

UC Berkeley

UC Berkeley Previously Published Works

Title

Double Robust Efficient Estimators of Longitudinal Treatment Effects: Comparative Performance in Simulations and a Case Study

Permalink

<https://escholarship.org/uc/item/5p76c6c0>

Journal

The International Journal of Biostatistics, 15(2)

ISSN

2194-573X

Authors

Tran, Linh
Yiannoutsos, Constantin
Wools-Kaloustian, Kara
[et al.](#)

Publication Date

2019-11-01

DOI

10.1515/ijb-2017-0054

Peer reviewed



HHS Public Access

Author manuscript

Int J Biostat. Author manuscript; available in PMC 2019 December 07.

Published in final edited form as:

Int J Biostat. ; 15(2): . doi:10.1515/ijb-2017-0054.

Double Robust Efficient Estimators of Longitudinal Treatment Effects: Comparative Performance in Simulations and a Case Study

Linh Tran¹, Constantin Yiannoutsos², Kara Wools-Kaloustian³, Abraham Siika⁴, Mark van der Laan⁵, Maya Petersen⁵

¹Department of Biostatistics, University of California Berkeley, Berkeley, CA, USA

²Department of Biostatistics, Indiana University Richard M Fairbanks School of Public Health, Indianapolis, IN, USA

³Infectious Diseases, Howard Hughes Medical Institute - Indiana University School of Medicine, Indianapolis, IN, USA

⁴Moi University, Eldoret, Kenya

⁵University of California at Berkeley, Berkeley, CA, USA

Abstract

A number of sophisticated estimators of longitudinal effects have been proposed for estimating the intervention-specific mean outcome. However, there is a relative paucity of research comparing these methods directly to one another. In this study, we compare various approaches to estimating a causal effect in a longitudinal treatment setting using both simulated data and data measured from a human immunodeficiency virus cohort. Six distinct estimators are considered: (i) an iterated conditional expectation representation, (ii) an inverse propensity weighted method, (iii) an augmented inverse propensity weighted method, (iv) a double robust iterated conditional expectation estimator, (v) a modified version of the double robust iterated conditional expectation estimator, and (vi) a targeted minimum loss-based estimator. The details of each estimator and its implementation are presented along with nuisance parameter estimation details, which include potentially pooling the observed data across all subjects regardless of treatment history and using data adaptive machine learning algorithms. Simulations are constructed over six time points, with each time point steadily increasing in positivity violations. Estimation is carried out for both the simulations and applied example using each of the six estimators under both stratified and pooled approaches of nuisance parameter estimation. Simulation results show that double robust estimators remained without meaningful bias as long as at least one of the two nuisance parameters were estimated with a correctly specified model. Under full misspecification, the bias of the double robust estimators remained better than that of the inverse propensity estimator under misspecification, but worse than the iterated conditional expectation estimator. Weighted estimators tended to show better performance than the covariate estimators. As positivity violations increased, the mean squared error and bias of all estimators considered became worse, with covariate-based double robust estimators especially susceptible. Applied analyses showed

similar estimates at most time points, with the important exception of the inverse propensity estimator which deviated markedly as positivity violations increased. Given its efficiency, ability to respect the parameter space, and observed performance, we recommend the pooled and weighted targeted minimum loss-based estimator.

Keywords

aiptw; causal inference; double robust; efficient influence function; iptw; longitudinal treatment; multiple testing; semiparametric models; tmle

1 Introduction

Estimating the effect of an exposure on an outcome is a common goal in research. In settings where the exposure of interest is longitudinal, or in other words, is comprised of multiple treatment decisions over time, identification of such causal parameters requires non-traditional statistical estimands, such as that provided by the longitudinal g-computation formula under an assumption of sequential randomization [1, 2]. Many estimators have been developed that target these longitudinal causally-motivated parameters, including multiple approaches to double robust estimation. One such example is a survival function at discrete time points, but under a specific treatment intervention, e.g. never treat. In this paper we provide an overview of estimators for longitudinal causal effects, with a focus on a particular class of double robust estimators. We compare the performance of alternative double robust estimators, as well as G-computation and inverse probability weighted estimators using simulated and real data.

Available estimators for longitudinal causal effects differ in their efficiency, in their nuisance parameters (and their choice of estimator of the nuisance parameter, such as data-adaptive or parametric-model based), and in their robustness. For example, the consistency of the “classical” longitudinal g computation approach [1] relies on consistent estimation of a series of conditional densities of non-exposure variables, while inverse probability weighted estimators [3, 4] rely on consistent estimation of the exposure mechanism.

A number of these estimators are double robust (DR). In other words, they have the appealing property that if either of two nuisance parameters are estimated consistently, then the resulting estimator will be consistent. The DR estimators are also efficient in a semi-parametric statistical model, in that they obtain the lowest possible asymptotic variance among regular asymptotically linear estimators if both nuisance parameters are estimated consistently at reasonable rates. Such DR semi-parametric efficient estimators include, among others, those based on estimating equations [5, 6] and sequential regression approaches based on iterated conditional expectations [7–12].

We focus here on a particular class of double robust efficient estimators based on the iterative conditional expectation representation of the longitudinal g-computation formula [7, 9, 10], an approach that can improve performance by dramatically reducing the dimensionality of one of the two nuisance parameters. Specifically, the approach avoids the need to estimate the entire conditional distribution of the outcome and instead relies upon

the conditional expectation. Within this class, DR efficient estimators may be defined either as solutions to estimating equations, or instead as substitution estimators (where we substitute the true distribution for the empirically observed one). The latter approach can improve stability in the face of data sparsity (in particular, near violations of the positivity assumption [13–15]) through ensuring that resulting estimates respect the parameter space.

While a number of sophisticated estimators of longitudinal effects have been proposed, there is a relative paucity of research comparing these methods directly to one another. The purpose of this study is to compare these various approaches to estimating a causal effect in a longitudinal treatment setting using both simulated data and data measured from a cohort of patients treated with human immunodeficiency virus (HIV). This HIV clinical cohort was established by the Academic Model Providing Access to Healthcare (AMPATH), one of the care and treatment programs that contributes data to the International Databases to Evaluate AIDS, East Africa Consortium. One of the goals of the consortium is to develop new methods to analyze data collected from large clinical cohorts, including the ability to assess programmatic interventions that are implemented organically. Among eligible patients in this cohort, scientists are interested in determining the causal effect of enrolling into a low-risk express care program on all cause mortality and loss to follow-up rates.

Six distinct estimators are considered here. They are (i) a simple substitution estimator, based on estimating the iterated conditional expectation (ICE) representation of the longitudinal g-computation formula [7, 9], (ii) an inverse propensity weighted (IPW) method [3, 4, 13], (iii) an augmented IPW (AIPW) estimator that directly solves the efficient influence function (EIF) estimating equation [9, 10, 16, 17], (iv) a double robust iterated conditional expectation (DRICE) estimator as presented by Bang and Robins [7], Robins [9], Scharfstein, Rotnitzky, and Robins [10] using a “clever covariate” as part of a parametric model, (v) a modified version of DRICE in which the estimated inverse probabilities are applied as observational weights [18–20], and (vi) and (vii) modified versions of the DRICEs (iv) and (v), in which the sequential outcome regressions are first fit using machine-learning and then updated in a second targeting step [9, 12]. Estimators (iv)–(vii) are specific implementations of the general targeted minimum-loss based estimation framework [8, 12, 21, 22]. Estimators (iii)–(vii) are all semiparametric efficient estimators, solving the estimating equation corresponding to the EIF. We focus on estimators (i) and (iii)–(vii) due to their relation to each other (shared basis on the ICE representation of the g computation formula and corresponding EIF), and additionally include IPW due to its significance in the causal literature.

Our paper builds on the published evidence base comparing similar methods in several important ways. First, much prior work has focused on the point treatment setting, in which both the statistical parameters and corresponding estimators are substantially simpler than their longitudinal counterparts (for example, Petersen et al. [14], Kang and Schafer [18], Gutman and Rubin [23], Han and Wang [24], Hattori and Henmi [25], Li, Kleinman, and Gillman [26], Zhou, Zhang, Li, and Zhang [27]). Second, most of the prior work comparing the performance of estimators of longitudinal treatment effects has generally been limited to comparison of a single double robust efficient estimator (such as a single DRICE or AIPW estimator) to a simpler alternative such as IPW or g-computation (for example, Decker,

Hubbard, Crespi, Seto, and Wang [28], Neugebauer, Schmittdiel, and van der Laan [29]). In contrast, direct comparisons, as provided in this study, of the performance of a range of double robust estimators including those incorporating machine learning are rare (although see Schnitzer, Lok, and Bosch [30], for a comparison of DRICE estimators of the parameters of non-saturated marginal logistic regression model in the presence of censoring). Furthermore, this study directly evaluates the comparative performance of a range of DRICE estimators under gradually increasing levels of positivity violations [14, 15, 31]. Lastly, for the applied setting (although not in the simulation), it considers estimates from both the generalized linear modelling approach and by incorporating Super Learning [32], an ensemble data-adaptive machine learning algorithm.

The remainder of the manuscript is structured as follows. Section 2 formalizes the data structure, statistical, and causal models. Section 3 defines the target parameter of interest and reviews its EIF. Section 4 presents each of the estimators being considered, details on the nuisance parameters required for estimation, and reviews Super Learning. Section 5 reviews practical implementation of the estimators and presents the results of the analyses for the simulations. Section 6 applies the estimators to non-simulated (i.e. real world) data. Lastly, Section 7 concludes with a discussion of the methods, results, and future directions. For reference, we include a notation list in Appendix A. Appendix B contains tables of all experimental results for the interested reader.

2 Data

Consider, as a working example, right censored survival data in which subjects are followed from a baseline time point $t = 0$ up to some final time point $K + 1$. At each time point t , subjects may enroll into a treatment program. Regardless of whether they enroll, each subject is followed until the first of either (i) some terminal event of interest is observed, (ii) the subject is right censored due to lost to follow-up, or (iii) the end of follow-up is reached. Time-varying covariates are measured at each time point t that may affect subsequent treatment, covariates, and the outcome. Additionally, baseline covariates are measured at baseline ($t = 0$).

More formally, we consider an independent and identically distributed (iid) statistical data structure. Let $Y(t)$ be a failure indicator, a counting process which takes value 0 until the outcome event of interest is observed and subsequently switches to and remains at 1 for all remaining time points. We assume that $Y(0) = 0$ for everyone, i.e. that no subjects have experienced the event at the beginning of follow-up. Let $L1(t)$ be the time-varying covariate values measured at each time point t . We define $L1(0)$ to additionally include all baseline covariates measured at $t = 0$. We refer to the joint outcome and covariate variables at time t as $L(t) = (Y(t), L1(t)) : t = 0, 1, \dots, K + 1$. Let $A1(t)$ be the indicator of enrollment into the treatment program and $A2(t)$ be the indicator of censoring due to loss to follow-up. Each of these variables take value 0 until an enrollment or censoring event is observed, respectively, at which point they become fixed at 1. We collectively refer to the treatment and censoring processes as $A(t) = (A1(t), A2(t)) : t = 0, 1, \dots, K$. For notational convenience, we define all variables after failure or right censoring occurs as deterministically equal to their last

observed value. Our data can now be represented as n independent and identically distributed (iid) copies of the longitudinal data structure

$$O = (L(0), A(0), L(1), \dots, A(K), L(K + 1)) \stackrel{iid}{\sim} P_0. \tag{1}$$

2.1 Likelihood

We first consider the likelihood $L(O)$ for the data structure specified above. In particular, we use the following factorization, implied by the assumed time ordering $L(t) \rightarrow A(t)$ such that the likelihood for subject i is

$$\begin{aligned} p_0(O_i) &= p_0(L_i(0), A_i(0), L_i(1), A_i(1), \dots, L_i(K + 1)) \\ &= p_0(L_i(K + 1) | \bar{L}_i(K), \bar{A}_i(K)) \cdot p_0(A_i(K) | \bar{L}_i(K), \bar{A}_i(K - 1)) \\ &\quad \cdot p_0(L_i(K) | \bar{L}_i(K - 1), \bar{A}_i(K - 1)) \cdot p_0(A_i(K - 1) | \bar{L}_i(K - 1), \bar{A}_i(K - 2)) \tag{2} \\ &\quad \dots p_0(L_i(0)) \\ &= \left[\prod_{t=0}^{K+1} \underbrace{p_0(L_i(t) | \bar{L}_i(t - 1), \bar{A}_i(t - 1))}_{q_{0, L(t)}(L(t), Pa(L(t)))} \right] \cdot \left[\prod_{t=0}^K \underbrace{p_0(A_i(t) | \bar{L}_i(t), \bar{A}_i(t - 1))}_{g_{0, A(t)}(A(t), Pa(A(t)))} \right] \end{aligned}$$

where we define $\bar{X}(t) \equiv (X(1), X(2), \dots, X(t))$ to denote the history of variable X up to time t , $p_0(\cdot)$ to denote the density of P_0 with respect to some dominating measure, and $A(-1) = L(-1) = \emptyset$. We follow convention in using $Pa(X)$ to denote the parents of the node defined as the variables which precede it, and denoting the conditional probabilities $p_0(L(t) | \cdot)$ and $p_0(A(t) | \cdot)$ as the q and g factors for the likelihood, respectively, such that $q_{0, L(t)}(L(t), Pa(L(t))) = p_0(L(t) | \bar{L}(t - 1), \bar{A}(t - 1))$ and $g_{0, A(t)}(A(t), Pa(A(t))) = p_0(A(t) | \bar{L}(t), \bar{A}(t - 1))$. For ease of notation, we collectively refer to the product of $p_0(L(t) | \cdot)$ and $p_0(A(t) | \cdot)$ over all t respectively as

$$\begin{aligned} q_0 &\equiv \prod_{t=0}^{K+1} q_{0, L(t)}(L(t) | Pa(L(t))) \\ g_0 &\equiv \prod_{t=0}^K g_{0, A(t)}(A(t) | Pa(A(t))). \end{aligned} \tag{3}$$

2.2 Statistical model

With our data and likelihood clear, we consider a statistical model \mathcal{M} for the true distribution P_0 , such that if \mathcal{Q} and \mathcal{S} are the variationally independent sets of all possible values for q_0 and g_0 respectively, then the statistical model can be represented as

$\mathcal{M} = \{P = q \cdot g : q \in \mathcal{Q}, g \in \mathcal{S}\}$. We assume a semiparametric model, which restricts the set of possible distributions for the g_0 and q_0 components of the likelihood. Specifically, to respect the factual details that we know about the data generating process, we force three model restrictions on the conditional distributions of $g_{0,A(t)}, (A(t), Pa(A(t))) : k = 0, 1, \dots, K$, and $q_{0,L(t)}, (L(t), Pa(L(t))) : k = 0, 1, \dots, K + 1$.

1. Once $A1(t) = 1$, we have that $\underline{A1}(t+1) = 1$.
2. Once $A2(t) = 1$, we have that $\underline{A2}(t+1) = 1$.
3. Once $Y(t) = 1$, we have that $\underline{Y}(t+1) = 1$.

where we define $\underline{X}(t+1) \equiv (X(t+1), X(t+2), \dots, X(K+1))$ to denote the remaining history of X from time $t+1$ to $K+1$.

2.3 Causal model

In addition to the statistical model \mathcal{M} presented above, we can additionally represent further causal assumptions about our data generating mechanism using a causal model $\mathcal{M}^{\mathcal{F}}$. The non-parametric structural equation model (NPSEM) [33] reflecting our beliefs about the time-ordering and relationships between the exposure, covariates, and outcome of interest is

$$\begin{aligned} L(t) &= f_{L(t)}(\bar{L}(t-1), \bar{A}(t-1), U_{L(t)}): t = 0, 1, \dots, K, K+1 \\ A(t) &= f_{A(t)}(\bar{L}(t), \bar{A}(t-1), U_{A(t)}): t = 0, 1, \dots, K \end{aligned} \tag{4}$$

where $U \equiv (U_{L(t)}, U_{A(t)}): t = 0, 1, \dots, K+1$ are unmeasured exogenous variables from some underlying probability distribution P_U and the function $f_{A(t)}$ is restricted according to 1 and 2 above. Here, $f_0 \equiv (f_{L(t)}, f_{A(t)} : t = 0, 1, \dots, K+1)$ are causal functions which deterministically assign the observed values of O based on the values of the arguments provided. This causal model assumes that each measured variable may be affected by all measured variables preceding it. While we initially make no assumptions on P_U , allowing any two measured variables to share an unmeasured common cause, some restrictions on this distribution will be needed to identify the causal effects of interventions on the treatment and censoring variables.

3 Target causal parameter and identifiability

As causal parameter we focus on the mean of the counterfactual outcome at specific time points under a specific treatment intervention; contrasts of this parameter under distinct treatment interventions can be used to summarize a wide range of casual effects. The intervention on the exposure(s) of interest at each time point corresponds to deterministically setting the values of $A(t)$ to some fixed value $a(t)$ for all t in the causal model $\mathcal{M}^{\mathcal{F}}$ specified above, resulting in a modified distribution P_a^0 . The counterfactual outcome $Y_a(t^*)$ under this

intervention can be interpreted as the value of $Y(t^*)$ at some specific time point $t^* \in K + 1$ that would have been generated had $A(t)$ been deterministically set to $a(t)$ for all t . We denote the set of possible regimes of interest through time $t^* - 1$ as $\mathcal{A}_{t^* - 1}$, and only consider regimes under interventions to prevent loss to follow up (set $\bar{a}2 = 0$). Under the untestable assumption of sequential randomization [1, 2] and the testable assumption of positivity [14, 15], this parameter $\mathbb{E}Y_{\bar{a}}(t^*)$ is identifiable from the observed data using the longitudinal g-computation formula [1]:

$$\mathbb{E}Y_{\bar{a}}(t^*) = \sum_{\bar{l}(t^* - 1)} \mathbb{E}(Y(t^*) | \bar{L}(t^* - 1) = \bar{l}(t^* - 1), \bar{A}(t^* - 1), = \bar{a}(t^* - 1)). \tag{5}$$

$$\prod_{j=0}^{t^* - 1} P(l(j) | \bar{L}(j - 1) = \bar{l}(j - 1), \bar{A}(j - 1) = \bar{a}(j - 1))$$

We note that issues related to measurement error and interference are beyond the scope of the current paper and will be addressed in future work.

Sequential randomization

Sequential randomization [1], assumes that

$$Y_{\bar{a}}(t^*) \perp\!\!\!\perp A(t) \Big| Pa(A(t)): t = 0, 1, \dots, t^* - 1, \bar{a}(t^* - 1) \in \mathcal{A}_{t^* - 1} \tag{6}$$

That is, our treatment is independent of the counterfactual outcome given its parents or informally, that measured covariates are sufficient to control for confounding of treatment and informative censoring. A sufficient condition for this assumption to be met is if all unmeasured exogenous variables affecting the treatment and censoring nodes $U_{A(t)}$ are independent of the exogenous variables affecting future $Y(t)$ nodes given the observed past up to time t .

Positivity

Our assumption of positivity [13–15, 31] states that:

$$P_0(A(t) = a(t) | \bar{L}(t - 1), \bar{A}(t - 1) = \bar{a}(t - 1)) > 0: t = 0, \dots, t^* - 1, \bar{a}(t^* - 1) \in \mathcal{A}_{t^* - 1} \text{ a.e.} \tag{7}$$

Informally, we require that there is adequate support for our intervention of interest regardless of covariate history, i.e. that there is no “finite sample” bias. As we demonstrate below in simulations, even in situations where this assumption on P_0 holds, in finite samples certain covariate histories and treatment combinations may be poorly supported, resulting in data sparsity and, consequently, potential increases in estimator bias and variance which threaten valid inference.

In the scenario where the required assumptions stated above are met, the longitudinal g-formula eq. (5) provides our target statistical parameter to be estimated from the observed data, denoted $\Psi(P_0)$. For reasons that will become clear later, and following Bang and Robins [7], Robins [9], we note that by the tower property of expectations our target statistical parameter $\Psi(P_0)$ can also be expressed in an iterated conditional expectation form, such that

$$\mathbb{E}[Y_{\bar{a}}(t^*)] = \mathbb{E}[\mathbb{E}[\dots\mathbb{E}[\mathbb{E}[Y(t^*)|\bar{L}^{\bar{a}}(t^* - 1)]|\bar{L}^{\bar{a}}(t^* - 2)]\dots|\bar{L}^{\bar{a}}(0)]] \tag{8}$$

where for notational convenience, we define $\bar{L}^{\bar{a}}(t) \equiv (\bar{L}(t), \bar{A}(t) = \bar{a}(t))$. We refer the interested reader to Robins [9] for further details on the derivation.

4 Estimation

Below, we first provide the efficient influence function, which underlies the four double robust estimators we consider in this paper. We next review two non-DR estimators (the ICE and IPW), followed by each of the DR estimators in turn. In estimating the nuisance parameters required for consistency of our target parameter, there are a number of approaches that can be taken. We review two specific details below. At a high level, the consistency of the IPW estimator relies on consistency of the estimator of the treatment mechanism g_0 , the consistency of the ICE estimator depends on the consistency of the estimator of the iterated outcome regressions in eq. (8), and the DR estimators remain consistent if either nuisance parameters are estimated consistently, but rely on consistent estimation of both for efficiency.

Throughout we use a subscript n to denote an estimator, and $\hat{\psi}_n$ to denote a point estimate of the true parameter value $\psi_0 = \Psi(P_0)$. For ease of notation, we focus on the target parameter defined at a single final time point: $t^* = K + 1$.

4.1 Efficient influence function (EIF)

An estimator is efficient if and only if it has an influence function (IF) equal to the EIF for $\Psi(P_0)$, endowing it with the lowest asymptotic variance among regular asymptotically linear estimators [34–36]. The EIF is thus the basis for constructing the four DR semiparametric efficient estimators described in this paper. The EIF for our target parameter, denoted $D^*(P)(O)$, has been derived previously [10, 12, 16]:

$$\begin{aligned} D^*(P_0)(O) &= \sum_{t=0}^{K+1} D_t^*(q_0, g_0)(O) \\ &= \sum_{t=1}^{K+1} \frac{\mathbb{I}(\bar{A}(t-1) = \bar{a}(t-1))}{g_{0,0:t-1}^{\bar{a}}} (\bar{Q}_{0,L(t+1)}^{\bar{a}} - \bar{Q}_{0,L(t)}^{\bar{a}}) + \bar{Q}_{0,L(1)}^{\bar{a}} - \Psi(\bar{Q}_0^{\bar{a}}) \end{aligned} \tag{9}$$

where we define $\bar{Q}_{0, L(t)}^{\bar{a}} \equiv \mathbb{E}_0[\bar{Q}_{0, L(t+1)}^{\bar{a}} | \bar{L}(t-1), \bar{A}(t-1) = \bar{a}(t-1)]$ and $g_{0, 0:t-1}^{\bar{a}} \equiv \prod_{k=0}^{t-1} P_0(A(k) = a(k) | \bar{L}(k), \bar{A}(k-1) = \bar{a}(k-1))$. For notational convenience, we use $\bar{Q}_{0, L(K+2)}^{\bar{a}}$ to represent $Y(K+1)$ under the true distribution P_0 .

There are a number of points worth reemphasizing here. Firstly, this EIF is simply a sum of time-specific IFs over all time points. Thus, estimators that solve the estimating equation corresponding to the EIF can be constructed such that they solve the estimating equations individually at each time point. Secondly, when $K+1$ is equal to 1, i.e. there is only one time point, this IF reduces to the well known EIF for the point treatment setting [5, 16, 37]. Lastly, the IF has, in the denominator of the first term, the cumulative probability of treatment up to each time point t . Consequently, positivity violations or near violations eq. (7), where the probability of treatment given the past is extremely low, can have large effects on the performance of estimators solving the estimating equation for this IF.

4.2 Iterated conditional expectation estimation (ICE)

As stated in eq. (8), our parameter of interest can be represented as a series of iterated conditional expectations. From this representation, as described by Scharfstein, Rotnitzky, and Robins [38] and Robins [9], we can form an estimator which starts by estimating the inner most conditional expectation and iterating outward until reaching the outermost conditional expectation. In other words, one first regresses the outcome (at time $K+1$) on past covariates among those who followed the treatment of interest until the outcome was measured. Estimation then steps backwards in time, where at each step, the fit from the prior regression serves as a new “outcome” and is regressed on past covariates among those who followed the treatment up to that time point. The parameter estimate $\hat{\psi}_n^{ICE}$ is then just the empirical mean of the final regression fit (which is now only a function of baseline covariates) over all the observations. The specific algorithm proceeds as follows:

1. Let T denote the failure time. Estimate the innermost conditional expectation $\bar{Q}_{0, L(K+1)}^{\bar{a}} = \mathbb{E}_0[Y(K+1) | \bar{L}(K), \bar{A}(K) = \bar{a}(K)]$, where the expectation is known to be equal to 1 if $T < K+1$. We denote this estimate as $\bar{Q}_{n, L(K+1)}^{\bar{a}}$.
2. Given $\bar{Q}_{n, L(K+1)}^{\bar{a}}$, we can recursively iterate outwards for $t = K, K-1, K-2, \dots, 1$, estimating $\bar{Q}_{0, L(t)}^{\bar{a}} = \mathbb{E}_0[\bar{Q}_{n, L(t+1)}^{\bar{a}} | \bar{L}(t-1), \bar{A}(t-1) = \bar{a}(t-1)]$, acknowledging our slight abuse of notation, where again the expectation is known to be equal to 1 if $T < t$. We denote these estimates as $\bar{Q}_{n, L(t)}^{\bar{a}}$.
3. At $t = 1$, we have the estimate $\bar{Q}_{n, L(1)}^{\bar{a}}$, which now is a function of only $L(0)$. As indicated in eq. (8), our parameter estimate is simply the empirical expectation over $L(0)$, i.e., $\hat{\psi}_n^{ICE} = \mathbb{E}_n \bar{Q}_{n, L(1)}^{\bar{a}}$.

As with the established parametric g-computation estimator (e.g. Robins [39], Taubman, Robins, Mittleman, and Hernán [40]), this estimator only relies upon the q_0 portion of the likelihood in estimating our target parameter. Unlike the parametric g-computation

estimator, however, which relies on estimating each of the conditional probability distributions given in the g-computation formula above eq. (5), this estimator relies on only the iterated conditional expectations $\bar{Q}_0^{\bar{a}} \equiv \bar{Q}_{0, L(t)}^{\bar{a}}, t = 1, \dots, K+1$. Thus, we only need to correctly model the first conditional moments of these distributions. Consequently, it provides a substantial advantage over the parametric g-computation approach in that it avoids the need to estimate the full the conditional distributions (or densities) of each of the non-intervention factors given the past (e.g. the second or higher conditional moments).

4.3 Inverse probability weighted estimation (IPW)

Our parameter can also be estimated by up-weighting subjects from $L1(t)$ that are under-represented compared to the representation they would have had under a randomized treatment assignment [3, 4]. This approach can be understood as creating a *pseudo*-population in which the measured covariates are balanced between treatment groups [41]. More formally, we implement the following estimator [4]:

$$\hat{\psi}_n^{HT} = \frac{\frac{1}{n} \sum_{i=1}^n \frac{\mathbb{I}(\bar{A}_i(K) = \bar{a}(K))}{g_{n, 0:K, i}^{\bar{a}}} Y_i(K+1)}{\frac{1}{n} \sum_{i=1}^n \frac{\mathbb{I}(\bar{A}_i(K) = \bar{a}(K))}{g_{n, 0:K, i}^{\bar{a}}}} \tag{10}$$

where $g_{n, 0:K, i}^{\bar{a}} = \prod_{k=0}^K P_n(A_i(k) = a(k) | \bar{L}_i(k), \bar{A}_i(k-1) = \bar{a}(k-1))$. This estimator relies upon the consistent estimation of g_0 for consistency.

4.4 Augmented inverse probability weighted estimation (AIPW)

Realizing that the EIF in eq. (9) has mean zero and is a function of our target parameter [34, 35], we can straight forwardly form an estimating equation and solve for our parameter [10, 42]. In other words, we simply replace g_0 and $\bar{Q}_0^{\bar{a}}$, the true treatment mechanism and iterated conditional expectations, respectively, of the EIF with their corresponding estimates, set the corresponding estimating equation equal to 0, and solve for ψ . This naturally results in the estimating equation estimator

$$\begin{aligned} \hat{\psi}_n^{AIPW} &= \mathbb{E}_n D^*(\bar{Q}_n^{\bar{a}}, g_n)(O_i) + \Psi(\bar{Q}_n^{\bar{a}}) \\ &= \mathbb{E}_n \left[\sum_{t=1}^{K+1} \frac{\mathbb{I}(\bar{A}_i(t-1) = \bar{a}(t-1))}{g_{n, 0:t-1, i}^{\bar{a}}} (\bar{Q}_{n, L(t+1), i}^{\bar{a}} - \bar{Q}_{n, L(t), i}^{\bar{a}}) + \bar{Q}_{n, L(1), i}^{\bar{a}} \right] \end{aligned} \tag{11}$$

where again, for notational convenience, we use $\bar{Q}_{n, L(K+2), i}^{\bar{a}}$ to denote $Y_i(K+1)$.

4.5 Double robust iterated conditional expectation (DRICE)

As shown by Bang and Robins [7], we can form a DR estimator as a sequential regression estimator quite similar to the ICE approach from Section 4.2. This approach, however, additionally uses the inverse cumulative probability of treatment estimate

$g_{n,0:t-1}(\bar{a}(t-1))^{-1}$ times the indicator of following treatment $\mathbb{1}(\bar{A}_t(K) = \bar{a}(K))$ as a covariate in estimating $\bar{Q}_{0,L(t)}^{\bar{a}}$, consequently augmenting the original estimate of the conditional mean. In other words, we use essentially the same algorithm as in the non-double robust ICE estimator, with the difference that each iterated outcome regression now includes as additional covariate an indicator of following the treatment regime of interest divided by the inverse of the estimated probability of following that regime. Our iterated conditional expectation algorithm is therefore updated as follows:

1. Estimate $g_{n,0:t-1}^{\bar{a}}: t = K, K-1, \dots, 0$. We denote the estimates as $g_{n,0:t-1}^{\bar{a}}$.
2. Let T denote the failure time. Similar to the ICE estimator, we estimate the innermost conditional expectation $\bar{Q}_{0,L(K+1)}^{\bar{a}} = \mathbb{E}_0[Y(K+1) | \bar{L}(K), \bar{A}(K) = \bar{a}(K)]$, where the expectation is known to be equal to 1 if $T < K+1$. Using generalized linear models, we augment this initial estimate by adding, as an extra covariate, $\mathbb{1}(\bar{A}(K) = \bar{a}(K)) \times g_{n,0:K}^{\bar{a}, -1}$. We denote this estimate as $\bar{Q}_{n,L(K+1)}^{\bar{a},g}$.
3. Given $\bar{Q}_{n,L(K+1)}^{\bar{a},g}$, we can recursively iterate outwards for $t = K, K-1, K-2, \dots, 1$ estimating $\bar{Q}_{0,L(t)}^{\bar{a}} = \mathbb{E}_0[\bar{Q}_{n,L(t+1)}^{\bar{a},g} | \bar{L}(t-1), \bar{A}(t-1) = \bar{a}(t-1)]$. The expectation is also known to be equal to 1 if $T < t$. Using generalized, each estimate is augmented by adding, as extra covariate, $\mathbb{1}(\bar{A}(t-1) = \bar{a}(t-1)) \times g_{n,0:t-1}^{\bar{a}, -1}$. We denote these estimates as $\bar{Q}_{n,L(t)}^{\bar{a},g}$.
4. At $t = 1$, we have the estimate $\bar{Q}_{n,L(1)}^{\bar{a},g}$, which now is a function of only $L(0)$. As indicated in eq. (8), our parameter estimate is simply the empirical expectation over $L(0)$, i.e. $\hat{\psi}_n^{DRICE} = \mathbb{E}_n[\bar{Q}_{n,L(1)}^{\bar{a},g}]$.

We note that use of the linear link function, even for a continuous outcome, can result in an unstable estimator as there is not guarantee that the estimate will respect the bounds of the parameter space. In contrast, use of the logit link for either a binary outcome or an appropriately transformed continuous outcome [12] ensures that the estimator respects the parameter space and is a substitution estimator. Although the general targeted maximum likelihood framework was not recognized at the time, the resulting estimator is a targeted maximum likelihood estimator.

4.5.1 DRICE weighted—Rather than using the inverse probability estimates $g_{n,0:t-1}^{\bar{a}}$ as part of a covariate as presented by Bang and Robins [7], an alternative approach uses inverse of $g_{n,0:t-1}^{\bar{a}}$ as observational weights and estimates $\bar{Q}_{0,L(t)}^{\bar{a}}$ in the same manner as the approach above, but instead using $\mathbb{1}(\bar{A}_t(t-1) = \bar{a}(t-1))$ as a covariate [18–20]. We refer to this estimator as the weighted DRICE and to the previous estimator as covariate DRICE.

The approach of applying the inverse of probability estimates (in generalized linear models) $g_{n,0:t-1}^{\bar{a}}$ as weights can potentially aid us in the presence of positivity violations, since

small values of $g_{n,0:t-1}^{\bar{a}}$ can potentially skew the estimates of $\bar{Q}_{0,L(t)}^{\bar{a}}$ as an outlier if included in the clever covariate.

4.6 DRICE with targeted update of initial fit (Targeted minimum loss-based estimation (TMLE))

While the two estimators described in Section 4.5 are TMLEs, van der Laan and Gruber[12] subsequently explicitly placed these estimators in the more general TMLE framework. The targeted minimum loss-based framework requires a number of ingredients, including (i) the EIF $D^*(q, g)(O)$ defined above, (ii) a generalized loss function possibly indexed by a nuisance parameter $\mathcal{L}_{t, \bar{Q}_{L(t+1)}^{\bar{a}}}$ ($\bar{Q}_{L(t)}^{\bar{a}}$), (iii) a least favorable parametric submodel $\bar{Q}_{L(t)}^{\bar{a}}(\epsilon_t)$

chosen such that the linear span of the generalized score at zero fluctuation spans the EIF, and (iv) an updating algorithm which iteratively minimizes the generalized loss-based empirical risk over the parameters of the least favorable parametric submodel. Using this framework, van der Laan and Gruber [12] described the following the algorithm for estimation:

1. Estimate $g_{0,0:t-1}^{\bar{a}}: t = K + 1, K, \dots, 1$. We denote the estimates as $g_{n,0:t-1}^{\bar{a}}$.
2. Let T denote the failure time. Estimate $\bar{Q}_{0,L(K+1)}^{\bar{a}} = \mathbb{E}_0[Y(K+1) | \bar{L}(K), \bar{A}(K) = \bar{a}(K)]$, where the expectation is known to be equal to 1 if $T < K + 1$. We denote this estimate as $\bar{Q}_{n,L(K+1)}^{\bar{a}}$.
3. Update the initial fit $\bar{Q}_{n,L(K+1)}^{\bar{a}}$ based on the $K + 1$ -th chosen loss function $\mathcal{L}_{K+1, \bar{Q}_{L(K+2)}^{\bar{a}}}(\bar{Q}_{L(K+1)}^{\bar{a}}(\epsilon_{K+1}))$ and using the parametric submodel $\bar{Q}_{L(K+1)}^{\bar{a}}(\epsilon_{K+1})$. By setting $\hat{\epsilon}_{n,K+1} = \underset{\epsilon}{\operatorname{argmin}} P_{n, \mathcal{L}_{K+1, \bar{Q}_{L(K+2)}^{\bar{a}}}}(\bar{Q}_{L(K+1)}^{\bar{a}}(\epsilon))$, an updated fit is formed $\bar{Q}_{n,L(K+1)}^{\bar{a},*} = \bar{Q}_{L(K+1)}^{\bar{a}}(\hat{\epsilon}_{n,K+1})$ that is targeted at the parameter $\Psi(P_0)$.
4. Given $\bar{Q}_{n,L(K+1)}^{\bar{a},*}$, we can recursively for $t = K, K - 1, K - 2, \dots, 1$:
 - a. Estimate the conditional expectation $\bar{Q}_{0,L(t)}^{\bar{a}} = \mathbb{E}_0[\bar{Q}_{n,L(K+1)}^{\bar{a},*} | \bar{L}(t-1), \bar{A}(t-1) = \bar{a}(t-1)]$, where again the expectation is known to be equal to 1 if $T < t$. We denote this estimate as $\bar{Q}_{n,L(t)}^{\bar{a}}$.
 - b. Similar to Step 3, update $\bar{Q}_{n,L(t)}^{\bar{a}}$ using the loss function $\mathcal{L}_{t, \bar{Q}_{L(t+1)}^{\bar{a}}}(\bar{Q}_{L(t)}^{\bar{a}})$ with the parametric submodel $\bar{Q}_{L(t)}^{\bar{a}}(\epsilon_t)$. Again, minimizing the empirical loss function

$$\hat{\epsilon}_{n,t} = \underset{\epsilon}{\operatorname{argmin}} P_n \mathcal{L}_{t, \bar{Q}_{L(t+1)}^{\bar{a}}}(\bar{Q}_{L(t)}^{\bar{a}}(\epsilon)) \text{ result in the updated fit}$$

$$\bar{Q}_{n,L(t)}^{\bar{a},*} = \bar{Q}_{L(t)}^{\bar{a}}(\hat{\epsilon}_{n,t}) \text{ for time } t.$$

5. At $t = 1$, we have the estimate $\bar{Q}_{n,L(1)}^{\bar{a},*}$, which now is a function of only $L(0)$. Our parameter estimate is simply the empirical expectation over $L(0)$, i.e.

$$\hat{\psi}_n^{TMLE} = \mathbb{E}_n \bar{Q}_{n,L(1)}^{\bar{a},*}.$$

In practice, the non-negative log likelihood loss is typically chosen as the generalized loss $\mathcal{L}_{t, \bar{Q}_{L(t+1)}^{\bar{a}}}(\bar{Q}_{L(t)}^{\bar{a}})$ along with the submodel

$$\operatorname{logit} \bar{Q}_{L(t)}^{\bar{a}}(\epsilon_t) = \operatorname{logit} \bar{Q}_{n,L(t)}^{\bar{a}} + \epsilon_t h(t)$$

where the covariate $h(t) = \mathbb{I}(\bar{A}(t) = \bar{a}(t)) / g_{n,0:t}^{\bar{a}}$.

Similar to the DRICE estimators described in Section 4.5, this approach generates a substitution estimator that solves the EIF estimating equation. In terms of practical implementation, the primary difference between the two approaches is that for each sequential regression in turn, TMLE first forms the initial fit and subsequently uses the inverse probability estimates to update this initial fit, resulting in a 2-step approach. Conversely, DRICE presented in Section 4.5 includes the clever covariate in the initial fit of a parametric model for each sequential regression. Under parametric models, meaningful differences would not be expected. However, the two step approach makes possible integration of machine learning at the stage of the initial fit of the sequential outcome regressions. Of note, outside the more general TMLE framework, Robins [9] and Rotnitzky, Lei, Sued, and Robins [20] also proposed such a two step fit. While both the one and two step DRICE estimators are TMLEs, we use “TMLE” for the remainder of this paper to refer to the two step estimator described in this section, as this corresponds to the longitudinal TMLE that to our knowledge is in most common use.

4.6.1 TMLE weighted—As with the one step DRICE of Section 4.5, the TMLE of Section 4.6 can also be implemented using an alternative choice for our loss function and submodel [12]. The submodel

$$\operatorname{logit} \bar{Q}_{L(t)}^{\bar{a}}(\epsilon_t) = \operatorname{logit} \bar{Q}_{n,L(t)}^{\bar{a}} + \epsilon_t$$

and loss function

$$\mathcal{L}_{t, \bar{Q}_{L(t+1)}^{\bar{a}}}(\bar{Q}_{L(t)}^{\bar{a}}) = - \frac{\mathbb{I}(\bar{A}(t) = \bar{a}(t))}{g_{n,0:t}^{\bar{a}}} \left(\bar{Q}_{n,L(t+1)}^{\bar{a}} \log \bar{Q}_{n,L(t)}^{\bar{a}} + (1 - \bar{Q}_{n,L(t+1)}^{\bar{a}}) \log(1 - \bar{Q}_{n,L(t)}^{\bar{a}}) \right)$$

also satisfy the generalized score condition required for the TMLE. For the remainder of this paper, we refer to the estimator as the weighted TMLE and the prior estimator as the covariate TMLE.

4.7 Conditioning nuisance parameter estimation on \bar{a}

One approach to estimating the nuisance parameters $g_{0,0:t-1}^{\bar{a}}$ and $\bar{Q}_{0,L(t)}^{\bar{a}}$, presented by Bang and Robins [7] and van der Laan and Gruber [12], conditions each factor on having followed the treatment regime of interest \bar{a} (up to time t) and uses only these subjects to estimate g_0 and $\bar{Q}_0^{\bar{a}}$.

An alternative approach is to instead pool across all subjects regardless of treatment history in estimating g_0 and $\bar{Q}_0^{\bar{a}}$ and then to evaluate each on at $\bar{A}(t) = \bar{a}(t): t = 0, 1, \dots, K$. We refer to this latter approach as the pooled estimator and the approach that conditions on following \bar{a} as the stratified estimator. The pooled approach provides potentially more stability due to the increased sample size used to estimate each nuisance parameter, albeit at a potential cost of increased bias in nuisance parameter estimates. We note that if, as in our working example, $A(t)$ is a counting process that jumps to one once and remains there deterministically, the stratified and pooled approaches to estimating g_0 will be identical (n.b. the deterministic nature of $A(t)$ naturally implies conditioning on prior $A(t-1) = 0$ when estimating g_0). Therefore, the stratified approach for the IPW estimator is equivalent to the pooled approach.

4.8 Data adaptive estimation

In estimating these nuisance parameters, both parametric generalized linear model approaches and data-adaptive machine learning approaches are possible. Potential machine learning approaches include gradient boosting machines [43, 44], support vector machines [45, 46], neural networks [47], k-nearest neighbors [48], and data-adaptive parametric models such as ridge regression [49], least absolute shrinkage and selection operator [50], and elastic nets [51]. To prevent overfitting of the data, cross validation is employed in determining the best fit. Additionally, these candidates can be combined to provide one overall fit in an approach known as stacking or model averaging [52]. One such approach which incorporates all of these is known as Super Learning [32]. This algorithm uses V-fold cross validation to select the best convex combination of conditional density or probability fits within a user-specified library of potential candidates. If none of the candidates in the library is a correctly specified parametric model, then Super Learner has been proven to perform at least as well asymptotically as an oracle selector that selects the best candidate from the library based on the (unknown in practice) true distribution P_0 . Otherwise, (e.g. if a correct parametric model is among the members of the Super Learner library) the Super Learner achieves the rate of convergence of $\log(n)/n$ [32], which is almost parametric (with rate $1/n$). We therefore relied on Super Learner in our applied analyses, and refer the interested reader to van der Laan, Polley, and Hubbard [32] for further details.

Most of the double robust estimators considered, with the exception of DRICE and weighted DRICE, can incorporate machine learning while retaining valid theory based inference

(assuming adequate rates of convergence for estimators of g_0 and \bar{Q}_0). The exception is the one step DRICES of Section 4.5, which are only defined for parametric models on the iterated conditional expectations. We therefore did not consider the data-adaptive approaches for DRICE and weighted DRICE. In contrast, theory-based inference is no longer supported when machine learning approaches are used for nuisance parameter estimation in the IPW and ICE estimators, representing an additional advantage of using double robust estimators such as AIPW or TMLE.

5 Simulations

To evaluate and compare the performance of these longitudinal causal estimators, we first implement a simulation with a known data generating distribution and evaluate the bias, variance, and mean squared error of each estimator. We note that, due to computational limitations for our simulations, estimators were implemented using correctly specified and misspecified parametric models, rather than machine learning. Specifically, the number of simulations required to accurately estimate performance metrics of the estimators made the use of machine learning infeasible. Our applied analysis in Section 6 only requires that we run the machine learning models once.

5.1 Data generating distribution P_0

Our data were generated for times $t = 0, 1, 2, \dots, 6$ using the data structure described in Section 2 under a sample size of $n = 500$ as follows:

$$\begin{aligned}
 &W_1, W_3 \sim N(0, 1) \\
 &W_2 \sim \text{Ber}(\text{logit}^{-1}(-1)) \\
 &Y(t) \mid Y(t-1) = 0 \sim \text{Ber}(\text{logit}^{-1}(-1.9 + 1.2W_1 - 2.4W_2 - 1.8L_1(t-1) \\
 &\quad - 1) \\
 &\quad - 1.6L_2(t-1) + L_1(t-1)L_2(t-1) - A_1(t-1))) \\
 &L_1(t) \mid Y(t-1) = 0 \sim N(0.1 + 0.4W_1 + 0.6L_1(t-1) - 0.7L_2(t-1) \\
 &\quad - 0.45A_1(t-1), 0.5) \\
 &L_2(t) \mid Y(t-1) = 0 \sim N(-0.55 + 0.5W_1 + 0.75W_2 + 0.1L_1(t-1) \\
 &\quad + 0.3L_2(t-1) - 0.75A_1(t-1), 0.5) \\
 &A_1(t) \mid Y(t-1) = 0, A_1(t-1) = 0 \sim \text{Ber}(\text{logit}^{-1}(-1 - 1W_1 + 0.75W_2 \\
 &\quad + 1.2L_1(t-1) - 1.8L_2(t-1) + 0.8L_1(t) * L_2(t))) \\
 &\quad)
 \end{aligned} \tag{12}$$

Let $L(t) = (Y(t), L_1(t), L_2(t))$ where at time $t=0$, we have that the baseline $L(0) = (W_1, W_2, W_3, Y(0), L_1(0), L_2(0))$ and $A(t) = (A_1(t))$ for all t . We defined $L_1(-1) = L_2(-1) = A_1(-1) = 0$ and fixed $Y(0) = 0$. Once a subject experiences a failure, i.e. $Y(t) = 1$, all the remaining values remain at 1. Subjects enrolling, i.e. having $A_1(t) = 1$, stayed enrolled for the remainder of follow-up. The resulting observed data consisted of 500 i.i.d. copies of $O_i = (\bar{A}_i(5), \bar{L}_i(6)) \sim P_0$.

Our justification for generating data this way is to create data that is similar to what we observe in our applied setting (presented below in Section 6). For example, our observed data shows a noticeable amount of positivity violations (or near violations). To mimic this, we consider large parameter values for the treatment conditional distribution $A1(t)$, i.e. 1.2 and -1.8 . Furthermore, our desired outcome is a counting process. We consequently set up our simulation such that it is also monotonic (i.e. it can only increase from 0 to 1 and stays at 1 once it occurs for some time point t^\dagger for all $t \geq t^\dagger$.) We also note that, as our applied setting (most) likely has confounding present, we simulate this in our simulation through the use of $W_1, W_2, W_3, L1_1, L1_2$.

5.2 Target parameter $\Psi(P_0)$

We considered the intervention of interest to be “*never enroll*”, i.e. $\bar{a}(t) = 0: t = 0, 1, \dots, K$, and estimated the target casual parameter $\mathbb{E}(Y_{\bar{a}}(t^*))$, or the counterfactual cumulative failure probability through time t^* under an intervention to never enroll, for $t^* = 1, 2, \dots, 6$. Because the needed identifying assumptions hold by design in this simulation, the target causal and statistical parameter values are identical.

We determined the true parameter value $\psi_0(t^*) = t^* = 1, 2, \dots, 6$ by drawing observations under the post-intervention data generating distribution $P_0^{\bar{a}}$ with a sample size of 8×10^6 , where data were generated according to eq. (12) but setting $\bar{A1}(t^*) = 0$. Defining our true parameter as $\psi_0 \equiv (\psi_0(1), \psi_0(2), \dots, \psi_0(6))$, this resulted in the true parameter value

$$\psi_0 \approx (0.232, 0.335, 0.390, 0.428, 0.460, 0.489). \tag{13}$$

Failing to adjust for confounders would result in underestimation of $\psi_0(t^*)$ for low values of t^* and overestimation for high values of t^* .

5.3 Positivity

Under the specified distribution P_0 , the degree of practical positivity violations increases with t^* . Figure 1 shows the marginal densities of $g_{0,0:t^*-1}^{\bar{a}}: t^* = 1, 2, \dots, 6$ for each final time point t^* under $\bar{a}(t^* - 1) = 0$, allowing us to obtain a sense of the severity of the violations. Because $g_{0,0:t^*-1}^{\bar{a}}: t^* = 1, 2, \dots, 6$ are functions of $\bar{L}(t^* - 1)$, the marginal densities were derived by taking the conditional probabilities marginally over the distributions of $\bar{L}(t^* - 1)$ for each final time point t^* . The density plot for time $t^* = 1$ shows a somewhat uniform distribution for $g_{0,0:0}^{\bar{a}=0}$, with only 4% of the marginal distribution below 0.01. As t^* increases, however, the cumulative probability of remaining unenrolled, i.e. having $\bar{a}(t^* - 1) = 0$ decreases significantly. For example, 40% of the marginal distribution $g_{0,0:5}^{\bar{a}=0}$ for time $t^* = 6$ was below 0.01. The resulting distributions become increasingly concentrated close to 0, indicating that the probability of remaining unenrolled is near 0 at later time points. Indeed, we saw in the realized simulations that on average, only 6% of observations were still unenrolled and at risk of failure at $t = 5$.

Following previous suggestions [17, 53], estimates of $g_{0,0:t^*-1}(\bar{a}(t^*))$ were truncated at 0.001 before being used in the estimators below.

We provide summaries below of how each estimator presented above is implemented in our simulation. For ease of presentation, we describe estimation of our target parameter for the final time point t^* of interest, i.e. $K + 1 = 6$.

5.4 Estimator implementation

5.4.1 ICE—We first condition our data on survival up to the penultimate time point $t = 5$, as well as optionally conditioning on the subset of subjects having $\bar{A}(K) = \bar{a}(K)$ (depending on whether we are using the stratified versus pooled approach described in Section 4.7), where as stated in Section 5.1.1, $\bar{a}(K) = 0$. With this subset, we carry out a logistic regression, regressing $Y(6)$ onto $\bar{A}(K)$ and $\bar{L}(K)$ under a correctly specified logistic model implied by the data generating distribution.

This logistic regression provides us with estimates $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)$, resulting in a fitted object $\bar{Q}_{n,L(K+1)}$ for time $t = 6$. With this fit, we estimate $\bar{Q}_{0,L(K+1)}^{\bar{a}}$ for everyone, where the conditional expectation is known to be equal to 1 if $T_i < 6$. We iterate this by now conditioning on survival up to time $t = 4$, using the estimates $\bar{Q}_{n,L(K+1),i}^{\bar{a}}$ as the outcome, and regressing it onto $\bar{A}(K-1)$ and $\bar{L}(K-1)$ using the same logistic model but replacing the predictors for time $t = 5$ with those of time $K - 1$. It is important to note that the use of the same logistic model at times $t < 6$ is technically incorrectly specified, as the conditional expectation is more complex than this. To reflect this, we refer to this specification as the correctly[†] specified model.

This gives us a fitted object $\bar{Q}_{n,L(K)}$ for time $t = 5$, which we use to estimate $\bar{Q}_{0,L(K)}^{\bar{a}}$ for each subject i by setting $A1(K-1) = 0$ for everyone and again calculating estimates under the fitted object, again setting it equal to 1 if $T_i < 5$. We continue iterating these steps backwards over t until we reach time $t = 1$, at which we will have the estimates $\bar{Q}_{n,L(1)}^{\bar{a}}$ as a function of only $L(0)$. The target parameter is then estimated by taking the empirical mean over the sample, $\hat{\psi}_n^{ICE} = \frac{1}{n} \sum_{i=1}^n \bar{Q}_{n,L(1),i}^{\bar{a}}$.

We also considered the performance of this estimator using the same algorithm above, but under the following mis-specified logistic model (acknowledging a slight abuse of notation):

$$\text{logit } \mathbb{E}[Y(6) | \bar{L}(5), \bar{A}1(5)] = \beta_0 + \beta_1 W_2 + \beta_2 L1_1(5) + \beta_3 A1(5) \quad (14)$$

5.4.2 IPW—At each time point $t = 0, 1, \dots, 5$, we condition our data on survival and having not enrolled yet, i.e. $\bar{Y}(t) = 0 \cap \bar{A}1(t-1) = 0$, and estimate g_0 by using maximum likelihood to fit a logistic regression of our treatment variable $A1(t)$ on $\bar{L}(t): t = 0, 1, \dots, K$ under a correctly specified logistic model $g_{A(t)}$ for each time point t . We separately used a mis-specified model in which we omitted $W_2, L1_2(t)$ and the interaction term $L1_1(t)L1_2(t)$.

We evaluate these fits at $A1(t) = 0$ for each time point t and subject i and take the cumulative product over all time points up to time K . The resulting subject specific estimates of following the longitudinal intervention “never enroll” ($g_{n,0:K,i}^{\bar{a}}$) are used in eq. (10) to form the estimate $\hat{\psi}_n^{IPW}$, with $Y_i = Y_i(K + 1)$.

5.4.3 AIPW—In implementing this estimator, g_0 was estimated as for the IPW estimator, using the same correctly[†] specified and misspecified parametric models. In contrast to the IPW estimator, which only requires an estimate of each subject’s cumulative probability of never enrolling through time (i.e. $g_{n,0:K}^{\bar{a}=0}$), for the AIPW estimator, we compute $g_{n,0:t}^{\bar{a}}$ for $t = 0, 1, \dots, K$, i.e. each subject i ’s predicted cumulative probability of not enrolling up to each time point t .

We first condition our data on survival up to the penultimate time point $t = 5$, as well as optionally conditioning on the subset of subjects having $\bar{A}(K) = \bar{a}(K)$ (depending on whether we are using the stratified versus pooled approach, as described in Section 4.7), and where as stated in Section 5.1.1, $\bar{a}(K) = 0$. We use this subset to estimate $\bar{Q}_{0,L(K+1)}^{\bar{a}}$ by carrying out a logistic regression of $Y(K + 1)$ on $\bar{L}(K)$. Both the correctly[†] specified model and the misspecified model from eq. (14) were considered (where under the stratified approach the coefficient on the $A1(t)$ covariate in the logistic model is absorbed into the intercept). The resulting logistic regression fit $\bar{Q}_{n,L(K+1)}^{\bar{a}}$ is then evaluated for each subject i in the study, setting the predicted value equal to 1 if $T_i < 6$. This allows us to evaluate the functional

$$D_{K+1}^*(\bar{Q}_n^{\bar{a}}, g_n)(O_i) = \frac{\mathbb{I}(\bar{A}_i(K) = \bar{a}(K))}{g_{n,0:K,i}^{\bar{a}}} (Y_i(K+1) - \bar{Q}_{n,L(K+1),i}^{\bar{a}}); i = 1, 2, \dots, n.$$

We iterate this by now conditioning on survival up to time $t = 4$ and on the subset of subjects having $\bar{A}(K - 1) = \bar{a}(K - 1)$, using the estimates $\bar{Q}_{n,L(K+1),i}^{\bar{a}}$ as the outcome, and regressing this outcome onto $\bar{L}(K - 1)$ using the correctly[†] specified or misspecified logistic model but replacing the predictors for time 5 with those of time $K - 1$. This gives us a fitted object $\bar{Q}_{n,L(K)}^{\bar{a}}$ for time 5, which we evaluate for each subject i setting the predicted value equal to 1 if $T_i < 5$. The fit is then used to evaluate

$$D_K^*(\bar{Q}_n^{\bar{a}}, g_n)(O_i) = \frac{\mathbb{I}(\bar{A}_i(K-1) = \bar{a}(K-1))}{g_{n,0:K-1,i}^{\bar{a}}} (\bar{Q}_{n,L(K+1),i}^{\bar{a}} - \bar{Q}_{n,L(K),i}^{\bar{a}}); i = 1, 2, \dots, n.$$

This procedure is iterated until all of $D_t^*(\bar{Q}_n^{\bar{a}}, g_n)(O_i); t = 1, \dots, K + 1$ are evaluated and $\bar{Q}_{n,L(1),i}^{\bar{a}}; i = 1, 2, \dots, n$ is estimated. The target parameter is then estimated by simply taking the empirical mean of the sum of $\bar{Q}_{n,L(1),i}^{\bar{a}}$ and $D_t^*(\bar{Q}_n^{\bar{a}}, g_n)(O_i)$ over t , i.e.

$$\hat{\psi}_n^{AIPW} = \frac{1}{n} \sum_{i=1}^n \left(\bar{Q}_{n, L(1), i}^{\bar{a}} + \sum_{i=1}^{K+1} D_i^* (\bar{Q}_n^{\bar{a}}, g_n)(O_i) \right).$$

5.4.4 DRICE—This estimator is very similar to the ICE approach, with the added step of using the inverse estimates $g_{n,0:t-1}^{\bar{a}, -1}$ times the indicator $\mathbb{I}(\bar{A}_i(t-1) = \bar{a}(t-1))$ as an additional predictor in estimating $\bar{Q}_{0, L(t)}^{\bar{a}}; t = 1, 2, \dots, K+1$. Similar to Section 5.1.5, We first compute $g_{n,0:t-1}^{\bar{a}}$ for $t = 1, 2, \dots, K+1$.

We then take the subset of subjects conditioned on survival up to the penultimate time point $t = 5$; if using the stratified approach we further conditioned on having $\bar{A}(K) = \bar{a}(K)$. With this subset, we carry out a logistic regression, regressing $Y(K+1)$ on $\bar{A}(K)$, $\bar{L}(K)$, and $\mathbb{I}(\bar{A}(K) = \bar{a}(K)) / g_{n,0:K}^{\bar{a}}$ such that the correctly[†] specified and misspecified models respectively are

$$\begin{aligned} \text{logit } E[Y(6) | \bar{L}(5), \bar{A}(5)] &= \beta_0 + \beta_1 W_1 + \beta_2 W_2 + \beta_3 L1_1(5) + \beta_4 L1_2(5) \\ &\quad + \beta_5 L1_1(5) L1_2(5) + \beta_6 A1(5) + \beta_7 \frac{\mathbb{I}(\bar{A}(5) = \bar{a}(5))}{g_{n,0:5}^{\bar{a}}} \end{aligned} \tag{15}$$

$$\begin{aligned} \text{logit } E[Y(6) | \bar{L}(5), \bar{A}(5)] &= \beta_0 + \beta_1 W_2 + \beta_2 L1_1(5) + \beta_3 A1(5) \\ &\quad + \beta_4 \frac{\mathbb{I}(\bar{A}(5) = \bar{a}(5))}{g_{n,0:5}^{\bar{a}}}, \end{aligned}$$

noting that in the stratified approach the coefficient on $A1(K)$ is simply absorbed into the intercept. Following this approach of fitting \bar{Q}_0 , the remaining steps are implemented identically to the ICE approach in Section 5.1.3 resulting in the estimate $\hat{\psi}_n^{DRICE(cov)}$.

We noticed a small proportion of the time that convergence was not achieved (at level 10^{-4}). This occurrence increased with the presence of practical positivity issues. To ensure that resulting estimates would not be biased, we therefore relied upon a customized optimization function which would directly solve the EIF in this situation. This function is included in the R package referred to below (uploaded to github).

5.4.5 DRICE weighted—As discussed in Section 4.5.1, it is also possible to form a DR estimator by instead using the inverse probability estimates $g_{n,0:t-1}^{\bar{a}, -1}$ as observational weights. This approach is similar to the DRICE approach above, with the lone exception that the inverse probability estimates $g_{n,0:t-1}^{\bar{a}, -1}$ are used as observational weights in the logistic regressions, leaving the indicator of treatment $\mathbb{I}(\bar{A}_i(t-1) = \bar{a}(t-1))$ as the model covariate. We denote estimates under this approach as $\hat{\psi}_n^{DRICE(wt)}$.

5.4.6 TMLE—This estimator differs from the DRICE approach in that at each sequential regression step, $\bar{Q}_{0,L(t)}^{\bar{a}}$ is first estimated without use of the estimated inverse probability, and then this initial fit is updated in a second regression step, using the initial fit as offset, and including the identical term $\mathbb{1}(\bar{A}_t(t) = \bar{a}(t))$ as a single covariate. As noted in Section 4.5.1, this modification facilitates the use of machine learning approaches for \bar{Q}_0 , implemented in the subsequent section.

As with the DRICE approach, we first take the subset of the data conditioned on survival up to the penultimate time point $t = 5$ (as well as optionally, $\bar{A}(K) = \bar{a}(K)$). With this subset, we carry out a logistic regression, regressing $Y(K + 1)$ onto $\bar{A}(K)$, $\bar{L}(K)$. Both the correctly[†] specified and misspecified models were considered. The transformed fit $\text{logit } \bar{Q}_{n,L(K+1)}^{\bar{a}}$ is then used as an offset in a univariate logistic regression with no intercept and the covariate $h(K) = \mathbb{1}(\bar{A}(K) = \bar{a}(K))/g_{n,0:K}^{\bar{a}}$ to form the following parametric submodel

$$\text{logit } \bar{Q}_{n,L(t)}^{\bar{a}}(\epsilon_t) = \text{logit } \bar{Q}_{n,L(t)}^{\bar{a}} + \epsilon_t h(t) \tag{16}$$

where $t = K + 1$. The sole parameter ϵ_{K+1} is estimated using maximum likelihood estimation, i.e. the negative log likelihood loss function, resulting in the estimate $\hat{\epsilon}_{n,K+1}$.

This allows the initial fit $\bar{Q}_{n,L(K+1)}^{\bar{a}}$ to be updated $\bar{Q}_{n,L(K+1)}^{\bar{a}}(\hat{\epsilon}_{n,K+1})$, which we denote $\bar{Q}_{n,L(K+1)}^{\bar{a},*}$. This updated fit is evaluated for each subject i in the study, setting equal to 1 if $T_i < K + 1$. We iterate by now conditioning on survival up to time $t = 4$, using the estimates $\bar{Q}_{n,L(K+1),i}^{\bar{a},*}$ as the outcome, and regressing it onto $\bar{A}(K - 1)$ and $\bar{L}(K - 1)$ using the logistic model but replacing the predictors for time K with those of time $K - 1$. This gives us an initial fitted object $\bar{Q}_{n,L(K)}^{\bar{a}}$ for time $t = 5$, which we again update using the parametric submodel specified in eq. 16, with $t = K$ covariate

$h(K - 1) = \mathbb{1}(\bar{A}(K - 1) = \bar{a}(K - 1))/g_{n,0:K-1}^{\bar{a}}$. The updated fit is evaluated for each subject i in the study setting equal to 1 if $T_i < 5$, and the procedure is iterated backwards over t until we reach time $t = 1$, at which we will have the estimate $\bar{Q}_{n,L(1)}^{\bar{a},*}$ as a function of only $L(0)$. The target parameter is then estimated by taking the empirical mean over the sample,

$\hat{\psi}_n^{TMLE(cov)} = \frac{1}{n} \sum_{i=1}^n \bar{Q}_{n,L(1),i}^{\bar{a},*}$. Similar to DRICE, we also noticed some convergence issues with this estimator as practical positivity issues increased and therefore implemented our own customized optimization function to directly solve the EIF in this situation.

5.4.7 TMLE weighted—We also implemented the weighted version of TMLE which uses the covariate $h_t = 1$ and observational weight $\mathbb{1}(\bar{A}(t) = \bar{a}(t))/g_{n,0:t}^{\bar{a}}$ as discussed in Section 4.6.1. We denote estimates under this approach as $\hat{\psi}_n^{TMLE(wt)}$.

5.4.8 R-packages—While TMLE is potentially complex in its execution, we note that an ltmle R-package has been developed for this estimator [8, 54, 55] and uploaded to The

Comprehensive R Archive Network (CRAN). Additionally, estimates using the ICE and IPW estimators can be obtained using this package. A built in option allows the user to either stratify or pool the subjects when estimating $\bar{Q}_{0,L(t)}^{\bar{a}}$ and $g_{n,0:t-1}^{\bar{a}}$ for $t = 1, 2, \dots, K+1$. Furthermore, it can conduct the estimations using either a parametric generalized linear model or the Super Learner estimation approach discussed in Section 4.8. We used version 0.9–6 of this package for this study.

We additionally developed a new `lrecCompare` R-package which was designed specifically for all of the remaining analyses for the current study. This package contains all of the functions to generate the simulation data, as well as code to perform the AIPW and DRICE computations. Similar to the `ltmle` R-package, a built in option allows the user to either stratify or pool the subjects and computations can be conducted using either a parametric generalized linear model or Super Learner. A further option allows the user to use the modified version of the DRICE estimator. This package has been uploaded to an online public repository at <http://www.github.com/tranlm/lrecCompare>.

5.4.9 Performance—Estimator performance was evaluated by comparing the bias, variance, and mean squared error of each estimator across 1000 iterations. While the sample variance of the estimated IF/ n can provide a straightforward variance estimator for most of the estimators considered (with the exception of the ICE estimator), IF based variance estimation has been shown in past work to result in anti-conservative confidence intervals in settings, such as the one deliberately studied here, with practical positivity violations (e.g. Petersen et al. [8]). While the comparative evaluation of variance estimators is an exciting area in its own right, we focused here on performance of estimators of the target parameter ψ_0 . We therefore omitted evaluation of the coverage of the estimators for the simulations.

5.5 Results

The mean squared error of each estimator is presented in Figure 2. Further results are also presented in Table 1 and Table 2 in Appendix B. for the interested reader. We note again that estimates of $g_{0,0:t^*-1}^{\bar{a}(t^*)}$ were truncated at 0.001.

5.5.1 No positivity violations—At the first time point, as predicted by theory, when nuisance parameters were estimated using correctly specified parametric models, all estimators were without meaningful bias. MSE was also similar across estimators, although the MSE of IPW and AIPW were slightly higher. No meaningful differences were observed when nuisance parameters were fit stratifying vs. pooling on treatment regime.

The non-DR estimators exhibited well-understood susceptibility to misspecification of nuisance parameter models. Recall that the misspecified g_0 and $\bar{Q}_0^{\bar{a}}$ models omit the W_1 and L_{12} variables. The bias of the ICE estimator increased substantially when $\bar{Q}_0^{\bar{a}}$ was estimated using a misspecified model, while the bias of the IPW estimator increased substantially when g_0 was estimated using a misspecified model.

As predicted by theory, all 5 DR estimators remained without meaningful bias as long as at least one of the two nuisance parameters, g_0 and \bar{Q}_0^a , were estimated with a correctly specified model, and all 5 exhibited lower bias than the IPW and ICE estimators under comparable misspecification. When both g_0 and \bar{Q}_0^a were estimated using misspecified models, the bias of the DR estimators remained better than that of the IPW estimator under misspecification, but worse than the ICE estimator. Weighted estimators tended to show better performance than the covariate estimators for both DRICE and TMLE. Pooled estimation of the nuisance parameters also showed better performance, particularly when using misspecified models.

5.5.2 Performance under increasing positivity violations—As the extent of positivity violation increased with increasing time points, the MSE and bias of all estimators considered became worse. As positivity violations increased, the ICE estimator maintained the lowest MSE of all estimators under correct[†] specification for g_0 and \bar{Q}_0^a , under both stratified and pooled approaches, with the expected susceptibility to bias from misspecification of the model used to estimate \bar{Q}_0^a .

As has been observed previously [14] the IPW estimator was particularly susceptible to increasing positivity violation. Under correct[†] specification for estimators of g_0 and \bar{Q}_0^a , the MSE of the IPW estimator at time point 6 exhibited the highest bias and among the highest MSE of all estimators considered. The MSE of the IPW estimator at time point 6 was 15 fold higher than at time point 1 under the correctly[†] specified model.

The covariate-based DR estimators also proved highly susceptible to positivity violations. When stratifying by treatment regime followed, under correct[†] specification, the covariate DRICE estimator at time point 6 had a bias almost as high as and an MSE even higher than IPW. The covariate-based TMLE had lower bias and MSE than the corresponding covariate-based DRICE, but still underperformed alternative DR estimators. Bias and MSE of the covariate-based DR estimators decreased, and the differences between the two estimators were less pronounced when estimators were implemented pooling across treatment regimes followed; however, both covariate-based DR estimators continued to have higher bias and lower MSE than alternative DR estimators.

In particular, despite being substitution estimators, and thus expected to exhibit improved robustness to data sparsity, when substantial positivity violations were present, both covariate based DR estimators exhibited higher bias and MSE than the estimating equation-based AIPW estimator. Importantly, however, a proportion of AIPW estimates were outside the parameter boundary of $[0,1]$, and this proportion increased with time and increasing positivity violation. For example, at $t = 1$ between 0.1% to 0.05% of the estimates were outside $[0,1]$, while this increased to as much as 2.1% at $t = 6$.

The weighted DR estimators consistently outperformed all other DR estimators in terms of both bias and MSE, whether or not stratification on treatment regime was used, and both when g_0 and \bar{Q}_0^a were correctly[†] specified as well as across most misspecification scenarios.

As expected given that these two estimators do not differ meaningfully, minimal differences were seen between the performance of the two weighted DR estimators. The improved robustness to positivity violations conferred by using a weighted vs covariate approach, demonstrated here in the longitudinal setting, echo earlier observations in the point treatment setting by Kang and Schafer [18] and Robins et al. [19].

6 Application: Impact of a task shifting program on the retention and mortality of HIV patients in Eastern Africa

We evaluate an HIV care program designed to shift patient care tasks to different clinic personnel for patients considered at low-risk, which has been described in detail previously [56]. To summarize, this task shifting program was implemented in 15 HIV clinics in Kenya between 2007 and 2009 from the Academic Model Providing Access to Healthcare (AMPATH). As part of the program, 75% of patient care tasks were shifted from doctors and clinical officers to nurses for patients deemed low risk. Once eligible, program enrolment occurred at the treating clinician's discretion, resulting in time-dependent confounding as patients who became sicker over the course of follow up were less likely to be enrolled. Patients enrolled in this low-risk express care (LREC) program subsequently had their care shifted for all subsequent clinical visits.

To examine all of the estimators considered, we focus on estimating the counterfactual probability of in-care survival (remaining alive and not lost to follow up) over time under a single hypothetical intervention to "never enroll in the LREC program".

Formally, our cohort data are defined using the same data structure defined in Section 2. The baseline time point, $t = 0$, is the first time point at which a patient is eligible for LREC enrollment after the LREC program has been initiated at the clinic in which the patient receives care. We discretize follow-up time into 90-day intervals. Let $Y(t)$ be an indicator of failure by the end of interval t (defined as an observed mortality or loss to follow up). Let $L1(t)$ consist of the following time-varying covariates: CD4 count, clinic type, calendar date, WHO HIV stage, ART use and adherence, pregnancy, and whether the subject had a clinical visit during the interval (noting that all patients had a visit during $t = 1$). We define $L1(0)$ to additionally include the following non-time varying covariates: age at baseline, gender, use of a second-line protease inhibitor based regimen, tuberculosis treatment, and clinic type. Let $A1(t)$ be an indicator of enrollment into the LREC program, $A2_1(t)$ be an indicator of censoring due to clinic transfers, and $A2_2(t)$ be an indicator of administrative censoring due to database closure. We refer to the outcome and covariate process as $L(t) = (Y(t), L1(t))$ and the treatment and censoring processes as $A(t) = (A1(t), A2_1(t), A2_2(t))$. Our cohort data are $O = (O_1, O_2, \dots, O_{15,225})$ for the 15,225 subjects in our data set, where for each subject i , we have

$$O_i = (L_i(0), A_i(0), L_i(1), A_i(1), \dots, A_i(K), L_i(K + 1)). \quad (17)$$

We assume here that the data are independent and identically distributed, such that for each subject i we have that $O_i \stackrel{iid}{\sim} P_0$. One approach to addressing potential dependence between

subjects due to shared clinic level characteristics would be to include clinic as an adjustment variable (a clinic fixed effect approach); here, we instead adjusted for clinic characteristics, which entails stronger assumptions. In addition, causal effects in this example may be subject to “interference”, in which the effect of an individual’s enrollment in the program depends on the number of other individuals in the same clinic concurrently enrolled; this issue, together with potential responses, is explored in detail in Miles, Petersen, and van der Laan [57].

Within the 15,225 subjects in our cohort, 2,011 immediately enrolled into the program at the start of followup (i.e. within the first 90 day interval). During follow-up, subjects continued to enroll into the LREC program. By 21 months from initial eligibility, 1,819 subjects were still alive and remained unenrolled. The patients cumulatively contributed a total of 229,941 person-months (Interquartile range: 6,18) of follow-up to the analyses; 1,440 experienced either loss-to-follow-up (1,362) or death (78) by 24 months, while 140 were censored due to clinic transfers. Subjects for whom the database closed less than 24 months after eligibility were administratively censored. We imputed values for missing adjustment covariates by using either their last observation carried forward (for time-varying covariates with prior measures) or by using the median of the value across all subjects. This did not noticeably change any of the overall summary statistics for the covariates when compared to a complete case analysis.

6.1 Target parameter

We are interested in estimating the probability of remaining alive and in-care over time under a hypothetical intervention to prevent enrollment at all time points ($\bar{a}1 = 0$) and enforce no censoring ($\bar{a}2 = 0$). Under the counterfactual intervention we estimate 1 minus the counterfactual probability of failure by time t^* , $\psi_{\bar{a}}(t^*) = \mathbb{E}_{P_{\bar{a}}^0}[Y_{\bar{a}}(t^*)]$ for each time point $t^* = 1, 2, \dots, 7$, where $t^* = 7$ corresponds to the 24th month of follow-up.

6.2 Estimation results

We first used parametric generalized linearmodels (GLMs) to estimate the g_0 and \bar{Q}_0 portions of the likelihood. Fits for g_0 were formed by pooling observations over the 7 time points, as opposed to stratifying by each time point and fitting separate models (as was done in the simulations). This aided us in estimating the clinic transfer mechanism, as the number of subjects transferring clinics was low ($n = 140$). Figure 3 shows the estimated marginal densities of $g_{n,0;t^*-1}^{\bar{a}} : t^* = 1, 2, \dots, 7$ at each time point t^* taken marginally over the 15,225 subjects. While the trend of shifting to lower probabilities over time is similar to that seen in the simulated data (Figure 1), the degree of positivity violations is less extreme.

Figure 4 shows the results of applying each estimator to our data using the GLM fits. 95% confidence intervals were calculated using standard error estimates based on each estimator’s estimated influence function (for the ICE estimator, the EIF was used for comparative purposes only, noting that it is not theoretically valid). At early time points, all approaches yielded very similar estimates. Most estimators remained in close agreement

over later time points, with the important exception of IPW, which deviated markedly as positivity violations increased. Confidence interval width also increased with time, consistent with decreasing data support. Pooling observations across treatment regimes versus stratifying by regime when fitting the nuisance parameters did not affect the estimates noticeably. The estimators all (within rounding errors) respected the monotonicity of the cumulative failure distribution over time.

As stated in Section 4.8, Super Learner was also used to estimate the conditional probability g_0 and conditional expectations \bar{Q}_0 . The library of potential candidates used here consisted of: a generalized linear model, Bayesian generalized linear model, multivariate adaptive regression spline, gradient boosting machine, support vector machine, neural network, LASSO, ridge regression, and a stepwise selected model using the Akaike information criterion. We took the linear combination which minimized the cross-validated (10 folds) non-negative binomial likelihood risk.

Figure 5 shows the estimated density for $g_{n,0:t^*-1}^{\bar{a}}$ for time point $t^* = 1, 2, \dots, 7$ using Super Learner. A comparison of the GLM and Super Learner plots implies that the fits between the two approaches are very similar, with Super Learner generally resulting in more concentrated distributions. Correlations in the estimated cumulative probabilities at each time point t^* resulting from the two approaches varied from 0.56 to 0.97, with higher correlations toward the later time points. Estimates of probabilities tended to be higher for $g_{n,0:t^*-1}^{\bar{a}=0}$ fits using Super Learner. Figure 6 shows the resulting estimates of our target parameter, with influence curve based 95% confidence intervals. We excluded DRICE here, as a number of the candidates we used in the Super Learner library were non-linear. While the estimates are similar to the approach using GLMs, a number of differences are seen. Firstly, the similarity of the estimates across the set of estimators (within each time point) is reduced. A noticeable reduction is seen in estimates from TMLE with inverse probability estimate as covariate at the later time points. Monotonicity is still preserved (within rounding errors). Confidence intervals were generally larger than those estimated using GLM fits.

7 Discussion

Numerous approaches have been proposed for estimating causal parameters in a longitudinal treatment setting, several of which have the attractive theoretical properties of double robustness and semi-parametric efficiency. However, there are few published direct comparisons of the relative performance of the multiple DR estimators and in particular, the various DR estimators based on the DRICE representation of the EIF. In this article, we presented and analyzed seven specific estimators, including 5 double robust estimators, in a longitudinal treatment setting using both simulated and real world data. We evaluated performance under model mis-specifications and steadily increasing levels of positivity violations for our simulated data. For the applied setting, we considered both a parametric and data-adaptive approach when estimating the treatment and censoring mechanism and the series of iterated outcome regressions. Simulation results showed decreasing estimator

performance with model mis-specifications or increasing positivity violations. IPW was the most severely affected estimator. Weighted DR estimators performed better than covariate-based DR estimators. Pooling observations tended to result in lower MSE, as well as reduced bias when at least one of the two nuisance parameters was estimated consistently. As stated in Section 4.7, this is likely due to the larger sample size available for estimating the nuisance parameters.

In the applied setting, similar estimates were observed across estimators when using GLM fits of the two nuisance parameters. The use of Super Learner to estimate g_0 and \bar{Q}_0 resulted estimates similar to the GLM approach, though with more variability between estimators and slightly wider confidence intervals. As seen in the simulations, integrating the inverse probability estimates as observational weights in the two DR substitution based estimators appeared to result in improved performance.

The estimators we evaluated can be grouped into two classes: estimating equation-based estimators (IPW, AIPW) and substitution-based estimators (ICE, DRICE, TMLE). In contrast to substitution-based estimators, estimating equation-based estimators do not always obey the constraints of the parameter space. Consequently, the use of these methods can, as seen here, lead to estimates and confidence intervals that are outside of the $[0,1]$ range for our parameter. In our simulations, we saw that up to 2% of the estimates from AIPW estimator were outside that range. Substitution estimators are defined as mappings applied to probability distributions and consequently will never result in values beyond the bounds of the parameter space.

There is also a debate regarding whether DR estimators really have an advantage, as in practice, both the outcome and treatment models are most likely mis-specified. Our simulations here and those from Kang and Schafer [18] in the point treatment setting show that when both are wrongly specified, the ICE approach can outperform the DR estimators in terms of MSE. However, in using machine-learning approaches to estimate both nuisance parameters, we largely side step this concern and gain several important advantages. By using machine learning approaches, which respect that our true statistical model is semi-parametric, we can support consistency of our nuisance parameters estimators, providing a basis for accurate inference as well as estimator efficiency. This further translates into finite sample gains in both bias and variance. Furthermore, in settings such as randomized controlled trials where g_0 is known, the DR estimators are guaranteed to be consistent and their use will only improve efficiency.

Our intention in comparing the various estimators was to provide a sense of the performance of these estimators in practice. Given its efficiency, ability to respect the parameter space, and observed performance, we recommend the pooled and weighted TMLE approach as the preferred estimator. This approach is essentially an identical estimator to the weighted DRICE estimator using parametric models for the iterated outcome regressions; however, it straightforwardly incorporates more flexible machine learning approaches to nuisance parameter estimation.

We end by stating that the approaches presented here can easily be generalized to a number of various other estimation problems. An example includes the estimation of causal effects of dynamic treatment regimes, where treatment decisions are functions of the time-dependent covariate process. Extensions involving marginal structural models may also be applied in summarizing the treatment effect over multiple time points, levels of treatments, or treatment effects that are also functions of baseline covariates [8, 58].

Acknowledgments

Funding

This work was supported by Doris Duke Clinical Scientist Development Award (NIH-NIAID U01AI069911), The National Institute of Allergy and Infectious Diseases of the National Institutes of Health (U01AI069911), The President's Emergency Plan for AIDS Relief (PEPFAR) (AID-623-A-12-0001), NIH (R01AI074345).

Appendix

A: Notation list

Variable	Description
Observed data	
$L1(t)$	covariates at time t
$Y(t)$	outcome at time t
$L(t) = (L1(t), Y(t))$	non-intervention nodes at time t
$A1(t)$	treatment at time t
$A2(t)$	indicator of right-censoring by time t
$A(t) = (A1(t), A2(t))$	intervention nodes at time t
$\bar{A}(k) = (A(0), \dots, A(k))$	history of intervention nodes $t = 0, \dots, k$
$\bar{L}(k) = (L(0), \dots, L(k))$	history of non-intervention nodes $t = 0, \dots, k$
$\underline{A}(k) = A(k+1), \dots, A(K+1)$	future of intervention nodes $t = k+1, \dots, K+1$
$\underline{L}(k) = L(k+1), \dots, L(K+1)$	future of non-intervention nodes $t = k+1, \dots, K+1$
$O = (L(0), A(0), \dots, L(K), A(K), L(K+1))$	observed data structure
$Pa(L(k)) = \bar{L}(k-1), \bar{A}(k-1)$	parents of non-intervention nodes $L(k)$
$Pa(A(k)) \subseteq \bar{L}(k), \bar{A}(k-1)$	parents of intervention nodes $A(k)$
Counterfactuals	
\mathcal{A}_t	set of interventions of interest to set $A(t)$ for $t = (0, \dots, t)$
$\bar{a}(t) \in \mathcal{A}_t$	specific static intervention to set $\bar{A}(t)$
$L_{\bar{a}}(t)$	Counterfactual $L(t)$ under intervention $\bar{a}(t-1)$
$Y_{\bar{a}}(t)$	counterfactual $Y(t)$ under intervention $\bar{a}(t-1)$
Distributions	
P_0	distribution of O
$P_{\bar{a}}^0$	distribution of counterfactual non-intervention nodes $\bar{L}_{\bar{a}}(K+1)$

Variable	Description
$q_0, L(t)$	distribution of $L(t) Pa(L(t))$
$q_0 \equiv \prod_{t=0}^{K+1} q_{0, L(t)}(L(t) Pa(L(t)))$	non-intervention factor of the likelihood
$g_0, A(k)$	distribution of $A(k) Pa(A(k))$
$g_0 \equiv \prod_{t=0}^K g_{0, A(t)}(A(t) Pa(A(t)))$	non-intervention factor of the likelihood
Models	
\mathcal{M}^F	Structural Causal Model: set of possible distributions for (O, U) (where U is unmeasured exogenous variables)
\mathcal{M}	statistical model for P_0 defines statistical assumptions: $P_0 \in \mathcal{M}$
\mathcal{G}	statistical model for g_0 defines statistical assumptions: $g_0 \in \mathcal{G}$
\mathcal{Q}	statistical model for q_0 defines statistical assumptions: $q_0 \in \mathcal{Q}$
Target Parameters	
$E[Y_{\bar{a}}(t^*)]$	casual target parameter, equal to the counterfactual mean outcome at some time t^*
$\Psi: \mathcal{M} \rightarrow \mathbb{R}$	statistical target parameter mapping
$\Psi(P_0) = \psi_0$	statistical target parameter value
Efficient influence function	
$D^*(P)(O) = D^*(q, g)(O)$	efficient influence curve of Ψ at P
Nuisance parameters	
$g_{0,t}^{\bar{a}} \equiv P_0(A(t) = a(t) \bar{L}(t), \bar{A}(t-1) = \bar{a}(t-1))$	time t -specific component of g_0 evaluated at $\bar{a}(t)$
$g_{0,0:k}^{\bar{a}} \equiv \prod_{t=0}^k g_{0,t}^{\bar{a}}$	cumulative probability of treatment up to time k
$\bar{Q}_{0, L(K+2)}^{\bar{a}} \equiv Y(K+1)$	alternative notation for final outcome
$\bar{Q}_{0, L(t)}^{\bar{a}}$	iterated conditional outcome regression for time t i.e. $E_0[\bar{Q}_{0, L(t+1)}^{\bar{a}} \bar{L}(t-1), \bar{A}(t-1) = \bar{a}(t-1)]$
$\bar{Q}_0^{\bar{a}} \equiv \bar{Q}_{0, L(t)}^{\bar{a}}, t = 1, \dots, K+1$	set of iterated outcome regression, $\Psi(P_0) = \Psi(\bar{Q}_0^{\bar{a}})$
Least favorable submodels	
$\{\bar{Q}_{L(t)}^{\bar{a}}(\epsilon_t); \epsilon_t\}$	submodel through $\bar{Q}_{L(t)}^{\bar{a}}$ at parameter value $\epsilon_t = 0$ used to update initial estimate of $\bar{Q}_{0, L(t)}^{\bar{a}}$ in TMLE
Loss functions	
$\mathcal{L}_{t, \bar{Q}_{L(t+1)}^{\bar{a}}}(\bar{Q}_{L(t)}^{\bar{a}})$	loss function for $\bar{Q}_{L(t)}^{\bar{a}}$ relies on estimator of previous $\bar{Q}_{L(t+1)}^{\bar{a}}$ used to fit ϵ of least-favorable submodel $\bar{Q}_{L(t)}^{\bar{a}}(\epsilon_t)$

Variable	Description
Estimators	
$\hat{g}_{n,0:k}^{\bar{a}}$	estimator of $g_{0,0:k}^{\bar{a}}$
$\bar{Q}_{n,L(t)}^{\bar{a}}$	initial (non-targeted) estimator of $\bar{Q}_{0,L(t)}^{\bar{a}}$
$\bar{Q}_{n,L(t)}^{\bar{a}}$	targeted estimator of $\bar{Q}_{0,L(t)}^{\bar{a}}$ in the DRICE estimator
$\bar{Q}_{n,L(t)}^{\bar{a},g}$	targeted estimator of $\bar{Q}_{0,L(t)}^{\bar{a}}$ used in TMLE
$\hat{\psi}_n^{ICE}$	iterated conditional expectation estimate of ψ_0
$\hat{\psi}_n^{HT}$	inverse probability weighted estimate of ψ_0
$\hat{\psi}_n^{AIPW}$	augmented inverse probability weighted estimate of ψ_0
$\hat{\psi}_n^{DRICE}$	double robust iterated conditional expectation estimate of ψ_0
$\hat{\psi}_n^{TMLE}$	targeted minimum loss based estimate of ψ_0

B: Simulation results

Table 1:

Simulation results for estimating $\psi_0(\tau^*) = \mathbb{E} \frac{Y(\tau^*)}{P_0^d} : \tau^* = 1, 2, \dots, 6$ stratifying data by treatment followed.

Time (τ^*)	1	2	3	4	5	6
$\psi_0(\tau^*)$	0.232	0.335	0.390	0.428	0.460	0.489
	Bias (Var)	Bias (Var)	Bias (Var)	Bias (Var)	Bias (Var)	Bias (Var)
Unadjusted	0.05 (0.001)	0.06 (0.001)	0.04 (0.001)	0.03 (0.002)	0.01 (0.002)	0.00 (0.003)
	[MSE]	[MSE]	[MSE]	[MSE]	[MSE]	[MSE]
	[0.003]	[0.003]	[0.004]	[0.003]	[0.002]	[0.002]
	[0.001]	[0.001]	[0.001]	[0.003]	[0.002]	[0.002]
Correct \bar{Q}_0	0.00 (0.001)	0.01 (0.002)	0.01 (0.003)	0.01 (0.005)	0.03 (0.007)	0.04 (0.010)
Mis- \bar{Q}_0	0.02 (0.001)	0.01 (0.001)	0.01 (0.001)	0.03 (0.002)	0.05 (0.002)	0.06 (0.003)
	[MSE]	[MSE]	[MSE]	[MSE]	[MSE]	[MSE]
	[0.001]	[0.001]	[0.001]	[0.001]	[0.002]	[0.005]
	[0.001]	[0.001]	[0.001]	[0.001]	[0.002]	[0.005]
	[0.001]	[0.001]	[0.001]	[0.001]	[0.002]	[0.005]
Correct \bar{g}_0	0.00 (0.003)	0.00 (0.006)	0.03 (0.010)	0.07 (0.014)	0.13 (0.017)	0.17 (0.016)
Mis- \bar{g}_0	0.07 (0.001)	0.08 (0.002)	0.08 (0.006)	0.05 (0.012)	0.00 (0.017)	0.05 (0.019)
	[MSE]	[MSE]	[MSE]	[MSE]	[MSE]	[MSE]
	[0.005]	[0.009]	[0.009]	[0.012]	[0.014]	[0.017]
	[0.001]	[0.002]	[0.006]	[0.012]	[0.014]	[0.017]
	[0.001]	[0.002]	[0.006]	[0.012]	[0.014]	[0.017]
Correct	0.00 (0.003)	0.00 (0.004)	0.00 (0.007)	0.01 (0.011)	0.03 (0.019)	0.04 (0.026)
\bar{Q}_0, \bar{g}_0	[MSE]	[MSE]	[MSE]	[MSE]	[MSE]	[MSE]
	[0.003]	[0.004]	[0.004]	[0.007]	[0.011]	[0.019]
	[0.003]	[0.004]	[0.004]	[0.007]	[0.011]	[0.019]
	[0.003]	[0.004]	[0.004]	[0.007]	[0.011]	[0.019]
Mis- \bar{Q}_0	0.00 (0.003)	0.00 (0.007)	0.00 (0.009)	0.00 (0.013)	0.01 (0.020)	0.03 (0.029)
Mis- \bar{g}_0	0.00 (0.001)	0.01 (0.002)	0.01 (0.006)	0.02 (0.014)	0.04 (0.024)	0.05 (0.031)
Mis- \bar{Q}_0, \bar{g}_0	0.04 (0.001)	0.03 (0.002)	0.02 (0.004)	0.01 (0.010)	0.01 (0.018)	0.03 (0.022)
	[MSE]	[MSE]	[MSE]	[MSE]	[MSE]	[MSE]
	[0.001]	[0.003]	[0.003]	[0.004]	[0.010]	[0.018]
	[0.001]	[0.003]	[0.003]	[0.004]	[0.010]	[0.018]
	[0.001]	[0.003]	[0.003]	[0.004]	[0.010]	[0.018]
Correct	0.00 (0.002)	0.01 (0.007)	0.04 (0.015)	0.07 (0.023)	0.11 (0.032)	0.11 (0.039)
\bar{Q}_0, \bar{g}_0	[MSE]	[MSE]	[MSE]	[MSE]	[MSE]	[MSE]
	[0.002]	[0.007]	[0.007]	[0.017]	[0.028]	[0.045]
	[0.002]	[0.007]	[0.007]	[0.017]	[0.028]	[0.045]
	[0.002]	[0.007]	[0.007]	[0.017]	[0.028]	[0.045]
Mis- \bar{Q}_0	0.00 (0.002)	0.02 (0.007)	0.02 (0.016)	0.04 (0.025)	0.07 (0.038)	0.08 (0.045)
Mis- \bar{g}_0	0.00 (0.001)	0.01 (0.005)	0.02 (0.011)	0.05 (0.020)	0.09 (0.028)	0.08 (0.034)
Mis- \bar{Q}_0, \bar{g}_0	0.05 (0.001)	0.05 (0.004)	0.03 (0.010)	0.00 (0.020)	0.03 (0.031)	0.05 (0.037)
	[MSE]	[MSE]	[MSE]	[MSE]	[MSE]	[MSE]
	[0.001]	[0.003]	[0.006]	[0.011]	[0.020]	[0.032]
	[0.001]	[0.003]	[0.006]	[0.011]	[0.020]	[0.032]
	[0.001]	[0.003]	[0.006]	[0.011]	[0.020]	[0.032]
Correct	0.00 (0.002)	0.01 (0.005)	0.02 (0.010)	0.05 (0.016)	0.08 (0.023)	0.09 (0.028)
\bar{Q}_0, \bar{g}_0	[MSE]	[MSE]	[MSE]	[MSE]	[MSE]	[MSE]
	[0.002]	[0.005]	[0.005]	[0.016]	[0.019]	[0.030]
	[0.002]	[0.005]	[0.005]	[0.016]	[0.019]	[0.030]
	[0.002]	[0.005]	[0.005]	[0.016]	[0.019]	[0.030]
Mis- \bar{Q}_0	0.00 (0.002)	0.01 (0.006)	0.02 (0.013)	0.04 (0.021)	0.07 (0.032)	0.08 (0.038)
	[MSE]	[MSE]	[MSE]	[MSE]	[MSE]	[MSE]
	[0.002]	[0.006]	[0.006]	[0.014]	[0.022]	[0.036]
	[0.002]	[0.006]	[0.006]	[0.014]	[0.022]	[0.036]
	[0.002]	[0.006]	[0.006]	[0.014]	[0.022]	[0.036]

Time (t^*)	1	2	3	4	5	6
$\psi_0(t^*)$	0.232	0.335	0.390	0.428	0.460	0.489
	Bias	[MSE]	Bias	[MSE]	Bias	[MSE]
	(Var)	(Var)	(Var)	(Var)	(Var)	(Var)
Mis- ξ_0	0.00 (0.001)	[0.001] 0.01 (0.003)	[0.003] 0.02 (0.008)	[0.008] 0.04 (0.013)	[0.014] 0.08 (0.020)	[0.025] 0.09 (0.024)
Mis- \bar{Q}_0, ξ_0	0.05 (0.001)	[0.003] 0.04 (0.003)	[0.005] 0.02 (0.008)	[0.008] 0.01 (0.016)	[0.016] 0.04 (0.025)	[0.027] 0.06 (0.030)
DRICE (weighted)	Correct	0.00 (0.001)	[0.001] 0.00 (0.003)	[0.003] 0.00 (0.004)	[0.005] 0.01 (0.007)	[0.007] 0.03 (0.011)
	\bar{Q}_0, ξ_0	0.00 (0.002)	[0.002] 0.00 (0.003)	[0.003] 0.00 (0.004)	[0.004] 0.00 (0.005)	[0.005] 0.02 (0.006)
	Mis- ξ_0	0.00 (0.001)	[0.001] 0.00 (0.002)	[0.002] 0.00 (0.004)	[0.004] 0.00 (0.007)	[0.007] 0.03 (0.012)
	Mis- \bar{Q}_0, ξ_0	0.04 (0.001)	[0.003] 0.04 (0.002)	[0.003] 0.02 (0.003)	[0.003] 0.01 (0.004)	[0.004] 0.01 (0.006)
TMLE (weighted)	Correct	0.00 (0.002)	[0.002] 0.00 (0.003)	[0.003] 0.01 (0.005)	[0.005] 0.01 (0.007)	[0.007] 0.03 (0.011)
	\bar{Q}_0, ξ_0	0.00 (0.002)	[0.002] 0.00 (0.003)	[0.003] 0.00 (0.005)	[0.005] 0.00 (0.006)	[0.006] 0.02 (0.008)
	Mis- ξ_0	0.00 (0.001)	[0.001] 0.01 (0.002)	[0.002] 0.01 (0.004)	[0.004] 0.02 (0.007)	[0.008] 0.04 (0.012)
	Mis- \bar{Q}_0, ξ_0	0.04 (0.001)	[0.003] 0.04 (0.002)	[0.003] 0.03 (0.003)	[0.004] 0.01 (0.005)	[0.006] 0.01 (0.008)

Estimates are based on 1000 simulations with a fixed sample size of $n=500$. $\psi_0(t^*)$ represents the correct parameter value at time t^* under the intervention $\bar{a}(K) = 0$.

Misspecified model for ξ_0 was $\xi_0 A(t) = \text{logit}^{-1}(\beta_0 + \beta_1 W_1 + \beta_2 W_3 + \beta_3 L_1(t-1))$. Misspecified model for \bar{Q}_0 was $\bar{Q}_0 L(t) = \text{logit}^{-1}(\beta_0 + \beta_1 W_2 + \beta_2 L_1(t-1) + \beta_3 A(t-1))$.

Stratified and pooled models are identical for IPW.

Table 2: Simulation results for estimating $\psi_0(\tau^*) = \mathbb{E} \frac{-Y(\tau^*)}{P_0^d} : \tau^* = 1, 2, \dots, 6$ pooling data over all treatment regimes.

Time (τ^*)	1	2	3	4	5	6
$\psi_0(\tau^*)$	0.232	0.335	0.390	0.428	0.460	0.489
	Bias (Var)	Bias (Var)	Bias (Var)	Bias (Var)	Bias (Var)	Bias (Var)
Unadjusted	0.05 (0.001)	0.06 (0.001)	0.04 (0.001)	0.03 (0.003)	0.01 (0.002)	0.00 (0.003)
	[MSE]	[MSE]	[MSE]	[MSE]	[MSE]	[MSE]
	[0.003]	[0.003]	[0.004]	[0.003]	[0.002]	[0.002]
	[0.001]	[0.001]	[0.001]	[0.003]	[0.002]	[0.002]
	[0.001]	[0.001]	[0.001]	[0.004]	[0.006]	[0.007]
ICE	Correct \bar{Q}_0	0.00 (0.001)	0.00 (0.001)	0.01 (0.002)	0.00 (0.004)	0.01 (0.007)
	Mis- \bar{Q}_0	0.02 (0.001)	0.01 (0.001)	0.01 (0.001)	0.02 (0.002)	0.06 (0.003)
IPW	Correct \bar{g}_0	0.00 (0.003)	0.00 (0.006)	0.03 (0.010)	0.07 (0.014)	0.13 (0.017)
	Mis- \bar{g}_0	0.07 (0.001)	0.08 (0.002)	0.08 (0.006)	0.05 (0.012)	0.00 (0.017)
AIPW	Correct	0.00 (0.003)	0.00 (0.004)	0.00 (0.007)	0.00 (0.011)	0.01 (0.022)
	\bar{Q}_0, \bar{g}_0					
	Mis- \bar{Q}_0	0.00 (0.003)	0.00 (0.006)	0.00 (0.009)	0.00 (0.013)	0.01 (0.019)
	Mis- \bar{g}_0	0.00 (0.001)	0.01 (0.002)	0.00 (0.005)	0.01 (0.008)	0.00 (0.023)
	Mis- \bar{Q}_0, \bar{g}_0	0.04 (0.003)	0.02 (0.006)	0.02 (0.009)	0.02 (0.013)	0.01 (0.019)
DRICE	Correct	0.00 (0.002)	0.00 (0.006)	0.01 (0.011)	0.04 (0.014)	0.07 (0.018)
	\bar{Q}_0, \bar{g}_0					
	Mis- \bar{Q}_0	0.00 (0.001)	0.02 (0.001)	0.03 (0.001)	0.03 (0.002)	0.06 (0.002)
	Mis- \bar{g}_0	0.00 (0.001)	0.01 (0.001)	0.02 (0.002)	0.04 (0.004)	0.07 (0.006)
	Mis- \bar{Q}_0, \bar{g}_0	0.05 (0.001)	0.04 (0.004)	0.03 (0.009)	0.01 (0.017)	0.01 (0.028)
TMLE	Correct	0.00 (0.002)	0.01 (0.005)	0.01 (0.009)	0.04 (0.013)	0.06 (0.017)
	\bar{Q}_0, \bar{g}_0					
	Mis- \bar{Q}_0	0.00 (0.002)	0.02 (0.006)	0.02 (0.014)	0.03 (0.021)	0.06 (0.031)

Time (t^*)	1	2	3	4	5	6
$\Psi_0(t^*)$	0.232	0.335	0.390	0.428	0.460	0.489
	Bias	[MSE]	Bias	[MSE]	Bias	[MSE]
	(Var)		(Var)		(Var)	
Mis- ξ_0	0.00 (0.001)	[0.001]	0.02 (0.003)	[0.003]	0.04 (0.010)	[0.020]
Mis- \bar{Q}_0, ξ_0	0.04 (0.001)	[0.003]	0.04 (0.003)	[0.004]	0.02 (0.023)	[0.030]
DRICE (weighted)	Correct	0.00 (0.001)	[0.001]	0.00 (0.003)	[0.003]	0.01 (0.012)
	\bar{Q}_0, ξ_0	0.00 (0.002)	[0.002]	0.00 (0.003)	[0.003]	0.04 (0.011)
	Mis- ξ_0	0.00 (0.001)	[0.001]	0.01 (0.002)	[0.002]	0.00 (0.013)
	Mis- \bar{Q}_0, ξ_0	0.04 (0.001)	[0.002]	0.03 (0.002)	[0.003]	0.02 (0.013)
TMLE (weighted)	Correct	0.00 (0.001)	[0.001]	0.00 (0.003)	[0.003]	0.01 (0.012)
	\bar{Q}_0, ξ_0	0.00 (0.002)	[0.002]	0.00 (0.003)	[0.003]	0.03 (0.009)
	Mis- ξ_0	0.00 (0.001)	[0.001]	0.01 (0.002)	[0.002]	0.00 (0.011)
	Mis- \bar{Q}_0, ξ_0	0.04 (0.001)	[0.002]	0.03 (0.002)	[0.003]	0.01 (0.010)

Estimates are based on 1000 simulations with a fixed sample size of $n=500$. $\Psi_0(t^*)$ represents the correct parameter value at time t^* under the intervention $\vec{a}(K) = 0$.

Misspecified model for ξ_0 was $\xi_0 A(t) = \text{logit}^{-1}(\beta_0 + \beta_1 W_1 + \beta_2 W_3 + \beta_3 L_1(t-1))$. Misspecified model for \bar{Q}_0 was $\bar{Q}_0 L(t) = \text{logit}^{-1}(\beta_0 + \beta_1 W_2 + \beta_2 L_1(t-1) + \beta_3 A(t-1))$.

Stratified and pooled models are identical for IPW.

References

- [1]. Robins JM. A new approach to causal inference in mortality studies with a sustained exposure period - application to control of the healthy worker survivor effect. *Math Modell* 1986;7:1393–512.
- [2]. Robins JM, Hernán MA. Estimation of the causal effects of time-varying exposures, chapter 1. In: Fitzmaurice GM, Davidian M, Verbeke G, Molenberghs G, editors. *Longitudinal Data Analysis I*. Hoboken, New Jersey: CRC Press, 2009.
- [3]. Horvitz D, Thompson D. A generalization of sampling without replacement from a finite universe. *J Am Stat Assoc* 1952;47:663–85.
- [4]. Robins JM. Marginal structural models. 1997 Proceedings of the american statistical association, section on bayesian statistical science, 1998:1–10. http://link.springer.com/chapter/10.1007/978-1-4419-9782-1_9.
- [5]. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *J Am Stat Assoc* 1994;89:846–67.
- [6]. Rotnitzky A, Robins J. Inverse probability weighted estimation in survival analysis. In: Armitage P and Colton T (editors). *Encyclopedia of Biostatistics*, 2nd edition. New York: Wiley, 2005;4:2619–25.
- [7]. Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. *Biometrics* 2005;61:962–73. <http://www.ncbi.nlm.nih.gov/pubmed/16401269>. [PubMed: 16401269]
- [8]. Petersen M, Schwab J, Gruber S, Blaser N, Schomaker M, van der Laan M. Targeted maximum likelihood estimation for dynamic and static longitudinal marginal structural working models hhs public access. *J Causal Inference* 2014;2:147–85. [PubMed: 25909047]
- [9]. Robins JM. Robust estimation in sequentially ignorable missing data and causal inference models. *Proceedings of the american statistical association section on bayesian statistical science* 2000:6–10.
- [10]. Scharfstein DO, Rotnitzky A, Robins JM. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J the Am Stat Assoc* 1999a;94:1096–120.
- [11]. Schnitzer ME, Moodie EE, van der Laan MJ, Platt RW, Klein MB. Modeling the impact of hepatitis C viral clearance on end-stage liver disease in an HIV co-infected cohort with targeted maximum likelihood estimation. *Biometrics* 2014;70:144–52. [PubMed: 24571372]
- [12]. van der Laan MJ, Gruber S. Targeted minimum loss based estimation of an intervention specific mean outcome Berkeley, CA: The Berkeley Electronic Press, 2011.
- [13]. Hernán Ma, Robins JM. Estimating causal effects from epidemiological data. *J Epidemiol Commun Health* 2006;60:578–86.
- [14]. Petersen ML, Porter KE, Gruber S, Wang Y, van der Laan MJ. Diagnosing and responding to violations in the positivity assumption. *Stat Meth Med Res* 2012;21:31–54.
- [15]. Cole SR, Hernan MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol* 2008;168:656–64. [PubMed: 18682488]
- [16]. Robins JM, Rotnitzky A, van der Laan MJ. Discussion of “On profile likelihood” by Murphy and van der Vaart. *J Am Stat Assoc* 2000b;95:477–82.
- [17]. van der Laan MJ, Robins JM. *Unified methods for censored longitudinal data and causality* New York: Springer, 2003.
- [18]. Kang JD, Schafer JL. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Stat Sci* 2007;22:523–39.
- [19]. Robins J, Sued M, Lei-gomez Q, Rotnitzky A. Comment : performance of double-robust estimators when “inverse probability” weights are highly variable. *Stat Sci* 2007;22:544–59.
- [20]. Rotnitzky A, Lei Q, Sued M, Robins JM. Improved double-robust estimation in missing data and causal inference models. *Biometrika* 2012:1–18.
- [21]. van der Laan MJ, Rose S, Gruber S. Readings on targeted maximum likelihood estimation Technical report, Bepress, 2009 <http://www.bepress.com/ucbbiostat/paper254>.

- [22]. van der Laan MJ, Rubin D. Targeted maximum likelihood learning U.C. Berkeley Division of Biostatistics Working Paper Series, 2006:1–87.
- [23]. Gutman R, Rubin DB. Estimation of causal effects of binary treatments in unconfounded studies. *Stat Med* 2015 <http://www.ncbi.nlm.nih.gov/pubmed/26013308>.
- [24]. Han P, Wang L. Estimation with missing data: beyond double robustness. *Biometrika* 2013;100:417–430. <http://biomet.oxfordjournals.org/cgi/doi/10.1093/biomet/ass087>.
- [25]. Hattori S, Henmi M. Stratified doubly robust estimators for the average causal effect. *Biometrics* 2014;70:270–7. <http://www.ncbi.nlm.nih.gov/pubmed/24571129>. [PubMed: 24571129]
- [26]. Li L, Kleinman K, Gillman MW. A comparison of confounding adjustment methods with an application to early life determinants of childhood obesity. *J Dev Origins Health Dis* 2014;5:435–47. <http://www.ncbi.nlm.nih.gov/pubmed/25171142>.
- [27]. Zhou J, Zhang Z, Li Z, Zhang J. Coarsened propensity scores and hybrid estimators for missing data and causal inference. *Int Stat Rev* 2014;n/a–n/a. <http://doi.wiley.com/10.1111/insr.12082>.
- [28]. Decker AL, Hubbard A, Crespi CM, Seto EY, Wang MC. Semiparametric estimation of the impacts of longitudinal interventions on adolescent obesity using targeted maximum-likelihood: accessible estimation with the ltmle package. *J Causal Inference* 2014;2:95–108. [PubMed: 26046009]
- [29]. Neugebauer R, Schmittiel JA, van der Laan MJ. Targeted learning in real-world comparative effectiveness research with time-varying interventions. *Stat Med* 2014;33:2480–520. [PubMed: 24535915]
- [30]. Schnitzer ME, Lok JJ, Bosch RJ. Double robust and efficient estimation of a prognostic model for events in the presence of dependent censoring. *Biostatistics* 2016;17, 165–77. [PubMed: 26224070]
- [31]. Westreich D, Cole SR. Invited commentary: positivity in practice. *Am J Epidemiol* 2010;171:674–7; discussion 678–81. [PubMed: 20139125]
- [32]. van der Laan MJ, Polley EC, Hubbard AE. Super learner U.C. Berkeley Division of Biostatistics Working Paper Series, 2007:1–20.
- [33]. Pearl J *Causality*, 2nd ed. New York: Cambridge University Press, 2009.
- [34]. Hampel FR. The influence curve and its role in robust estimation. *J Am Stat Assoc* 1974;69:383–93. <http://www.tandfonline.com/doi/abs/10.1080/01621459.1974.10482962>.
- [35]. Tsiatis A *Semiparametric theory and missing data* New York: Springer, 2006.
- [36]. Vermeulen K *Semiparametric efficiency* Gent, Belgium: University of Ghent: 2011.
- [37]. Robins JM, Rotnitzky A. Recovery of Information and adjustment for dependent censoring using surrogate markers, chapter 3. In: Jewell NP, Dietz K, Farewell VT, editors. *AIDS Epidemiology* Boston: Birkhäuser, 1992:297–331.
- [38]. Scharfstein DO, Rotnitzky A, Robins JM. Adjusting for nonignorable drop-out using semiparametric nonresponse models: rejoinder. *J Am Stat Assoc* 1999b;94:1135–46.
- [39]. Robins JM. A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *J Chronic Dis* 1987;40:139S–61S. [PubMed: 3667861]
- [40]. Taubman SL, Robins JM, Mittleman Ma, Hernán Ma. Intervening on risk factors for coronary heart disease: an application of the parametric g-formula. *Int J Epidemiol* 2009;38:1599–611. [PubMed: 19389875]
- [41]. Ma Hernán, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology (Cambridge, Mass.)* 2000;11:561–70.
- [42]. Robins J *Marginal structural models versus structural nested models as tools for causal inference. Stat Models Epidemiol Environ ...* 1999:1–30. http://link.springer.com/chapter/10.1007/978-1-4612-1284-3_2.
- [43]. Friedman J Greedy function approximation: a gradient boosting machine. *Ann Stat* 2001;29:1189–232.
- [44]. Friedman JH. Stochastic gradient boosting. *Comput Stat Data Anal* 2002;38:367–78.

- [45]. Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. Proceedings of the 5th annual ACM workshop on computational learning theory, 1992:144–52. <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.21.3818>.
- [46]. Cortes C, Vapnik VN. Support-vector networks. Mach Learn 1995;20:273–97.
- [47]. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys 1943;5:115–33.
- [48]. Altman N An introduction to kernel and nearest-neighbor nonparametric regression. Am Stat 1992;46:175–85.
- [49]. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. Technometrics 1970;12:55–67.
- [50]. Tibshirani R Regression Shrinkage and Selection via the Lasso. J Royal Stat Soc 58:267–88.
- [51]. Hastie T, Rosset S, Tibshirani R, Zhu J. The entire regularization path for the support vector machine. Test 2004;5:1391–415. <http://portal.acm.org/citation.cfm?id=1005332.1044706>.
- [52]. Hastie T, Tibshirani R, Friedman J. Elements of statistical learning, 2nd ed. Stanford, CA: Springer, 2008.
- [53]. van der Laan MJ, Rose S. Targeted learning New York: Springer, 2011.
- [54]. Lendle S, Schwab J, Petersen ML, van der Laan MJ. ltmle: an R package implementing targeted minimum loss-based estimation for longitudinal data. J Stat Softw 2016;81:1–21.
- [55]. Schwab J, Lendle S, Petersen M, van der Laan M. ltmle: Longitudinal Targeted Maximum Likelihood Estimation, 2015. R package version 0.9–6 <https://CRAN.R-project.org/package=ltmle>. CRAN.
- [56]. Tran L, Yiannoutsos CT, Musick BS, Wools-Kaloustian KK, Siika A, Kimaiyo S, van der Laan MJ, Petersen ML. Evaluating the impact of a HIV low-risk express care task-shifting program: a case study of the targeted learning roadmap The Berkeley Electronic Press, 2016.
- [57]. Miles CH, Petersen M, van der Laan MJ. Causal inference for a single group of causally-connected units under stratified interference. arXiv, 2017:1–38. <http://arxiv.org/abs/1710.09588>.
- [58]. Robins JM, Hernán MÁ, Brumback B Marginal structural models and causal inference in epidemiology. Epidemiology 2000;11:550–60. [PubMed: 10955408]

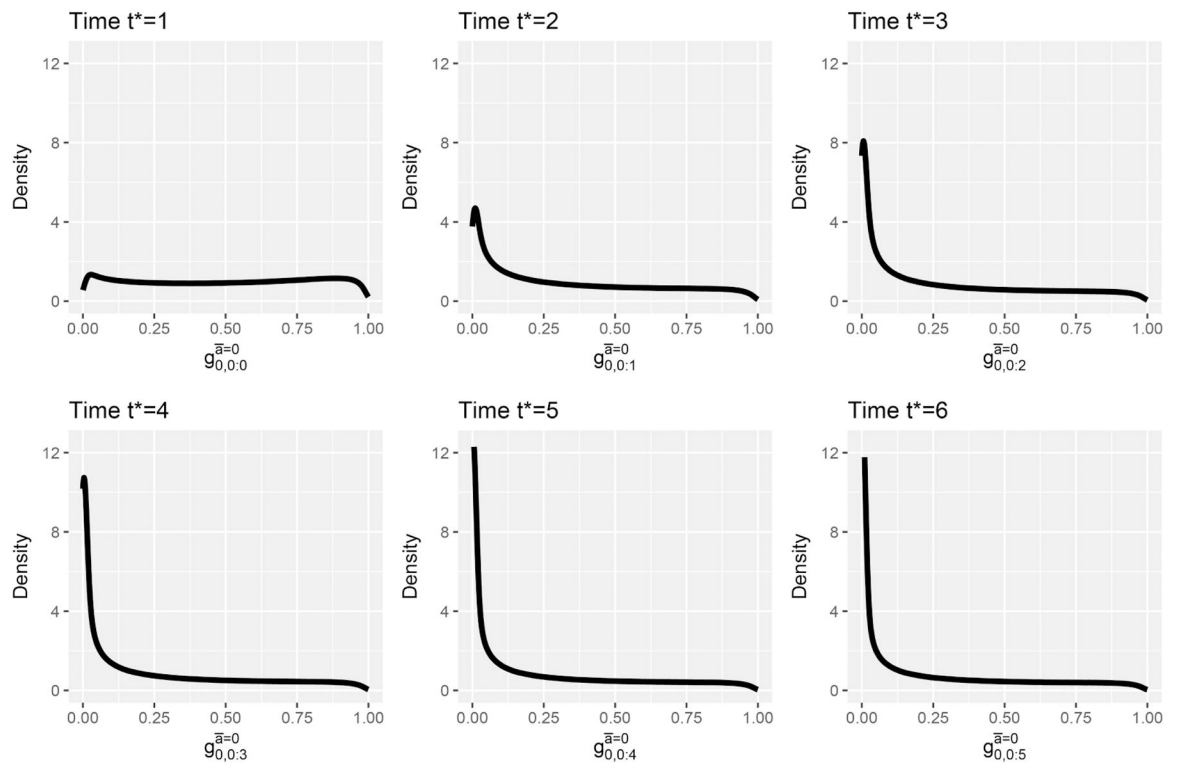


Figure 1. Marginal densities of $g_{0,0;t^*-1}^{\bar{a}=0}$ for each time point t^* . The densities become more concentrated close to 0 as t increases.

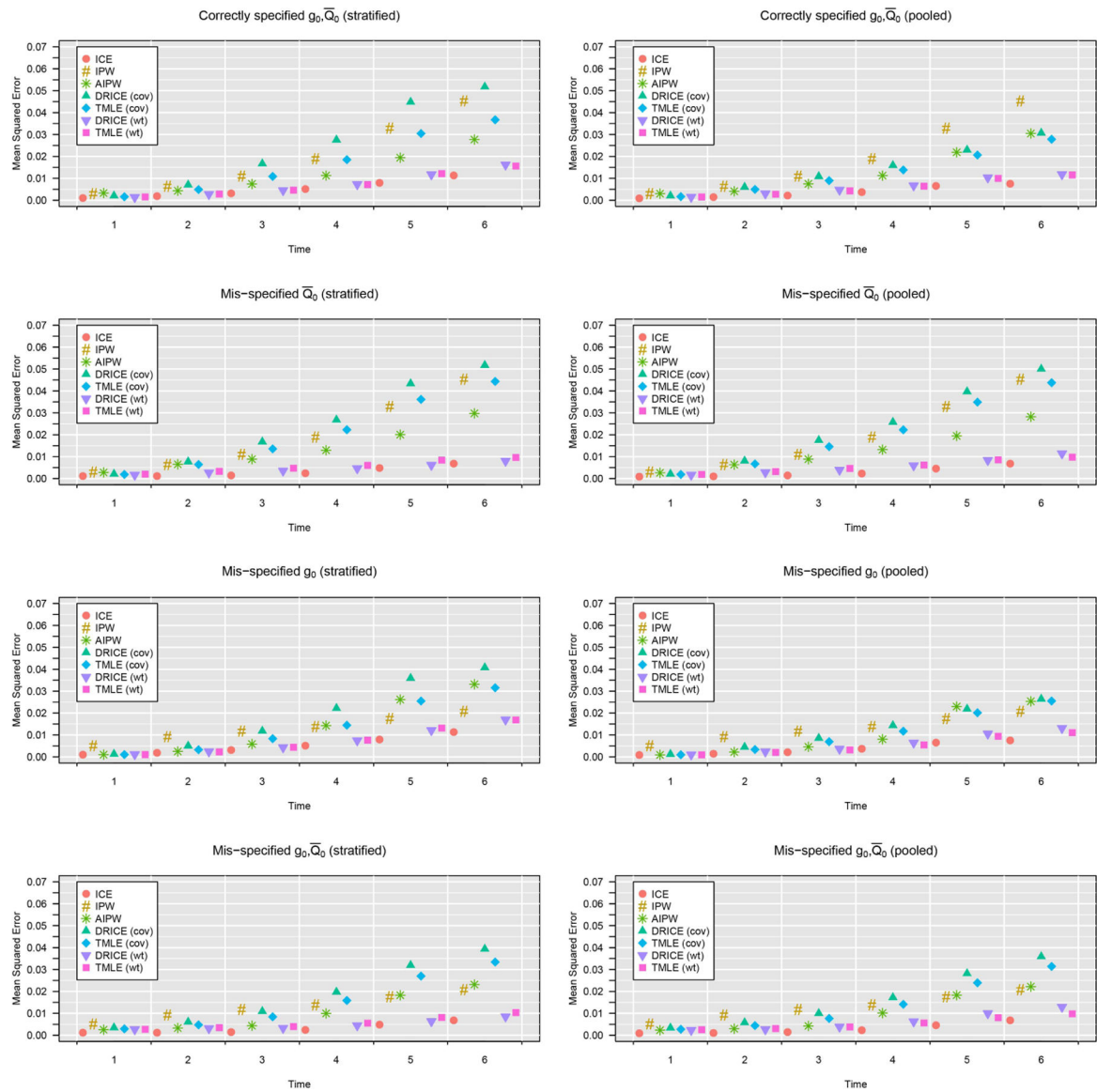


Figure 2. Simulated mean squared error of each estimator under correct[†] and mis-specification of g_0 and \bar{Q}_0 . ICE, weighted DRICE, and weighted TMLE have among the lowest mean squared error.

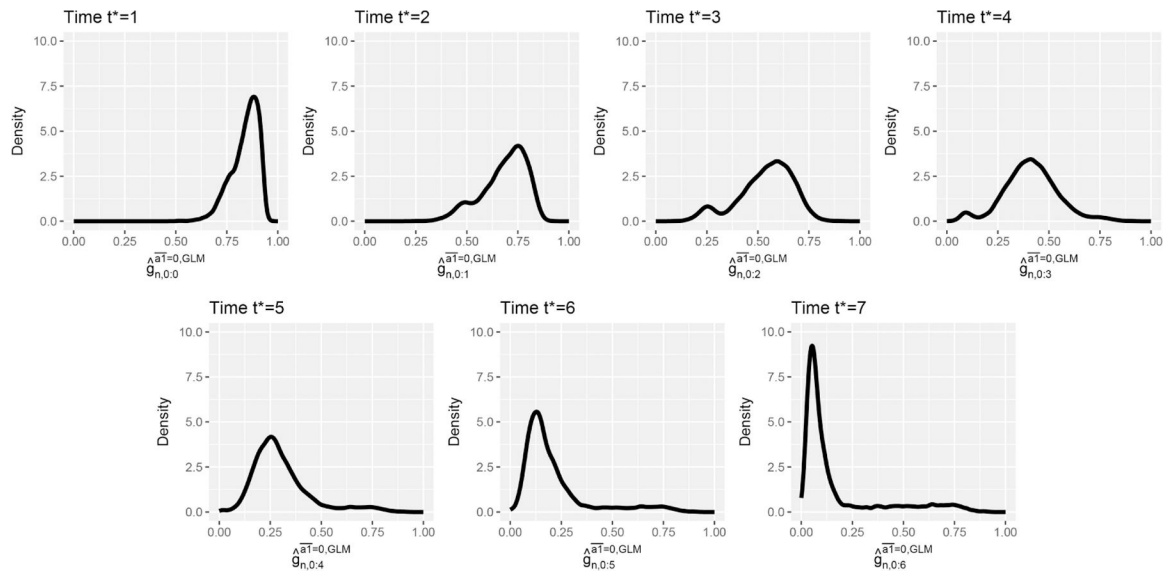


Figure 3. Marginal densities of $g_{n,0:t^*-1}^{\bar{a}}$ for each time point t^* taken over the 15,225 subjects using GLMs in applied example, where $\bar{a}1 = 0$.

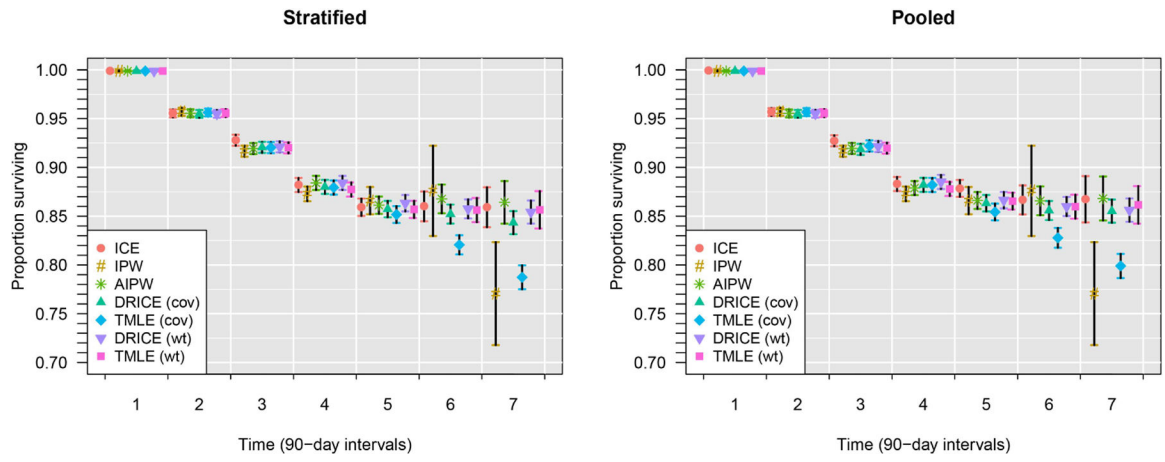


Figure 4. Applied example GLM estimates and 95% confidence intervals for estimating $\psi_{\bar{a}=0}(t^*) = \mathbb{E}[Y_{\bar{a}=0}(t^*)]: t^* = 1, 2, \dots, 7.$

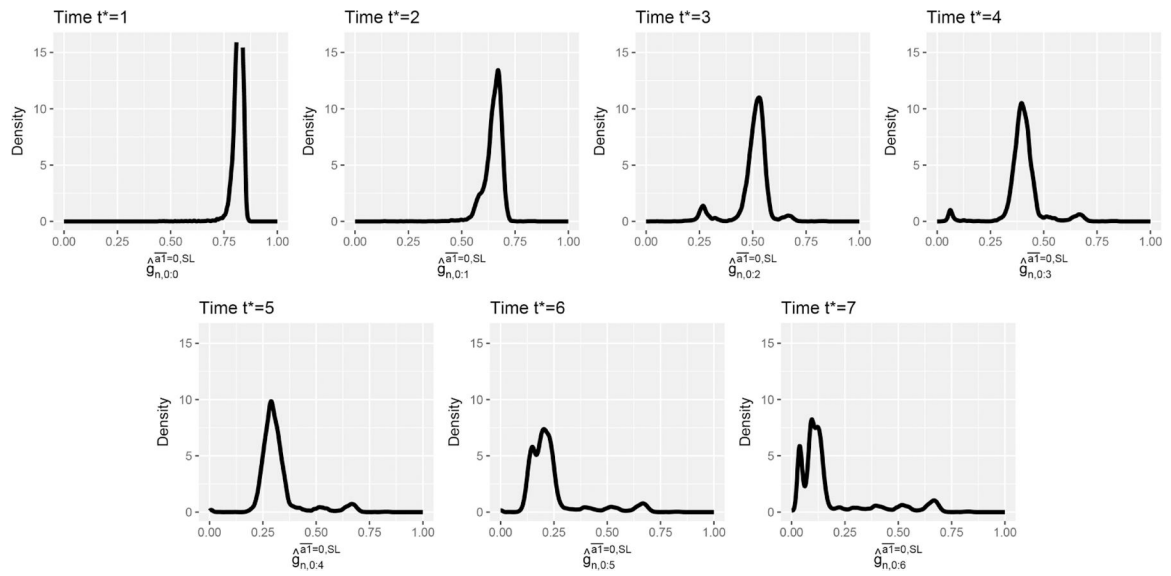


Figure 5. Marginal densities of $g_{n,0:t^*-1}^{\bar{a}_1}$ for each time point t^* taken over the 15,225 subjects using Super Learner from applied example, where $\bar{a}_1 = 0$.

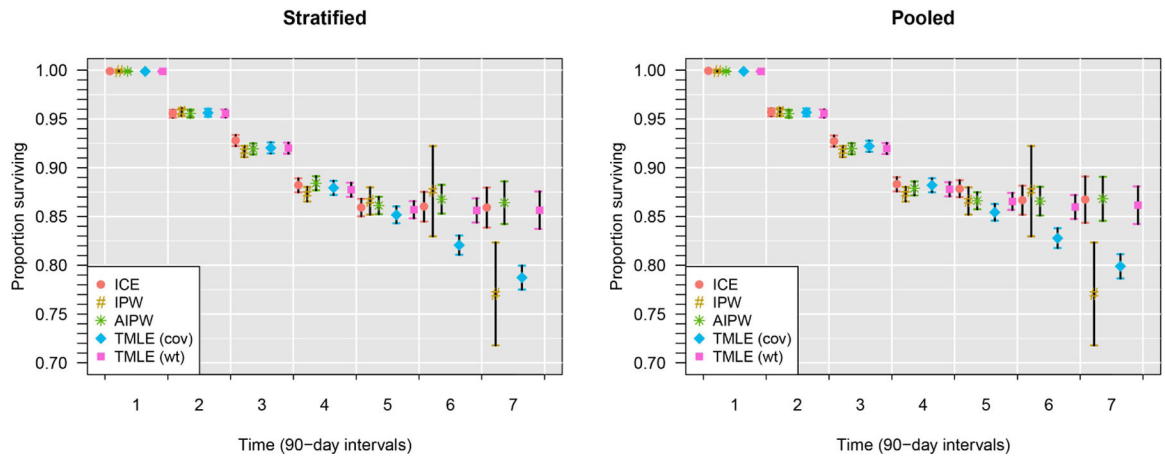


Figure 6. Applied example Super Learner estimates and 95% confidence intervals for estimating $\psi_{\bar{a}=0}(t^*) = \mathbb{E}[Y_{\bar{a}=0}(t^*)]$; $t^* = 1, 2, \dots, 7$. n. b. DRICE estimators were not used in this setting.