Targeted Maximum Likelihood Estimation for Longitudinal Data

Mireille Elisa Schnitzer

Doctor of Philosophy

Department of Epidemiology, Biostatistics and Occupational Health

McGill University

Montréal, Québec

2012-08-15

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

DEDICATION

This thesis is dedicated to my parents, Sue Blais and Lonnie Schnitzer.

ACKNOWLEDGEMENTS

I am very pleased to be able to thank the many people and institutions that have assisted me in producing this dissertation.

I am sincerely grateful for my supervisor Dr. Erica Moodie for her consistent support, advice (both personal and career-oriented!), and insight throughout these three years, in particular during my time at U.C. Berkeley. I greatly appreciate her dedication to her students and how she always made herself very available, even at challenging times. Her encouragement to pursue greater ambitions and her confidence in me has given me the strength to continue on what I find to be a very demanding path. To my co-supervisor Dr. Robert Platt, I am deeply thankful for his dedication, patience and good ideas. His supportive attitude and openness were very much appreciated, especially when I was still integrating into the department and learning causal inference for the first time. I would also like to express great thanks to both of my supervisors for financially supporting my extensive travel schedule during the past two years.

This dissertation would not have been possible without the collaboration and teachings of Dr. Mark van der Laan, who generously invited me to U.C. Berkeley and donated much of his time to advance my studies in semiparametric efficiency theory and Targeted Maximum Likelihood Estimation. I would also like to express appreciation to my friends at U.C. Berkeley for accepting me into their TMLE brotherhood.

I am continuously grateful for my mentor, Dr. Russell Steele, for placing me

on my career path and planting notions of greatness inside my head. I deeply value both his guidance and confidence in me.

I would like to express sincere thanks to the faculty, post-doctoral scholars and fellow students in the Department of Epidemiology, Biostatistics and Occupational Health at McGill University for fostering a welcoming and engaging atmosphere for study. I would also like to thank our wonderful department administrators who help create a warm and friendly environment in our workplace.

The data that we analyzed in this thesis was kindly made available by Dr. Michael Kramer and the investigators of the Promotion of Breastfeeding Intervention Trial, and by Dr. Marina Klein and the investigators of the Canadian Co-Infection Cohort. I would like to extend my thanks to Dr. Kramer and Dr. Klein for also sharing their clinical expertise.

I am happily indebted to Dr. Anne-Sophie Charest, who translated the thesis abstract, and Ms. Laura De Benedetti, who copy-edited several chapters of the thesis.

I acknowledge the usage of Consortium Laval, Université du Québec, McGill and Eastern Quebec computational facilities, an invaluable resource that enabled the computation in all three studies.

This work would not have been possible without the financial support of the Natural Sciences and Engineering Research Council of Canada (NSERC). Support for travel to Berkeley, CA to further my education in the primary matter of this thesis was provided by the Fonds de recherche du Québec – Nature et technologies (FQRNT) foreign travel award with help from the Centre de recherches mathématiques

(CRM), and by the McGill University Faculty of Medicine Graduate Program for International Travel.

Finally, I thank my family and friends for always providing support. My grandparents, Moomoo and Grandpa Schnitzer for so much love and confidence. And my great-uncle Dr. Morris Schnitzer for his well-received encouragement at a key time.

ABSTRACT

Semiparametric efficient methods in causal inference have been developed to robustly and efficiently estimate causal parameters. As in general causal estimation, the methods rely on a set of mathematical assumptions that translate into requirements of causal knowledge and confounder identification. Targeted maximum likelihood estimation (TMLE) [78] methodology has been developed as a potential improvement on efficient estimating equations, in that it shares the qualities of double robustness (unbiasedness under partial misspecification) and semiparametric efficiency, but can be constructed to provide boundedness of parameter estimates, robustness to data sparsity, and a unique estimate.

This thesis, composed primarily of three manuscripts, presents new research on the analysis of longitudinal and survival data with time-dependent confounders using TMLE. The first manuscript describes the construction of a two time-point TMLE using a generalized exponential distribution family member as the loss function for the outcome model. It demonstrates the robustness of the continuous version of this TMLE algorithm in a simulation study, and uses a modified version of the method in a simplified analysis of the PROmotion of Breastfeeding Intervention Trial (PROBIT) [27] where evidence for a protective causal effect of breastfeeding on gastrointestinal infection is obtained.

The second manuscript presents a description of several substitution estimators for longitudinal data, a specialized implementation of a longitudinal TMLE method, and a case study using the full PROBIT dataset. The K time point sequential

TMLE algorithm employed (theory developed in [74]), implemented nonparametrically using Super Learner [36], differs fundamentally from the strategy used in the first manuscript, and offers some benefits in computation and ease of implementation. The analysis compares different durations of breastfeeding and the related exposure-specific (and censoring-free) mean counts of gastrointestinal infections over the first year of an infant's life and concludes that a protective effect is present. Simulated data mirroring the PROBIT dataset was generated, and the performance of TMLE was again assessed.

The third manuscript develops a methodology to estimate marginal structural models for survival data. Utilizing the sequential longitudinal TMLE algorithm [74] to estimate the exposure-specific survival curves for all exposure patterns, it demonstrates a way to combine inference in order to model the outcome using a linear specification. This article presents the theoretical construction of two different types of marginal structural models (modeling the log-odds survival and the hazard) and presents a simulation study demonstrating the unbiasedness of the technique. It then describes an analysis of the Canadian Co-infection Cohort study [26] undertaken with one of the TMLE methods to fit survival curves and a model for the hazard function of development of end-stage liver disease (ESLD) conditional on time and clearance of the Hepatitis C virus.

ABRÉGÉ

Des méthodes d'analyse causale semi-paramétriques et efficaces ont été développées pour estimer les paramètres causaux efficacement et de façon robuste. Comme c'est le cas en général pour l'estimation causale, ces méthodes se basent sur un ensemble d'hypothèses mathématiques qui impliquent que la structure causale et les facteurs de confusion doivent être connus. La méthode d'estimation par le maximum de vraisemblance ciblé (TMLE) [78] se veut une amélioration des équations d'estimation efficaces: elle a les propriétés de double robustesse (sans biais même avec une erreur de spécification partielle) et d'efficacité semi-paramétrique, mais peut également garantir des estimés finis pour les paramètres et la production d'un seul estimé en plus d'être robuste si les données sont éparses.

Cette thèse, composée essentiellement de trois manuscrits, présente de nouvelles recherches sur l'analyse avec le TMLE de données longitudinales et de données de survie avec des facteurs de confusion variant dans le temps. Le premier manuscrit décrit la construction d'un TMLE à deux points dans le temps avec une distribution de la famille exponentielle généralisée comme fonction de perte du modèle de la réponse. Il démontre à l'aide d'une étude de simulation la robustesse de la version continue de cet algorithme TMLE, et utilise une version Poisson de la méthode pour une analyse simplifiée de l'étude PROmotion of Breastfeeding Intervention Trial (PROBIT) qui donne des signes d'un effet causal protecteur de l'allaitement sur les infections gastrointestinales.

Le deuxième manuscrit présente une description de plusieurs estimateurs de substitution pour données longitudinales, une implémentation spéciale de la méthode TMLE longitudinale et une étude de cas du jeu de données PROBIT entier. Un algorithme TMLE séquentiel à K points dans le temps est utilisé (théorie développée dans [74]), lequel est implémenté de façon non-paramétrique avec le Super Learner [36]. Cet algorithme diffère fondamentalement de la stratégie utilisée dans le premier manuscrit et offre des avantages en terme de calcul et de facilité d'implémentation. L'analyse compare les moyennes de dénombrements du nombre d'infections gastro-intestinales dans la première année de vie d'un nouveau-né par durée d'allaitement et avec aucune censure, et conclut à la présence d'un effet protecteur. Des données simulées semblables au jeu de données PROBIT sont également générées, et la performance du TMLE de nouveau étudiée.

Le troisième manuscrit développe une méthodologie pour estimer des modèles structurels marginaux pour données de survie. En utilisant l'algorithme séquentiel du TMLE longitudinal pour estimer des courbes de survie spécifiques à l'exposition pour tous les patrons d'exposition, il montre une façon de combiner les inférences pour modéliser la réponse à l'aide d'une spécification linéaire. Cet article présente la construction théorique de deux différents types de modèles structurels marginaux (modélisant le log du rapport des chances de survie et le risque) et présente une étude de simulation démontrant l'absence de biais de la technique. Il décrit ensuite une analyse de l'Étude de la Cohorte Canadienne de Co-Infection [26] à l'aide d'une des méthodes TMLE pour ajuster des courbes de survie et un modèle pour la fonction de

risque du développement de la maladie chronique du foie (ESLD) conditionnellement au temps et à l'élimination du virus de l'hépatite C.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AKME: Adjusted Kaplan-Meier Estimator

ARV: Antiretroviral therapy

BR: Bang and Robins' estimator

CCC: Canadian Co-infection Cohort

CI: Confidence interval

COV: Percent coverage

Cover: Percent coverage

EE: Estimating equation

ESLD: End-stage liver disease

G-COMP: G-Computation

HCV: Hepatitis C virus

HIV: Human immunodeficiency virus

IPTW: Inverse probability of treatment weighting

MAR: Missing at random

MSE: Mean-squared error

MSM: Marginal structural model

NNT: Number needed to treat

PROBIT: PROmotion of Breastfeeding Intervention Trial

RAL: Regular and asymptotically linear

rMSE: Root mean-squared error

SE: Standard error

TMLE: Targeted maximum likelihood estimation/estimator

PREFACE: CONTRIBUTIONS OF AUTHORS

This thesis contains new research in addition to an introductory overview of certain topics in causal inference.

Chapters 2–3 contain an original summary of selected topics in causal inference. These chapters were written entirely by Mireille Schnitzer (MS) and edited and corrected by Erica Moodie (EM) and Robert Platt (RP).

Chapters 4–6 of this thesis are comprised of original research carried out by the authors specified on the cover page for each chapter. A description of the unique contribution of each chapter is included in its preamble.

The idea for the study in Chapter 4 was conceived by MS, EM and RP. All methodological work, writing, programming, and analysis was carried out by MS with EM and RP as advisors, editors (also contributing small amounts of content directly) and troubleshooters. RP provided guidance on the use of the PROBIT dataset and contextual interpretations of the results.

Chapter 5 was conceptualized by MS, who carried out all writing, programming, and analysis. Support for the theoretical write-up and implementation was given by Mark van der Laan (MvdL). MvdL, EM and RP all served as advisors and editors. RP guided the use of the PROBIT dataset and acted as liaison between the authors and the PROBIT study director.

The methodological topic of Chapter 6 was chosen by MS, who was guided through the theory and general methodology by MvdL. MS worked out the mathematics required for the implementation and performed all necessary programming and analysis. EM, MvdL, and RP advised MS for the application involving the CCC data and guided the choice of modeling methods. Marina Klein contributed the data and EM facilitated and offered guidance related to the usage of the dataset. All authors contributed as editors.

Chapter 7, which concludes the thesis, was written by MS and edited by EM and RP.

PREFACE: ETHICS APPROVAL

The manuscripts in this thesis include analyses of previously collected data from human subjects. Ethics approval for the collection of the data was obtained by the original studies. The first two manuscripts use data from the PROmotion of Breastfeeding Intervention Trial (PROBIT) [27], and the third uses data from the Canadian Co-Infection Cohort (CCC) [26]. The PROBIT study received ethical approval from McGill University Health Center Research Ethics Board, the Human Subjects Committees at Harvard Pilgrim Health Care, and the Avon Longitudinal Study of Parents and Children Law and Ethics Committee. The CCC study has been approved by all the research ethics boards of the participating institutions and the Community Advisory Committee (CAC) of the CIHR Canadian HIV Trials Network (CTN).

CHAPTER 1
Introduction

Causal inference, the statistical analysis of cause and effect, seeks to pinpoint the true effect of a treatment or exposure (depending on which characterizes the source of the effect of interest) on an outcome. In a longitudinal setting, where both treatment and outcomes are observed over a period of time, the interest lies in determining the effect of a *treatment regimen* (a sequence of point-treatments over time) on a final outcome. Such an effect can be extracted by correctly implementing a randomized trial where different groups, made approximately comparable by randomization, are each blindly issued different treatment regimens. Under ideal circumstances, including full compliance with the assigned treatment, no dropout, the absence of statistical uncertainty, and successful blinding of both participants and outcome assessors where necessary, the difference in outcome between these groups can be attributed solely to the effect of the treatment regimen [54]. Unfortunately, a randomized trial is not always feasible. For example, ethical concerns may arise from lack of clinical equipoise [14], or blinding may be impossible due to an invasive treatment being tested such as a new surgical technique [29]. In such cases, it might be that only non-experimental data can be made available to the researcher.

When it is impossible to randomize a treatment (as in observational data), and the resulting data are used in a statistical analysis, confounding of the treatment

effect may be present. *Confounding* can be intuitively described as the incomparability between two exposure groups in terms of a given outcome, meaning that when confounding is present, the true effect of the treatment in question cannot be derived simply from the difference in outcome between the group that underwent treatment, and the group that did not [30]. Confounding by a baseline variable may arise when there is a patient characteristic that affects both the probability of being treated and the outcome. As a basic example of confounding, pulmonary abnormalities are common in patients with Scleroderma (SSc or Systemic Sclerosis) [68] and immunosuppressant drugs are often prescribed to improve lung function (among other things) [60]. A simple comparison of the measure of lung function between scleroderma patients in the Canadian population who are taking immunosuppressants versus those who are not would initially seem to indicate that the drugs are harmful to lung function. Upon consideration, however, we note that this conclusion may not be warranted as those patients who are most affected by their disease are more likely to have been prescribed immunosuppressants and also have reduced lung function. That is, the comparison is primarily between two very different groups of patients: the desired effect measurement is confounded by disease progression and severity.

In longitudinal data, an additional challenge may arise due to the presence of time-dependent confounders, which affect both later exposure and outcome. When a time-dependent confounder is also affected by previous treatment, traditional methods of controlling for confounding using regression models (such as mixed models) are biased whether or not one chooses to control for the time-dependent confounders

[43, 46]. Because of this, several methods have been produced to properly analyze longitudinal data when time-dependent confounders are thought to be present. Two of these methods are inverse probability of treatment weighting (IPTW) [8], which requires a model for the probability of treatment, and G-computation [44], which models the outcome and time-dependent variables.

Both IPTW and G-computation produce unbiased estimates of a causal parameter under standard causal assumptions (see Section 2.1.1). However, they both also assume that the required models are correctly specified (i.e. that the form chosen for the model captures the true causal structure). Van der Laan and Robins [76] used semiparametric efficiency theory to identify causal estimators that improved upon IPTW and G-computation. Their theoretical framework led to the construction of regular, asymptotically linear estimators that have optimally low asymptotic variance in their class when correctly specified. Some of these estimators only require correct specification of either the treatment (and censoring) model or the model for outcome and time-varying covariates in order to obtain asymptotically unbiased inference. This property is called double robustness [25]. Examples of double robust, semiparametric efficient estimators are augmented IPTW [45] for cross-sectional data, and the method proposed by Bang and Robins for longitudinal data [1].

Targeted Maximum Likelihood Estimation (TMLE) [78] is a framework for building doubly robust, semiparametric efficient estimators that have additional benefits over previously developed semiparametric efficient estimators. In particular, TMLE offers additional flexibility in constructing the estimator by allowing for a choice of loss function to minimize, it naturally produces bounded estimation

for bounded parameters, it can be used to construct estimators for a wider range of causal parameters, and it does not have the problem of multiple solutions that exists for other types of semiparametric efficient methods. Finally, TMLE allows for nonparametric estimation of the required density components (for example, the probability of treatment, and the models for outcome and time-dependent variables), removing the requirement of choosing a correct parametric model.

In this thesis, we focus on the use of TMLE for estimating different measures of causal effect of time-dependent outcomes in longitudinal settings. In Chapter 2, we provide a critical review of the literature on causal inference for time-dependent outcomes, and Chapter 3 enumerates our research objectives. In Chapter 4, we demonstrated how a two time point TMLE can be constructed with a generalized exponential family member loss function, and apply the method to an extensive simulation study and a case study. In Chapter 5, we describe the theory and specialized implementation of an alternative longitudinal TMLE method in the context of a case study in Paediatrics. Chapter 6 describes a method of fitting double robust marginal structural models for survival data using TMLE, and the application of the method in a data analysis involving an HIV/Hepetitus C co-infection cohort. Chapters 4, 5, and 6 were originally written as stand-alone papers. Consequently there is some overlap and inconsistency in notation (sometimes due to journal formatting) between these chapters. Chapter 4 will appear in *Biostatistics*. Chapter 5 has been submitted for publication, and is under review. Chapter 6 will be submitted to a statistical journal shortly after the submission of the thesis. In Chapter 7, we conclude with an overview of the three manuscripts.

CHAPTER 2
Literature Review

The Literature Review is comprised of three sections. The first is a description of the Rubin-Neyman counterfactual model of causal inference, and the causal assumptions and identifiability results required for the estimation of a large class of causal parameters with a particular focus on longitudinal parameters. The second section involves inferential methods for the estimation of causal parameters defined for longitudinal and survival data. In the third section, semiparametric efficiency theory is summarized, and a description of efficient estimating equations is given. This is followed by an introduction to TMLE, the general construction and previous work on estimation in longitudinal and survival contexts.

2.1 Causal inference and the counterfactual model

The Rubin-Neyman counterfactual model [54] can be used to describe causal effects and confounding in a fairly intuitive and direct way. For an individual subject indexed by i and potentially exposed to two different treatments, let outcome under treatments $a = 0, 1$ be denoted Y_i^0 and Y_i^1, respectively. The individual-level causal effect is the difference between one person's outcome under treatment and non-treatment in a specific context and time-frame. It can be denoted in this counterfactual framework as $Y_i^0 - Y_i^1$. This effect can almost never be directly observed because one person can only be treated or not treated at a given time or under identical circumstances, and therefore we can only observe one result. The

marginal population-level difference is defined as the difference in mean outcome of a given population when then entire population has received treatment versus when they have not. If subject i was drawn from population D, then the causal difference between treatments $a = 0$ and 1 on (random) outcome Y can be defined as $E_D(Y^0 - Y^1) = E_D(Y^0) - E_D(Y^1)$. Just as in the individual case, we can never observe both outcomes of the same population under different treatment statuses but otherwise the exact same conditions.

A randomized trial is the most natural way to make causal comparisons, as random treatment assignment in a given population provides easy access to an unbiased estimate of the population-level causal effect of treatment. Ideally, a simple randomized trial would take a random sample of participants from the population, D, of interest, and then randomly allocate levels of treatment to the participants. Let D_0 be the group of size n_0 that received treatment $A = 0$, and D_1 be the group of size n_1 that received $A = 1$. Then, an unbiased estimate of the population-level difference is $1/n_0 \sum_{i \in D_0} Y_i - 1/n_1 \sum_{i \in D_1} Y_i$.

When randomized trials are unavailable, infeasible, or suffer from non-compliance, the goal is then to derive unbiased and efficient causal estimates from observational or other data where the treatment of interest is unrandomized. Since the two levels of exposure cannot both be applied and evaluated on the same population at the same time, we may alternatively use information on surrogate populations who have been observed under the different levels of the exposure of interest. Suppose we observe group F of size n_F with exposure $A = 0$ and group G of size n_G with exposure $A = 1$. An estimate of the causal difference of population D might then be

$1/n_F \sum_{i \in F} Y_i - 1/n_G \sum_{i \in G} Y_i$ which unbiasedly estimates $E_F(Y^0) - E_G(Y^1)$. *Confounding bias* is present when $E_D(Y^0) - E_D(Y^1) \neq E_F(Y^0) - E_G(Y^1)$, meaning that the populations used for inference do not represent the population of interest [30, 16].

In an unrandomized study, the legitimacy of this type of simple inference is very often contestable due to comparisons between incompatible populations. Modeling methods are often required to correct for relevant differences between populations. Ultimately, however, extracting a fair causal comparison relies on extensive knowledge of the data application in question, and a collection of assumptions.

2.1.1 Causal assumptions and identifiability of the causal effect

For a causal effect to be identifiable, several data-generating assumptions are required. For the methods presented in this thesis, the necessary assumptions are i) no interference between subjects, ii) consistency, iii) positivity (also called the experimental treatment assumption), and iv) unconfounded treatment assignment (also called no unmeasured confounders or sequential randomization).

Assumptions i) and ii) are together called the Stable Unit Treatment Value Assumption (SUTVA) by Rubin, 1980 [57]. The absence of interference means that one subject's potential outcome is independent of the other subjects' exposures. Consistency, as refined by VanderWeele, 2009 [81], is the assumption that the observed outcome of a subject who experienced a given exposure is the same as the potential outcome under the exposure. Essentially, for observed exposure $A_i = a$, $Y_i^a = Y_i^{obs}$ where Y_i^{obs} is the actual realization of subject i. The consistency assumption therefore incorporates the requirement that there is only one version of treatment a. Together,

these assumptions imply that subject i exposed according to a will obtain outcome Y_i^a, independent of other subjects' exposures.

Assumption iii) requires that every unit has a non-zero probability of being exposed at every treatment level. For treatment at a single time point, if A is the variable indicating treatment received, and X is a set of variables occurring prior to treatment, the theoretical positivity assumption states that $Pr(A = a \mid X) > 0$ for any treatment a and all possible realizations of X. As an example of a contradiction to theoretical positivity, doctors may determine that patients can be too sick to receive an experimental surgery. In this case, there is a subpopulation of patients that could never have received surgery (the intervention of interest). For this subpopulation (of very sick patients), the expected outcome under surgery is undefined. In general, even if the theoretical positivity assumption holds, the practical positivity assumption may still be violated. This occurs when a set of data is obtained in which certain covariate patterns are not represented for (at least) one exposure status, perhaps due to small sample size. In this situation, no outcome information on this subgroup is available for this exposure status. Therefore, comparisons for different exposures within this subgroup become impossible unless the analyst decides that it is defensible to smooth between groups (for instance, by decreasing the dimension of X). The theoretical positivity assumption is required for definition of the parameter, and the practical assumption is required for estimation. A dataset will be described as "sparse" when very few subjects of certain covariate patterns are available for any one of the exposure groups.

Assumption iv) was described notably by Rubin [55] in the related context of missing data. Ignorability is the minimal sufficient criteria for inference of missing data, which can be translated to the causal scenario as

$$Pr(A \mid X, Y^0, Y^1) = Pr(A \mid X, Y^{obs}),$$

where A is the indicator of treatment, and X is an observed set of covariates (occurring prior to treatment, or representative of a pre-treatment variable), and Y^{obs} is the observed outcome. "Unbiased" treatment assignment or conditional randomization occurs when the conditional probability of A is also independent of the observed outcome [17].

A precise definition of a "confounder" is generally not provided, but for the purposes of this thesis and intuition, it will be described as a pre-treatment variable which, relative to a set of pre-treatment covariates W, removes a portion of the confounding bias when conditioned on in addition to the set W. This allows for redundancy between confounders (two confounders relative to the same set may not both be needed to adjust for confounding) and relativity (a variable may be described as a confounder with respect to one set of covariates, but not another). A confounding variable can also be (insufficiently) described as a variable that causally affects the outcome and is differentially distributed in the exposure groups [17].

2.1.2 Longitudinal data, confounding, and causal parameters

Suppose that we observe data of the general longitudinal form
$O = (L_0, A_0, L_1, A_1, ..., A_{K-1}, L_K = Y)$, where each of these variables may be multivariate. The baseline variable L_0 includes all potentially confounding variables.

The "intervention nodes" $A_t, t = 0, ..., K-1$ include the exposure status at each time point. For a longitudinal cohort study where patients may become lost to follow-up, the intervention nodes may also include censoring status. The variables $L_t, t = 1, ..., K-1$ include any time-dependent confounding variables (affecting both subsequent exposure and outcome) that are also affected by previous exposures. If the outcome of interest is survival at the final end-point, then the time-dependent variables must include survival status at each time point.

Our data O consist of n independently and identically distributed draws from a true underlying distribution P_0. Let $\bar{L}_t = (L_0, ..., L_t)$ be a history of the time-dependent variables up until time t, and let $\bar{A}_t$ be similarly defined for the exposure history. The true density may be decomposed corresponding to the time-dependent structure of the data as

$$P_0 = \prod_{t=0}^{K-1} g_t(A_t \mid \bar{A}_{t-1}, \bar{L}_t) \underbrace{\prod_{t=0}^{K} f_t(L_t \mid \bar{A}_{t-1}, \bar{L}_{t-1})}_{Q_0} \tag{2.1}$$

(where A_{-1} and L_{-1} should be taken as the empty set).

When longitudinal data arise with time-dependent confounding of the effect of exposure on outcome, standard regression techniques (including mixed effects modeling, linear regression, and standard Cox-proportional hazards modeling in the survival setting) will yield biased results whether or not they control for the time-dependent confounders [43]. Careful identification and estimation of the target parameter is required.

Defining the parameter of interest

The primary parameter of interest is the marginal exposure-specific mean, $\psi^{\bar{a}} = E(Y^{\bar{a}})$, described as the expected counterfactual outcome under no censoring and fixed exposure regimen (which we will also call *exposure pattern*) $\bar{a} = (a_0, a_1, ..., a_{K-1})$ corresponding to a specific history of exposure. Different parameters of interest describing the exposure effects can be defined as functions of the marginal exposure-specific mean, for different exposure patterns. An example of such a function is the difference in means for exposure patterns $\bar{a}^{(0)}$ and $\bar{a}^{(1)}$, $E(Y^{\bar{a}^{(0)}}) - E(Y^{\bar{a}^{(1)}})$.

A larger set of causal parameters that might be of interest comes from modeling the counterfactual outcomes conditional on time-varying exposure pattern, time and subgroup status. Often this is accomplished by assuming a linear model for the outcome. A *marginal structural model* (MSM) is a model of the causal effect of a time-dependent exposure on an outcome, when time-dependent confounding also affected by previous treatment is present [46]. The parameters of interest are the coefficients of the covariates of the linear model. Note that modeling the difference in marginal exposure-specific means is an example of a saturated MSM equivalent to:

$$E(Y) = \beta_0 + \beta_1 I(\bar{A} = \bar{a}^{(0)}) + \beta_2 I(\bar{A} = \bar{a}^{(1)})$$

where $\psi^{\bar{a}^{(0)}} = \beta_0 + \beta_1$ and $\psi^{\bar{a}^{(1)}} = \beta_0 + \beta_2$ so that $\beta_1 - \beta_2$ is equal to the difference between the marginal exposure-specific means. Also note that parametric assumptions are not imposed in this equation. An unsaturated example given in Robins, Hernán, and Brumback (2000) [46] describes a binary outcome being modeled as a

function of cumulative exposure:

$$\text{logit}[E(Y^{\bar{a}})] = \gamma_0 + \gamma_1 cum(\bar{a})$$

where $cum(\bar{a}) = \sum_{t=0}^{K-1} a_t$ for $\bar{a} = (a_0, ..., a_{K-1})$.

Hernán, Brumback and Robins (2000) [21] choose to fit a marginal Cox-proportional hazards model for a survival outcome. Suppose for each subject there exists a survival time T and censoring time C measured from a fixed baseline so that T is only observed if $T < C$. The counterfactual survival time $T^{\bar{a}}$ is the survival time that would have been observed had the subject been exposed according to $\bar{a}$ and uncensored. Drawing a link to the longitudinal context, let $L_{1,t}$ indicate survival at time point t, so that the marginal exposure-specific survival curve can be defined as $S^{\bar{a}}(t) = Pr(T^{\bar{a}} > t) = Pr(L^{\bar{a}}_{1,t} = 1)$, where $L^{\bar{a}}_{1,t}$ is the counterfactual survival status at time t. For a baseline covariate, $W \subseteq L_0$, Hernán et al model the counterfactual hazard of mortality as

$$\lambda^{\bar{a}}(t \mid W) = \lambda_0(t) \exp(\beta_1 a_t + \beta_2 W),$$

where the discrete hazard function can be defined as $\lambda^{\bar{a}} = Pr(T^{\bar{a}} = t \mid T^{\bar{a}} \geq t)$.

Identification of the exposure-specific mean

The marginal exposure-specific mean can be identified through Robin's G-computation formula [43] conditional on the assumption of sequential randomization. Similar to the conditional randomization assumption in the cross-sectional setting (also known as "no unmeasured confounders"), the identification of the longitudinal parameter

requires that

$$Pr(A_t \mid \{\bar{L}_K^{\bar{a}}, \bar{A}_{K-1}; \text{for all } \bar{a} \in \bar{\mathcal{A}}\}) = Pr(A_t \mid \bar{L}_t, \bar{A}_{t-1}), t = 0, ..., K-1.$$

In words, this requires that the distribution of exposure only relies on the observable past. In particular, it also assumes that there are no unmeasured time-dependent confounders.

Let $Q_t(l_t \mid \bar{l}_{t-1}, \bar{a}_{t-1})$ be the conditional distribution of the time-dependent variable L_t evaluated at realization l_t and conditional on fixed exposure history $\bar{a}_{t-1} = (a_0, a_1, ..., a_{t-1})$ and given time-dependent variables path $\bar{l}_{t-1} = (l_0, l_1, ..., l_{t-1})$. Then, the marginal exposure-specific mean can be identified through the G-computation formula using nested integrals:

$$\psi^{\bar{a}} = \int_{l_0} \cdots \int_{l_{K-1}} E(Y \mid \bar{l}_{K-1}, \bar{a}_{K-1}) \times Q_{K-1}(l_{K-1} \mid \bar{l}_{K-2}, \bar{a}_{K-2}) dl_{K-1} \cdots Q_0(l_0) dl_0.$$

Here, the Lebesgue integrals are taken over the supports of $L_0, L_1, ..., L_{K-1}$, respectively (and each represents multiple integrals in the case where the time-dependent variables are multivariate). For the simplified scenario where the time-dependent variables are univariate and binary (but the structure of the baseline variable is unconstrained), this formula reduces to nested summations:

$$\psi^{\bar{a}} = \int_{l_0} \cdots \sum_{l_{K-2}=\{0,1\}} \sum_{l_{K-1}=\{0,1\}} E(Y \mid \bar{l}_{K-1}, \bar{a}_{K-1}) \bar{Q}^{\bar{a}}_{L_{K-1}}(\bar{l}_{K-1}) \bar{Q}^{\bar{a}}_{L_{K-2}}(\bar{l}_{K-2}) \cdots f_0(l_0) dl_0, \tag{2.2}$$

where $\bar{Q}^{\bar{a}}_{L_t}(\bar{l}_t) = Pr(L_t = l_t \mid \bar{L}_{t-1} = \bar{l}_{t-1}, \bar{A}_{t-1} = \bar{a}_{t-1}), t = 1, ..., K-1$ is the conditional probability that L_t takes the value of l_t conditional on the fixed exposure history and time-dependent variable path.

Identification of a parameter of a marginal structural model

In general, the identification of a marginal structural model parameter can be made through the choice of a loss function, $\mathcal{L}(\theta)$, in addition to the choice of the marginal mean model. An example of such a loss function is the negative of the log of an exponential family member distribution:

$$\mathcal{L}_Y(\theta) = -\left\{\frac{Y\theta - b(\theta)}{a(\eta)}\right\}$$

where $a(\eta)$ is the family-specific dispersion factor that depends on the nuisance parameter η. In the corresponding density, the mean of Y is $E(Y|\theta) = b'(\theta)$. Let g be the canonical link function such that $g\{E(Y|\theta)\} = \theta$. Using a Gaussian family member, for example, this loss function simplifies to a squared-error loss. Suppose we specify a marginal model $E(Y^{\bar{a}}) = \mu(\bar{a}, W, \boldsymbol{\beta})$ where W is a baseline variable included in the design matrix of the MSM and $\boldsymbol{\beta}$ is the vector of model coefficients. Then, the parameters of the MSM can be defined to minimize the loss function [33]:

$$\boldsymbol{\beta} = \arg\min_{\beta'} \sum_{\bar{a}\in\bar{\mathcal{A}}} E\mathcal{L}[\mu(\bar{a}, W, \boldsymbol{\beta}')]$$

which, for an exponential family member loss function, simplifies to

$$\boldsymbol{\beta} = \arg\max_{\beta'} \sum_{\bar{a}\in\bar{\mathcal{A}}} E\left\{\frac{Y^{\bar{a}}\mu(\bar{a}, W, \boldsymbol{\beta}') - b(\mu(\bar{a}, W, \boldsymbol{\beta}'))}{a(\eta)}\right\}.$$

2.1.3 A note about marginal versus conditional parameters

Specialists in causal inference often prefer to estimate marginal parameters, such as the marginal exposure-specific mean that was described above. For many parameters and models, the marginal parameter is not the same as the one obtained when conditioning on a set of covariates. This is called *noncollapsibility* of the parameter. As a specific example, consider n independent, identically distributed observations (W, A, Y). The marginal exposure-specific difference (for two single time point exposures, denoted $A = 0, 1$) of a binary outcome Y may be estimated through logistic regression. If the variable W is sufficient to control for confounding of the effect of interest, a logistic regression could be fit, with the regression model $\text{logit}[E(Y \mid A, W)] = \beta_0 + \beta_1 A + \beta_2 W$. An estimate of the marginal difference could be obtained using the resulting fit

$$E_n(Y^0) - E_n(Y^1) = \frac{1}{n}\sum_{i=1}^{n}\left\{\text{expit}(\hat{\beta}_0 + \hat{\beta}_2 w_i) - \text{expit}(\hat{\beta}_0 + \hat{\beta}_1 \times 1 + \hat{\beta}_2 w_i)\right\},$$

where w_i is the ith subject's realization of the baseline variable. This effect estimate is not necessarily equal to the $\hat{\beta}_1$ coefficient, which is often used as an estimate of the treatment effect.

For the general regression case, consider two regressions fit – one controlling for both W and Z, and the other controlling only for W. The two models are: $g[E(Y \mid A, W, Z)] = \beta_0 + \beta_1 A + \beta_2 W + \beta_3 Z$ and $g[E(Y \mid A, W)] = \beta_0^* + \beta_1^* A + \beta_2^* W$. The conditional parameter is said to be noncollapsible for β_1 over Z if $\beta_1 \neq \beta_1^*$ [17]. Therefore, if two investigators modeled the same data but controlled for

different variables and *even if confounding control was adequate in both analyses*, the parameters estimated using such a regression method would not be the same.

The estimation of marginal parameters is nevertheless sensitive to the choice of population and dependent on appropriate confounder control which makes marginal parameters somewhat difficult to compare across studies [82]. This is why epidemiological studies must take care to carefully define their population of interest, and describe the parameter as specific to the chosen population. While strictly observed for randomized trials, controlling the entrance criteria in order to be able to clearly define the parameter of interest is also necessary for observational studies [6].

2.2 Inference

The previous section described conditions under which causal parameters are estimable, and the identification of the longitudinal parameter that is of interest in this thesis. This section contains an overview of several estimation methods defined for parameters related to a longitudinal data structure.

2.2.1 Inverse probability of treatment weighting

One class of inferential methods utilizes measured confounders to estimate subjects' differing levels of probability of obtaining treatment (or the probability of leaving the study). They involve calculating the probability of receiving treatment given a set of confounders. If missing visits or censoring also occur in the study, a conditional probability of missingness can be calculated. Once the treatment and missingness conditional probabilities are calculated from the observed data, they can be included as a covariate in a regression analysis [50], used to weight each participant in estimating the mean outcome, or in a model for the outcome [46, 42], or used

to match subjects who are comparable with respect to their covariate values [12]. Bias due to the measured confounders is removed because the levels of confounders will be similar within strata defined by these probabilities.

Inverse probability of treatment weighting (IPTW) (which will also be used to refer to weighting to adjust for censoring or missingness) calculates the conditional probabilities of treatment and censoring, and uses their inverses to weight each participant. This procedure produces reweighted exposure groups that are comparable in terms of the confounders included in the weighting model. For the longitudinal data structure described in Section 2.1.2, IPTW can be used to estimate the marginal exposure-specific mean by reweighting each subject when calculating the mean outcome. This procedure can also be used to reweight non-monotone missing visits (i.e. intermittently missing visits) [42] but this section only demonstrates the method for censored data.

Let A_t now refer to the exposure status at time t, with $\bar{A}_{K-1}$ describing the complete history of exposure, and C_t, $t = 1, ..., K-1$, referring to the monotone censoring status. An estimate of the exposure-specific mean under $\bar{a}$ can be obtained by weighting the outcomes and using

$$\hat{\psi}^{\bar{a}}_{IPTW} = \frac{1}{n}\sum_{i=1}^{n} Y \frac{I(\bar{A}_{K-1} = \bar{a})I(C_{K-1} = 0)}{\bar{g}_{A_0,n}(a_0)\bar{g}_{C_0,n}(a_0)\prod_{t=1}^{K-1}\bar{g}_{A_t,n}(\bar{a}_t)\bar{g}_{C_t,n}(\bar{a}_t)}, \tag{2.3}$$

where

$$\bar{g}_{A_0,n}(a_0) = Pr_n(A_0 = a_0 \mid L_0),$$

$$\bar{g}_{C_0,n}(a_0) = Pr_n(C_0 = 0 \mid A_0 = a_0, L_0),$$

$$\bar{g}_{A_t,n}(\bar{a}_t) = Pr_n(A_t = a_t \mid \bar{A}_{t-1} = \bar{a}_{t-1}, C_{t-1} = 0, \bar{L}_t), \text{ and,}$$

$$\bar{g}_{C_t,n}(\bar{a}_t) = Pr_n(C_t = 0 \mid \bar{A}_t = \bar{a}_t, C_{t-1} = 0, \bar{L}_t).$$

These weights can be constructed for each subject. Once estimated, they can also be used to weight each observation in a MSM. For example, in a time-dependent MSM, the observations used are each subject at each time point, which are modeled conditional on the covariates in the MSM. In the saturated MSM case where the parameter of interest is truly the exposure-specific mean, the IPTW procedure is the same as taking a weighted mean of the outcome of interest, as in Equation (2.3).

When data are sparse, it is often the case that few subjects were observed to have been treated or untreated for a range of covariate values in some of the conditional models. When this occurs, the affected conditional models will predict very small probabilities for being exposured or unexposed for subjects at those covariate levels. The inverse of the probabilities will therefore be very large, sometimes approaching computational infinitude. One solution is to stabilize the weights by including any function of the baseline and exposure variables (but not the time-dependent confounders) in the numerator of the weights. For instance, the weights are often altered to be:

$$w(\bar{L}_{K-1}; \bar{a}) = \frac{h(\bar{A}_{K-1}, \bar{C}_{K-1}; \bar{a})}{\bar{g}_{A_0,n}(a_0)\bar{g}_{C_0,n}(a_0)\prod_{t=1}^{K-1}\bar{g}_{A_t,n}(\bar{a}_t)\bar{g}_{C_t,n}(\bar{a}_t)}$$

where the function h can be set to

$$h(\bar{A}_{K-1}, \bar{C}_{K-1}; \bar{a}) = Pr_n(A_0 = a_0) Pr_n(C_0 = 0 \mid A_0 = a_0) \times \prod_{t=1}^{K-1} Pr_n(A_t = a_t \mid \bar{A}_{t-1} = \bar{a}_{t-1}, C_{t-1} = 0) Pr_n(C_t = 0 \mid \bar{A}_t = \bar{a}_t, C_{t-1} = 0).$$

To maintain confounding control, variables that are included in the conditional probabilities in the numerator must be included in the outcome model of the MSM. Therefore, when fitting an MSM that doesn't include any confounding variables, the numerator also cannot be a function of any confounding variables.

When stabilization also fails to control the size of the inverse weights, analysts may choose to perform an ad hoc procedure such as weight trimming or weight truncation where the top mth percent of weights is either removed (i.e. the corresponding subjects are removed from the calculation) or reduced to the values of the weight at the $1 - m$th percentile [8, 84].

The success of these techniques again relies on the causal assumptions specified in Section 2.1.1. In particular, the set of suspected confounders must be sufficient to adjust for confounding. Similarly, these methods also require that the probability of dropout or missing values only depends on observed variables (i.e. that they are *missing at random* [55]) and variables that do not affect the outcome.

IPTW for survival data

IPTW can be used to fit MSMs for general longitudinal and survival data equally [46, 21]. The exact same weights are used (the inverse conditional probabilities, possibly stabilized), but they are used to weight each uncensored subject-specific

visit in a regression corresponding to the marginal model of interest (and the chosen loss function).

Often, it is of interest to estimate the marginal exposure-specific probability of survival at each time point, $S^{\bar{a}}(t) = Pr(T^{\bar{a}} > t)$, in order to directly visualize or statistically compare the survival curves. The nonparametric Kaplan-Meier method of constructing the survival curve is straight-forward but assumes no confounding and independence between censoring and failure time [23]. Let $\delta_{j,\bar{a}}$ be the number of deaths at time point j corresponding to exposure pattern $\bar{a}_j$ and $Y_{j,\bar{a}}$ the total number of subjects with exposure pattern $\bar{a}_j$ at risk at time j. Then, the Kaplan-Meier estimate for the survival curve at time t for subjects with exposure pattern $\bar{a}_t$ is given by $\prod_{j=1}^{t}\left(1-\frac{\delta_{j,\bar{a}}}{Y_{j,\bar{a}}}\right)$.

IPTW has also been used to reweight the Kaplan-Meier curve in the case of baseline and time-dependent confounding [9]. Xie and Liu [86] developed the Adjusted Kaplan-Meier estimate (AKME), a type of IPTW estimator. For each subject observed at a time point j, let $p_{j,\bar{a}} = Pr(\bar{A}_j = \bar{a}_j \mid \bar{L}_{j-1})$ be the probability of that subject having a given exposure pattern, $\bar{a}_j$, possibly conditional on a history of time-dependent and baseline covariates, $\bar{L}_{j-1}$. Let δ be the indicator of whether a subject was observed to fail (as opposed to having been censored). Then, $\delta_{j,\bar{a}} = \sum_{T=j} I(\bar{A}_j = \bar{a}_j)\delta$, where the sum is taken over all subjects who fail exactly at time j. Similarly, $Y_{j,\bar{a}} = \sum_{T\geq j} I(\bar{A}_j = \bar{a}_j)$ where the sum is taken over all subjects at risk at time j. The weighted counterparts of these measures can be described as $\delta^W_{j,\bar{a}} = \sum_{T=j} I(\bar{A}_j = \bar{a}_j)\delta/p_{j,\bar{a}}$ and $Y^W_{j,\bar{a}} = \sum_{T\geq j} I(\bar{A}_j = \bar{a}_j)/p_{j,\bar{a}}$ and fit for all subjects.

The AKME is equal to

$$\hat{S}^{\bar{a}}_{AKME}(t) = \begin{cases} 1 & \text{if no failures are observed prior to } t \\ \prod_{j=1}^{t} \left(1 - \frac{\delta^W_{j,\bar{a}}}{Y^W_{j,\bar{a}}}\right) & \text{otherwise.} \end{cases}$$

The AKME is a method that is unbiased as long as the probabilities of exposure are correctly specified. This estimator is also unbiased under MAR censoring if a model for censoring can also be correctly specified and included in the weights.

2.2.2 Plug-in estimation and G-computation

In Section 2.1.2, the marginal exposure-specific mean was identified using Robin's G-computation formula. Substitution or "plug-in" estimation of this parameter [43] can be produced by estimation of each of the components in the formula (or through Monte-Carlo sampling of the densities in the formula for the general structure) that are then inserted into the formula. If the time-dependent variables are binary, the formula simplifies to Equation (2.2), and estimation of the target parameter can be produced through 1) estimation of the expectation of the outcome conditional on the past and the fixed exposure pattern,

$$\bar{Q}^{\bar{a}}_{n,Y}(\bar{l}_{K-1}) = E_n(Y \mid \bar{L}_{K-1} = \bar{l}_{K-1}, \bar{A}_{K-1} = \bar{a}_{K-1}),$$

2) estimation of the conditional probabilities for each time-dependent variable,

$$\bar{Q}^{\bar{a}}_{n,L_t}(\bar{l}_{t-1}) = Pr_n(L_t = l_t \mid \bar{L}_{t-1} = \bar{l}_{t-1}, \bar{A}_{t-1} = \bar{a}_{t-1}), \text{for all } \bar{l}_t,$$

and 3) fitting the empirical distribution for the baseline covariate(s), $Q_n(l_0) = 1/n$ for each subject's baseline covariates. Then, an estimate of $\psi^{\bar{a}} = E(Y^{\bar{a}})$ can be

calculated using

$$\hat{\psi}^{\bar{a}}_{GCOMP} = \frac{1}{n}\sum_{i=1}^{n}\sum_{l_1=0,1}\cdots\sum_{l_{K_1}}\bar{Q}^{\bar{a}}_{n,Y}(\bar{l}_{K-1})\bar{Q}^{\bar{a}}_{n,L_t}(\bar{l}_{t-1})\cdots\bar{Q}^{\bar{a}}_{n,L_1}(l^i_0)$$

Note that when fitting the G-computation formula, the conditional expectations are calculated over each combination of the time-dependent variable path $\bar{l}_{K-1} = (l^i_0, l_1, ..., l_{K-1})$ where the baseline covariates are subject-specific.

A variant of the G-computation estimator may be used to estimate the parameters of a marginal structural model as well [67, 80] but is rarely used in practice, with IPTW being the prominent method for fitting MSMs.

In general, plug-in estimators can be described for any parameter that can be defined as a smooth function of a component of the underlying data-generating function, $Q_0 \subset P_0$. Let Ψ be one such function that takes an argument in the model space $\mathcal{M}$ and returns a value in the space of real numbers. For a parameter that can be described as $\psi = \Psi(Q_0)$, plug-in estimation of this parameter is available by first fitting the component of the necessary underlying distribution, Q_n, and then plugging this estimate into the function so that $\hat{\psi} = \Psi(Q_n)$.

Unbiasedness of a plug-in estimator relies on the correct modeling of the required components of the data-generating distribution. Correct parametric specification of the density models will result in efficient estimation. However, the general procedure of optimizing a density estimate (possibly by maximizing a likelihood) and then plugging it into a function doesn't necessarily optimize the estimation of the target parameter. Such a procedure may lead to arbitrary trade-offs between bias

and variance and produce a slowly-converging estimator (even when asymptotically unbiased) [78].

No closed form variance is available for G-computation or for a general plug-in estimator. Often, the bootstrap procedure is used [67, 80]. Possibly because of this, and because G-computation rapidly becomes exceedingly computationally complex for more time points and time-varying exposures, the method remains underutilized in epidemiological studies [82].

2.3 Efficient estimation

The literature on efficient estimating equations goes back many decades, with roots in Fisher's definition of the information of an estimator, sufficiency, and minimal asymptotic variance [24, 4]. The following theory on minimal variance estimation is a closely related generalization of the Cramér-Rao lower bound [40, 10, 41] applicable to all regular, asymptotically linear semiparametric estimators. Van der Laan and Robins (2003) [76] extended efficient methods to the general causal inference and missing data settings, and described a general theory of semiparametric efficiency applicable to causal parameters.

The first part of this section briefly describes elements of the theory that led up to the development of Targeted Maximum Likelihood Estimation. The second discusses a relevant estimator for the marginal exposure-specific mean of longitudinal data that was developed by Bang and Robins (2005) [1] using this theory. The final part describes the general framework of TMLE as proposed by Van der Laan and Rubin (2006) [78].

2.3.1 The efficient influence function for plug-in estimation

Suppose we observe n independent and identically distributed observations $O = \{O_i, i = 1, ..., n\}$ with joint probability distribution $P_0 \in \mathcal{M}$. Let the R-dimensional parameter ψ be defined as a pathwise differentiable function of a component of the underlying density, $Q_0 \subset P_0$. That is, let $\psi = \Psi(Q_0)$ for some function Ψ that takes an argument in a statistical model space, $\mathcal{M}$, and returns a real value (possibly a real vector). For instance, Ψ could be defined as an expected value operator on a single variable, or as the G-computation formula that identified the exposure-specific mean in Section 2.1.2. As described in Section 2.2.2, plug-in estimators for the parameter ψ may be constructed using density estimator Q_n and evaluating the expression $\Psi(Q_n)$.

Let $Q_0(\epsilon) \in \mathcal{M}$ be a regular parametric submodel of the density component Q_0 (such that $Q_0(0) = Q_0$) with score function S_ϵ. Under pathwise differentiability,

$$\frac{d}{d\epsilon}\Psi(Q_0(\epsilon)) \mid_{\epsilon=0} = E\left[D(P_0)(O)S_\epsilon(O)\right] \tag{2.4}$$

for some R-dimensional function $D(P_0)$ with finite variance and zero mean. This function, called a gradient of the pathwise derivative at P_0 [76], and elsewhere termed an *influence function* [71], can be seen as a member of a Hilbert space equipped with inner product $< d_1, d_2 > = E(d_1(O)d_2(O))$ [78]. $S_\epsilon(O)$ is the score function with respect to parameter ϵ for the parametric submodel, $Q_0(\epsilon)$, i.e. $S_\epsilon(O) = d/d\epsilon \ \log\{ \ Q_0(\epsilon)(O) \ \} \mid_{\epsilon=0}$.

Let a tangent sub-space $T(P_0)$ of the Hilbert space be the closure of the linear span of score functions $S_\epsilon(O)$ for each parametric submodel, $Q(\epsilon)$. Then, the *canonical gradient* $D^*(P_0)$ is defined as the unique gradient contained within $T(P_0)$ (see Tsiatis, 2006, ch. 3 [71] for the geometry of gradients).

For each observation, let $D(P_0)(O_i)$ be the subject-specific component of the gradient $D(P_0)(O)$. Given a gradient $D(P_0)$, there exists a regular asymptotically linear estimator that is associated with it. Let P_n be a fit of the underlying probability distribution P_0. And let Q_n be the fit of the component of the distribution required for the parameter identification $Q_0 \subseteq P_0$. Then, an estimator $\Psi(Q_n)$ can be found using the definition

$$n^{1/2}[\Psi(Q_n) - \psi] = n^{-1/2} \sum_{i=1}^{n} D(P_0)(O_i) + o_P(1) \tag{2.5}$$

where $o_P(1)$ is the component that converges to zero in probability as n increases. By an application of the Central Limit Theorem, this implies that

$$n^{1/2}[\Psi(Q_n) - \psi] \rightarrow^{\mathcal{D}} N\{0, E[D(P_0)(O)D(P_0)(O)^T]\}$$

when $P_n(O)$ is asymptotically unbiased (a full description is available in Van der Laan and Rose, 2011, Appendix A.2 [77]). From Equation (2.5), it is apparent that the asymptotic behaviour of the estimator is completely determined by its associated gradient. In particular, the large sample variance of the estimator can be approximated by the variance of the gradient scaled by $1/n$. Efficiency in estimation can therefore be improved by finding estimators associated with gradients that have the minimal variance.

It can be shown that the gradient with minimal variance is identical to the canonical gradient, commonly referred to as the *efficient influence function* (Van der Laan and Rose, 2011, Appendix A.4 [77]). This generalization of the Cramér-Rao lower bound theorem implies that there is a lower variance bound amongst the class of regular, semiparametric estimators that can only be attained through estimation with the efficient influence function.

The efficient influence function can be found by performing a linear projection of any gradient onto the tangent space $T(P_0)$. Van der Laan and Robins [76] derived the efficient influence functions for many causal parameters. As will be described in more detail in Section 2.3.2, Bang and Robins [1] derived the efficient influence function for the exposure-specific mean parameter for a longitudinal exposure. Under a different factorization of the data-generating distribution, Van der Laan [73] derived a simplified expression of the efficient influence function for the same parameter when time-dependent confounders can be factorized into binary variables.

A simple example

As a simple example of the derivation of an influence function (taken from the 2011 course notes of Van der Laan at University of California, Berkeley), consider a data set with N independent and identically distributed observations, each with baseline variable W, discrete exposure variable A, and outcome Y. Under the time-ordering assumption, the full likelihood is given as $P_0(O) = Q_Y(Y \mid A, W)g_A(A \mid W)Q_W(W)$, where Q_Y represents the conditional distribution of the outcome variable, g_A represents the conditional distribution of the exposure variable, and Q_W is the distribution of the baseline variable. Under the counterfactual framework

where single exposure A is set to a, the likelihood of the data factorizes into $Q_0 = Q_Y(Y \mid A = a, W)Q_W(W)$. Correspondingly, the exposure-specific mean $E(Y^a)$ is identifiable through the G-computation formula as

$$\begin{aligned} E(Y^a) = \Psi(Q_0) &= \int_w E(Y \mid A = a, W = w) Pr(W = w) dw \\ &= \int_w \int_y y Pr(Y = y \mid A = a, W = w) Pr(W = w) dy dw \end{aligned}$$

where E and Pr are an expectation and probability function, respectively, taken with respect to the probability distribution Q_Y. Note that, as usual, the exposure-specific mean can be identified independently of the likelihood component g_A. The integrals are taken over the support of Y and W, respectively. The parameter can be estimated by choosing and fitting a model for $Pr(y \mid a, w) = Pr(Y = y \mid A = a, W = w)$, using the empirical fit for Q_W (so that $Pr_n(w) = Pr_n(W = w) = 1/n$ for each subject), and carrying out an empirical average of the conditional probability for Y over all subjects.

Suppose we have defined some parametric submodel of Q_0, which implies a submodel for both components, say $Q(\epsilon) = Q_W(\epsilon)Q_Y(\epsilon)$. Let $S_{\epsilon,W}(w)$ be the score function for the submodel $Q_W(\epsilon)$. Similarly, let $S_{\epsilon,Y}(y \mid a, w)$ be the score function for the conditional distribution of Y. Lower-case letters indicate evaluation at a realization. By properties of the score function, it follows that $E_{P_0}\{S_{\epsilon,W}(W)\} = E_{P_0}\{S_{\epsilon,Y}(Y \mid A, W)\} = 0$.

The first goal is to identify a gradient of the parameter, $\psi^a = E(Y^a)$. Construct a regular parametric submodel of the density Q_0 with respect to some parameter ϵ

by defining

$$\Psi(Q(\epsilon)) = \int_w \int_y y\{1 + \epsilon S_{\epsilon,Y}(y \mid a, w)\} Pr(y \mid a, w)\{1 + \epsilon S_{\epsilon,W}(w)\} Pr(w) dy dw.$$

Taking a derivative with respect to ϵ (and assuming that the integral and derivative and be interchanged) yields

$$\begin{aligned}\frac{d}{d\epsilon}\Psi(Q(\epsilon)) = &\int_w \int_y y S_{\epsilon,Y}(y \mid a, w) Pr(y \mid a, w) Pr(w) dy dw \\ &+ \int_w \int_y y S_{\epsilon,W}(w) Pr(y \mid a, w) Pr(w) dy dw.\end{aligned}$$

Multiply and divide both components by $Pr(a \mid w)$, the probability of obtaining exposure a conditional on baseline w. Then, using a summation manipulation (and because A has a discrete support), this is equal to

$$\begin{aligned}\frac{d}{d\epsilon}\Psi(Q(\epsilon)) = &\int_w \sum_{a^*} \int_y y \frac{I(a^* = a)}{Pr(a \mid w)} S_{\epsilon,Y}(y \mid a, w) Pr(y \mid a, w) dy Pr(a \mid w) Pr(w) dw \\ &+ \int_w \sum_{a^*} \int_y y \frac{I(a^* = a)}{Pr(a \mid w)} S_{\epsilon,W}(w) Pr(y \mid a, w) dy Pr(a \mid w) Pr(w) dw \\ =&E_{P_0}\left[Y \frac{I(A = a)}{Pr(A \mid W)} S_{\epsilon,Y}(Y \mid A, W)\right] + E_{P_0}\left[Y \frac{I(A = a)}{Pr(a \mid W)} S_{\epsilon,W}(W)\right] \\ =&E_{P_0}\left[Y \frac{I(A = a)}{Pr(A \mid W)} \{S_{\epsilon,Y}(Y \mid A, W) + S_{\epsilon,W}(W)\}\right].\end{aligned}$$

The above expectation is taken over P_0, the complete distribution. To see how this equation relates to the definition of the gradient in Equation (2.4), note that the score function for the parametric subspace $Q_0(\epsilon)$ separates as $S_{\epsilon,Y}(Y \mid A, W) + S_{\epsilon,W}(W)$, and that its expectation under P_0 is zero. The above expectation is therefore equal

to

$$\frac{d}{d\epsilon}\Psi(Q(\epsilon)) = E_{P_0}\left[\left\{Y\frac{I(A=a)}{Pr(a \mid W)} - \psi^a\right\}S_\epsilon(O)\right].$$

Then, by the definition, $D(O) = Y\frac{I(A=a)}{Pr(A|W)} - \psi^a$, which has mean zero as required and is equal to a gradient of the parameter $\psi^a = E(Y^a)$. Note that when empirically evaluated by taking a mean over N subjects, this is equivalent to the unstabilized IPTW method, resulting in an unbiased estimator of $E(Y^a)$ when $Pr(a \mid W)$ is correctly specified.

The above gradient is not the canonical gradient as it is not an element of the semiparametric tangent space associated with Q_0, which we shall denote $\mathcal{T}_Q$. Since $Q_0 = Q_Y(Y \mid A = a, W)Q_W(W)$ is decomposed as a product of orthogonal densities, this semiparametric tangent space is defined as the mean square closure of the space spanned by the score functions, $S_{\epsilon,Y}$ and $S_{\epsilon,W}$, for all parametric submodels, $Q(\epsilon)$. This can be written using the direct summation $\mathcal{T}_Q = \mathcal{T}_Y \oplus \mathcal{T}_W$. To obtain the canonical gradient, the gradient $D(O)$ can be projected onto each orthogonal tangent space, resulting in a gradient with the smallest norm as defined by the inner product metric, or equivalently, the smallest variance. By Tsiatis' Theorem 4.5 [71],

this projection is:

$$\begin{aligned}
\prod(D \mid \mathcal{T}_Q) &= \prod(D \mid \mathcal{T}_Y \oplus \mathcal{T}_W) \\
&= \prod(D \mid \mathcal{T}_Y) + \prod(D \mid \mathcal{T}_W) \\
&= \{E(D \mid Y, A = a, W) - E(D \mid A = a, W)\} + E(D \mid W) \\
&= \left\{ Y \frac{I(A = a)}{Pr(a \mid W)} - E(Y \mid A = a, W) \frac{I(A = a)}{Pr(a \mid W)} \right\} + \{E(Y \mid A, W) - \psi^a\} \\
&= \frac{I(A = a)}{Pr(a \mid W)} \{Y - E(Y \mid A = a, W)\} + E(Y \mid A = a, W) - \psi^a. \quad (2.6)
\end{aligned}$$

Therefore, Equation (2.6) is $D^*(O)$, the canonical gradient, or the efficient influence function for ψ.

2.3.2 Efficient and double robust inference for longitudinal parameters

One strategy to produce an estimator connected to the efficient influence function is to use the efficient influence function to define an estimating equation (so that P_n^* is and m-estimator [71]) where P_n^* solves $\sum_{i=1}^n D^*(P_n^*) = 0$. Extended to the causal inference setting by Robins and Rotnitzky [47], such an estimator may have a closed form solution or require an optimization algorithm to solve the equation.

One example derived from the general efficient estimating equation framework is the augmented inverse probability of treatment weighted estimator (AIPTW) [47, 45, 25], demonstrated originally for censoring, but later applied equivalently to estimating causal effects of exposure. For the simple dataset $O = (W, A, Y)$, the efficient influence function of the exposure-specific mean, $\psi^a = E(Y^a)$, was derived in Equation (2.6). Treated as an estimating equation, this equation can be set equal to zero and solved for the parameter ψ^a. This can easily be accomplished in closed

form, resulting in the AIPTW estimator

$$\hat{\psi}^a_{AIPT} = \frac{I(A=a)}{Pr(a \mid W)} \{Y - E(Y \mid A=a, W)\} + E(Y \mid A=a, W).$$

Similarly, Robins, Rotnitzky and Zhao (1995) [49, 48] produced semiparametric efficient estimators for the censoring-free mean outcome in the context of repeated measures. Robins and Rotnitzky developed an efficient Cox proportional hazards model for censored survival data [47].

For the longitudinal setting with time-dependent confounders $O = (L_0, A_0, L_1, A_1, ..., A_{K-1}, L_K = Y)$, described in Section 2.1.2, Bang and Robins [1] described an alternative decomposition of the likelihood of the marginal exposure-specific mean and thereby derived a closed-form efficient influence function for $\psi^{\bar{a}} = E(Y^{\bar{a}})$. Recall that $\bar{a} = (a_0, a_1, ..., a_{K-1})$ denotes a fixed, longitudinal exposure pattern and $\bar{a}_t = (a_0, a_1, ..., a_t)$ denotes the pattern truncated at time t. Define the exposure-specific conditional expectation

$$\bar{Q}_K = E(Y \mid \bar{L}_{K-1}, \bar{A}_{K-1} = \bar{a}_{K-1}).$$

Then, iteratively define

$$\bar{Q}_t = E(\bar{Q}_{t+1} \mid \bar{L}_{t-1}, \bar{A}_{t-1} = \bar{a}_{t-1}), t = K, ..., 1.$$

Using the property of iterated expectation, the parameter of interest can then be written as $E(Y^{\bar{a}}) = E(\bar{Q}_1)$. To use this decomposition in a type of G-computation estimation, each of the $\bar{Q}_t$'s can be fit with statistical models and the parameter

estimate obtained by taking the empirical mean of the fit of $\bar{Q}_1$. The estimator can then be described as a function of $\bar{Q}_t, t = K, ..., 1$.

Van der Laan and Gruber [74] demonstrate that one way to obtain the Bang and Robins efficient influence function for this parameter is to project the IPTW gradient $\frac{I(\bar{A}=\bar{a})}{Pr(\bar{a}|\bar{L}_{K-1},\bar{a}_{K-1})}Y - \psi^{\bar{a}}$ onto each component of the density Q_0. Q_0 is comprised of the conditional density of each $L_t, t = 0, ..., K$ (as described in Equation (2.1)). These projections result in the efficient influence function (or canonical gradient) defined in terms of the $\bar{Q}_t$'s, resulting in a summation of the components

$$D^*(t)(O) = \frac{I(\bar{A} = \bar{a})}{Pr(\bar{A} = \bar{a} \mid \bar{L}_{t-1}, \bar{A}_{t-2})}(\bar{Q}_{t+1} - \bar{Q}_t), t = 1, ..., K, \text{ and}$$
$$D^*(0)(O) = \bar{Q}_1 - \psi^{\bar{a}}$$

where $\bar{Q}_{t+1}$ is defined as Y for notational simplicity and A_{-1} is null for the same reason.

Bang and Robins present their estimator in terms of a series of regressions (sequentially solving each of the components $D^*(t)$), until finally the estimate for the parameter is set as the empirical mean of $\bar{Q}_1$.

While semiparametric efficient and double robust, efficient estimating equation-based estimators can sometimes be unstable under small misspecifications of the density models [25] because they are unbounded and use inverse probabilities that can become arbitrarily small in certain situations. This sensitivity to modeling violations demonstrated the need for stability in double robust estimation.

2.3.3 Targeted Maximum Likelihood Estimation

Targeted Maximum Likelihood Estimation (TMLE) [78], described in the original paper as the unification of maximum likelihood estimation and estimating function based estimation, is a method that constructs efficient plug-in estimators. Suppose, as before, that the parameter ψ can be defined as a function $\psi = \Psi(P_0)$ where P_0 is a member of a statistical model space $\mathcal{M}$. (To be specific, Ψ may take a subcomponent of P_0 as an argument.)

First, the initial density estimate P_n is fit using a choice of modeling method. The TMLE procedure creates a parametric submodel $P_0(\epsilon)$ so that 1) $P_0(0) = P_0$, and 2) the score (which can be defined more generally in terms of a loss function) is proportional to the efficient influence function, i.e. $d/d\epsilon\mathcal{L}\{P_0(\epsilon)\} \mid_{\epsilon=0} \propto D^*(P_0)$. The fit for ϵ is obtained by a maximum likelihood procedure, or more generally, by minimizing a loss function, so that $\hat{\epsilon}^{(1)} = \text{argmin}_\epsilon d/d\epsilon\mathcal{L}\{P_n(\epsilon)\}$. The updated fit is then $P_n^{(1)} = P_n^{(1)}(\hat{\epsilon})$.

This procedure might need to be iterated (by creating a fluctuation of $P_n^{(1)}$ and updating in the same way to get $P_n^{(2)}$, and so on) until the iteration produces no update (i.e. $\epsilon^{(k)} = 0$). If this series of updates converges, the limiting distribution $P_n^{(\text{inf})}$ is a solution of the efficient influence function set to zero, $D^*(P_n^{(\text{inf})}) = 0$ [78]. Consequently, the plug-in estimator $\Psi(P_n^{(\text{inf})})$ is semiparametric efficient.

The same simple example

Section 2.3.1 described the derivation of the efficient influence function for n independent, identically distributed data of form (W, A, Y) for the exposure-specific

mean parameter, $\psi^a = E(Y^a)$. The TMLE construction of this estimator is given in [51].

A G-computation plug-in estimator for this parameter is

$$\hat{\psi}^a_{GCOMP} = \Psi(P_n^{(1)}) = \frac{1}{n}\sum_{i=1}^{n} E_n(Y \mid A = a, W = w^i)$$

where $\bar{Q}^{(1)}_{n,Y} = E_n(Y \mid A = a, W = w^i)$ (a component of $P_n^{(1)}$) can be estimated with a generalized linear model, for instance. An example of a fluctuation that can be used in the update is

$$\text{logit}\{\bar{Q}^{(2)}_{n,Y}(\epsilon^{(1)})\} = \text{logit}\{\bar{Q}^{(1)}_{n,Y}\} + \epsilon^{(1)} C_n(P_n).$$

We have that this fluctuation satisfies the first criterion because $\bar{Q}^{(2)}_{n,Y}(0) = \bar{Q}^{(1)}_{n,Y}$. The second criterion requires that the score of the fluctuated density (with respect to a loss function) be proportional to the efficient influence function. Using the logistic loss function (and relying on the smoothness of the function of ϵ and the positivity assumption), the score is

$$\begin{aligned}
\frac{d}{d\epsilon}\mathcal{L}\{P_0(\epsilon)\} \mid_{\epsilon=0} &= -\frac{d}{d\epsilon}\left\{Y \log \bar{Q}^{(2)}_{n,Y}(\epsilon) + (1-Y)[1 - \log\{1 - \bar{Q}^{(2)}_{n,Y}(\epsilon)\}]\right\} \\
&\propto Y \frac{1}{\bar{Q}^{(1)}_{n,Y}} \frac{\exp\{\text{logit}\bar{Q}^{(1)}_{n,Y}\}}{[1 + \exp\{\text{logit}\bar{Q}^{(1)}_{n,Y}\}]^2} C_n - \\
&\quad (1-Y)(1 - \bar{Q}^{(1)}_{n,Y}) \frac{\exp\{\text{logit}\bar{Q}^{(1)}_{n,Y}\}}{[1 + \exp\{\text{logit}\bar{Q}^{(1)}_{n,Y}\}]^2} C_n \\
&= C_n\left[Y\{1 - \bar{Q}^{(1)}_{n,Y}\} - (1-Y)\bar{Q}^{(1)}_{n,Y}\right] \\
&= C_n\{Y - \bar{Q}^{(1)}_{n,Y}\}.
\end{aligned}$$

The score can therefore be made proportional to the first component of the efficient influence function in Equation (2.6) by setting $C_n = I(A = a)/Pr_n(a \mid W)$ which had previously been left undefined. The second component of the efficient influence function is proportional to the score of the baseline variable as a result of the use of the empirical distribution. Rosenblum and Van der Laan [51] show clearly why no additional update is required. Intuitively, this is because the empirical distribution used in the initial fit is the nonparametric maximum likelihood estimate for the unconditional distribution of the baseline covariate, and therefore the best possible fit.

The next step in the procedure is to minimize

$$\mathcal{L}\{P_0(\epsilon^{(1)})\} \mid_{\epsilon=0} = -Y \log \bar{Q}^{(2)}_{n,Y}(\epsilon^{(1)}) + (1-Y)[1 - \log\{1 - \bar{Q}^{(2)}_{n,Y}(\epsilon^{(1)})\}]$$

with respect to $\epsilon^{(1)}$ in order to obtain the fit $\hat{\epsilon}^{(1)}$. This minimization can be conveniently performed using an intercept-free logistic regression fit, taking Y as the outcome and C_n as the lone covariate. Then, the updated density estimate is

$$\bar{Q}^{(2)}_{n,Y}(\hat{\epsilon}^{(1)}) = \text{expit}\left[\text{logit}\{\bar{Q}^{(1)}_{n,Y}\} + \hat{\epsilon}^{(1)} \frac{I(A=a)}{Pr_n(a \mid W)}\right].$$

As the variable C_n remains the same, no further updates are required (i.e. convergence is immediate). The final estimate of the parameter is obtained by plugging $\bar{Q}^{(2)}_{n,Y}$ into the function Ψ.

TMLE methods for longitudinal parameters

Van der Laan (2010) [73] developed a TMLE method for the estimation of causal effects of multiple time point exposures. Assuming a binary decomposition of the

time-dependent variables (which must include all time-dependent confounders), he derived a simplified form of the efficient influence function. This could naturally be used in a one-step TMLE procedure (i.e. with only one update step needed for each density) using G-computation as the plug-in estimator. A two time point version of this method is demonstrated and implemented in Chapter 4 of this thesis for a single binary time-dependent confounder.

The difficulty with this method is primarily that it relies upon the binary decomposition of potentially non-factor time-dependent variables. This problem can be partially overcome by transforming each time-dependent variable into (possibly ordered) factor levels. However, since the method requires that each factor level be modeled, a large number of time-dependent variables, factor levels, or time points can all exponentially inflate the computational complexity of the procedure (similar to the computational problems involved in using the discrete G-computation estimator for large numbers of time points).

Van der Laan and Gruber (2012) [74] produced a more computationally convenient longitudinal TMLE estimator using the Bang and Robins (2005) [1] sequential decomposition of the efficient influence function (see Section 2.3.2) and a corresponding G-computation-type plug-in estimator. A detailed explanation of this type of TMLE estimator is given in Chapter 5 of this thesis. It provides an improvement on the original longitudinal TMLE because it can naturally incorporate a large number of time-dependent confounders without requiring additional models to be fit. In addition, the number of models fit increases linearly with the number of time points.

Super Learner applied to TMLE

Computing a TMLE requires fitting all density components present in the efficient influence function, but it does not prescribe a method for doing so. Even when all confounders have been identified, parametric methods like main terms generalized linear models may produce misspecified fits for the baseline density components. Therefore, Van der Laan [73] suggests using nonparametric method Super Learner [75, 36, 37] in order to produce fits of any required density components without relying on the assumption that a specific parametric model holds true.

Briefly, Super Learner is a nonparametric predictive method that involves fitting a library of user-specified models to the data. For each model, the cross-validated estimate and model loss are calculated (using a randomized partitioning the dataset). The Discrete Super Learner then takes the estimate with lowest cross-validated loss and uses it as the final estimate. The (non-discrete) Super Learner instead uses a logistic model to combine the estimates from the different models in a way that minimizes the cross-validated risk over all possible weighted combinations. Van der Laan, Polley and Hubbard [75] show that Super Learner will do no worse in terms of cross-validated loss (such as mean-squared error) than the most successful method in the library, and in practice produces fits with less error (calculated on a validation set or with cross-validation). In a later paper, Polley and Van der Laan [36] demonstrate the finite-sample performance of Super Learner for both simulated and real data in which they conclude that the Discrete Super Learner is "an adaptive and robust estimator selection procedure for small samples" and that Super Learner minimizes over-fitting even over a large library.

Super Learner allows a user to implement TMLE while removing parametric modeling assumptions, and therefore assists in creating a fully nonparametric causal inference method. In addition it removes the obligation to choose a modeling method for the density components, which can be beneficial in that it is unknown a priori which method will perform best in a given scenario.

CHAPTER 3
Objectives

Three studies were undertaken whose manuscripts form this thesis. The primary objectives in pursuing these projects were to further the development of TMLE methodology in the longitudinal context, evaluate recent methodology, and utilize these methods appropriately in real data analysis. A secondary goal was to clearly explain the context, methods, and benefits of TMLE to a growing and uninitiated audience.

In the first manuscript (Chapter 4), we apply a property of the score function of generalized linear models to enhance a targeted longitudinal estimator so that it can be constructed for any generalized exponential family member loss function. When the modeling is performed parametrically, this method naturally incorporates any outcome that can be assigned an exponential family member distribution.

In the second manuscript (Chapter 5), we describe several different substitution estimators for longitudinal data and produce a guide for implementation of the sequential longitudinal TMLE produced by Van der Laan and Gruber [74]. We focus on a case study of the PROmotion of Breastfeeding Intervention Trial (PROBIT) data [27, 28] involving estimation of the causal effect of breastfeeding on gastrointestinal infections in infants.

The third manuscript (Chapter 6) develops the methodology for fitting a marginal structural model to a survival outcome using TMLE, building on the general longitudinal method for the intervention specific mean outcome of Van der Laan and Gruber [74].

CHAPTER 4
Targeted Maximum Likelihood Estimation for Marginal Time-Dependent Treatment Effects under Density Misspecification

Preamble to Manuscript 1. The first general theoretical framework for semiparametric efficient estimation with TMLE was proposed by Van der Laan [73]. The construction of the two time point case of this TMLE was demonstrated for a binary outcome by Rosenblum and Van der Laan [53]. The theoretical contribution of the following manuscript, published in *Biostatistics*, is a careful description of the construction (and implementation) of the two time point TMLE with an outcome that can naturally be prescribed a generalized exponential family member distribution. This is done using a well-known property of the generalized linear model with the canonical link function. Specifically, it demonstrates how one can choose a link function that is related to the distribution of an exponential family member in the TMLE procedure, but can otherwise perform all density estimation using non-parametric methods. This manuscripts also contains one of the first applications of TMLE in a longitudinal setting, and the only application using a count variable as the outcome of interest. In addition, the extensive simulation study presented in this manuscript provides compelling evidence of the benefits of TMLE and a practical demonstration of its theoretical properties.

Targeted Maximum Likelihood Estimation for Marginal Time-Dependent Treatment Effects under Density Misspecification

Mireille E Schnitzer, Erica E M Moodie, and Robert W Platt

Department of Epidemiology, Biostatistics, & Occupational Health, McGill University, Montréal, Québec, Canada

Paper in press in *Biostatistics*, July 2012, DOI: KXS024.

Abstract. Targeted maximum likelihood methods have been proposed to estimate treatment effects for longitudinal data in the presence of time-dependent confounders. This class of methods has been mathematically proven to be doubly robust and to optimize the asymptotic estimating efficiency among the class of regular, semiparametric estimators when all estimated density components are correctly specified. We show that methods previously proposed to build a one-step estimator with a logistic loss function generalize to a generalized linear loss function, and so may be applied naturally to an outcome that can be described by any exponential family member. We evaluate several methods for estimating unstructured marginal treatment effects for data with two time intervals in a simulation study, showing that these estimators have competitively low bias and variance in an array of misspecified situations, and can be made to perform well under near-positivity violations. We apply the methods to the PROmotion of Breastfeeding Intervention Trial data, demonstrating that longer term breastfeeding can protect infants from gastrointestinal infection.

4.1 Introduction

In medical and public health settings, the marginal effect of a treatment, intervention or exposure can often be of interest. In the presence of time-dependent confounders that are affected by previous treatment, it becomes necessary to adjust for this confounding in order to produce an unbiased estimate of the desired causal parameter. The traditional approaches of either including the time-dependent confounders in a conditional model or ignoring them altogether both lead to biased estimation [43]. Unbiased estimates of marginal treatment effects can be produced, for example, using inverse probability weighted estimators [46, 21], a type of estimating equation, and G-computation [44], a maximum-likelihood approach. Both of these methods consistently estimate marginal effects under correct specification of certain components of the underlying density.

Under near-positivity violations (or sparse data), inverse probability weighting is notorious for producing very large weights, leading to highly variable effect estimators. Methods exist to control the size of these weights [2, 85], the most popular of which are simple, but ad hoc [8, 84] and arbitrarily trade a small increase in bias for what are typically large gains in efficiency.

G-computation is a type of substitution (or plug-in) estimator. Substitution estimators use an estimate of the underlying data-generating density in order to make inference about a parameter [5].

Targeted maximum likelihood estimation (TMLE) yields a new class of substitution estimators that offer several attractive properties. The procedure modifies the underlying density estimate in a specific way to produce asymptotic efficiency of

estimation in a class of semiparametric estimators of the parameter of interest, and to potentially reduce the bias arising from partially misspecified density models. We use TMLE to refer to either targeted maximum likelihood estimation or estimator where it is possible to do so without ambiguity.

In this paper, we demonstrate the construction of the longitudinal TMLE for two time intervals for an outcome that can be ascribed a loss function associated with a generalized exponential family. In Section 4.2, we summarize the framework and properties of TMLE. In Section 4.3, we extend the methodology of [53] for two time points so as to incorporate this type of loss function for the final time point. To evaluate the proposed benefits of this TMLE method for the estimation of a marginal expected outcome under given exposure, we conduct a simulation study in Section 4.4, comparing commonly used methods under various types of misspecification of the models for the underlying data-generating mechanism. Finally, in Section 4.5, we apply the methods to study the effect of breastfeeding on the number of gastrointestinal tract infections in infants.

4.2 Background: targeted estimation

An "influence function" governs the asymptotic behaviour of an asymptotically linear estimator. The efficient influence function is the influence function that has minimal variance in the class of regular semiparametric estimators of a given parameter. Therefore, inference conducted with an estimator associated with the efficient influence function will be optimally efficient in this class [71]. Certain types of efficient estimators in causal inference will also inherit the double robustness property,

only requiring that one of the exposure or outcome models is correctly specified in order to guarantee consistent estimation [25].

TMLE is characterized by modifications of the density estimates used in the plug-in estimation. This is carried out so that the resulting estimate is the root of the efficient influence function. This procedure results in doubly robust, locally efficient estimators and can be applied to all path-wise differentiable parameters. The estimators respect the constraints of the parameter of interest (as they are substitution estimators) and, unlike efficient estimating equations, are guaranteed to produce only one solution [73].

We reiterate the steps of the general TMLE procedure. Suppose that we observe identically and independently distributed data from n subjects. The TMLE is constructed by first obtaining an initial data density or likelihood fit, p_n^0. This estimate is then updated by creating a fluctuation $p_n^0(\epsilon)$ of the original density fit, parametrized by ϵ, ensuring that 1) $p_n^0(\epsilon)$ is equal to p_n^0 when $\epsilon = 0$, and 2) given a loss function, the score of the fluctuation function (with parameter ϵ) linearly spans the efficient influence function of the target parameter. Then, given a fixed initial density estimate p_n^0 and a loss function $\mathcal{L}$, the submodel loss is minimized with respect to ϵ so that $\hat{\epsilon} = \text{argmin}_\epsilon \frac{1}{n} \sum_{i=1}^n \mathcal{L}\{p_n^0(\epsilon)\}$ (an example of this would be likelihood maximization where the negative log-likelihood is used as the loss function). The estimate $\hat{\epsilon}$ is then plugged into the fluctuation function to define the updated density p_n^1. These updates are iterated until convergence (although the examples given in [51] specifically demonstrate useful applications of TMLE when only one update is needed for convergence). Finally, the resulting density estimate is used to

calculate an estimate of the targeted parameter. As shown in [78], if this procedure converges, it results in an estimator with the appealing properties described above.

For longitudinal data with binary intermediate variables, the TMLE was demonstrated with logistic working models with a binary outcome by [53], and for a survival outcome by [69]. Practical benefits of TMLE procedures for a point-source treatment (competitive mean-squared errors and robustness to misspecification and positivity violations) have been observed in simulation (e.g., [18, 38, 69]). However, the performance of the longitudinal TMLE has yet to be compared with prevailing longitudinal methods (in particular, the efficient and doubly robust estimator proposed by [1]). In addition, its construction has not yet been demonstrated with the incorporation of a generalized linear loss function.

4.3 Construction

4.3.1 Data and efficient influence function

We consider data with a longitudinal structure. Each subject i contributes an observation of the form $O = (L_0, A_0, L_1, A_1, Y)$ where A_t is a binary covariate at time t indicating whether or not a subject was treated/exposed, and L_1 represents a binary intermediate covariate. Y is the final outcome of interest measured at time $t = 2$.

For this application, we are interested in evaluating the marginal effects of exposure on the final outcome. For now, the parameter of interest is considered to be $\psi_{a_0,a_1} = E(Y_{a_0,a_1})$, the marginal mean of the final outcome under the fixed regime (a_0, a_1). The exposure pattern $(a_0, a_1) = (1, 1)$ would indicate exposure at both time intervals.

The form of the efficient influence function for a fixed regime of $A_0 = a_0$ and $A_1 = a_1$ (as constructed in [73] for any outcome) is the sum of the three components

$$\begin{aligned}
D_0(O) =& E(Y \mid L_0, A_0 = a_0, A_1 = a_1) - \psi, \\
D_1(O) =& \frac{I(A_0 = a_0)}{p(A_0 = a_0 \mid L_0)} \{L_1 - E(L_1 \mid L_0, A_0 = a_0)\} \times \\
& \{E(Y \mid L_0, A_0 = a_0, L_1 = 1, A_1 = a_1) \\
& - E(Y \mid L_0, A_0 = a_0, L_1 = 0, A_1 = a_1)\}, \text{ and} \\
D_2(O) =& \frac{I(A_0 = a_0) I(A_1 = a_1)}{p(A_0 = a_0 \mid L_0) p(A_1 = a_1 \mid L_0, A_0 = a_0, L_1)} \times \\
& \{Y - E(Y \mid L_0, A_0 = a_0, L_1, A_1 = a_1)\}.
\end{aligned}$$

Van der Laan [73] derived this efficient influence function as the projection of the IPTW gradient onto the tangent space orthogonal to the Hilbert space of the nuissance parameter. Each of the components of this efficient influence function corresponds to a mean-zero projection of the gradient in a different dimension of this tangent space. General theory for finding semiparametric efficient estimators and their related influence functions can be found in Tsiatis [71].

4.3.2 Specifying the initial density estimate

The underlying data-generating density must first be estimated. The joint density $p(Y, A_1, L_1, A_0, L_0)$ factors into a product of conditional distributions:

$$p(L_0) p(A_0 \mid L_0) p(L_1 \mid L_0, A_0) p(A_1 \mid L_0, A_0, L_1) p(Y \mid L_0, A_0, L_1, A_1).$$

Each of the conditional components is fit using a model of choice. As a simple example, a logistic regression may be used to fit the density components with binary

outcomes, and a generalized linear model may be used to estimate the conditional density of Y. Finally, $p(L_0)$ is fit with an empirical distribution.

4.3.3 Determining loss functions and clever covariates

The so-called clever covariate method of constructing TMLEs was demonstrated by [78]. The resulting estimators for the effect of an intervention at a single time point are equivalent to those first created by [61]. [52] demonstrated the implementation of generalized linear loss functions in the context of single time point randomized trials, and [18] implemented a logistic loss function for a continuous outcome.

In the longitudinal context developed in [73], the update to the estimate of $E(Y \mid L_0, A_0 = a_0, L_1, A_1 = a_1)$ is made first. The subsequent intermediate variable(s) are then updated backwards through time, each using the updated estimates from the future. This produces a closed form for the fluctuation functions, only requiring one round of updates.

The first clever covariate is defined to update the conditional expected outcome $E_Y^0 = E(Y \mid L_0, A_0, L_1, A_1)$. A generalized linear loss can be expressed as

$$\mathcal{L}_Y(\theta) = -\left\{ \frac{Y\theta - b(\theta)}{a(\eta)} \right\}$$

where $a(\eta)$ is the family-specific dispersion factor that depends on the nuisance parameter η. Note that this is simply the log-likelihood of an exponential family member (minus a term that is independent of θ). In the corresponding density, the mean of Y is $E(Y|\theta) = b'(\theta)$. Let g be the canonical link function such that $g\{E(Y|\theta)\} = \theta$. Using a Gaussian family member, for example, this loss function simplifies to a squared-error loss.

Allow the fluctuation of the conditional mean of the outcome to take the form

$$E_Y^0(\epsilon_1) = g^{-1}\{g(E_Y^0) + C_1\epsilon_1\}. \tag{4.1}$$

This fluctuation produces no update when $\epsilon_1 = 0$, as required.

We wish to determine the form of the clever covariate C_1 in order for the score (derivative of the loss function) to be proportional to the last component of the efficient influence function, D_2. First, we insert the fluctuated mean in Equation (4.1) into the loss function (note that we plug in $\theta = gE_Y^1(\epsilon_1) = g(E_Y^0) + C_1\epsilon_1$) and obtain

$$\mathcal{L}_Y(\epsilon_1) = -\left[\frac{Y\{g(E_Y^0) + C_1\epsilon_1\} - b\{g(E_Y^0) + C_1\epsilon_1\}}{a(\eta)}\right].$$

Then, the loss-based score at zero is

$$\begin{aligned}\left.\frac{d\mathcal{L}_Y(\epsilon_1)}{d\epsilon_1}\right|_{\epsilon_1=0} &= -\left(\frac{C_1}{a(\eta)}[Y - b'\{g(E_Y^0)\}]\right), \\ &= -\left\{\frac{C_1}{a(\eta)}(Y - E_Y^0)\right\}.\end{aligned}$$

Setting

$$C_1 = C_1(L_0, A_0, L_1, A_1) = \frac{I(A_0 = a_0)I(A_1 = a_1)}{p(A_0 = a_0 \mid L_0)p(A_1 = a_1 \mid L_0, A_0 = a_0, L_1)},$$

we have that the score at zero is indeed proportional to the last component of the efficient influence function. Note that the clever covariate takes the form of the IPTW estimate for the mean of the potential outcome under exposure pattern (a_0, a_1). In

particular, a subject who does not follow the fixed exposure pattern would have a zero value for this covariate.

The dispersion factor of a general exponential family may only be dependent on the nuisance parameter, η. Therefore, it need not be estimated as its value will be absorbed into the estimate of the coefficient ϵ_1.

Because minimizing the loss function is equivalent to maximizing a likelihood in this example, the coefficient ϵ_1 can be estimated using a generalized linear model with single covariate $C_1(L_0, A_0, L_1, A_1)$ and no intercept, taking the estimate of $g(E_Y^0)$ as the offset. Once the estimate $\hat{\epsilon}_1$ is obtained, define $E_Y^1 = E_Y^0(\hat{\epsilon}) = g^{-1}\{g(E_Y^0) + C_1(L_0, A_0, L_1, A_1)\hat{\epsilon}_1\}$ as the updated expectation for Y (i.e. plug in the estimates of C_1 and ϵ_1 into Equation (4.1)).

For the update to the intermediate variable fit, the fluctuation function for the conditional density is similarly described by a fluctuation of the probability $p_{L_1}^0 = p^0(L_1 = 1|L_0, A_0)$, that was previously estimated. The fluctuation is given as

$$p_{L_1}^0(\epsilon_2) = \text{logit}^{-1}\{\text{logit}(p_{L_1}^0) + C_2\epsilon_2\}. \tag{4.2}$$

The intermediate variable is binary, and so we can use a logistic loss function to determine the update, which is a special case of the generalized linear loss function with $\theta = \text{logit}(p_{L_1}^0)$ and $b(\theta) = \log(1 - p_{L_1}^0)$. The loss function simplifies to $-\left[L_1 \log p_{L_1}^0 + (1 - L_1)\{1 - \log(1 - p_{L_1}^0)\}\right]$.

The choice for clever covariate C_2 becomes apparent when the score is calculated at $\epsilon_2 = 0$. This score takes the local form $\frac{d\mathcal{L}_{L_1}(\epsilon_2)}{d\epsilon_2} \mid_{\epsilon_2=0} = C_2\{L_1 - E(L_1 \mid L_0, A_0)\}$. In order to have this score equal the efficient influence component $D_1(O)$, C_2 can be

defined as

$$C_2(L_0, A_0) = \frac{I(A_0 = a_0)}{p(A_0 = a_0 \mid L_0)} \times \{E^1(\epsilon_1)(Y \mid L_0, A_0 = a_0, L_1 = 1, A_1 = a_1) - E^1(\epsilon_1)(Y \mid L_0, A_0 = a_0, L_1 = 0, A_1 = a_1)\}.$$

Note that the conditional expectations of Y are calculated under the updated density (using the clever covariate and $\hat{\epsilon}_1$ found in the previous update), and calculated conditional on $L_1 = l_1$ for $l_1 = \{0, 1\}$, $A_0 = a_0$ and $A_1 = a_1$.

The intermediate variable density component $p^0_{L_1}(\epsilon_2)$ is updated as before by minimizing the loss function with respect to ϵ_2. This can be done by fitting a no-intercept logistic regression using $p^0_{L_1}$ as an offset and the estimate of $C_2(L_0, A_0)$ as the lone covariate. The estimate of the coefficient of C_2 is $\hat{\epsilon}_2$, which is then used to update the density. Therefore, let

$$p^1_{L_1} = p^0_{L_1}(\hat{\epsilon}_2) = \text{logit}^{-1}\{\text{logit}(p^0_{L_1}) + C_2(L_0, A_0)\hat{\epsilon}_2\}$$

.

For $D_0(O)$, the development in Section 3 of [51] shows that specifying a particular fluctuation function for the baseline density ($p(L_0)$ in our case) results in no update.

4.3.4 Using the updated density to estimate the final parameter

After estimating the two clever covariates (C_1, C_2) and using maximum likelihood to solve for $\hat{\epsilon} = (\hat{\epsilon}_1, \hat{\epsilon}_2)$, the updated density $p^1_n(\hat{\epsilon})$ is obtained. If these updating steps are used, convergence occurs in the first iteration.

The targeted parameter ψ_{a_0,a_1} can then be calculated using G-computation [15]) with the updated density. For this simple example, the targeted maximum likelihood G-computation is

$$\Psi\{p_n^1(\hat{\epsilon})\} = \frac{1}{n}\sum_{i=1}^{n}\sum_{l=\{0,1\}} E^1(\hat{\epsilon}_1)(Y \mid L_0 = l_0^i, A_0 = a_0, L_1 = l, A_1 = a_1)\times$$
$$p_n^1(\hat{\epsilon})(L_1 = l \mid L_0 = l_0^i, A_0 = a_0).$$

It is important to note here that while the $\hat{\epsilon}$ coefficients are constant, the clever covariates are functions of the observed variables. When conditioning on $A_0 = a_0$, for instance, these values must therefore also be altered in the clever covariates.

4.3.5 Modification for continuous outcome under positivity violations

In a situation where data sparsity leading to near-positivity violations exists, the denominators of the clever covariates may become very small for certain individuals, leading to inflated clever covariate values which may in turn produce unstable mean estimates. This has been shown to be particularly true when using squared loss functions with linear models. [18] have shown that the use of scaling and a logistic-loss function for the update step can result in improved estimation under these conditions. While not an issue in the applied analysis of this paper, this modification proved to be essential in the simulation study where near-positivity violations were produced.

4.4 Simulation

4.4.1 Methods

For this simulation study, four estimators of the parameter $\psi_{1,1}$ were compared: (i) the TMLE, (ii) an untargeted G-computation, (iii) an inverse probability of treatment weighted (IPTW) estimator with stabilized weights, and (iv) Bang and Robins' (2005) doubly robust estimator. We considered parametric models that were correctly specified in accordance with the true data-generating distributions, as well as model specification or data generation that was varied to represent three different types of misspecification that arise in practice: 1) missing confounders, 2) nonlinear dependence on covariates, and 3) data sparsity. In the data sparsity scenario, modified versions of both the TMLE and Bang and Robins' estimator were also used in the estimation. No truncation of inverse weights was used in any of the methods.

For each scenario, we tested the performance of the models for sample sizes of $n = 200$, 1,000 and 10,000. Details of the data generation are given in the Supplementary material (available at *Biostatistics* online). We present measures of the bias, standard error, root mean-squared error, and percent confidence interval coverage. The standard error was calculated using a nonparametric bootstrap. The root mean-squared error was calculated using the squared errors of the difference between the model estimate and the true parameter value for each of the 1,000 generated datasets. The percent coverage refers to the proportion of runs where the 2.5th and 97.5th percentiles of the bootstrap estimates contained the true parameter value.

We also simulated data with a Poisson outcome and model misspecification due to omitted confounders. We refer the interested reader to the Supplementary material (available at *Biostatistics* online) for the Poisson outcome results, as well as those for the Normal outcome with incorrectly-specified nonlinear effects.

4.4.2 Simulation results: omitted confounder

For the first situation, data were generated from a density that had a confounding variable acting at both time points. In order to see how misspecification affects the different models, we demonstrated the effect of omitting the confounder from different parts of the models (the exposure, the outcome or both, as applicable). Each of the three models assumed linear dependence on the covariates.

Table 4–1 presents the results for each of the four estimators under correct specification, or with the misspecification of the exposure, the outcome (and intermediate), or both. A given model was misspecified by omitting the confounder from the model. Since G-computation relies only on outcome specification, misspecifying the outcome was the same as total misspecification. Similarly, inverse probability weighting depends only on the exposure model.

Under correct specification, the TMLE had a magnitude of bias and coverage similar to inverse probability weighting, G-computation and Bang and Robins' estimator. For $n = 200$, 1,000 and 10,000, both TMLE and Bang and Robins' estimator consistently had higher standard errors. When the exposure model was misspecified, only the IPTW estimator exhibited a large degree of bias for all values of n. For the misspecified outcome model, G-computation suffered the most with large asymptotic bias. Bang and Robins' estimator seemed to converge slower than the TMLE and

Table 4–1: Simulation results for various omitted confounder scenarios. Each estimate is calculated over 1,000 datasets. The true value of the parameter is $\psi_{1,1} = 4.35$.

	Correct Specification				**Misspecified Exposure**			
	% Bias	SE*	rMSE	% Cover†	% Bias	SE	rMSE	% Cover
$n = 200$								
TMLE	-13	27	27	94	-22	25	25	95
G-COMP	-10	23	23	95	-10	23	23	95
IPTW	-2	25	25	94	440	23	30	84
BR	-12	33	28	95	-11	28	27	95
$n = 1,000$								
TMLE	-12	12	12	93	-11	11	11	93
G-COMP	-14	10	10	94	-14	10	10	94
IPTW	-14	11	11	93	430	10	21	53
BR	-13	12	12	94	-11	12	12	94
$n = 10,000$								
TMLE	-3	4	4	93	-3	4	4	92
G-COMP	-2	3	3	94	-2	3	3	94
IPTW	-3	3	4	94	443	3	20	0
BR	-3	4	4	92	-4	4	4	93
	Misspecified Outcome				**Total Misspecification**			
	% Bias	SE	rMSE	% Cover	% Bias	SE	rMSE	% Cover
$n = 200$								
TMLE	-10	27	27	93	436	25	31	86
G-COMP	-437	23	30	84	-437	23	30	84
IPTW	-2	25	25	94	440	23	30	84
BR	-20	28	27	94	437	26	32	87
$n = 1,000$								
TMLE	-12	12	12	93	437	11	22	59
G-COMP	426	10	21	54	426	10	21	54
IPTW	-14	11	11	93	430	10	21	53
BR	-12	12	12	93	438	11	22	58
$n = 10,000$								
TMLE	-3	4	4	93	451	4	20	0
G-COMP	440	3	19	0	440	3	19	0
IPTW	-3	3	4	94	443	3	20	0
BR	-3	4	4	93	448	4	20	0

All values except for coverage given as $\times 10^2$ the original value.
SE, The nonparametric bootstrap standard error is computed using 200 resamples for each drawn dataset of size n, the mean of the SE is then taken over the 1,000 generated datasets; †Cover, The coverage is by bootstrap 2.5th and 97.5th percentiles.
rMSE, root-mean-squared error; TMLE, targeted maximum likelihood estimation; G-COMP, G-computation; IPTW, inverse probability of treatment weighting; BR, Bang and Robins' estimator.

the IPTW estimator. When both models were misspecified all four models were similarly biased.

4.4.3 Simulation results: data sparsity

We considered two different levels of data sparsity leading to near-positivity violations resulting from heavy dependence on the covariates in the treatment decision. The TMLE method with a Gaussian loss function (i.e. minimizing mean-squared error) was expected to perform very poorly in this case (biased estimation in addition to inflated variance) so we also tested the abilities of the TMLE using a logistic loss function for a continuous outcome. To do so, we correctly specified the outcome density using a linear regression model. We then shifted and scaled the prediction from this model,

$$E(Y^*) = \{E(Y) - \min(Y)\}/\{\max(Y) - \min(Y)\}$$

and used a logistic loss function for the update step, following the approach of [18] for unbounded data generation. Bang and Robins' estimator was also given added robustness to the data sparsity by scaling the outcome to $[0, 1]$ and predicting the outcome using logistic regression rather than linear regression (and thereby misspecifying the estimation of the conditional outcome density).

The results of the model fits on the data with near-positivity violations are recorded in Table 4–2. As anticipated, the TMLE procedure with a squared error link did very poorly for small samples in particular, as did the Bang and Robins' estimator with linear outcome model. In the severe case, they both performed very poorly for all values of n. G-computation also performed as expected, producing the

best inference by far over all sample sizes, with negligible bias and ideal coverage. The logistic TMLE at smaller samples was less sensitive to data sparsity than the logistic Bang and Robins' estimator, resulting in lower bias and mean-squared error. The stabilized IPTW performed worse than logistic TMLE in terms of bias, but had lower standard errors, resulting in lower mean-squared error. At $n = 200$, extreme data sparsity in a small percentage of generated datasets produced inflated standard error estimates for the linear TMLE and Bang and Robins' estimator, creating an inflated and unreliable mean standard error. For instance, with mild data sparsity at $n = 200$, Bang and Robins' estimator's mean bootstrap-estimated standard error was approximately 21, but the median was 7 (and similarly for the TMLE).

4.4.4 Other results

Data with a normal outcome and nonlinear dependence on covariates were generated and then modeled incorrectly using only linear terms. The targeted maximum likelihood estimator maintained low bias for misspecification of either data density component, and generally performed similar to Bang and Robins' estimator. When data were generated with a Poisson outcome and a similar omitted confounder scenario was evaluated, the results closely reflected those from the omitted confounder scenario in Section 4.4.2. Additional details regarding the full simulation study are available in the Supplementary material (available at *Biostatistics* online).

4.5 Example

4.5.1 The PROBIT Trial

The PROmotion of Breastfeeding Intervention Trial (PROBIT) was a cluster-randomized trial that introduced the WHO/UNICEF Baby Friendly Initiative, a

Table 4–2: Simulation results for two levels of near-positivity violations. Each estimate is calculated over 1,000 datasets. The true value of the parameter is $\psi_{1,1} = 4.338$.

	Mild Data Sparsity				**Severe Data Sparsity**			
	% Bias	SE*	rMSE	% Cover†	% Bias	SE	rMSE	% Cover
$n = 200$								
TMLE	233	3,110	441	93	-	-	-	-§
TMLE_{log}	114	91	95	93	18	83	99	93
G-COMP	13	39	39	94	9	31	31	95
IPTW	113	49	57	92	91	41	52	92
BR	1,128	2,119	913	96	-	-	-	-§
BR_{log}	196	112	114	95	85	97	112	95
$n = 1,000$								
TMLE	20	76	68	93	857	20,131	8,708	93
TMLE_{log}	-8	40	43	93	-52	49	56	92
G-COMP	3	18	18	94	-4	14	14	95
IPTW	11	26	33	92	63	25	37	91
BR	-79	100	103	92	-4,068	868	4,182	93
BR_{log}	-1	47	49	92	-58	56	63	93
$n = 10,000$								
TMLE	-7	15	15	92	-5,809	1,611	8,427	94
TMLE_{log}	-3	13	14	93	-1	23	26	94
G-COMP	-1	6	6	94	-3	4	4	94
IPTW	-12	10	11	92	11	12	19	91
BR	-9	17	18	93	-175	144	212	93
BR_{log}	1	14	15	93	24	25	27	93

All values given as $\times 10^2$ the original value.

*SE, The nonparametric bootstrap standard error is computed using 200 resamples for each drawn dataset of size n, the mean of the SE is then taken over the 1,000 generated datasets; †Cover, The coverage is by bootstrap 2.5th and 97.5th percentiles.

rMSE, root-mean-squared error; TMLE, targeted maximum likelihood estimator; TMLE_{log}, targeted maximum likelihood estimator with logistic model for outcome; G-COMP, G-computation; IPTW, inverse probability of treatment weighting; BR, Bang and Robins' estimator; BR_{log}, Bang and Robins' estimator with logistic model for outcome. §Model results in extremely high bias and standard error.

breastfeeding promotion program, to selected hospitals in the republic of Belarus [27]. The purpose of the trial was to evaluate the effect of the intervention on health outcomes including gastrointestinal tract infection. Healthy, full-term, singleton breastfed infants of mothers who intended to breastfeed (n = 17,044) weighing $\geq$2,500 g were enrolled soon after birth and followed up at 1, 2, 3, 6, 9, and 12 months of age for various measures of parental behaviours, size and health, including number of gastrointestinal infections over each time interval.

We perform a simplified analysis using the data of the form $O = (L_0, A_0, L_1, A_1, Y)$. The variable L_0 is a vector-valued covariate containing all suspected baseline confounders of the duration of breastfeeding and infection. The exposures A_0 and A_1 indicate whether the mother is still breastfeeding at 3 months and at 6 months, respectively. The intermediate variable L_1 is whether the infant had an infection between 3 and 6 months. The outcome variable Y is the number of infections counted between 6 and 12 months.

We wish to examine whether the duration of breastfeeding has an effect on the number of gastrointestinal tract infections reported between 6 and 12 months. A potentially important binary intermediate variable is whether or not the infant had any infections between 3 and 6 months of age. Observed baseline confounders are: mother's education, mother's smoking status, mother's age, family history of allergy, number of previous children, whether the birth was by cesarean section, gender of child, gestational age, Apgar score for health of the newborn, presence of infection before 3 months, geographic region, and the weight, height and head circumference at birth.

Only three regimes (a_0, a_1) exist because breastfeeding is generally a monotone process. We are interested in three different effects: 1) the effect of breastfeeding up to 3 months (but not until 6 months) versus not attaining 3 months, 2) the effect of breastfeeding up to 6 months versus stopping between 3 and 6 months, and 3) the effect of breastfeeding up to 6 months versus not attaining 3 months. These effects correspond to the estimates $\delta_1 = \psi_{1,0} - \psi_{0,0}$, $\delta_2 = \psi_{1,1} - \psi_{1,0}$, and $\delta_3 = \psi_{1,1} - \psi_{0,0}$ where $\psi_{a_0,a_1} = E(Y_{a_0,a_1})$ is the marginal mean outcome under specified treatment.

We used generalized linear models to fit all density components for each method. The TMLE was implemented with a Poisson loss function, corresponding with the count outcome, and both the G-computation and Bang and Robins' estimator were fit using Poisson distributions to model the mean outcome.

4.5.2 Results

Out of the 17,044 enrollments, 15,642 (92%) had complete data for the two time intervals. For simplicity, we performed a complete case analysis, discarding the 8% of observations with missing data. Characteristics of the cohort including missing data summaries are provided in Table 4–3. Most notably, the infection counts are very low, with only 828 (4.9% of the full cohort) with one infection, 56 (0.3%) with two infections, 3 (0%) with three infections, and 1 (0%) with six infections between 6 and 12 months.

Three of the four methods use some variety of inverse probability weights (G-computation being the exception). Using histograms and univariate summaries, each set of calculated weights was examined in order to assess whether the positivity assumption was violated. No excessively large weights were noted.

Table 4–3: Characteristics of the 17,044 mother-infant pairs in the PROBIT dataset.

Characteristic	Observations		Missing	
	N	%	N	%
Median age of mother (years)	23	(21,27)*		
Male child	8,827	52		
Median gestational age (months)	40	(39,40)		
Cesarean	1,974	12		
Median infant weight (kg)	3.4	(3.2,3.7)		
Median Apgar score†	9	(8,9)	5	0.0
(A_0) Breastfed at 3 months	11,101	65		
(A_1) Breastfed at 6 months	7,176	42		
(L_1) Infection by 3 months	593	3·5	1,087	6.4
(Y) Infection count between 6 and 12 months			1396	8.2
1 infection	828	4.9		
2 infections	56	0.3		
3 infections	3	0.0		
6 infections	1	0.0		

*For numeric variables, the inter-quartile range is given.

†The Apgar score is an assessment of newborn health (range 1–10) where 8+ is vigorous, 5–7 is mildly depressed and 4 or less is severely depressed [13]. We observed a range of (5–10) due to entry restrictions on weight.

The nonparametric bootstrap was used to estimate the standard error and confidence intervals for each estimator. This was accomplished by taking 200 resamples with replacement of sample size 15,642. The endpoints of the bootstrap 95% confidence intervals were the 2.5th and 97.5th quantiles of the bootstrap resampled estimates.

Table 4–4 shows the estimates, standard errors and confidence intervals for each method applied to each of the three parameters of interest. The estimates for the effects of breastfeeding exposure are all negative, indicating that breastfeeding has a preventative impact on gastrointestinal infection. The effect of breastfeeding up to 3 months (compared with not reaching 3 months) was not found to be significantly different from zero at the 95% confidence level by all methods. The effect of breastfeeding for 6 months compared with feeding for at least 3 months but less than 6 was estimated at -0.019 by TMLE (the largest estimate by magnitude). This means that the expected number infection counts is decreased by 0.019 if an infant's breastfeeding is extended up to 6 months. Here, all methods agreed on the direction and significance of the difference. The third parameter is the effect of breastfeeding for over 6 months, compared to less than three months. This effect is the strongest (it is the sum of the previous two parameters), and is found to be significant by all methods. The TMLE results suggest that breastfeeding for 6 months when compared with breastfeeding for fewer than three months decreases the expected number of gastrointestinal infections experienced between 9 and 12 months of age by 0.026. This effect estimate corresponds with a number needed to treat (NNT) of 38.

Table 4–4: Breastfeeding effect estimates at 3 and 6 months for each model.

Model	Estimate	SE*	95% CI†
Effect of breastfeeding for 3 months vs. <3 months			
TMLE	-7	6	(-17,5)
G-comp	-9	6	(-21,6)
IPTW	-7	5	(-18,5)
BR	-8	7	(-19,5)
Effect of breastfeeding for 6+ months vs. 3 months			
TMLE	-19	6	(-31,-8)
G-comp	-16	5	(-26,-7)
IPTW	-15	5	(-24,-7)
BR	-17	6	(-27,-5)
Effect of breastfeeding for 6+ months vs. <3 months			
TMLE	-26	5	(-33,-17)
G-comp	-25	5	(-33,-15)
IPTW	-22	4	(-30,-14)
BR	-25	5	(-33,-14)

All values given as $\times 10^3$ the original value.
*SE: The bootstrap standard error was computed using 200 resamples from the data set of size n=15 642.
†The estimated confidence interval is the interval between the 2.5th and 97.5th bootstrap percentiles.
TMLE, targeted maximum likelihood estimator; G-COMP, G-computation; IPTW: inverse probability of treatment weighting; BR, Bang and Robins' estimator.

The TMLE estimated an expected infections count of 0.072 for infants breastfed less than 3 months. A reduction in the expected number of infections of 0.007 for infants breastfed for 3-6 months compared with those breastfed for less than 3 months therefore corresponded with a 10% reduction. The mean count for 3-6 months of breastfeeding was estimated at 0.066, so the risk difference corresponded with an estimated 30% expected reduction. In addition, the expected reduction when comparing <3 months to 6+ months of breastfeeding was estimated at 36%.

4.6 Discussion

In this paper, we have carefully demonstrated the construction of the two time interval TMLE using a generalized linear loss function. While we implemented the method using parametric models, all conditional densities may alternatively be fit through any means desired (including regression and nonparametric methods) while the update step is performed with respect to the chosen generalized linear loss function. We have thus shown how a longitudinal TMLE with a chosen loss function can be fit without resorting to parametric modeling assumptions. We have performed a systematic comparison of the performance of the longitudinal TMLE to competing methods under several challenging data scenarios. In addition, we applied the TMLE methodology to estimate the impact of breastfeeding on gastrointestinal infections. To the best of our knowledge, this is the first application of TMLE to a longitudinal estimation problem with a count outcome.

In our simulation study, TMLE did not produce a reduction in finite-sample bias or variance for correctly specified densities compared with the G-computation substitution estimator. The two doubly robust methods performed comparably in general, but the logistic TMLE proved more stable under near-positivity violations than Bang and Robins' estimator with the logistic outcome. However, we were able to stabilize the TMLE without misspecifying the underlying densities (which was not true for Bang and Robins' estimator).

[38] contributed to the debate initiated by [25] by demonstrating that certain versions of the TMLE procedure for continuous outcomes can control bias better than traditional doubly robust methods in such cases as near-positivity violations

and model misspecification. We show correspondingly that this TMLE performs very well in the two time intervals case and does no worse than two non-doubly robust methods under dual misspecification. With a small variation in implementation, TMLE can also be made to be stable under near-positivity violations.

We also acknowledge the very recent development of a new TMLE for longitudinal data by [74], which is an extension of the [1] estimator. This method is computationally efficient, requires fewer data structure constraints than the longitudinal TMLE that we have evaluated, and can also be implemented to respect the global bounds of the parameter of interest. Future work will involve extensive comparisons of this alternative TMLE.

4.7 Supplementary material for Manuscript 1: Simulation details

4.7.1 Comment on estimation of the standard error by bootstrap

In our simulation study, we generated data that came close to violating the practical positivity assumption required when using inverse probability of treatment (IPTW), targeted maximum likelihood estimation (TMLE), and Bang and Robins' estimator (to varying degrees of sensitivity). In several scenarios, the bootstrap resamples created resampled datasets that had more extreme levels of data sparsity. The estimates of these resampled datasets (and subsequently the bootstrapped standard error estimates) were subsequently inflated.

For a data analysis problem, however, it is important to diagnose data sparsity in the bootstrap resamples, even if they don't exist in the full data. And it is best to avoid bootstrap resampling in the face of any data sparsity. Other methods

to estimate the standard error exist, including a large-sample estimate using the influence curve [76], but we did not utilize them in this study.

4.7.2 Normal outcome, omitted confounder

The data were generated in the form (U,L_0,A_0,L_1,A_1,Y) so that, as before, A_t was treatment at time t. The confounding variable U is a Bernoulli random variable, taking the value of one with a probability of 1/2. The baseline variable L was generated as a Normal variate with mean 1 and variance 1/16. At $t = 0$, treatment was obtained with probability

$$p_{a_0} = \text{logit}^{-1}(1/10L_0 + U).$$

The intermediate binary variable L_1 was one with probability

$$p_{l_1} = \text{logit}^{-1}(1/2 - 2L_0 + 1/2A_0 + 2U).$$

The second treatment, at $t = 1$, was obtained with probability

$$p_{a_1} = \text{logit}^{-1}(1/2 - L_0 + 1/10A_0 + 1/10L_1 + U),$$

and the final outcome was generated as

$$Y = 3 + 1/10A_0 + 1/2A_1 + 1/2L_1 + U + \xi$$

where $\xi \sim \mathcal{N}(0, 4)$ is noise.

4.7.3 Normal outcome, nonlinearity

In this situation, we created four different types of data scenarios, and tested the four models. The scenarios were: linearity, nonlinearity in the exposure, nonlinearity in the outcome (and intermediate), and nonlinearity in both the exposure and outcome (referred to as total nonlinearity). All of the models that were fit assumed linearity in the covariates, so they were misspecified when any nonlinearity was present.

The basic linear data generation for this situation was specified as follows:

$$
\begin{aligned}
L_0 &\sim \mathcal{N}(1, 1/16) \\
A_0 &\sim \text{Bernoulli}\{p_{a_0} = \text{logit}^{-1}(1/2L_0)\} \\
L_1 &\sim \text{Bernoulli}\{p_{l_1} = \text{logit}^{-1}(1/2 - 2L_0 + 1/2A_0)\} \\
A_1 &\sim \text{Bernoulli}\{p_{a_1} = \text{logit}^{-1}(1/2 - 4/5L_0 + 1/10A_0 + 1/10L_1)\} \\
Y &= 1/2 + L_0 + 1/2A_0 + 3/2A_1 + 3L_1 + \xi
\end{aligned}
$$

where $\xi \sim \mathcal{N}(0, 4)$ is random error.

Nonlinear misspecification of the exposure was imposed by changing p_{a_0} to

$$\tilde{p}_{a_0} = \text{logit}^{-1}\{\cos(L_0 + 3/5)^3\}$$

and p_{a_1} to

$$\tilde{p}_{a_1} = \text{logit}^{-1}[2/5(1 - A_0)\cos\{1/6(L_0 + 2)^2\} + 2/5A_0 \sin L^2].$$

The trigonometric functions were used to induce polynomial-type slopes on the usual range of the baseline covariate while keeping the range of probabilities bounded in

order to avoid positivity violations. The other variables (baseline, intermediate and outcome) were generated linearly (as in Situation 1).

Nonlinear misspecification of the intermediate and outcome variables was imposed by changing p_{l_0} to

$$\tilde{p}_{l_0} = \text{logit}^{-1}\{1 + 1/2A_0L_0^2 + 1/2A_0 - \exp(L_0/3)\}$$

and the outcome generation to

$$Y = -6/5 + 9/2L_1L_0^2 + A_0L_0^2 + 2A_1L_1 + \xi.$$

The remaining variables (baseline and treatments) were generated in the same way as in the linear scenario.

Finally, the model was totally misspecified when both the exposure, the intermediate and outcome variables were all generated as nonlinear in their covariates. We combined all of the misspecifications described above for this scenario.

The results in Table 4–5 provides the estimates, standard errors, root-mean-squared errors and percent coverage for each model in the various data-generating scenarios. As in the previous situations, both the TMLE and Bang and Robins' estimator appeared to have larger small-sample standard errors. For the linear (correctly specified) data generation, all four models performed comparably in terms of bias and coverage over various values of n. Generating treatment using nonlinear models did not substantially affect estimation, although the IPTW estimator remained biased for $n = 10,000$. When the outcome and intermediate variables were generated nonlinearly, G-computation was the most affected, resulting in large bias

and coverage which decreased with increasing n. Bang and Robins' estimator and TMLE both performed well. The IPTW estimator had smaller standard errors, but remained unbiased for all values of n. When all of exposure, intermediate and outcome were generated nonlinearly, inverse propensity weighting produced much larger bias and coverage was sometimes as low as 0%. Targeted maximum likelihood and G-computation remained biased and behaved similarly. Bang and Robins' estimation had half the bias for $n = 1,000$ and 10,000, and had consistently high coverage over the different sample sizes.

4.7.4 Normal outcome, data sparsity

The baseline, intermediate and outcome variables were generated as in the linear specification in the nonlinear case. For mild data sparsity, the exposure status was generated by two Bernoulli variables with means

$$p_{a_0} = \text{logit}^{-1}(-5 + 4L_0) \quad \text{and}$$
$$p_{a_1} = \text{logit}^{-1}(-6 + 4L_0 + A_0 + 2L_1),$$

respectively. For severe data sparsity, the means were changed to

$$p_{a_0} = \text{logit}^{-1}(-8 + 8L_0) \quad \text{and}$$
$$p_{a_1} = \text{logit}^{-1}(-10 + 8L_0 + 2A_0 + 4L_1).$$

4.7.5 Poisson outcome, omitted confounder

We evaluated the performance of the TMLE with a Poisson loss function for a Poisson-generated outcome under unmeasured confounding. As in the omitted confounder case, the exposure, intermediate and outcome variables were generated

conditional on a confounder that was sometimes not included in the model. The confounder was omitted from different parts of estimation to misspecify the exposure models, outcome models, or all components.

The data for this simulation were generated in the same way as described in the scenario for an omitted confounder with a Gaussian outcome, except the outcome was generated as a random draw from a Poisson distribution with mean

$$\lambda_Y = 3 + 1/10A_0 + 1/2A_1 + 1/2L_1 + U,$$

where, as before, U is the sometimes-omitted confounder. In the correct specification scenario, both exposure and outcome models (when used in the estimator) correctly include the variable U. Misspecified exposure means that U was incorrectly omitted from the exposure model, misspecified outcome means it was omitted from both the intermediate and outcome models, and the scenario "total misspecification" indicates that U was not used at all in the estimation.

The results of this simulation scenario are displayed in Table 4–6. Under correct specification, TMLE competed very closely with the other estimators. When the exposure was misspecified, only inverse probability weighting was highly biased, with substantially greater standard errors as well. Under a misspecified outcome model, G-computation is biased and produced higher standard errors. It was clear that the misspecified outcome caused higher standard errors for TMLE as compared to its previous performance. Inverse probability weighting therefore outperformed TMLE in terms of standard error. Both inverse probability of treatment weighting and TMLE were unbiased as expected. Finally, when the confounder was never included

and all models were incorrectly specified, all four methods were biased. The targeted maximum likelihood estimator had slightly less bias and marginally higher standard errors, resulting in better coverage for the smallest sample size and generally better root mean-squared error. The results from Bang and Robins' estimator were nearly identical to those from the TMLE, for all situations and sample sizes. For the correctly specified scenario, the mean value of the bootstrap-estimated standard errors differed unexpectedly from the root-mean square error for Bang and Robins' estimator. Upon investigation, this was found to be due to data sparsity at some levels of the covariates for fewer than 2% of the generated datasets. This data sparsity caused occasional large weights which subsequently led to the unstable bootstrap standard error estimates.

Table 4–5: Simulation results for various scenarios involving nonlinear covariates. Each estimate is calculated over 1,000 datasets.

	Correct Specification				**Nonlinear Exposure**			
	$\psi_{1,1} = 4.338$				$\psi_{1,1} = 4.338$			
	% Bias	SE*	rMSE	% Cover†	% Bias	SE	rMSE	% Cover
$n = 200$								
TMLE	-19	32	33	94	-7	32	32	94
G-COMP	-6	28	28	94	-4	28	28	94
IPTW	-10	28	29	94	1	29	29	94
BR	-20	34	33	95	-5	34	32	95
$n = 1,000$								
TMLE	1	14	14	95	14	14	14	94
G-COMP	1	14	14	95	17	13	13	94
IPTW	-9	13	12	95	27	13	13	94
BR	0	14	14	95	9	14	14	94
$n = 10,000$								
TMLE	5	4	4	94	-2	4	4	95
G-COMP	2	4	4	94	-1	4	4	94
IPTW	2	4	4	94	10	4	4	94
BR	4	5	4	94	-7	4	4	94
	Nonlinear Outcome				**Total Nonlinearity**			
	$\psi_{1,1} = 4.353$				$\psi_{1,1} = 4.353$			
	% Bias	SE	rMSE	% Cover	% Bias	SE	rMSE	% Cover
$n = 200$								
TMLE	-53	49	50	94	60	48	49	94
G-COMP	-279	44	46	93	62	46	46	95
IPTW	-44	28	29	94	3,402	28	150	00
BR	-42	51	50	95	-96	51	50	95
$n = 1,000$								
TMLE	22	22	22	93	108	22	22	92
G-COMP	-236	20	22	92	103	20	21	94
IPTW	-43	13	12	95	3,403	13	149	00
BR	20	22	21	93	-56	22	22	94
$n = 10,000$								
TMLE	3	7	7	94	104	7	8	90
G-COMP	-247	6	12	57	101	6	8	88
IPTW	-32	4	4	92	3414	4	149	00
BR	4	7	7	94	-55	7	7	93

All values given as $\times 10^2$ the original value.

*SE: The nonparametric bootstrap standard error is computed using 200 resamples for each drawn dataset of size n, the mean of the SE is then taken over the 1,000 generated datasets; †Cover: The coverage is by bootstrap 2.5th and 97.5th percentiles.

rMSE, root-mean-squared error; TMLE, targeted maximum likelihood estimator; G-COMP, G-computation; IPTW, inverse probability of treatment weighting; BR, Bang and Robins' estimator.

Table 4–6: Poisson outcome: simulation results for various omitted confounding scenarios. Each estimate is calculated over 1,000 datasets. The true value of the parameter is $\psi_{1,1} = 94.611$.

	Correct Specification				**Misspecified Exposure**			
	% Bias	SE*	rMSE	% Cover†	% Bias	SE	rMSE	% Cover
$n = 200$								
TMLE	-12	419	412	95	-12	419	411	95
G-COMP	-9	416	407	95	-9	416	407	95
IPTW	-2	453	438	95	1,726	557	1,721	17
BR	-11	579	413	95	-13	420	412	95
$n = 1,000$								
TMLE	-10	188	187	94	-10	188	187	94
G-COMP	-10	186	187	94	-10	186	187	94
IPTW	-10	196	196	94	-1,715	248	1,641	00
BR	-10	188	187	94	-10	188	187	94
$n = 10,000$								
TMLE	5	59	62	93	5	59	62	93
G-COMP	6	59	61	93	6	59	61	93
IPTW	6	62	64	93	1,735	78	1,643	00
BR	5	59	62	93	5	59	62	93
	Misspecified Outcome				**Total Misspecification**			
	% Bias	SE	rMSE	% Cover	% Bias	SE	rMSE	% Cover
$n = 200$								
TMLE	3	472	467	95	1,692	561	1,694	18
G-COMP	1,698	545	1,692	17	1,698	545	1,692	17
IPTW	-2	453	438	95	1,726	557	1,721	17
BR	-16	453	426	96	1,697	610	1,701	19
$n = 1,000$								
TMLE	-10	209	208	94	1,675	250	1,605	00
G-COMP	1,687	243	1,614	00	1,687	243	1,614	00
IPTW	-10	196	196	94	1,715	248	1,641	00
BR	-10	192	190	94	1,676	250	1,605	00
$n = 10,000$								
TMLE	6	66	69	92	1,697	79	1,608	00
G-COMP	1,708	77	1,618	00	1,708	77	1,618	00
IPTW	6	62	64	93	1,735	78	1,643	00
BR	5	60	62	93	1,697	79	1,607	00

All values given as $\times 10^2$ the original value.

*SE, The nonparametric bootstrap standard error is computed using 200 resamples for each drawn dataset of size n, the mean of the S.E. is then taken over the 1,000 generated datasets; †Cover, The coverage is by bootstrap 2.5th and 97.5th percentiles.

rMSE, root-mean-squared error; TMLE, targeted maximum likelihood estimator; G-COMP, G-computation; IPTW, inverse probability of treatment weighting; BR, Bang and Robins' estimator.

CHAPTER 5
Effect of Breastfeeding on Gastrointestinal Infection in Infants: A Targeted Maximum Likelihood Approach for Longitudinal Data With Censoring

Preamble to Manuscript 2. Previous applications of TMLE for longitudinal data (e.g. as implemented in Manuscript 1), initially used only in simple contexts, produces computational challenges for more complex data structures [69]. It is based on G-computation which is computationally complex for many time points and intermediate variables [82]. Van der Laan and Gruber [74] therefore produced a computationally simpler TMLE method for longitudinal data. In this manuscript, the theoretical background for this estimator is described, and an implementation of their method is presented in the context of a case study. This is the first demonstration of this longitudinal method in an applied data analysis, where it is readily implemented in a dataset with six time points and censored observations. In addition, the simulation study in this manuscript showcases the robustness of this TMLE method under near-positivity violations where it outperforms a related efficient estimating equation approach.

Effect of Breastfeeding on Gastric Infection in Infants: A Targeted Maximum Likelihood Approach for Longitudinal Data With Censoring

Mireille E Schnitzer*, Mark J van der Laan†, Erica E M Moodie*, and Robert W Platt*

**Department of Epidemiology, Biostatistics, & Occupational Health, McGill University, Montréal, Québec, Canada*

†Division of Biostatistics, School of Public Health, University of California, Berkeley, USA

Submitted July 2012

Abstract. The PROmotion of Breastfeeding Intervention Trial (PROBIT) randomized a program encouraging breastfeeding to new mothers. The original studies indicated that this intervention successfully increased duration of breastfeeding, and lowered rates of gastrointestinal tract infections in newborns. Additional scientific interest lies in determining the causal effect of extending breastfeeding duration on the number of gastrointestinal infections. In this study, we estimate the marginal exposure-specific mean infection count for various lengths of breastfeeding. We demonstrate the method of Targeted Maximum Likelihood Estimation (TMLE) for time-dependent exposure-specific means in the context of this application. We compare this method (implemented both parametrically and using a data-adaptive algorithm) to other causal methods for this situational example. In addition, a simulation study was conducted with data generation structurally similar to the PROBIT example. We varied the specification of the data generation to demonstrate the abilities of this TMLE method under several scenarios, including unmeasured confounding and near positivity violations. TMLE was compared to G-computation, inverse probability of treatment weighting, and efficient estimating equations.

5.1 Introduction

The PROmotion of Breastfeeding Intervention Trial (PROBIT) [27, 28] was undertaken in order to obtain evidence from a randomized trial of the effect of longer duration of breastfeeding on infection in newborns. This was done by randomizing an intervention that supported breastfeeding by encouraging exclusivity and duration. In the PROBIT study, the relationship between this breastfeeding intervention and gastrointestinal tract infection was originally evaluated using an intention-to-treat analysis with results indicating a significant reduction in infection incidence for infants whose mothers had been assigned to the intervention group [27]. While it's reasonable to assume that the effect of intervention was due to breastfeeding, the estimated effect is clearly biased due to the intent-to-treat analysis and "noncompliance" in the study.

However, real scientific interest also lies in the causal effect of breastfeeding on gastrointestinal infection. One of the challenges involved in determining this effect is the confounding effect of infection, the presence of which may be associated with future discontinuation of breastfeeding. Since the probability of the presence of infection at each time point might also be reduced through breastfeeding, presence of infection can be described as a potential *time-dependent confounder*. As infection is also hypothesized to be affected by previous breastfeeding status, standard regression methods (including or excluding the time-dependent confounder) may produce a biased estimate of the causal parameter [43]. Causal methods are therefore required to isolate the desired effect, which is also obscured by multivariate baseline confounding, and by participant dropout.

Modeling for longitudinal data that takes into account time-dependent confounders predicted by past exposure often involves weighting methods, such as inverse-probability-of-treatment weighting for marginal structural models [21, 46]. These weighting methods can be inefficient and often unstable, and their shortcomings have since spurred the development of new estimators with better properties. Estimating equation methodology using the efficient influence curve for the parameter of interest [76, 71] produces estimators that are doubly robust (only requiring the correct specification of the exposure models or the outcome models for unbiased estimation) and regular semiparametric efficient when correctly specified. Targeted Maximum Likelihood Estimation (TMLE) [78] rivals efficient estimating equation methods as it inherits the stability and boundedness properties (respect for global constraints) of substitution estimation as well as the double robustness and efficiency that comes from estimation using the efficient influence curve.

[73] established a method of targeted estimation for longitudinal data based on a formulation of the efficient influence curve that relies on a binary decomposition of the intermediate variables (the time-dependent confounders). This method has been described and implemented by [53], [62] and [7] for two time points, and [69] for a survival outcome. However, the implementation of this method for large numbers of time points results in heavy computational requirements and a restriction on the form of the data. More recently, [74] revisited an alternative decomposition of the efficient influence curve, first proposed by [1] that allows for a more flexible and simpler implementation of TMLE for longitudinal data.

This paper demonstrates a sequential implementation of the TMLE procedure first proposed in [74] for estimation of a marginal treatment effect for longitudinal data. The sequential TMLE approach is used to estimate the causal effect of breastfeeding duration on gastrointestinal tract infection in infants using data from the PROBIT. In addition, we compare the sequential TMLE approach to other causal techniques for longitudinal data in a simulation study.

5.2 The PROBIT data

In the PROBIT, healthy, full-term, singleton infants of mothers who intended to breastfeed, weighing at least 2500g, were enrolled soon after birth and followed up at 1, 2, 3, 6, 9, and 12 months of age for various measures of health and size, including number of gastrointestinal infections over each time-interval. At each follow-up visit, it was established whether the mother continued to breastfeed.

17,044 mother/infant pairs were recruited into the trial. Of these, eight were missing some necessary baseline information, and were removed from the analysis. The remaining 17,036 subject pairs were used in the analysis. Characteristics of the complete dataset (including missing data summaries) are presented in Table 5–1.

Measured baseline potential confounders of the effect of breastfeeding on infection (and predictors of outcome) were chosen to be mother's education, mother's smoking status during pregnancy, mother's age, family history of allergy, number of previous children, whether the birth was by cesarean section, gender of child, gestational age, Apgar score for health of the newborn, geographic region, and the weight, height and head circumference at birth.

Table 5–1: Characteristics at baseline of the 17,044 mother-infant pairs in the PROBIT dataset.

Characteristic	Summary		N. Missing
Numeric variables	Median	IQR[a]	
Age of mother (years)	23	(21,27)	
N. previous children	0	(0,1)	
Gestational age (months)	40	(39,40)	
Infant weight (kg)	3.4	(3.2,3.7)	
Apgar score[b]	9	(8,9)	5
Head length (cm)	35	(34,36)	3
Binary variables	N.	%	
Smoked during pregnancy	389	2.28	
History of allergy	750	4.40	
Male child	8827	52	
Cesarean	1974	12	

NOTE: [b]The Apgar score is an assessment of newborn health (range 1-10) where 8+ is vigorous, 5-7 is mildly depressed and 4- is severely depressed [13]. A range of 5-10 was observed in PROBIT due to entry restrictions on weight and health at baseline.

[a]IQR: inter-quartile range.

Exposure at a given time point was whether the child had been breastfed up until that time. The binary intermediate variable at a given time was whether or not gastrointestinal infection occurred in the interval immediately preceding the time point. The outcome is the total number of infections occurring up until 12 months of age.

Aside from a minimal amount of missing baseline information, information was lost due to participant dropout, which was observed to occur at various times after the baseline visit. The number of censored subjects at each time point is described in Table 5–2. Mothers may have left the study for individual reasons that depended on subject-specific characteristics, health and experience.

Table 5–2: Censoring, number of infections and mothers still breastfeeding by time point

Time point	1	2	3	4	5	6
Month	1	2	3	6	9	12
N. censored	284	500	326	491	717	139
Cumulative N.	284	784	1110	1601	2318	2457
Cumulative N. %	1.66	4.60	6.52	9.40	13.61	14.42
N. with infections	171	232	230	443	518	408
N. of infections	173	235	236	472	544	439
N. breastfeeding	15,392	13,128	10,765	6,893	4,717	-

At each visit, the number of gastrointestinal infections since the last visit were counted. In addition, breastfeeding status at that time was obtained. There is therefore uncertainty about exact time-ordering of each infection and breastfeeding cessation within a time interval. By defining the exposure as breastfeeding status at time point t, we can consider that this intervention point occurs after infection counts

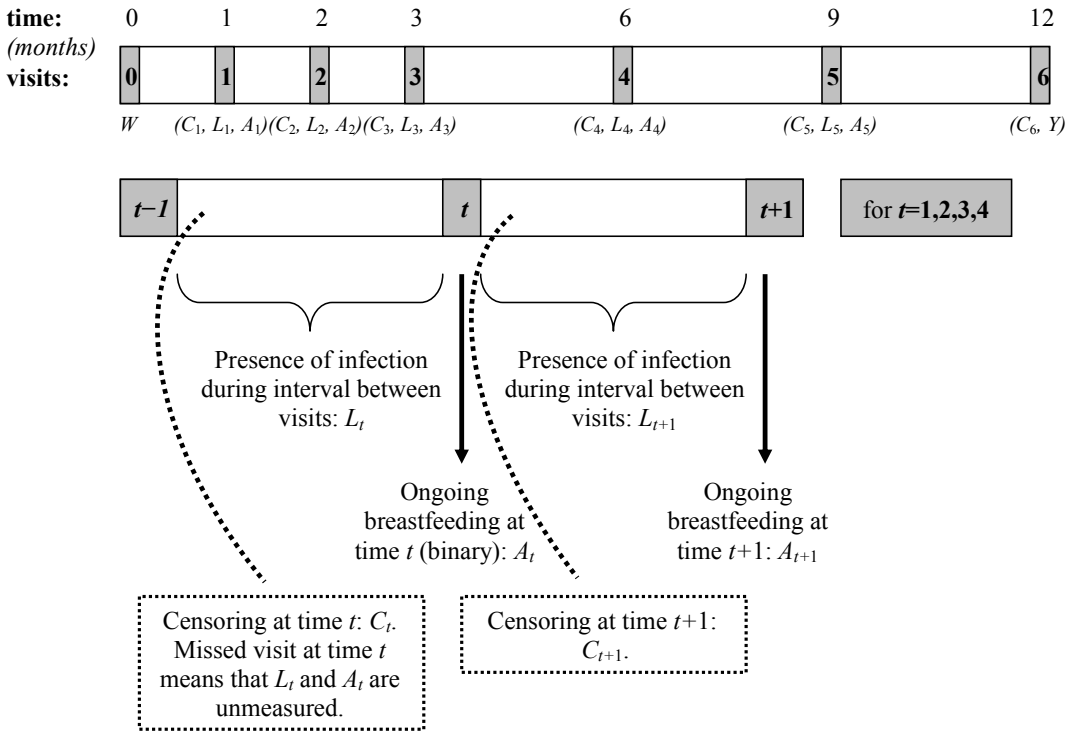


Figure 5–1: Time-ordering of the variables in the PROBIT study. Data were collected at baseline and six follow-up times. At each follow-up time point, breastfeeding status (A_t) and presence of infection over the past interval (L_t) were noted. Censoring occurring at time t ($C_t = 1$) indicates that later breastfeeding and infection status were not observed.

measured over the previous interval. With six visits, and the outcome assessed at the sixth visit, this means that only the first five exposure nodes are considered in the analysis. However, we observe six censoring times (occurring before each of the six follow-up times). Figure 5–1 gives a graphic display of the time-ordering of the observed data.

See Table 5–2 for a summary of the infection counts and exposure status at each time point. It is clear from this table that a child having more than one infection during a given time-interval was somewhat uncommon.

5.3 Targeted estimation for longitudinal data

Suppose we observe longitudinal data (identical to the PROBIT data structure) of the form $O = (W, C_1, L_1, A_1, C_2, L_2, ..., A_{K-1}, C_K, L_K = Y)$. Let A_t denote exposure at a time point t, and the censoring indicator C_t indicate whether subjects have dropped out of the study before the t^{th} time point. The vector $C = (C_1, ..., C_K)$ indicates a subject's monotone censoring pattern, so that we let $C = 0$ mean that a subject was never censored. W is the collection of potentially confounding variables at baseline, L_t are intermediate measurements taken at time t, and Y is the outcome of interest.

Following the Neyman-Rubin model [54], define the counterfactual $L_t^{\bar{a},C=0}$ as the observation, L_t, an individual would have produced under fixed exposure history $\bar{a} = (a_1, ..., a_{K-1})$, having been fully observed. The target of inference is the marginal mean counterfactual outcome, $\psi_{\bar{a},C=0} = E(Y^{\bar{a},C=0})$. The standard causal missing data problem arises from only observing individuals under one exposure pattern over a subset of the time period.

For the example at hand, exposure A_t is whether breastfeeding was ongoing at time point t, intermediate measurements $L_t, t = 1, ..., K-1$ indicate whether the infant had any gastrointestinal infections between time points $t-1$ and t, and the outcome Y is the number of infections that occurred up until time point K. W is the set of baseline potential confounders of the effect of breastfeeding on gastrointestinal infection. For this application, this set is also chosen to be sufficient to adjust for censoring.

5.3.1 The G-computation method

G-computation [15] is a likelihood-based approach to estimating a causal parameter. Suppose our data O consist of n independently and identically distributed draws from a true underlying distribution P_0. This density may be decomposed corresponding to the time-dependent structure of the data as

$$P_0 = \underbrace{\prod_{t=1}^{K-1} P_0(A_t \mid Pa(A_t)) \prod_{t=1}^{K} P_0(C_t \mid Pa(C_t))}_{g_0} \underbrace{\prod_{t=1}^{K} P_0(L_t \mid Pa(L_t))}_{Q_0}$$

where $Pa(X)$ represents all variables preceding X in time (the graph-theoretical concept of "parents"; [34]). Q_0 is the density component of the L-variables, and g_0 is the distribution of the exposure and censoring variables.

Give a choice of fixed longitudinal exposure, $\bar{a}$, we can define the distribution of the corresponding counterfactual variables $\bar{L}_K^{\bar{a},C=0}$ as

$$P_0^{\bar{a},C=0} = \prod_{t=1}^{K} P_0(L_t \mid \bar{C}_t = 0, \bar{A}_{t-1} = \bar{a}_{t-1}, \bar{L}_{t-1}),$$

where $\bar{A}_t = (A_1, ..., A_t)$ and $\bar{L}_t = (W, L_1, ..., L_t)$ represent the respective histories of these variables up until time point t (and correspondingly define $\bar{C}_t = (C_1, ..., C_t)$). Similarly, $\bar{a}_t = (a_1, ..., a_t)$ is the component of the fixed regime up until time point t. The targeted parameter of interest can then be described as $E_{P_0^{\bar{a},C=0}} Y^{\bar{a},C=0}$ where the expectation is taken under this $P_0^{\bar{a},C=0}$.

G-computation is a substitution (or plug-in) estimator, defined by a smooth functional, Ψ, that takes a density estimate from the set $\mathcal{M}$ and returns a real-valued parameter estimate. A fit of the density, $p_n^{\bar{a},C=0}$, is obtained and substituted

into the function Ψ to produce parameter estimate $\hat{\psi}_{\bar{a},C=0} = \Psi(p_n^{\bar{a},C=0})$. When the density is correctly specified, this is an unbiased estimate of the true parameter $\psi_{\bar{a},C=0} = \Psi(P_0^{\bar{a},C=0})$.

For the given data (with binary intermediate variables $L_t, 1 \leq t \leq K-1$), the G-computation estimator for parameter $\psi_{\bar{a},C=0}$ is

$$\begin{aligned}\hat{\psi}_{\bar{a},C=0} = &\sum_W \sum_{l_1=\{0,1\}} \cdots \sum_{l_{K-1}=\{0,1\}} E_n(Y|\bar{L}_{K-1} = \bar{l}_{K-1}, \bar{A}_{K-1} = \bar{a}, C=0)\times \\ &p_n(L_{K-1} = l_{K-1} \mid \bar{L}_{K-2} = \bar{l}_{K-2}, \bar{A}_{K-2} = \bar{a}_{K-2}, \bar{C}_{K-1} = 0)\times \\ &\cdots p_n(L_1 = l_1 \mid W, C_1 = 0) p_n(W),\end{aligned}$$

where p_n and E_n represent empirically derived fits of the conditional expectations.

Therefore, conditional probabilities for $L_t, 1 \leq t \leq K$ must be fit to produce a G-computation estimate. This can be done using any parametric or nonparametric method as desired. The density for the baseline variables W can be estimated using the empirical density estimate so that $p_n(W = w_i) = 1/n$ for each subject (with realization w_i).

To estimate the above, only a complete baseline vector W is needed for a subject to be included in the analysis (without requiring the use of additional missing data likelihood augmentation or imputations).

5.3.2 Sequential G-computation formulation

As suggested by [1] and used by [74], an alternative formulation of the likelihood, and therefore an alternative to the full likelihood G-computation, can be constructed by taking sequential expectations of the outcome. Their result is an application of

the probabilistic property of iterated expectation, specifically

$$E(Y) = E\{E(Y|X)\}.$$

Calculation of the above mean allows the underlying density to be broken down into two components: the conditional expectation of counterfactual Y given X and the subsequent marginal expectation taken over the X values. The marginal mean under exposure and full observation can then be reexpressed as

$$\begin{aligned}
\psi_{\bar{a},C=0} &= E(Y^{\bar{a},C=0}) \\
&= E\{E(Y^{\bar{a},C=0} \mid \bar{L}_{K-1}^{\bar{a}_{K-2},C_{K-1}=0})\} \\
&= E[E\{E(Y^{\bar{a},C=0} \mid \bar{L}_{K-1}^{\bar{a}_{K-2},C_{K-1}=0}) \mid \bar{L}_{K-2}^{\bar{a}_{K-3},C_{K-2}=0}\}] \\
&= E[...E\{E(Y^{\bar{a},C=0} \mid \bar{L}_{K-1}^{\bar{a}_{K-2},C_{K-1}=0}) \mid \bar{L}_{K-2}^{\bar{a}_{K-3},C_{K-2}=0}\}... \mid W] \qquad (5.1)
\end{aligned}$$

where $L_t^{\bar{a}_{t-1},C_t=0}$ represents the potential outcome of intermediate variable L_t under no censoring up until time point t and the fixed intervention $\bar{a}_{t-1}$.

To effectuate this calculation and obtain an estimate of the parameter, a model must be fit for each level of conditioning, beginning with the innermost expectation. To more easily refer to each model fit, [74] described the conditional models of the counterfactuals using Q-notation. Let

$$Q_K = E(Y^{\bar{a},C=0} \mid \bar{L}_{K-1}^{\bar{a}_{K-2},C_{K-1}=0})$$

be the outcome expectation conditional on the full history, for those who are exposed according to the regime $\bar{a}$ and fully observed. The fit Q_K is obtained using a

conditional modeling method. Then, recursively define

$$Q_t = E(Q_{t+1} \mid \bar{L}_{t-1}^{\bar{a}_{t-2}, C_{t-1}=0}), \quad t = K-1, ..., 2$$
$$Q_1 = E(Q_2|W)$$

for each successive nested expectation (letting $\bar{a}_0$ be the null set).

This alternative decomposition of the parameter space can be used to compute the parameter of interest using the following algorithm. It is done by producing model fits for each of the Q's and taking a mean over all participants of the final Q_1. Specifically, the estimation algorithm proceeds as follows:

1. First, model the outcome Y given all of the covariate history, for only those completely uncensored subjects with observed intervention $\bar{A}_{K-1} = \bar{a}$. This can be done using logistic regression or any appropriate prediction method. (Alternatively, a general conditional expectation conditional on $\bar{A}_{K-1}$ can be fit using all uncensored subjects and then evaluated at $\bar{a}_{K-1}$ in order to smooth over all observations.)
2. Then, using the model produced in 1), predict the conditional outcome for all subjects (including those censored), producing the fit $\hat{Q}_K$. This step will require imputing a value into the unobserved intermediate variables for those who were censored.
3. Model the conditional outcome from the previous step given the covariate ancestors of L_{K-1}, only for those uncensored up until time $K-1$ (i.e., subjects with $C_{K-1} = 0$) and for those with observed intervention $\bar{A}_{K-2} = \bar{a}_{K-2}$.

(Again, this model can be alternatively fit using all uncensored subjects using an expectation conditional on $\bar{A}_{K-2}$, and then evaluating at $\bar{a}_{K-2}$.)

4. Predict a new conditional outcome from this last model for all subjects, producing the fit $\hat{Q}_{K-1}$.

Repeat steps 3 and 4 for each time point (going backwards in time) until a model is obtained for the outcome conditional on only the baseline covariates, W. The parameter estimate is then obtained by taking a mean of the final conditional outcome $\hat{Q}_1$ for all observations or, analogously, over all values of W (this is equivalent to modeling the baseline covariates with empirical distributions).

Example using two time points

To clarify the above procedure, consider the simplified algorithm for the reduced dataset with structure $O = (W, C_1, L_1, A_1, C_2, Y)$. Suppose Y is a count outcome, measured after the second time-interval ($t = 2$). The single intervention A_1 is binary, the baseline W is continuous, and $C_t, t = 1, 2$ is the indicator of whether the subject has been censored at time 1 or 2, respectively. This data structure, O, mirrors the first two time points of the PROBIT data. Let the parameter of interest be $\psi_{1,C=0} = E(Y^{a_1=1,C=0})$, the expected outcome under intervention and no censoring. This corresponds with the mean population outcome when setting the two intervention nodes to ($C_1 = 0$) and ($A_1 = 1$, $C_2 = 0$).

A parametric version of the algorithm proceeds as follows:

1. Use a Poisson regression to model Y as a function of W and L_1, for only subjects who were never censored ($C_2 = 0$) and experienced the intervention ($A_1 = 1$).

2. Impute values of L_1 for those who were censored at time $t = 1$. Then, using the model fit in step 1, calculate the conditional expectations of the outcome for all subjects given their observed or imputed values of W and L_1.
3. Model the expected outcomes from step 2 using a Poisson regression conditional on W. Fit this model only using subjects with $C_1 = 0$.
4. Using the model from step 3, predict the conditional outcomes for all subjects given their observed value of W.
5. Take a mean of the predicted conditional outcomes of step 4, over all subjects. This is the estimate of the parameter of interest.

Note that the above process only fits two models regardless of the form or dimension of the intermediate variable L_t. For a general longitudinal dataset, this procedure fits one model per time point (where there is an intervention or censoring), so that the number of modeling steps is independent on the form of the intervention, the baseline variables and the intermediate variables.

5.3.3 Efficient estimation for longitudinal data

Both G-computation algorithms described here require correct specification of different decompositions of the underlying data-generating form. Working instead with a semiparametric efficient estimator would produce estimators with asymptotically minimal variance among semiparametric estimators and give the added benefit of double robustness [76, 71]. A simple way of obtaining such beneficial properties is to estimate a formulation of the efficient influence curve for the parameter of interest, and solve it as an estimating equation by setting it equal to zero.

Corresponding to the original G-computation factorization of the likelihood, [73] presented a representation of the efficient influence curve for a longitudinal form with binary intermediate variables. Similarly, [69] modified the corresponding theory for survival data. The alternative formulation for the efficient influence curve is given by [1] and [74], allowing for a general longitudinal form and much easier estimation procedures for higher dimensional or more complex longitudinal data.

Let $g_{\bar{a},t}(\bar{L}_{t-1})$ be the estimate of the probability associated with obtaining a given history of exposure up until time $t-1$, and no censoring up until time point t, as a function of the observed counterfactual history, $\bar{L}_{t-1}^{\bar{a}_{t-2},C_{t-1}=0}$, for $t = 2, ..., K$ (letting a_0 be the empty set). Exceptionally, let $g_1(W)$ be the probability of being uncensored at the first time point, conditional on baseline covariates, W (and not dependent on $\bar{a}$). These probabilities can be estimated as, for instance, a product of conditional probabilities (for exposure and non-censoring at each time point conditional on the history) estimated using logistic regression. As derived and explained for a general longitudinal structure in Van der Laan & Gruber, the components of the efficient influence curve can then be written recursively for the PROBIT data as

$$\begin{aligned} D_t &= \frac{I(\bar{A}_{t-1} = \bar{a}_{t-1}, C_t = 0)}{g_{\bar{a},t}}(Q_{t+1} - Q_t) \quad \text{for } t = K, ..., 2, \\ D_1 &= \frac{I(C_1 = 0)}{g_1}(Q_2 - Q_1), \text{ and} \\ D_0 &= (Q_1 - \hat{\psi}_{\bar{a},C=0}). \end{aligned} \tag{5.2}$$

where $Q_{K+1} = Y$ is defined for notational convenience (and the dependencies of some components repressed).

With each of the g and Q components estimated using any given prediction method, the parameter $\psi_{\bar{a},C=0}$ can be estimated by setting the sum of the $K+1$ components equal to 0 and solving for $\hat{\psi}_{\bar{a},C=0}$.

5.3.4 TMLE using the alternative G-computation formulation

TMLE for the point treatment mean in a longitudinal setting

The sequential G-computation method described in Section 5.3.2 is a plug-in estimator because it is a function of a underlying densities, in this case formulated as a sequence of conditional expectations. The general TMLE procedure begins with some choice of plug-in estimator, but improves upon this estimator by updating the density estimates according to specific rules which we detail below. This produces efficient, doubly robust estimators. This general procedure has been described previously, for example, by [78, 18, 51].

Details regarding the construction of the sequential longitudinal estimator are given by [74]. The first step in the TMLE procedure is to fit the conditional densities $\{Q_t, t = 1, ..., K\}$ using a method of choice. For the update step, the logistic loss function is chosen even for our case of integer-valued outcome (reduced to proportions by shifting and scaling to [0,1]) due to the boundedness properties of the inverse of its canonical link function. The logistic loss becomes particularly valuable when there is sparsity at certain levels of the covariates or exposure [18].

The next step is to fluctuate each of the density estimates $\{\hat{Q}_t, t = K, ..., 1\}$, going backwards through time, with respect to a new parameter, ϵ_t. The fluctuation function for each $\hat{Q}_t(\epsilon_t)$ can be described as

$$\text{logit}\hat{Q}_t^1(\epsilon_t) = \text{logit}\hat{Q}_t + \epsilon_t G_t, \quad t = 1, ..., K$$

for some expression G_t. Again letting $\hat{Q}_{K+1} = Y$, the optimal value for ϵ_t is found by minimizing the empirical mean of the logistic loss function

$$\mathcal{L}\{\hat{Q}_t^1(\epsilon_t)\} = -[\hat{Q}_{t+1}\log\{\hat{Q}_t^1(\epsilon_t)\} + (1-\hat{Q}_{t+1})\log\{1-\hat{Q}_t^1(\epsilon_t)\}],$$

which is equivalent to solving the empirical mean score (or derivative of the loss function) at zero. This requires that the function G_t is defined and estimated.

The above fluctuation function is required to satisfy two conditions: 1) it must equal the original when $\epsilon_t = 0$, and 2) the derivative with respect to ϵ_t of the loss function at $\epsilon_t = 0$ must span the efficient influence curve. The first condition is clearly satisfied when $\epsilon_t = 0$. Taking the derivative of the loss function with respect to ϵ_t gives:

$$\left.\frac{d\mathcal{L}(\hat{Q}_t^1(\epsilon_t))}{d\epsilon_t}\right|_{\epsilon_t=0} = G_t(\hat{Q}_{t+1} - \hat{Q}_t), \quad t = 1, ..., K.$$

Therefore, the score spans the efficient influence curve when G_t is defined as

$$G_t(\bar{A}_{t-1}, C_t, \bar{L}_{t-1}) = \frac{I(\bar{A}_{t-1} = \bar{a}_{t-1}, C_t = 0)}{g_{\bar{a},t}}.$$

The fluctuation step is carried out by minimizing the loss function, $\mathcal{L}\{\hat{Q}_t^1(\epsilon_t)\}$, with respect to ϵ_t. This is equivalent to running a no-intercept logistic regression with offset $\hat{Q}_t$ and unique covariate $G_t(\bar{A}_{t-1}, C_t, \bar{L}_{t-1})$. Let $\hat{\epsilon}_t$ be the estimate of the coefficient for G_t, which is the maximum likelihood estimate (or equivalently, the minimum loss-based estimate) for ϵ_t.

Once all of the densities have been updated to give $\{\hat{Q}_t^1, t = K, ..., 1\}$, the parameter $\psi_{\bar{a},C=0}$ is estimated as the mean of $\hat{Q}_1^1$ over all subjects, i.e. $\hat{\psi}_{\bar{a},C=0} = \frac{1}{n}\sum_i Q_1^1(W = w_i)$ (where w_i is the observed baseline vector for subject i).

Procedure for the PROBIT data

We observed the following procedure in our estimation of the parameter $\psi_{\bar{a},C=0}$, for a given exposure history $\bar{a}$. As described above, our interpretation of the structure of the PROBIT dataset is $O = (W, C_1, L_1, A_1, C_2, L_2, ..., A_5, C_6, L_6 = Y)$. There are six intervention nodes: censoring can occur at any of them, and exposure is measured at the later five. All subjects are breastfeeding at the baseline, so that exposure pattern is uniquely determined by breastfeeding cessation by a given time point.

1. Fit models for the exposure and censoring indicators at each time point, given all history up until that time point. Obtain predicted values for each subject's probability of obtaining exposure status a_t, and each subject's probability of being observed at each time point. These must be calculated conditional on fixed exposure history $\bar{a}_{t-1}$ and having been observed up until the given time point.
    - In particular, given the monotone nature of breastfeeding exposure, if $\bar{a} = (1, 0, 0, 0, 0)$, for instance, the predicted probability of *not* breastfeeding at time 3 will be one for all participants, since it's conditional on stopping before time 2.
2. Using the predictions from step 1, fit the propensities,

$$g_1(W) = p_n(C_1 = 0 \mid W), \text{ and}$$
$$g_{a,t}(\bar{L}_{t-1}) = p_n(C_1 = 0 \mid W) \prod_{k=2}^{t} \{p_n(C_k = 0 \mid \bar{A}_{k-1} = \bar{a}_{k-1}, C_{k-1} = 0, \bar{L}_{k-1}) \times$$
$$p_n(A_{k-1} = a_{k-1} \mid \bar{A}_{k-2} = \bar{a}_{k-2}, C_{k-1} = 0, \bar{L}_{k-1})\}$$

for $t = 2, ..., 6$, and where A_0 and a_0 should be considered the null set.

3. Set $\hat{Q}_7 = Y$, where Y is rescaled to [0,1]. Then, for $t = 6, ..., 1$,
   - For the subset of subjects with $\bar{A}_{t-1} = \bar{a}_{t-1}$ and $C_t = 0$, fit a model for $E(\hat{Q}_{t+1} \mid \bar{L}_{t-1})$. Using this model, predict the conditional outcome for all subjects and let this vector be denoted $\hat{Q}_t$ (this may require imputing values in $\bar{L}_{t-1}$ for censored subjects).
   - Construct "clever covariate" $G_t(\bar{A}_{t-1}, \bar{L}_{t-1}) = I(\bar{A}_{t-1} = \bar{a}_{t-1}, C_t = 0)/g_{\bar{a},t}$.
   - Update the expectation by running a no-intercept logistic regression with the fit $\text{logit}(\hat{Q}_t)$ as an offset, and clever covariate G_t as the unique covariate. Let $\hat{\epsilon}_t$ be the estimated coefficient of G_t.
   - Update the fit of Q_t by setting

$$\hat{Q}_t^1 = \text{expit}\{\text{logit}(\hat{Q}_t) + \hat{\epsilon}_t G_t(\bar{A}_{t-1} = \bar{a}_{t-1}, C_t = 0, \bar{L}_{t-1})\}$$

     and obtaining a predicted value of $\hat{Q}_t^1$ for all subjects (which may again require filling in values for $\bar{L}_{t-1}$ for those who were censored earlier).
   - Note that for $t = 1$, $\hat{Q}_t$ is only conditional on W, and it is initially modeled only for subjects with $C_1 = 0$.
4. Having fit $\hat{Q}_1^1$, take the mean of this vector of values over all subjects. This is a targeted estimator for $\psi_{\bar{a},C=0}$.

5.4 Analysis of the PROBIT

The PROBIT data were analyzed by both G-computation methods, TMLE with parametric modeling of the sequential conditional means and conditional probabilities of exposure and censoring (logistic main terms regression for binary exposure

and censoring, and for the outcome shifted and scaled to [0,1]), TMLE with Super Learner to model the underlying densities, and a stabilized inverse probability of treatment weighted (IPTW) estimator. All models were implemented directly in R Statistical Software [39] with the exception of Super Learner which we fit using the R library `SuperLearner` [37]. Super Learner produces fits for each method in a library, and then estimates the ideal combination of these results based on the k-fold cross-validated error. The library we utilized included main terms logistic regression, generalized additive modeling [19], the mean estimate, nearest neighbour algorithm [35], multivariate adaptive regression spline models [31], and a stepwise AIC procedure (`stepAIC` from [83]).

A stabilized IPTW estimator was computed by obtaining the solution of

$$E\left\{(Y-\hat{\psi}^{IPTW}_{\bar{a},C=0})\frac{I(\bar{A}_5=\bar{a},C=0)}{g_{\bar{a},6}}\right\}.$$

This is the influence curve of the stabilized IPTW.

The standard errors for all methods except the G-computations were calculated using the sandwich estimator, which uses the form of the influence curve to approximate the asymptotic variance. The standard error of the estimate is found by estimating the influence curve value for each subject and then taking the empirical standard error of the 17,036 squared values. Confidence intervals were calculated assuming Normality of the estimator, taking limits to be a distance of 1.96 times the estimated standard error from the estimate. The standard errors for the G-computation methods were estimated using nonparametric bootstrap by resampling the full dataset with replacement 200 times, recalculating the estimates, and taking

the standard error of the estimates. Confidence intervals were calculated by taking the 2.5th and 97.5th quantiles of the resampled estimates.

The estimates of the marginal expected number of infections up until one year of age, for the six different breastfeeding patterns, are presented in Table 5–3. Specifically, the exposure patterns considered were exposure to breastfeeding terminated in the interval preceding the given followup date. The table presents the estimate under each method for each follow-up time, along with the 95% confidence interval. The different methods give roughly similar results, with the notable exception of sequential G-computation. This method deviates in particular in its results for breastfeeding duration of over nine months. A chart with a visual display of the results for TMLE with Super Learner is presented in Figure 5–2(a). This method (along with all of the other methods with the exception of sequential G-computation) gives decreasing point-estimates for the number of infections as breastfeeding duration is increased.

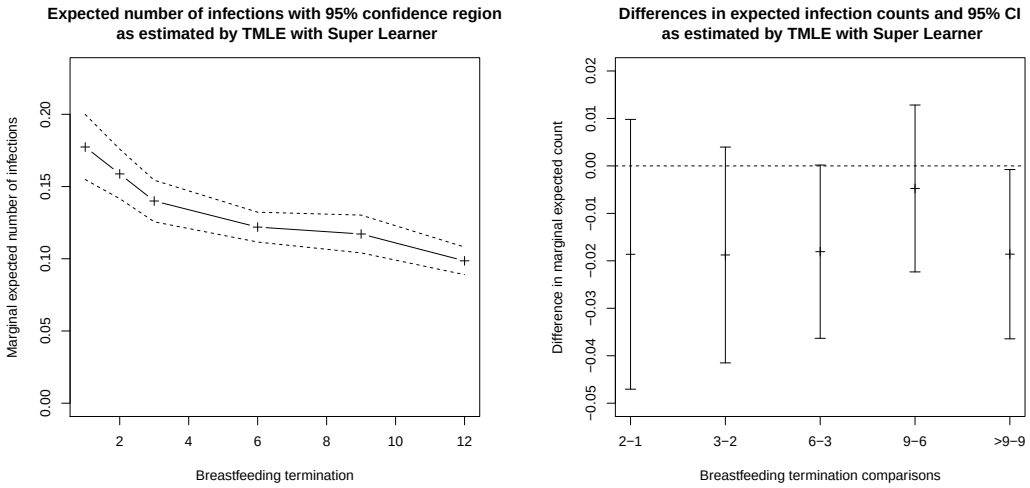
Inference regarding the treatment differences of sequential pairwise comparisons estimated using TMLE with Super Learner is summarized in Figure 5–2(b). The differences relating pairwise comparisons of breastfeeding termination between immediately subsequent study intervals are consistently estimated as negative. The corresponds with decreasing expected infection counts for longer duration of breastfeeding. However, the TMLE with Super Learner estimation method only finds the last pairwise comparison to be significantly different from zero at the 95% confidence level.

Additional interesting comparisons can be made. In particular, the expected difference between ceasing breastfeeding before one month, compared to between

Table 5–3: Marginal mean number of infections by duration of breastfeeding.

Method	Estimate	95% C.I.	Estimate	95% C.I
	0-1 months		*1-2 months*	
G-Comp (likelihood)	0.20	(0.16,0.22)	0.15	(0.14,0.18)
G-Comp (sequential)	0.19	(0.16,0.25)	0.17	(0.15,0.20)
TMLE with SL	0.18	(0.15,0.20)	0.16	(0.14,0.18)
parametric TMLE	0.18	(0.15,0.20)	0.16	(0.14,0.18)
IPTW	0.20	(0.16,0.23)	0.16	(0.14,0.18)
	2-3 months		*3-6 months*	
G-Comp (likelihood)	0.13	(0.12,0.14)	0.11	(0.10,0.13)
G-Comp (sequential)	0.15	(0.13,0.19)	0.12	(0.12,0.14)
TMLE with SL	0.14	(0.13,0.15)	0.12	(0.11,0.13)
parametric TMLE	0.14	(0.12,0.15)	0.12	(0.11,0.13)
IPTW	0.14	(0.13,0.15)	0.12	(0.11,0.13)
	6-9 months		*over 9 months*	
G-Comp (likelihood)	0.10	(0.09,0.12)	0.10	(0.09,0.11)
G-Comp (sequential)	0.12	(0.10,0.22)	0.16	(0.11,0.46)
TMLE with SL	0.12	(0.10,0.13)	0.10	(0.09,0.11)
parametric TMLE	0.11	(0.10,0.13)	0.10	(0.08,0.12)
IPTW	0.12	(0.10,0.13)	0.11	(0.09,0.13)

NOTE: G-Comp: G-computation, using both methods described in the text: likelihood in Section 5.3.1, sequential in Section 5.3.2; TMLE: Targeted Maximum Likelihood Estimation; SL: Super Learner; IPTW: inverse probability of treatment weighting (stabilized).



(a) (b)

Figure 5–2: (a) Plot of marginal expected counts for termination of breastfeeding occurring in the interval preceding the time point. (b) Expected differences between exposure patterns with 95% confidence intervals. The pairwise exposure patterns compared are termination of breastfeeding in one interval compared to termination in the immediately following interval. Both summaries of the results are obtained using TMLE with Super Learner.

three and six months was estimated as -0.06 (95% CI: -0.08,-0.03). The expected difference between breastfeeding for between three and six months compared to over 9 months was -0.02 (95% CI: -0.04,-0.01). Finally, the overall difference between breastfeeding for less than one months versus more than nine months was estimated by TMLE with Super Learner as -0.08 (95% CI: -0.10,-0.05). This overall difference in the effect corresponds with a Number Needed to Treat (NNT) of 13 to avoid one gastrointestinal infection during the first year of life. This can roughly be compared with the intention-to-treat result in the original PROBIT study [27], where they obtained a NNT of 24 for the presence of *any* gastrointestinal infection over the

first year when contrasting subjects who did and did not receive the breastfeeding intervention.

5.5 Simulation study

5.5.1 Data generation and modeling

A simulation study was performed where data were generated to have the same structure as the PROBIT dataset. Specifically, the simulated data were of the form $O = (W, U, C_1, L_1, A_1, C_2, L_2, ..., A_5, C_6, L_6)$ where exposure, $A_t, t = 1, .., 5$ is binary, $C_t, t = 1, .., 6$ is the censoring indicator (and therefore also monotone), $L_t, t = 1, .., 5$ is binary and $Y = \sum_{t=1}^{6} L_t$ is a count variable. W and U are one-dimensional Gaussian random variables, representing baseline confounders. Exposure was generated as conditional on the baseline variables and immediate preceding covariates at every time point. In particular, breastfeeding was specifically made to be less likely to continue when infection was indicated at the current time point. Breastfeeding exposure is also monotone, and so it was only possible at a given time point if the subject was still breastfeeding at the previous one. Censoring was missing at random, conditional on baseline covariates and most recent infection status; censoring was less likely if breastfeeding continued at the previous time point and more likely if an infection occurred at the previous time point. Infections were generated as being dependant on baselines, and indicators of exposure for the past two visits, so that longer duration of breastfeeding decreased the probability of infection. Finally, The count outcome was created as the sum over six binary variables indicating infection at each time-interval (each of the first time-intervals L_t plus an additional one at time 6).

The parameter $\psi_{\bar{a}} = E(Y_{\bar{a},C=0})$ was estimated for $\bar{a} = (1,1,1,1,1)$. In other words, we estimated the marginal expected value under full exposure and without censoring. This was done under three scenarios: with no unmeasured confounders (and correct propensity model when estimated), unmeasured confounding, and near positivity violations. The model contained unmeasured confounders when U was left out of the estimation procedure. Near positivity violations were generated by making the covariates highly predictive of the exposure and censoring. Details of the data generation can be found in the Supplementary Materials.

For computational efficiency, we restricted ourselves to a smaller sample size for the simulation study. Therefore, to make our models estimable, we made the probability of infection at each time point greater than what was observed in the PROBIT. One result of this decision was that the parameter of interest had a true value of 2.01, much higher than what was estimated for the application.

Five hundred datasets of 1,000 observations were generated for each of the three data-generating scenarios. The performance of the TMLE for each of these datasets was compared to a correctly specified G-computation, an incorrectly specified sequential formulation of the G-computation formula, a stabilized IPTW estimator, and the estimate found by solving the efficient influence curve in Equation (5.2) like an estimating equation (EE). Standard errors were computed using both non-parametric bootstrap resampling (details in the footnote of Table 5–4) and influence curve inference where available. These methods for estimating the standard errors were compared, as well as the mean coverage obtained under their use.

As a small departure from the real data, the simulated data allowed only one infection at each time interval (as opposed to more than one event). The G-computation used the information that the outcome was a sum of the first five binary infection variables, and the additional binary variable, L_6, measured at time $t = 6$. Thus $Y = \sum_{t=1}^{5} L_t + L_6$, so that the G-computation simplified to

$$\frac{1}{n} \sum_{l_1=\{0,1\}} \cdots \sum_{l_K=\{0,1\}} \left[\left\{ \sum_{t=1}^{5} (L_t) + E(L_6 | \bar{L}_5 = \bar{l}_5, \bar{A}_5 = \bar{a}_5, C = 0) \right\} \times \prod_{t=2}^{5} \left\{ p(L_t = l_t | \bar{L}_{t-1} = \bar{l}_{t-1}, \bar{A}_{t-1} = \bar{a}_{t-1}, \bar{C}_t = 0) \right\} p(L_1 = l_1 | W, C_1 = 0) \right].$$

Note that using the information regarding the number of infections at each time-interval for the PROBIT data analysis would have required fitting multinomial models in the likelihood G-computation. With so few subjects having more than one infection at any given time, we did not feel that substantial information could be added by increasing the complexity of the model for the applied example by using a similar approach.

5.5.2 Simulation results

The results of each of the models run on the datasets simulated under the three data-generating scenarios are displayed in Table 5–4. With no unmeasured confounders, likelihood G-computation and IPTW were both fully correctly specified. Likelihood G-computation performed better than IPTW in terms of mean-squared error, standard error, and bias. The IPTW influence curve produced a conservative estimate of the standard error, and therefore an inflated confidence interval. Because of the way the data were generated, the Q-components used in the sequential

G-computation, TMLE and EE were incorrectly specified. Due to double robustness, the TMLE and EE were both unbiased with comparatively small standard errors, and mean-squared errors. For the TMLE and EE, the influence curve and bootstrapped estimates of the standard error were very similar. Sequential G-computation performed worst of all, with the highest bias, standard error, and mean-squared error.

When a baseline confounder was omitted from the analysis (in the second scenario), all of the models were misspecified in all components (including the exposure models and the estimates of the underlying densities). Likelihood G-computation produced far more bias, but similar standard error, mean-squared error and coverage when compared to the previous scenario. Sequential G-computation suffered in terms of bias, mean-squared error, and coverage. In terms of MSE, the TMLE and EE again performed the best of all the estimators, and did not do much worse with one confounder unmeasured as compared to the previous scenario where no confounders were omitted. IPTW also did not do much worse than in the previous scenario, though the influence curve estimate of the standard error was again found to be conservative.

Under near positivity violations, the correctly specified likelihood G-computation performs very well as it does not estimate a probability of exposure. The incorrectly specified sequential G-computation performed poorly with high bias and low coverage. The stabilized IPTW was unbiased, but had higher standard errors than in previous scenarios, and was also affected in terms of coverage. The EE was sensitive to the positivity violations that we generated, and produced a biased estimate with a

Table 5–4: Marginal mean outcome under always-exposed, by scenario. True value = 2.01

Method	% bias	SE (IC)	SE (BS)[a]	MSE	% COV (IC)	% COV (BS)[b]
			no unmeasured confounders			
G-Comp (likelihood)	-0.1	-	0.07	0.003	-	94.4
G-Comp (sequential)	2.4	-	0.12	0.016	-	96.0
parametric TMLE	-0.3	0.06	0.06	0.003	96.6◇	94.2
IPTW	-0.4	0.13	0.10	0.010	97.8◇	92.2◇
Efficient EE	-0.2	0.06	0.06	0.003	94.2	94.0
			unmeasured confounder			
G-Comp (likelihood)	1.1	-	0.07	0.004	-	94.4
G-Comp (sequential)	4.4	-	0.04	0.023	-	81.4◇
parametric TMLE	0.8	0.07	0.06	0.004	96.0	93.0
IPTW	0.6	0.12	0.10	0.010	98.6◇	94.2
Efficient EE	-0.9	0.06	0.06	0.004	93.8	93.4
			positivity violations			
G-Comp (likelihood)	0.0	-	0.06	0.003	-	96.2
G-Comp (sequential)	5.8	-	0.13	0.027	-	76.4
parametric TMLE	-0.8	0.06	0.08	0.005	92.2◇	95.4
IPTW	-0.7	0.14	0.13	0.019	94.2	90.6◇
Efficient EE	2.1	0.16	0.29	0.057	93.8	92.2◇

NOTE: SE (IC): standard error calculated using the influence curve; SE (BS): standard error calculated using the nonparametric boostrap; MSE: mean-squared error calculated over the simulated datasets; COV: mean coverage; TMLE: Targeted Maximum Likelihood Estimator; G-Comp: G-computation; IPTW: (stabilized) inverse probability of treatment weighting; Efficient EE: estimating equation using the efficient influence curve.

[a]The bootstrap standard error was computed using 200 resamples from the data set of size n=17,036; [b]The estimated coverage is the % of times that the true value falls between the 2.5th and 97.5th bootstrap percentiles. ◇ Indicates a coverage significantly different from 95% (2-sided z-test, 0.05 significance level).

means squared error that was over twice the size of the incorrectly specified sequential G-computation. Contrastingly, the TMLE produced a low-bias, low-standard error estimate, with a mean-squared error comparable to that of the likelihood G-computation.

5.6 Discussion

In this article, we applied five different causal methods to the PROBIT data to obtain estimates of the marginal expected number of infection counts under the six possible breastfeeding patterns. TMLE with parametric and Super Learner density estimation produced the smallest standard errors. Sequential G-computation seemed unstable in this example, with occasionally high standard errors that resulted in very large confidence intervals. The difference between TMLE fit with parametric models and Super Learner was not great, potentially due to the limitations of our chosen library, or a good fit from the parametric regressions. We did however see a slight decrease in the standard error of the parameter estimate when the TMLE was fit with Super Learner, which may be due to the better initial fit of the underlying density. To our knowledge, this is the first time this sequential TMLE method has been applied to a real data example.

Our original goal for the PROBIT analysis was to investigate whether longer-term breastfeeding would result in lower expected numbers of infections for the first 12 months after birth. To this end, we presented the results of TMLE with Super Learner for comparing different exposure patterns. We did not observe significant results for extending breastfeeding by one time interval (with the exception of stopping before 9 months vs. extending past 9 months). However, we did see that comparing

larger differences in breastfeeding duration led to statistically and clinically significant results, corresponding to the protective effect of breastfeeding observed in the original PROBIT intention-to-treat analysis.

A causal interpretation of the analysis of the PROBIT data requires the usual causal assumptions, including the sequential randomization assumption (no unmeasured confounding). This pinpoints a limitation in the causal interpretability of our results, as the complexities of the substantive matter make it challenging to believe that we identified all the common causes of breastfeeding cessation and infections [29]. In addition, we must assume that there is no interference between study units (mother and infant pairs) and that only one version of the treatment is applied to all units (together referred to as the stable unit treatment variable assumption, or SUTVA; [56]).

In the simulation study, we generated three different types of data scenarios, and tested the performance of five reasonable longitudinal estimators. While a correctly specified likelihood computation (G-computation) will perform optimally, estimators based on the efficient influence curve are more stable under misspecification. Recall that both of the doubly robust estimators presented in the simulation study were always misspecified in the Q-components for the data generated. And yet, they performed comparatively to the correctly specified G-computation in the first two scenarios. In the scenario where we generated near positivity violations, efficient estimation without the use of plug-in estimation suffered from instability, while the TMLE remained stable, and still comparable to the optimal performance of the likelihood G-computation.

5.7 Supplementary material for Manuscript 2: Simulation details

The data in the simulation study was generated in order to resemble the PROBIT data, with structure $O = (W, U, C_1, L_1, A_1, C_2, L_2, ..., A_5, C_6, L_6)$ where exposure, $A_t, t = 1, .., 5$ is binary, $C_t, t = 1, .., 6$ is the censoring indicator (and therefore also monotone), $L_t, t = 1, .., 5$ is binary and $Y = \sum_{t=1}^{6} L_t$ is a count variable. W and U are one-dimensional Gaussian random variables, representing baseline confounders. In Section 5 of the main text, we heuristically summarize the data generation. The major differences between the generated data and the PROBIT data is that the generated outcome is a summation over the binary intermediate variables, and the sample size for the generated data was made to be smaller for computational efficiency, leading to the computational necessity of making unrealistically large probabilities of infection at each time point.

We generated slightly altered data for each of the three scenarios, namely, correct propensity estimation, unmeasured confounding, and positivity violations. We used the following functions (written in R Statistical Software version 2.13.2, [39]) to generate the data for the correct and unmeasured confounding (`data_cor`) and near positivity violations (`data_pos`) scenarios:

```
##########################################################
#Robust expit function (allowing for large values of x)
expit<-function(x){
z<-exp(x)/(1+exp(x))
z[is.na(z)]<-1
return(z) }
##########################################################

##########################################################
#CORRECT/UNMEASURED CONFOUNDING DATA GENERATION
#bf decreases prob of infection (2 time points)
```

```
#infection increases prob of censoring
#infection increases prob of later infection
#bf decreases prob of censoring
#censoring decreases prob of subsequent bf
#br at time t only possible if bf at time t-1

data_cor<-function(i,ssize){
set.seed(i*5436)

W<-rnorm(n=ssize)/4+1
U<-rbinom(n=ssize,size=1,prob=0.5)

c1<-expit(-3+0.01*W+0.5*U)
C1<-rbinom(n=ssize,size=1,prob=c1)

mu1<-expit(1.5-2.5*W+U)
L1<-rbinom(n=ssize,prob=mu1,size=1)

p1<-expit(1+W+1.5*U-0.5*C1-1*L1)
A1<-rbinom(n=ssize,size=1,prob=p1)

C2<-rep(1,ssize)
c2<-expit(-3+0.5*W+2*L1+0.5*U-1.2*A1)[C1==0]
C2[C1==0]<-rbinom(n=length(c2),size=1,prob=c2)

mu2<-expit(1-1*W+0.1*L1+0.5*U-0.5*A1)
L2<-rbinom(n=ssize,prob=mu2,size=1)

A2<-rep(0,length=ssize)
p2<-expit(1.5*W[A1=1]-1*L2[A1==1]+U[A1==1]-0.5*C2[A1==1])
A2[A1==1]<-rbinom(n=length(p2),size=1,prob=p2)

C3<-rep(1,ssize)
c3<-expit(-2+0.07*W+2*L2+0.5*U-1.2*A2)[C2==0]
C3[C2==0]<-rbinom(n=length(c3),size=1,prob=c3)

mu3<-expit(2-1*W+0.1*L2+0.5*U-A2-0.5*A1)
L3<-rbinom(n=ssize,prob=mu3,size=1)

A3<-rep(0,length=ssize)
p3<-expit(1+W[A2==1]-1*L3[A2==1]+U[A2==1]-0.5*C3[A2==1])
A3[A2==1]<-rbinom(n=length(p3),size=1,prob=p3)

C4<-rep(1,ssize)
c4<-expit(-2.5+0.07*W+2*L3+0.5*U-1.2*A3)[C3==0]
C4[C3==0]<-rbinom(n=length(c4),size=1,prob=c4)
```

```
mu4<-expit(2-1*W+0.1*L3+0.5*U-A3-0.5*A2)
L4<-rbinom(n=ssize,prob=mu4,size=1)

A4<-rep(0,length=ssize)
p4<-expit(1+1*W[A3==1]-1.5*L4[A3==1]+U[A3==1]-0.5*C4[A3==1])
A4[A3==1]<-rbinom(n=length(p4),size=1,prob=p4)

C5<-rep(1,ssize)
c5<-expit(-2+0.07*W+2*L4+0.5*U-1.2*A4)[C4==0]
C5[C4==0]<-rbinom(n=length(c5),size=1,prob=c5)

mu5<-expit(1-1*W+0.2*L4+0.5*U-A4-0.5*A3)
L5<-rbinom(n=ssize,prob=mu5,size=1)

A5<-rep(0,length=ssize)
p5<-expit(0.5+W[A4==1]-1.5*L5[A4==1]+U[A4==1]-0.5*C5[A4==1])
A5[A4==1]<-rbinom(n=length(p5),size=1,prob=p5)

C6<-rep(1,ssize)
c6<-expit(-1.5+0.07*W+2*L5+0.5*U-1.2*A5)[C5==0]
C6[C5==0]<-rbinom(n=length(c6),size=1,prob=c6)

#use to get Y
mu6<-expit(-1.8-1*W+0.7*L5+0.5*U-A5-0.5*A4)
L6<-rbinom(n=ssize,prob=mu6,size=1)

Y<-L1+L2+L3+L4+L5+L6

L1[C1==1]<-NA
A1[C1==1]<-NA
L2[C2==1]<-NA
A2[C2==1]<-NA
L3[C3==1]<-NA
A3[C3==1]<-NA
L4[C4==1]<-NA
A4[C4==1]<-NA
L5[C5==1]<-NA
A5[C5==1]<-NA
Y[C6==1]<-NA

return(as.data.frame(cbind(W,L1,L2,L3,L4,L5,Y,A1,A2,A3,A4,A5,C1,C2,C3,C4,C5,C6,U)))

}
############################################################
```

```
#########################################################
#DATA GENERATION FOR NEAR POSITIVITY VIOLATIONS
data_pos<-function(i,ssize){
set.seed(i*5436)

W<-rnorm(n=ssize)/4+1
U<-rbinom(n=ssize,size=1,prob=0.5)

c1<-expit(-3+0.01*W+0.5*U)
C1<-rbinom(n=ssize,size=1,prob=c1)

mu1<-expit(1.5-2.5*W+U)
L1<-rbinom(n=ssize,prob=mu1,size=1)

p1<-expit(-5+8*W+1.5*U-0.5*C1-1*L1)
A1<-rbinom(n=ssize,size=1,prob=p1)

C2<-rep(1,ssize)
c2<-expit(-3+0.5*W+2*L1+0.5*U-1.2*A1)[C1==0]
C2[C1==0]<-rbinom(n=length(c2),size=1,prob=c2)

mu2<-expit(1-1*W+0.1*L1+0.5*U-0.5*A1)
L2<-rbinom(n=ssize,prob=mu2,size=1)

A2<-rep(0,length=ssize)
p2<-expit(1+1.5*W[A1=1]-1*L2[A1==1]+U[A1==1]-0.5*C2[A1==1])
A2[A1==1]<-rbinom(n=length(p2),size=1,prob=p2)

C3<-rep(1,ssize)
c3<-expit(-2+0.07*W+2*L2+0.5*U-1.2*A2)[C2==0]
C3[C2==0]<-rbinom(n=length(c3),size=1,prob=c3)

mu3<-expit(2-1*W+0.1*L2+0.5*U-A2-0.5*A1)
L3<-rbinom(n=ssize,prob=mu3,size=1)

A3<-rep(0,length=ssize)
p3<-expit(2*W[A2==1]-1*L3[A2==1]+U[A2==1]-0.5*C3[A2==1])
A3[A2==1]<-rbinom(n=length(p3),size=1,prob=p3)

C4<-rep(1,ssize)
c4<-expit(-2+0.07*W+2*L3+0.5*U-1.2*A3)[C3==0]
C4[C3==0]<-rbinom(n=length(c4),size=1,prob=c4)

mu4<-expit(2-1*W+0.1*L3+0.5*U-A3-0.5*A2)
L4<-rbinom(n=ssize,prob=mu4,size=1)
```

```
A4<-rep(0,length=ssize)
p4<-expit(1+1*W[A3==1]-1.5*L4[A3==1]+U[A3==1]-0.5*C4[A3==1])
A4[A3==1]<-rbinom(n=length(p4),size=1,prob=p4)

C5<-rep(1,ssize)
c5<-expit(-2+0.07*W+2*L4+0.5*U-1.2*A4)[C4==0]
C5[C4==0]<-rbinom(n=length(c5),size=1,prob=c5)

mu5<-expit(1-1*W+0.2*L4+0.5*U-A4-0.5*A3)
L5<-rbinom(n=ssize,prob=mu5,size=1)

A5<-rep(0,length=ssize)
p5<-expit(2*W[A4==1]-1.5*L5[A4==1]+U[A4==1]-0.5*C5[A4==1])
A5[A4==1]<-rbinom(n=length(p5),size=1,prob=p5)

C6<-rep(1,ssize)
c6<-expit(-18+8*W+4*L5+8*U-1.2*A5)[C5==0]
C6[C5==0]<-rbinom(n=length(c6),size=1,prob=c6)

#use to get Y
mu6<-expit(-1.8-1*W+0.7*L5+0.5*U-A5-0.5*A4)
L6<-rbinom(n=ssize,prob=mu6,size=1)

Y<-L1+L2+L3+L4+L5+L6

L1[C1==1]<-NA
A1[C1==1]<-NA
L2[C2==1]<-NA
A2[C2==1]<-NA
L3[C3==1]<-NA
A3[C3==1]<-NA
L4[C4==1]<-NA
A4[C4==1]<-NA
L5[C5==1]<-NA
A5[C5==1]<-NA
Y[C6==1]<-NA

return(as.data.frame(cbind(W,L1,L2,L3,L4,L5,Y,A1,A2,A3,A4,A5,C1,C2,C3,C4,C5,C6,U)))
}
```

The only differences between the two functions given above are the coefficients given to the some of the independent variables at each step.

Several methods were tested for each of the data generation scenarios described. The methods are G-computation, sequential G-computation, parametric TMLE, inverse probability of treatment weighting (IPTW), and estimating equations using the efficient influence function (EE). For the correct propensity and near positivity violations scenarios, both of the confounders U and W were included in the estimation procedures. For the unmeasured confounding scenario, only W was included.

CHAPTER 6
Marginal Structural Modeling of a Survival Outcome with Targeted Maximum Likelihood Estimation

Preamble to Manuscript 3. This manuscript continues with the development of the longitudinal TMLE method described in Manuscript 2. This study demonstrates the modification of this method for survival analysis, both for the exposure-specific marginal mean parameter (equivalent to a saturated MSM) and for two different types of MSMs. It presents a different modeling approach than that taken by Rosenblum and Van der Laan [50] who also demonstrate a method for estimating the parameters of a MSM with TMLE. The theory and the simulation study in this manuscript show that the logistic MSM for the hazard (which can be estimated using TMLE) is equivalent to the IPTW MSM described in [21] which is also used to estimate the parameters of a Cox proportional hazards MSM. TMLE is compared to IPTW and standard techniques in a case study of HIV and Hepatitis C virus (HCV) co-infected patients where the effect of HCV clearance on end-stage liver disease (ESLD) is estimated. The dataset used in the analysis had different types of missing data and data sparsity which inspired the use of multiple imputations to help adjust for missing information. While the clinical results of this analysis are inconclusive, the use of TMLE proved to be beneficial as this method resulted in lower variance estimation than IPTW.i

Marginal Structural Modeling of a Survival Outcome with Targeted Maximum Likelihood Estimation

Mireille E Schnitzer*, Erica E M Moodie*, Mark J van der Laan†, Robert W Platt*, and Marina B Klein‡

**Department of Epidemiology, Biostatistics, & Occupational Health, McGill University, Montréal, Québec, Canada*

†Division of Biostatistics, School of Public Health, University of California, Berkeley, USA

‡Department of Medicine, McGill University, Montréal, Québec, Canada

Abstract. When estimating causal parameters in survival analysis, the analyst must take into consideration the presence of baseline confounders, loss to follow-up, and time-dependent confounders in the longitudinal data structure. Marginal structural models (MSM) can be used to model the effect of an exposure on a survival outcome. We demonstrate the doubly robust and semiparametric-efficient method of Targeted Maximum Likelihood Estimation (TMLE) applied to estimating both the marginal exposure-specific probability of survival and the parameters of a MSM with a survival-type outcome. We show the theoretical derivation of the efficient influence functions for the parameters of two different MSMs and how they can be used to produce variance approximations for parameter estimates. A simulation study demonstrates the unbiasedness of TMLE for estimating the survival curve and a hazard MSM and compares the method to inverse probability of treatment weighting methods. Finally, we undertake an analysis of the Canadian Co-infection Cohort Study where we used TMLE to estimate the impact of clearance of the Hepatitis C virus on the time to end-stage liver disease in subjects infected with both Hepatitis C and HIV.

6.1 Introduction

Standard survival modeling (such as Kaplan-Meier or Cox proportional hazards modeling) relies on the assumption that censoring and survival are independent, which is a highly unlikely supposition in many scenarios. In addition, standard methods ignore or over-adjust time-dependent confounders [43].

Marginal structural models (MSM) [46] have been developed to estimate the effect of time-dependent exposure on the outcome in the presence of time-dependent confounders that are affected by previous treatment. In the survival context, MSMs have been developed to estimate the parameters of a Cox proportional hazards model [21]. Despite their mathematically convenient form and widespread usage, the interpretation of the parameters of a Cox proportional hazards model is often challenging [69, 20] leading to the desire for alternative semiparametric survival models.

Weighting methods such as inverse probability of treatment weighting (IPTW) [8, 86] and substitution estimators such as G-computation [44] have been developed to overcome this problem, and correctly incorporate baseline and time-dependent covariates when the models are correctly specified.

Semiparametric efficient estimators in causal inference have also been produced for the survival context [47, 61, 1] giving the added advantage of double-robustness where only a component of the underlying density must be correctly specified for asymptotic unbiasedness [25]. These methods are sometimes called doubly-robust IPTW estimators or estimating equations, but all of them produce estimates that solve a corresponding efficient influence function (equivalently, efficient influence curve) for the target parameter set equal to zero.

Targeted Maximum Likelihood Estimation (TMLE) is a framework that produces semiparametric efficient estimators, and can be used for all pathwise differentiable parameters. TMLE [78] creates a substitution estimator, computed as a function of estimated components of the underlying data-generating function. However, TMLE differs from standard substitution estimation as the underlying density fits are updated in order to solve the equation of the efficient influence curve set equal to zero, and thereby produce semiparametric efficient inference in the class of regular asymptotically linear estimators [76, 71]. Like IPTW and G-computation, TMLE allows for confounding adjustment, but unlike either of these methods, it is also double-robust. In addition, TMLE offers improvements over efficient estimating equation methodology, in that it will never produce multiple solutions, and that it preserves the natural bounds of the targeted parameter in estimation. The flexibility of the estimating framework allows improvements in estimation that can give TMLE an additional advantage in challenging situations such as data sparsity [18]. TMLE has been used to produce semiparametric efficient, double-robust estimators for survival parameters [74, 69], general longitudinal parameters [73, 53, 62, 74], and MSMs for longitudinal data [51].

In this paper we explore several ways of modeling survival with TMLE: directly constructing survival curves, and using marginal structural models to summarize the log-odds of survival and the hazard function, respectively. Variance estimates for each of these methods will be made possible using efficient influence curve inference, which allows us to construct a closed-form solution for the large-sample approximations of the variance of the different estimators. We illustrate and assess the feasibility of

our methods by performing an analysis on a cohort of patients co-infected with HIV and the Hepatitis C Virus (HCV), investigating the effect of the clearance of HCV on risk for end-stage liver disease.

6.2 Background

Estimation with TMLE requires that the target parameter be identified as a differentiable function of a component of the underlying data density. Let Ψ be a differentiable function that takes an argument in a model space $\mathcal{M}$ and value in the space of real numbers (or vectors). Then, our parameter can be defined as $\psi = \Psi(Q)$, for some Ψ as described, and where $Q \in \mathcal{M}$ is some component of the underlying data density, P. Treating the function Ψ as a substitution estimator, we can estimate Q using a dataset, O, and plug it into the function Ψ so that the estimate of the target parameter becomes $\hat{\psi} = \Psi(\hat{Q})$.

Loosely, the influence curve of a regular, asymptotically linear (RAL) estimator is the component of the estimator that determines its asymptotic properties. Suppose one observes n sets of independent, identically distributed subject-specific data, $O = \{O_i, i = 1, .., n\}$. The influence curve of the estimator $\Psi(\hat{Q})$ is a function of the data and denoted $D(P)(O)$ (with subject-specific components $D(P)(O_i)$). The influence curve can be defined as

$$n^{1/2}[\Psi(\hat{Q}) - \psi] = n^{-1/2} \sum_{i=1}^{n} D(P)(O_i) + o_P(1)$$

where $o_P(i)$ is a random term than converges to zero in probability [76, 71]. By an application of the Central Limit Theorem, this implies that

$$n^{1/2}[\Psi(\hat{Q}) - \psi] \rightarrow^{\mathcal{D}} N\{0, E[D(P)D(P)^T]\}$$

so that the influence curve provides a large-sample approximation for the variance of the estimator. Specifically, $Var[\Psi(\hat{Q})] \approx 1/nVar[D(P)]$. There is a lower variance bound in the class of influence curves of RAL estimators, and the unique influence curve that attains this bound is called the efficient influence curve [71].

The general TMLE procedure is described in [78]. TMLE is defined by the procedure of updating the estimate, $\hat{Q}$, of the component of the data density used in the substitution estimator $\Psi(\hat{Q})$ in order to produce inference using the efficient influence curve. For example, TMLE is often implemented using the G-computation formula with carefully constructed updates for the density estimates [53].

Estimators based on the efficient influence curve are favoured for having low variance. For many causal parameters, estimation with the efficient influence curve is also doubly-robust [25]. This means that, while two components of the underlying density P must be fit, the estimator is asymptotically unbiased if either of the components is correctly specified. In the examples of the longitudinal TMLEs presented in this paper, only the exposure mechanism or the outcome models must be correctly specified in order for the method to be asymptotically unbiased.

6.3 Modeling theory and procedures

A general strategy for fitting a MSM for survival with TMLE is as follows:

1. Estimate the marginal probability of survival for each defined exposure category at each time point using a TMLE procedure. These probabilities describe the values of the survival curves for different exposure groups at different time points.

2. If the survival curve is not monotonic decreasing in t, one can perform a least-squares projection onto the space of monotone functions.
3. Estimate the efficient influence curve for the probability of survival for each exposure group at each time point. This can be used to build point-wise confidence intervals for the survival curves.
4. Choose a MSM specification: Choose a function of the counterfactual outcome and a linear specification for the mean. In this paper, we give examples using the log-odds of survival and hazard function. Fit the chosen MSM by performing a regression of the function of the estimated probabilities of survival for each unique time point/exposure pattern (e.g. the log-odds or the hazard function at each different exposure pattern at each time point) conditional on the mean model. This produces estimates of the MSM parameters.
5. Obtain variance estimates of the marginal structural parameters by calculating the efficient influence curve for each parameter. This is done using the estimates of the efficient influence curve for the survival curves.

6.3.1 TMLE for a survival outcome

Van der Laan and Gruber [74] developed a TMLE for the marginal exposure-specific mean outcome in a general longitudinal setting. In this section, we use their methodology to obtain an estimate of the survival curve and the influence curve of the estimator (and therefore, the approximate variance of the estimator).

Let $T^{\bar{a}}$ denote the survival time obtained under exposure pattern $\bar{a} = (a_0, a_1, ..., a_{K-1})$, defined according to the Neyman-Rubin counterfactual model [54]. The parameters of interest are the exposure-specific, censoring-free survival

probabilities $S_{\bar{a}}(t) = P(T^{\bar{a}} > t)$ for a fixed exposure pattern $\bar{a}$ at discrete time points $t = 1, ..., K$ (which we will also refer to as the survival curve under exposure $\bar{a}$ at time t). The survival curve can also be constructed separately for individual subgroups, V, if that is of interest.

Suppose we observe independent and identically distributed discretized survival times, T, and censoring times, C, for n subjects. In addition, we have information about a time-dependent exposure of interest and potentially confounding covariates at each time point. A corresponding censored data structure (without other variable missingness) can be described as $O = (\mathbf{L_0}, \mathbf{A_0}, Y_1, \mathbf{L_1}, \mathbf{A_1}, Y_2, ..., \mathbf{A_{K-1}}, Y_K)$, where the subscripts indicate a time-ordering. The vector-variable $\mathbf{L_0}$ contains baseline variables (potential confounders). The bivariate intervention nodes $\mathbf{A_t} = (A_t(1), A_t(2)), t = 0, ..., K-1$ indicate categorical exposure and censoring status, respectively, at each time point. Specifically, $A_t(2) = 0$ indicates that a subject is uncensored at time t (i.e. $C > t$), and $A_t(2) = 1$ indicates censoring prior to or at time t ($C \leq t$). Time-dependent variables $\mathbf{L_t}, t = 0, ..., K-1$ contain information about any time-dependent confounders. $Y_t, t = 1, ..., K$ is the survival status at time t where $Y_t = 1$ indicates continued survival (so that $Y_t = 1$ if and only if $T > t$). We also let $\bar{A}_t$, $\bar{L}_t$ and $\bar{Y}_t$ indicate the variable history up to and including time t.

In order to describe the formulation of the efficient influence curve first developed by Bang and Robins [1] and used by Van der Laan and Gruber [74] to develop the TMLE, fix a time $t \leq K$ and define

$$Q_t^{\bar{a}}(t) = P(T^{\bar{a}} > t \mid \bar{A}_{t-1}(1) = \bar{a}_{t-1}, A_{t-1}(2) = 0, \bar{L}_{t-1}, Y_{t-1}).$$

Note that this conditional probability is zero if there was failure at the previous time point, i.e. if $Y_{t-1} = 0$. Recursively define (going backwards starting with $j = t$)

$$Q_t^{\bar{a}}(j-1) = E(Q_t^{\bar{a}}(j) \mid \bar{A}_{j-2}(1) = \bar{a}_{j-2}, A_{j-2}(2) = 0, \bar{L}_{j-2}, Y_{j-2}), j = t, ..., 2.$$

Each $Q_t^{\bar{a}}(j-1)$ is therefore defined by taking the conditional mean, $Q_t^{\bar{a}}(j)$ and marginalizing over the intermediate covariate L_{j-1}. Finally, the parameter $S_{\bar{a}}(t)$ can be identified as $E(Q_t^{\bar{a}}(1))$. Therefore, this target parameter is defined as a function of the sequential conditional means. Each of these Q's can be modeled, and an estimate of the target parameter can be produced by taking an empirical mean of the fit $\hat{Q}_t^{\bar{a}}(1)$ (calculated for each subject) over the baseline confounders.

Define exposure and censoring-free probabilities for a given exposure pattern among the at-risk population as

$$g_{\bar{a}}(t) = \prod_{j=1}^{t} Pr(A_j(1) = a_j \mid \bar{A}_{j-1}(1) = \bar{a}_{j-1}, A_j(2) = 0, \bar{L}_{j-1}, Y_{j-1} = 1)$$
$$Pr(A_j(2) = 0 \mid \bar{A}_{j-1}(1) = \bar{a}_{j-1}, A_{j-1}(2) = 0, \bar{L}_{j-1}, Y_{j-1} = 1).$$

The efficient influence curve, $D_{\bar{a},t}$ for the parameter $S_{\bar{a}}(t) = P(T^{\bar{a}} > t)$ can be written as the sum of the $t+1$ components

$$\begin{aligned}
D_{\bar{a},t}(j+1) &= \frac{I(\bar{A}_{t-1} = \bar{a}_{t-1}, A_{t-1}(2) = 0)}{g_{\bar{a}}(t-1)}(Y_t - Q_t^{\bar{a}}(t)), \\
D_{\bar{a},t}(j) &= \frac{I(\bar{A}_{j-2} = \bar{a}_{j-2}, A_{j-2}(2) = 0)}{g_{\bar{a}}(j-2)}(Q_t^{\bar{a}}(j) - Q_t^{\bar{a}}(j-1)) \quad \text{for } j = t, ..., 2, \quad (6.1) \\
D_{\bar{a},t}(1) &= Q_t(1) - S_{\bar{a}}(t).
\end{aligned}$$

Fitting procedure

For a given t and $\bar{a}$, a TMLE estimate for the parameter $S_{\bar{a}}(t)$ can be obtained using the procedure given in [74]. Start with $j = t$. For convenience of notation, set $Q_t^{\bar{a},*}(j+1) = Y_t = I(T > t)$ (generally, the *-notation will indicate an updated fit produced according to the TMLE methodology). Fit the conditional expectation $Q_t^{\bar{a}}(j) = E(Q_t^{\bar{a},*}(j+1) \mid \bar{A}_{j-1}(1) = \bar{a}_{j-1}, A_{j-1}(2) = 0, \bar{L}_{j-1}, Y_{j-1})$ as the initial fit, $\hat{Q}_t^{\bar{a}}(j)$, calculated for all subjects (zero for those not at-risk). For example, this fit could be produced using logistic regression for all at-risk subjects. To update the fit for those at-risk, set $Q^{\bar{a},*}$ to a fluctuation of $\hat{Q}_t^{\bar{a}}(j)$ with respect to a parameter, $\epsilon_t(j)$:

$$\text{logit}\{Q_t^{\bar{a},*}(j)\} = \text{logit}\{\hat{Q}_t^{\bar{a}}(j)\} + \epsilon_t(j)\frac{1}{g_{\bar{a}}(j-1)}. \tag{6.2}$$

To fit the update by obtaining an estimate for $\epsilon_t(j)$, perform a regression, amongst those at-risk, of $\hat{Q}_t^{\bar{a},*}(j+1)$ with offset $\hat{Q}_t^{\bar{a}}(j)$ and unique covariate $I[\bar{A}_{j-1}(1) = \bar{a}_{j-1}, A_{j-1}(2) = 0]/\hat{g}_{\bar{a}}(j-1)$. Set $\hat{\epsilon}_t(j)$ to be the estimate of the coefficient of this covariate. Then update the original fit by plugging $\hat{\epsilon}_t(j)$ into Equation (6.2) and obtain a fit for all at-risk subjects (the fit for those who previously failed remains zero). We will then refer to the updated conditional expectation for all subjects as $\hat{Q}_t^{\bar{a},*}(j)$.

Repeat the above procedure for $j = t-1, ..., 1$. After the last iteration, the fit $\hat{Q}_t^{\bar{a},*}(1)$ is predicted for all subjects. The parameter estimate $\hat{S}_{\bar{a}}(t)$ is obtained by taking the mean of $\hat{Q}_t^{\bar{a},*}(1)$ over all subjects.

This procedure can be repeated for each time point $t = 1, ..., K$ to obtain an estimate of the survival curve $S^{\bar{a}}(t)$ for all values of t. One can then estimate different

survival curves for each fixed exposure pattern of interest. Let each full exposure pattern, defined up until the maximum time $K-1$, be denoted $\bar{a}^l$. Let a given exposure pattern up until time t be denoted $\bar{a}_t^l$. We will use M to represent the number of unique truncated exposure patterns $\bar{a}_t^l$. We can calculate M survival estimates, one for each truncated exposure pattern, $\bar{a}_t^l$.

6.3.2 MSM for the log-odds of survival

A model for the log-odds of survival can be described as

$$\log \frac{S_{\bar{a}^l}(t)}{1 - S_{\bar{a}^l}(t)} = X_{l,t}^T \boldsymbol{\beta}, \text{ for all unique patterns } \bar{a}_t^l$$

where $X_{l,t}^T \boldsymbol{\beta}$ represents the form of the linear specification of the model. Let $\mathbf{X}$ be the design matrix, potentially including functions of $\bar{a}^l$ and t. Let $X_{l,t}$ represent the R-dimensional row of the design matrix corresponding with exposure $\bar{a}^l$ and time t, represented as a column vector. For example, if the MSM was a linear model with an intercept and a linear term for time, then for each unique pattern $\bar{a}_{t^*}^l$ for the time point t^*, $X_{l,t^*} = (1, t^*)^T$. The design matrix can also contain subgroups if $\mathbf{S}$ was calculated separately for the components of a categorical variable, V, and although we do not include conditioning in our notation for simplicity, the following development easily extends to such a case. Finally, let $\boldsymbol{\beta}$ denote the vector of coefficients corresponding with the columns of the design matrix. Therefore, since there are M estimates for the survival function, the dimension of the matrix $\mathbf{X}$ is R by M, corresponding with a $\boldsymbol{\beta}$-vector of length R.

The parameter $\boldsymbol{\beta}$ can be defined as

$$\text{argmax}_{\boldsymbol{\beta}} E \sum_{l,t} \log \left\{ [\text{expit}(X_{l,t}^T \boldsymbol{\beta})]^{I(T^{\bar{a}^l} > t)} [1 - \text{expit}(X_{l,t}^T \boldsymbol{\beta})]^{I(T^{\bar{a}^l} \leq t)} \right\},$$

i.e. the maximum log-likelihood for the logistic model with marginal mean specification $\text{expit}(X_{l,t}^T \boldsymbol{\beta})$.

We are interested in estimation of $\boldsymbol{\beta}$, which can be implicitly written as a function of the parameters $\mathbf{S} = (S_{\bar{a}^l}(t), \text{ for all unique values of } \bar{a}_t^l)$ through the score equation:

$$0 = U(\mathbf{S}, \boldsymbol{\beta}) = \sum_{l,t} X_{l,t} \left(S_{\bar{a}^l}(t) - \text{expit}(X_{l,t}^T \boldsymbol{\beta}) \right); \quad S_{\bar{a}^l}(0) = 1.$$

In order to derive the efficient influence function for $\boldsymbol{\beta}$, we will use the Functional Delta Method [79]. In this context, it states that for a parameter $\boldsymbol{\beta} = \boldsymbol{\beta}(\mathbf{S})$ that can be written as a function of other parameters whose efficient influence functions, $D_{\bar{a}^l,t}$ are already known, the efficient influence function for $\boldsymbol{\beta}$ is equal to

$$D_{\boldsymbol{\beta}} = \sum_{l,t} \frac{d\boldsymbol{\beta}(\mathbf{S})}{dS_{\bar{a}^l}(t)} D_{\bar{a}^l,t}. \tag{6.3}$$

By the implicit function theorem, the derivative in Equation (6.3) can be obtained using

$$\frac{d\boldsymbol{\beta}(\mathbf{S})}{dS_{\bar{a}^l}(t)} = -\left[\frac{dU(\mathbf{S}, \boldsymbol{\beta})}{d\boldsymbol{\beta}} \right]^{-1} \frac{dU(\mathbf{S}, \boldsymbol{\beta})}{dS_{\bar{a}^l}(t)}. \tag{6.4}$$

We have that

$$\frac{dU(\mathbf{S}, \boldsymbol{\beta})}{d\boldsymbol{\beta}} = -\sum_{l,t} X_{l,t} X_{l,t}^T \frac{\exp(X_{l,t}^T \boldsymbol{\beta})}{(1 + \exp(X_{l,t}^T \boldsymbol{\beta}))^2}$$

is a matrix with dimension $R \times R$, and

$$\frac{dU(\mathbf{S}, \boldsymbol{\beta})}{dS_{\bar{a}^l}(t)} = X_{l,t}$$

is a column vector of length R. The two above components can be numerically evaluated and combined to form a column vector of length R using Equation (6.4).

The efficient influence function can be derived by combining Equation (6.4) with Equation (6.3) and simplifying slightly:

$$D_{\boldsymbol{\beta}} = \left[\sum_{l,t} \frac{\exp(X_{l,t}^T \boldsymbol{\beta})}{(1 + \exp(X_{l,t}^T \boldsymbol{\beta}))^2} X_{l,t} X_{l,t}^T \right]^{-1} \sum_{l,t} X_{l,t} D_{\bar{a}^l,t}.$$

Since the influence curve $D_{\bar{a}^l,t}$ can be numerically evaluated for each of the n subjects, we obtain a matrix of dimension $n \times R$, representing the joint influence components for $\boldsymbol{\beta}$.

Fitting procedure

Treating each of the functions $S_{\bar{a}^l}(t)$ as an outcome (so that there are M "observations", one for each unique exposure pattern and time), fit a logistic regression with a chosen linear specification and a logit link. This will produce the point estimate of $\boldsymbol{\beta}$. To obtain variance estimates, fit the efficient influence curve for $\boldsymbol{\beta}$ for each subject by estimating each of the components as described in Section 6.3.2 and combining them as indicated. Then, for each of the R columns of the resulting matrix the empirical variance is the estimated variance for the corresponding MSM coefficient estimate of $\boldsymbol{\beta}$.

6.3.3 MSM for the hazard function

If desired, it is also possible to model the discrete hazard function, $\lambda_{\bar{a}}(t) = P(T^{\bar{a}} = t \mid T^{\bar{a}} \geq t)$ using a logistic model. As shown in [11], when the hazard at all time points is small, this model is approximately equivalent to a Cox model (as the estimated odds ratio provides a good approximation of the hazard ratio).

As before, let $X_{l,t}^T\boldsymbol{\beta}$ denote the linear specification of the MSM where $\mathbf{X}$ is the design matrix and $\boldsymbol{\beta}$ is the vector of regression coefficients of length R and the parameter of interest. The parameter $\boldsymbol{\beta}$ can be defined as the value that maximizes the log-likelihood of a logistic model

$$\boldsymbol{\beta} = \text{argmax}_{\boldsymbol{\beta}} E \sum_{l,t} \log \left\{ [\text{expit}(X_{l,t}^T\boldsymbol{\beta})]^{I(T^{\bar{a}^l}=t)} [1 - \text{expit}(X_{l,t}^T\boldsymbol{\beta})]^{I(T^{\bar{a}^l}>t)} \right\}$$

so that only subjects with $T^{\bar{a}^l} > t$ contribute to the likelihood at a given time point. By passing the expectation through the linear expression and noting that $P(T^{\bar{a}^l} = t) = P(T^{\bar{a}^l} = t, T^{\bar{a}^l} \geq t) = P(T^{\bar{a}^l} = t \mid T^{\bar{a}^l} \geq t) P(T^{\bar{a}^l} \geq t)$ (and similarly $P(T^{\bar{a}^l} > t) = P(T^{\bar{a}^l} > t \mid T^{\bar{a}^l} \geq t) P(T^{\bar{a}^l} \geq t)$), this expression simplifies to

$$\text{argmax}_{\boldsymbol{\beta}} \sum_{l,t} S_{\bar{a}^l}(t-1) \left\{ \lambda_{\bar{a}^l}(t) \log[\text{expit}(X_{l,t}^T\boldsymbol{\beta})] + [1 - \lambda_{\bar{a}^l}(t)] \log[1 - \text{expit}(X_{l,t}^T\boldsymbol{\beta})] \right\}$$

where $S_{\bar{a}^l}(0) = 1$. This corresponds to the maximum log-likelihood for a logistic regression with outcome $\lambda_{\bar{a}^l}(t)$ and weights $S_{\bar{a}^l}(t-1)$.

Once again, $\boldsymbol{\beta}$ can be written implicitly as a function of the marginal survival parameters through the score equation corresponding to the above expression:

$$0 = U(\mathbf{S}, \boldsymbol{\beta}) = \sum_{l,t} S_{\bar{a}^l}(t-1) X_{l,t} \left(\frac{S_{\bar{a}^l}(t) - S_{\bar{a}^l}(t-1)}{S_{\bar{a}^l}(t-1)} - \text{expit}(X_{l,t}^T\boldsymbol{\beta}) \right).$$

We then obtain

$$\frac{dU(\mathbf{S},\boldsymbol{\beta})}{d\boldsymbol{\beta}} = -\sum_{l,t} S_{\bar{a}^l}(t-1) X_{l,t} X_{l,t}^T \frac{\exp(X_{l,t}^T\boldsymbol{\beta})}{(1+\exp(X_{l,t}^T\boldsymbol{\beta}))^2}$$

and that for each unique exposure patterns $\bar{a}_t^l$,

$$\frac{dU(\mathbf{S},\boldsymbol{\beta})}{dS_{\bar{a}^l}(t)} = X_{l,t} - \sum_{m:\{\bar{a}_t^l \subset \bar{a}_{t+1}^m\}} X_{m,t+1}[1+\text{expit}(X_{m,t+1}^T\boldsymbol{\beta})].$$

The above summation is taken over all m for which the truncated exposure pattern $\bar{a}_t^l$ is a subset of the pattern $\bar{a}_{t+1}^m$ (or, equivalently, $\bar{a}_t^m = \bar{a}_t^l$) so that in particular, $S_{\bar{a}^m}(t) = S_{\bar{a}^l}(t)$.

Substituting these expressions into Equation (6.4) gives a form for $d\boldsymbol{\beta}(\mathbf{S})/d\boldsymbol{\beta}$ which can then be substituted into Equation (6.3) to produce the form of the efficient influence function for the parameters of the MSM:

$$D_{\boldsymbol{\beta}} = \left[\sum_{l,t} S_{\bar{a}^l}(t-1) X_{l,t} X_{l,t}^T \frac{\exp(X_{l,t}^T\boldsymbol{\beta})}{(1+\exp(X_{l,t}^T\boldsymbol{\beta}))^2}\right]^{-1}$$
$$\sum_{l,t}\left\{X_{l,t} - \sum_{m:\{\bar{a}_t^l \subset \bar{a}_{t+1}^m\}} X_{m,t+1}[1+\text{expit}(X_{m,t+1}^T\boldsymbol{\beta})]\right\} D_{\bar{a}^l,t}.$$

The efficient influence function components can be calculated for each subject, producing an influence matrix of dimension $n \times R$.

Fitting procedure

To obtain the point estimates of the MSM parameters, first calculate the hazard functions for each exposure pattern and time using $\lambda_{\bar{a}^l}(t) = \{S_{\bar{a}^l}(t) - S_{\bar{a}^l}(t-1)\}/S_{\bar{a}^l}(t-1)$. Then, using these values as outcome measurements, fit the logistic regression with a choice of linear specification, with weights equal to $S_{\bar{a}^l}(t-1)$. The

variance estimates are obtained as in the previous model, by fitting the efficient influence function matrix for the $\boldsymbol{\beta}$ parameter and taking the empirical variance of each column.

6.4 Simulations

To demonstrate the performance of the TMLE for survival data with time-dependent confounders, we generated data of the form
$(W, A_1, L_1, Y_1, ..., A_5, L_5, Y_5)$ using known data-generating functions. W is a continuous baseline confounder, $A_t, t = 1, ..., 5$ are the binary exposure variables and $Y_t, t = 1, ..., 5$ are the survival indicators at each time point. The exposure is monotone (once exposed, always exposed). L_t is a binary variable that acts as a time-varying confounder. Each variable (unless determined by the monotonicity of exposure and survival) was generated dependent on the baseline and the covariate values at the previous time point according to the general rule that exposure reduces the probability of survival at the next time point as do higher values of L_t. Censoring was not included in the simulation study (with the exception of administrative censoring from the end-of-study, which occurred after the fifth time point for all subjects). Code for the data generation is provided in the supplementary materials.

The TMLE method described in the previous sections was evaluated in its ability to predict $S^{\bar{a}=1}(5)$, the probability of survival at the fifth time point under the counterfactual condition of having all subjects exposed at the first time point. The TMLE was compared to the Adjusted Kaplan-Meier Estimator (AKME), the inverse probability of treatment weighting method for the Kaplan-Meier curve described in [86]. Both methods were implemented using logistic regressions to estimate all

probabilities. The standard error for TMLE was estimated using its efficient influence function, and for AKME using the non-parametric bootstrap. The non-parametric bootstrap was performed by taking 500 resampled data sets with replacement from the complete data set. Each resampled data set was the same size as the original. The standard error was found by taking the standard deviation of the estimates calculated from the resampled data sets. The 95% confidence intervals for TMLE were estimated using the Normal approximation and the standard error from the efficient influence function. The confidence intervals for AKME used the 2.5th and 97.5th quantiles of the estimates from 500 bootstrap resamples.

Because of the way the data were generated, the models for each of the $Q_t^{\bar{a}}(j)$'s in the TMLE procedure were always misspecified (even when they included the correct set of confounders). It was possible to correctly specify the exposure model, so the unbiasedness of the TMLE in this simulation study is a result of the method's double-robustness.

In the simulation study, 1,000 data sets were drawn with sample sizes 2,500 and 5,000. AKME and TMLE were both implemented so that the exposure models were correctly specified. Table 6–1 (top) shows the simulation results for the estimation of the counterfactual probability of survival at the final time point under a history of always being exposed. For both sample sizes, AKME and TMLE perform very similarly, both producing unbiased estimates, and identical mean-squared errors (MSE) and standard errors (SE). Coverage was also close to 95% for both methods.

TMLE was then evaluated in its ability to estimate the parameters of a marginal structural model for the hazard (Section 6.3.3). The model evaluated was $\text{logit}\lambda_{\bar{a}^l}(t) =$

Table 6–1: Simulation results for (above) the probability of survival at time five under always-exposed, and (below) the coefficient of cumulative exposure in the hazard model (β_1). Correct exposure model used. Estimates taken over 1,000 generated datasets. True value for survival = 0.274; true value for MSM = 0.099

Method	Bias	MSE	SE	% COV
Survival				
		$n = 2500$		
TMLE	<0.001	0.001	0.024	94.3
AKME	<0.001	0.001	0.024	94.7
		$n = 5000$		
TMLE	<0.001	<0.001	0.017	96.0
AKME	<0.001	<0.001	0.017	95.9
MSM				
		$n = 2500$		
TMLE	<0.001	<0.001	0.028	96.3
IPTW-MSM	<0.001	<0.001	0.025	94.4
		$n = 5000$		
TMLE	<0.001	<0.001	0.019	96.2
IPTW-MSM	<0.001	<0.001	0.018	94.5

MSE: mean squared error calculated over the 1000 datasets; SE: influence curve estimate of the standard error for TMLE, and bootstrap resampled estimate for AKME and IPTW-MSM; % COV: percent coverage, calculated for IPTW using quantiles of the bootstrap-resampled estimates; TMLE: Targeted Maximum Likelihood Estimator; AKME: adjusted (inverse-weighted) Kaplan-Meier curve; IPTW-MSM: inverse probability of treatment weighted marginal structural model.

$\beta_0+\beta_1\text{cum}(\bar{a}_t^l)+\beta_2 t$ where β_1, the coefficient for the cumulative number of past times exposed, was the parameter of interest. Since the data was not generated from this MSM, the β parameters represent a likelihood projection of the survival probabilities at each time point onto a linear model [33]. TMLE was compared to the IPTW method for fitting the hazard MSM described in [21]. The IPTW was fit with unstabilized weights and its standard error was estimated using the nonparametric bootstrap (with the same specifications as for AKME). The MSM results in Table 6–1 (bottom) indicate that while both methods produced unbiased inference, for the lower sample size, TMLE had a slightly higher estimated standard error, resulting in slightly inflated confidence intervals and coverage. The standard errors for $n = 5,000$ coincided for the two methods.

6.5 The impact of HCV clearance on ESLD

The Canadian Co-Infection Cohort (CCC) study [26] follows a population of patients co-infected with HIV and HCV recruited from 16 Canadian centres. Participants are scheduled for appointments every six months, with data collected on status of treatments, lab tests describing disease progression, and drug and alcohol use at each follow up visit. While all patients were exposed to HCV, which can produce permanent damage to the liver when the infection becomes chronic, some clear the virus either through natural immunity or after HCV treatment. Our scientific question of interest is whether the clearance of HCV reduces the rate of onset of end-stage liver disease (ESLD).

At the time of data extraction, the study had collected data on 1,055 individuals. Ten were described as transgendered and were removed from the analysis (due to

their small number). Of those remaining, 778 had not cleared HCV at the time of cohort entry and had not yet been diagnosed with ESLD. Among those still actively co-infected with HCV (as determined by the presence of HCV RNA in plasma) and at-risk for ESLD, 38 had Hepatitis B and were excluded from the analysis as chronic Hepatitis B is itself a very strong risk factor for progressive liver disease. Therefore, 740 subjects were included in the analysis. The median follow-up in this subgroup was two years after baseline, sometimes including missed visits.

Potential baseline confounders were considered to be age, HIV duration, HCV duration, gender, and education. Potential time-dependent confounders (collected at baseline and at subsequent visits) were CD4 cell count, whether the participant was receiving antiretroviral therapy, HCV treatment status, and whether the participant had reported drinking alcohol in the past six months.

Characteristics of the sample used in the analysis are given in Table 6–2. Missing data were present, including the baseline covariates. The population generally consisted of patients who had been infected with HCV and HIV for a long duration. While most were receiving antiretroviral therapy to control their HIV infection, few received treatment for HCV. Approximately 25% of the sample was female.

We chose to perform the analysis using six visits after the baseline visit (equivalent to a follow-up of three years) due to the data becoming excessively sparse afterwards. Subjects often missed their biannual visits, and in addition, the time-varying covariates, exposure and development of ESLD were all subject to irregular (i.e. non-monotone) missingness. We assumed that exposure was monotone: if a

Table 6–2: Characteristics at baseline of the 740 co-infected subjects.

Characteristic	Summary		N. Missing
Numeric variables	Median	IQR	
Age (years)	44	(39,50)	2
HIV duration (years)	11	(6,16)	20
HCV duration (years)	18	(11,25)	4
CD4 cell count	380	(242,540)	16
Binary variables	N.	%	
Female	227	25	1
Education: $\geq$ high school	760	83	0
Taking ARVs	735	80	1
Currently treated for HCV	28	3	0
Drank alcohol in past 6 months	455	50	3

ARV: Antiretroviral therapy; IQR: inter-quartile range.

patient had cleared HCV, we considered them permanently clear. Using the assumption of monotonicity, we were able to complete some of the missing exposure data. Similarly, the outcome event (diagnosis of ESLD) carries the assumption of monotonicity, which also allowed us to logically impute some values. Subjects often dropped out of the study, without documentation. A subject was assumed to be censored if they missed three visits in a row, or died from a cause unrelated to ESLD. If a subject died from liver complications, they were considered to have experienced the event. Table 6–3 reports the counts for the number of subjects at risk and the failure incidence at each time point, by exposure status (when known, and when unknown). The time-dependent exposure status is defined as having cleared HCV at some previous time. Therefore, a subject would be included in the "unexposed" group until clearing HCV, at which point they would be considered part of the "exposed" group. Exposure status is unknown if the subject had not yet been observed

to have cleared HCV and no test was done at that time point (often due to a missed visit).

Table 6–3: Number at-risk and failure incidence by time point and exposure status (when known)

Status	Visit	1	2	3	4	5	6
Unexposed	N. at-risk	380	294	214	159	102	78
	N. failed	22	16	12	4	4	4
Exposed	N. at-risk	29	62	80	85	84	76
	N. failed	0	3	1	1	1	2
Unknown	N. at-risk	320	325	216	195	197	162
	N. failed	14	9	10	4	5	3

Many subjects had an unknown exposure status at various time points. These subjects were omitted from the table calculations for any time points that their exposure status was uncertain, but were included in the analysis with the help of multiple imputation.

The data structure can be described as $O = (W, L_1, A_1(1), A_1(2), Y_1, ..., L_6, C_6, A_6(1), A_6(2), Y_6)$ where W is the collection of baseline covariates, L_t is the multivariate time-dependent confounders, $A_t(1)$ is whether or not HCV has been cleared, $A_t(2)$ is a censoring indicator, and Y_t is an indicator for ESLD at time t.

Due to the large amount of missing data in the data set (in particular, due to many missed visits), we chose to employ multiple imputations [59] as part of our analytical strategy to account for non-censoring missingness. The validity of multiple imputations relies on the assumption that the data are missing at random conditional on the variables used to impute. In addition, it relies on correct specification of the imputation models. We built the imputation models using all of the variables included in the analysis (time-varying confounders, baseline confounders, exposure

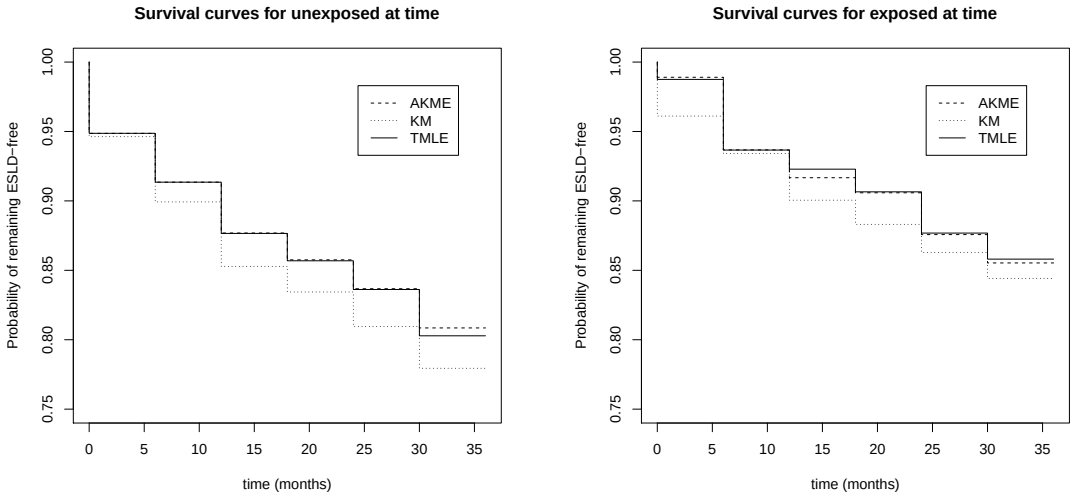
and outcome). The imputation models chosen allowed each variable to be imputed conditional on previously or simultaneously collected variables so that future information was never used. Multivariate Imputation by Chained Equations (MICE) was performed using the R package `mice` [72]. After a burn-in of 20 draws, 50 imputations were drawn with 20 lagged iterations each. Logistic regression was used to impute all binary variables, and Bayesian linear regression was used for all normal variables (including CD4 cell count, which was imputed on a logistic scale). Each analytical method used was performed on each imputed data set, and the estimates and standard errors obtained were combined according to the usual methodology introduced in [59] to produce the final inference.

The probabilities of survival for both exposure states at each time point were calculated using the Kaplan-Meier estimator, the Adjusted Kaplan-Meier Estimator (AKME) [86], and a version of the TMLE described above. AKME incorporated inverse probability weighting to adjust for censoring and non-randomized exposure. All probabilities were estimated with logistic regression using the baseline, previous time point covariates, and an indicator of whether or not the visit was missing in each model. Unlike the simulation study, exposure patterns were not considered in this analysis, so the probabilities of exposure at times t (used in both AKME and TMLE) were modified to be the probabilities of *ever* being exposed prior to time t. Due to the sparsity of failures among the exposed subjects, we implemented a reduced model for the outcome (but not the exposure or censoring) in the TMLE procedure. The reduced model was selected primarily based on substantive knowledge and computational feasibility.

The survival curves estimated with Kaplan-Meier, AKME, and TMLE, all using multiple imputations, are presented in Figure 6–1 for both exposure groups. Each point on the curves can be interpreted as the estimate of the counterfactual survival at time t for a subject who has been exposed prior to time t. The curve representing the mean probabilities of remaining ESLD-free when exposed appears to be underestimated at each time point by the unadjusted Kaplan-Meier estimator. TMLE and AKME give very similar estimates, with TMLE often estimating a slightly higher probability of remaining ESLD-free. For the counterfactual survival curve for an unexposed population, the unadjusted Kaplan-Meier again appears to underestimate the probabilities of remaining free of ESLD compared to the adjusted methods. Subfigure 6–1(c) compares the survival curves for the exposed and unexposed with TMLE. From this graph, we observe that the estimated probability of remaining ESLD-free under HCV clearance is higher than without clearance.

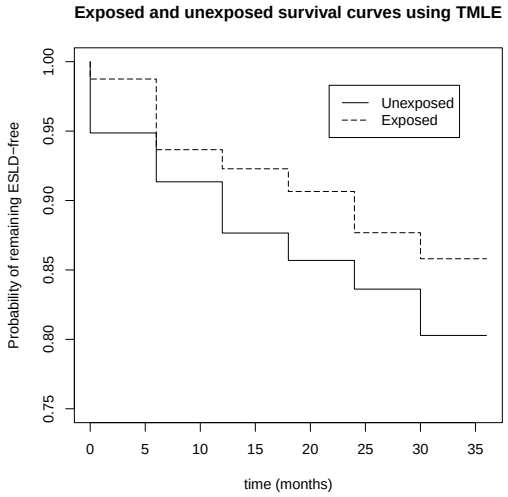
Figure 6–2 presents the TMLE estimates and 95% confidence intervals of the survival curves for exposed and unexposed at each time point. The large confidence intervals are a consequence of the sparse data and exceed the parameter bounds due to the assumption of normality, which could be corrected with the development of an exact confidence interval or usage of the nonparametric bootstrap.

A MSM for the hazard of developing ESLD was defined using a logistic mean model: $\text{logit}\lambda_a(t) = \gamma_0 + \gamma_1 a(t) + \gamma_2 t$ where $a(t)$ is the binary exposure status at time t, and $\lambda_a(t)$ is the counterfactual hazard at time t for exposure status $a(t)$. A negative value for γ_1 would indicate that clearing HCV has a protective effect against



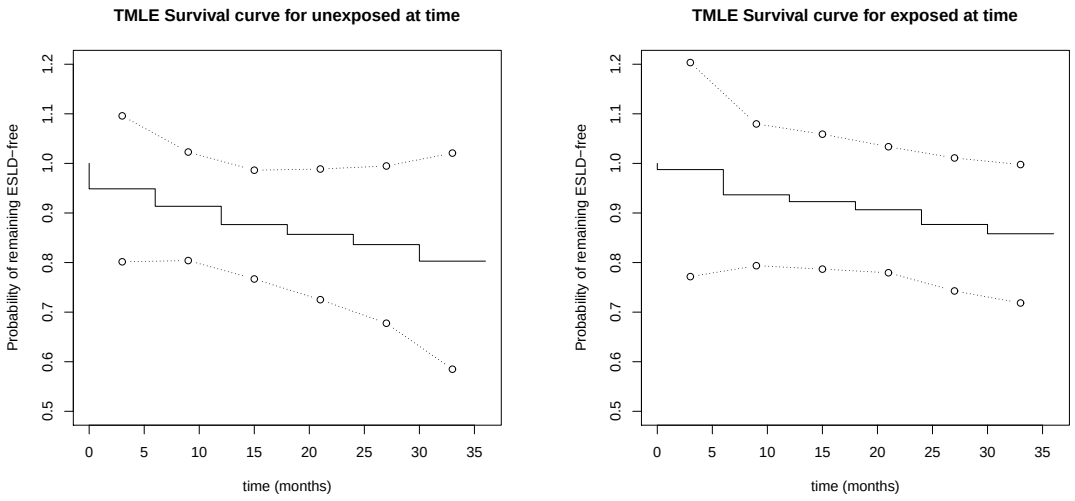
(a)

(b)



(c)

Figure 6–1: Survival curves for subjects (a) unexposed and (b) exposed at given time. Curves were calculated with inverse probability of treatment weighting (IPTW), unadjusted Kaplan-Meier (K-M), and Targeted Maximum Likelihood Estimation (TMLE).



(a) (b)

Figure 6–2: Survival curves for subjects (a) unexposed and (b) exposed at given time. Curves were calculated with Targeted Maximum Likelihood Estimation (TMLE). For each imputed dataset, the variance was estimated using the influence curve. The total variance was calculated by combining the variance of the estimate for each imputed dataset and the variance between the estimates from the imputed datasets. Pointwise 95% confidence intervals were calculated using estimate ± 1.96*SE.

developing ESLD (as it decreases the hazard of failing). A positive value for γ_2 would mean that hazard increases over time.

This MSM was fit using the three different methods shown in Table 6–4, each incorporating the multiple imputations. The unadjusted logistic regression was fit for all subject-times, with a robust sandwich estimator to estimate the standard error of each coefficient (using R library `sandwich` [87, 88]). The MSM was also fit using IPTW (adjusting for both non-randomized exposure and censoring) with stabilized weights. The standard errors were estimated using the same robust sandwich estimator. The TMLE was fit by combining the estimates for the survival curves in order to model the hazard as described in Section 6.3.3.

The results for γ_1 indicate that the coefficient for exposure status was estimated as -0.34 but not significantly different than zero at the 0.05 level when using the naive method (which does not adjust for confounding or dependent dropout). TMLE and IPTW estimated an effect magnitude of -0.38. TMLE had a 34% smaller standard error than IPTW, but neither estimator found a significant effect of interest. Including the indicator for a missing visit in the exposure model made an important difference in this analysis, because excluding the indicator resulted in a TMLE estimate for γ_1 of -0.50 (SE=0.33).

All of the models yielded a negative parameter estimate for γ_2, but only the unadjusted method concluded that it was statistically significant. TMLE and IPTW produced estimates with much smaller magnitudes which were more plausible results as a marginal hazard of ESLD that decreases with time is unlikely. The incorrect estimate produced by the unadjusted model may result from sicker patients leaving

the study so that the cohort appears to be improving in health, and therefore, the hazard appears to decrease over time. Here, the standard error for TMLE was approximately half the size of the IPTW standard error.

Table 6–4: MSM results: Logistic model for hazard of developing end-stage liver disease.

Method	Est	SE	95% CI	p-val
γ_1 *Coefficient of exposure status*				
Unadjusted	-0.34	0.33	(-0.95,0.28)	0.28
IPTW	-0.38	0.49	(-1.34,0.58)	0.49
TMLE	-0.38	0.32	(-1.01,0.25)	0.24
γ_2 *Coefficient of time*				
Unadjusted	-0.13	0.07	(-0.26,-0.00)	0.04
IPTW	-0.05	0.20	(-0.44,0.34)	0.80
TMLE	-0.06	0.09	(-0.25,0.12)	0.58

Unadjusted: Unweighted logistic regression, standard error calculated using robust sandwich estimator; IPTW: Inverse probability of treatment weighted logistic regression, standard error calculated using robust sandwich estimator; TMLE: Targeted Maximum Likelihood Estimation for survival data, standard error calculated from efficient influence curve. Each method was performed on 50 multiply-imputed datasets and the results or each analysis combined according to [59].

6.6 Discussion

Many parameters and types of marginal structural models can be defined to compare survival curves, and we have demonstrated how to construct a TMLE for two different model types. The model for the hazard that we implemented for both the simulation and the example can be directly compared with the well-known IPTW method for estimating a MSM with pooled logistic regression, and when the hazard at all time points is small, to a marginal structural Cox model [21, 11].

In this paper we also used this method of estimating survival curves and a MSM for the hazard in an example where we evaluated the effect of clearing HCV on time-to-ESLD. A major challenge particular to the TMLE was the necessity to fit outcome

models for failure at every time point. The rarity of events in the CCC data made this very difficult, even after using multiple imputations to fill in the missing values. In particular, our original goal was to compare different exposure patterns (much like in the simulation study and described in the methods) but were prevented by the sparsity in the outcome. IPTW does not require fitting an outcome model, and may therefore be more flexible in similar situations. However, from the results of the example, even our reduced outcome model appeared to produce an improvement in the estimation of the MSM by TMLE, so it may generally be worthwhile to attempt to fit a TMLE even in a challenging data environment.

Clearance of HCV occurs both spontaneously and due to HCV treatment, and the subsequent risk of liver damage might differ depending on the situation. However, subgroup analysis is prohibited by the data sparsity described and is therefore not possible with the currently available information. Partially due to different types of viral clearance, the causal relationship of viral clearance on ESLD may indeed be more complicated than was represented in our simple MSM. Further analyses should also consider the different ethnic groups participating in the study, including the Aboriginal subpopulation (representing 15% of our sample) who may clear HCV more easily than the general population [32, 64]. Due to the potentially complex confounding due to ethnicity that was not fully addressed (primarily due to data sparsity), this analysis should be taken primarily as an illustration of the TMLE method.

Multiple imputations were used to adjust for missing values. The validity of this method relies on the assumption that the missing values were missing at random [55]

(only dependant on observed variables) conditional on the variables used in the imputation models. This methodology was fundamental in allowing us to use as much of the information in the data set as possible, as the missingness was often irregular (comprising of both intermittently missing visits and additional missingness in the covariates) and the usable sample size was not large. We preferred to use multiple imputations over simpler imputation methods which require stronger (and in our opinion, untenable) assumptions about the nature of the missing data [3]. We found complete case analysis to be impossible as very few subjects had complete data. Multiple imputations have been proposed for and used in causal inference studies [58, 70, 66, 65].

Assessments of the performance of the TMLE for estimating marginal longitudinal or survival parameters described in this paper and comparisons to other causal methods have also been obtained through simulation study in [74] and [63]. In our simulation study, we confirmed the unbiasedness of this TMLE under misspecification of the outcome model when estimating the survival curve (a partial demonstration of its double-robustness). We also numerically confirmed the unbiasedness and efficiency of the extension of the method for estimating the parameters of a marginal structural model, again under misspecification of the outcome models. The slightly higher standard errors obtained for TMLE when compared to IPTW can be explained by the different methods used to estimate the variance. For IPTW, we used nonparametric bootstrap resampling. For TMLE, we used the influence curve-based sandwich estimator, which is known to be conservative for misspecified Q-models [77].

6.7 Supplementary material for Manuscript 3: Simulation details

6.7.1 Target parameter of the IPTW

Here we show that the inverse probability of treatment weighted (IPTW) estimator for the hazard model used in the simulation study and the example targets the same parameter of interest as the TMLE.

Let $\lambda_\beta(\bar{a}, t) = \text{expit}(X_{l,t}^T \boldsymbol{\beta})$ be the logit-linear model for the hazard. The parameter estimated by the Targeted Maximum Likelihood Estimator was defined as

$$\text{argmax}_{\boldsymbol{\beta}} \sum_{l,t} S_{\bar{a}^l}(t-1) \left\{ \lambda_{\bar{a}^l}(t) \log(\lambda_\beta) + [1 - \lambda_{\bar{a}^l}(t)] \log(1 - \lambda_\beta) \right\}.$$

Factoring out $\lambda_{\bar{a}^l}(t)$ gives

$$\begin{aligned}
&\text{argmax}_{\boldsymbol{\beta}} \sum_{l,t} S_{\bar{a}^l}(t-1) \lambda_{\bar{a}^l}(t) \left\{ [\log \lambda_\beta - \log(1 - \lambda_\beta)] + \log(1 - \lambda_\beta) \right\} \\
&= \text{argmax}_{\boldsymbol{\beta}} \sum_{l,t} S_{\bar{a}^l}(t-1) \lambda_{\bar{a}^l}(t) \left\{ \log \left[\frac{\lambda_\beta}{(1 - \lambda_\beta)} \right] + \log(1 - \lambda_\beta) \right\}.
\end{aligned}$$

To maximize this expression, the derivative with respect to $\boldsymbol{\beta}$ can be taken and the resulting expression set to zero. This results in

$$\begin{aligned}
&\sum_{l,t} S_{\bar{a}^l}(t-1) \lambda_{\bar{a}^l}(t) \left\{ \frac{d}{d\boldsymbol{\beta}} \log \left[\frac{\lambda_\beta}{(1 - \lambda_\beta)} \right] + \frac{d}{d\boldsymbol{\beta}} \log(1 - \lambda_\beta) \right\} \\
&= \frac{d}{d\boldsymbol{\beta}}(\lambda_\beta) \frac{1}{\lambda_\beta (1 - \lambda_\beta)} \left[\lambda_{\bar{a}^l}(t) - \lambda_\beta \right] \\
&= \frac{d}{d\boldsymbol{\beta}} \text{logit}(\lambda_\beta) \left[\lambda_{\bar{a}^l}(t) - \lambda_\beta \right] = 0.
\end{aligned}$$

Noting that $\text{logit}(\lambda_\beta) = X_{l,t}^T \boldsymbol{\beta}$ is the linear specification so that the derivative with respect to $\boldsymbol{\beta}$ is the vector of variables in the marginal structural model. The above score equation is therefore the logistic regression defined on the counterfactuals that is solved by the IPTW-MSM.

6.7.2 Data simulation

We generated data of the form $(W, A_1, L_1, S_1, ..., A_5, L_5, S_5)$ using known data generating functions. W is a continuous baseline confounder, $A_t, t = 1, ..., 5$ are the binary exposure variables and $S_t, t = 1, ..., 5$ are the survival indicators at each time point. The exposure generated was monotone (once exposed, always exposed). L_t is a binary variable that acts as a time-varying confounder. Each variable (unless determined by the monotonicity of exposure and survival) was generated dependent on the baseline and the covariate values at the previous time point according to the general rule that exposure reduces the probability of survival at the next time point as do higher values of L_t. Censoring was not included in the simulation study.

We used the following function (written in R Statistical Software version 2.13.2, [39]) to generate the data:

```
data_surv_new<-function(i,ssize){
set.seed(i*5436)

W<-rnorm(n=ssize)/4+1

#TP1

p1<-expit(-4.2+2.5*W)
A1<-rbinom(n=ssize,size=1,prob=p1)
```

```
mu1<-expit(1+W+0.5*A1)
L1<-rbinom(n=ssize,size=1,prob=mu1)

s1<-expit(2+W-0.7*L1-0.5*A1)
S1<-rbinom(n=ssize,size=1,prob=s1)

#TP2

A2<-rep(0,length=ssize)
p2<-expit(-3.2+1*W+1.2*L1)
A2[A1==0&S1==1]<-rbinom(n=sum(A1==0&S1==1),size=1,prob=p2[A1==0&S1==1])
A2[A1==1&S1==1]<-1

mu2<-expit(1+L1+0.5*A2)
L2<-rbinom(n=ssize,size=1,prob=mu2)

S2<-rep(1,ssize)
s2<-expit(1.6+W-0.7*L2-0.5*A2)
S2[S1==1]<-rbinom(n=sum(S1==1),size=1,prob=s2[S1==1])
S2[S1==0]<-0

#TP3

A3<-rep(0,length=ssize)
p3<-expit(-2.9+1*W+1.2*L2)
A3[A2==0&S2==1]<-rbinom(n=sum(A2==0&S2==1),size=1,prob=p3[A2==0&S2==1])
A3[A2==1&S2==1]<-1
```

```
mu3<-expit(1+L2+0.5*A3)
L3<-rbinom(size=1,prob=mu3,n=ssize)

S3<-rep(0,ssize)
s3<-expit(2.5+0.8*W-0.7*L3-0.5*A3)
S3[S2==1]<-rbinom(n=sum(S2==1),size=1,prob=s3[S2==1])
S3[S2==0]<-0

#TP4

A4<-rep(0,length=ssize)
p4<-expit(-2+0.5*W+1.2*L3)
A4[A3==0&S3==1]<-rbinom(n=sum(A3==0&S3==1),size=1,prob=p4[A3==0&S3==1])
A4[A3==1&S3==1]<-1

mu4<-expit(1+L3+0.5*A4)
L4<-rbinom(prob=mu4,size=1,n=ssize)

S4<-rep(0,ssize)
s4<-expit(1.2+W-0.7*L4-0.5*A4)
S4[S3==1]<-rbinom(n=sum(S3==1),size=1,prob=s4[S3==1])
S4[S3==0]<-0

#TP5

A5<-rep(0,length=ssize)
p5<-expit(-1+0.5*W+1.2*L4)
```

```
A5[A4==0&S4==1]<-rbinom(n=sum(A4==0&S4==1),size=1,prob=p5[A4==0&S4==1])
A5[A4==1&S4==1]<-1

mu5<-expit(1+L4+0.5*A5)
L5<-rbinom(prob=mu5,size=1,n=ssize)

S5<-rep(0,ssize)
s5<-expit(1+0.5*W-0.7*L5-0.5*A5)
S5[S4==1]<-rbinom(n=sum(S4==1),size=1,prob=s5[S4==1])
S5[S4==0]<-0

#If dead, make missing
A2[S1==0]<-NA
L2[S1==0]<-NA
A3[S2==0]<-NA
L3[S2==0]<-NA
A4[S3==0]<-NA
L4[S3==0]<-NA
A5[S4==0]<-NA
L5[S4==0]<-NA


return(as.data.frame(cbind(W,L1,L2,L3,L4,L5,A1,A2,A3,A4,A5,S1,S2,S3,S4,S5)))

}
```

CHAPTER 7
Conclusions

7.1 Summary

The three manuscripts (Chapters 4–6) presented in this thesis describe a body of work all within the context of the estimation of causal parameters in longitudinal data settings using Targeted Maximum Likelihood Estimation. The first work (Chapter 4) demonstrates the flexibility of construction in a simplified longitudinal context, obtained by varying the TMLE loss and fluctuation function in a coordinated manner. The second work (Chapter 5) shows how the construction of a TMLE is also flexible in the choice of plug-in estimator used, and investigates a different type of longitudinal TMLE that can be used easily for more complicated data structures. This study goes a step further and also incorporates Super Learning, a nonparametric multi-library machine learning method, into the estimation procedure, consequently illustrating how TMLE can be implemented as a fully nonparametric method. The final manuscript (Chapter 6) demonstrates how survival data with time-dependent confounding may be analyzed in the same way as longitudinal data. It shows how TMLE can be used to estimate the parameters of an unsaturated marginal structural model. This last study applies the longitudinal estimator described in the second manuscript to a survival context and describes a challenging analysis that fully takes advantage of the flexibility of this estimator to naturally incorporate many time-dependent confounders.

The simulation studies in each chapter add to a growing understanding of the finite sample performance of TMLE. In the first manuscript, the simulation study exhibits the benefits of TMLE over other causal estimators, in particular in terms of double robustness and stability under misspecification. The simulation study in the second manuscript demonstrates a key benefit of using TMLE over efficient estimating equations; in the situation with near-positivity violations, the TMLE produced far less bias and smaller standard errors than the related estimating equation. In the third manuscript, the simulation study is used to demonstrate the unbiasedness of the TMLE for estimation of both a saturated and unsaturated MSM in the survival context when partially misspecified.

The first and second manuscripts also contain different analyses of the PROBIT study, with the objective of estimating the impact of breastfeeding on gastrointestinal tract infections in infants. The first PROBIT analysis utilizes a simplified version of the dataset (two time-intervals and only subjects with complete data) in order to demonstrate the TMLE method described in the first manuscript. The second, more sophisticated, analysis uses the full six follow-up times and allows for censoring. The TMLE method described in the second manuscript is used to estimate the expected number of infections if the population of mothers had been breastfeeding for different durations of time, adjusting for loss to follow-up in addition to both baseline and time-dependent confounding. The comparison between these different regimens suggests that a longer duration of breastfeeding might decrease the number of infections in the defined population of infants.

The third manuscript presents an analysis of the effect of clearing the Hepatitis C virus on the occurrence of end-stage liver disease. The TMLE method described in the second manuscript was implemented for this case study, chosen for its ability to easily include many time-dependent confounders without added modeling steps. The large amount of missing data in the exposure, intermediate variables and outcome (the majority of which came from missed visits) led to the incorporation of multiple imputation by chained equations within the TMLE procedure. This analysis did not find a statistically significant effect of viral clearance on the time-dependent hazard of ESLD, but this could be attributed to the large amount of missing data, the relatively small number of subjects who had cleared the virus, or the various modeling choices. Additional investigation involving different choices of analytic methods may be beneficial.

7.2 Future work

The application in the third manuscript demonstrated how useful the method of multiple imputations can be when performing an analysis in the presence of non-dropout missing data. In this manuscript, we employed the method in a somewhat ad-hoc fashion. Further work on this topic would involve formalizing the inclusion of multiple imputations in the TMLE framework and then comparing different methods for missing data under scenarios with missing information on the baseline confounders, exposure, time-dependent confounders, or outcome. This work would assist analysts using TMLE in deciding which missing data method would be most appropriate when faced with different types of missing data.

The effects of clearance of HCV on risk of ESLD could also be investigated further by considering more carefully the different ways that the virus can be cleared, which is also related to the different ethnic groups participating in the study. The Canadian Co-Infection Cohort is an ongoing study, so further substantive analyses could benefit from the collection of additional data. In particular, different MSMs could be investigated with the help of clinical investigators. As a methodological improvement, it would potentially be beneficial to reduce the parametric assumptions by incorporating a Super Learner to fit the exposure, censoring and outcome densities (as was done in the second manuscript).

On a larger scale, I plan on proceeding with additional work on different TMLE methods for structural nested models in the context of longitudinal data where the sequential randomization assumption is not expected to hold [22].

7.3 Concluding remarks

Targeted maximum likelihood estimation is a beneficial modeling choice for a number of reasons, most notably due to gains in efficiency, double-robustness and flexibility of implementation (i.e. usage of Super Learner, choice of loss function, etc). The study of TMLE is also instructive because of the importance of selecting the target parameter. This emphasizes the need to understand and carefully identify the parameter that is being estimated in the analysis. Better understanding of this modeling option should prove useful for causal inference specialists, and more widespread usage of TMLE could result in better estimation and improvements in statistical practice.

References

[1] H Bang and J M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61:962–972, 2005.

[2] O Bembom and M J van der Laan. Data-adaptive selection of the truncation level for inverse-probability-of-treatment-weighted estimators. *U.C. Berkeley Division of Biostatistics Working Paper Series*, (Working Paper 230), 2008.

[3] C Beunckens, G Molenberghs, and M G Kenward. Direct likelihood analysis versus simple forms of imputation for missing data in randomized clinical trials. *Clinical Trials*, 2:379–386, 2005.

[4] V P Bhapkar. On a measure of efficiency of an estimating equation. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 34(4):467–472, 1972.

[5] P J Bickel and K A Doksum. *Mathematical Statistics*, volume 1. Upper Saddle River, N.J.: Prentice Hall, 2nd edition, 2001. Section 2.1.2, p. 104.

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

[7] P Chaffee and M J van der Laan. Targeted maximum likelihood estimation for dynamic treatment regimes in sequential randomized controlled trials. *U.C. Berkeley Division of Biostatistics Working Paper Series*, (Working Paper 277), 2011.

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

[9] S R Cole, M A Hernán, J B Margolick, M H Cohen, and J A Robins. Marginal structural models for estimating the effect of highly active antiretroviral therapy initiation on cd4 cell count. *American Journal of Epidemiology*, 162(5):471–478, 2005.

[10] H Cramér. A contribution to the theory of statistical estimation. *Scandinavian Actuarial Journal*, 1:85–94, 1946.

[11] R B D'Agostino, M Lee, A J Belanger, L A Cupples, K Anderson, and W B Kannel. Relation of pooled logistic regression to time dependent cox regression analysis: The framingham heart study. *Statistics in Medicine*, 9:1501–1515, 1990.

[12] R H Dehejia and S Wahba. Propensity score-matching methods for nonexperimental causal studies. *The Review of Economics and Statistics*, 84(1):151–161, 2002.

[13] M Finster and M Wood. The apgar score has survived the test of time. *Anesthesiology*, 102(4):855–857, 2005.

[14] B Freedman. Equipoise and the ethics of clinical research. *The New England Journal of Medicine*, 317(3):141–145, 1987.

[15] R D Gill and J M Robins. Causal inference for complex longitudinal data: The continuous case. *The Annals of Statistics*, 29(6):1785–1811, 2001.

[16] S Greenland and H Morgenstern. Confounding in health research. *Annual Review of Public Health*, 22:189–212, 2001.

[17] S Greenland, J M Robins, and J Pearl. Confounding and collapsibility in causal inference. *Statistical Science*, 14(1):29–46, 1999.

[18] S Gruber and M J van der Laan. A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. *The International Journal of Biostatistics*, 6(1):Article 26, 2010.

[19] T Hastie. *gam: Generalized Additive Models*, 2011. R package version 1.04.1.

[20] M A Hernán. The hazards of hazard ratios. *Epidemiology*, 21(1):13–15, 2010.

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

[22] M M Joffe, W P Yang, and H I Feldman. Selective ignorability assumptions in causal inference. *The International Journal of Biostatistics*, 6(2):Article 11, 2010.

[23] J D Kalbfleisch and R L Prentice. *The Statistical Analysis of Failure Time Data*. Wiley Series in Probability and Statistics. Wiley-Interscience, 2 edition, 2002.

[24] G Kallianpur and C R Rao. On Fisher's lower bound to asymptotic variance of a consistent estimate. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 15(4):331–342, 1955.

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

[26] MB Klein, S Saeed, H Yang, J Cohen, B Conway, C Cooper, P Côté, J Cox, J Gill, S Haider, M Harris, D Hasse, J Montaner, N Pick, A Rachlis, D Rouleau, R Sandre, M Tyndall, and S Walmsley. Cohort profile: The Canadian HIV-Hepatitis C Co-Infections Cohort study (CCC; CTN 222 study). *International Journal of Epidemiology*, 39(5):1162–1169, 2010.

[27] M S Kramer, B Chalmers, E D Hodnett, Z Sevkovskaya, I Dzikovich, S Shapiro, J P Collet, I Vanilovich, I Mezen, T Ducruet, G Shishko, V Zubovich, D Mknuik, E Gluchanina, V Dombrovskiy, A Ustinovitch, T Kot, N Bogdanovich, L Ovchinikova, and E Helsing. PROmotion of Breastfeeding Intervention Trial (PROBIT). *The Journal of the American Medical Association*, 285(4):413–420, 2001.

[28] M S Kramer, T Guo, R W Platt, S Shapiro, J P Collet, B Chalmers, E Hodnett, Z Sevkovskaya, I Dzikovich, and I Vanilovich. Breastfeeding and infant growth: Biology or bias? *Pediatrics*, 110(2):343–347, 2002.

[29] M S Kramer and S H Shapiro. Scientific challenges in the application of randomized trials. *Jounrnal of the American Medical Association*, 252(19):2739–2745, November 1984.

[30] G Maldonado and S Greenland. Estimating causal effects. *International Journal of Epidemiology*, 31:422–429, 2002.

[31] S Milborrow. *earth: Multivariate Adaptive Regression Spline Models*, 2011. Derived from mda:mars by Trevor Hastie and Rob Tibshirani.

[32] G Y Minuk, M Zhang, S G Wong, J Uhanova, C N Bernstein, B Martin, M R Dawood, L Vardy, and A Giulvi. Viral hepatitis in a canadian first nations community. *Canadian Journal of Gastroenterology*, 17:593596, 2003.

[33] R Neugebauer and M J van der Laan. Nonparametric causal effects based on marginal structural models. *Journal of Statistical Planning and Inference*, 137(2):419–434, 2007.

[34] J Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2 edition, 2009.

[35] A Peters and T Hothorn. *ipred: Improved Predictors*, 2011. R package version 0.8-11.

[36] E C Polley and M J van der Laan. Super Learner in prediction. *U.C. Berkeley Division of Biostatistics Working Paper Series*, (Working Paper 266), 2010.

[37] E C Polley and M J van der Laan. *Package 'SuperLearner'*. CRAN, 2.0-4 edition, 2011.

[38] K E Porter, S Gruber, M J van der Laan, and J S Sekhon. The relative performance of targeted maximum likelihood estimators. *U.C. Berkeley Division of Biostatistics Working Paper Series*, (Working Paper 279), 2011.

[39] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.

[40] C R Rao. Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37(3):81–91, 1945.

[41] C R Rao and F J Anscombe. Sufficient statistics and minimum variance estimates. *Proceedings of the Cambridge Philosphical Society*, 45(2):213–218, 1949.

[42] J M Robin, A Rotnitzky, and Zhao L P. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121, 1995.

[43] J M Robins. A new approach to causal inference in mortality studies with a sustained exposure period – application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7:1393–1512, 1986.

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

[45] J M Robins. Robust estimation in sequentially ignorable missing data and causal inference models. *Proceedings of the American Statistical Association Section on Bayesian Statistical Science*, pages 6–10, 2000.

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

[47] J M Robins and A Rotnitzky. Recovery of information and adjustment for dependent censoring using surrogate markers. *AIDS Epidemiology - Methodological Issues*, pages 297–331, 1992.

[48] J M Robins and A Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

[49] J M Robins, A Rotnitzky, and L P Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121, 1995.

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

[51] M Rosenblum and J M van der Laan. Targeted maximum likelihood estimation of the parameter of a marginal structural model. *The International Journal of Biostatistics*, 6(2):Article 19, 2010.

[52] M Rosenblum and M J van der Laan. Simple, efficient estimators of treatment effects in randomized trials using generalized linear models to leverage baseline variables. *The International Journal of Biostatistics*, 6(1):Article 13, 2010.

[53] M Rosenblum and M J van der Laan. Simple examples of estimating causal effects using targeted maximum likelihood estimation. *U.C. Berkeley Division of Biostatistics Working Paper Series*, 2010.

[54] D B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 65(5):688–701, 1974.

[55] D B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

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

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

[58] D B Rubin. Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics*, 31(2):161170, 2004.

[59] D B Rubin and N Schenker. Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81(394):366–374, 1986.

[60] A N Sapadin and R Fleischmajer. Treatment of scleroderma. *Archives of Dermatology*, 138(1):99–105, January 2002.

[61] D O Scharfstein, A Rotnitzky, and J M Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120, 1999.

[62] M E Schnitzer, E E M Moodie, and R W Platt. Targeted maximum likelihood estimation for marginal time-dependent treatment effects under density misspecification. *Biostatistics*, 2012. doi: 10.1093/biostatistics/kxs024.

[63] M E Schnitzer, M J van der Laan, E E M Moodie, and R W Platt. Effect of breastfeeding on gastrointestinal infection in infants: A targeted maximum likelihood approach for longitudinal data with censoring. *Journal of the American Statistical Society – Case Studies*, 2012. Submitted July 2012.

[64] J D Scott, B J McMahon, D Bruden, D Sullivan, C Homan, C Christensen, and D R. Gretch. High rate of spontaneous negativity for hepatitis c virus rna after establishment of chronic infection in alaska natives. *Clinical Infectious Diseases*, 42(7):945–952, 2006.

[65] S M Shortreed and A B Forbes. Missing data in the exposure of interest and marginal structural models: A simulation study based on the Framingham Heart Study. *Statistics in Medicine*, 29(4):431443, 2009.

[66] S M Shortreed and E E M Moodie. Estimating the optimal dynamic antipsychotic treatment regime: evidence from the sequential multiple-assignment randomized clinical antipsychotic trials of intervention and effectiveness schizophrenia study. *Journal of the American Statistical Society*, 2012. Accepted.

[67] J M Snowden, S Rose, and K M Mortimer. Implementation of g-computation on a simulated data set: Demonstration of a causal inference technique. *American Journal of Epidemiology*, 173(7):731–738, 2011.

[68] V D Steen, G R Owens, G J Fino, G P Rodnan, and T A Medsger Jr. Pulmonary involvement in systemic sclerosis. *Arthritis and Rheumatism*, 28(7):759–767, July 1985.

[69] O M Stitelman, V De Gruttola, and M J van der Laan. A general implementation of tmle for longitudinal data applied to causal inference in survival analysis. *U.C. Berkeley Division of Biostatistics Working Paper Series*, (Working Paper 281), 2011.

[70] L Taylor and X H Zhou. Multiple imputation methods for treatment noncompliance and nonresponse in randomized clinical trials. *Biometrics*, 65:88–95, 2009.

[71] A A Tsiatis. *Semiparametric Theory and Missing Data*. Springer Series in Statistics. Springer, 2006.

[72] S van Buuren and K Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67, 2011.

[73] M J van der Laan. Targeted maximum likelihood based causal inference: Part I. *The International Journal of Biostatistics*, 6(2):Article 2, January 2010.

[74] M J van der Laan and S Gruber. Targeted minimum loss based estimation of an intervention specific mean outcome. *U.C. Berkeley Division of Biostatistics Working Paper Series*, (Working Paper 290), 2011.

[75] M J van der Laan, E C Polley, and A E Hubbard. Super Learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1), 2007.

[76] M J van der Laan and J M Robins. *Unified Methods for Censored Longitudinal Data and Causality*. Springer Series in Statistics. Springer Verlag: New York, 2003.

[77] M J van der Laan and S Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer Series in Statistics. Springer, 2011.

[78] M J van der Laan and D Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1):Article 11, 2006.

[79] A W van der Vaart and J A Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer, 1996.

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

[81] T VanderWeele. Concerning the consistency assumption in causal inference. *Epidemiology*, 20(6):880–883, 2009.

[82] S Vansteelandt and N Keiding. Invited commentary: G-computationlost in translation? *American Journal of Epidemiology*, 173(7):739–742, 2011.

[83] W N Venables and B D Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.

[84] D Westreich and S R Cole. Invited commentary: Positivity in practice. *American Journal of Epidemiology*, 171(6):674–677, 2010.

[85] Y Xiao, M Abrahamowicz, and E E M Moodie. Accuracy of conventional and marginal structural cox model estimators: A simulation study. *The International Journal of Biostatistics*, 6(2):Article 13, 2010.

[86] J Xie and C Liu. Adjusted kaplan-meier estimator and log-rank test with inverse probability of treatment weighting for survival data. *Statistics in Medicine*, 24(20):30893110, 2005.

[87] A Zeileis. Econometric computing with hc and hac covariance matrix estimators. *Journal of Statistical Software*, 11(10):1–17, 2004.

[88] A Zeileis. Object-oriented computation of sandwich estimators. *Journal of Statistical Software*, 16(9):1–16, 2006.