University of California, Berkeley

U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2003 *Paper* 132

Double Robust Estimation in Longitudinal Marginal Structural Models

Zhuo Yu[*] Mark J. van der Laan[†]

[*]Department of Statistics, University of California, Berkeley, zyu@stat.berkeley.edu

[†]Division of Biostatistics, School of Public Health, University of California, Berkeley

Double Robust Estimation in Longitudinal Marginal Structural Models

Zhuo Yu and Mark J. van der Laan

Abstract

Consider estimation of causal parameters in a marginal structural model for the discrete intensity of the treatment specific counting process (e.g. hazard of a treatment specific survival time) based on longitudinal observational data on treatment, covariates and survival. We assume the sequential randomization assumption (SRA) on the treatment assignment mechanism and the so called experimental treatment assignment assumption which is needed to identify the causal parameters from the observed data distribution. Under SRA, the likelihood of the observed data structure factorizes in the auxiliary treatment mechanism and the partial likelihood consisting of the product over time of conditional distributions of covariate and survival at a specific time, given the past. Due to the curse of dimensionality, without assuming lower dimensional models for either the partial likelihood or the treatment mechanism, there exist no regular asymptotically linear estimators with reasonable practical performance (van der Laan and Robins [2002]). In this article, we define three estimators the Inverse Probability of Treatment weighted (IPTW) estimator based on a maximum likelihood estimator of the treatment mechanism according to a model, the maximum likelihood estimator (MLE) based on a maximum likelihood estimator of the partial likleihood according to a model, and a double robust (DL) estimator based on the maximum likelihood estimator of the treatment mechanism and the maximum likelihood estimator of the partial likelihood. The double robust estimator is obtained by following a general methodology for constructing double robust estimating functions in censored data models as described in van der Laan and Robins [2002]. We propose specific implementation of this estimator based on Monte-Carlo simulation methods, which makes the estimator computationally tractable and maximally robust. The double-robust estimator is consistent and asymptotically linear when either the treatment mechanism or the partial likelihood of the observed data is consistently estimated. We extend the estimator to handle informative censoring. We illustrate the practical performance of the DR estimator relative to the IPTW and ML estimators in a simulation study. The proposed methodology is also applied to estimate the causal effect of exercise on physical functioning in a longitudinal study of seniors in Sonoma County.

1 Introduction

It is important to point out why conventional approaches for estimating causal effects of time-dependent treatment will not produce valid results for longitudinal data structures. The most common method for handling confounders is to adjust for them or, in other words, to include all confounders in the regression model. In a point-treatment study, the resulting regression coefficient for the treatment does indeed have a causal interpretation. However, in a longitudinal study where the treatment changes over time – possibly in response to observed confounders, which are also affected by past treatment – such a regression has no causal interpretation, even if the assumption of no unmeasured confounders holds and the regression model is correctly specified. Another reason to look beyond the usual approach is the need to describe treatment effects for large diverse populations, for example, in policy-making. We may be truly interested the **marginal** effect of treatment on a population, as opposed to the treatment effect conditional on the values of certain covariates.

Marginal Structural models (MSMs) model causal effect of a time-dependent treatment on the distribution of treatment specific counterfactual outcomes of interest. Under the sequential randomization and the experimental treatment assignment assumption, these marginal structural models are identified from the observed data distribution. Robins et al. [2000] proposes inverse-probability-of-treatment weighted (IPTW) estimators of the unknown parameters of the MSM. The IPTW estimator is consistent if the treatment mechanism is correctly specified. In subsequent work Robins [2000b], proposes a double robust estimation methodology to estimate causal parameters in marginal structural models, whose consistency relies on either consistent estimation of the treatment mechanism or on consistent estimation of regressions. The assumed working models for the regressions cannot be expected to be compatible with a data generating distribution, and therefore Robins [2000b] refers to the corresponding estimators as generalized double robust estimators. In this article we develop a double robust (DR) estimator whose consistency relies on consistent estimation of either the treatment mechanism or the partial likelihood.

If both treatment mechanism and partial likelihood are correctly specified, the DR estimator is more efficient than the IPTW estimator. In practice, even when the treatment mechanism is consistently estimated, if one uses a small working model for the partial likelihood, then the double robust estimator is typically more efficient, except at extreme misspecification of the working model. We carry out a simulation study to demonstrate the performance of the DR estimator relative to the inverse-probability-of-treatment (IPTW) and the maximum likelihood estimator (MLE) maximizing the partial likelihood. The method is also generalized to the case where there is informative censoring. We use the methodology to estimate the effect of an activity score on the physical functioning for women enrolled in an observational study entitled "Study of Physical Performance and Age Related Changes in Sonomans" (SPARCS) Tager et al. [2000a].

1.1 The observed data model

The theory of counterfactual causal inference in longitudinal studies is first laid out in Robins [1986, 1987b, 1989, 1997]. Let time t take values in $\tau = [0, T]$, where T is a finite fixed number. For a time-dependent process $t \rightarrow Z(t)$, we denote its sample path up to time t^* with $\bar{Z}(t^*) = \{Z(t) : t \leq t^*\}$ and its complete sample path by $\bar{Z} = \bar{Z}(T) = \{Z(t) : t \in \tau\}$. In this paper, we divide $\tau = [0, T]$ into $K + 1$ intervals of equal length. Therefore, time is discrete and takes values in $\tau = \{0 = t_0, t_1, \ldots, t_K, t_{K+1} = T\}$. For any time-dependent process, we will at times use the abbreviated notation Z_k and $\bar{Z}_k$ in place of $Z(t_k)$ and $\bar{Z}(t_k)$.

Let $t \rightarrow A(t)$ be a time-dependent treatment process on $[0, t_K]$ and let $\mathcal{A}$ be the set of possible sample paths of $\bar{A}$, where we assume that $\mathcal{A}$ is finite. Let $X_{\bar{a}}(\cdot) \equiv (Y_{\bar{a}}(\cdot), L_{\bar{a}}(\cdot))$ be the counterfactual outcome and covariate process over time $t \in [0, T]$ under treatment regime $\bar{a}$. Let $S_{\bar{a}}$ be a treatment specific survival time which is part of $\bar{X}_{\bar{a}}$. For example, if survival is the outcome of interest, then one will have $Y_{\bar{a}}(t) = I(S_{\bar{a}} \leq t)$. We truncate all counterfactual processes at $S_{\bar{a}}$, that is, $X_{\bar{a}}(t) = X_{\bar{a}}(\min(t, S_{\bar{a}}))$. In time sequence the complete history of the subject counterfactual results is given by

$$L_{\bar{a}}(t_0), Y_{\bar{a}}(t_0), a_0 \ldots, L_{\bar{a}}(t_K), Y_{\bar{a}}(t_K), a_K, L_{\bar{a}}(t_{K+1}), Y_{\bar{a}}(t_{K+1}).$$

We will denote the baseline (pretreatment) covariates with $W = L_{\bar{a}}(t_0) = L(t_0)$ which are not affected by the treatment regime. For each possible treatment regime $\bar{a}$, $\bar{X}_{\bar{a}}(t_k)$ represents the data one would observe on the subject up to time t_k, if the subject were to follow treatment regime $\bar{a}$. The complete sample path $\bar{X}_{\bar{a}}$ is a counterfactual and is comprised of the paths of the outcome process $\bar{Y}_{\bar{a}}$, the covariate process $\bar{L}_{\bar{a}}$. We assume $\bar{X}_{\bar{a}(t_k)} = \bar{X}_{\bar{a}(t_{k-1})}(t_k)$. The full data for a subject is the collection of counterfactuals $X_{\bar{a}}$ generated by allowing treatment to range over the entire space $\mathcal{A}$, that is:

$$X = (\bar{Y}_{\bar{a}}, \bar{L}_{\bar{a}} : \bar{a} \in \mathcal{A})$$

Let $\bar{A}$, given X, follow a conditional density which is such that

$$g(A(t_k) \mid \bar{A}(t_{k-1}), X) = I(S_{\bar{A}} \leq t_k) I(A(t_k) = A(t_{k-1})) + I(S_{\bar{A}} > t_k) g(A(t_k) \mid \bar{A}(t_{k-1}), X).$$

In words, we truncate the treatment process at death.

The observed data is a missing data structure given by

$$O = (\bar{A}, \bar{X}_{\bar{A}}) = (\bar{A}, \bar{Y}_{\bar{A}}, \bar{L}_{\bar{A}}).$$

We see that the observed data is the observed treatment regime $\bar{A}$ and the corresponding single counterfactual $X_{\bar{A}}$. If $X \sim F_X$ and $A \mid X$ has density $g(\cdot \mid X)$, then we denote the corresponding distribution of O with $P_{F_X, g}$.

To acknowledge that the data collection stops deterministically after $S = S_{\bar{A}}$, one can also represent O as

$$O = (\bar{A}(S), \bar{Y}(S), \bar{L}(S).$$

If S exceeds T, then we note that each of these processes is truncated at T in the sense that (e.g.) $L(t) = L(\min(t, T))$.

Let $V \subset W$ be a selected subset of the baseline covariates for which one wishes to adjust the treatment specific intensity of $Y_{\bar{a}}$. We assume a marginal structural intensity model for the process $Y_{\bar{a}}(\cdot)$:

$$E(dY_{\bar{a}}(t_k)|\bar{Y}_{\bar{a}}(t_{k-1}), V) = \lambda(t_k, \bar{a}_{k-1}, \bar{Y}_{\bar{a}}(t_{k-1}), V|\boldsymbol{\alpha}), \tag{1}$$

where $\lambda(t_k, \bar{a}_{k-1}, \bar{Y}_{\bar{a}}(t_{k-1}), V|\boldsymbol{\alpha})$ is a known function, up to a p-dimensional parameter $\boldsymbol{\alpha}$. Our goal is to estimate $\boldsymbol{\alpha}$ based on n i.i.d. copies $(O_1, O_2, \ldots, O_n)$ of O.

As an example consider a study designed to determine the causal effect of a time-dependent treatment on survival or, alternatively, the hazard of mortality. Let $Y_{\bar{a}}(t_k)$ be a counting process that indicates the occurence of death in the interval (t_{k-1}, t_k), where $Y_{\bar{a}}(t_0)$ is 0 for all $\bar{a} \in \mathcal{A}$ by definition. The survival time can be recovered from the path $\bar{Y}_{\bar{a}}$, up to the resolution permitted by discrete time. A subject's record continues until failure. In this case $\lambda(\cdot|\boldsymbol{\alpha})$ is the product of an indicator that the subject is at risk for the event of interest and a function π:

$$\lambda(t_k, \bar{a}_{k-1}, \bar{Y}_{\bar{a}}(t_{k-1}), V|\boldsymbol{\alpha}) = I(Y_{\bar{a}}(t_{k-1}) = 0) \times \pi(t_k, \bar{a}_{k-1}, V|\boldsymbol{\alpha}), \tag{2}$$

where $\pi(t_k, \bar{a}_{k-1}, V|\boldsymbol{\alpha})$ might be a logistic function $\text{logit}^{-1}(\alpha_0 + \alpha_1 t_k + \alpha_2 a_{k-1} + \alpha_3 V)$. The choice of V (e.g. $V =$) depends on what the parameter of interest is and is therefore determined by the scientific question. For example, in policy making one might elect not to adjust for baseline covariates.

We adopt the framework of Bryan, Yu, van der Laan (2002), by not treating all timepoints t_k as equal, but in terms of the actions taken and the information collected. Continuing the above example, the potential failure times t_k will be called monitoring times. The interval length $t_k - t_{k-1}$ will generally be quite small and corresponds to the resolution with which we record survival time, for example, up to the month of death. At a given subset of the monitoring times, which we call the measuring times, we observe the covariate process $L(t_k)$. These measuring times generally coincide with a regular assessment such as a medical check-up. Typically, the treatment A_k can change at these measuring times, but one can also imagine situations in which the treatment changes at even fewer time points. We call these treatment times, which are a subset of the measuring times, which are a subset of the monitoring times. Schematically, at a time t_k, here is what happens for a subject:

1. Determine the outcome Y_k, i.e. confirm that subject survived the interval (t_{k-1}, t_k).

2. If t_k is a measuring time, measure the covariate L_k, otherwise assign it the same value as $L(t_{k-1})$.

3. If t_k is a treatment time, assign the treatment A_k for the next time interval, otherwise assign it the same value as $A(t_{k-1})$.

In order to identify causal effects, we must assume that the probability of a particular treatment decision at a treatment time t_k only depends on the observed history $(\bar{A}_{k-1}, \bar{Y}_k, \bar{L}_k, W) = (\bar{A}_{k-1}, \bar{X}_{\bar{A}}(t_k))$ of the subject. This assumption is called the sequential randomization assumption (SRA). To formally define the SRA, we recall the full data for a subject: $X = (\bar{X}_{\bar{a}} : \bar{a} \in \mathcal{A})$. The treatment mechanism satisfies SRA if

$$\begin{aligned} g(\bar{A}|X) &= g(A_0|X) \prod_{k=1}^{K} g(A_k|\bar{A}_{k-1}, X) \\ &= g(A_0|\bar{X}_{\bar{A}}(t_0)) \prod_{k=1}^{K} g(A_k|\bar{A}_{k-1}, \bar{X}_{\bar{A}}(t_k)) \end{aligned} \tag{3}$$

In other words, conditional on the observed past, the treatment decision at time t_k is independent of the full set of counterfactual data X [Robins, 1997]. This assumption is also referred to as the assumption of no unmeasured confounders.

Under SRA, the observed data likelihood w.r.t. an appropriate dominating measure is given by

$$\begin{aligned} p(O) =& \prod_{j=1}^{K} p(Y_j|\bar{Y}_{j-1}, \bar{L}_j, \bar{A}_j) p(L_j|\bar{Y}_{j-1}, \bar{L}_{j-1}, \bar{A}_{j-1}) \times g(\bar{A}|X) \\ \equiv& \mathcal{Q}_X(O) g(\bar{A}|X), \end{aligned} \tag{4}$$

where p represents a conditional density. We denote the first part of $p(O)$ with $\mathcal{Q}_X$ which is the F_X part of the likelihood since

$$\mathcal{Q}_X(\bar{A}, X_{\bar{A}}) = p_{X_{\bar{a}}}(X_{\bar{A}}))_{\bar{a}=\bar{A}}$$

is the marginal density of the counterfactual distribution of $X_{\bar{a}}$ with $\bar{a} = \bar{A}$. This relation between the observed data likelihood Q_X and the counterfactual distributions is also referred to as the G-computation formula for the density of $\bar{X}_{\bar{a}}$ Robins [1987a]. More general, it is a consequence of the general factorization of the likelihood of a censored data structure under coarsening at random (as implied by SRA), as proved by Heitjan and Rubin [1991]), Jacobsen and Keiding [1995], and Gill et al. [1997] in increasing generality. The discrete version of the G-computational formula is first given by Robins [1987a]. Continuous versions of the G-computation formula are proved in Gill and Robins [2001] and Yu and van der Laan [2002].

Consistent estimation of the causal parameter α requires consistent estimators of either the partial likelihood $\mathcal{Q}_X$ or the treatment mechanism g (van der Laan, Robins, 2002). Due to the curse of dimensionality, this means that we will also need to assume a lower dimensional model for either the partial likelihood $\mathcal{Q}_X$ or the treatment mechanism or both.

1.2 Overview.

In the next section we define three estimators: the Inverse Probability of Treatment weighted (IPTW) estimator based on a maximum likelihood estimator of the treatment mechanism according to a model, the maximum likelihood estimator (MLE) based on a maximum likelihood estimator of the partial likelihood according to a model, and a double robust (DL) estimator based on the maximum likelihood estimator of the treatment mechanism and the maximum likelihood estimator of the partial likelihood. The double robust estimator is obtained by following a general methodology for constructing double robust estimating functions in censored data models as described in van der Laan and Robins [2002]. We propose specific implementation of this estimator based on Monte-Carlo simulation methods, which make the estimator computationally tractable and maximally robust. The double-robust estimator is consistent and asymptotically linear when either the treatment mechanism or the partial likelihood of the observed data is consistently estimated. For both the IPTW and the DR estimators, we provide confidence intervals for $\boldsymbol{\alpha}$. In section 3 we extend the approach to handle right censored data.

Bryan, Yu and Laan [2002] also studied a one-step estimator and showed that it is superior to the IPTW estimator in terms of efficiency. The one-step estimator corresponds with the first step of the Newton-Raphson algorithm for solving the double robust estimating function when starting with the IPTW estimator. Consequently, though, the one-step estimator is very easy to compute, it is not doubly robust.

In order to compare the practical performance of the estimators under consideration, we present the results of a simulation study in section 4. Our results show that the DR estimator can be far more efficient than the IPTW estimator and that it is far more robust than the IPTW and maximum likelihood estimator. Finally, in section 5, we apply the extended methodology to estimate the causal effect of exercise on mortality in a longitudinal study of seniors in Sonoma County.

2 Estimation and Inference

2.1 The IPTW estimator

In this subsection we describe the first of three estimators of $\boldsymbol{\alpha}$: the IPTW estimator. An IPTW estimator is obtained as the solution of an estimating equation and the relevant estimating function results from a mapping of full data estimating functions into observed data estimating functions. The details of the IPTW mapping and a proof that the T_{SRA} orthogonalized IPTW mapping, as presented in the next subsection, produces the class of all observed data estimating functions is provided in van der Laan and Robins [2002].

Let

$$\epsilon_{\bar{a}}(t_k \mid \boldsymbol{\alpha}) \equiv Y_{\bar{a}}(t_k) - \lambda(t_k, \bar{a}_{k-1}, \bar{Y_{\bar{a}}}(t_{k-1}), V|\boldsymbol{\alpha})$$

and let $\epsilon_{\bar{A}}(\cdot \mid \boldsymbol{\alpha})$ be the observed vector of residuals. Let $h(\cdot)$ be any function of time, the selected baseline covariates V, and the observed history of the treatment and outcome processes; a typical choice of h is given by (8). For every h, we can define an

IPTW estimating function $IC_{iptw}(O|g,\boldsymbol{\alpha},h)$:

$$\begin{aligned} IC_{iptw}(O|g,\boldsymbol{\alpha},h) &= \sum_k sw(t_k)\times\ h(t_k,\bar{A}_{k-1},\bar{Y}_{k-1},V)\times\ \epsilon_{\bar{A}}(t_k\mid\boldsymbol{\alpha}) \\ &= \sum_k \left(\prod_{j=0}^{k} \frac{g(A_j|\bar{A}_{j-1},V)}{g(A_j|\bar{A}_{j-1},\bar{X}_{\bar{A}}(t_j))}\right)\times\ h(t_k,\bar{A}_{k-1},\bar{Y}_{k-1},V)\times\ \epsilon_{\bar{A}}(t_k\mid\boldsymbol{\alpha}). \end{aligned} \tag{5}$$

This has the familiar form of generalized estimating functions for the regression model corresponding with the marginal structural model, namely, a sum over time of time-specific products of a residual and a function of the covariates. However in this case, we additionally have stabilized time-specific weights $sw(t_k)$ that capture the probability of the observed treatment given the past. In practical terms, any IPTW-type estimator works by upweighting (downweighting) subjects that, given their observed past, have received an unusual (typical) treatment. This is achieved through the use of weights inversely proportional to the probability of the observed treatment, given the covariate. The stabilized weights $sw(t_k)$ in (5) [Robins, 1998] include a numerator term that, in the absence of time-dependent confounding, will equal the denominator and will produce an unweighted estimating function. In the presence of confounding the stabilized weights cause the estimating function to remain unbiased. Formally, if

$$g(a^*|\bar{A}_{j-1},\bar{X}(j))>0 \text{ for all } a^*\in\{a_j:\bar{a}_{j-1}=\bar{A}_{j-1},\bar{a}\in\mathcal{A}\}, \tag{6}$$

then $E_{P_{F_X},g}IC_{iptw}(O\mid g,\boldsymbol{\alpha}(F_X),h)=0$.

This latter (so called) experimental treatment assignment assumption is a condition on the support of $g(\cdot\mid X)$. It could be seriously violated and thereby make the IPTW estimating function heavily biased. In addition, even when this assumption holds theoretically, but the probabilities on certain a^* are almost equal to zero, then this so called practical violation of the ETA (that is, given the finite sample size, the support of $g(\cdot\mid X)$ is truly restricted) will cause serious finite sample bias of the corresponding IPTW estimator. We refer to Neugebauer and van der Laan [2002], and van der Laan and Robins [2002] for a detailed explanation and practical illustration. For example, suppose a_j is a dichotomous variable and $P(A_j=1|A_{j-1},\bar{X}(j-1))=expit(\theta_0+\theta_1A_{j-1}+\theta_2X(j-1))$, where $expit(x)\equiv exp(x)/(1+exp(x))$. If θ_2 is very large, then for certain values of $X(j-1)$, the probability that A_j is zero is so small that it does not happen for the given sample size.

Since the treatment mechanism g is typically unknown, it represents a nuisance parameter of the IPTW estimating function. We can estimate g with maximum likeihood estimation according to a lower dimensional parametric or semiparametric model for $g(A_j|\bar{A}_{j-1},\bar{X}_{\bar{A}}(t_j))$. Let g_n denote such a maximum likelihood estimator of the treatment mechanism. Let $\widehat{\boldsymbol{\alpha}}_n^{iptw}$ (the IPTW estimator) be defined as the solution of the following estimating equation

$$\sum_{i=1}^{n} IC_{iptw}(O_i|g_n,\widehat{\boldsymbol{\alpha}}_n^{iptw},h)=0. \tag{7}$$

Regarding the choice of h, we will make the usual choice for h, namely, the optimal choice for the regression model corresponding with the MSM. [Robins, 2000a, Hernan et al., 2000]:

$$h(t_k, \bar{A}_{k-1}, \bar{Y}_{k-1}, V) = \frac{d}{d\boldsymbol{\alpha}} \lambda(t_k, \bar{A}_{k-1}, \bar{Y}_{\bar{a}}(t_{k-1}), V|\boldsymbol{\alpha}). \tag{8}$$

Standard software can be employed to solve the weighted estimating equation implied by (8). Practical details for implementing this particular IPTW estimator are provided in Bryan et al. [2002].

2.2 Maximum Likelihood Estimator

In this section we provide a second estimator of $\boldsymbol{\alpha}$, a Maximum Likelihood (ML) estimator assuming a model for the partial likelihood $\mathcal{Q}_X(\bar{A}, X_{\bar{A}})$. Let $P_{L_{k+1}|\bar{A}_k,\bar{Y}_k,\bar{L}_k}(dl_{k+1}; \bar{a}_k, \bar{y}_k, \bar{l}_k)$ and $P_{Y_{k+1}|\bar{A}_{k+1},\bar{Y}_k,\bar{L}_{k+1}}(dy_{k+1}; \bar{a}_k, \bar{y}_k, \bar{l}_{k+1})$ be the regular conditional distributions for L_{k+1} given $(\bar{A}_k, \bar{Y}_k, \bar{L}_k)$ and Y_{k+1} given $(\bar{A}_{k+1}, \bar{Y}_k, \bar{L}_{k+1})$ respectively. If the experimental treatment assumption (6) holds, then

$$\begin{aligned} P(Y_{\bar{a}}(t_j) \in dy_j) = \int_{l_1}\int_{y_1} \cdots \int_{y_{j-1}} \int_{l_j} & \prod_{k=0}^{j-1} P_{Y_{k+1}|\bar{A}_{k+1},\bar{Y}_k,\bar{L}_{k+1}}(dy_{k+1}; \bar{a}_{k+1}, \bar{y}_k, \bar{l}_{k+1}) \\ & \times \prod_{k=0}^{j-1} P_{L_{k+1}|\bar{A}_k,\bar{Y}_k,\bar{L}_k}(dl_{k+1}; \bar{a}_k, \bar{y}_k, \bar{l}_k). \end{aligned} \tag{9}$$

Given the partial likelihood $\mathcal{Q}_X$, Robins (1987a) proposes to evaluate this distribution of the counterfactual process $Y_{\bar{a}}(\cdot)$ by drawing a large sample with the following Monte-Carlo simulation algorithm. An asterix is used to denote simulated variables. Given a treatment regime $\bar{a} \in \mathcal{A}$, one first generates $L_1^* = l_1^*$ from the marginal distribution of L_1. Subsequently, one generates $Y_1^* = y_1^*$ from the conditional distribution of Y_1 given $L_1 = l_1^*, A_1 = a_1$. Then one generates l_2^* from the conditional distribution of L_2 given $L_1 = l_1^*, A_1 = a_1, Y_1 = y_1^*$; and so on. This provides us with a draw $(y_1^*, \ldots, y_K^*)$ from the distribution of $\bar{Y}_{\bar{a}}$ described by (9). By drawing a large number of realizations, one obtains an arbitrarily good approximation of the distribution of $\bar{Y}_{\bar{a}}$.

Since $\mathcal{Q}_X$ is unknown, we will estimate it with maximum likelihood estimation according to a model. The conditional density of Y_j given the past can be estimated by assuming a parametric model $p_\theta(Y_j|\bar{Y}_{j-1}, \bar{L}_j, \bar{A}_j)$ and then estimating θ with the MLE. Similary we can estimate the conditional density for L_j given the past. Substituting this fit of the partial likelihood $\mathcal{Q}_X$ into the formula (9) provides us now with an estimate of the counterfactual distribution of $Y_{\bar{a}}$ for each $\bar{a} \in \mathcal{A}$.To determine the corresponding estimate of the causal parameter $\boldsymbol{\alpha}$, we now apply the above Monte-Carlo algorithm to generating a large sample of $\hat{Y}_{\bar{a}}(t_K)$ for a rich collection of treatment regimes $\bar{a}$'s. Finally, we fit our marginal structural model (1) to this large sample on $(Y_{\bar{a}}, \bar{a})$ with standard software. For example, if our MSM is the logistic regression model, then by regressing simulated $y_{\bar{a}}^*(t_j)$ on t_j and a_j using logistic regression, treating the pooled sample as an i.i.d. sample, we obtain an estimate of $\boldsymbol{\alpha}$. We note that the model we assume for $\mathcal{Q}_X$ may not be compatible with our MSM. In that case, the above method determines the fit of the MSM most compatible with $\mathcal{Q}_X$.

2.3 Double Robust Estimator

In this section we provide the third estimator of $\boldsymbol{\alpha}$, namely, the Double Robust (DR) estimator $\widehat{\boldsymbol{\alpha}}_n^{dr}$. The general proposal for Double Robust estimation in censored data models such as our causal inference model is provided in Secion 1.6 of van der Laan and Robins [2002]. The double robust estimating functions are defined by subtracting from the IPTW estimating functions the projection onto a nuisance tangent space $T_{SRA}$of the treatment mechanism in the Hilbert space $L_0^2(P_{F_X,g})$. Formally, T_{SRA} is defined as the Hilbert space of all the nuisance scores corresponding with one-dimensional submodels through the true treatment mechanism with the only restriction being that these submodels need to satisfy the SRA. Though not important for following this section, we also remind the reader that $L_0^2(P_{F_X,g})$ is the Hilbert space consisting of all functions of the observed data structure $O = (\bar{A}, X_{\bar{A}}) \sim P_{F_X,g}$ with mean zero and finite variance endowed with inner product $\langle h_1, h_2 \rangle_{P_{F_X,g}} = E_{F_X,g} h_1(O) h_2(O)$.

It will be convenient in this and later sections to define $\mathcal{F}_k \equiv (\bar{A}_{k-1}, \bar{Y}_k, \bar{L}_k)$; in words, $\mathcal{F}_k$ is the observed past just prior to the treatment assignment A_k. The tangent space for the treatment mechanism nuisance parameter at time t_k, denoted $T_{SRA,k}$, is the space of scores obtained by varying $g(A_k|\mathcal{F}_k)$. This is the space of all functions of A_k and $\mathcal{F}_k$ that have conditional mean zero, given the observed past $\mathcal{F}_k$. That is, for d_k ranging over all functions of A_k and $\mathcal{F}_k$ we have that

$$T_{SRA,k} = \{ d_k(A_k, \mathcal{F}_k) - E_g(d_k(A_k, \mathcal{F}_k)|\mathcal{F}_k) : d_k \}, \quad k = 0, \ldots, K \tag{10}$$

The factorization of $g(\bar{A}|X)$ into time-specific terms implies that

$$\begin{aligned} T_{SRA} &= T_{SRA,0} \oplus T_{SRA,1} \oplus \ldots \oplus T_{SRA,K} \\ &= \left\{ \sum_{k=0}^{K} d_k(A_k, \mathcal{F}_k) - E(d_k(A_k, \mathcal{F}_k)|\mathcal{F}_k)(d_1, \ldots, d_K) \right\}. \end{aligned} \tag{11}$$

That is T_{SRA} is the orthogonal sum of $T_{SRA,k}, k = 0, \ldots, K$

We obtain a DR estimating function from the IPTW estimating function $IC_{iptw}(O|g, \boldsymbol{\alpha}, h)$ by subtracting its projection onto T_{SRA}. The projection of $IC_{iptw}(O|g, \boldsymbol{\alpha}, h)$ onto T_{SRA} is given by

$$\begin{aligned} IC_{SRA}(O|\mathcal{Q}_X, g) = \sum_{k=0}^{K} & E_{\mathcal{Q}_X,g}(IC_{iptw}(O|g, \boldsymbol{\alpha}(\mathcal{Q}_X), h)|A_k, \mathcal{F}_k) \\ & - E_{\mathcal{Q}_X,g}(IC_{iptw}(O|g, \boldsymbol{\alpha}(\mathcal{Q}_X), h)|\mathcal{F}_k), \end{aligned} \tag{12}$$

We note that $IC_{SRA}(O|\mathcal{Q}_X, g)$ is only a function of $\mathcal{Q}_X$ and g since the observed data likelihood factorizes into $\mathcal{Q}_X$ and g, and consequently, the conditional expectations are completely determined by $\mathcal{Q}_X$ and g. We also note that we consider here $\boldsymbol{\alpha} = \boldsymbol{\alpha}(\mathcal{Q}_X)$ as a function of $\mathcal{Q}_X$. The DR estimating function is given by

$$IC_{dr}(O|\mathcal{Q}_X, g, h, \boldsymbol{\alpha}) = IC_{iptw}(O|g, \boldsymbol{\alpha}, h) - IC_{SRA}(O|\mathcal{Q}_X, g). \tag{13}$$

It is shown in Section 1.6 of van der Laan and Robins [2002] that the following double robustness result holds. This result can also be verified directly as in Yu [2002].

Lemma 2.1. *We have $E_{\mathcal{Q}_X,g} IC_{dr}(O|\mathcal{Q}^1, g_1, h, \boldsymbol{\alpha}) = 0$ if either ($g_1 = g$ and (6) holds at g) or ($\mathcal{Q}^1 = \mathcal{Q}_X$ and (6) holds at g_1).*

We note that the estimating function is a function of the observed data O, the causal parameter $\boldsymbol{\alpha}$, the choice of h and the nuisance parameters g and $\mathcal{Q}_X$. As described above, h can be chosen general, but we will work with the typical choice as provided in (8). Maximum likelihood estimation of g and $\mathcal{Q}_X$ according to models has been discussed in previous sections. Given maximum likelihood estimates $\mathcal{Q}_n$, h_n, and g_n of $\mathcal{Q}_X$, h, and g, respectively, we define the double robust estimator $\widehat{\boldsymbol{\alpha}}_n^{dr}$ as the solution of

$$0 = \frac{1}{n}\sum_{i=1}^{n} IC_{dr}(O_i|\mathcal{Q}_n, g_n, h_n, \boldsymbol{\alpha}) \tag{14}$$

By the double robust property Lemma (2.1), the DR estimator will remain consistent if either the F_X part $\mathcal{Q}_X$ of the likelihood $p(O)$ or g is correctly specified and (6) holds at the estimated g_n.

Implementation of the double robust estimator.

We solve this equation in α with the Newton-Raphson algorithm. In the NR algorithm the derivative is given by $c_n = \frac{d}{d\alpha}\frac{1}{n}\sum_{i=1}^{n} IC_{iptw}(O_i|g_n, \boldsymbol{\alpha}, h)$ since IC_{SRA} does not depend on $\boldsymbol{\alpha}$. Notice that evaluation of $IC_{dr}(O_i|\mathcal{Q}_n, g_n, h_n, \boldsymbol{\alpha})$ requires evaluating $E_{\mathcal{Q}_n,g_n}(IC_{iptw}(O|g_n, \boldsymbol{\alpha}, h)|A_k, \mathcal{F}_k)$ and $E_{\mathcal{Q}_n,g_n}(IC_{iptw}(O|g_n, \boldsymbol{\alpha}, h)|\mathcal{F}_k)$. We propose to evaluate these conditional expectations with the following Monte-Carlo simulation algorithm. Given a subject's observed history $\mathcal{F}_k = (\bar{a}_{k-1}, \bar{y}_k, \bar{l}_k)$, we fix $\bar{a}^*_{k-1} = \bar{a}_{k-1}, \bar{y}^*_k = \bar{y}_k, \bar{l}^*_k = \bar{l}_k$ and now generate the future by sequentially generating from the factors in the observed data likelihood. Specifically, we generate the future observation $(A(k), Y(k+1), L(k+1))$ by 1) generating $A^*_k = a^*_k$ from the conditional distribution of A_k given $\bar{A}_{k-1} = \bar{a}^*_{k-1}, \bar{Y}_k = \bar{y}^*_k \bar{L}_k = \bar{l}^*_k$, 2) generating $Y^*_{k+1} = y^*_{k+1}$ from the conditional distribution of Y_{k+1} given $\bar{A}_k = \bar{a}^*_k, \bar{Y}_k = \bar{y}^*_k \bar{L}_k = \bar{l}^*_k$, and 3) generating L^*_{k+1} from the conditional distribution of L_{k+1}, given $\bar{A}_k = \bar{a}^*_k, \bar{L}^*_k, \bar{Y}^*_{k+1}$. Now, set $k = k+1$ and repeat till the complete future is observed and thereby the observed data structure O^*. We can now evaluate $IC_{iptw}(O^*|g_n, \widehat{\boldsymbol{\alpha}}_n^{ml}, h_n)$ using this simulated O^*, estimated g_n and MLE $\widehat{\boldsymbol{\alpha}}_n^{ml}$. We repeat this N times and we use the empirical mean of $IC_{iptw}(O^*_b|g_n, \widehat{\boldsymbol{\alpha}}_n^{ml}, h_n), b = 1, \ldots, N$ as our evaluation of $E_{\mathcal{Q}_n,g_n}(IC_{iptw}(O|g_n, \boldsymbol{\alpha}, h)|\mathcal{F}_k)$. The conditional mean $E_{\mathcal{Q}_n,g_n}(IC_{iptw}(O|g_n, \boldsymbol{\alpha}, h)|A_k, \mathcal{F}_k)$ can be evaluated similarly, except that, we start from at the observed past $(A_k, \mathcal{F}_k)$. We conclude that, given a fit of the observed data likelihood $\mathcal{Q}_{Xn}(\bar{A}, X_{\bar{A}})g_n(\bar{A} \mid X)$, the implementation of our double robust estimator is straightforward.

A method for making the double robust estimator more robust.

As we see, in order to calculate the DR estimator, we need, in particular, to model the distribution of $L(t_k)$ given the past and estimate it with MLE. Since L_k is typically a high dimensional vector in practice, it would be tedious and computationally intensive to form and fit such a high dimensional likelihood. To simplify the estimation of the time-specific factors $p(L(t_k) \mid \bar{L}(t_{k-1}), \bar{Y}(t_{k-1}, \bar{A}(t_k))$ in the partial likelihood $\mathcal{Q}_X$ at each time point t_k, we propose to reduce the multivariate time-dependent covariate L_k to a univariate time-dependent covariate $L_{trt,k}$ extracted from the fitted treatment mechanism. For example, suppose that the treament value is dichotomous and satisfies the following logistic regression model:

$$P(A^t(t_k) = 1|\bar{A}(t_{k-1}), \bar{Y}(t_k), \bar{L}(t_k)) = I(Y(t_k) = 0)\lambda_{t_k}(\boldsymbol{\gamma}), \tag{15}$$

where $\lambda_{t_k}(\boldsymbol{\gamma}) = \text{logit}^{-1}\left(\gamma_{0,k} + \gamma_{1,k}A^t_{k-1} + \gamma^T_{2,k}L_k + \gamma^T_{3,k}W\right)$. Then we set $L_{trt,k} \equiv \hat{\gamma}_{0,k} + \hat{\gamma}_{1,k}A_{k-1} + \hat{\gamma}^T_{2,k}L_k + \hat{\gamma}^T_{3,k}W$, where $\hat{\gamma}$ is the maximum likelihood estimator of γ. Thus $L_{trt,k}$ includes all the information about the observed covariate process which was considered predictive of treatment assignment. Consider the reduced data structure $O' \equiv (\bar{A}_K, \bar{L}_{trt,K+1}, \bar{Y}_{K+1})$. This reduction of the data does not come at cost of a (increased) violation of SRA since the fit of the treatment model is not affected by this reduction: we have

$$P(A^t(t_k) = 1|\bar{A}^t(t_{k-1}), \bar{Y}(t_k), \bar{L}(t_k)) = \text{logit}^{-1}(L_{trt,k}), \tag{16}$$

We proceed to calculate the DR estimator using the Monte-Carlo simulation described above for this reduced data structure. Because of the data reduction, calculation of the maximum likelihood estimator of $\mathcal{Q}_X$ is now strongly simplified. We note that this data reduction does not sacrifice consistency but does reduce the efficiency. However, possibly much more important in practice, the reduction makes the nuisance parameter $\mathcal{Q}_X$ much easier to estimate and thereby potentially strongly improves the robustness of the double robus estimator of $\boldsymbol{\alpha}$ relative to the double robust estimator based on the complete observed data structure. Having said this, to improve efficiency at cost of increased computational burden and decreased robustness, one could also include another univariate covariate extracted from a fitted model of Y_j given the past.

2.4 Confidence Intervals

Firstly, we discuss inference when one assumes a correctly specified model for $g(\bar{A}|X)$ so that $\mathcal{Q}_n \to \mathcal{Q}^1_X$ and $g_n \to g$, where $\mathcal{Q}^1_X \neq \mathcal{Q}_X$ is allowed. Under regularity conditions, the DR estimator $\widehat{\boldsymbol{\alpha}}^{dr}_n$, defined in equation (14), is a regular asymptotically linear estimator with influence curve

$$IC(\cdot) = -c^{-1}\left[IC_{dr}(\cdot|\mathcal{Q}^1, g, h, \boldsymbol{\alpha}) - \Pi(IC_{dr}(\cdot|\mathcal{Q}^1, g, h, \boldsymbol{\alpha})|T_g)\right] \tag{17}$$

where $T_g \subset T_{SRA}$ is the tangent space of g under the assumed model for the treatment mechanism g. For more details we refer to van der Laan and Robins [2002]. We note that, if $\mathcal{Q}_X^1 = \mathcal{Q}_X$, then the projection term in (17) equals zero.

Since the DR estimator $\widehat{\boldsymbol{\alpha}}_n^{dr}$ is asymptotically linear with influence curve (17), the asymptotic covariance matrix of $\widehat{\boldsymbol{\alpha}}_n^{dr}$ is given by

$$\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \widehat{IC}(O_i)^{\otimes 2}. \tag{18}$$

One can avoid the calculation of the projection operator on T_g in (17), by estimating the asymptotic covariance matrix of $\widehat{\boldsymbol{\alpha}}_n^{dr}$ *conservatively* with

$$\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \widehat{IC}_{dr}(O_i | \mathcal{Q}_n, g_n, h_n, \boldsymbol{\alpha})^{\otimes 2}. \tag{19}$$

If $\mathcal{Q}_X^1 = \mathcal{Q}_X$, then the covariance estimates are asymptotically identical.

The above variance estimates can be used to construct a 95% confidence interval for the j-th component of $\boldsymbol{\alpha}$ given by

$$\widehat{\boldsymbol{\alpha}}_n^{dr} \pm 1.96 \frac{\widehat{\Sigma}_{jj}}{\sqrt{n}}. \tag{20}$$

If we only assume that either $\mathcal{Q}_n \rightarrow \mathcal{Q}_X$ or $g_n \rightarrow g$, then we recommend using the Bootstrap to compute the C.I.

3 Causal Inference for Censored Data

Suppose that the observed data is subject to right censoring and D is the censoring time. Let $A^t(\cdot)$ denote the usual treatment process; the superscript t has been added to emphasize the treatment action. We define a censoring process $A^c(t_k) = I(D \leq t_k)$ and the process $A = (A^t, A^c)$ will now be defined more generally and refer to both treatment and censoring. The observed data structure can now be represented as

$$O = (R = S \wedge D, \bar{A}(R) = (\bar{A}^t(R), \bar{A}^c(R)), \bar{X}_{\bar{A}^t}(R)), \tag{21}$$

where $\bar{X}_{\bar{A}^t}(\cdot) \equiv (\bar{L}_{\bar{A}^t}(\cdot), \bar{Y}_{\bar{A}^t}(\cdot))$ is the observed history up to R.

We will represent this data structure as a missing data structure on counterfactual indexed by the joint-action regime $\bar{a}$. Here it is assumed that the uncensored counterfactual response and covariate processes are not affected by the actual realization of the censoring process: that is, in terms of counterfactuals $X_{\bar{a}}(t) = X_{\bar{a}^t}(\min(t, d))$, where d is the censoring time corresponding with $\bar{a}^c$. Let $\bar{A}$, given X, follow a conditional density which is such that

$$\begin{aligned} g(A(t_k) \mid \bar{A}(t_{k-1}), X) \;&=\; I(S_{\bar{A}} \leq t_k, D_{\bar{A}} \leq t_{k-1}) I(A(t_k) = A(t_{k-1}) \\ &\quad + (1 - I(S_{\bar{A}} \leq t_k, D_{\bar{A}} \leq t_{k-1})) g(A(t_k) \mid \bar{A}(t_{k-1}), X). \end{aligned}$$

In words, we truncate the action process at death. Thus $A^t(t) = A^t(\min(t, R))$, $A^c(t) = A^c(\min(t, R))$ so that the process $A = (A^t, A^c)$ is observable. In this manner, we can represent the observed data structure (21) as a general censored data structure $(\bar{A}, X_{\bar{A}})$, where now the joint regime $A = (A^c, A^t)$ represents the censoring variable.

The sequential randomization assumption is now given by

$$g(A_k|\bar{A}_{k-1}, X) = g(A_k|\bar{X}_{\bar{A}^t}(t_k), \bar{A}_{k-1}). \tag{22}$$

We have

$$\begin{aligned} g(\bar{A}|X) &= \Pi_k g\left(A_k|\bar{X}_{\bar{A}^t}(t_k), \bar{A}_{k-1}\right) = \Pi_k g\left(A^c_k, A^t_k|\bar{X}_{\bar{A}^t}(t_k), \bar{A}_{k-1}\right) \\ &= \Pi_k g\left(A^c_k|\bar{X}_{\bar{A}^t}(t_k), \bar{A}_{k-1}, A^t_k\right) \times \Pi_k g\left(A^t_k|\bar{X}_{\bar{A}^t}(t_k), \bar{A}_{k-1})\right). \end{aligned} \tag{23}$$

We construct an Inverse Probability of Action (IPAW) estimator in the same manner as in section 2.1. Here we used the term “Action” to refer to both censoring and treatment assignments. We simply must extend the stabilized weights to include information on the censoring process as well as the treatment process. The IPAW-estimating function is given by

$$IC_{ipaw}(O|g, \boldsymbol{\alpha}, h) = \sum_k sw(t_k) \times \ I(A^c_k = 0) \times \ h(t_k, \bar{A}^t_{k-1}, \bar{Y}_k, V) \times \ \epsilon_{\bar{A}^t}(t_k \mid \boldsymbol{\alpha}), \tag{24}$$

where the stabilized weight is given by

$$\begin{aligned} sw(t_k) &= \frac{g\left(\bar{A}^t_k, \bar{A}^c_k = 0|V\right)}{g\left(\bar{A}^t_k, \bar{A}^c_k = 0|X\right)} \\ &= \frac{\prod_{j=0}^k g\left(A^c_j = 0|\bar{A}^t_j, \bar{A}^c_{j-1} = 0\right)}{\prod_{j=0}^k g\left(A^c_j = 0|\bar{X}_{\bar{A}^t}(t_j), \bar{A}^t_j, \bar{A}^c_{j-1} = 0\right)} \\ &\quad \times \frac{\prod_{j=0}^k g\left(A^t_j|\bar{A}^t_{j-1}, \bar{A}^c_{j-1} = 0, V\right)}{\prod_{j=0}^k g\left(A^t_j|\bar{X}_{\bar{A}^t}(t_j), \bar{A}^t_{j-1}, \bar{A}^c_{j-1} = 0\right)}. \end{aligned} \tag{25}$$

Once again, we can use standard models, such as logistic regression, to fit the censoring mechanism $g\left(A^c_j = 0|\bar{X}_{\bar{A}^t}(t_j), \bar{A}^t_j, \bar{A}^c_{j-1} = 0\right)$. Given a choice of h, such as the usual choice, we can then solve the corresponding estimating equation for $\boldsymbol{\alpha}$ as in section 2. For example, we can use the S-plus function `glm()` with weights $I(A^c_k = 0)\ sw(t_k)$ at time t_k. For details on the implementation we refer to Bryan, Yu, van der Laan (2002).

The action-mechanism orthogonalized estimating function is constructed by subtracting from IC_{ipaw} the projection on the tangent space T_{SRA} of the nonparametric model for $g(\bar{A} \mid X)$ defined by (23). Let $\mathcal{F}_{k,c} \equiv (\bar{X}_{\bar{A}^t}(t_k), \bar{A}_{k-1}, A^t_k)$ and $\mathcal{F}_{k,t} \equiv (\bar{X}_{\bar{A}^t}(t_k), \bar{A}_{k-1})$. We have

$$\begin{aligned} IC_{SRA} &\equiv \Pi\left(IC_{ipaw}|T_{SRA}\right) \\ &= \sum_k \left(E\left(IC_{ipaw}|\mathcal{F}_{k,c}, A^c_k\right) - E\left(IC_{ipaw}|\mathcal{F}_{k,c}\right)\right) \\ &\quad + \sum_k \left(E\left(IC_{ipaw}|\mathcal{F}_{k,t}, A^t(t_k)\right) - E\left(IC_{ipaw}|\mathcal{F}_{k,t}\right)\right) \end{aligned}$$

In the same manner as we did for the treatment orthogonalized estimating function, we can consider $IC_{SRA}(O \mid g, Q)$ as a function of the data O, the action mechanism g and the partial likelihood $\mathcal{Q}_X$. The action orthogonalized estimating function is now defined by $IC_{ao}(O \mid g, \alpha, h, Q) = IC_{ipaw}(O \mid g, \alpha, h) - IC_{SRA}(O \mid g, Q)$. Given maximum likelihood estimators of g and $\mathcal{Q}_X$, all the condition expectations can be evaluated with the Monte-Carlo simulation described in the previous section. Again, to simplify the DR estimation we can reduce L to two time-dependent covariates extracted from the fitted treatment and fitted censoring mechanisms, as we described in section 2.3. Details are omitted and presented in our data analysis.

4 A Simulation Study

In this section, we carry out a simulation study to compare various estimators of the causal parameter $\boldsymbol{\alpha}$. We consider the IPTW estimator $\widehat{\boldsymbol{\alpha}}_n^{iptw}$, the MLE estimator $\widehat{\boldsymbol{\alpha}}_n^{ml}$ and the DR estimator $\widehat{\boldsymbol{\alpha}}_n^{dr}$. We compare the estimates by the mean squared error (MSE), bias and variance across 50 samples of size 250. Corresponding the theory, the simulations show the following:

- In case that confounding is so severe that the identifiability assumption (6) is practically violated or the treatment mechanism is misspecified, the finite sample bias of the DR estimator is much smaller than that of the IPTW estimator.
- If g is correctly specified, the MLE is very sensitive to likelihood misspecification where as the finite sample bias of the DR estimator remains small.
- If g is correctly specified, the confounding is extreme (so that (6) is practically violated) *and* the likelihood is misspecified, then the finite sample bias of the DR estimator is much smaller than that of the MLE and slightly smaller than that of IPTW estimator.
- In the case that g is correctly specified, confounding is not extreme, *and* the likelihood is correctly specified , the MSE of the DR estimator is smaller than that of IPTW estimator.

Section 4.1 describes the causal marginal structural model and the treatment mechanism we use in the simulation study. Section 4.2 describes how we analyze the simulated data and thus includes concrete details on implementing all three estimators. Secton 4.3 compares the MLE based on different models for the likelihood $\mathcal{Q}_X$. Section 4.4 gives the data-generating parameter values and all of the simulation results.

4.1 Data Generating Model

We use the same simulation setting as in Bryan et al. [2002]. We continue to work in a discrete time setting, with a finite number of monitoring times. We also use the same survival counting process introduced earlier. At a given subset of the monitoring times, we measure a covariate process. The covariate process $L(t_k)$ is real-valued and always positive. We assume that the covariate grows linearly with time until initiation of treatment and remains unchanged thereafter. The generation of subject-specific slopes and intercepts is discussed below, as is the effect of treatment on the covariate. At these measuring times, we may make a change in the treatment and this decision may be a function of the covariate and treatment history of the subject. The treatment A_k takes on the values 0 and 1, corresponding to 'off' and 'on' treatment, respectively. Once a subject is on treatment, s/he will remain so until failure. And, of course, a subject will not be considered for treatment after failure. Given the above considerations, we are left with a relatively small set of times at which treatment can change. At a treatment time t_k, the probability of initiating treatment is given by the following logistic model:

$$\operatorname{logit} P(A_k = 1 \mid \bar{L}(t_k), \bar{A}_{k-1} = 0) = \theta_0 + \theta_1 t_k + \theta_2 L(t_k). \tag{26}$$

This is equivalent to the intensity model:

$$E(dA_k | \bar{L}(t_k), \bar{A}_{k-1}) = I(\bar{A}_{k-1} = 0) \times \operatorname{logit}^{-1}(\theta_0 + \theta_1 t_k + \theta_2 L(t_k)).$$

Therefore partial likelihood estimation can be used to estimate $\boldsymbol{\theta}$, as was mentioned in section 2. If a subject goes on treatment, we refer to the treatment initiation time as t^*.

We can now state the MSM. Given a subject that has not failed, the probability of failure in the upcoming interval is given by

$$\operatorname{logit}(P(Y_{\bar{a}}(t_k) = 1 \mid Y_{\bar{a}}(t_{k-1}) = 0) = \alpha_0 + d_1(t_k)\alpha_1 + a_k\alpha_2 + d_3(t_k)\alpha_3, \tag{27}$$

where $d_1(t_k) = (1 - a_k)t_k + a_k t^*$ and $d_3(t_k) = a_k(t_k - t^*)$. This corresponds to the intensity model:

$$\lambda(t_k, \bar{a}, \bar{Y}_{\bar{a}}(t_{k-1}) | \boldsymbol{\alpha}) = I\left(\bar{Y}_{\bar{a}}(t_{k-1}) = 0\right) \times \operatorname{logit}^{-1}\left(\alpha_0 + d_1(t_k)\alpha_1 + a_k\alpha_2 + d_3(t_k)\alpha_3\right). \tag{28}$$

We see that the probability of failure depends on the current treatment status a_k, the treatment initiation time t^* (subjects on treatment), and on either the study time elapsed t_k (subjects off treatment) or the time since treatment initiation $t_k - t^*$ (subjects on treatment). If there is no treatment effect, $\alpha_2 = 0$ and $\alpha_1 = \alpha_3$. All other things held equal, $\alpha_2 < 0$ corresponds to a positive treatment effect, i.e. treatment causes a persistent decrease in the hazard. The case $\alpha_3 < \alpha_1$ also corresponds to a positive treatment effect, i.e. treatment causes the hazard to grow more slowly as a function of time. In a situation where $\alpha_2 < 0$, but $\alpha_3 > \alpha_1$, the effect of treatment is ambiguous. At certain times, it is less hazardous to be on treatment, while at others, it is less hazardous to be off treatment. In reality, the outcome of interest is survival time $S_{\bar{a}}$ and, therefore, we want to find the treatment regime that will maximize, for example, median survival. Therefore, we will estimate the causal parameter $\boldsymbol{\alpha}$ of the MSM and estimate median survival for each possible treatment regime. We note that a 'treatment regime' in this setting is completely specified by the treatment initiation time. We refer to Bryan et al. [2002] for a description on how to simulate data from the above model.

4.2 Analysis of Simulated Data

First we describe the observed data structure. Each subject will exhibit an outcome of the survival process Y that is a vector of zeros followed by exactly one one, i.e. something of the form (0, 0, ... , 1). Of the same length as this vector, we will have the outcome of the covariate and treatment processes. By concatenating this subject-specific data, we create a dataset from which to estimate the treatment mechanism and the causal parameter $\boldsymbol{\alpha}$.

For the details on how to calculate the IPTW estimator based on the simulated data, we refer to Bryan et al. [2002]. In order to calculate the DR estimator, we need to estimate the projection term $IC_{SRA}(O|\mathcal{Q}_X, g)$ in the estimating function (13). Given a fitted $\mathcal{Q}_n, g_n$, $IC_{SRA}(O|\mathcal{Q}_n, g_n)$ can be calculated using a Monte-Carlo simulation as we described in section 2.3. $\mathcal{Q}_X$ includes two parts. The first part is the conditional distribution of Y_j given the past and the second part is the conditional distribution of L_j given the past. In this simulation, given the baseline covariate L_0, the future covariates are degenerate depending on known parameters. So we only need to model the condition distribution of Y_j given the past. We will discuss this in the next section.

4.3 Modelling the Partial Likelihood $\mathcal{Q}_X$

To compute the DR estimator, we need to assume a model for the distribution of $Y(t_k)$ given $(A_k, \mathcal{F}_k)$. In our simulation, we generate $Y_{\bar{a}}$ satisfying the MSM and set $Y \equiv Y_{\bar{A}}$. So we actually don't know the distribution of $Y(t_k)$ given $(A_k, \mathcal{F}_k)$ for the simulated data. We also need the MLE $\boldsymbol{\alpha}(\widehat{\mathcal{Q}}_X)$ of $\boldsymbol{\alpha}$ corresponding to the MLE $\widehat{\mathcal{Q}}_X)$ of $\mathcal{Q}_X$ to calculate the projection term in estimating function (13). In this section, we consider several candidate models for the law of $Y(t_k)$ given $(A_k, \mathcal{F}_k)$ and calculate the MLE of α using the method described in section 2.2. In our simulation model 3 produced the best MLE for the causal parameter α_2. We will use Model 3 to calculate the DR estimator in our simulation. Note that our model is thus misspecified which represents a realistic scenario.

Model 1:

$$\begin{aligned}&\operatorname{logit} P(Y_k = 1|\bar{Y}_{k-1} = 0, \bar{L}_k, \bar{A}_k)\\&\quad = \beta_0 + \beta_1 t_k + \beta_2 A_k + \beta_3 L_k + \beta_4(t_k \times A_k) + \beta_5(A_k \times L_k)\end{aligned} \tag{29}$$

Model 2:

$$\begin{aligned}&\operatorname{logit} P(Y_k = 1|\bar{Y}_{k-1} = 0, \bar{L}_k, \bar{A}_k)\\&\quad = \beta_0 + \beta_1 t_k + \beta_2 A_k + \beta_3 L_k + \beta_4(A_k \times L_k)\end{aligned} \tag{30}$$

Model 3:

$$\operatorname{logit} P(Y_k = 1|\bar{Y}_{k-1} = 0, \bar{L}_k, \bar{A}_k) = \beta_0 + \beta_1 t_k + \beta_2 L_k + \beta_3(A_k \times L_k) \tag{31}$$

Table 1 summarizes the bias and MSE of the maximum likelihood estimates for $\boldsymbol{\alpha}$ based on these three models.

		Bias	Var	MSE
α_0	MLE1	0.292	0.014	0.099
	MLE2	0.316	0.010	0.109
	MLE3	0.625	0.006	0.397
α_1	MLE1	0.063	0.001	0.005
	MLE2	0.046	0.001	0.003
	MLE3	0.023	0.001	0.001
α_2	MLE1	1.654	0.134	2.870
	MLE2	1.677	0.143	2.954
	MLE3	0.133	0.041	0.058
α_3	MLE1	0.175	0.003	0.033
	MLE2	0.185	0.003	0.037
	MLE3	0.078	0.001	0.007

Table 1: Bias, Variance and MSE of MLE's (200 replicates)

4.4 Simulation Results

In all of the simulations described below, the study takes place over time interval [0,20]. The monitoring times are $\{t_0 = 0, t_1 = 1, \ldots, t_{19} = 19\}$. We measure the covariate once in every five intervals, therefore the measuring times are $\{t_0, t_5, t_{10}, t_{15}\}$. The treatment times are t_0 and t_5. Therefore the set of possible treatment paths is given by

$$\mathcal{A} = \{(a_0, a_5) : a_0, a_5 \in \{0, 1\}, a_0 \leq a_5\} = \{(0,0), (0,1), (1,1)\}. \tag{32}$$

The value of $\boldsymbol{\alpha}$, the parameter of the logistic MSM given in equation (27), is given by

α_0	α_1	α_2	α_3
-3.04	0.175	-1.5	0.388

Since $\alpha_2 < 0$ and $\alpha_1 > \alpha_3$, there is a treatment effect. In fact, we are in one of the ambiguous situations described in section 4.1, in which it is not immediately apparent which treatment strategy is best. In the implementation of the IPTW estimator we truncate the weights. That is, if the weight is greater than 3 we set it to be 3. We report the bias and mean squared error (MSE) of α_2 for Naive, IPTW MLE and DR estimators in table 2. Recall that in our marginal structura model α_2 is our primary parameter of interest.

	DR	IPTW	MLE	Naive
Simulation 1: confounding is extreme, truncation of weights of weights at 3 *and* Model 3 for the likelihood				
Bias	0.06	0.649	0.143	2.059
MSE	0.627	1.216	0.058	4.362
Simulation 2: g is misspecified *and* Model 3 for the likelihood				
Bias	0.004	0.323	0.188	0.353
MSE	0.025	0.191	0.063	0.209
Simulation 3: g is correctly specified *and* Model 4 for the likelihood				
Bias	0.002	0.032	1.464	0.377
MSE	0.054	0.059	2.167	0.227
Simulation 4: confounding is extreme *and* Model 4 for the likelihood				
Bias	0.499	0.768	1.901	1.940
MSE	0.645	0.855	3.642	3.858
Simulation 4: confounding is normal *and* Model 3 for the likelihood				
Bias	0.0054	0.0214	0.1550	0.3377
MSE	0.0251	0.0413	0.0456	0.1660

Table 2: Finite sample bias and MSE of α_2 for all the estimators in the 5 simulation settings

4.4.1 Simulation 1

The main point of this simulation is to show the relative performance of the IPTW, double robust and ML estimators in case that the confounding is so severe that the identifiability assumption is practically violated. The presence and degree of confounding is completely determined by the parameter $\boldsymbol{\theta}$ in model (26).

The first value of $\boldsymbol{\theta}$ we consider in this simulation is $\boldsymbol{\theta} = (-6.7, 0, 0.5)$. (if $w = 30$ which is the maximum baseline covariate seen in one simulation, then the probability of getting treatment 0 is $1/(1 + exp(8.3)) = 0.00025$). We truncate the weights at 3.

The finite sample bias of the IPTW estimate is large because the truncation of the weights makes the IPTW estimating function biased even though we use a correctly specified model for g. Because of the truncation, this case corresponds with actually choosing a wrong model for the treatment mechanism g. We will see in the other simulations that the truncation does not affect the unbiasedness of the IPTW estimate when the confounding is not extreme. The finite sample bias of the DR estimate remains small since the DR estimating function remains unbiased at correctly specified likelihood and misspecified g. In fact, the finite sample bias of the DR is significantly smaller than the bias of DR. It was interesting to see that the finite sample bias of the DR estimate is much smaller than that of the MLE for α_0 and α_3, not reported here.

The difference in MSE between the IPTW, the DR estimate, and the MLE is significant, with the MLE being a clear winner. We note that the variance of the MLE is supposed to be smaller than the variance of the DR, since the MLE assumes a smaller model.

4.4.2 Simulation 2

In this simulation we will show the relative performance of the four estimators when the treatment model is misspecified. We use the following misspecified treatment model:

$$\operatorname{logit} P(A_k = 1 \mid \bar{L}_k, \bar{A}_{k-1} = 0) = \theta_0 + \theta_1 t_k + \theta_2 \sin(L_k). \tag{33}$$

In other words, we regress on $\sin(L_k)$ instead of L_k. The value of $\boldsymbol{\theta}$ we use for this simulation is $\boldsymbol{\theta} = (-0.9, 0, 0.04)$.

We note that the finite sample bias of the DR estimate is much smaller than that of the IPTW estimate. This is due to the protection against misspecification of g at a correctly specified likelihood $\mathcal{Q}_X$. The bias of DR is also smaller than the bias of the MLE.

The difference in MSE for α_2 between the IPTW and DR estimates is dramatic in this case, while the DR estimator now heavily outperforms the MLE.

4.4.3 Simulation 3

In this simulation, we will show the relative performance of the four estimators when the likelihood for Y_k given the past is misspecified. We use the following misspecified model to estimate the distribution of Y_k given the past.

Model 4

$$\text{logit}\, P(Y_k = 1|\bar{Y}_{k-1} = 0, \bar{L}_k, \bar{A}_k) = \beta_0 + \beta_1 A_k + \beta_2 L_k + \beta_3 (A_k \times L_k) \qquad (34)$$

Note that we deleted the t_k term in Model 4 from Model 2. The value of $\boldsymbol{\theta}$ for this simulation is same as simulation 2: $\boldsymbol{\theta} = (-0.9, 0, 0.04)$.

We note that the finite sample bias of the DR estimate is much smaller than that of the MLE. This is due to the fact that, at correctly specified treatment mechanism, the consistency of the DR estimator is protected against misspecification of the likelihood

The finite sample MSE of the DR estimate for α_2 is slightly smaller than that of the IPTW estimate while it is 40 times as small than the MSE of the MLE.

4.4.4 Simulation 4

In this simulation, we consider the case where confounding is extreme and the likelihood $\mathcal{Q}_X$ is misspecified. The value we choose for $\boldsymbol{\theta}$ is the same as in simulation 1, $\boldsymbol{\theta} = (-6.7, 0, 0.5)$. We use model (34) to estimate the distribution of Y_k given the past.

In this case, all the estimates are supposed to be inconsistent and thus biased. We note that the DR estimate is the least biased and most efficient, and the MLE is heavily biased.

4.4.5 Simulation 5

In this last simulation we set the confounding at a normal level $\boldsymbol{\theta} = (-0.9, 0, 0.04)$ and assume Model 3 for the likelihood so that it is approximately correctly specified.

In this case, both the IPTW and the DR estimates are known to be asymptotically consistent. We note that the finite sample MSE of the DR estimate is smaller than that of the IPTW estimate, and the MLE is by far the most biased estimator. This shows that a bias in the MLE due to slight misspecification of the model for $\mathcal{Q}_X$ does result in much less of a bias in the DR estimator.

5 Analysis of SPARCS Data

Here we apply the methodology developed in previous sections to analyze data from a project entitled "Study of Physical Performance and Age Related Changes in Sonomans" (SPARCS) [Tager et al., 2000b]. SPARCS is a community-based longitudinal study of physical activity and fitness in people at least 55 years of age who live in Sonoma, California. One of the goals of SPARCS and the primary goal of the current analysis is to estimate the causal effect of increased physical activity on physical functioning.

5.1 Data Structure

The subset of the data that we examine here was collected in the first three home evaluations of female SPARCS participants $n = 947$, over the time period May, 1993 - 1999.

Our measure of physical activity (a^t) is based on an *activity score* that is recorded for each subject at each evaluation. The activity score takes values in the set $\{1, 2, 3, 4\}$, where 4 corresponds to the highest level of activity. We define a time-dependent treatment process $A^t(t_k)$ that is an indicator for an activity score of 3 or 4 during the interval (t_k, t_{k+1}), which implies that the subject is engaging in moderately vigorous activity. Note that, although subjects are not being actively treated in any way, we can simply handle a subject's self-chosen activity level as an intervention whose efficacy we wish to measure.

The outcome of interest (y), physical functioning, is recorded for each subject at each evaluation. It is a dichotomous variable with 1 indicating "disabled".

At the initial evaluation, information on the the following baseline covariates is obtained: age in years (age), indicator of activity decline in past 5 - 10 years, Bmass2, Bmass3. This collection of variables is referred to collectively as W. At each evaluation, including the baseline evaluation, information on the following time-dependent covariates is obtained: indicator of other health conditions, Ln2Fat, Q1D, Speed, Dps2, Along, other These variables will be referred to collectively as L_k.

One expects that the variables W and L_k can influence both the activity level and the physical functioning. Therefore, we must regard W and L_k as potential confounders in our study of the causal relationship between activity level and physical functioning.

The data we study here is different from what we study in the simulation. There are only three time point: t_0, t_1 and t_2. The treatments, covariates and outcomes are measured at each time point until the subject drops out or dies. So there are two types of censoring. We refer to the censoring time as D which is the minumum of death and dropout time. Of the 1197 participants, each subject accumulates a history until the earliest of these events: end-of-study $K = 2$, or censoring D.

We use the following rules to deal with missing values: the subjects were excluded from the analysis as a result of missing baseline info for 1 or more of the following: Ln2Fat (extreme outliers), decline, bmass, overall health, Q1D. Other subjects were treated as right censored in the analysis from the moment there was missing information.

5.2 Causal Models

The full data for a subject includes the uncensored outcome (physical functioning) and covariate processes for every possible treatment regime is

$$X = (\bar{Y}_{\bar{a}^t}(t_K), \bar{L}_{\bar{a}^t}(t_K), W; \bar{a}^t \in \mathcal{A}^t),$$

where $K = 2$. In contrast, the observed data includes the outcome and covariate processes corresponding to the subject's actual treatment history possibly subject to censoring:

$$O = (R = t_K \wedge D, \bar{A}^t(R), \bar{Y}(R), \bar{L}(R), W)$$

We consider the following Marginal Structural Model for $\bar{Y}_{\bar{a}^t}$

$$P(Y_{\bar{a}^t}(t_k) = 1) = \alpha_0 + \alpha_1 t_k + \alpha_2 a_k^t$$

In order to compute the IPTW estimator, we must model the treatment and censoring mechanisms to form the weights. We assume the following logistic regression model for the treatment mechanism:

$$P(A_k^t = 1 | \bar{A}_{k-1}^t, \bar{L}_k, W) = \lambda_{t_k}(\boldsymbol{\gamma}), \tag{35}$$

where $\lambda_{t_k}(\boldsymbol{\gamma}) = \text{logit}^{-1}\left(\gamma_{0,k} + \gamma_{1,k} A_{k-1}^t + \gamma_{2,k}^T L_k + \gamma_{3,k}^T W\right)$. The usual partial mle $\boldsymbol{\gamma}$ can be computed using standard software and we compute the treatment contribution to the estimated weights as described before in section 4.2.

We do not assume that the drop-out time D is independent of survival, but we do assume that D is independent conditional on the observed covariate history as in section 3. Therefore, our weights will include the probability of not being censored due to dropout, in addition to the above probability on the treatment mechanism. We define a drop-out censoring process $A_k^c = I(D \in (t_k, t_{k+1}))$. If the subject drops out before death or end of the study, $\bar{A}^c = (0, \ldots, 0, 1)$; otherwise $\bar{A}^c = (0, \ldots, 0, 0)$. We assume the following intensity model for the drop-out process:

$$E(dA_k^c | \bar{L}_k, W, \bar{A}_k^t, \bar{A}_{k-1}^c) = I(\bar{A}_{k-1}^c = 0) \times \pi_{t_k}(\boldsymbol{\beta}), \tag{36}$$

where $\pi_k(\boldsymbol{\beta}) = \text{logit}^{-1}\left(\beta_{0,k} + \beta_{1,k} A_k^t + \beta_{2,k}^T L_k + \beta_{3,k}^T W\right)$. Just as with the treatment process, we fit the regression implied by (36). Denote the estimator by $\widehat{\boldsymbol{\beta}}$ and the implied probability of dropout at time t_k for subject i by $\widehat{q}_{ik}$. The same quantities, but based on a regression in which $\bar{L}_k$ is omitted as a predictor, are denoted by $\widetilde{\boldsymbol{\beta}}$ and $\widetilde{q}_{ik}$. The censoring contribution to the estimated stabilized weight is then

$$\prod_{j=0}^{k} \frac{(1 - \widetilde{q}_{ij})}{(1 - \widehat{q}_{ij})}$$

The treatment contribution is calculated as it was earlier in the simulations and the estimated stabilized weights are an element-wise product of treatment and censoring terms.

To simplify the estimation of IC_{SRA}, as we described in section 2.3, at each time point, we extract two covariates from the fitted treatment and censoring mechanisms respectively. The first covariate $L_{trt,k} \equiv \hat{\gamma}_{0,k} + \hat{\gamma}_{1,k} A_{k-1}^t + \hat{\gamma}_{2,k}^T L_k + \hat{\gamma}_{3,k}^T W$ includes all the information contributing to the treatment mechanism. The treatment model now simply becomes

$$P(A_k^t = 1 | \bar{A}_{k-1}^t, \bar{L}_k, W) = \text{logit}^{-1}(L_{trt,k}), \tag{37}$$

	Intercept	Time	Treatment
$\widehat{\boldsymbol{\alpha}}_n^{dr}$	-0.873	0.071	-0.126
	(0.097)	(0.059)	(0.109)
$\widehat{\boldsymbol{\alpha}}_n^{iptw}$	-0.908	0.182	-0.089
	(0.100)	(0.053)	(0.117)
$\widehat{\boldsymbol{\alpha}}_n^{ml}$	-1.339	0.624	-0.139
$\widetilde{\boldsymbol{\alpha}}_n$	-0.753	0.116	-0.364

Table 3: Data analysis results of four different estimators

The other covariate $L_{cen,k} \equiv \hat{\beta}_{0,k} + \hat{\beta}_{1,k} A_k^t + \hat{\beta}_{2,k}^T L_k + \hat{\beta}_{3,k}^T W$, includes the information contributing to the censoring covariates. Now the censoring model becomes

$$E(dA_k^c | \bar{L}_k, W, \bar{A}_k^t, \bar{A}_{k-1}^c) = I(\bar{A}_{k-1}^c = 0) \times \text{logit}^{-1}(L_{cen,k}). \tag{38}$$

In the reduced data, at each time point there are only two covariates $L_{trt,k}$ and $L_{cen,k}$. To compute the MLE and IC_{SRA}, we need the conditional distributions of Y_k given the past, $L_{cen,k}$ given the past and $L_{trt,k}$ given the past. At each time point, we assume the distribution of $L_{trt,k}$ given the past is normally distributed with mean equal to a linear function of $Y_{k-1}, L_{trt,k-1}$ and A_{k-1}^t, the distribution of $Lcen, k$ given the past is normal with mean equal to a linear function of $L_{trt,k}, L_{cen,k-1}$ and A_{k-1}^t, the distribution of Y_k given the past is linear logistic function with covariates $A_k^t, L_{trt,k}, L_{cen,k}$ and Y_{k-1}. We use maximum likelihood to estimate the parameters. The details on computing $\widehat{\boldsymbol{\alpha}}_n^{ml}$ and IC_{SRA} using Monte-Carlo simulation were described previously.

5.3 Results

Table 3 reports the estimate of $\boldsymbol{\alpha}$ based on the DR, IPTW, MLE and Naive estimators. The conservatively estimated standard errors described in Section 2.4 of the DR and the IPTW estimators are reported in parentheses.

The data analysis suggest that physical activity decreases the probability of being disabled in physical functioning. This effect is seen in all of the four estimators. But for both the DR and the IPTW estimates, the effect is not statistically significant. We also note that the DR estimator is closer to the IPTW than to the MLE. This might be due to misspecification of the partial likelihood $\mathcal{Q}_X$.

References

J. Bryan, Z. Yu, and M. van der Laan. Analysis of longitudinal marginal structural models. Submitted to Biostatistics, 2002.

R. Gill and J. Robins. Causal inference in complex longitudinal studies: continuous case. *Ann. Stat.*, 29(6), 2001.

R.D. Gill, M.J. van der Laan, and J.M. Robins. Coarsening at random, characterizations, conjectures and counter examples. In *Proceedings of the First Seattle Symposium in Biostatistics 1995*, pages 255–294. Springer Verlag, New York, 1997.

D.F. Heitjan and D.B. Rubin. Ignorability and coarse data. *Annals of Statistics*, 19: 2244–2253, 1991.

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

M. Jacobsen and N. Keiding. Coarsening at random in general sample spaces and random censoring in continuous time. *Annals of Statistics*, 23:774–786, 1995.

R. Neugebauer and M.J. van der Laan. Why prefering doubly robust estimation? application to point treatment marginal structural models. Submitted to Journal of Statistical Planning and Inference, 2002.

J. M. Robins. Errata to: "A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect" [Math. Modelling **7** (1986), no. 9-12, 1393–1512; MR 87m:92078]. *Comput. Math. Appl.*, 14(9-12):917–921, 1987a. ISSN 0898-1221.

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

James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Math. Modelling*, 7(9-12):1393–1512, 1986. ISSN 0270-0255. Mathematical models in medicine: diseases and epidemics, Part 2.

James M. Robins. Marginal structural models versus structural nested models as tools for causal inference. In *Statistical models in epidemiology, the environment, and clinical trials (Minneapolis, MN, 1997)*, pages 95–133. Springer, New York, 2000a.

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" [Math. Modelling **7** (1986), no. 9-12, 1393–1512; MR 87m:92078]. *Comput. Math. Appl.*, 14(9-12):923–945, 1987b. ISSN 0097-4943.

J.M. Robins. The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In L. Sechrest, H. Freeman, and A Mulley, editors, *Health Service Research Methdology: A Focus on AIDS*, pages 113–159. NCHSR, U.S. Public Health Service, Dordrecht, 1989.

J.M. Robins. Causal inference from complex longitudinal data. In *Latent variable modeling and applications to causality (Los Angeles, CA, 1994)*, pages 69–117. Springer, New York, 1997.

J.M. Robins. Marginal structural models. In *1997 Proceedings of the Section on Bayesian Statistical Science*, pages 1 – 10, Alexandria, VA, 1998. American Statistical Association.

J.M. Robins. Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association.* American Statistical Association, Alexandria, VA, 2000b.

I. Tager, T. Haight, Hollenberg, and Satariano. Physical functioning and mortality in elderly females. Unpublished manual, University of California, Berkeley, School of Public Health., 2000a.

I. Tager, T. Haight, Hollenberg, and Satariano. Physical functioning and mortality in elderly females. Unpublished manual, University of California, Berkeley, School of Public Health., 2000b.

M.J. van der Laan and J. M. Robins. *Unified methods for censored longitudinal data and causality.* Springer-Verlag, New York, 2002.

Z. Yu. Causal inference in longitudinal studies. In *Ph.D. thesis.* University of California, Berkeley, 2002.

Z. Yu and M. van der Laan. Construction of counterfactuals and the g-computation formula. Submitted to Scandinavian Journal of Statistics, 2002.