Inverse Probability of Treatment Weighting Under Violations of Positivity

Tomas D. Morley

August 15, 2018

Abstract

Aknowledgements

Contents

Introduction

- set up problem and why we need to study positivty - give this motivation before anything else
- exhchangeability/confounder control vs positivity
- thresholding/positivty compliant doctors.
- in a survival context, need to choose simulation algorithm carefull because survival models are noncollapsible.

Marginal structural models (MSMs) are a popular class of models for performing causal inference in the presence of time dependent confounders. These models have an important application in areas of research such as epidemiology, social sciences and economics where randomised trials are prohibited by ethical or financial considerations, and hence confounding cannot be ruled out by randomization. Under these circumstances confounding can obscure the causal effect of treatment on outcome. An example of this, common in epidemiological studies, occurs when prognostic variables inform treatment decisions while at same time being predictors of the outcome of interest. In a longitudinal setting this is further complicated when the confounder itself is determined by earlier treatment. One consequence is that regression adjustment methods do not control for confounding in the longitudinal case and other techniques are required.

The Inverse probability of treatment weighting (IPTW) estimator is a technique which leads to consistent estimates in the presence of has been applied to censoring, missing data and survey design problems. The central idea is that by weighting the observed data in order to create a pseudo population is constructed in which treatemnt is assigned at random. Subsequent analysis where we ignore the confounder is then possible. which inference on the target population can be achieved. For example, when there is missing data weights can be used to create a pseudo-population in which there is no missingness. In the context of MSMs, the IPT weights relate to a pseudo-population in which there is no longer any confounding between the confounder and treatment and causal inferences can be made.

Underlying the IPTW method for estimating MSMs are four assumptions: 1) consistency 2) exchangeability 3) positivity 4) and correct model specification. Exchangeability, also known as the no unmeasured confounding assumption, is closely linked to causality?? Several studies have considered violations of exchangeability and corrected model specification. Positivity has received less attention because in typical observational study positivity violations are not suspected explain why. In the clinical context that we consider, protocols (give some examples, like Platt 2012) threaten to violate the positivity assumption and we investigate whether MSMs are robust against positivity. The focus of this thesis will be on violations of the positivity assumption. Positivity means that within every strata spanned by the confounders, there must be a positive probability of patients being exposed or unexposed to treatment. For example, in a medical context, if treatment protocols demand that treatment is initiated whenever a prognostic variable falls below a pre-defined threshold, there will only be exposed and no unexposed patients in this strata of the confounding prognostic variable. make decisions based on protocols positivity can be. In the absence of structural positivity violations, there is always the threat that random zeroes arise in some strata of the confounder especially when the sample size is small or the number of confounding variables is large. In each case the sparsity of data within the strata of the confounder results in a high chance that positivity is violated. Positivity violations increase the bias and variance of estimates of the causal effect but the extent of the damage is not well known. The central

aim of this thesis will be to investigate positivity violations when fitting MSMs to longitudinal data. To our knowledge positivity violations have not been systematically studied in the literature from a simulation point of view. We quantify the bias and variance introduced due to positivity violations and hope to provide practical advice to researchers tempted to fit MSMs to overcome confounding without realising the potential consequences of positivity violations in their data.

Throughout this thesis we focus on clinical applications as examples. In the literature on marginal structural models the causal effect of Zidovudine on the survival of HIV positive men is often cited as an example. In this example a patients white blood cell (CD4) count is a prognostic variable that influences a doctor's decision to initiate treatment while at the same time being a predictor of survival. As a result CD4 count is a confounder. In the longitudinal setting previous treatments influence CD4 count. As such studies often depend on protocols which means that poistivity in some levels of the confounder make this a suitable example for our purposes.

The structure of this thesis is as follows. In section 2 of part 1, the model considered in this thesis and its important aspects are explained. In part 2 simulating from this statistical model is discussed in detail. In part 3 the model under dynamic strategies is considered and comparisons are drawn with the static case. In part 4 we entertain violations of positivity in the data, this section represents the novelty in this thesis. Part 5 conducts a simulation study. Part 6 includes a discussion, conclusions and suggestions for future work.

A second consequence is that simulating data from a specific marginal structural models is more challenging when the data is to exhibit time dependent confounding.

Look through literature for applications of MSMs

Allow for the joint determination of outcomes and treatment status or omitted variables related to both treatment status and outcomes (Angrist 2001).

A covariate L is a confounder if it predicts the event of interest and also predicts subsequent exposure. Explain how this actually happens, as U0 is a common ancestor of A through L and also Y, also that there is selection bias, and L is sufficient to adjust for confounding see Havercroft algorithm code page bottom.

0.1 Outline

In this thesis we assess the impact of violations of the positivity assumption on the performance of marginal structural models.

The first chapter introduces marginal structural models and inverse probability of treatment weighting. Particular attention is given to the role of the positivity assumption in MSMs and the trade-off between finer confounding control and positivity. Chapter 2 explains the issues surrounding simulating data from a specific MSM with a longitudinal structure that captures the issues which arise in time dependent confounding. This chapter includes a literature review of algorithms that have been developed to simulate from a given marginal structural model. A particular simulation algorithm which is versatile enough that it can be used to introduce violations of positivity is then selected and explained in more detail. Chapter 3 presents simulation results and key findings. Chapter 4 uses real world data in which positivity violations arise as a result of treatment protocols in a chemotherapy trial. The last chapter concludes and provides limitations and directions for future work.

0.2 Software

All simulations and analysis carried out in this thesis use the Python programming language and are provided with this thesis. Several modules were used to extend the base Python language and these are highlighted in the code where appropriate. The *survey* and *ipw* packages written in the R programming language were used to provide functionality not currently available as a Python module. These packages are freely avilable through the Comprehensive R Archive Network (CRAN). Combining R functionality in Python code is made possible through the *rpy2* Python module.

All functions used for this thesis are provided in appendices. Appendix ? contains the code for generating data from the chosen marginal structural model and performing monte carlo simulations. Appendix ? contains the code used to generate the results and graphs in this thesis.

1 Marginal structural models

Marginal structural models (MSMs) are a class of models for the estimation of causal effects from observational data Robins et al. (2000). In this chapter we introduce the central concepts and assumptions behind MSMs as well as the notation that will be used throughout this thesis.

- describe other methods like g-formula as alternative to MSMs
- MSMs are models for some aspect (like the mean) of the distribution of counterfactuals.
- Marginal structural models use only observed data and a set of assumptions to investigate causal effects.

1.1 Counterfactuals and causality

In the counterfactual framework (Neyman (1923), Rubin (1978), Robins (1986)) the causal effect of treatment X on outcome Y for a subject can be defined as the difference between that subject's outcome when they are exposed to X and the same subject's outcome when they are unexposed to X. In other words, one outcome is necessarily counterfactual because the same subject cannot be both exposed and unexposed. If we denote the outcome for subject i when exposed as $Y_i^{x=1}$ and the outcome when not exposed as $Y_i^{x=0}$ then the causal effect can be expressed as $Y_i^{x=1} - Y_i^{x=0}$. For example, suppose a subject with a headache takes ibuprofen ($X = 1$), a popular treatment for headaches. After a suitable amount of time, say one hour, the headache either remains $Y_i^{x=1} = 1$ or has passed $Y_i^{x=1} = 0$. If the subject had not taken ibuprofen the outcome would be either $Y_i^{x=0} = 1$ or $Y_i^{x=0} = 0$.

Often we are interested in the average causal effect for a population rather than for one subject. Suppose sixty subjects are suffering from a headache and everyone was given ibuprofen. After one hour each subject i will either have a headache ($Y_i^{x=1} = 1$) or their headache will have passed ($Y_i^{x=1} = 0$). The average outcome across all subjects is $\mathbb{E}(Y_i^{x=1})$ or equivalently when Y is a dichotomous variable, $\mathbb{P}(Y_i^{x=1})$. The relevant causal comparison is now between $\mathbb{P}(Y_i^{x=1})$ and $\mathbb{P}(Y_i^{x=0})$. As all sixty subjects were exposed we do not observe the quantity $Y_{x=0}$ for any subject, and consequently we do not observe the quantity $\mathbb{P}(Y_{x=0})$. The fact that the counterfactual outcome for a subject is never observed has led several authors to cast causal problems in terms of missing data problems where the counterfactual outcome is viewed as missing (Ding and Li (2017), Howe et al. (2015), Edwards et al. (2015)).

1.2 Confounding

Continuing the headache example, if group A consists of thirty of the headache sufferers who all took ibuprofen, we would ideally compare the quantity $\mathbb{P}(Y_{x=1}|X = 1) = \mu_{A_{x=1}}$ with the quantity $\mathbb{P}(Y_{x=0}|X = 1) = \mu_{A_{x=0}}$. As $\mu_{A_{x=0}}$ is not actually observed, we could instead compute the observable quantity $\mathbb{P}(Y_{x=0}|X = 0) = \mu_{B_{x=0}}$ from the remaining thirty subjects who did not use ibuprofen and belong to group B. Replacing the comparison between $\mu_{A_{x=1}}$ and $\mu_{A_{x=0}}$ with the comparison between $\mu_{A_{x=1}}$ and $\mu_{B_{x=0}}$ will have a causal interpretation if $\mu_{A_{x=0}} = \mu_{B_{x=0}}$. In other words, if a subject from group B can be viewed as an analogue of a subject from group A had they, contrary to fact, not received ibuprofen.

If $\mu_{A_{x=1}} \neq \mu_{B_{x=0}}$ then the comparison $\mu_{A_{x=1}} - \mu_{B_{x=0}}$, a measure of association, is confounded for $\mu_{A_{x=1}} - \mu_{A_{x=0}}$, a measure of causal effect (Greenland et al. (1999)). For example, if all the

subjects in group A are male it would be reasonable to ask whether their sex influenced their decision to take ibuprofen. Suppose that males also tend to have headaches of a shorter duration so that at the end of one hour they are less likely to have a headache than females. The result is that both the decision to take ibuprofen and the probability of having a headache at the end of one hour are dependent on the sex of the subject. This obscures the causal effect of ibuprofen on headaches because there is a spurious association between X and Y through the subject's sex. We cannot establish whether the outcome is due to a causal relationship between ibuprofen and headache alleviation, a relationship between sex and headache alleviation or a mixture of the two.

One explanation for confounding is .... missing covariates (refer to greenland paper where confounding variables are one reason why there is confounding.)

Closely related to confounding, exchangeability is the assumption that the distribution of the counterfactual outcomes Y_x is independent of the actually observed treatment X. When exchangeability holds, subjects from group A and group B are exchangeable in the sense that were they all to remain untreated the distribution of the counterfactual outcomes Y_x would be the same in the two groups Daniel et al. (2013). Imagine exchanging a subject from group A with a subject from group B where both recieve the treatment prevailing in their new group. Under exchangeability, the average outcome in the two groups is unchanged Imbens and Rubin (2008). However, exchanging subjects between group A and group B introduces females into group A and males into group B. As males have a higher probability that $Y = 0$, exchanging subjects changes the distribution of the counterfactuals. The relationship between confounding and exchangeability is why the assumption of exchangeability is also called the assumption of "no unmeasured confounding".

- set-up why comparisons are possible within strata and then averages across strata and why this means that naive methods for addresing time fixed confounding work
- Hint that the finer the confounding control the more accurate the analysis but this has consequences for positivty.
- from pearl 2001: namely, that if we compare treated vs. untreated subjects having the same
- values of the selected factors, we get the correct treatment effect in that subpopulation of subjects.
- Explain clearly why it is a bias.
- Also explain why we want to be judicious in our choice of number confounders to control for. Can't include everything (Dawid 1979 on this)
- Explain that structural parameters only coincide with associational parameters under exchangeability.
- Define what the naive analysis is - analysis without adjustment.
- knowing the value of Z gives us no more information about the distribution of the counterfactuals Y_x
- Explain that in a randomized experiment exchangeability is guaranteed because X is automatically noy related to any other variables.
- Randomization ensures that missing values occur by chance. So the counterfactual values that we don't see for some observations are missing randomly and not due to confounding through a covariate.
- Any residual confounding cannot be due to the variables that we have conditioned on.
- hernan 2011 "we say that positivity does not hold because for some confounder values there are no treated and untreated subjects to be compared"
- link to splines as a way of reducing residual confounding see cole 2008

- When confounding is present we cannot simply substitute or exchange the exposed cohorts experience for the unexposed cohort.
- confounders are simply covariates which explain why confounding is present (see Greenland 1996)
- By conditioning on a variable (or a set of variables) C we will mean examining relations within levels of C (i.e. within strata defined by single values of C) (see Greenland 2011)
- conditioning and adjustment not the same thing. Control is used as a synonym for asjutment.
- in randomization, confounding is absent in expectation
- explain how we standardize measures across strata but weighted combination and link to standardization itself
- use conditional exchangeability to show why we can standardize across populations
- Also allows a way in to positivity by looking at Technical Point 3.1 in hernan robins causality book.
- explain residual confounding so that it can be reffered to later when estimating weights - for example explaining why splines help to remove any residual confounding.
- explain that we can use other effect measures not just the difference.

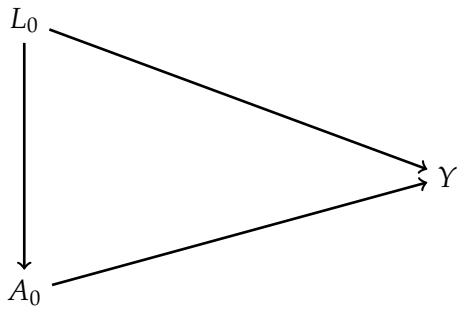
1.3 Directed Acyclic Graphs: graphical representations of causality

Causal relationships, like those described in the previous section, 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 Pearl (2009). The absence of an edge between two variables indicates that the variables are independent of one another. 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 directly or indirectly Hernán et al. (2004).

For example, figure ? represents the case where interest is in the causal relationship between treatment X and outcome Y. Treatment is assigned according to conditional distributions $P(treatment|male)$ and $P(treatment|female)$. Once treatment has been assigned, the outcome Y is determined by both X and Z by the conditional distribution P(Y | X, Z). **?**Pearl2001}, **?**Pearl2014}. Blocking or screening off Z has the same inuition as explained in the section on confounding. The causal effect of X and Y cannot be different between two subjects because of Z when everyone is that strata has the same value of Z. Blocking is the same as holding Z constant. Intuition to drive forward is that difference in outcome cannot be due to strata when everyone shares that strata.

Show the same graph without causal relationship between X and Y. There is a marginal dependence between X and Y through Z, but once we condition on Z this dependence dissapears as shown by the lack of an edge between X and Y. Once we condition on Z (i.e. we know that the subject was male of female.) then the margial dependence dissapears. Introduce idea of common cause here as well.

Both treatment and outcome are determined by sex leading to a spurious association between X and Y through Z. This is called a "back door" path between X and Y. Conditioning on Z is reprsented graphically by blocking the back door and any spurious assocations to allow causal estimation



Causal graph

- use example of cause by just removing an arrow from the DAG to illustrate the point that there must be a cause so adding back the causal arrow of interest does not change the fact that part of the cause comes through the confounder.
- common cause and structural approach to selection bias paper.
- the absense of an arrow means no direct effect between two variables cole 2009 illustrating bias paper.
- Didelez 2010 and Pearl 2009 for connecting DAGs to probability distributions and factorizations.
- explain how colliders block relationships so that variables are independent if there is a collider on their path

1.4 Time dependent confounding

So far we have considered the time fixed context in which treatment and confounders take on a single value. It was sufficient to block the "back door" path between the treatment and outcome by conditioning on the confounding variable(s). In the headache example, the causal effect of ibuprofen on headache alleviation was confounded by sex. For most people, sex is a time-fixed covariate because it does not change value over time. To broaden the setting to a time dependent context, we adopt the canonical example of the causal effect of Zidovudine (AZT) on mortality amongst human immunodeficiency virus (HIV)-infected subjects Hernan et al. (2000). In this example, subjects are measured at baseline $t = 0$ and at subsequent visits. In each visit the patient's CD4 lymphocyte count is measured and a treatment decision made. Survival at the end of follow-up is a binary outcome equal to 1 if the patient has died and 0 otherwise.

The time-fixed notation can be extended to include subject histories for time varying variables. Treatment and covariate histories up to visit k are can be represented by an overhead bar. For example, $\bar{X}_k = \{X_0, \ldots, X_k\}$ represented the vector of treatment decisions while $\bar{Z}_k = \{Z_0, \ldots, Z_k\}$ represents the vector of measurements on the time dependent-confounder Z. Time-fixed covariates like sex, or covariates which change linearly over time like age are tyically 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 patient is dead and 0 otherwise.

Just as in the time-fixed case, time-dependent confounders lead to spurious associations between X and Y through a "back door" path between X and Y through L. To estimate a

causal effect it is necessary to block this path by conditioning on the confounding variables. Figure ? gives an example of this in the time dependent case for two periods ($t = 0, 1$). In the first period a treatment decision is made based on the measured confounder Z_0. In the second period ($t = 1$) a new treatment decision is made based on both Z_0 and Z_1. Conditioning on $\bar{Z}$ under this DAG leads to a consistent estimate of the causal effect because doing so blocks all paths between X_0 and X_1 and Y except the causal path of interest.

However, the time-dependent context also admits structures like the middle pane of figure ? with the addition of a causal relationship between X_0 and Z_1. It is now possible for current treatments to be a determinant of future confounders which are in turn determinants of future treatment Robins (2000). As a result the effect of A_0 on Y is mediated through L_1 in the path $A_0 \rightarrow L_1 \rightarrow Y$. Blocking this path by conditioning on Z also blocks some portion of the effect of A_0 on Y and will lead to a biased estimate.

A second danger in the time-dependent context arises when Z is a common effect of treatment and an unmeasured variable U which also influences the outcome Y. There is no direct association Figure ? shows the same two structures with the addition of a an unmeasured variable U which influences Z and Y. Conditioning on Z. Selection bias precludes unbiased estimation Hernán et al. (2004). There is a mediating relationship between A and Z in which case there is a spurious relationship between A and Y again? This is less inuitive and so examples are best according to Cole et al. (2010). We can say that Z is a common effect of A and U, once we condition on Z we create a dependence of A on U. U is a cause of Y and hence there is an association between A and Y. This association is present even when there is no direct causal path between A and Y.

Hazard ratios and selection bias Hernán (2010). Actual application will look at toxicity of treatment. Some people will be suceptible and drop out leaving more people in the untreated arm of the study. presumably because in any population some people are more suceptible than others.

general point about selection bias is that the general population is not a valid control group. This is interesting because it links very closely with the counterfactual approach which defines the causal effect, not with references to a population but to an individual.

Conclusion, 1) clearly a different technique is required for analysis 2) the nature of time dependent case needs to be described fully enough to explain why we choose the simulation algorithm that we choose and any holes in it. In subsequent sections our choice of simulation algorithm will be motivated by the structure of time dependent confounding as well as the viability of introducing positivity violations which are propogated through the time dependent structure. Explain meaning of a collider and that a collider that is conditioned on will not block confounding. Essentially with this kind of data we cannot use confoudning or stratification methods.

- Intuition from Pearl 2009 book pp. 17 also on schools. More intuition in cole 2010
- Simpson's paradox linked to making comparisons within strata - collapsibility
- explain that we are often interested in parsimnonious models so cannot have all covariates U that will create associations between X and Y
- explain why we do not need to worry about the path between A_0 and A_1
- explain why mediation is likely to occur in example.
- explain why saturated models cannot be used because they will have
- inituitive examples of selection bias.
- Saturated models are not an option because they would be computationally intensive and so we use parametric models which also links to positivity because we smooth over zeroes

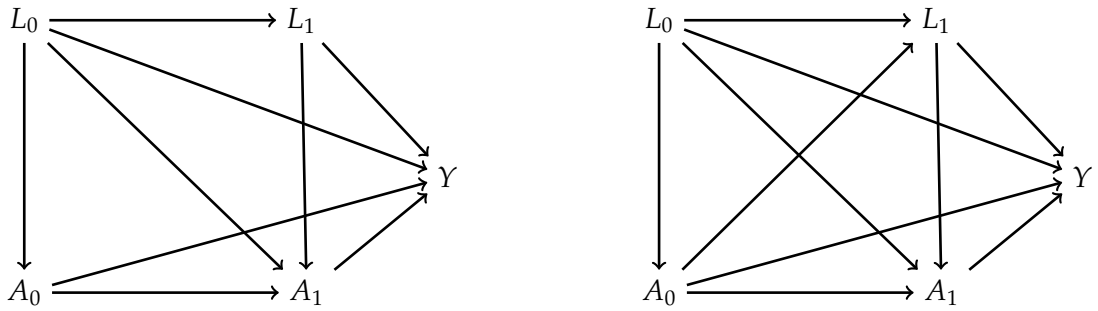


Figure 1 DAG

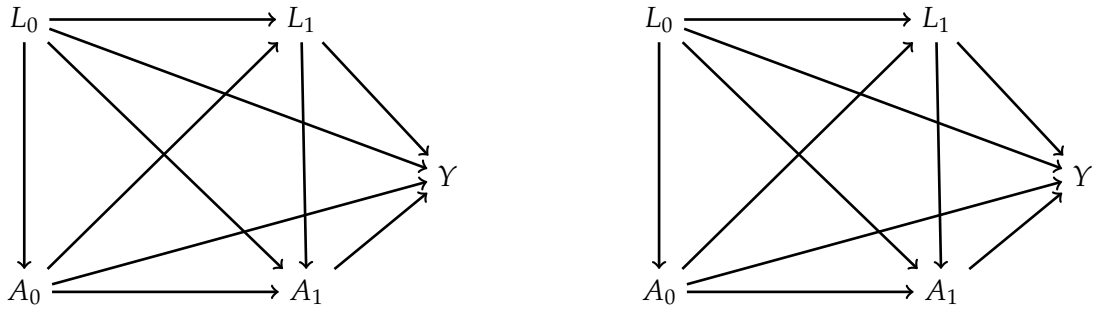


Figure 1 DAG

in certain strata.

- Explain why hazard ratios have a built in selection bias after giving some examples of why selection bias arises. It is because it is selective on patients reaching the time period in question. Is this also the reason why summary methods create selection bias.
- give an inuitive explanation for why CD4 count is a predictor of subsequent treatment and of death.
- Because treatment is randomized (at baseline) in expectation the proportion of men and women in each group is the same.
- The reason why summary values are a problem for selection bias is in Robins et al. (1992) is similar/the same as hernan on hazards of hazard ratios.
- explain that Z is a mediator variable and also explain colliders and why conditioning on a collider creates an association between A and U and hence A and Y

1.5 Inverse Probability of Treatment Weighting

The previous section has highlighted how standard approaches for controlling for confounding in a time dependent context may lead to biased estimates. In this section we describe a technique called inverse probability of treatment weighting that can be used to obtain unbiased estimates of the causal effect of treatment on outcome in the presence of time dependent confounding.

Inverse probability of treatment weighting is a technique that can be used to obtain unbiased estimates of the causal effect of treatment on outcome in the presence of time dependent confounding. The intuition behind the technique is that by re-weighting the data a pseudopopulation is created in which the treatment is independent of any measured confounders. Regression analysis on the pseudopopulation can be carried out without the need to control for confounders

eliminating the problems which arose in the previous section due to conditioning on Z. Crucially, in the pseudopopulation, the causal effect of X on Y remains unchanged. As a result, it is possible to estimate the true causal effect of X on Y.

Construction of weights

- why does it work and naive methods do not?
- How does it break the link between X and Z?
- Dangers associated with startification and controlling for methods highlighted already explain why we need other methods - explain a few of these like G-estimation, SNTM etc.
- creates a pseudo-population in which we have something similar to an experimental setting
- Because we need weights this means we need a model for the weights - model can be non parametric or parametric depending on data used.
- Contrast IPTW methods with stratification methods.

Horvitz and Thompson (1952)

Areas where IPTW has been used (time dependent confounding, comparing dynamic regimes, missing data)

Inverse probability of treatment weighting is a technique that re-weights subject observations to a population where assignment of treatment is at random. An early example of this technique is the **?** weighted estimator of the mean. In the context of marginal structural models, a weight is calculated for each subject which can be thought of informally as the inverse of the probability that a subject receives their own treatment Robins et al. (2000). The result of applying these weights is to re-weight the data to create a pseudo-population in which treatment is independent of measured confounders Cole and Hernán (2008). Crucially, in the pseudo population the counterfactual probabilities are the same as in the true study population so that the causal RD, RR or OR are the same in both populations Robins et al. (2000).

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

stabilized weights

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

- why can weights be very unstable?
- show why the stabilized weights have a mean of 1 by applying law of iterated expectations. Hernán and Robins (2006)
- see appendix 1 of cole 2008 for good informal/intuitive explanation of stabilized weights.
- describe no-parametric way of estimating numerator P(A=1) by cases/total subjects or saturated model with just an intercept.

The use of IPTW is valid under the four assumptions of consistency, exchangeability, positivity and no misspecification of the model Cole and Hernán (2008).

Informally a patients weight through visit k is proportional to the inverse of the probability of having her own exposure history through visit k (Cole and Hernan 2008)

The weight is informally proportional to the participants probability of receiving her own exposure history

As these weights have high instability we need to stabilize them. The unstabilized weights can be driven by only a small number of observations. Why are they unstable?

- true weights are unknown but can be estimated from the data.
- A_t is no longer affected by L_t, and crucially the causal effect of $\bar{A}$ on Y remains unchanged

Be more specific about what is contained in the weights. The denominator depends on the measured confounders L the numerator does not.

- weighted regression and MSM are equivalent.

Point out that we need baseline variables in the conditional statments in the num and denom of the weights otherwise we break the relationship between outcome and baselines in the new pseudo-population. If the baseline variables are not confounders, then we do not want to break this relationship. Baseline covariates also help to stabalize the weights (how?)

importantly, changing the relationship between L and A, won;t change the relationship between L and Y. This means that an intervention in A does not affect the relationship between L and Y. So we remove the link between L and A and assign to A the value of treatment on or off. Once we place the patient on treatment, regardless of the relationship which had existed before hand between the covariate and treatment, a new relationship between A and Y exists in which the covariate has no say.

- stabilized weights should have a mean of 1

1.5.1 Applications of IPTW

Different from application of MSMs - IPW is a technique which has been applied to MSMs - i.e. MSMs are on example of how IPTW can be used.

- Crucial that the relationship between Y and X remains the same in the new population. I.e. the marginal structural model is the same.

Have been used for missing data problems. see pp.442 of Hernan, Brumback, Robins 2001 for a list of papers linked to this

- many types like a marginal structural cox model (maybe let this follow on after weights part.

1.6 Assumptions

This section formalises the five assumptions under which inverse probability weighting can be used to correctly estimate MSMs. The first three assumptions of consistency, correct model specification and no measurement error are dealt with briefly. The exchangeability . There is a trade off between finer confounder control and positovity. As a result, this section will mainly focus on exchangeability and positivity assumptions with only brief exposition of the consistency, correct model specification and measurement bias assumptions. Where necessary the reader is referred to further work on these assumptions. The assumption of no unmeasured confounders has received the most attention in the literature (). Most attention will be given to positivity. Conditions under which IPTW work are largely untestable (westreich 2012)

1.6.1 No unmeasured confounders

The assumption of no unmeasured confounders has already been discussed with respect to confounding

$$Y_x \perp\!\!\!\perp X$$

- necessity of identifying the most important confounders ,
- conditional exchangeability.
- this assumption is not empirically verifiable.
- no unmeasured confounding in the simulation

1.6.2 Consistency

The consistency assumption states that the actual outcome Y_i^{obs} for a subject i is equal to the potential outcome Y_i^x when the treatment received by subject i is x, that is $X_i = x$ (Cole and Frangakis (2009)). The consistency assumption is required to make inferences about y^x using observational data because it connects the observational data to the potential outcomes. For example, if it was known for a subject i that ibuprofen did not alleviate headaches, then $Y_i^{x=1} = 0$ is the potential outcome associated with these events. However, if the same subject drank a glass of water along with the ibuprofen and this action did alleviate the headache, then $Y_i^{obs} = 1$ and the potential outcome is not equal to the observed outcome despite the fact that $X_i = 1$. In other words, the consistency assumption rules out side effects of exposure and anchors the observational outcome to the potential outcome framework. As a result, expressions involving probabilities of counterfactuals can be cast in terms of ordinary conditional probabilities of measured variables Pearl (2010) and equated as follows.

$$P(Y^x = y \mid Z = z, X = x) = P(Y = y \mid Z = z, X = x)$$

In words, the probability that $Y^x = y$ when $Z = z$ and $X = x$ are observed is equal to the conditional probability that $Y = y$. The consistency assumption has generated discussion over whether it is really an assumption an axiom or definition. The interested reader is referred to Vander Weele (2009), Cole and Frangakis (2009) and Pearl (2010).

1.6.3 No model mispecification

Estimating MSMs with a continuous confounder involves specifying a model for the weights and a structural model relating exposure to outcome. In both models, correct specification is required 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 Cole and Frangakis (2009). The weight model also needs to be correctly specified in order to consistenly estimate the weights.

For the weight model, the stabilized weights should be one. Parametric models are unlikely to be perfectly specified but should provide a good approximation of the true model. Deviations from one are an indication that the weight model is mispecified.

Several studies have examined the effect of model mispecification in the estimatuion of MSMs including Cole and Frangakis (2009). Some key findings. Broadly, the idea is to simulate data from a known MSM weight model and introduce realistic deviations from the model.

Correct specifcation of the inverse probability weighting (IPW) model is necessary for consistent inference from a marginal structural Cox model (MSCM).

- link to why simulation studies make it possible to isolate model specification as an error.
- stabilized weights should have a mean of 1 - indicative of model mispecification if they do not
- results in some confounding if the model is mispecified, think about using splines to mop up residual confounding
- IPW cannot correctly adjust for all confounding under model miscpecification

In this thesis we simulate data from a known weight and structural model. In other words, we simulate data in the absence of model mispecification elimiating this as a source of bias. Importantly, this makes it possible to isolate effects of interest without worrying about model miscpecification.

1.6.4 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 carlessnes 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. Analgous to the case of no model miscpecification, 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 Hernán and Cole (2009).

1.6.5 Positivity.

To conclude this chapter ....

- use intuition of making comparisons within strata as a means to introduce positivity and link to weight model and standardization as examples.

The final assumption underlying MSMs, and the central topic of this thesis, is the positivity assumption. MSMs are used to estimate average causal effects in the study population, and one must therefore be able to estimate the average causal effect in every subset of the population defined by the confounders Cole and Hernán (2008). The positivity assumption requires that there be exposed and unexposed individuals in every strata of the confounding covariates. For example, when treatment is Zidovudine and CD4 count is the confounder, there must be a positive probability of some patients being exposed and unexposed at every level of CD4 count. Positivity can be expressed formally as $Pr(A = a \mid L) > 0$ for all $a \in A$, which extends straightforwardly to the time dependent case where the positivity assumption must hold at every time step conditonal on previous treatment, time dependent confounders and any baseline covaraiates:

$$Pr(A_{it} = a_{it} \mid L_{it}, A_{i,t-1}, V_{i0}) > 0$$

Models for the risk $P(Y = 1 \mid A = a, L = l)$ are commonly studied in epidemiological applications. Applying basic probability rules reveals that the risk can be re-written with the term $Pr(A = a \mid L = l)$ in the denominator:

$$P(Y = 1 \mid A = a, L = l) = \frac{P(Y = 1, A = a, L = l)}{Pr(A = a, L = l)} = \frac{P(Y = 1, A = a, L = l)}{Pr(A = a \mid L = l)Pr(L = l)}$$

This model is only estimable when $Pr(A = a \mid L = l) \neq 0$. Therefore, when positivity does not hold it is not possible to estimate the model. In the context of MSMs a similar problem emerges. Although weighting via IPTW allows naive estimation of (?) without including the confounders, the weights in (?) involved the term $Pr(A = a \mid L = l)$ in the denominator. This means that the weights are inestimable whenever positivity is violated. In order to estimate the causal effect of A on Y, weights must be estimable in every subset of the population otherwise the average causal effect in the study population cannot be estimated.

In practice, positivity can arise when random zeroes or structural zeroes are present in some levels of the confounding covariates. Random zeroes arise when, by chance, no individuals or all individuals, receive treatment within a certain strata as defined by the covariates. For example, Cole and Hernán (2008) studies positivity violations in individuals in strata defined by CD4 count and viral load. By increasing the levels of CD4 count the chances of random zeroes also increases and Cole and Hernán (2008) show that the IPT weights rapidly lose their stability with the consequence that causal effects are no longer estimable. Researchers applying IPTW methods must actively check that there are both treated and untreated individuals at every level of their covariates within cells defined by their covariates because parametric methods will smooth over positivity violations and not provide any indication of nonpositivity. Increasingly refined covariates are attractive because they provide better control of confounding, but the point that Cole and Hernán (2008) make is that this control needs to be traded off against increased occurence of random zeroes and subsequent instability of IPT weights.

More relevant to this thesis are violations of the positivity assumption due to structural zeroes. These occur when an individual cannot possibly be treated or if an individual is always treated within some levels of the confounding covariate, as is the case in the clinical protocol example motivating this thesis. Several studies give examples of structural violations of the positivity assumption in epidemiological contexts. In Cole and Hernán (2008) structural zeroes arise when the health effects due to exposure to a chemical are confounded by health status proxied by being at work. If individuals can only be exposed to the chemical at work then all individuals not at work will be unexposed. A second example is liver disease as a contraindication of treatment. If individuals with liver disease cannot be treated then all individuals in the "liver disease = 1" strata will be untreated. In Messer et al. (2010) structural zeroes arise in the context of rates of preterm birth and racial segregation, whereas Cheng et al. (2010) find structural zeroes in the context of fetal position and perinatal outcomes. Our motivating example is most closely related to liver disease as a contraindication, except that the clinical protocols require that patients with low CD4 count always be treated instead of never being treated, as in the case in the liver disease example.

Although in many epidemiological settings the positivity assumption is guaranteed by experimental design, studying positivity violations is relevant because, as our own motivating example and the examples above suggest, structural violations do occur, and random zeroes are always possible especially at finer levels of confounding covariates. Studying the finite sample propoerties of MSMs under violations to positivity is therefore an important issue which is yet to be

dealt with systematically in the literature. As Westreich and Cole (2010) points out, positivity violations, positivity violations by a time varying confounder pose an analytic challenge and they suggest g-estimation or g-computation may be a way forward. A good start to dealing with the time varying confounder case is to see how well MSMs work when positivity is violated. This is also a novelty of this thesis.

6. estimated weights with a mean far from one, or very extreme values indicate either non-positivity or model mispecification of the weight model.
7. It is not always true that we want more finely tuned covariates for confounder control because the bias and variance of the effect estimate may increase with the number of categories. This is similar to the positivity masking example.
8. Our results are equally valid for other circumstances in which positivity may arise.
9. Also think about how the number of categories of exposure increases the chance that one level of exposure will have a positivity.
10. Westreich and Cole 2010 have suggested that methodological approaches are needed to weigh the resultant biases incurred when trading of confounding and positivity. The framework we use is flexible enough to allow this in a simulation setting.

If the structural bias occurs within levels of a time-dependent confounder then restriction or censoring may lead to bias whether one uses weighting or other methods (Cole and Hernan 2008). In fact, weighted estimates are more sensitive to random zeroes (Cole, Hernan, 2008) Introducing violations of positivity can be achieved by censoring observations.

But to give an intuitive example, think about how it links back to a situation where sicker patients receive treatment compared to others. So in the "sick" strata of the CD4 count **ALL** patients receive treatment which inflates the IPTW. This also affects how we think about the associational versus causal models. The causal effect might be 50/50 but because sicker patients get treatment the mortality ratio in the treated group is likely to be higher.

The trade-off between positivity and confounding bias is emphasized in Cole2008

Why is practicality important? Cole paper highlights practical advice to practictioners. positivity can be violated in a practical setting because of two few strata, it can be the result of protocols in a clinical setting and it can be seen as a trade-off between exchangeability (and we need more measured predictors to maintain exchangeability) and positivity where more predictors leads to more likely a zero problem.

- Dynamic stratgies evaluated using MSMs will have rules like, start treatment if CD4 falls below a certain threshold. See Didelez presentation on this
- Explain that there is positivity in estimation of the structural model and also in the weight model. The reason why positivity is more important in the weight model is because when the weights are unstable the estimates can be very wrong as a result.
- read and add Naimi et al. (2011) who have already done a simulation study of non-positivity
  - in context of healthy worker effect
  - explain why this doesn't get to the heart of the issue of non-positivity
  - a relevant question highlighted by Naimi et al. (2011) is whether the effect of non-positivity is amplified with more than two time periods.
  - explain why healthy worker effect and positivty go neatly together because as a result of people dropping out there will be empty cells/strata.

A model that parameterises $P(Y \mid do(A = a))$ is called a marginal structural model (MSM) as it is marginal over any covariates and structural in the sense that it represents an interventional rather than observational model.

2 Key concepts in survival analysis

This chapter briefly reviews several important concepts in survival analysis which are pertinent to this thesis and specifically the simulation algorithm that will be used to generate data in later chapters. Survival analysis is the study of the distribution of life times. For example, in. Often we are interested in comparing survival in two or more groups. For example a group which is exposed to treatment and a group which is not. Confounding is also important. Survival analysis is by nature time dependent and hence time dependent confounding is particularly studies in a survival context. The basic concepts reviewed here are a condensed version of those presented in John P. Klein (2003).

- Explain why we do not look at censoring and truncation.
- emphasis on discrete time which is the application.

2.0.1 Survival function

The survival time T is the time between a well defined start point and a well defined end point. For example the time between birth and death. The survival function is the probability that a certain individual survives until time t or equivalently the probability that the survival time T is greater than t,

$$S(t) = P(T > t)$$

. In the continuous case the survival function can be written

$$S(t) = P(T > t) = \int_x^\infty f(t)dt$$

In the discrete case the survival function

$$S(t) = P(T > t) = \sum_{x_j > x} p(x)$$

2.0.2 Hazard function

The hazard function or hazard rate expresses *the "approximate" probability of an individual of age x experiencing the event in the next instant.*

$$h(x) = \lim_{\Delta x \to 0} \frac{P(x \leq X \leq x + \Delta x \mid X \geq x)}{\Delta x}$$

In the discrete case, the probability that the event occurs between $j-1$ and j is equal to the difference in survival of these times $p(x_j) = S(x_j) - S(x_{j-1})$. The hazard function is

$$h(x) = Pr(X = x_j \mid X \geq x_j) = \frac{p(x_j)}{S(x_{j-1})}$$

$$h(x) = 1 - \frac{S(x_j)}{S(x_{j-1})}$$

$$S(x) = \prod \frac{S(x_j)}{S(x_{j-1})} = \prod 1 - h(x_j)$$

Showing that the survival function is determined by the hazard rates.

In words, the discrete hazard function. There are two probabilities, the probability that the death occurs

$$Pr(Y = 1 \mid X = x) = \frac{e^{\beta_0+\beta_1\times x}}{e^{\beta_0+\beta_1\times x}+1} = \frac{1}{1+e^{-(\beta_0+\beta_1\times x)}}$$

- show that hazard function is not collapsible or hazard ratio.
- condition on survival up to a certain point?
- Hazard ratios differ from relative risks and odds ratios in that RRs and ORs are cumulative over an entire study, using a defined endpoint, while HRs represent instantaneous risk over the study time period, or some subset thereof.
- compare hazard rate and hazard function with the hazard ratio. The hazard ratio could be between men and women for example.
- relate survival to hazard function in order to show that survival is completely determined by hazard rate.
- Naturally survival is measured at each follow up point, but as the person is alive until their final follow up then we are always conditioning on Y = 0 until the final follow up.
- a logistic model is one where the log-odds of the probability of an event is a linear combination of independent or predictor variables.
- need to make clear that we could split groups into a control and treatment group if they are exchangeable.

2.0.3 Proportional hazards

- explain multiplicative model
- causal model
- discrete time relationship between cox model and logistic model

Often we are interested in how survival differs between subjects who have been exposed to a treatment versus those who remain unexposed. In a randomised trial exchangeability is guaranteed and the survival function can be calculated. In observational trials the same confounding issues described in the previous chapter require a different approach. Adjusting for other variables is important to remove bias and provide a more accurate fit.

In a typical study we want to adjust for confounders and other variables in order to get a better estimate than just looking at treated versus non-treated patients.

- explain how we let the survival time depend on covariates.
- link to causal framework, introduce marginal structural logistic regression or marginal structural cox proportional hazards model.
- test for equality between (all) groups.

3 Simulating from marginal structural models

In order to assess the impact of violations of the positivity assumption on the performance of the IPTW estimator we simulate data from a specific marginal structural model in a series of monte carlo simulations. Several algorithms for simulating from marginal structural models have been suggested in the literature. To test the effect of positivity violations on the performance of marginal structural models we use the algorithm of **?**. This algorithm has several key features which are described in detail in this chapter. We start with an overview of the logic behind monte carlo simulations in general terms.

In this section we first describe in general terms the logic behind monte carlo simulations. The specific In this chapter we start by describing the logic behin monte carlo simulations in general terms. Next, we consider several important criteria that a simulation model must exhibit in the context of MSMs. In particular we require an algorithm that can simulate from a specific MSM, has the observational structure described earlier and we also define noncollapsibility. Several algorithms have been proposed in the literature and these are briefly discussed and compared. We then focus on the algorithm suggested in and explain why it satisfies our requirements. The most salient aspects of this algorithm for the purposes of this thesis are described.

Monte Carlo Simulations In statistical research, interest often lies in the estimation of a population parameter θ. When only a sample X_1 from the population is available, statistical procedures are applied to estimate the population parameter $\hat{\theta}_1$ based on that sample. The same procedure applied to a second sample X_2 drawn from the same population will result in second estimate $\hat{\theta}_2$, and so on for more samples. We rely on the sampling distribution of the $\hat{\theta}_i$ to draw statistical inferences.

Monte Carlo simulations flip this process on its head. We start with a known true model governed by parameters θ and generate a sample of data according to this model. We then apply a statistical procedure to rediscover the true parameters governing the data generating process. Most often we are interested in the finite sample properties of techniques. In other words, we are interested in how the technique behaves with, say, only one thousand observations rather than one million or one billion. A single sample of one thousand observations we never truly simulated data and check how closely the to Where statistical inference is a process of discovering The properties of a statistical method Rather than being a process of dicoverey we can use simulations as a process of rediscovery and appraise a method by its ability to correctly rediscover the parameter. For example, the effect of a ibuprofe on headaches can be expressed in by an effect measure such as the odds ratio

$$OR = \frac{P(Y = 1 \mid X = 1)/(P(Y = 1 \mid X = 0))}{P(Y = 0 \mid X = 1)/(P(Y = 0 \mid X = 0))}$$

Or the log odds ratio

$$log(OR) = log(P(Y = 1 \mid X = 1)/(P(Y = 1 \mid X = 0)) - log(P(Y = 0 \mid X = 1)/(P(Y = 0 \mid X = 0))$$

When $P(Y = y \mid X = x) = $

$$log\frac{P(Y = 1 \mid X = 1)}{P(Y = 1 \mid X = 0)} = \alpha + \beta \times x$$

These steps can be summarised

1. Specify the articifial population
2. Sample from that population
3. Calculate the parameter of interest
4. repeat steps 2 and 3 a certain number of times
5. draw conclusions.

- show that the logistic model is the log odds model or odds ratio model.
- explain that the cox prop hazard model discrete time equivalent is the logistic model.

So that the log odds ratio between those who received treatment and those who did not is given as

$$log\frac{P(Y=1 \mid X=1)}{P(Y=1 \mid X=0)} = \beta_0 + \beta_1 \times x$$

By specifying β_0, β_1 we can simulate outcome from this model by first generating values for x and then using the values of x to generate Y using calcualting y by rounding probabilities. Each time we sample we get (Y, X) and we can perform logistic regression to see whether the correct parameters are resolved. We could then change something like the model form and see what happens to the model under changes. Or we may just want to see how accurate the model is using finite a finite sample rather than asymptotics, say a sample of n = 1000 observations.

More generally, the steps taken in a simulation study can be summarized:

Explain why simulations still need to be treated critically. Greenland and Maldonado (1997)

- advantage of no measurement error, no model misspecification
- Add in confidence intervals.
- end with why we need to sample from a specific MSM

Noncollapsibility In the example of the effect of ibuprofen on headache alleviation it was reasonable to ask whether that effect differed between male and female subjects. That effect could be captured by effect measures such as the risk difference, risk ratio or odds ratio. If treatment affects the outcome in males and females differently then the effect in the male strata will be different from the effect in the female strata. Marginalizing over the two strata will lead to biased effect measure because comparisons will not be between exchangeable subjects which will propogate up to the effect measure of interest.

On the other hand, if treatment affects males and females equally then it seems reasonable to expect that the marginal effect (after marginalizing over sex - pooling subjects) would also be equal to the male and female specific effects. In other words, if the effect in males is the same as in females then it is reasonable to think that the effect is the same in the total population, which is made up of males and females. In that case, the effect measure is collapsible because sex is unrelated to exposure to ibuprofen and pooling males and females together and collapsing over sex, correct analysis can be performed using a marginal or naive analysis. This extends to the case of many subgroups defined by, for example, age or race. Collapsibility is useful because it helps to reduce the dimensionality and computational effort which arises when many subgroups need to be taken account of in the analysis Didelez et al. (2010).

However, some effect measures are noncollapsible, in the sense that even when their strata specific effects are equal, collapsing over strata the marginal or naive effect measure does not

equal the strata specific effect measures. Simpson's paradox (). Examples of non-collapsible effect measures include the odds ratio and rate difference. The odds ratio is estimated using logistic regression

In a practical setting, the odds ratio may be used and there may be confounding and there may be noncollapsibility. In fact, the ability of the inverse probability to correctly estimate the causal effect of treatment on outcome in the presence of time dependent confounding make it possible to separate the confounding and collapsibility elements and quantfy the extend to which collapsibility affects results Pang et al. (2016).

The relevance of noncollapsibility to this thesis is the need to simulate data from a known MSM model. The conditionals from which we draw data will not typically be collapsible.

is the same it seems reasonable that the effect in the population would also be the same. Naturally the difference could be due to the differening effect in the population in which case confounding is present. the same we could conclude that sex made no difference and look at marginal associations instead. Reducing the number of variables required in analysis is often useful because it reduces dimensionality and computational effort allowing subgroups to be pooled together

An effect measure like a rate difference, rate ratio or odds ratio is collapsible when the effect is equal in both strata and equal to marginal (over C) effect. Collapsibility is a useful property because it means that analysis can be carried out on a subset of variables after marginalizing over the others . In this case that could mean marginalizing over sex and looking at just the relationship between Y and X.

- see greenland 1999 collapsibility regression formulation for an example in regression context.
- mixed up with confounding and concluded that there is no confounding when the conditional odds ratio equals the conditional odds ratio.

An effect measure for the association between Y and X such as an rate ratio or an odds ratio is noncollapsible when conditioning on a covariate a to Y changes the size of the odds ratio even when . Collapsibility is useful when

- Explain that the effect of interest depends on the question (also cite Pearl (2014) on this)

An effect measure is non-collapsible across strata defined in the analysis if the constant effect measure does not equal the strata specific effect measure. greenland 1996

example of randomized trial, there being no confounding because it is a randomized trial, same result in men and women, different result in whole population. Not confounding because we eliminate confounding by randomization.

Greenland 1999 - if the model with Z is correct then the model without Z is unlikely to have the same coefficient

Noncollapsibility arises when the marginal effect measure (marginal over any covariates, i.e. unstratified or with no confounder control, crude) is not equal to the strata specific effect measure.

- explain why it is a probelm for marginals versus conditionals with (i.e. won;t be the same MSM)
- use cox ph model as example.

This is a problem when simulating from marginal structural models because the correct marginal structural model

$$correct marginal structural model$$

$$model with covariates$$

$$collapsed over covariates and not equal to marginal model$$

Collapsibility starts with the notion of confounders. We assume that within strata of confounders that the effect of the confounder is homogenous. I.e. in the female strata, the effect of being female is homogenous.

Greenland (1996), Greenland et al. (1999), Greenland and Pearl (2011), Sjölander et al. (2016)

The effect of treatment on diesease outcome may be unconfounded but noncollapsible

Collapsibility is the same as simpson's paradox if we adopt the definition that without the conditional variable they can be equal.

collapsibility depends on the measure used. Some are collapsible and some are not

Could arise in two ways 1) within strata effect measures may not be the same 2) Even if they are the same they may not equal the marginal effect measure (marginal over any covariates Z)

Collapsibility means there is no incompatibility between the marginal model and the conditional distributions used to simulate the data. Provide example of this. Explain how this affects the simulation algorithm. Especially hazard ratios which are non-collapsible.

Models are noncollapsible when conditioning on a covariate **related to the outcome** changes the size of the estimate even when the covariate is unrelated to the exposure. Illustrate why this happens with survival models.

Survival models are non-collapsible. Hence we cannot eaily simulate from them. Instead we use U as a sneaky trick. Explain why survival models are non-collapsible - through the hazard function.

This is particularly important because collapsibility and confounding are often treated as identical concepts when in fact they are not. Greenland et al. (1999)

- relevance to the algorithm? How does this work with a specific MSM.
- Show why hazard ratios are not collapsible.
- Explain why models with product terms are clearly not collapsible
- relation to a hazard function
- link to lack of exchangeability

Observational structure The previous chapter included a discussion of the role of confounding covariates and time dependent confounding. In the time fixed context, blocking the path from a confounding covariate by conditioning was sufficient to consistently estimate the effect of treatment on outcome in the absence of unmeasured confounders. In the time dependent context, conditioning on confounding covariates may biases estimates in two ways. First, in a time dependent context, both treatment and covariates can change over time. Some effect of treatment may be transmitted through the covariate to the outcome or future treatment. Blocking this path also blocks the indirect effect of treatment on outcome through the confounding covariate. At the same time, selection bias arises due to conditioning on a colider. The hazard ratio, the ratio of the hazard function/rate at two levels of an explanatory variable, typically exhibits selection bias. Failing to condition

- why important for observational studies
- examples of it in observational studies
- explain why we need algorithm with this structure.

Simulation algorithms literature review. Several algorithms for simulating data from a specific marginal structural model have been suggested in the data. Here we briefly summarise the discussion in Havercroft and Didelez (2012) on competeing algorithms and then highlight a number of algorithms suggested subsequently to that paper.

- briefly summarise the litrature in Havercroft and Didelez (2012)
- update with more recent algorithms like Young and Tchetgen Tchetgen (2014)
- discuss Naimi et al. (2011) limitations of simulation algorithm. look into more detail for how this algoirthm works.

Several algorithms for simulating from a specific MSM have ben suggested in the literature.

- Havercroft and Didelez (2012)
- Bryan et al. (2004)
- Westreich et al. (2012) not in review by Havercroft and Didelez (2012)
- Young and Tchetgen Tchetgen (2014)
- Bryan 2004: fixes the vector L at the beginning which means A never affects L which don't make no sense. In other words it doesn't have the observational structure we are looking for.
- Do we introduce a form of selection bias into the data when we force positivity according to protocols? At baseline the proportion of people in any CD4 strata was unknown but randomized and hence in expectation it should be the same in treated and untreated.
- Young 2014 - Law of the observed outcome conditional on the measured past. What this paper shows is that the regression results are not correct but the IPTW ones are fine. The issues is comparing IPTW to normal regression results. As we compare IPTW results to IPTW results under positivity it should be fine to use the Havercroft algorithm.

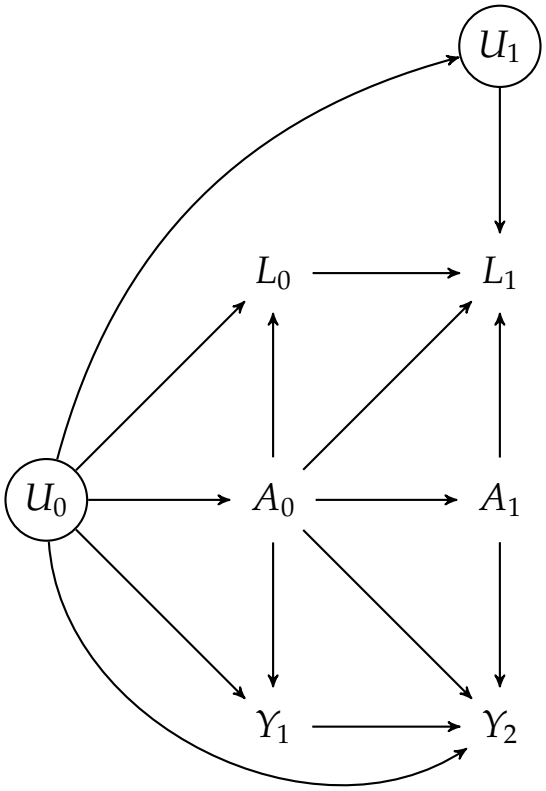
Havercroft and Didelez (2012) also make these comparisons in their paper in a few paragraphs. What we add here is how control over L allows us to introduce positivity violations.

- should explain all the reasons why we choose Havercroft algorithm over others.

3.1 Simulation algorithm

The base algorithm used in this paper is drawn from Havercroft and Didelez (2012). In that paper, an algorithm for simulating data from a specific MSM is derived from the joint factorization of a DAG which exhibits time dependent confounding. The algorithm generates subject data for multiple visits. The central concepts of the algorithm can be explained using a simplifies two period version of the algorithm, see figure ?.

- stress that we simulate from conditionals.



Causal graph

The DAG corresponding to their simulation algorithm is shown in figure ?. This DAG who derive their algorithm according to a specific DAG. Here a sketch of the key points in this algorithm is presented, the full proof demonstrating that the algorithm simulates data from a specific MSM

is presented in Appendix B of Havercroft and Didelez (2012).

Algorithm 1: Simulation Algoirthm MSM

Result: Marginal Structural Model Under Time Dependent Confounding

for i *in* $1, \dots, n$ **do**
  $U_{0,i} \sim U[0,1]$
  $\epsilon_{0,i} \sim N(\mu, \sigma^2)$
  $L_{0,i} \leftarrow F^{-1}_{\Gamma(k,\theta)}(U_{i,0}) + \epsilon_{0,i}$
  $A_{-1,i} \leftarrow 0$
  $A_{0,i} \leftarrow Bern(expit(\theta_0 + \theta_2(L_{0,i} - 500)))$
  if $A_{0,i} = 1$ **then**
    $T^* \leftarrow 0$;
  end
  $\lambda_{0,i} \leftarrow expit(\gamma_0 + \gamma_2 A_{0,i})$
  if $\lambda_{0,i} \geq U_{0,i}$ **then**
    $Y_{1,i} \leftarrow 0$
  else
    $Y_{1,i} \leftarrow 1$
  end
  for k *in* $1, \dots, T$ **do**
    if $Y_{t,i} = 0$ **then**
      $\Delta_{t,i} \sim N(\mu_2, \sigma_2^2)$
      $U_{t,i} \leftarrow min(1, max(0, U_{t-1,i} + \Delta_{t,i}))$
      if $t \neq 0 \ (mod\ k)$ **then**
        $L_{t,i} \leftarrow L_{t-1,i}$
        $A_{t,i} \leftarrow A_{t-1,i}$
      else
        $\epsilon_{t,i} \sim N(100(U_{t,i} - 2), \sigma^2)$
        $L_{t,i} \leftarrow max(0, L_{t-1,i} + 150 A_{t-k,i}(1 - A_{t-k-1,i}) + \epsilon_{t,i})$
        if $A_{t-1,i} = 0$ **then**
          $A_{t,i} \sim Bern(expit(\theta_0 + \theta_1 t + \theta_2(L_{t,i} - 500)))$
        else
          $A_{t,i} \leftarrow 1$
        end
        if $A_{t,i} = 1 A_{t-k,i} = 0$ **then**
          $T^* \leftarrow t$
        end
      end
      $\lambda_{t,i} \leftarrow expit()\gamma_0 + \gamma_1[(1 - A_{t,i})t + A_{t,i}T^*] + \gamma_2 A_{t,i} + \gamma_3 A_{t,i}(t - T^*))$
      if $1 - \prod_{\tau=0}^{t}(1 - \lambda_{\tau,i}) \geq U_{0,i}$ **then**
        $Y_{t+1,i} = 1$
      else
        $Y_{t+1,i} = 0$
      end
    end
  end
end

- describe how the algorithm works, is derived and how it acheives a specific MSM, the observational structure and a specific MSM (collapsibility)

3.2 Algorithm with positivity violations.

- show parts of algorithm which change
- present full algorithm with all changes in appendix
- interesting question is how the positivity violations are propogated through time. This can be checked by lengthening the number of visits sequentially. For example, do positivty violations with a time horizon of 5 visits have less of an effect than 10 visits.

The parameter of interest could be expected survival or the five year survival probability

Need to specify what the MSM is, give an example of it as a hazard function. Survival is completeley determined by the hazard function.

U allows us to get any distribution we like for Y marginal over covariates, WOuld L itself allows this? probably not. We can somehow get from this the subjects counterfactual survival time.

Importantly, this expression on the left hand side has unobserved counterfactuals, but the right hand side has only observed quantities which would be observed in an actual observational study.

non-collapsibility is an unresolved issue here. So even if we investigate positivity we can still only do so for collapsible models?

Equivalent way of motivating dividing the joint distribution by Pr(A | L) is through IPTW.

1. derive relationship between MSM and DAG and the correct conditional distributions. Follows from truncated factorisation why we can et P(Y | do(a))

2. think of this process as if we had fixed a treatment vector in advance. consistency assumption.

3. HD 2012, with Pearl and truncated factorization formula, show that it is possible to link the counterfactual represented by $P(Y \mid do(a))$ to observattional data generated in an observational way. But the problem arises when the model is non-collapsible or non-linear.

In the one shot case we set A = 1/0 because we are interested in the outcome under either of these treatment scenarios. In the time dependent case, A is a vector of 0s and 1s and we want to pretend that we decide in advance that the whole vector A is specified. But A and L have a complex interplay in an observational setting. So we want to pretend that A (a vector of 1s and 0s) is set in advance but at the same time have the observational structure for A and L.

The relationship between Y and L is then dependent on A. There is no relationship between A and U because of the set-up in the DAG. The variable L blocks this relationship.

Figure 1 represents the system under consideration. The DAG in figure 1 represents the one-shot non-longitudinal case. Factorising the joint distributions of the variables in figure 1 yields

$$P(U,\ L,\ W,\ A,\ Y) = P(W)P(U)P(W)P(L \mid U)P(A \mid L, W)P(Y \mid U, A)$$

Where, following definition 1.1 we delete $P(A \mid L, W)$, a probability function corresponding to A, and replace $A = a$ in all remaining functions

$$P(U, L, W, Y \mid do(A = a)) = \begin{cases} P(U)P(L \mid U)P(Y \mid U, A = a) & \text{if } A = a \\ 0 & \text{if } A \neq a \end{cases}$$

The goal is to simulate from a particular MSM. This means parameterising $P(Y \mid do(A = a))$. Applying the law of total probability over W, U and L yields

$$P(Y \mid do(A = a) = \sum_{w,u,l} P(W)P(U)P(L \mid U)P(Y \mid U, L, A = a) = \sum_{u,l} P(U)P(L \mid U)P(Y \mid U, L, A = a)$$

Making use of the fact that $P(L, U) = P(L \mid U)P(U) = P(U \mid L)P(L)$ and summing over either W and U or W and L yields

$$P(Y \mid do(A = a) = \sum_{l} P(Y \mid L, A = a)P(L) = \sum_{u} P(Y \mid U, A = a)P(U))$$

If we can find suitable forms for either $P(Y \mid L, A = a)$ and $P(L)$ or $P(Y \mid U, A = a)$ and $P(U)$ that correspond to the MSM $P(Y \mid do(A = a)$, then, given suitable values for A, L, U it will be possible to simulate from the chosen MSM.

Choosing a functional form for

$$P(Y \mid do(A = a)$$

depends on convenience. We need a functional form that can be easily represented by $P(Y \mid L, A = a)P(L)$. non-linear functions will be hard to work into the analysis.

U ~ U[0, 1] is a good choice because we can usethe CDF of Y because U[0, 1] is always between 0 and 1

General health is patient specific but comes from a clear distribution and has a nice medical interpretation. In contrast L would be more difficult to include. It is better as a function of U than a value in of itself.

- Explain issue that survival models are not collapsible which is why most algorithms don't work. Big reason we choose HD2012 is because of this
- no model mispecification in the HD2012 algorithm
- stay on treatment after treatment starts
- they motivate a logistic model for the haxard function, they use a discrete equivalent to the hazrd function (link to citation about farington study.)
- treatment regime is determined by t* (starting point of treatment because it is a vector of {0, 0, 0, 1, 1, 1}
- Explain why we can introduce positivity in this algorithm but not in others as easily.

4 Violations of Positivity

The motivation for using the algorith of Havercroft and Didelez (2012) is that we have control over how L affects Y whereas in other algorithms L would be fixed in advance., so we can itroduce positivity using a threshold. In other algorithms there would be a direct link between L and Y, this would be a problem because altering treatment decisions based on L would affect Y directly.
- creating an artificial population in which positivity is violated in specific ways.

4.1 Extended discussion of algorithm linking to positivity

As described in the introduction, one assumption of the model is that there is a non-zero probability of the event occuring at every startum of the covariate.

- When previous covariates like CD4 count are strongly associated with treatment the probabilities in the denominator of the ustabilized weights may vary greatly. Because we are foricing positivity by using a treatment rule when L falls below a threshold and A is then eaual to one, we create a strong association between A and L -> hence the unstabilized weights would vary. (Robins et al 2000 pp. 553)
- present the algorithm again with positivity violations.

4.2 Simulation scenarios

thresholding

percentage of compliant doctors.

propogation through time, longer time periods

- table of weights mean, min, max at T = 5, 10, 15 etc. plot with bias against time.

5 Simulation study

5.1 Data Structure

Include an example simulation graph plot showing the data colored to show where positivity would arise.

5.2 Number of positivity compliant doctors.

5.3 Varying levels of threshold.

We wish to simulate survival data in discrete time $t = 0, \ldots, T$ for n subjects. At baseline $t = 0$ all subjects are assumed to be at risk of failure so that $Y_0 = 0$. For each time period $t = 0, \ldots, T$ a subject may either be on treatment, $A_t = 1$, or not on treatment, $A_t = 0$. All patients are assumed to be not on treatment before the study begins. Once a patient commences treatment, they remain on treatment in all subsequent periods until failure or the end of follow-up. In each time period L_t is the value of a covariate measured at time t. In the simulated data, L_t behaves in a similar manner to CD4 counts such that a low value of L_t represents a more severe illness and hence a higher probability of both tratemnt and failure in the following period. In addition to L_t, the variable U_t represents subject specific general health at time t.

Each time period is either a check up visit or is between two check up visits. If t is a check-up visit and treatment has not yet commenced, L_t is measured and a decision is made on whether to commence treatment. Between visits, treatment remains unchanged at the value recorded at the previous visit. Similarly, L_t which is only measured when t is a visit, alos remains unchanged.

We represent the history of a random variable with an over bar. For example, the vector representing the treatment history of the variable A is represented by $\bar{A} = [a_0, a_1, \ldots, a_m]$ where $m = T$ if the subject survives until the end of follow-up, or $m < T$ otherwise. Prior to basline both $A = 0$ for all subjects.

- explain what U is and how it relates to the simulation design/algorithm
- Be more specific on Y

- L_t is a measured confounder
- U_t is an unmeasured confounder.

5.4 Simulation Algorithm

5.4.1 Algorithm

Next, we describe the algorithm used to simulate data from our chosen marginal structural model under time dependent confounding. In the following section we discuss in detail how the algorithm works and the salient features for this thesis. The algorithm is taken from Havercroft and Didelez (2012) who generate data on n patients, for k time periods. The outer loop in the following algorithm $i \in 1, \ldots, n$, refers to the patients while the inner loop $t \in 1, \ldots, T$ refers to the subsject specific time periods from baseline to failure or the end of the study. There will be at least one, and at most T records for each patient.

Within the inner loop ($t \in 1, \ldots, T$) we see that the data is only updated at time $t \neq 0 \ (mod\ k)$, where k refers to evenly spaced check-up visits. If t is not a check-up visit the values of A_t and L_t are the same as in $t-1$. When t is a visit A_t and L_t are updated.

- if treatment has been commenced then a subject may feel extra benefit if more time has elapsed since treatment began
- L_t affects A_t and also Y_t
- explain starting values for A and Y are all zero (except L maybe)

In order to operationalize the Algorithm 1 we need to choose parameters for $()$. In their paper Havercroft and Didelez (2012) use values that simulate data with a close resemblance to the Swiss HIV Cohort Study. We postpone disussion of the patameters in Algorithm 1 to section 2.4. We just need to state that we follow their parameters because this is not the focus of this thesis.

5.4.2 Discussion of how algorithm works

The algorithm of Havercroft and Didelez (2012) works by factorizing the joint density of the histories of the four variables in the analysis.

- Important is that the form of the MSM is not specified intil the last stage
- role of $U_{0,i}$
- How does positivity enter the analysis?
- Why this model is important in terms of positivity.

5.5 Constructing IPT weights

Inverse Probability of Treatment weights can be used to adjust for measured confounding and selection bias in marginal structural models. Link back to pseudo population idea in previous section. This method relies on four assumptions consistency, exchangeability, positivity and no mispecification of the model used to estimate the weights Cole and Hernán (2008). Unstabilized weights are defined as:

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

Where the denominator is the probability that the subject received the particular treatment history that they were observed to receive up to time t, given their prior observed treatment and

covariate histories (Havercroft, Didelez, 2012). The probabilities $p_\tau(A_{\tau,i} \mid \bar{A}_{\tau-1,i}, \bar{L}_{\tau,i})$ may vary greatly between subjects when the covariate history is strongly asscoaited with treatment. In terms of the resulting pseudopopulation, very small values of the unstabilized weights for some subjects would result in a small number of observations dominating the weighted analysis. The result is that the IPTW estimator of the coefficients will have a large variance, and will fail to be normally distributed. This variability can be mitigated by using the following stabilized weights

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

In the case that there is no confounding the denominator probabiliies in the stabilized weights reduce to $p_\tau(A_{\tau,i} \mid \bar{A}_{\tau-1,i})$ and $sw_{it} = 1$ so that each subject contributes the same weight. In the case of confounding this will not be the case and the stabilized weight will vary around 1.

In practice, we estimate the weights from the data using a pooled logistic model for the numerator and denominator probabilities. The histories of the treatment and covariates are included in the probabilities. In practice Specifically, following Havercroft and Didelez (2012), we estimate the model where the visit is only the visits every check up time. Between check ups both the treatment and covariate remain the same. Other ways of doing this include a spline function over the months to create a smooth function between the visits. Another difference might be to use a coxph function instead of logistic function

$$logit\ p_\tau(A_{\tau,i} \mid \bar{A}_{\tau-1,i}, \bar{L}_{\tau,i}) = \alpha_0 + \alpha_1 k + \alpha_2 a_{k-1} + \cdots + \alpha_k a_0 +$$

We have several options for estimating these weights. We could use a coxph model, or a logistic model.

5.6 Simulation Set-up

We follow the simulation set-up of Havercroft, Didelez (2012) which is based on parameters that closely match the Swiss HIV Cohort Study (HAART).

5.7 Results

- check the distribution of the weights that come out of the model (see Cole 2008). This would allow us to see weight model mispecifications. Not a problem in the simuation case.
- compare the bias, se, MSE, and 95% confidence interval
- compare all of these in the positivity violation and non-positivty violation case.
- explain to some extent monte-carlo standard error.
- we don't confirm the results of the havercroft of Bryan papers, instead refer readers to these papers to see how IPTW outperforms the naive estimators.
- Explain why we use MSE or other measures to assess simulation results.

6 Discussion and Conclusion

The focus of this thesis was the effect of positivity violations on

6.1 Limitations

References

Jenny Bryan, Zhuo Yu, and Mark J. Van Der Laan. Analysis of longitudinal marginal structural models. *Biostatistics*, 2004. ISSN 14654644. doi: 10.1093/biostatistics/kxg041.

Yvonne W Cheng, Alan Hubbard, Aaron B Caughey, and Ira B Tager. The association between persistent fetal occiput posterior position and perinatal outcomes: an example of propensity score and covariate distance matching. *American journal of epidemiology*, 171(6):656–663, 2010.

Stephen R. Cole and Constantine E. Frangakis. The consistency statement in causal inference: A definition or an assumption?, 2009. ISSN 10443983.

Stephen R Cole and Miguel A Hernán. Constructing inverse probability weights for marginal structural models. *American journal of epidemiology*, 168(6):656–664, 2008.

Stephen R. Cole, Robert W. Platt, Enrique F. Schisterman, Haitao Chu, Daniel Westreich, David Richardson, and Charles Poole. Illustrating bias due to conditioning on a collider. *International Journal of Epidemiology*, 2010. ISSN 03005771. doi: 10.1093/ije/dyp334.

R. M. Daniel, S. N. Cousens, B. L. De Stavola, M. G. Kenward, and J. A.C. Sterne. Methods for dealing with time-dependent confounding. *Statistics in Medicine*, 2013. ISSN 02776715. doi: 10.1002/sim.5686.

Vanessa Didelez, Svend Kreiner, and Niels Keiding. Graphical Models for Inference Under Outcome-Dependent Sampling. *Statistical Science*, 2010. ISSN 0883-4237. doi: 10.1214/10-STS340.

Peng Ding and Fan Li. Causal Inference: A Missing Data Perspective. pages 1–47, 2017. ISSN 0883-4237. doi: 10.1214/18-STS645. URL http://arxiv.org/abs/1712.06170.

Jessie K Edwards, Stephen R Cole, and Daniel Westreich. All your data are always missing: incorporating bias due to measurement error into the potential outcomes framework. *International journal of epidemiology*, 44(4):1452–1459, 2015.

Sander Greenland. Absence of confounding does not correspond to collapsibility of the rate ratio or rate difference. *Epidemiology*, 1996. ISSN 10443983. doi: 10.1097/00001648-199609000-00008.

Sander Greenland and G Maldonado. The importance of critically interpreting simulation studies. *Epidemiology (Cambridge, Mass.)*, 1997. ISSN 1044-3983.

Sander Greenland and Judea Pearl. Adjustments and their Consequences-Collapsibility Analysis using Graphical Models. *International Statistical Review*, 2011. ISSN 03067734. doi: 10.1111/j.1751-5823.2011.00158.x.

Sander Greenland, James M Robins, and Judea Pearl. Confounding and collapsibility in causal inference. *Statistical science*, pages 29–46, 1999.

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

M A Hernan, B Brumback, and J M Robins. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*, 2000. ISSN 1044-3983. doi: 10.1097/00001648-200009000-00012.

Miguel A. Hernán. The hazards of hazard ratios, 2010. ISSN 10443983.

Miguel A. Hernán and Stephen R. Cole. Invited commentary: Causal diagrams and measurement bias, 2009. ISSN 00029262.

Miguel A. Hernán and James M. Robins. Estimating causal effects from epidemiological data, 2006. ISSN 0143005X.

Miguel A Hernán, Sonia Hernández-D\'\iaz, and James M Robins. A structural approach to selection bias. *Epidemiology*, 15(5):615–625, 2004.

D. G. Horvitz and D. J. Thompson. A Generalization of Sampling Without Replacement from a Finite Universe. *Journal of the American Statistical Association*, 1952. ISSN 1537274X. doi: 10.1080/01621459.1952.10483446.

Chanelle J. Howe, Lauren E. Cain, and Joseph W. Hogan. Are All Biases Missing Data Problems? *Current Epidemiology Reports*, 2015. ISSN 2196-2995. doi: 10.1007/s40471-015-0050-8.

Guido W Imbens and Donald B Rubin. Causal Inference. *Boca Raton: Chapman & Hall/CRC, forthcoming.*, (June):39–58, 2008.

Melvin L. Moeschberger John P. Klein. *Survival Analysis - Techniques for Ceosnred and Truncated Data - 2nd Edition*. 2003. ISBN 038795399X. doi: 10.1145/390011.808243.

Lynne C Messer, J Michael Oakes, and Susan Mason. Effects of socioeconomic and racial residential segregation on preterm birth: a cautionary tale of structural confounding. *American journal of epidemiology*, 171(6):664–673, 2010.

Ashley I. Naimi, Stephen R. Cole, Daniel J. Westreich, and David B. Richardson. A comparison of methods to estimate the hazard ratio under conditions of time-varying confounding and nonpositivity. *Epidemiology*, 2011. ISSN 10443983. doi: 10.1097/EDE.0b013e31822549e8.

J Neyman. On the application of probability theory to agricultural experiments: principles (in Polish with German summary). *Roczniki Nauk Rolniczch*, 10(1):21–51, 1923. ISSN 0883-4237. doi: 10.1214/ss/1177012031. URL http://projecteuclid.org/euclid.ss/1177010123.

Menglan Pang, Jay S. Kaufman, and Robert W. Platt. Studying noncollapsibility of the odds ratio with marginal structural and logistic regression models. *Statistical Methods in Medical Research*, 2016. ISSN 14770334. doi: 10.1177/0962280213505804.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Judea Pearl. On the consistency rule in causal inference: Axiom, definition, assumption, or theorem? *Epidemiology*, 2010. ISSN 10443983. doi: 10.1097/EDE.0b013e3181f5d3fd.

Judea Pearl. Comment: Understanding Simpson's paradox, 2014. ISSN 15372731.

James Robins. 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. ISSN 02700255. doi: 10.1016/0270-0255(86)90088-6.

James M Robins. Marginal structural models versus structural nested models as tools for causal inference. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 95–133. Springer, 2000.

James M. Robins, Donald Blevins, Grant Ritter, and Michael Wulfsoh. G-estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia on the survival of AIDS patients. *Epidemiology*, 1992. ISSN 15315487. doi: 10.1097/00001648-199207000-00007.

James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology, 2000.

Donald B. Rubin. Bayesian Inference for Causal Effects: The Role of Randomization. *The Annals of Statistics*, 1978. ISSN 0090-5364. doi: 10.1214/aos/1176344064.

Arvid Sjölander, Elisabeth Dahlqwist, and Johan Zetterqvist. A note on the noncollapsibility of rate differences and rate ratios. *Epidemiology*, 27(3):356–359, 2016.

Tyler J. Vander Weele. Concerning the consistency assumption in causal inference. *Epidemiology*, 2009. ISSN 10443983. doi: 10.1097/EDE.0b013e3181bd5638.

Daniel Westreich and Stephen R Cole. Invited commentary: positivity in practice. *American journal of epidemiology*, 171(6):674–677, 2010.

Daniel Westreich, Stephen R. Cole, Enrique F. Schisterman, and Robert W. Platt. A simulation study of finite-sample properties of marginal structural Cox proportional hazards models. *Statistics in Medicine*, 2012. ISSN 02776715. doi: 10.1002/sim.5317.

Jessica G. Young and Eric J. Tchetgen Tchetgen. Simulation from a known Cox MSM using standard parametric models for the g-formula. *Statistics in Medicine*, 2014. ISSN 02776715. doi: 10.1002/sim.5994.