Simulation and Causal Inference Methods for Repeated Measures or Longitudinal Data

Alexander Levis

Department of Epidemiology, Biostatistics & Occupational Health

McGill University
Montreal, Canada

July 2017

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science (M.Sc.) in Biostatistics.

2017/07/25

Abstract

Several methods for estimating treatment effects in the presence of time-dependent confounding have been proposed for which asymptotic consistency results have been proven. To investigate finite-sample properties of these complex estimators, it is often necessary to use simulated data to benchmark performance. However, simulations in the literature have typically been simplistic, with no exhaustive comparison of available methods. We propose and implement a complex simulation study that compares a variety of popular causal methods in the longitudinal, repeated-measures setting. The causal methods investigated are then applied to data from a large cluster-randomized trial, the PROmotion of Breastfeeding Intervention Trial (PROBIT). Implications of the simulation study are discussed in the context of the current literature.

Résumé

De nombreuses méthodes pour estimer les effets de traitements, en présence de facteurs confusionnels qui varient avec le temps, ont jusqu'à présent été proposées pour lesquelles des résultats asymptotiques ont été prouvés. Pour étudier les caractéristiques et comparer les performances de ces estimateurs complexes appliqués à des échantillons finis, il est souvent nécessaire d'utiliser des données simulées. Cependant, les études de simulation actuellement publiées sont typiquement simplistes, et ne comparent pas de manière exhaustive toutes les méthodes disponibles. Ainsi, nous proposons et effectuons une étude de simulation complexe qui compare une collection de méthodes causales populaires dans le cadre de données longitudinales et de mesures répétées. Ensuite, ces méthodes sont appliquées aux données provenant d'un grand essai randomisé en grappes: le PROmotion of Breastfeeding Intervention Trial (PROBIT). Finalement, les implications de l'étude de simulation sont discutées en fonction de leur lien avec la littérature actuelle.

Acknowledgments

I first want to extend the warmest appreciation to Dr. Robert Platt for supervising this master's thesis. His guidance, time, and support were invaluable, and his encouragement helped me immensely in making progress as I was getting to grips with this field that I initially knew next to nothing about. I owe Dr. Brett Thombs a world of thanks for first introducing me to conducting rigorous scientific research, and for helping me find the field of biostatistics that was and is truly a perfect match for me. I am forever grateful to Brett for kickstarting my career, and for being a constant source of encouragement and support. Thanks also to all the biostatistics professors in the department here at McGill for getting us excited about this amazing field, sharing their time and ideas, and bringing in fantastic speakers to widen our horizons. A special thanks to Drs. Erica Moodie and Jim Hanley. On top of being a committee member for this thesis, Erica's superb teaching inspired me to conduct my master's research in the area of causal inference. Her persistent belief in us students, and the time she provides beyond class hours, are truly instrumental to our success. Jim's one-of-a-kind teaching style got me hooked on statistics; I will forever hear his voice pressing us to keep things simple, and I hope to always keep his elegant foundational examples in my toolbox. To Guanbo, Hao, Shomoita, and Stan, it was incredible to share much of this two-year journey with all of you—thanks for being there for me, and I very much hope we remain lifelong friends. I especially thank Guanbo for helping me to understand the particularly thorny area of targeted learning.

All my friends and family deserve my deepest gratitude. Thanks to my mom and dad, Reesa and Victor, for believing in me, supporting me, and making sure I take care of myself. I thank my sister Brooke for getting me into research in the first place, and my brother

Mitch for cheering me on and always being there for me. My siblings have always been the best role models for what to do (and what not to do...). To all my friends, thanks for being so awesome and always putting up with me. I am especially grateful to my roommates Jamie and Nick, who made this past year a blast and kept me sane. Finally, I am so fortunate and thankful to have Kira Riehm in my life. I can always count on her for encouragement, advice, or just to lend an ear. She never ceases to amaze me with her hard work and incredible accomplishments, and I can't wait to see what the future has in store for her.

I would additionally like to express my appreciation for the Fonds de recherche du Québec and the Canadian Institutes of Health Research for funding my master's degree, and thank Dr. Michael Kramer for letting us have unrestricted access to the PROBIT dataset.

Contribution of Authors

I am the sole author of this thesis, and am responsible for formulating the research objectives, developing the simulation study, analyzing the simulated and real data, compiling results, and writing and revising the thesis. My supervisor, Dr. Robert Platt, assisted and provided guidance for all parts of the research, and reviewed drafts of this thesis and provided comments. My other committee member, Dr. Erica Moodie, was involved in initial discussions of the simulation study design, and provided advice throughout the process.

Contents

List of Figures

List of Tables

List of Acronyms

AIDS	Acquired Immune Deficiency Syndrome
ATE	Average Treatment Effect
DAG	Directed Acyclic Graph
GEE	Generalized Estimating Equations
HIV	Human Immunodeficiency Virus
IPTW	Inverse Probability of Treatment Weighting
MSM	Marginal Structural Model
TMLE	Targeted Maximum Likelihood Estimation/Estimator
SWIG	Single World Intervention Graph

Chapter 1

Introduction

The problem of estimating causal treatment effects from observational data is well studied. Beginning with Rubin (1974, 1978), the language of potential outcomes (or counterfactuals) was introduced in the context of observational data, after first being laid out for randomized studies (Neyman, 1923). This formalization gave more precise meaning to the term 'confounding', and permitted sufficient conditions for valid treatment effect estimation to be written succinctly. It also immediately suggested how to construct consistent estimators.

Subsequently, Robins (1986) realized that in the longitudinal observational setting, where treatment and covariates are measured at several time points before an outcome is measured, the traditional methods investigators were using to estimate treatment effects while controlling for the covariates were problematic in most cases. Specifically, he showed that if a time-varying covariate is a confounder (it has an effect on subsequent treatment assignment as well as the outcome) and also a mediator (it is affected by prior treatment), then regardless of whether or not one controls for this time-varying covariate in a standard regression analysis, the estimated effect of the treatment on the outcome would be biased.

Such a covariate is known as a time-dependent confounder. Consequently, Robins (1986) extended the point treatment potential outcome language of Rubin (1974) to the longitudinal setting, and showed how to properly estimate treatment effects while accounting for any time-dependent confounders with the g-computation algorithm formula (or g-formula).

Meanwhile, the analysis of repeated measures data—a longitudinal data setting in which the outcome is also measured at each time point—was being revolutionized due to work of Liang & Zeger (1986); Zeger & Liang (1986). Their generalized estimating equation methodology was applicable to settings where one wished to estimate associations of covariates with a vector of correlated outcome variables. For instance, in repeated measures data, the sequence of outcomes for one subject is such a correlated vector, and one might use their methods for estimating time-varying treatment effects on the outcome. This works particularly well for randomized experiments in which there are no confounders, but when data are observational, it turns out that the same methodology cannot deal directly with time-dependent confounders. Specifically, the same arguments of Robins (1986) apply, and estimates obtained in the presence of time-dependent confounders would not reflect causal effects, only associations that might lead to spurious conclusions. Years later, Robins (1997, 1999) developed marginal structural models (MSMs), a general set of methods for causal inference for point treatment or longitudinal data. Hernán *et al.* (2002) described one such MSM that is a simple extension of traditional generalized estimating equations, that could be used for the analysis of repeated measures data to obtain valid treatment effect estimates when there are time-dependent confounders.

Since, there have been several advances in statistical and causal inference theory, which have led to the development of "double robust" estimators (e.g., see Kang & Schafer,

2007). These are estimators that are somewhat more robust to model misspecification than a g-formula or MSM approach. A particularly promising double robust estimator is the targeted maximum likelihood estimator (van der Laan & Rubin, 2006). Schnitzer *et al.* (2013) and Petersen *et al.* (2014) developed a longitudinal version of the targeted estimator, that could be used to estimate treatment effects controlling properly for time-dependent confounders.

When assumptions are satisfied and models correctly specified, the g-formula, MSMs, and longitudinal targeted maximum likelihood estimator have all been proven to be asymptotically consistent for estimating treatment effects in longitudinal or repeated measures data. Nonetheless, to show that they perform well in realistic sample sizes necessitates the use of simulation studies. Thus far, most simulations have been relatively simplistic, considering either only one or two covariates at baseline, or only two or three time points. Thus, the objective of this thesis is to address this gap, and conduct a complex simulation study that includes a larger number of covariates, and simulates several time points. This will hopefully give a better idea of how these estimators fare in more realistic data settings.

The remainder of the thesis is organized in the following manner. Chapter 2 provides an overview of potential-outcome-based causal inference theory. The notation for point treatment and longitudinal data settings is introduced, and the g-formula, MSMs, and targeted estimation are elaborated in terms of the causal theory. Chapter 3 presents a repeated measures simulation study in a setting with many baseline covariates and many follow-up visits, and we assess and compare the performance of the different causal estimators under a suite of simulation scenarios and modeling assumptions for the estimators. In Chapter 4, we analyze a longitudinal dataset, applying the causal estimators evaluated

in the simulation study in the context of estimating the effect of breastfeeding on infant weight. Lastly, in Chapter 5, we discuss implications and conclude.

Chapter 2

Literature and Theory Review

There is a vast literature surrounding causal inference and methods for dealing with longitudinal data. In this chapter, we review the major developments of causal theory relevant to this thesis. We show the foundations of each method of interest by first dealing with the time-fixed case, and describe extensions for measuring treatment effects in longitudinal settings when there are time-varying confounders that are also mediators (which we henceforth refer to as 'time-dependent confounders'). To provide a concrete example of such a covariate, consider a longitudinal study that evaluates, among HIV-positive subjects, the effect of a treatment on progression to AIDS or death. In this setting, CD4 count would be considered a time-dependent confounder. The variable is a mediator because CD4 count measured at a given visit would plausibly have been affected by the patient's treatment status in the recent past, and it could also directly affect the outcome. Similarly, CD4 count is a confounder because it influences which treatment is used in subsequent periods and, again, possibly the outcome variable as well.

2.1 The Neyman-Rubin model for point treatment effects

Although more thoroughly developed and popularized by Rubin (Rosenbaum & Rubin, 1983; Rubin, 1974, 1978), the language of potential outcomes was first introduced by Neyman (1923) in the context of randomized experiments in agriculture. Thus, the paradigm to be described is often referred to as the Neyman-Rubin causal model (e.g., in Sekhon, 2008).

Assume that we have $(W_i, A_i, Y_i), i = 1, \ldots, n$, independent and identically distributed copies of (W, A, Y) on n subjects, where W denotes a vector of covariates, A a binary treatment taking values 0 or 1 (for simplicity; this is easily generalized to multiple treatment levels, or continuous treatment), and Y the outcome of interest (may be discrete or continuous). We presume the existence of two variables called *potential outcomes* or *counterfactuals*: $Y(a)$, the value of the outcome had, possibly counter to fact, the subject been exposed to treatment level a (i.e., $A = a$), for $a = 0, 1$. It is assumed that we are able to observe the potential outcome for the treatment that was actually observed, that is:

$$Y = (1 - A)Y(0) + AY(1), \tag{2.1}$$

an assumption known as *consistency*. It is worth mentioning that the assumptions of existence of potential outcome variables and consistency are tied toegether implicitly with an assumption that there is only one version of each treatment—this has been referred to as the "stable unit-treatment value assumption" (SUTVA) (e.g., see Rosenbaum & Rubin, 1983). One can imagine this criterion being violated if, for instance, treatment was a surgery for which different surgeons were more or less capable, as then a given patient's potential outcome would be different depending on the surgeon (VanderWeele & Hernan,

2013). This violation thus leads to a breakdown of the definition of potential outcomes and consistency, and makes causal inference very challenging, which explains why SUTVA is so often implicitly assumed. See VanderWeele & Hernan (2013) for methods that can be used when there are in fact multiple versions of treatment and SUTVA does not hold. If we are willing to grant the existence of potential outcomes and consistency, notice that we obtain

$$AY(1) = AY;\ (1-A)Y(0) = (1-A)Y, \tag{2.2}$$

directly from (2.1).

It would be of great interest to know the causal effect of treatment on the outcome, $Y(1) - Y(0)$, however in reality we may only observe the outcome under one treatment for any given individual under study. This basic limitation is referred to as the "fundamental problem of causal inference" (Holland, 1986). Thus, in practice, the focus becomes identifiability and estimation of average causal contrasts, for example the average treatment effect (ATE):

$$\mathbb{E}[Y(1) - Y(0)].$$

We interpret this quantity as the average difference in the outcome if everyone in the population were given treatment $a = 1$ versus if everyone were given $a = 0$. In the case of a completely randomized controlled trial, it is very reasonable to assume that

$$Y(0), Y(1) \perp\!\!\!\perp A,$$

that is, the treatment effect is unconfounded. We interpret this independence as asserting that the treatment assignment is totally exogenous to how an individual would respond

to the different levels of treatment. A straightforward consequence of this independence together with consistency is that

$$\frac{1}{n}\left[\sum_{i=1}^{n}\left(\frac{A_iY_i}{\mathrm{P}(A_i=1)}-\frac{(1-A_i)Y_i}{\mathrm{P}(A_i=0)}\right)\right]$$

is an unbiased estimate of the ATE, and therefore the difference in sample means in the two groups, $\frac{\sum_{i=1}^{n}A_iY_i}{\sum_{i=1}^{n}A_i}-\frac{\sum_{i=1}^{n}(1-A_i)Y_i}{\sum_{i=1}^{n}(1-A_i)}$, consistently estimates the ATE. If treatment is not unconfounded, but we assume that the treatment mechanism is determined only by measured covariates in W (i.e., there are *no unmeasured confounders*),

$$Y(0), Y(1) \perp\!\!\!\perp A|W, \tag{2.3}$$

and that all subjects have a positive probability of being under any treatment level (called a *positivity* assymption),

$$\mathrm{P}(A=a|W=w)>0, \text{for } a=0,1, \text{and } w \text{ in the support of } W, \tag{2.4}$$

then again the ATE may be consistently estimated from the data. In particular,

$$\frac{1}{n}\left[\sum_{i=1}^{n}\left(\frac{A_iY_i}{\mathrm{P}(A_i=1|W_i)}-\frac{(1-A_i)Y_i}{\mathrm{P}(A_i=0|W_i)}\right)\right] \tag{2.5}$$

is an unbiased estimator of the ATE. To see this, first note that:

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{A_iY_i}{\mathrm{P}(A_i=1|W_i)}\right)\right] &= \mathbb{E}\left[\frac{AY}{\mathrm{P}(A=1|W)}\right], \text{ by i.i.d. assumption,}\\
&= \mathbb{E}\left[\frac{AY(1)}{\mathrm{P}(A=1|W)}\right], \text{ by consistency (2.2),}\\
&= \mathbb{E}_W\left[\mathbb{E}_{Y(1),A|W}\left(\frac{AY(1)}{\mathrm{P}(A=1|W)}\middle|W\right)\right], \text{ iterating expectations,}\\
&= \mathbb{E}_W\left[\frac{\mathbb{E}_{Y(1),A|W}(AY(1)|W)}{\mathrm{P}(A=1|W)}\right], \text{ since } \mathrm{P}(A=1|W) \text{ is a function of } W,\\
&= \mathbb{E}_W\left[\frac{\mathbb{E}_{A|W}(A|W)\mathbb{E}_{Y(1)|W}(Y(1)|W)}{\mathrm{P}(A=1|W)}\right],\\
&\qquad \text{by no unmeasured confounders (2.3),}\\
&= \mathbb{E}_W\left[\frac{\mathrm{P}(A=1|W)\mathbb{E}_{Y(1)|W}(Y(1)|W)}{\mathrm{P}(A=1|W)}\right], \text{ since } A \text{ is binary,}\\
&= \mathbb{E}_W\left[\mathbb{E}_{Y(1)|W}(Y(1)|W)\right], \text{ simplifying,}\\
&= \mathbb{E}[Y(1)], \text{ by iterated expectation identity.}
\end{aligned}$$

By an analogous argument, it is straightforward to show that

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{(1-A_i)Y_i}{\mathrm{P}(A_i=0|W_i)}\right)\right] = \mathbb{E}[Y(0)],$$

from which unbiasedness of (2.5) follows. The assumption of no unmeasured confounders holds necessarily in single time point randomized experiments where patients are randomly assigned treatment with probabilities determined by their measured covariates. In this case, $\mathrm{P}(A_i = a_i|W_i = w_i)$ is known by design. In an observational study, however, the conditional independence assumption of no unmeasured confounders is untestable, and probabilities $\mathrm{P}(A_i = a_i|W_i = w_i)$ must be estimated. If we are willing to grant all of the above assumptions (i.e., existence of potential outcomes, consistency, no unmeasured

confounders, and positivity), (2.5) allows for consistent estimation of the ATE even in observational studies, so long as the treatment mechanism can be consistently estimated (this asymptotic consistency follows from standard convergence theorems, and from our above derivation that (2.5) is unbiased when the treatment mechanism $\mathrm{P}(A_i = 1|W_i)$ is known). Notice in (2.5) that all subjects are weighted by the inverse of the probability of receiving the treatment they did receive given their covariates, thus leading to the terms *inverse probability of treatment* (IPT) and *inverse probability of treatment weighting* (IPTW) that are used to describe this approach and its generalizations.

Rosenbaum & Rubin (1983) refer to $e(W_i) \equiv \mathrm{P}(A_i = 1|W_i)$ as the *propensity score*, an individual's probability of being under treatment level $a = 1$ given their covariates. They show that the propensity score has many nice properties, and that it can be used in its own right (as opposed to its use for inverse weighting as in (2.5)) to control for measured confounding in observational studies. Propensity score methods condition on $e(W_i)$, through matching, stratification, or including it as a covariate in a regression analysis, in order to identify and estimate $\mathbb{E}[Y(1) - Y(0)]$, the ATE. Our focus will not be on these propensity score methods, however, and will remain similar in spirit to the IPTW approach shown above.

Before proceeding with more recent developments, it should be noted that the use of potential outcomes for causal inference is contentious to some degree. For example, Dawid (2000) raises several philosophical points of criticism. In particular, this paper notes that potential outcomes are unobservable, and any analysis will depend on the validity of untestable assumptions concerning these potential outcomes. The other very popular approach to causal inference—particularly in the computer science literature—involves the use of non-

parametric structural equation models represented by directed acyclic graphs (DAGs), due in large part to Pearl (see Pearl (2009) for a thorough treatment of the topic). However, the potential outcome and DAG formalisms have recently been unified neatly by Richardson & Robins (2013), through the so-called Single-World Intervention Graph (SWIG) construction. The main practical limitation to any methods that estimate average treatment effects, as in (2.5), centres around the untestable assumption that the variables included in W are sufficient to control for confounding, as this is considered rarely if ever plausible to be true in real-world examples, but remains necessary for the validity of inference.

2.2 Longitudinal data setting and the g-formula

Assume we have data of the form $(W_0, A_0, W_1, A_1, \ldots, W_K, A_K, Y)$, ordered by time, where W_k is a vector of covariates measured at visit k, A_k is the treatment assigned at visit k, and Y is an outcome measured at the end of follow up (i.e., at visit $(K+1)$). Note that one special case of this is repeated measures data, where at each time point, one measures the outcome variable, i.e., Y_k is a component of W_k. In that case $Y \equiv Y_{K+1}$. Returning to the general setting, let $\overline{W}_k \equiv (W_0, \ldots, W_k)$ and $\overline{A}_k \equiv (A_0, \ldots, A_k)$ denote the covariate and treatment histories, respectively, up to time k. Let the complete covariate and treatment histories up to time K be denoted by $\overline{W} \equiv \overline{W}_K$ and $\overline{A} \equiv \overline{A}_K$. Let $\overline{w}_k \equiv (w_0, \ldots, w_k)$, $\bar{a}_k \equiv (a_0, \ldots, a_k)$, $\overline{w}$ and $\bar{a}$ denote specific instantiations of these histories. We also assert the existence of potential outcome variables $Y(\bar{a})$, which represent the value that the outcome would take had the subject been exposed to treatment sequence $\bar{a}$. Assume now that we have sequential randomization based only on measured covariates (or 'no unmeasured

confounders', a generalization of (2.3)), that is,

$$Y(\bar{a}) \perp\!\!\!\perp A_k | \overline{W}_k, \overline{A}_{k-1} = \bar{a}_{k-1}, \text{ for } k = 0, \ldots, K, \tag{2.6}$$

as well as a positivity assumption (a generalization of (2.4)),

$$f(\bar{a}_{k-1}, \overline{w}_k) > 0 \implies f(a_k | \bar{a}_{k-1}, \overline{w}_k) > 0, \tag{2.7}$$

where $f(\cdot)$ is the conditional or joint density implied by its arguments. One instance where the sequential randomization assumption would hold is in a sequentially randomized experiment, where the probability of being assigned treatment at each time point is an investigator-defined function of prior treatment and covariate history. Under these assumptions, Robins (1986) showed that the distribution of the potential outcome $Y(\bar{a})$ is identified via the so-called g-computation algorithm formula, or *g-formula* (Robins, 1986). Specifically, we have that for any treatment sequence $\bar{a}$,

$$\begin{aligned}
\mathbb{E}[Y(\bar{a})] &= \int_{w_0} \mathbb{E}[Y(\bar{a})|w_0] f(w_0)\, dw_0, \text{iterating expectations} \\
&= \int_{w_0} \mathbb{E}[Y(\bar{a})|w_0, a_0] f(w_0)\, dw_0, \text{by sequential randomization} \\
&= \int_{w_0} \int_{w_1} \mathbb{E}[Y(\bar{a})|\overline{w}_1, \bar{a}_1] f(w_1|w_0, a_0) f(w_0)\, dw_1 dw_0, \text{repeating the above two steps} \\
&\vdots \\
&= \int_{w_0} \int_{w_1} \ldots \int_{w_K} \mathbb{E}[Y(\bar{a})|\overline{w}_K, \bar{a}_K] f(w_K|\overline{w}_{K-1}, \bar{a}_{K-1}) \cdots f(w_1|w_0, a_0) f(w_0)\, dw_K \ldots dw_1 dw_0, \\
&= \int_{w_0} \int_{w_1} \ldots \int_{w_K} \mathbb{E}[Y|\overline{w}, \bar{a}] \prod_{k=0}^{K} f(w_k|\overline{w}_{k-1}, \bar{a}_{k-1})\, dw_K \ldots dw_1 dw_0, \text{with } W_{-1} = A_{-1} = \varnothing
\end{aligned}$$

where $f(\cdot)$ denotes the density implied by its arguments. The last equality in the above derivation follows from the assumption of consistency (which in this case may be expressed as $Y = \sum_{\bar{a}' \in \mathcal{A}} I(\overline{A} = \bar{a}')Y(\bar{a}')$, $\mathcal{A}$ the support of $\overline{A}$, thus generalizing (2.1)). The g-formula expresses a causal quantity—above, the mean of the potential outcome $Y(\bar{a})$—in terms of the distribution of observable variables W_k, A_k, and Y. It thus allows estimation of the causal parameter from data, assuming that sequential randomization and positivity hold. In the time-fixed setting, the g-formula simplifies to

$$\mathbb{E}[Y(a)] = \int_w \mathbb{E}[Y|w, a] f(w)\, dw,$$

if W is a continuous random vector, or

$$\mathbb{E}[Y(a)] = \sum_w \mathbb{E}[Y|w, a] \mathrm{P}(W = w),$$

if W is discrete, where the integral or sum is over the support of W. In both cases, this is the same as this identity:

$$\mathbb{E}[Y(a)] = \mathbb{E}_W \left[\mathbb{E}_{Y|A=a,W}[Y|A = a, W] \right]. \tag{2.8}$$

In full generality, the longitudinal g-formula is equivalently expressed as:

$$\mathbb{E}[Y(\bar{a})] = \mathbb{E}_{W_0} \left[\mathbb{E}_{W_1|W_0,A_0=a_0} \left[\cdots \left[\mathbb{E}_{W_K|\overline{W}_{K-1},\overline{A}_{K-1}=\bar{a}_{K-1}} \left[\mathbb{E}_{Y|\overline{W},\overline{A}=\bar{a}}[Y|\overline{W}, \overline{A} = \bar{a}] \right] \right] \cdots \right] \right]. \tag{2.9}$$

Viewed as this chained conditional expectation, we can estimate $\mathbb{E}[Y(\bar{a})]$ for any treatment sequence $\bar{a}$ of interest by first estimating the the distributions of W_k given $\overline{W}_{k-1}$ and $\overline{A}_{k-1}$, for $k = 0, \ldots, K$, as well as the outcome distribution Y given $\overline{W}$ and $\overline{A}$. This may be done

nonparametrically when all variables are discrete (particularly appropriate if all variables are binary). However, when some variables in W_k vary continuously, it becomes necessary to specify a parametric model for each of these conditional distributions. One can then estimate $\mathbb{E}[Y(\bar{a})]$ for a given $\bar{a}$ via Monte Carlo simulation, iteratively generating W_0 (via bootstrap for example), W_1 given the generated W_0 and setting $A_0 = a_0$, ..., W_K given the generated $\overline{W}_{K-1}$ and setting $\overline{A}_{K-1} = \bar{a}_{K-1}$, via K estimated parametric models, and finally simulating Y given the simulated $\overline{W}$ and setting $\overline{A} = \bar{a}$ using an estimated parametric model. Taking the mean across a large number of these simulated outcomes yields an estimate of $\mathbb{E}[Y(\bar{a})]$. This procedure can be repeated for all treatment sequences $\bar{a}'$ of interest (e.g., always treated: $\bar{a}' = (1, 1, \ldots, 1)$; never treated: $\bar{a}' = (0, 0, \ldots, 0)$), and the difference or ratio of estimated potential outcome means would yield treatment effect estimates. The method just outlined is known as the parametric g-formula. Despite its relative simplicity, the parametric g-formula has only rarely been used for dealing with confounding in either single time point or longitudinal settings, as the more popular marginal structual models are often used instead. Robins *et al.* (2004) describe the parametric g-formula and the necessary underlying assumptions for valid estimation of the effect of time-dependent interventions, and apply it in a coronary heart disease example. Westreich *et al.* (2012a) applied this method for assessing the effect of antiretroviral therapy on incident AIDS; van der Wal *et al.* (2009) for evaluating the effect of the presence of tuberculosis on AIDS-related mortality; Young *et al.* (2011) for comparing the effectiveness of several dynamic regimes of antiretroviral therapy for reducing AIDS-related mortality; Taubman *et al.* (2009) for assessing the effect of a range of lifestyle interventions on the risk of coronary heart disease. Finally, Snowden *et al.* (2011) give a step-by-step walkthrough of the method in a very simple, single time point simulated example (one outcome, one exposure and two covariates). To our knowledge, there has not been a systematic evaluation

of the parametric g-formula in a complex, longitudinal simulation study, where the effect of parametric model misspecification could be studied, for instance. In addition, it is of interest to compare the method to other popular estimators.

One theoretical drawback of the parametric g-formula—which might be partly responsible for the method's lack of representation in the literature—is the so-called "g-null paradox" (Robins & Wasserman, 1997). This is a technical issue (described in much greater detail in Robins & Wasserman, 1997) that arises when using standard parametrizations of the outcome and time-varying covariate mean models (linear, or generalized linear models; see McCullagh & Nelder, 1989). Specifically, if there are unmeasured variables (not confounders) that affect a time-varying covariate as well as the outcome, then it becomes increasingly probable as sample size increases that the null hypothesis of no treatment effect is rejected, even when the null hypothesis holds. It is, however, unclear in practice how often these sorts of problematic missing variables are present in the causal structure, as well as the magnitude of the induced bias. Regardless, it is important to assess the performance of this approach under a variety of simulation scenarios and to compare with other modern techniques.

2.3 Repeated measures & Generalized Estimating Equations

In order to introduce the marginal structural model for repeated measures data in the next section, let us first review the traditional method for correlated data upon which it is based. Assume again, borrowing the notation from section 2.2, that we have data of the form $(W_0, A_0, W_1, A_1, \ldots, W_K, A_K, Y)$, where the variables are ordered by time, but in particular where the outcome variable Y_t is a component of W_t, for $t = 0, \ldots, K$, and

$Y \equiv Y_{K+1}$ is the outcome measured at the final visit. A common example of this data structure is in studies of HIV/AIDS, where the CD4 lymphocyte count is measured on each patient at each study visit. One may wish to estimate and compare the effect of different time-varying antiretroviral treatment strategies on the evolution of CD4. That is, one may be interested in estimating the sequence of mean potential outcomes under a treatment regime $\bar{a}$, $\mathbb{E}[Y_{t+1}(\bar{a})]$, for $t = 0, \ldots, K$, to know how this sequence changes as a function of different possible treatment sequences, or assess the effect of one treatment sequence $\bar{a}$ versus another $\bar{a}'$, $\mathbb{E}[Y_{t+1}(\bar{a}) - Y_{t+1}(\bar{a}')]$, either over time or at a particular time point. Three apparent challenges that come with these data are to borrow information across observed treatment sequences to make inference for the rarer instances, account for the correlation in succesive measures of the outcome variable, as well as appropriately adjust for possible confounders in W_t. In sequentially randomized controlled trials, the confounding issue is alleviated, as treatment assignment is independent of potential outcomes either marginally (as in a completely randomized trial) or conditional on measured covariates, by design. A natural way to analyze data in a completely randomized trial, so as to borrow information across treatment sequences, would be to use a parametric model

$$\mathbb{E}[Y_{t+1}|\overline{A}_t] = g(\overline{A}_t; \boldsymbol{\gamma}), \tag{2.10}$$

where g is some function of treatment history and unknown parameters. Treating $\mathbf{Y} := (Y_0, Y_1, \ldots, Y_{K+1})$ as a random vector, *generalized estimating equations* (GEE) (Liang & Zeger, 1986; Zeger & Liang, 1986) are a tool that results in consistent estimates of the parameters $\boldsymbol{\gamma}$ of such a model (assuming correct specification), acknowledging and accounting for the variance matrix of $\mathbf{Y}$. In broad terms, GEE proceeds by choosing a working covariance model (e.g., independence: $\text{Var}(\mathbf{Y}) = \tau^2 I$; exchangeable: $\text{Cor}(Y_t, Y_{t^*}) = \rho$ for

$t \neq t^*$, etc.), and solving a score equation for the model, alternately solving for the variance parameters and mean parameters, and iterating until convergence. Liang & Zeger (1986) show the consistency and asymptotic normality of this estimator when the mean model is correct (even when the working variance model is wrong), and derive the sandwich variance estimator, which provides valid asympotic standard errors for the parameters. Note, though, that (2.10) is an associational model, and in the end we wish to make conclusions about causal parameters.

By consistency, we know that $\mathbb{E}[Y_{t+1}|\overline{A}_t = \bar{a}_t] = \mathbb{E}[Y_{t+1}(\bar{a})|\overline{A}_t = \bar{a}_t]$, but it is not guaranteed that $\mathbb{E}[Y_{t+1}(\bar{a})|\overline{A}_t = \bar{a}_t] = \mathbb{E}[Y_{t+1}(\bar{a})]$. It is said that there is *no confounding* if for $t = 0, \ldots, K$,

$$Y_{t+1}(\bar{a}) \perp\!\!\!\perp I(\overline{A}_t = \bar{a}_t), \tag{2.11}$$

and that there is *no unmeasured confounding* (a further restriction of (2.6)) when, for $t = 0, \ldots, K$,

$$Y_{t+1}(\bar{a}) \perp\!\!\!\perp A_j|\overline{W}_j, \overline{A}_{j-1} = \bar{a}_{j-1}, \text{ for } j = 0, \ldots, t. \tag{2.12}$$

Clearly, the independence (2.11) implies that $\mathbb{E}[Y_{t+1}(\bar{a})|\overline{A}_t = \bar{a}_t] = \mathbb{E}[Y_{t+1}(\bar{a})]$, which means that for a completely randomized trial (where there is no confounding by design), the parameters of the GEE have causal interpretation, and no further method is required. In observational studies, there are often *time-dependent confounders*: variables affected by previous treatment that also influence both future treatment assignment and the outcome. It may be the case though that (2.12) holds in an observational study with time-dependent confounders. If so, then the standard GEE analysis with (2.10) would be biased, whether or not one also included the confounders as covariates (due to classic graphical arguments, see Robins (1986) for example). By contrast, with a simple extension to the standard GEE,

the marginal structural model for repeated measures presented in the next section does provide causally interpretable, consistent estimates when (2.12) holds.

2.4 Marginal structural models

The *marginal structural model* (MSM) was introduced in Robins (1997, 1999) as a method for obtaining estimates of time-varying treatment effects in observational studies that accounts properly for time-dependent confounders. Simply put, a MSM is a regression model for the expectation of potential outcome random variables, the parameters of which are estimated with the use of IPT weights. In the point treatment case described in section 2.1, where the covariate-treatment-outcome vector (W, A, Y) is observed for all individuals, one might stipulate the model:

$$\mathbb{E}[Y(a)] = \beta_0 + \beta_1 a \tag{2.13}$$

that would give the value of the mean potential outcome over all possible treatment levels a. If treatment A is binary, then this model is saturated and hence correctly specified (there are two parameters and two unknown mean potential outcomes), such that $\mathbb{E}[Y(0)] = \beta_0$, $\mathbb{E}[Y(1)] = \beta_0 + \beta_1$, and the ATE is given by $\mathbb{E}[Y(1) - Y(0)] = \beta_1$. The MSM is 'marginal', since it models the marginal potential outcome mean, not conditional on any covariates—although structural models that do condition on covariates are possible using the same methodology. If Y were a binary outcome (e.g., indicator of death, coronary event, etc.), then instead one might consider a logistic model:

$$\text{logit}(\mathbb{E}[Y(a)]) = \log\left(\frac{\text{P}(Y(a)=1)}{\text{P}(Y(a)=0)}\right) = \beta_0 + \beta_1 a,$$

so that e^{β_1} would be the causal odds ratio for a unit change in A. Regardless of whether Y is continuous or discrete, when treatment is binary one could use the basic IPTW estimator (2.5) to estimate treatment effects. MSMs become more useful when treatment is continuous by allowing for borrowing of strength across data for different treatment levels. That is, if we wanted to estimate $\mathbb{E}[Y(a)]$ across a continuous interval of possible values of a, it would be tedious to repeatedly calculate estimates using the basic IPTW approach (2.5) (with densities replacing probability mass), and in the end we would not have a very succint summary of how treatment generally relates to the potential outcome mean. MSMs are more parsimonious than non-parametric alternatives, as each parameter can encode causal information across a wide range of possible treatment values. MSMs also can be used for estimating time-varying treatment effects of both continuous and binary exposures. The parameters of MSMs are estimated with generalized IPT weights, which we now describe in the fully general case.

MSMs were originally formulated in the general longitudinal setting of section 2.2, where we observe the time-ordered data vector $(W_0, A_0, W_1, A_1, \ldots, W_K, A_K, Y)$ for all subjects. A MSM is specified by choosing a function g of treatment history (usually chosen as inverse canonical link functions of exponential families composed with linear functions of treatment so as to mirror generalized linear models) and asserting

$$\mathbb{E}[Y(\bar{a})] = g(\bar{a}; \boldsymbol{\beta}). \tag{2.14}$$

For example, if Y is continuous, one MSM recommended in Robins *et al.* (2000) is the linear model,

$$\mathbb{E}[Y(\bar{a})] = \beta_0 + \beta_1 \sum_{t=0}^{K} a_t,$$

or if Y is binary, to use

$$\text{logit}\left(\text{P}(Y(\bar{a})=1)\right) = \beta_0 + \beta_1 \sum_{t=0}^{K} a_t.$$

In both these models, the treatment affects the outcome only in its cumulative amount, such that the particular order or structure of the treatment sequence plays no role. In practice, the analyst may use substantive knowledge to specify a more apt model, perhaps where only the last two treatment assignments a_{K-1}, a_K are involved in $g(\bar{a};\boldsymbol{\beta})$, for instance.

By construction, the parameters $\boldsymbol{\beta}$ directly encode causal meaning. Since—as dictated by the fundamental problem of causal inference—we do not have access to the potential outcome variables $Y(\bar{a})$, however, we must use the observable variables to make inference about $\boldsymbol{\beta}$. This is accomplished with IPT weights. To illustrate, suppose that the treatment A_t at each time point $t = 0, \ldots, K$ is binary, and that the standard no unmeasured confounders (2.6) and positivity (2.7) assumptions hold. Let

$$w_i = \left\{ \prod_{t=0}^{K} \text{P}(A_{t,i} = a_{t,i} | \overline{A}_{t-1,i} = \bar{a}_{t-1,i}, \overline{W}_{t,i} = \overline{w}_{t,i}) \right\}^{-1},$$

where double subscript t, i denotes the variable for the i-th individual measured at time t, and by convention, $A_{-1,i} \equiv \varnothing$. The quantity w_i is the subject-specific IPT weight for subject i, which involves the product of conditional probabilities of receiving the treatment they did receive, conditional on their past treatment and time-dependent covariate history. The corresponding *stabilized weight* is defined as

$$sw_i = w_i \times \left\{ \prod_{t=0}^{K} \text{P}(A_{t,i} = a_{t,i} | \overline{A}_{t-1,i} = \bar{a}_{t-1,i}) \right\},$$

which, by multiplying by a product of conditional probabilities of treatment given only treatment history, results in much less variables weights. These probabilities are not known in the case of an observational study, and thus must be estimated, which is typically done with a series of binomial generalized linear models when treatment is binary. By the results proven in Robins (1997, 1999), by fitting the associational version of the MSM (2.14),

$$\mathbb{E}[Y|\overline{A}] = g(\overline{A};\boldsymbol{\gamma})$$

via standard generalized linear model algorithms, but including estimated subject-specific weights $\widehat{w}_i$ or $\widehat{sw}_i$, then the resulting estimates $\widehat{\boldsymbol{\gamma}}$ are consistent estimates of the causal parameters $\boldsymbol{\beta}$ when the MSM (2.14) is correctly specified and the models for the treatment distribution for estimating w_i are correct. Although both valid, the estimates obtained using the stabilized weights sw_i are less variable than when their unstabilized counterparts w_i are used, and hence the stabilized weights are favoured in the longitudinal case (Robins *et al.* , 2000).

The seminal work of Robins (1997, 1999) also suggested an extension of the MSM to the repeated measures setting introduced in the last section 2.3, and this extension was more fully elaborated in Hernán *et al.* (2002). Now with a sequence of outcomes $\mathbf{Y} = (Y_0, \ldots, Y_K, Y_{K+1})$, the MSM has the form:

$$\mathbb{E}[Y_{t+1}(\bar{a})] = g(\bar{a}_t;\boldsymbol{\beta}). \tag{2.15}$$

For example, one could use the linear cumulative treatment model,

$$\mathbb{E}[Y_{t+1}(\bar{a})] = \beta_0 + \beta_1 \sum_{k=0}^{t} a_k + \beta_2 t.$$

To estimate the parameters $\boldsymbol{\beta}$, one now estimates subject-time-specific IPT weights. The unstabilized weight for subject i at visit t is

$$w_{t,i} = \left\{ \prod_{k=0}^{t} \mathrm{P}(A_{k,i} = a_{k,i} | \overline{A}_{k-1,i} = \bar{a}_{k-1,i}, \overline{W}_{k,i} = \overline{w}_{k,i}) \right\}^{-1},$$

and the corresponding stabilized weight is

$$sw_{t,i} = w_{t,i} \times \left\{ \prod_{k=0}^{t} \mathrm{P}(A_{k,i} = a_{k,i} | \overline{A}_{k-1,i} = \bar{a}_{k-1,i}) \right\}.$$

When the MSM (2.15) and treatment models for $w_{t,i}$ are correctly specified, and the positivity (2.7) and no unmeasured confounders (2.12) assumptions hold, then one can fit the associational GEE model (2.10), with subject-time-specific weights $\widehat{w}_{t,i}$ or $\widehat{sw}_{t,i}$. In doing this, Robins (1997) proved that the resulting estimators $\widehat{\boldsymbol{\gamma}}$ are consistent estimators of $\boldsymbol{\beta}$, the causal parameters of the MSM. Again, the stabilized weights $sw_{t,i}$ result in lower variance estimates of $\boldsymbol{\beta}$, and are thus preferred.

Perhaps due to their simplicity relative to other methods, and their resemblance to standard regression models, MSMs are a very popular tool in epidemiology for controlling for time-dependent confounders in longitudinal observational studies. They have been used in many different areas of public health and epidemiological research. A couple examples are Petersen *et al.* (2007) who use an MSM to assess the effect of using pillbox organizers on adherence to HIV antiretroviral therapy, controlling for time-dependent confounders,

and Tager *et al.* (2004) use an MSM to assess how physical activity and relative muscle mass affect the functional limitation of elderly subjects. On top of the numerous applied examples in the literature, there are a few simulation studies that have been conducted that involved MSMs: Lefebvre *et al.* (2008) simulate a two time point longitudinal study and evaluate the effect of misspecifying treatment models for the IPT weights by including different sets of variables in the models; Talbot *et al.* (2015) simulate longitudinal data with a couple covariates and time points and demonstrate scenarios in which, contrary to the dogma, unstabilized weights outperform stabilized weights for the IPTW estimation of an MSM; and Westreich *et al.* (2012b) simulated survival data over a long follow-up period of (maximum) 10 visits, included one baseline and time-varying covariate, and compared the performance of marginal structural Cox models under a variety of modeling scenarios.

In general, because a model (2.14) or (2.15) must be stipulated in addition to the treatment models for estimating the IPT weights, the MSM approach does have the disadvantage that in the nearly inevitable situation where a model is misspecified, the resulting estimators of treatment effects are no longer consistent. Nonetheless, there is little other option when one wants to gain information for a large number of different treatment sequences while still maintaining parsimony. Given that there is relatively little knowledge of the performance of the repeated measures MSM in estimating specific treatment effects compared to other approaches for dealing with repeated measures data, the simulation study in the next chapter will provide valuable knowledge to fill this gap.

2.5 Targeted estimation

Both the g-formula and marginal structural models described in the previous sections depend on correct model specification: the parametric g-formula requires correct outcome models, and MSMs require correct treatment models for the IPT weights. A class of estimators have been proposed that are so-called "double robust" (e.g., see Kang & Schafer, 2007), which incorporate both treatment and outcome models, and are consistent if at least one of these is correctly specified. A more recent addition to this class of estimators, first described in van der Laan & Rubin (2006), is the *targeted maximum likelihood estimator*, along with its generalization, the *targeted minimum loss-based estimator* (both TMLE). Targeted estimation is an increasingly popular, very general semiparametric framework for estimating statistical parameters, boasting favourable asymptotic properties. Broadly speaking, TMLE proceeds in two steps: (i) direct initial estimation of the distribution of the data (e.g., via maximum likelihood, or machine learning algorithm), and (ii) a bias-reducing/targeting step in which the initial fit is fluctuated in such a manner as to produce a substitution estimator with reduced bias for the parameter of interest.

We will introduce TMLE first in the point treatment context described in section 2.1: let (W, A, Y) be observed for every individual in a sample, where W denotes a vector of covariates, A a binary treatment taking values 0 or 1, and Y an outcome variable. Suppose our target parameter is the ATE, $\psi := \mathbb{E}[Y(1) - Y(0)]$. Under the standard assumptions, (2.8) tells us that

$$\psi = \mathbb{E}_W \left[\mathbb{E}_{Y|A=1,W}(Y|A=1,W) - \mathbb{E}_{Y|A=0,W}(Y|A=0,W) \right].$$

With a parametric g-formula, one would thus estimate ψ by fitting a parametric model for $\mathbb{E}[Y|A,W]$, and then substituting to get

$$\widehat{\psi} = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\mathbb{E}}(Y_i|A_i=1,W_i) - \widehat{\mathbb{E}}(Y_i|A_i=0,W_i)\right),$$

or alternatively, bootstrap resample from $\{W_i\}_{i=1}^n$ and compute the analogous mean difference. The first step of TMLE proceeds in the same way, by getting an initial fit for the model $\mathbb{E}[Y|A,W]$. It should be noted that, since only fitted/predicted values are required for this first step, we are not restricted to standard parametric models; indeed, the developers of TMLE advocate for the use of data-adaptive estimation, particularly with their powerful ensemble method, *Super Learner* (van der Laan *et al.* , 2007).

The initial estimate $\widehat{\mathbb{E}}[Y|A,W]$ is obtained by minimizing some global loss function with respect to the distribution of Y given A and W (e.g., the negative log-likelihood). However, it is desirable to sacrifice some bias or variance of what might be considered nuisance parameters of this distribution, in order to get better estimates of the target parameter ψ. The second step of TMLE achieves this goal by updating the initial fit attained in the first step. Irrespective of the initial estimation method, the final TMLE estimator is

$$\widehat{\psi}^{\text{TMLE}} = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\mathbb{E}}^*(Y_i|A_i=1,W_i) - \widehat{\mathbb{E}}^*(Y_i|A_i=0,W_i)\right), \tag{2.16}$$

where

$$\widehat{\mathbb{E}}^*(Y_i|A_i=a,W_i) := \widehat{\mathbb{E}}(Y_i|A_i=a,W_i) + \widehat{\epsilon}\widehat{h}(a,W_i), \text{ for } a=0,1. \tag{2.17}$$

That is, roughly speaking, the second step of TMLE fluctuates the initial fits $\widehat{\mathbb{E}}$ by a factor $\widehat{\epsilon}\widehat{h}$ to obtain the targeted $\widehat{\mathbb{E}}^*$. The notation $\widehat{\epsilon}$, $\widehat{h}$ is used to make clear that these are

empirical (i.e., estimated from data) counterparts of ϵ, h respectively, which themselves are functions of the true underlying distribution of (W, A, Y). The function h depends on the treatment mechanism (a nuisance parameter in the context of estimating ψ), and is related to the efficient influence function of ψ (see Bickel *et al.* , 1998; Tsiatis, 2007 for a comprehensive treatment of semiparametric theory), whereas ϵ determines the magnitude of the fluctuation, which naturally depends on the amount of confounding remaining after the initial fit. In the case of a binary treatment which we have assumed here,

$$\begin{aligned} h(a, W) &= \frac{a}{\mathrm{P}(A = 1|W)} - \frac{1 - a}{\mathrm{P}(A = 0|W)} \\ &= \begin{cases} \{\mathrm{P}(A = 1|W)\}^{-1}, & \text{if } a = 1 \\ -\{\mathrm{P}(A = 0|W)\}^{-1}, & \text{if } a = 0 \end{cases} \end{aligned}$$

The estimate $\widehat{h}$ is obtained by substituting fitted values $\widehat{\mathrm{P}}(A = a|W)$ from a treatment model into the above expression, which requires the analyst to fit a parametric model or learning algorithm to the treatment mechanism. The efficient influence curve for the ATE, ψ, denoted $IC^*(\psi; h, \mathbb{E}[Y|A, W])$, is, as proven in Rotnitzky *et al.* (1998),

$$\begin{aligned} IC^*(\psi; h, \mathbb{E}[Y|A, W]) = h(A, W)(Y - \mathbb{E}[Y|A, W]) \\ + \mathbb{E}[Y|A = 1, W] - \mathbb{E}[Y|A = 0, W] - \psi. \end{aligned}$$

Let $\widehat{IC}_i^*(\psi) := IC^*(\psi; \widehat{h}, \widehat{\mathbb{E}}[Y_i|A_i, W_i])$, where we substitute in the data for the i-th subject, some $\widehat{h}$ which depends on an estimated treatment model, and some $\widehat{\mathbb{E}}[Y|A, W]$, an estimated outcome model. By standard results then, an estimator $\widehat{\psi}$ that solves the efficient influence function estimating equation

$$\frac{1}{n} \sum_{i=1}^{n} \widehat{IC}_i^*(\widehat{\psi}) = 0 \tag{2.18}$$

has minimal asymptotic variance (i.e., is locally semiparametric efficient) when the treatment and outcome models are correctly specified (Bickel *et al.* , 1998). Further, given the nice properties of influence functions, an estimator that solves this estimating equation (2.18) comes with Wald-type confidence intervals $\widehat{\psi} \pm 1.96\frac{\widehat{\sigma}}{\sqrt{n}}$, where

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\{\widehat{IC}_i^*(\widehat{\psi})\}^2, \tag{2.19}$$

which are easy to compute as well as asymptotically valid.

The last piece of the puzzle is the fluctuation variable ϵ used to compute the TMLE in the second step. The maximum likelihood estimate $\widehat{\epsilon}$ for ϵ is obtained by regressing the outcome Y on $\widehat{h}(A, W)$, with no intercept, and an offset of $\widehat{E}[Y|A, W]$ (from the inital outcome model fit in the first step). From this model, $\widehat{\epsilon}$ is just the coefficient corresponding to $\widehat{h}(A, W)$. For instance, if Y is continuous, one would fit the Gaussian linear model:

$$Y = \widehat{\mathbb{E}}[Y|A, W] + \alpha\widehat{h}(A, W) + \delta, \tag{2.20}$$

where $\delta \sim \mathcal{N}(0, \tau^2)$ is an error term. Given the model, we then assign $\widehat{\epsilon} := \widehat{\alpha}$. Finally, the value for $\widehat{\epsilon}$ is plugged into (2.17) to update the initial outcome model, and $\widehat{\psi}^{\text{TMLE}}$ is obtained via (2.16). The algorithm can iterate over several targeting steps, but it turns out that in most cases (including the case shown in this example with a linear fluctuation function (2.17)), one application of this second step suffices to achieve convergence (Rosenblum & van der Laan, 2010).

Crucially, it has been shown (e.g., Rose & van der Laan, 2011) that the TMLE solves

the efficient influence function estimating equation, (2.18). We will now demonstrate this fact in the example we have developed here for a continuous outcome and linear fluctuation function as in (2.17).

Since $\widehat{\epsilon}$ is the maximum likelihood estimator of the linear model (2.20), it solves the Gaussian score equation. The log-likelihood for (2.20) is

$$\ell(\alpha) = -\frac{1}{2\tau^2}\sum_{i=1}^{n}\left(Y_i - \widehat{\mathbb{E}}[Y_i|A_i, W_i] - \alpha\,\widehat{h}(A_i, W_i)\right)^2.$$

Differentiating and setting to zero, the score equation for the parameter α is

$$0 = \sum_{i=1}^{n}\left(\widehat{h}(A_i, W_i)\left[Y_i - \widehat{\mathbb{E}}[Y_i|A_i, W_i] - \alpha\,\widehat{h}(A_i, W_i)\right]\right)$$

which, solving for α, implies that

$$\widehat{\epsilon} = \frac{\sum_{i=1}^{n}\widehat{h}(A_i, W_i)\left(Y_i - \widehat{\mathbb{E}}[Y_i|A_i, W_i]\right)}{\sum_{i=1}^{n}\widehat{h}(A_i, W_i)^2}. \tag{2.21}$$

Now, we know by (2.16) and (2.17) that

$$\begin{aligned}\widehat{\psi}^{\text{TMLE}} = \frac{1}{n}\sum_{i=1}^{n}\Bigg(&\widehat{\mathbb{E}}(Y_i|A_i = 1, W_i) - \mathbb{E}(Y_i|A_i = 0, W_i) \\ &+ \widehat{\epsilon}\left(\widehat{h}(1, W_i) - \widehat{h}(0, W_i)\right)\Bigg)\end{aligned} \tag{2.22}$$

Let $\widehat{IC}_i^*(\widehat{\psi}^{\text{TMLE}}) = IC^*(\widehat{\psi}^{\text{TMLE}}; \widehat{h}, \widehat{\mathbb{E}}^*[Y_i|A_i, W_i])$. In order to show that the TMLE solves the efficient influence function $\frac{1}{n}\sum_{i=1}^n \widehat{IC}_i^*(\widehat{\psi}^{\text{TMLE}}) = 0$, we need to establish that

$$\begin{aligned} 0 = \frac{1}{n}\sum_{i=1}^{n} \Bigg(& \widehat{h}(A_i, W_i)(Y_i - \widehat{\mathbb{E}}^*[Y_i|A_i, W_i]) \\ & + \widehat{\mathbb{E}}^*[Y|A_i = 1, W_i] - \widehat{\mathbb{E}}^*[Y|A_i = 0, W_i] \Bigg) - \widehat{\psi}^{\text{TMLE}} \end{aligned} \tag{2.23}$$

Notice that

$$\begin{aligned} & \widehat{h}(A_i, W_i)(Y_i - \widehat{\mathbb{E}}^*[Y_i|A_i, W_i]) + \widehat{\mathbb{E}}^*[Y_i|A_i = 1, W_i] - \widehat{\mathbb{E}}^*[Y_i|A_i = 0, W_i] \\ & = \widehat{h}(A_i, W_i)\left(Y_i - \widehat{\mathbb{E}}[Y_i|A_i, W_i] - \widehat{\epsilon}\widehat{h}(A_i, W_i)\right) \\ & \quad + \widehat{\mathbb{E}}[Y_i|A_i = 1, W_i] + \widehat{\epsilon}\widehat{h}(1, W_i) - \widehat{\mathbb{E}}[Y_i|A_i = 0, W_i] - \widehat{\epsilon}\widehat{h}(0, W_i) \\ & = \widehat{h}(A_i, W_i)(Y_i - \widehat{E}[Y_i|A_i, W_i]) + \widehat{\epsilon}\left(\widehat{h}(1, W_i) - \widehat{h}(0, W_i) - \widehat{h}(A_i, W)^2\right) \\ & \quad + \widehat{\mathbb{E}}[Y_i|A_i = 1, W_i] - \widehat{\mathbb{E}}[Y_i|A_i = 0, W_i] \end{aligned}$$

from which it follows that

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{h}(A_i,W_i)(Y_i-\widehat{\mathbb{E}}^*[Y_i|A_i,W_i])+\widehat{\mathbb{E}}^*[Y_i|A_i=1,W_i]-\widehat{\mathbb{E}}^*[Y_i|A_i=0,W_i]\right)\\
&=\frac{1}{n}\sum_{i=1}^{n}\Bigg(\widehat{h}(A_i,W_i)(Y_i-\widehat{E}[Y_i|A_i,W_i])+\widehat{\epsilon}\left(\widehat{h}(1,W_i)-\widehat{h}(0,W_i)-\widehat{h}(A_i,W)^2\right)\\
&\qquad\qquad+\widehat{\mathbb{E}}[Y_i|A_i=1,W_i]-\widehat{\mathbb{E}}[Y_i|A_i=0,W_i]\Bigg)\\
&=\widehat{\psi}^{\text{TMLE}}+\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{h}(A_i,W_i)(Y_i-\widehat{E}[Y_i|A_i,W_i])-\widehat{\epsilon}\widehat{h}(A_i,W_i)^2\right),\text{ by (2.22)}\\
&=\widehat{\psi}^{\text{TMLE}}+\frac{1}{n}\sum_{i=1}^{n}\widehat{h}(A_i,W_i)(Y_i-\widehat{E}[Y_i|A_i,W_i])\\
&\quad-\frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{k=1}^{n}\widehat{h}(A_k,W_k)\left(Y_k-\widehat{\mathbb{E}}[Y_k|A_k,W_k]\right)}{\sum_{j=1}^{n}\widehat{h}(A_j,W_j)^2}\widehat{h}(A_i,W_i)^2,\text{ plugging in (2.21)}\\
&=\widehat{\psi}^{\text{TMLE}}+\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{h}(A_i,W_i)(Y_i-\widehat{E}[Y_i|A_i,W_i])\right)\left(1-\sum_{i=1}^{n}\frac{\widehat{h}(A_i,W_i)^2}{\sum_{j=1}^{n}\widehat{h}(A_j,W_j)^2}\right)\\
&=\widehat{\psi}^{\text{TMLE}}+\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{h}(A_i,W_i)(Y_i-\widehat{E}[Y_i|A_i,W_i])\right)(1-1)\\
&=\widehat{\psi}^{\text{TMLE}},
\end{aligned}
$$

thus establishing (2.23).

Hence, by standard semiparametric theory (Bickel *et al.* , 1998), we have that the TMLE estimator is asymptotically linear (i.e., $\widehat{\psi}^{\text{TMLE}} \xrightarrow{d} \mathcal{N}(\psi, \frac{1}{n}\mathbb{E}[\{IC^*\}^2])$) and is maximally efficient when the treatment and outcome models are correctly specified. It also has valid standard errors, confidence intervals and p-values deriving from (2.19). Furthermore, the TMLE has a "double robustness" property (van der Laan, 2010; van der Laan & Gruber, 2010), which, as mentioned earlier, means that so long as either the treatment model or

outcome model is correctly specified (i.e., can be consistently estimated), then the estimator is consistent for its estimand, even if one model is badly misspecified.

Since its inception, the TMLE algorithm has been extended in several ways. Two major advances have been the *collaborative targeted maximum likelihood estimator* (C-TMLE) (van der Laan & Gruber, 2010) and the extension of TMLE to estimate time-dependent treatment effects in the longitudinal or repeated measures data setting presented in section 2.2 (Petersen *et al.* , 2014; Schnitzer *et al.* , 2013; van der Laan, 2010). In brief, C-TMLE uses cross-validation to iteratively improve the estimate of the nuisance parameter (in the case described above, the treatment mechanism) so as to better target the parameter of interest ψ. The longitudinal TMLE proceeds analogously to the point-treatment case elaborated in this section: in the first step, the distribution of the outcome is fit via the full longitudinal g-formula (2.9), either using parametric models or machine learning methods; secondly, moving recursively backwards through the longitudinal causal structure, the factors in the g-formula are fluctuated to target the ATE, finally yielding a subsitution estimator as in the point-treatment case. The same semiparametric theory is used to prove consistency and double-robustness properties of the longitudinal TMLE. We will see an example of its application in both our simulation study and PROBIT dataset.

As mentioned at the beginning of this section, TMLE has become a steadily more popular estimation paradigm. It has been used in a wide variety of applications and across multiple scientific domains. Just to name a few examples, Spertus *et al.* (2016) used TMLE to assess the performance of hospitals after coronary intervention in terms of excess mortality; Decker *et al.* (2014) applied the longitudinal TMLE to compare the effect of different interventions for adolescent obesity; and Wang *et al.* (2011) use C-TMLE in a

genetics dataset to find quantitative trait loci. In addition to these applications, there have been a few simple simulation studies to verify the properties of the targeted estimators. For example, Porter *et al.* (2011) carried out a point-treatment simulation study to compare the performance of TMLE and C-TMLE to other popular methods, and demonstrate double-robustness; Neugebauer *et al.* (2014) compare the longitudinal TMLE to an inverse probability-weighted estimator of treatment effects of dynamic treatment regimes in a very simple longitudinal simulation study. We are not aware of any simulation studies in the literature evaluating targeted methods that involve more than a few covariates as well as time points. It is therefore important to evaluate the finite sample properties of the longitudinal TMLE and compare it to other methods, in a challenging simulation scenario with many covariates and time points.

2.6 Further Resources

Before moving on to describe our simulation study, it should be noted that although sufficient for the purposes of providing background and motivating our analysis, the literature and theory reviewed in this section only scratch the surface of scholarly work in causal methods. For instance, all of the methods described (g-formula, MSMs, and TMLE) have extensions for dealing with time-to-event data, censoring and missingness, which will be largely ignored in this thesis for brevity. The reader is referred to Hernan & Robins (2010) for a conceptual and theoretical introduction to potential-outcome-based causal inference methods; Robins *et al.* (2004) for a deeper view of the g-formula, an overview of marginal structural Cox models for time-to-event data, and an explanation of how to account for censored data; and Rose & van der Laan (2011) for a thorough treatment of targeted learning methods, and a great starting point for finding resources for particular areas of interest

within the TMLE framework.

Finally, although not discussed much here, the other popular school of potential-outcome-based causal inference is spearheaded by the work of Donald Rubin and colleagues. This group advocates for different methods, including propensity score stratification, weighting and regression (Rosenbaum & Rubin, 1983), as well as principal stratification (Frangakis & Rubin, 2002), and generally favors more Bayesian approaches. We refer the reader to Imbens & Rubin (2015) for an expansive description of these methods, and for references to related published research.

Chapter 3

Simulation Study

It is especially critical in medical statistics that the relative strengths and weaknesses of different available data analytic approaches for addressing certain types of statistical questions are well understood: downstream, this will determine which methods are used in applied research, and ultimately will affect public health screening and prevention strategies, as well as the diagnosis, treatment, and prognosis of patients in the clinic. In general, simulation studies are a valuable and necessary tool for evaluating and comparing statistical methods. Simulations are a tricky business, however, and it can be challenging if not impossible to strike an appropriate balance of representing realistically complex data generation paradigms comparable to what might be seen in practice, while still maintaining conciseness and control over parameters of interest. We do not pretend that the simulations presented in this thesis achieve a perfect balance, but we do aim to uphold best practices (Burton *et al.* , 2006), which include but are not limited to pre-specifying the objectives, providing a reproducible and clear protocol, thoroughly detailing methods used, and justifying the choices made at each step of the design. Complying with these guidelines should facilitate critical appraisal of our methods and the possibility of extending the framework

we used in the future.

3.1 Objectives

As is evident from Chapter 2, there have been several methods developed for the purposes of estimating causally interpretable time-varying treatment effects for longitudinal data, and these often have specializations for the case of repeated measures. Specifically, the parametric g-formula, repeated measures MSM, and longitudinal TMLE, have all been proven to provide consistent results under standard causal assumptions and correct model specification. Although the finite-sample properties of these methods have generally been validated in simple simulations, there has not, to our knowledge, been any systematic comparison of all state-of-the-art methods, in a complex setting with many correlated baseline covariates, many simulated time points where the outcome and time-dependent covariate(s) are measured, and an interesting underlying causal structure. Our objective is to address this gap. In particular, we simulate repeated measures data for which an average treatment effect (ATE) ψ is known, and wish to assess for a collection of estimators:

(a) bias, variance, power, and coverage

(b) impact of model misspecification

(c) impact of different magnitudes of treatment effect

(d) robustness to skewed conditional outcome distribution

(e) effect of increasing sample size

3.2 Simulation protocol

3.2.1 Overview

To achieve our objectives, we simulate under several variations of a common data structure. The variations will be elaborated in the next subsection, but in all cases, we generate data of the form:

(i) $\mathbf{W}_0$: a vector of 30 correlated baseline covariates (includes binary and continuous variables),

(ii) Y_0: a baseline measurement of the continuous outcome variable,

(iii) A_0: the binary treatment assigned at the baseline visit, depending on the observed $(\mathbf{W}_0, Y_0)$,

(iv) (W_1', Y_1): one binary time-varying covariate and the measurement of the outcome at visit 1,

(v) A_1: treatment assigned following visit 1,

(vi) $(W_2', Y_2), A_2, \ldots, (W_{11}', Y_{11}), A_{11}$: time-varying covariate, repeated-measures outcome, and subsequent treatment assignment for visits 2 to 11,

(vii) Y_{12}: the outcome observed at the twelfth and final visit.

We decided that 12 time points is a good target as it is on the larger end of the number of follow-up visits in real-world longitudinal studies and thus would provide a challenge and a good benchmark for our methods. We take the convention (as in Gill & Robins, 2001; Hernán *et al.* , 2002) that at visit t, the time-varying covariate and outcome are measured, and then A_t is determined on the basis of W_t, Y_t and all other past history. In the clinical

research setting, for instance, this occurs when the doctor decides the treatment for the patient based on their current health and their prior history. Thus, for an arbitrary subject, the full history of the subject is represented by the following ordered sequence:

$$(\mathbf{W}_0, Y_0), A_0, (W_1', Y_1), A_1, \ldots, (W_{11}', Y_{11}), A_{11}, Y_{12}.$$

For the sake of simplicity, we assume that the visits $t = 0, 1, \ldots, 12$, are equally spaced in time, and there is no censoring or missing data. As will be seen in the data-generating process specified in the next subsection, the underlying causal DAG is shown in Figure 3.1.

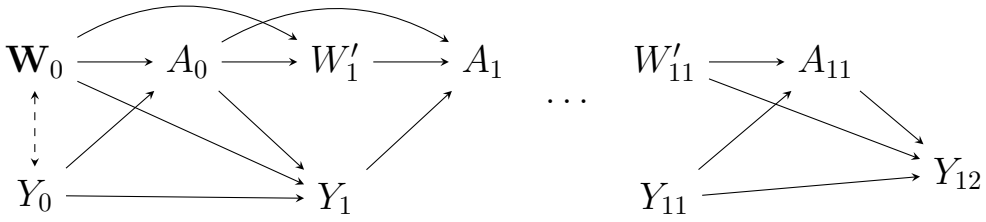

Fig. 3.1 Causal directed acyclic graph of data generating process

Let $\bar{a}^{(1)} = (1, 1, \ldots, 1)$ (a 12-vector of 1's) and $\bar{a}^{(0)} = (0, 0, \ldots, 0)$ (a 12-vector of 0's). In order to make a fair comparison of the different methods, we select one estimand, the ATE at the final visit comparing an 'always treated' intervention (intervening to force $\overline{A} \equiv \overline{A}_{11} = (A_0, A_1, \ldots, A_{11}) = \bar{a}^{(1)}$) to a 'never treated' intervention (setting $\overline{A} = \bar{a}^{(0)}$). That is, we are interested in $\psi := \mathbb{E}[Y_{12}(\bar{a}^{(1)}) - Y_{12}(\bar{a}^{(0)})]$.

3.2.2 Simulation algorithm

To simulate our data, we made use of the relatively new `simcausal` package in `R` (Sofrygin *et al.* , 2015), which facilitates and streamlines generating longitudinal data with many

time points. See Appendix A for the exact `R` code that was used to generate our simulated data. we now describe in detail the simulation procedure.

First, we simulate the 30 baseline covariates $\mathbf{W}_0$, in an approach derived from Setoguchi *et al.* (2008), which generates correlated variables first using the multivariate normal distribution (with mean 0, and marginal variances of 1), and dichotomizes some components of this random vector to produce binary covariates. We also simultaneously generate the baseline outcome variable so as to produce correlation with the baseline covariates, which likely occurs in most real-world data. We decided to include 30 covariates as this was considerably closer to a realistic number of variables to measure in a given study than the 1 or 2 used in most simulation studies, and that it was still small enough to keep the computation time manageable. Specifically, we first generate a collection of hidden continuous 'base' variables $\mathbf{V}$, a (30×1) random vector, with the multivariate normal distribution:

$$\begin{bmatrix} \mathbf{V} \\ Y_0 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_V & \boldsymbol{\alpha} \\ \boldsymbol{\alpha}^\top & \sigma^2 \end{bmatrix} \right),$$

where Σ_V, the covariance matrix of $\mathbf{V}$, is set to be a (30×30) matrix with 1's on the diagonal entries and various other correlations between -1 and 1 in the off-diagonal entries, and the variance of baseline outcome Y_0 is set to $\sigma^2 = 1$. Here, $\boldsymbol{\alpha}$ is a vector of correlation coefficients, some of which are 0 ($\alpha_1, \ldots, \alpha_{10}$ and $\alpha_{21}, \ldots, \alpha_{30}$ are non-zero, so the corresponding variables in $\mathbf{V}$ are associated with the baseline outcome). We then create the observed baseline covariates $\mathbf{W}_0$ by dichotomizing some of the components of V. For components $j \in \{1, 5, 6, 8, 9, 12, 13, 15, 18, 19, 22, 24, 26, 28, 30\}$ of the random vector $\mathbf{V}$, we

dichotomized as follows:

$$W_{j,0} := \begin{cases} 1, & \text{if } V_j \geq 0 \\ -1, & \text{otherwise} \end{cases}$$

which meant each of these binary factors would still have mean zero. For the baseline covariate that would become the binary time-varying covariate, $W_0' \equiv W_{3,0}$, we dichotomized slightly differently:

$$W_{3,0} := \begin{cases} 1, & \text{if } V_3 \geq 1.65 \\ 0, & \text{otherwise} \end{cases}$$

so that W_0' would take values in $\{0, 1\}$ like an indicator variable, and would be equal to 1 in roughly 5% of subjects. The remaining components j of $\mathbf{W}_0$ were kept continuous and set exactly equal to the corresponding component of $\mathbf{V}$. See Appendix A for the code and specific parameter settings (e.g., correlation structure) for the baseline covariate generation.

The baseline exposure A_0 is then generated as follows:

$$\operatorname{logit} \mathrm{P}(A_0 = 1) = \mathbf{W}_0^\top \boldsymbol{\beta} + 3 \times I(Y_0 \leq -3.2),$$

where $\boldsymbol{\beta}$ is a vector of coefficients for the treatment distribution ($\beta_1, \beta_2, \ldots, \beta_{20}$ are non-zero, so these are predictors of baseline treatment status). A very significant predictor of baseline treatment status is whether Y_0 is below the threshold shown above. This is meant to mimic how in treatment of HIV, one strategy is to begin treatment (A_0) when CD4 count (Y_0) is below a threshold (note that the particular value -3.2, and indeed the entire scale on which $Y_0, Y_1, \ldots, Y_{12}$ are measured is entirely arbitrary, and is not meant to accurately represent any real-world variable).

For the remaining repeated measures outcomes $Y_1, \ldots, Y_{12}$, we examine two conditional outcome distributions, one symmetric and one skewed:

$$\text{Symmetric: } Y_t \sim \mathcal{N}(Y_{t-1} + \alpha' W'_{t-1} + \gamma A_{t-1}, \sigma_Y^2 = 0.25^2),$$

$$\text{Skewed: } Y_t = Y_{t-1} + \alpha' W'_{t-1} + \gamma A_{t-1} + (Z - \mathbb{E}[Z]), \text{ where } Z \sim \text{Gamma}(a = 1.5, b = 0.204).$$

The conditional mean of Y_t, $(Y_{t-1} + \alpha' W'_{t-1} + \gamma A_{t-1})$, is the same in both cases, and the shape (a) and scale (b) parameters of the gamma distribution were chosen to match the variance of the continuous case (i.e., $ab^2 = \sigma_Y^2 = 0.25^2$). This allows us to isolate the effect of the skewness of the distribution. To visualize these two distributions, see Figure 3.2, which overlays the symmetric and skewed densities when the mean is 0.

Once the baseline treatment A_0 is simulated as described above, we then generate the data for the remaining visits as follows:

for $t = 1, \ldots, 11$ **do**

Simulate Y_t via

$$Y_t \sim \mathcal{N}(Y_{t-1} + \alpha' W'_{t-1} + \gamma A_{t-1}, \sigma_Y^2 = 0.25^2),$$

for symmetric outcome data, or

$$Y_t = Y_{t-1} + \alpha' W'_{t-1} + \gamma A_{t-1} + (Z - \mathbb{E}[Z]), \text{ where } Z \sim \text{Gamma}(a = 1.5, b = 0.204),$$

for skewed outcome data.

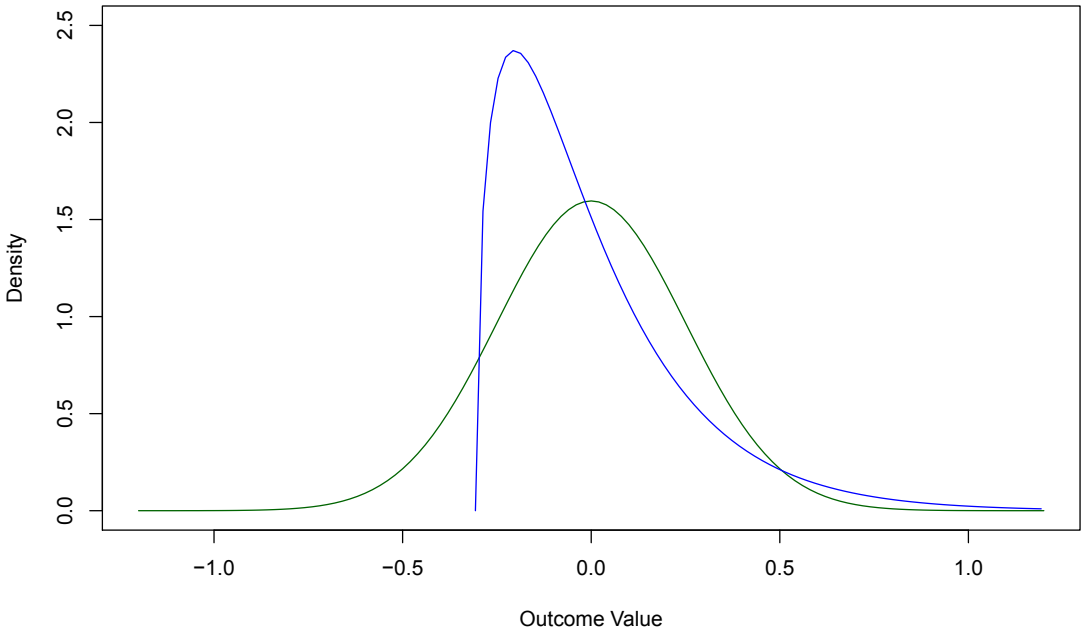

Fig. 3.2 Conditional densities for symmetric (green line) and skewed (blue line) outcome distributions

Simulate W'_t via

$$\mathrm{P}(W'_t = 1) = \begin{cases} p_{1,1} & \text{if } A_{t-1} = 1, W_{t-1} = 1 \\ p_{1,0,t} & \text{if } A_{t-1} = 1, W_{t-1} = 0 \\ p_{0,1} & \text{if } A_{t-1} = 0, W_{t-1} = 1 \\ p_{0,0,t} & \text{if } A_{t-1} = 0, W_{t-1} = 0 \end{cases}$$

Simulate A_t via

$$\mathrm{P}(A_t = 1) = 1, \text{ if } A_{t-1} = 1, \text{ and otherwise:}$$

$$\mathrm{P}(A_t = 1) = \begin{cases} 0.8 & \text{if } Y_t \leq -3.2, W_t = 1 \\ 0.6 & \text{if } Y_t \leq -3.2, W_t = 0 \\ 0.1 & \text{if } Y_t > -3.2, W_t = 1 \\ 0.02 & \text{if } Y_t > -3.2, W_t = 0 \end{cases}$$

end for

Finally, we generate the outcome at the final visit:

$$Y_{12} \sim \mathcal{N}(Y_{11} + \alpha' W'_{11} + \gamma A_{11}, \sigma_Y^2 = 0.25^2),$$

for symmetric outcome data, or

$$Y_{12} = Y_{11} + \alpha' W'_{11} + \gamma A_{11} + (Z - \mathbb{E}[Z]), \text{ where } Z \sim \text{Gamma}(a = 1.5, b = 0.204),$$

for skewed outcome data.

This data generation procedure is represented by the DAG in Figure 3.1. Notice that, as in Hernán *et al.* (2002), once treatment is initiated, the subject invariably remains exposed for the remainder of the study. By definition, W_t' is a time-dependent confounder, as it is affected by A_{t-1}, and influences both A_t and Y_{t+1}. In our simulated data, if we consider higher values of the outcome to be preferable (e.g., CD4 count), then W_t' is a negative trait (e.g., indicator of detectable viral load) that, say, causes physicians to be more likely to initiate treatment, and causes lower values of the next measurement of the outcome.

An important component of our simulation is to represent three levels of treatment efficacy: (i) no effect, (ii) small effect, and (iii) large effect. The 'no effect' scenario is achieved by setting

$$\gamma = 0, p_{1,1} = p_{0,1} = 0.4, p_{1,0,t} = p_{0,0,t} = 0.1361 + \frac{t}{72};$$

the 'small effect' scenario is

$$\gamma = 0.0642, p_{1,1} = 0.4, p_{0,1} = 0.5, p_{1,0,t} = 0.1361 + \frac{t}{96}, p_{0,0,t} = 0.1361 + \frac{t}{72};$$

and the 'large effect' scenario is:

$$\gamma = 0.3278, p_{1,1} = 0.4, p_{0,1} = 0.75, p_{1,0,t} = 0.1361, p_{0,0,t} = 0.1361 + \frac{t}{72}.$$

These parameter settings were chosen so that under no effect, $\psi = \mathbb{E}[Y_{12}(\bar{a}^{(1)}) - Y_{12}(\bar{a}^{(0)})] = 0$, while for a small effect, $\psi = 1$, and for a large effect, $\psi = 5$. We interpret γ as the di-

rect effect of treatment at time t on the outcome at time $(t+1)$, and the probabilities $\{p_{1,1}, p_{0,1}, p_{1,0,t}, p_{0,0,t}\}$ as determining the chance of the time-varying covariate at time t being 1. Thus, the positive effect of treatment on the outcome in the small and large effect settings come from both decreasing the probability of the negative trait W_t' occuring, as well as a direct effect on the outcome Y_{t+1} through γ. To visualize the sequence of potential outcome means in one of these three scenarios, see Figure 3.3, which shows how the expected potential outcome evolves over time for the two interventions of interest when treatment has a large positive effect. The mean potential outcomes in this graph—and for the other treatment effect settings—were calculated by recursively applying the iterated expectation identity and using the known underlying distributions of the observed variables. Notice in the figure that at visit 12, the difference between the expected potential outcomes is exactly $\psi = 5$.

The final tuning parameter of interest in our study is the sample size. We consider three different sample sizes, $n = 1000, 2000, 5000$. We do not consider sample sizes less than 1000 as our methods failed to converge in test runs of our simulation when the sample size was small. On the other hand, we chose 5000 as the upper limit as it is a respectably large sample in most applications, and was not so large as to make the computation infeasible. The simulation study then consists of simulating data as described here for 1000 replications each, for every combination of effect size (no effect, small effect, large effect), outcome distribution type (symmetric, skewed), and sample size (1000, 2000, 5000). This allows us to study components (c), (d), and (e) of our objectives outlined in the previous section.

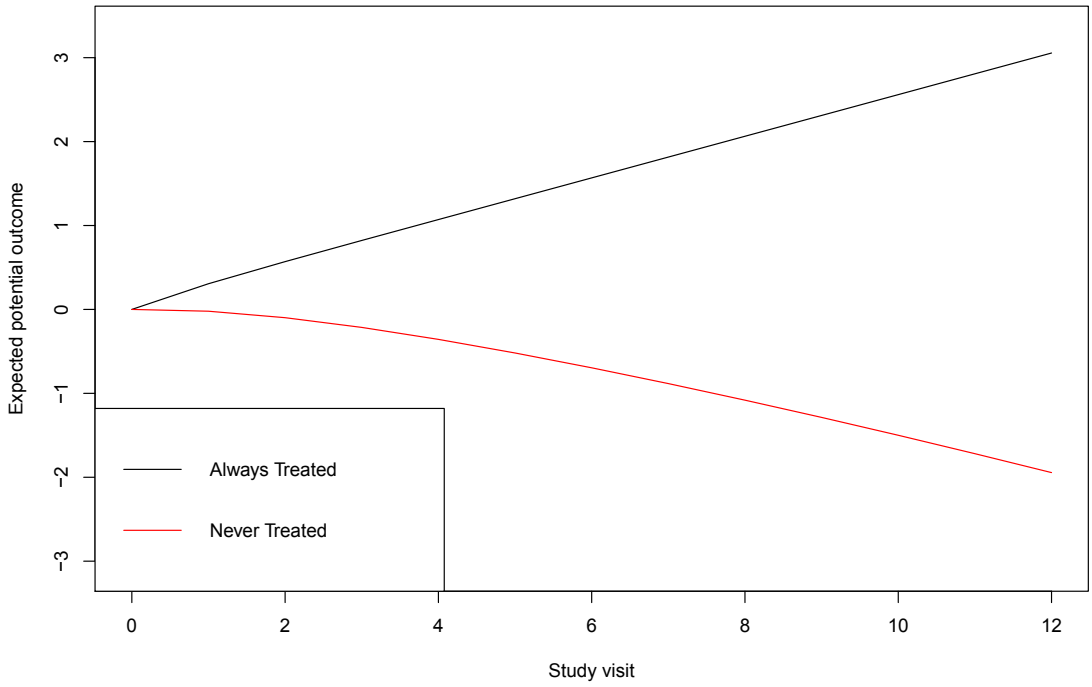

Fig. 3.3 Potential outcome mean of repeated measures outcome across time, for the 'always treated' and 'never treated' interventions, and a large treatment effect

3.3 Estimators

Before describing the methods we compared, it is important to note that, by construction, our simulation approach guarantees that all the standard causal inference assumptions hold. We can generate potential outcomes under any treatment scenario by running our simulation but forcing treatment to be a particular value at each time point. Further, positivity holds (among interventions that satisfy the restriction of remaining on treatment once begun) since the conditional probability of treatment is always positive regardless of past variables. Finally, the no unmeasured confounders criterion follows from the causal structure of the data illustrated in Figure 3.1, as well as the fact that all simulated variables are observed.

We now briefly review the methods described in Chapter 2 and how we applied them to our simulated data. As mentioned, we want to assess statistical performance of each estimator, and also evaluate the impact of model misspecification (objectives (a) and (b) from the overview). Thus, we will apply each method when the corresponding models are correctly specified, as well as when misspecified, and see how statistical performance is affected.

3.3.1 G-formula

The parametric g-formula (see section 2.2) fits parametric models for the outcome and time-varying covariates in order to estimate ψ. For each simulated dataset, we first sampled with replacement from the rows to obtain baseline covariates and outcome $(\mathbf{W}_0, Y_0)$. We considered three modeling scenarios for the remaining time-varying covariate and outcome variables. The first was *correct specification*, a simple linear regression of outcome

given treatment, time-varying confounder, and outcome measured at the previous time point, and a logistic regression of the time-varying confounder given its previous value, the last treatment value and an interaction term. The *misspecified* scenario left the time-varying confounder model correctly specified, but modeled the outcome given only its prior value, leaving out treatment and the time-varying confounder. Finally, we considered a modeling approach where both outcome and time-varying confounder were modeled at each time point given *all previous variables*, including baseline variables, which therefore would include more variables than is necessary and could introduce noise, for instance. We simulated $m = 5,000$ potential outcomes for both 'always treat' $\bar{a}^{(1)}$ and 'never treat' $\bar{a}^{(0)}$ treatment sequences and took the difference in means to obtain $\widehat{\psi}$. On a subset of simulated datasets (to save computational time), a 95% confidence interval for the estimate $\widehat{\psi}$ was obtained by bootstrap resampling 500 times from the data, each time simulating $m = 5,000$ potential outcomes for both treatment sequences, and then taking the 2.5 and 97.5 percentiles of the resulting 500 estimates.

3.3.2 Marginal Structural Models

We considered both a linear and quadratic specification of a repeated measures marginal structural model (MSM; see section 2.4). The linear MSM is

$$\mathbb{E}[Y_{t+1}(\bar{a})] = \beta_0 + \beta_1 \sum_{k=0}^{t} a_k + \beta_2 t,$$

and the quadratic model is

$$\mathbb{E}[Y_{t+1}(\bar{a})] = \beta_0 + \beta_1 \sum_{k=0}^{t} a_k + \beta_2 t + \beta_3 \left(\sum_{k=0}^{t} a_k \right)^2 + \beta_4 t^2.$$

As seen in Figure 3.3, the actual sequences of potential outcome means are curved, thus the linear MSM is misspecified. In fact, the true function that describes these sequences is an 11-th degree polynomial of treatment (follows from recursive application of iterating expectations and using information from the underlying data generating process), and could never actually be correctly specified in practice. However, the quadratic model is closer to the truth than the linear MSM, and so might be expected to perform better.

For the IPT weight treatment models, we considered two scenarios, *correct specification*, and *misspecified models*. The correctly specified denominator treatment model for the IPT weights was a logistic regression including all the variables that are parents of the treatment node in the true DAG, and used an indicator for whether the outcome was below the threshold described in the simulation protocol (i.e., $I(Y_t \leq -3.2)$). The misspecified model included all past variables, but left the outcome as a linear term (i.e., Y_t). The MSMs were fit using the `geepack` package in `R` (Halekoh *et al.* , 2006).

3.3.3 Targeted maximum likelihood estimation

TMLE depends on both outcome and treatment models (see 2.5) in order to estimate ψ. We examine several scenarios—combining the outcome models used for the g-formula described above, and the treatment models used for the MSMs described above—that allow us to assess the impact of model misspecification as well as demonstrate the double robustness property of TMLE. The *correct specification* scenario uses the correct outcome and treatment models; we *misspecify the treatment model*, but keep the outcome model correct; we also *misspecify the outcome model*, but keep the treatment model correct; we *misspecify both the treatment and outcome models*; and finally we include *all past variables* in both treatment and outcome models. The confidence intervals for the ATE estimates

were obtained via influence-curve based methods. We used the `ltmle` package in `R` to carry out the longitudinal TMLE algorithm (Schwab *et al.* , 2016).

3.4 Results

We now summarize the results of our simulation. In the interest of brevity and clarity, we present in this section the results for $n = 2000$, null and large treatment effects, and leave the remaining numerical results in Appendix B. The results for $n = 1000$ and $n = 5000$ are similar, with expected differences: unbiased methods have a smaller ($n = 5000$) or larger ($n = 1000$) spread around ψ, and biased methods become less ($n = 5000$) or more ($n = 1000$) spread around the wrong target.

The simulation results for symmetric outcome data, and where $\psi = 0$ (no treatment effect), are shown in Table 3.1. Looking at the correctly specified models, we see that the g-formula and TMLE are unbiased for ψ, and relatively efficient. The influence-curve based confidence intervals for TMLE perform well, as coverage is near 95%. The bootstrap-based confidence intervals for the g-formula, however, are overly conservative, as coverage is 100%. There is a small bias in the MSMs even for correctly specified treatment models, and the quadratic MSM is slightly less biased and has confidence intervals with better coverage (but overly conservative) than the linear MSM. As a follow-up analysis (not shown here), we simulated 250,000 observations from our model (for null, small, large treatment effects and symmetric outcome data), and the quadratic MSM fit nearly perfectly to the true sequence of expected potential outcomes. Therefore, the bias of the MSMs when the treatment models are correct is likely due to the inevitable slight misspecification of the form of the MSM combined with highly variable estimated IPT weights resulting from the finite

Table 3.1 Simulation results: $n = 2000$, symmetric outcome, null treatment effect

(a) Correct Model Specification

	Bias	SE	Coverage (%)
G-formula	0.002	0.068	100.0
Linear MSM	-0.089	0.141	87.5
Quadratic MSM	-0.074	0.145	100.0
TMLE	0.003	0.087	94.3

(b) Treatment Model Misspecified

	Bias	SE	Coverage (%)
Linear MSM	-0.062	0.105	97.7
Quadratic MSM	-0.107	0.145	100.0
TMLE	0.002	0.084	94.2

(c) Outcome Model Misspecified

	Bias	SE	Coverage (%)
G-formula	0.001	0.031	100.0
TMLE	-0.002	0.089	94.7

(d) Outcome and Treatment Model Misspecified

	Bias	SE	Coverage (%)
TMLE	-0.089	0.084	81.6

(e) Models Controlling for All Past Variables

	Bias	SE	Coverage (%)
G-formula	0.003	0.086	100.0
TMLE	0.002	0.086	94.1

sample size. As seen in Table 3.1, the MSMs and g-formula perform nearly the same for the null effect setting even when the treatment models and outcome models, respectively, are misspecified. Double robustness of the TMLE is clearly demonstrated by our results, as estimates are unbiased and have valid confidence intervals when either the treatment model, outcome model, or both, are correctly specified. Performance of the TMLE does drop when both models are misspecified. Finally, the g-formula and TMLE using naive models where the outcome model depends on all previously measured variables perform

essentially identically to when models are correctly specified with no unnecessary variables, but with a small amount of efficiency lost (increased standard error) for the g-formula, and a slight efficiency gain for the TMLE.

Table 3.2 shows the results for the same setting, but where treatment has a large positive effect on the outcome ($\psi = 5$). The results are mostly similar to the null effect case with a few notable distinctions. First, the quadratic MSM is much better on average in

Table 3.2 Simulation results: $n = 2000$, symmetric outcome, large treatment effect

(a) Correct Model Specification

	Bias	SE	Coverage (%)
G-formula	-0.000	0.071	100.0
Linear MSM	0.237	0.367	48.5
Quadratic MSM	0.085	0.335	100.0
TMLE	-0.001	0.111	95.1

(b) Treatment Model Misspecified

	Bias	SE	Coverage (%)
Linear MSM	0.334	0.110	29.0
Quadratic MSM	0.359	0.167	100.0
TMLE	0.021	0.087	93.7

(c) Outcome Model Misspecified

	Bias	SE	Coverage (%)
G-formula	4.998	0.046	0.0
TMLE	-0.025	0.127	93.8

(d) Outcome and Treatment Model Misspecified

	Bias	SE	Coverage (%)
TMLE	0.169	0.088	59.7

(e) Models Controlling for All Past Variables

	Bias	SE	Coverage (%)
G-formula	0.010	0.087	100.0
TMLE	-0.003	0.089	95.0

this setting than its linear counterpart when the treatment models are correct. Second, there is a substantial bias in both the MSMs and g-formula when the treatment model and outcome model, respectively, are misspecified. This shows that the results in the null effect setting were spurious, and that these methods are not, in general, robust to model misspecification. Third, although again TMLE is double robust in the large effect case, the performance when both treatment and outcome models are misspecified seems to be worse than in the null effect setting.

The results for null and large treatment effects in the skewed outcome setting are shown in Tables 3.3 and 3.4, respectively. As a whole, the results for skewed conditional outcome data are entirely analogous to the symmetric data case. Our findings suggest, therefore that all of these methods are robust to skewness in the outcome, and that this skewness need not be directly modeled.

Table 3.3 Simulation results: $n = 2000$, skewed outcome, null treatment effect

(a) Correct Model Specification

	Bias	SE	Coverage (%)
G-formula	0.000	0.067	99.0
Linear MSM	0.085	0.166	87.2
Quadratic MSM	0.068	0.174	100.0
TMLE	0.003	0.090	95.2

(b) Treatment Model Misspecified

	Bias	SE	Coverage (%)
Linear MSM	0.061	0.107	97.9
Quadratic MSM	0.101	0.146	100.0
TMLE	0.001	0.083	95.4

(c) Outcome Model Misspecified

	Bias	SE	Coverage (%)
G-formula	-0.001	0.031	100.0
TMLE	0.007	0.095	94.7

(d) Outcome and Treatment Model Misspecified

	Bias	SE	Coverage (%)
TMLE	0.092	0.083	79.6

(e) Models Controlling for All Past Variables

	Bias	SE	Coverage (%)
G-formula	-0.000	0.087	99.0
TMLE	0.001	0.085	94.4

Table 3.4 Simulation results: $n = 2000$, skewed outcome, large treatment effect

(a) Correct Model Specification

	Bias	SE	Coverage (%)
G-formula	-0.001	0.070	99.0
Linear MSM	0.259	0.319	42.5
Quadratic MSM	0.095	0.302	100.0
TMLE	-0.001	0.102	95.4

(b) Treatment Model Misspecified

	Bias	SE	Coverage (%)
Linear MSM	0.344	0.110	25.6
Quadratic MSM	0.369	0.197	100.0
TMLE	0.019	0.081	96.0

(c) Outcome Model Misspecified

	Bias	SE	Coverage (%)
G-formula	4.998	0.046	0.0
TMLE	-0.026	0.114	94.9

(d) Outcome and Treatment Model Misspecified

	Bias	SE	Coverage (%)
TMLE	0.168	0.083	62.4

(e) Models Controlling for All Past Variables

	Bias	SE	Coverage (%)
G-formula	0.008	0.085	99.0
TMLE	-0.003	0.084	96.5

Chapter 4

PROBIT Data Analysis

In this chapter, we analyze the PROmotion of Breastfeeding Intervention Trial (PROBIT), a cluster-randomized study which, among other variables, measured infant weight and breastfeeding status across multiple follow-up visits. We apply the methods for analyzing repeated measures data described in the previous chapters.

4.1 The PROBIT study

The PROBIT study, first analyzed in Kramer *et al.* (2001), was a large cluster-randomized trial conducted in Belarus from June 1996 to December 1997, following participants over a period of 12 months. The clusters consisted of 31 maternity hospitals and their polyclinics, each of which randomized to an intervention (promoting breastfeeding practice), or to control (standard care). The original aim of the study was to assess the effect of promoting breastfeeding on the duration and exclusivity of breastfeeding, as well as infant infection and eczema.

The data consisted of 17,046 mother-infant pairs, and was restricted to pairs with a healthy mother initially intending to breastfeed, and a full-term singleton newborn weighing at least 2.5 kg (Kramer *et al.* , 2001). The pairs were assessed at baseline and subsequent visits at 1, 2, 3, 6, 9, and 12 months thereafter. Baseline variables included hospital region, child's sex, maternal age, past smoking status and breastfeeding history of the mother, maternal education, and family history of atopy. Variables measured at follow-up visits included breastfeeding status, the child's weight and time-varying health status, and smoking and drinking habits of the mother.

4.2 Notation and assumptions

We will use the PROBIT data to estimate the effect of breastfeeding on infant weight. Let (W_{ij}, A_{ij}, Y_{ij}) denote, for mother-infant pair i, the vector of covariates (W_{ij}) measured at month j after baseline, breastfeeding indicator (A_{ij}) for the period preceeding month j, and infant weight (Y_{ij}) measured at month j after baseline. The observed data consists therefore of these variables, for $j \in \mathcal{T} = \{0, 1, 2, 3, 6, 9, 12\}$, and for $i = 1, \ldots, 17,046$. In order to proceed, we make the standard causal assumptions. First, we assert the existence of potential outcomes $Y_{ij}(\bar{a})$, the outcome for mother-infant pair i, at month j, if possibly counter to fact they had breastfeeding sequence $\overline{A}_i \equiv (A_{i0}, A_{i1}, A_{i2}, A_{i3}, A_{i6}, A_{i9}, A_{i12}) = \bar{a}$. We also assume consistency—that if in fact the $\overline{A}_i = \bar{a}$ was observed, that $Y_{ij}(\bar{a}) = Y_{ij}$—as well as positivity,

$$f(\bar{a}_{k_{\text{prev}}}, \overline{w}_k) > 0 \implies f(a_k | \bar{a}_{k_{\text{prev}}}, \overline{w}_k) > 0,$$

and no unmeasured confounders

$$Y_{ij}(\bar{a}) \perp\!\!\!\perp A_{ik} | \overline{W}_{ik}, \overline{A}_{ik_{\text{prev}}} = \bar{a}_{k_{\text{prev}}}, \text{ for } k = 0, \ldots, j,$$

where k_{prev} denotes the month of the visit prior to that at month k. By design, $A_{i0} = 1$ for all mother-infant pairs, i.e., all mothers breastfed their infants at baseline. In addition, if $A_{ij} = 0$, then breastfeeding status was 0 at all future follow-up visits, as no mothers reinitiated breastfeeding after stopping (similar to the simulation study, this restriction in itself does not constitute a violation of positivity, since the treatment sequences of interest defined below satisfy this constraint). Given these restrictions, there are only seven possible breastfeeding exposure sequences; let $\bar{a}^{(j)}$ denote the 7-vector of binary indicators of breastfeeding corresponding to the months in $\mathcal{T}$ which is 1 up to and including the component corresponding to month j, and 0 onwards. Our estimand of interest is the average treatment effect (ATE) at 12 months of an intervention where mothers breastfeed throughout the entire study period versus weaning immediately after baseline: $\psi := \mathbb{E}[Y_{i12}(\bar{a}^{(12)}) - Y_{i12}(\bar{a}^{(0)})]$, where $\bar{a}^{(12)} = (1, 1, \ldots, 1)$ and $\bar{a}^{(0)} = (1, 0, \ldots, 0)$.

We note that our approach is analogous to a per-protocol analysis as opposed to an intention-to-treat approach (Hernán & Hernández-Díaz, 2012), since we use the actual observed breastfeeding behaviour instead of the 'intended' exposure (i.e., that everyone in the breastfeeding promotion intervention group would adhere to the recommended breastfeeding duration). There was a small amount of intermittently missing data among the covariates (ranged from 0-17% missing) and outcome (ranged from 0-6% missing at different time points), which was handled as in Platt *et al.* (2009) using median imputation. Finally, we note that the choice of models outlined below (i.e., linear models for outcome, logistic models for treatment and time-varying covariates, restricting to controlling for confounders measured at previous visit, inclusion of squared terms in models) was based on the PROBIT analysis of Platt *et al.* (2009), who carefully selected their models and ensured adequacy of the fit.

4.3 Methods

4.3.1 G-formula

As elaborated in section 2.2, the parametric g-formula estimates an ATE by simulating potential outcomes via fitted parametric models. This is done by simulating forward in time, generating the data for each visit given past variables and setting exposure to the desired values (e.g., breastfeed for entire study duration or stop after baseline). Here, we bootstrap resampled from the PROBIT dataset to get baseline covariates (hospital and region; intervention group; maternal smoking history, breastfeeding history, age, number of children, and education; family history of atopy; child sex, indicator of cesarean birth, and baseline weight). Models were then fit for time-varying covariates and the repeated measures outcome given past variables. The four time-varying covariates were binary indicators of maternal smoking status, drinking status, child illness in the last period, and child hospitalization in the last period. The models for each of these at month 1 after baseline were logistic regression models with all baseline covariates as main effects, and at months 2, 3, 6, 9, 12 were logistic regression models controlling for the time-varying covariates and child's weight measured at the previous visit. The outcome (weight of infant) model at month 1 was a linear regression including all baseline variables as main effects, and at months 2, 3, 6, 9, 12 were linear regression models again controlling for the time-varying covariates and weight measured at the previous visit, and also included baseline variables. We simulated $m = 500,000$ potential outcomes for both 'breastfeed to 12 months' $\bar{a}^{(12)}$ and 'wean after baseline' $\bar{a}^{(0)}$ breastfeeding sequences and took the difference in means to obtain $\widehat{\psi}$. The 95% confidence interval for the estimate $\widehat{\psi}$ was obtained by bootstrap resampling 1,000 times from the full data, each time simulating $m = 5,000$ potential outcomes for both breastfeeding sequences, and then taking the 2.5 and 97.5 percentiles of the resulting

1,000 estimates.

4.3.2 Marginal Structural Models

We fit both a longitudinal MSM, considering only the outcome at month 12, as well as a repeated measures MSM that models the sequence of correlated outcomes. The longitudinal MSM is

$$\mathbb{E}[Y_{12}(\bar{a})] = \beta_0 + \sum_{j \in \mathcal{T} \setminus \{0\}} \beta_j I(\bar{a} = \bar{a}^{(j)}),$$

and the repeated measures MSM is:

$$\mathbb{E}[Y_j(\bar{a})] = \beta_0^* + \sum_{k \leq j, k \in \mathcal{T} \setminus \{0\}} \beta_k^* I(\bar{a} = \bar{a}^{(k)}) + \alpha t.$$

That is, the longitudinal MSM is a saturated model, modeling the expected potential outcome at month 12 for each possible breastfeeding sequence separately, and the repeated measures MSM models the potential outcome means at each visit, using a linear term for the effect of time (e.g., we would expect infant weight to increase over time in general, so α should be positive). Of course, as described in section 2.4, we need to estimate IPT weights in order to estimate the parameters of these MSMs. The denominator treatment models for the IPT weights used here are identical models to those described in Platt *et al.* (2009). In particular, the probability of treatment at each month j is modeled with a logistic regression including linear terms for previous weight and squared weight, maternal age and squared maternal age, the current value of the four time-varying covariates, and the remaining baseline covariates. To create stabilized weights, the marginal probability of remaining breastfed (in the subset of mother-infant pairs breastfeeding at the past visit) is modeled with the empirical proportion, and used for the numerators of the IPT weights.

The MSMs were fit using the `geepack` package in `R` (Halekoh *et al.* , 2006).

4.3.3 TMLE

The TMLE approach, described in section 2.5, requires specifying both outcome models and treatment models, and uses both to attempt to more efficiently target the parameter ψ. The infant weight outcome models for the TMLE here are the same as was described above for the g-formula, and the treatment models are the same as the denominator models described for the MSMs above. The confidence interval for the ATE estimate is obtained via influence-curve based methods. We used the `ltmle` package in `R` to carry out the longitudinal TMLE algorithm (Schwab *et al.* , 2016).

4.4 Results and interpretation

The results of the main analysis are summarized in table 4.1. Overall, the different methods seem to agree that the average causal effect of breastfeeding to 12 months versus early weaning is significantly negative, i.e., that there is evidence that $\mathbb{E}[Y_{i12}(\bar{a}^{(12)})] < \mathbb{E}[Y_{i12}(\bar{a}^{(0)})]$.

Table 4.1 PROBIT data analysis results: estimates of the average effect of breastfeeding on infant weight at month 12, comparing a 'breastfeed to 12 months' intervention to a 'wean after baseline' intervention

Method	$\widehat{\psi}$	95% CI
G-formula	-0.159	(-0.214, -0.136)
Longitudinal MSM	-0.206	(-0.272, -0.139)
Repeated Measures MSM	-0.203	(-0.263, -0.142)
Longitudinal TMLE	-0.208	(-0.273, -0.142)

We can conclude that at 12 months, an 'always breastfeeding' intervention would result in a decrease in weight of between 0.14 kg and 0.27 kg, on average, compared to an 'early weaning' intervention (if indeed the confounder set controlled for was sufficient). The agreement

among the analytical approaches may be evidence for the robustness of this finding. We can compare this to the intention-to-treat analysis of Kramer *et al.* (2002), which found that at 12 months, the breastfeeding promotion intervention group was 0.007 kg lighter, on average, than the standard care group, but that this was not statistically significant. The crude observational per-protocol difference estimate between those that breastfed to 12 months versus those that stopped after baseline was $\widehat{\psi} = -0.088$ kg. Thus, all methods at least give the same sign. The increased magnitude of the controlled (between -0.139 and -0.273) and uncontrolled (-0.088) per-protocol estimates as compared with the intention-to-treat estimate (-0.007) may be a result of residual unmeasured confounding, or due to a real biological consequence of breastfeeding that was underestimated by the intention-to-treat analysis (e.g., see Hernán & Hernández-Díaz, 2012 for a discussion of this phenomenon).

Note that although the point effect at month 12 is of interest, it is also important to analyze how the two interventions affect infant weight across time. Only the repeated measures MSM can provide simultaneous estimates of this process (the other methods can be used also by considering the outcome at each visit to be the final measurement, but it would require a separate analysis for each time point). Figure 4.1 shows the distribution of the stabilized IPT weights for each study visit. As in Hernán *et al.* (2002), as well as our simulation study in the previous chapter, we see that the weights are centred around 1 (their logarithm is centred around 0), and they increase in spread across time. Further, the median weight drops slightly below 1, and decreases at later study visits, a pattern that also appears to be persistent in such repeated measures settings.

In figure 4.2, we see the modeled expected potential infant weights across time for the two interventions of interest. The pattern matches the observational analysis of the PROBIT

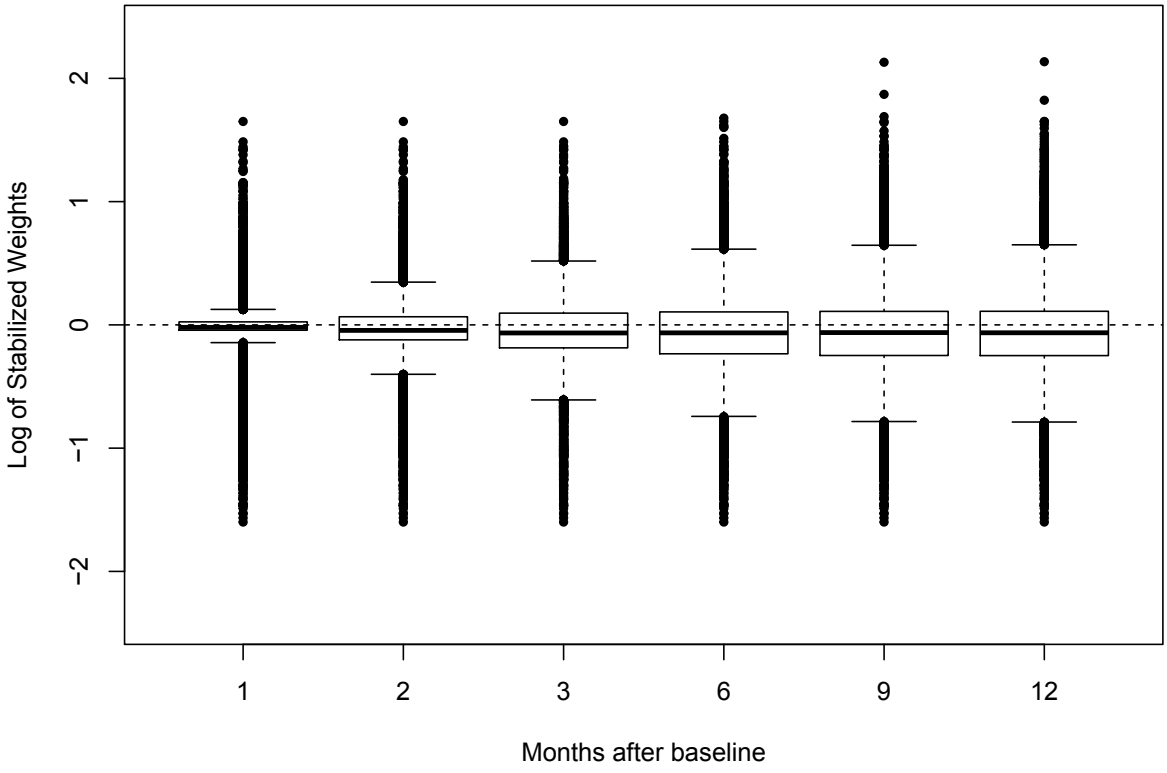

Fig. 4.1 Distribution of stabilized inverse probability weights by follow-up visit

data by Kramer *et al.* (2002), as they conclude that early weaners had lower weight at month 1, but they caught up and eventually outgrew the persistent breastfeeders by the end of the study period.

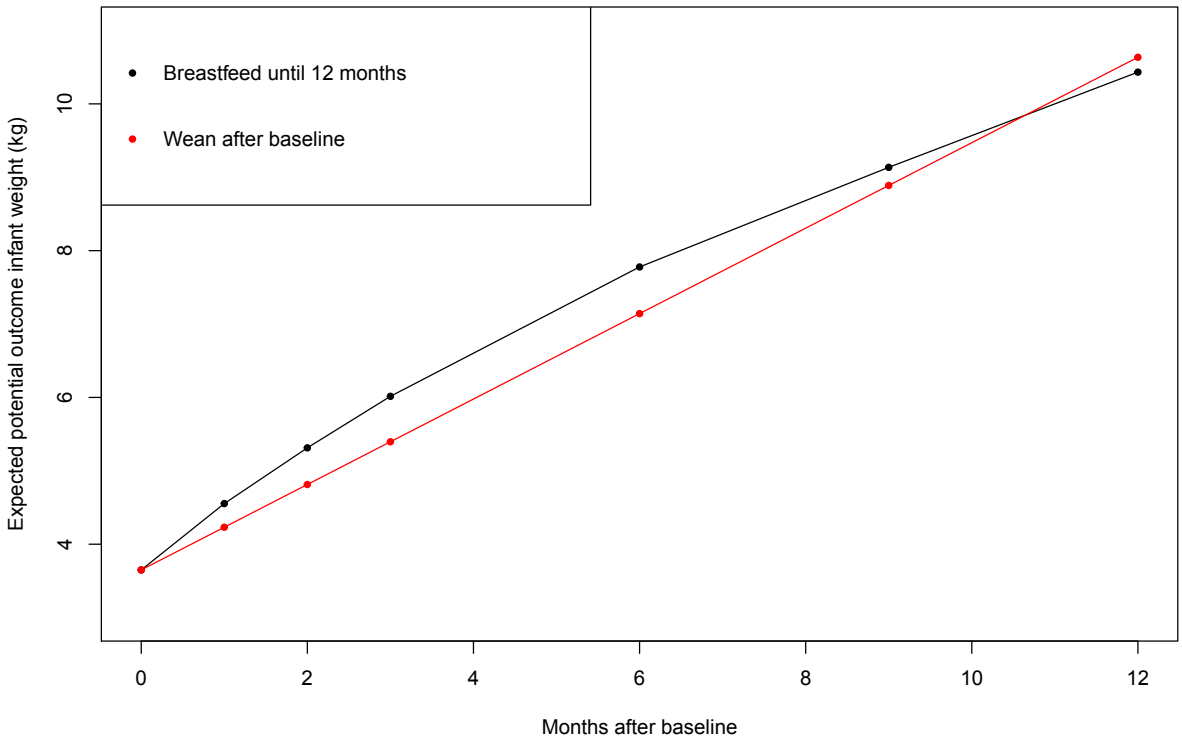

Fig. 4.2 Estimates of potential outcome means from repeated measures MSM for 'breastfeed to 12 months' and 'wean after baseline' interventions

Chapter 5

Conclusion

This thesis presented a simulation study comparing the most promising methods for estimating causal treatment effects in longitudinal or repeated measures observational data, and applied these methods to a real dataset. The simulation study is notable for including both a substantial number of baseline covariates in addition to simulating many follow-up measurements. Further, our simulation procedure may be a valuable starting point for future studies of causal methods for longitudinal data. The results of our PROBIT analysis are consistent with published analyses of this dataset (Platt *et al.* , 2009; Schnitzer *et al.* , 2013).

To place the findings of our simulation study in the broader context of the related literature, it is instructive to reference the findings of Schnitzer *et al.* (2013) and Pang *et al.* (2016a,b). The paper by Schnitzer *et al.* (2013) showed a construction of the longitudinal TMLE in a two time point setting, and evaluated its performance relative to the g-formula and inverse probability weighted estimator, among other methods. They found that TMLE and the g-formula were comparable when all models were correctly specified,

with no advantage for the TMLE from an additional correctly specified treatment model. Our results show the same phenomenon. They also found that TMLE did no worse than other methods when all models are misspecified, which can also be said in our simulation study. The work of Pang *et al.* (2016a) was a large-scale simulation study meant to mimic a high-dimensional, point-exposure pharmacoepidemiologic observational study with a binary outcome. They found that including a rich collection of covariates in the outcome model resulted in unbiased estimation for the TMLE. In our simulation study, we found that when overspecifying the outcome model—that is, including more covariates than is truly necessary—by conditioning on all past variables resulted in good performance of the TMLE. We further found that the g-formula also performed well under this overspecification.

On top of agreeing with past research, we found that in our simulation, the methods were robust in dealing with relatively skewed outcome data, which we do not believe has been assessed previously. In addition, notably, we found that in our data generating paradigm, the functional form of the marginal structural model was next to impossible to correctly specify. This led to the MSMs yielding substantially more biased estimates than the g-formula or TMLE even when the parametric models for the treatment mechanism were correct. We note that performance was somewhat better when the functional form was better approximated, as the quadratic MSM more closely matched the true sequence of expected potential outcomes than the linear MSM. Lastly, we also demonstrated the double robustness property of the TMLE estimator in a complex, longitudinal setting.

Our simulation study has several limitations. First, as with any simulation study, we are only limited to assessing a very restricted class of data generating models, and thus

it is difficult if not impossible to generalize findings to all possible real data generating mechanisms. In real applications, the underlying data generating process may be much more complicated than the standard parametric models used to simulate our data, and may not even adhere to all the necessary assumptions for valid estimation (e.g., no unmeasured confounders or positivity). Second, we did not include censoring or missingness in our simulations, although our framework can easily be extended to incorporate these. Third, we restricted our attention to continuous outcomes; it is also of great interest to study methods for survival data, as well as count or binary data. In general, longitudinal methods for binary or count data are more complicated, but it is important to study these cases nonetheless. Our simulation framework can also easily be extended to generate these types of data, for which different available methods can then be compared. Fourth, we only assessed one pair of treatment sequences, one in which patients are treated at all time points, and one in which they never receive treatment. This is a very simple case and it would be of interest to also investigate treatment sequences that vary across time, or more ambitiously, to compare the effects of different dynamic treatment regimes.

In brief, we developed and presented a complex simulation study of continuous longitudinal observational data, and contrasted the performance of different popular methods for estimating causal treatment effects under a variety of modeling scenarios. Our results may inform the use of these methods in practice, and our simulation framework has the potential to be used as a springboard for future studies.

Appendix A

Data Generation

In this appendix, we provide the R code that was used to generate data for our simulation study. We show here the scenario of skewed conditional outcome data, and a large treatment effect. The remaining scenarios are easily attained by changing the outcome distribution to a Gaussian, and by changing the parameters to the values specified in the text for the different treatment effects.

```
library(simcausal)
options(simcausal.verbose=FALSE)
library(mvtnorm)

p <- 30 # number of covariates measured at baseline, 1 of which will
        # also be time-varying
t.end <- 12 # number of time points at which outcome is observed

### constants used in Setoguchi et al. paper + arbitrary
### constants added by me, specified here
## beta: exposure coefficients
## alpha: outcome coefficients
```

```r
## gamma: treatment effect
beta0 <- -2; alpha0 <- -0.85;
gamma <- 0.3275538
beta <- 0.15*c(0.8, -0.25, 0.6, -0.4, -0.8, -0.5, 0.7, 0.35, -0.62,
               0.43, 0.22, -0.55, 0.29, -0.48, -0.31, 0.36, -0.72,
               0.59, 0.26, -0.2)
alpha <- 0.2*c(0.3, 0.36, -0.43, -0.2, 0.71, -0.19, 0.26, 0.42, 0.21,
               -0.6, -0.28, -0.33, 0.22, -0.56, 0.4, -0.46, -0.29,
               0.37, 0.8, -0.31)
alpha3 <- alpha[3] <- -0.43

# idea: covariates W1, ..., W30, exposure A, outcome Y
# W1 - W10 confounders, associated with A & Y
# W11 - W20 instruments, associated only with A
# W21 - W30 pure outcome predictors, associated only with Y
# the following subset of variables are binary, rest are continuous
binary <- c(1,3,5,6,8,9,12,13,15,18,19,22,24,26,28,30)

### correlation matrix of "base covariates"
cor.mat <- outer(1:(p+1), 1:(p+1), Vectorize(function(x, y) {
  if(x == y) {
    1
  } else if((x == 5 && y == 1) || (x == 1 && y == 5)) {
    0.2
  } else if((x == 6 && y == 2) || (x == 2 && y == 6)) {
    0.9
  } else if((x == 8 && y == 3) || (x == 3 && y == 8)) {
    0.2
  } else if((x == 9 && y == 4) || (x == 4 && y == 9)) {
    0.9
```

```
43    } else if((x == 15 && y == 11) || (x == 11 && y == 15)) {
44      0.3
45    } else if((x == 16 && y == 12) || (x == 12 && y == 16)) {
46      0.6
47    } else if((x == 18 && y == 13) || (x == 13 && y == 18)) {
48      0.3
49    } else if((x == 19 && y == 14) || (x == 14 && y == 19)) {
50      0.6
51    } else if((x == 25 && y == 21) || (x == 21 && y == 25)) {
52      0.4
53    } else if((x == 26 && y == 22) || (x == 22 && y == 26)) {
54      0.5
55    } else if((x == 28 && y == 23) || (x == 23 && y == 28)) {
56      0.4
57    } else if((x == 29 && y == 24) || (x == 24 && y == 29)) {
58      0.5
59    } else if(x == 31 && y %in% c(1:10,21:30)) {
60      if (y <= 10) {
61        alpha[y]
62      } else {
63        alpha[y-10]
64      }
65    } else if(y == 31 && x %in% c(1:10,21:30)) {
66      if (x <= 10) {
67        alpha[x]
68      } else {
69        alpha[x-10]
70      }
71    } else {
72      0
```

```
  }
}))

### simulate from model
# create empty DAG
D <- DAG.empty()

## add baseline covariates
# "base covariates"
V <- paste("V",1:(p+1),sep=""); W <- c(paste("W",1:p,sep=""),"Y")
D <- D + node(V, t=0, distr = "rmvnorm",
              asis.params = list(mean = "rep(0, p+1)",
                                 sigma="cor.mat"))
# "final covariates"
W.from.V <- vapply(c(1:2,4:(p+1)), function(x) {
  ifelse(x %in% binary,
         paste("ifelse(V", x, "_0 >= 0, 1, -1)", sep=""),
         paste("V", x, "_0", sep=""))
}, "")
all.W <- paste('c(',paste(c(W.from.V[1:2],
                            "ifelse(V3_0 >= 1.65, 1, 0)",
                            W.from.V[3:p]),
                          collapse=", "), ')',sep="")
eval(parse(text=paste('D <- D + node(W, t=0, distr = "rconst",
                      const = ', all.W, ')',sep="")))


## 'interleave' is a helper function
## useful for creating linear combination strings
interleave <- function(s1,s2,insert="*") {
```

```
103   ord1 <- 2*(1:length(s1))-1
104   ord2 <- 2*(1:length(s2))
105   woven <- c(s1,s2)[order(c(ord1,ord2))]
106   paste(woven[ord1],woven[ord2],sep=insert)
107 }
108
109 ## linear predictor string, given model paramaters
110 ## and covariate names
111 lin.pred <- function(model.params, covariates) {
112   paste(model.params[1]," + ",
113         paste(interleave(covariates, model.params[-1]),
114               collapse="+"),sep="")
115 }
116
117 ## add random skewed data generator (custom gamma distribution)
118 rcgamma <- function(n, shape, scale, mean) {
119   rgamma(n = n, shape = shape, scale = scale) - shape*scale + mean
120 }
121 vecfun.add("rcgamma")
122
123 ## add baseline exposure
124 # linear logistic model for now: beta0 + W1*beta1 + W2*beta2 + ...
125 trt.fn <- lin.pred(c(beta0,beta), paste("W",1:20,"_0",sep=""))
126 eval(parse(text=paste('D <- D + node("A", t=0, distr = "rbern",
127                       prob = plogis(',
128                       trt.fn, ' + 3*(Y[0] <= -3.2)))', sep="")))
129
130 ## add time-varying response (e.g. CD4 count)
131 D <- D + node("Y", t=1:t.end, distr = "rcgamma",
132               shape = 1.5, scale = 0.204124,
```

```
133                 mean = Y[t-1] + alpha3*W3[t-1] + gamma*A[t-1])
134
135 ## define time-varying confounder
136 D <- D + node("W3", t=1:(t.end-1), distr="rbern",
137                 prob = ifelse(A[t-1] == 1 & W3[t-1] == 1, 0.4,
138                               ifelse(A[t-1] == 0 & W3[t-1] == 1, 0.75,
139                                      ifelse(A[t-1] == 1 & W3[t-1] == 0, 0.1361,
140                                             0.1361 + t / (6*t.end)))))
141
142 ## add time-varying exposure
143 D <- D + node("A", t=1:(t.end-1), distr = "rbern",
144                 prob = ifelse(A[t-1] == 1, 1,
145                               ifelse(Y[t] <= -3.2 & W3[t] == 1, 0.8,
146                                      ifelse(Y[t] <= -3.2 & W3[t] == 0, 0.6,
147                                             ifelse(Y[t] > -3.2 & W3[t] == 1,
148                                                    0.1, 0.02)))))
149
150 D <- set.DAG(D)
151
152 ## we can now simulate from the structural equation model
153 ## we developed: for example,
154
155 Odat <- sim(D, n=5000, rndseed = 123)
156
157 ## this generates 5,000 observations from our model
```

Appendix B

Simulation Results

This appendix presents the remaining results tables that were not included in the main text so as to keep it brief. In particular, the scenarios not shown in the text are: small effect size for $n = 2000$ (symmetric and skewed outcome data), and all effect sizes (null, small, and large) for $n = 1000$ and $n = 5000$ (for each setting, using either symmetric or skewed outcome data). One will see in the tables that the g-formula does not have associated coverage values—we did not compute bootstrap confidence intervals for these simulation scenarios as they are extremely computationally intensive. Our hope is that the coverage patterns for the scenarios shown in the main text apply to the settings shown here as well.

Table B.1 Simulation results: $n = 2000$, symmetric outcome, small treatment effect

(a) Correct Model Specification

	Bias	SE	Coverage (%)
G-formula	0.001	0.070	—
Linear MSM	0.127	0.159	80.0
Quadratic MSM	0.079	0.169	100.0
TMLE	0.001	0.093	93.4

(b) Treatment Model Misspecified

	Bias	SE	Coverage (%)
Linear MSM	0.112	0.107	93.0
Quadratic MSM	0.155	0.153	100.0
TMLE	0.005	0.089	92.9

(c) Outcome Model Misspecified

	Bias	SE	Coverage (%)
G-formula	0.999	0.031	—
TMLE	0.005	0.096	93.3

(d) Outcome and Treatment Model Misspecified

	Bias	SE	Coverage (%)
TMLE	0.112	0.089	72.4

(e) Models Controlling for All Past Variables

	Bias	SE	Coverage (%)
G-formula	0.002	0.089	—
TMLE	0.001	0.091	92.8

Table B.2 Simulation results: $n = 2000$, skewed outcome, small treatment effect

(a) Correct Model Specification

	Bias	SE	Coverage (%)
G-formula	-0.001	0.067	—
Linear MSM	0.087	0.197	76.8
Quadratic MSM	0.042	0.196	100.0
TMLE	-0.001	0.098	93.9

(b) Treatment Model Misspecified

	Bias	SE	Coverage (%)
Linear MSM	0.181	0.098	60.4
Quadratic MSM	0.335	0.181	100.0
TMLE	0.003	0.088	94.8

(c) Outcome Model Misspecified

	Bias	SE	Coverage (%)
G-formula	0.999	0.025	—
TMLE	0.003	0.101	94.7

(d) Outcome and Treatment Model Misspecified

	Bias	SE	Coverage (%)
TMLE	0.117	0.087	70.3

(e) Models Controlling for All Past Variables

	Bias	SE	Coverage (%)
G-formula	0.002	0.085	—
TMLE	0.000	0.089	94.6

Table B.3 Simulation results: $n = 1000$, symmetric outcome, null treatment effect

(a) Correct Model Specification

	Bias	SE	Coverage (%)
G-formula	0.002	0.090	—
Linear MSM	0.110	0.183	89.8
Quadratic MSM	0.097	0.189	100.0
TMLE	-0.000	0.126	95.1

(b) Treatment Model Misspecified

	Bias	SE	Coverage (%)
Linear MSM	0.073	0.254	97.0
Quadratic MSM	0.208	0.430	100.0
TMLE	-0.000	0.123	95.2

(c) Outcome Model Misspecified

	Bias	SE	Coverage (%)
G-formula	0.000	0.032	—
TMLE	0.006	0.127	94.2

(d) Outcome and Treatment Model Misspecified

	Bias	SE	Coverage (%)
TMLE	0.090	0.122	86.8

(e) Models Controlling for All Past Variables

	Bias	SE	Coverage (%)
G-formula	-0.003	0.116	—
TMLE	0.001	0.127	94.7

Table B.4 Simulation results: $n = 1000$, skewed outcome, null treatment effect

(a) Correct Model Specification

	Bias	SE	Coverage (%)
G-formula	-0.000	0.090	—
Linear MSM	0.113	0.180	89.4
Quadratic MSM	0.102	0.193	100.0
TMLE	0.007	0.124	95.0

(b) Treatment Model Misspecified

	Bias	SE	Coverage (%)
Linear MSM	0.060	0.225	96.8
Quadratic MSM	0.154	0.361	100.0
TMLE	0.006	0.123	94.8

(c) Outcome Model Misspecified

	Bias	SE	Coverage (%)
G-formula	-0.002	0.031	—
TMLE	0.013	0.125	94.6

(d) Outcome and Treatment Model Misspecified

	Bias	SE	Coverage (%)
TMLE	0.096	0.123	85.8

(e) Models Controlling for All Past Variables

	Bias	SE	Coverage (%)
G-formula	0.004	0.115	—
TMLE	0.005	0.129	94.2

Table B.5 Simulation results: $n = 1000$, symmetric outcome, small treatment effect

(a) Correct Model Specification

	Bias	SE	Coverage (%)
G-formula	0.000	0.091	—
Linear MSM	0.159	0.186	84.6
Quadratic MSM	0.118	0.201	100.0
TMLE	0.000	0.131	93.6

(b) Treatment Model Misspecified

	Bias	SE	Coverage (%)
Linear MSM	0.123	0.235	94.9
Quadratic MSM	0.220	0.329	100.0
TMLE	0.003	0.127	93.5

(c) Outcome Model Misspecified

	Bias	SE	Coverage (%)
G-formula	0.999	0.030	—
TMLE	0.008	0.132	93.6

(d) Outcome and Treatment Model Misspecified

	Bias	SE	Coverage (%)
TMLE	0.109	0.127	85.2

(e) Models Controlling for All Past Variables

	Bias	SE	Coverage (%)
G-formula	0.004	0.124	—
TMLE	-0.000	0.136	93.6

Table B.6 Simulation results: $n = 1000$, skewed outcome, small treatment effect

(a) Correct Model Specification

	Bias	SE	Coverage (%)
G-formula	0.001	0.091	—
Linear MSM	0.125	0.160	81.9
Quadratic MSM	0.079	0.185	100.0
TMLE	0.006	0.134	94.4

(b) Treatment Model Misspecified

	Bias	SE	Coverage (%)
Linear MSM	0.198	0.216	79.2
Quadratic MSM	0.383	0.418	100.0
TMLE	0.004	0.126	93.5

(c) Outcome Model Misspecified

	Bias	SE	Coverage (%)
G-formula	0.999	0.025	—
TMLE	0.013	0.136	94.5

(d) Outcome and Treatment Model Misspecified

	Bias	SE	Coverage (%)
TMLE	0.116	0.125	84.3

(e) Models Controlling for All Past Variables

	Bias	SE	Coverage (%)
G-formula	0.003	0.118	—
TMLE	-0.001	0.129	94.5

Table B.7 Simulation results: $n = 1000$, symmetric outcome, large treatment effect

(a) Correct Model Specification

	Bias	SE	Coverage (%)
G-formula	0.006	0.097	—
Linear MSM	0.318	0.338	54.7
Quadratic MSM	0.162	0.321	100.0
TMLE	0.006	0.141	94.5

(b) Treatment Model Misspecified

	Bias	SE	Coverage (%)
Linear MSM	0.381	0.263	63.4
Quadratic MSM	0.501	0.470	100.0
TMLE	0.028	0.128	93.7

(c) Outcome Model Misspecified

	Bias	SE	Coverage (%)
G-formula	5.003	0.045	—
TMLE	-0.025	0.162	94.1

(d) Outcome and Treatment Model Misspecified

	Bias	SE	Coverage (%)
TMLE	0.174	0.130	78.5

(e) Models Controlling for All Past Variables

	Bias	SE	Coverage (%)
G-formula	0.021	0.117	—
TMLE	0.006	0.133	94.4

Table B.8 Simulation results: $n = 1000$, skewed outcome, large treatment effect

(a) Correct Model Specification

	Bias	SE	Coverage (%)
G-formula	-0.001	0.101	—
Linear MSM	0.333	0.281	53.3
Quadratic MSM	0.174	0.286	100.0
TMLE	0.002	0.134	94.9

(b) Treatment Model Misspecified

	Bias	SE	Coverage (%)
Linear MSM	0.363	0.252	65.2
Quadratic MSM	0.459	0.405	100.0
TMLE	0.023	0.123	94.9

(c) Outcome Model Misspecified

	Bias	SE	Coverage (%)
G-formula	4.999	0.045	—
TMLE	-0.030	0.150	94.7

(d) Outcome and Treatment Model Misspecified

	Bias	SE	Coverage (%)
TMLE	0.170	0.127	78.7

(e) Models Controlling for All Past Variables

	Bias	SE	Coverage (%)
G-formula	0.012	0.118	—
TMLE	0.001	0.130	94.9

Table B.9 Simulation results: $n = 5000$, symmetric outcome, null treatment effect

(a) Correct Model Specification

	Bias	SE	Coverage (%)
G-formula	0.001	0.046	—
Linear MSM	0.046	0.192	86.6
Quadratic MSM	0.030	0.193	100.0
TMLE	0.001	0.058	96.0

(b) Treatment Model Misspecified

	Bias	SE	Coverage (%)
Linear MSM	0.063	0.054	95.4
Quadratic MSM	0.095	0.070	100.0
TMLE	0.001	0.050	94.8

(c) Outcome Model Misspecified

	Bias	SE	Coverage (%)
G-formula	-0.000	0.030	—
TMLE	0.004	0.060	95.6

(d) Outcome and Treatment Model Misspecified

	Bias	SE	Coverage (%)
TMLE	0.092	0.050	56.9

(e) Models Controlling for All Past Variables

	Bias	SE	Coverage (%)
G-formula	0.001	0.058	—
TMLE	0.001	0.051	95.1

Table B.10 Simulation results: $n = 5000$, skewed outcome, null treatment effect

(a) Correct Model Specification

	Bias	SE	Coverage (%)
G-formula	0.002	0.048	—
Linear MSM	0.046	0.164	86.6
Quadratic MSM	0.028	0.160	100.0
TMLE	-0.000	0.059	95.0

(b) Treatment Model Misspecified

	Bias	SE	Coverage (%)
Linear MSM	0.062	0.054	95.3
Quadratic MSM	0.084	0.074	100.0
TMLE	0.001	0.053	94.9

(c) Outcome Model Misspecified

	Bias	SE	Coverage (%)
G-formula	-0.000	0.030	—
TMLE	0.002	0.060	95.2

(d) Outcome and Treatment Model Misspecified

	Bias	SE	Coverage (%)
TMLE	0.092	0.053	57.6

(e) Models Controlling for All Past Variables

	Bias	SE	Coverage (%)
G-formula	0.000	0.060	—
TMLE	0.001	0.054	94.9

Table B.11 Simulation results: $n = 5000$, symmetric outcome, small treatment effect

(a) Correct Model Specification

	Bias	SE	Coverage (%)
G-formula	0.001	0.048	—
Linear MSM	0.079	0.177	72.6
Quadratic MSM	0.034	0.172	100.0
TMLE	0.001	0.062	95.7

(b) Treatment Model Misspecified

	Bias	SE	Coverage (%)
Linear MSM	0.113	0.055	77.4
Quadratic MSM	0.139	0.068	100.0
TMLE	0.004	0.051	95.8

(c) Outcome Model Misspecified

	Bias	SE	Coverage (%)
G-formula	1.000	0.031	—
TMLE	0.005	0.062	96.3

(d) Outcome and Treatment Model Misspecified

	Bias	SE	Coverage (%)
TMLE	0.111	0.051	43.6

(e) Models Controlling for All Past Variables

	Bias	SE	Coverage (%)
G-formula	0.002	0.058	—
TMLE	0.001	0.051	96.2

Table B.12 Simulation results: $n = 5000$, skewed outcome, small treatment effect

(a) Correct Model Specification

	Bias	SE	Coverage (%)
G-formula	0.001	0.044	—
Linear MSM	0.061	0.157	71.2
Quadratic MSM	0.013	0.155	100.0
TMLE	0.001	0.067	95.7

(b) Treatment Model Misspecified

	Bias	SE	Coverage (%)
Linear MSM	0.176	0.050	11.3
Quadratic MSM	0.317	0.088	100.0
TMLE	0.005	0.051	95.7

(c) Outcome Model Misspecified

	Bias	SE	Coverage (%)
G-formula	0.999	0.025	—
TMLE	0.003	0.071	96.2

(d) Outcome and Treatment Model Misspecified

	Bias	SE	Coverage (%)
TMLE	0.119	0.051	38.0

(e) Models Controlling for All Past Variables

	Bias	SE	Coverage (%)
G-formula	0.003	0.055	—
TMLE	0.002	0.051	95.7

Table B.13 Simulation results: $n = 5000$, symmetric outcome, large treatment effect

(a) Correct Model Specification

	Bias	SE	Coverage (%)
G-formula	0.002	0.051	—
Linear MSM	0.229	0.305	32.4
Quadratic MSM	0.072	0.283	100.0
TMLE	0.005	0.084	94.1

(b) Treatment Model Misspecified

	Bias	SE	Coverage (%)
Linear MSM	0.340	0.058	0.1
Quadratic MSM	0.352	0.086	100.0
TMLE	0.026	0.055	92.1

(c) Outcome Model Misspecified

	Bias	SE	Coverage (%)
G-formula	4.999	0.045	—
TMLE	-0.010	0.097	94.4

(d) Outcome and Treatment Model Misspecified

	Bias	SE	Coverage (%)
TMLE	0.174	0.055	15.9

(e) Models Controlling for All Past Variables

	Bias	SE	Coverage (%)
G-formula	0.015	0.063	—
TMLE	0.005	0.055	95.0

Table B.14 Simulation results: $n = 5000$, skewed outcome, large treatment effect

(a) Correct Model Specification

	Bias	SE	Coverage (%)
G-formula	0.002	0.052	—
Linear MSM	0.220	0.261	38.0
Quadratic MSM	0.062	0.256	100.0
TMLE	-0.001	0.072	95.3

(b) Treatment Model Misspecified

	Bias	SE	Coverage (%)
Linear MSM	0.339	0.056	0.3
Quadratic MSM	0.348	0.087	100.0
TMLE	0.022	0.053	92.8

(c) Outcome Model Misspecified

	Bias	SE	Coverage (%)
G-formula	4.999	0.045	—
TMLE	-0.011	0.094	94.2

(d) Outcome and Treatment Model Misspecified

	Bias	SE	Coverage (%)
TMLE	0.172	0.054	16.7

(e) Models Controlling for All Past Variables

	Bias	SE	Coverage (%)
G-formula	0.010	0.061	—
TMLE	0.001	0.053	95.7

References

Bickel, Peter J, Klaassen, Chris AJ, Ritov, Yaacov, & Wellner, Jon A. 1998. *Efficient and adaptive estimation for semiparametric models*. Springer-Verlag.

Burton, Andrea, Altman, Douglas G, Royston, Patrick, & Holder, Roger L. 2006. The design of simulation studies in medical statistics. *Statistics in Medicine*, **25**(24), 4279–4292.

Dawid, A Philip. 2000. Causal inference without counterfactuals. *Journal of the American Statistical Association*, **95**(450), 407–424.

Decker, Anna L, Hubbard, Alan, Crespi, Catherine M, Seto, Edmund YW, & Wang, May C. 2014. Semiparametric estimation of the impacts of longitudinal interventions on adolescent obesity using targeted maximum-likelihood: Accessible estimation with the LTMLE package. *Journal of Causal Inference*, **2**(1), 95–108.

Frangakis, Constantine E, & Rubin, Donald B. 2002. Principal stratification in causal inference. *Biometrics*, **58**(1), 21–29.

Gill, Richard D, & Robins, James M. 2001. Causal inference for complex longitudinal data: the continuous case. *The Annals of Statistics*, **29**(6), 1785–1811.

Halekoh, Ulrich, Højsgaard, Søren, & Yan, Jun. 2006. The R package geepack for generalized estimating equations. *Journal of Statistical Software*, **15**(2), 1–11.

Hernán, Miguel A, & Hernández-Díaz, Sonia. 2012. Beyond the intention-to-treat in comparative effectiveness research. *Clinical Trials*, **9**(1), 48–55.

Hernan, Miguel A, & Robins, James M. 2010. *Causal inference*. CRC Boca Raton, FL:.

Hernán, Miguel A, Brumback, Babette A, & Robins, James M. 2002. Estimating the causal effect of zidovudine on CD4 count with a marginal structural model for repeated measures. *Statistics in Medicine*, **21**(12), 1689–1709.

Holland, Paul W. 1986. Statistics and causal inference. *Journal of the American Statistical Association*, **81**(396), 945–960.

Imbens, Guido W, & Rubin, Donald B. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Kang, Joseph DY, & Schafer, Joseph L. 2007. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 523–539.

Kramer, Michael S, Chalmers, Beverley, Hodnett, Ellen D, Sevkovskaya, Zinaida, Dzikovich, Irina, Shapiro, Stanley, Collet, Jean-Paul, Vanilovich, Irina, Mezen, Irina, Ducruet, Thierry, *et al.* . 2001. Promotion of Breastfeeding Intervention Trial (PROBIT): a randomized trial in the Republic of Belarus. *JAMA*, **285**(4), 413–420.

Kramer, Michael S, Guo, Tong, Platt, Robert W, Shapiro, Stanley, Collet, Jean-Paul, Chalmers, Beverley, Hodnett, Ellen, Sevkovskaya, Zinaida, Dzikovich, Irina, Vanilovich, Irina, *et al.* . 2002. Breastfeeding and infant growth: biology or bias? *Pediatrics*, **110**(2), 343–347.

Lefebvre, Geneviève, Delaney, Joseph AC, & Platt, Robert W. 2008. Impact of misspecification of the treatment model on estimates from a marginal structural model. *Statistics in Medicine*, **27**(18), 3629–3642.

Liang, Kung-Yee, & Zeger, Scott L. 1986. Longitudinal data analysis using generalized linear models. *Biometrika*, **73**(1), 13–22.

McCullagh, Peter, & Nelder, James A. 1989. *Generalized Linear Models, no. 37 in Monograph on Statistics and Applied Probability*.

Neugebauer, Romain, Schmittdiel, Julie A, & Laan, Mark J. 2014. Targeted learning in real-world comparative effectiveness research with time-varying interventions. *Statistics in Medicine*, **33**(14), 2480–2520.

Neyman, Jerzy. 1923. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, **5**(4), 465–472. Translated from 1923 original Polish article and edited by Dabrowska DM and Speed TP, 1990.

Pang, Menglan, Schuster, Tibor, Filion, Kristian B, Schnitzer, Mireille E, Eberg, Maria, & Platt, Robert W. 2016a. Effect Estimation in point-exposure studies with binary outcomes and high-dimensional covariate data–a comparison of targeted maximum likelihood estimation and inverse probability of treatment weighting. *The International Journal of Biostatistics*, **12**(2).

Pang, Menglan, Schuster, Tibor, Filion, Kristian B, Eberg, Maria, & Platt, Robert W. 2016b. Targeted maximum likelihood estimation for pharmacoepidemiologic research. *Epidemiology*, **27**(4), 570.

Pearl, Judea. 2009. *Causality*. Cambridge University Press.

Petersen, Maya, Schwab, Joshua, Gruber, Susan, Blaser, Nello, Schomaker, Michael, & van der Laan, Mark. 2014. Targeted maximum likelihood estimation for dynamic and static longitudinal marginal structural working models. *Journal of Causal Inference*, **2**(2), 147–185.

Petersen, Maya L, Wang, Yue, van der Laan, Mark J, Guzman, David, Riley, Elise, & Bangsberg, David R. 2007. Pillbox organizers are associated with improved adherence to HIV antiretroviral therapy and viral suppression: a marginal structural model analysis. *Clinical Infectious Diseases*, **45**(7), 908–915.

Platt, Robert W, Schisterman, Enrique F, & Cole, Stephen R. 2009. Time-modified confounding. *American Journal of Epidemiology*, **170**(6), 687–694.

Porter, Kristin E, Gruber, Susan, Van Der Laan, Mark J, & Sekhon, Jasjeet S. 2011. The relative performance of targeted maximum likelihood estimators. *The International Journal of Biostatistics*, **7**(1), 1–34.

Richardson, Thomas S, & Robins, James M. 2013. Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, **128**(30).

Robins, J, Hernan, M, & Siebert, U. 2004. Effects of multiple interventions. *In:* Ezzati, M, Lopez, AD, Rodgers, A, & Murray, CJ (eds), *Comparative Quantification of Health Risks*, vol. 2. Geneva: World Health Organization.

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

Robins, James M. 1997. Marginal structural models. *Pages 1–10 of: Proceedings of the Section on Bayesian Statistical Science*. Alexandria: American Statistical Association.

Robins, James M. 1999. Association, causation, and marginal structural models. *Synthese*, **121**(1), 151–179.

Robins, James M, & Wasserman, Larry. 1997. Estimation of effects of sequential treatments by reparameterizing directed acyclic graphs. *In: Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc.

Robins, James M, Hernan, Miguel A, & Brumback, Babette. 2000. Marginal structural models and causal inference in epidemiology. *Epidemiology*, **11**(5), 550–560.

Rose, Sherri, & van der Laan, Mark J. 2011. *Targeted learning: Causal inference for observational and experimental data*. New York: Springer.

Rosenbaum, Paul R, & Rubin, Donald B. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**(1), 41–55.

Rosenblum, Michael, & van der Laan, Mark J. 2010. Targeted maximum likelihood estimation of the parameter of a marginal structural model. *The International Journal of Biostatistics*, **6**(2).

Rotnitzky, Andrea, Robins, James M, & Scharfstein, Daniel O. 1998. Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association*, **93**(444), 1321–1339.

Rubin, Donald B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, **66**(5), 688–701.

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

Schnitzer, Mireille E, Moodie, Erica EM, & Platt, Robert W. 2013. Targeted maximum likelihood estimation for marginal time-dependent treatment effects under density misspecification. *Biostatistics*, **14**(1), 1–14.

Schwab, Joshua, Lendle, Samuel, Petersen, Maya, & van der Laan, Mark. 2016. *ltmle: Longitudinal Targeted Maximum Likelihood Estimation*. R package version 0.9-9.

Sekhon, Jasjeet S. 2008. The Neyman-Rubin model of causal inference and estimation via matching methods. *Pages 271–299 of:* Box-Steffensmeier, Janet, Brady, Henry, & Collier, David (eds), *The Oxford Handbook of Political Methodology*. New York: Oxford University Press.

Setoguchi, Soko, Schneeweiss, Sebastian, Brookhart, M Alan, Glynn, Robert J, & Cook, E Francis. 2008. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and Drug Safety*, **17**(6), 546–555.

Snowden, Jonathan, Rose, Sherri, & Mortimer, Kathleen M. 2011. Implementation of G-computation on a simulated data set: demonstration of a causal inference technique. *American Journal of Epidemiology*, **173**(7), 731–738.

Sofrygin, Oleg, van der Laan, Mark J, & Neugebauer, Romain. 2015. simcausal: Simulating Longitudinal Data with Causal Inference Applications. *R package version 0.4, URL http://CRAN.R-project.org/package=simcausal*.

Spertus, Jacob V, T Normand, Sharon-Lise, Wolf, Robert, Cioffi, Matt, Lovett, Ann, & Rose, Sherri. 2016. Assessing Hospital Performance After Percutaneous Coronary Intervention Using Big Data. *Circulation: Cardiovascular Quality and Outcomes*, **9**(6), 659–669.

Tager, Ira B, Haight, Thaddeus, Sternfeld, Barbara, Yu, Zhuo, & van der Laan, Mark. 2004. Effects of physical activity and body composition on functional limitation in the elderly: application of the marginal structural model. *Epidemiology*, **15**(4), 479–493.

Talbot, Denis, Atherton, Juli, Rossi, Amanda M, Bacon, Simon L, & Lefebvre, Geneviève. 2015. A cautionary note concerning the use of stabilized weights in marginal structural models. *Statistics in Medicine*, **34**(5), 812–823.

Taubman, Sarah L, Robins, James M, Mittleman, Murray A, & Hernán, Miguel A. 2009. Intervening on risk factors for coronary heart disease: an application of the parametric g-formula. *International Journal of Epidemiology*, **38**(6), 1599–1611.

Tsiatis, Anastasios. 2007. *Semiparametric theory and missing data*. Springer Science & Business Media.

van der Laan, Mark J. 2010. Targeted maximum likelihood based causal inference: Part I. *The International Journal of Biostatistics*, **6**(2), 1–45.

van der Laan, Mark J, & Gruber, Susan. 2010. Collaborative double robust targeted maximum likelihood estimation. *The International Journal of Biostatistics*, **6**(1), 1–71.

van der Laan, Mark J, & Rubin, Daniel. 2006. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, **2**(1), 1–40.

van der Laan, Mark J, Polley, Eric C, & Hubbard, Alan E. 2007. Super Learner. *Statistical Applications in Genetics and Molecular Biology*, **6**(1), 1–23.

van der Wal, WM, Prins, M, Lumbreras, B, & Geskus, RB. 2009. A simple G-computation algorithm to quantify the causal effect of a secondary illness on the progression of a chronic disease. *Statistics in Medicine*, **28**(18), 2325–2337.

VanderWeele, Tyler J, & Hernan, Miguel A. 2013. Causal inference under multiple versions of treatment. *Journal of causal inference*, **1**(1), 1–20.

Wang, Hui, Rose, Sherri, & van der Laan, Mark J. 2011. Finding quantitative trait loci genes with collaborative targeted maximum likelihood learning. *Statistics & Probability Letters*, **81**(7), 792–796.

Westreich, Daniel, Cole, Stephen R, Young, Jessica G, Palella, Frank, Tien, Phyllis C, Kingsley, Lawrence, Gange, Stephen J, & Hernán, Miguel A. 2012a. The parametric g-formula to estimate the effect of highly active antiretroviral therapy on incident AIDS or death. *Statistics in Medicine*, **31**(18), 2000–2009.

Westreich, Daniel, Cole, Stephen R, Schisterman, Enrique F, & Platt, Robert W. 2012b. A simulation study of finite-sample properties of marginal structural Cox proportional hazards models. *Statistics in medicine*, **31**(19), 2098–2109.

Young, Jessica G, Cain, Lauren E, Robins, James M, OReilly, Eilis J, & Hernán, Miguel A. 2011. Comparative effectiveness of dynamic treatment regimes: an application of the parametric g-formula. *Statistics in Biosciences*, **3**(1), 119.

Zeger, Scott L, & Liang, Kung-Yee. 1986. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, **42**(1), 121–130.