:10.1214/ss/1177012031.

5. Rubin DB. Bayesian Inference for Causal Effects: The Role of Randomization. The Annals of Statistics. 1978.

6. Robins J. 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. 1986.

7. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. 2000.

8. Holland PW. Statistics and causal inference. Journal of the American Statistical Association. 1986.

9. Cole SR, Frangakis CE. The consistency statement in causal inference: A definition or an assumption? 2009.

10. Ding P, Li F. Causal Inference: A Missing Data Perspective. 2017;1–47. doi:10.1214/18-STS645.

11. Howe CJ, Cain LE, Hogan JW. Are All Biases Missing Data Problems? Current Epidemiology Reports. 2015.

12. Edwards JK, Cole SR, Westreich D. All your data are always missing: incorporating bias due to measurement error into the potential outcomes framework. International journal of epidemiology. 2015;44:1452–9.

13. Greenland S, Robins JM, Pearl J. Confounding and collapsibility in causal inference. Statistical science. 1999;29–46.

14. Daniel RM, Cousens SN, De Stavola BL, Kenward MG, Sterne JA. Methods for dealing with time-dependent confounding. Statistics in Medicine. 2013.

15. Hernán MA, Robins JM. Estimating causal effects from epidemiological data. 2006.

16. Pearl J. On the consistency rule in causal inference: Axiom, definition, assumption, or theorem? Epidemiology. 2010.

17. Vander Weele TJ. Concerning the consistency assumption in causal inference. Epidemiology. 2009.

18. Pearl J. Causality. Cambridge university press; 2009.

19. Greenland S, Pearl J. Adjustments and their Consequences-Collapsibility Analysis using Graphical Models. International Statistical Review. 2011.

20. Pearl J. Causal inference in the health sciences: A conceptual introduction. Health Services and

Outcomes Research Methodology. 2001.

21. Hernán MA, Hernández-Diaz S, Robins JM. A structural approach to selection bias. Epidemiology. 2004;15:615–25.

22. Didelez V, Kreiner S, Keiding N. Graphical Models for Inference Under Outcome-Dependent Sampling. Statistical Science. 2010.

23. Gill RD, Robins JM. Causal inference for complex longitudinal data: The continuous case. Annals of Statistics. 2001.

24. Robins JM. Marginal structural models versus structural nested models as tools for causal inference. In: Statistical models in epidemiology, the environment, and clinical trials. Springer; 2000. pp. 95–133.

25. Hernán MA. The hazards of hazard ratios. 2010.

26. Cole SR, Platt RW, Schisterman EF, Chu H, Westreich D, Richardson D, et al. Illustrating bias due to conditioning on a collider. International Journal of Epidemiology. 2010.

27. Sjölander A, Dahlqwist E, Zetterqvist J. A note on the noncollapsibility of rate differences and rate ratios. Epidemiology. 2016;27:356–9.

28. Nelder JA, Wedderburn RWM, Nelder AJ a, Wedderburn RWM. Generalized Linear Models. J R Statist Soc A. 1972.

29. Hernan MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. Epidemiology. 2000.

30. Robins JM. Data, design, and background knowledge in etiologic inference. Epidemiology. 2001.

31. Hernán MA, Hernández-Diaz S, Werler MM, Mitchell AA. Causal knowledge as a prerequisite for confounding evaluation: An application to birth defects epidemiology. American Journal of Epidemiology. 2002.

32. Hernán MA, Cole SR. Invited commentary: Causal diagrams and measurement bias. 2009.

33. John P. Klein MLM. Survival Analysis - Techniques for Ceosnred and Truncated Data - 2nd Edition. 2003.

34. Cox DR. Regression Models and Life-Tables. Journal of the Royal Statistical Society Series B (Methodological) Journal of the Royal Statistical Society Series B Research Section, on Wednesday. 1972.

35. Bull K, Spiegelhalter DJ. Survival Models: Survival Analysis in Observational Studies. In: Tutorials in biostatistics, statistical methods in clinical studies. 2005.

36. D'Agostino RB, Lee M-, Belanger AJ, Cupples LA, Anderson K, Kannel WB. Relation of pooled logistic regression to time dependent cox regression analysis: The framingham heart study. Statistics in Medicine. 1990.

37. Hernán MA RJ. Causal Inference. Boca Raton: Chapman & Hall/CRC, forthcoming.; 2018. https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/.

38. Greenland S, Maldonado G. The importance of critically interpreting simulation studies. Epidemiology (Cambridge, Mass). 1997.

39. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. Statistics in Medicine. 2006.

40. Pearl J. Comment: Understanding Simpson's paradox. 2014.

41. Pang M, Kaufman JS, Platt RW. Studying noncollapsibility of the odds ratio with marginal structural and logistic regression models. Statistical Methods in Medical Research. 2016.

42. Westreich D, Cole SR, Schisterman EF, Platt RW. A simulation study of finite-sample properties of

marginal structural Cox proportional hazards models. Statistics in Medicine. 2012.

43. Young JG, Tchetgen Tchetgen EJ. Simulation from a known Cox MSM using standard parametric models for the g-formula. Statistics in Medicine. 2014.

44. Rizzo ML. Statistical Computing with R. Computing. 2008.

45. Havercroft WG, Didelez V. Simulating from marginal structural models with time-dependent confounding. Statistics in medicine. 2012;31:4190–206.

46. Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. American journal of epidemiology. 2008;168:656–64.

47. Messer LC, Oakes JM, Mason S. Effects of socioeconomic and racial residential segregation on preterm birth: a cautionary tale of structural confounding. American journal of epidemiology. 2010;171:664–73.

48. Cheng YW, Hubbard A, Caughey AB, Tager IB. The association between persistent fetal occiput posterior position and perinatal outcomes: an example of propensity score and covariate distance matching. American journal of epidemiology. 2010;171:656–63.

49. Molina J, Sued M, Valdora M. Models for the propensity score that contemplate the positivity assumption and their application to missing data and causality. Statistics in Medicine. 2018.

50. Bembom O, Van Der Laan MJ. A practical illustration of the importance of realistic individualized treatment rules in causal inference. Electronic Journal of Statistics. 2007.

51. Mortimer KM, Neugebauer R, Van Der Laan M, Tager IB. An application of model-fitting procedures for marginal structural models. American Journal of Epidemiology. 2005;162:382–8.

52. Neugebauer R, Laan M van der. Why prefer double robust estimators in causal inference? Journal of Statistical Planning and Inference. 2005.

53. Wang Y, Petersen M, Bangsberg D, Laan MJ van der. Diagnosing bias in the inverse probability of treatment weighted estimator resulting from violation of experimental treatment assignment. UC Berkeley Division of Biostatistics working paper series. 2006.

54. Petersen ML, Porter KE, Gruber S, Wang Y, Van Der Laan MJ. Diagnosing and responding to violations in the positivity assumption. Statistical Methods in Medical Research. 2012.

55. Naimi AI, Cole SR, Westreich DJ, Richardson DB. A comparison of methods to estimate the hazard ratio under conditions of time-varying confounding and nonpositivity. Epidemiology. 2011.

56. Kang JDY, Schafer JL. Rejoinder: Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. Statistical Science. 2007.