Biometrics 61, 962–972
December 2005

DOI: 10.1111/j.1541-0420.2005.00377.x

Doubly Robust Estimation in Missing Data and Causal Inference Models

Heejung Bang

Division of Biostatistics and Epidemiology, Department of Public Health, Weill Medical College of Cornell University, New York, New York 10021, U.S.A.
email: heb2013@med.cornell.edu

and

James M. Robins

Departments of Biostatistics and Epidemiology, Harvard School of Public Health, Boston, Massachusetts 02115, U.S.A.

Summary. The goal of this article is to construct doubly robust (DR) estimators in ignorable missing data and causal inference models. In a missing data model, an estimator is DR if it remains consistent when either (but not necessarily both) a model for the missingness mechanism or a model for the distribution of the complete data is correctly specified. Because with observational data one can never be sure that either a missingness model or a complete data model is correct, perhaps the best that can be hoped for is to find a DR estimator. DR estimators, in contrast to standard likelihood-based or (nonaugmented) inverse probability-weighted estimators, give the analyst two chances, instead of only one, to make a valid inference. In a causal inference model, an estimator is DR if it remains consistent when either a model for the treatment assignment mechanism or a model for the distribution of the counterfactual data is correctly specified. Because with observational data one can never be sure that a model for the treatment assignment mechanism or a model for the counterfactual data is correct, inference based on DR estimators should improve upon previous approaches. Indeed, we present the results of simulation studies which demonstrate that the finite sample performance of DR estimators is as impressive as theory would predict. The proposed method is applied to a cardiovascular clinical trial.

Key words: Causal inference; Doubly robust estimation; Longitudinal data; Marginal structural model; Missing data; Semiparametrics.

1. Introduction

In a missing data model, an estimator is doubly robust (DR) or doubly protected if it remains consistent when either a model for the missingness mechanism or a model for the distribution of the complete data is correctly specified. In a causal inference model, an estimator is DR if it remains consistent when either a model for the treatment assignment mechanism or a model for counterfactual data is correctly specified. Because of the frequency and near inevitability of model misspecification, double robustness is a highly desirable property.

Robins, Rotnitzky, and Zhao (1994) and Rotnitzky, Robins, and Scharfstein (1998) proposed augmented orthogonal inverse probability-weighted (AIPW) estimators in missing data models. Scharfstein, Rotnitzky, and Robins (1999) showed the orthogonal AIPW estimator had an alternative "regression representation." More importantly, they showed this estimator was DR and developed a general method to construct DR estimators in missing data models when the data are missing at random (MAR). They also showed how to construct DR estimators in causal inference models under the assumption of no unmeasured confounders. This methodology was further extended in Robins (2000), Robins, Rotnitzky, and Van der Laan (2000), Lunceford and Davidian (2004), Neugebauer and Van der Laan (2005), Lipsitz, Ibrahim, and Zhao (1999), Robins and Rotnitzky (2001), and Van der Laan and Robins (2003); the last two references provide the detailed mathematical theory underlying the methodology.

In this article, we review previously developed methods and algorithms for constructing DR estimators in nonlongitudinal missing data and causal inference models and extend them to longitudinal monotone missing data models and longitudinal causal models, specifically to longitudinal marginal structural models (MSMs). Although algebraically equivalent, we represent our DR estimators as sequential regression estimators rather than as AIPW estimators, because the former representation leads to a computational algorithm that can be easily implemented using standard off-the-shelf regression software.

This article is organized as follows. We begin, in Section 2, by considering estimation of the mean of an outcome variable from nonlongitudinal data when the outcome is MAR. We next consider estimation of the effect of a binary treatment in the presence of high-dimensional baseline covariate data under the assumption of no unmeasured confounders (i.e., ignorability). In Section 3, we obtain a DR estimator for monotone MAR longitudinal data. In Section 4, we construct DR estimators of the parameters of an MSM under the assumption of no unmeasured confounders. In each section, results of simulations illustrate the finite sample efficiency and robustness of our DR estimators. The method is illustrated with the data from a multicenter cardiovascular clinical trial in Section 5. Some concluding remarks and discussion are provided. In the Appendix, we show how to represent our sequential regression estimators as AIPW estimators.

2. Cross-Sectional Models

In this section, we show how to construct DR estimators in two nonlongitudinal models: the first a model with missing outcome data and the second a model for a treatment effect.

2.1 *A Missing Data Model*

Consider an observational follow-up study with full data $\mathbf{L} = (\mathbf{V}', Y)'$, where $\mathbf{V}$ is an always observed vector of baseline variables and Y is a scalar outcome which is missing by happenstance on some subjects. Let Δ be the indicator of whether Y is missing. Then the observed data are $\mathbf{O} = (\Delta, \mathbf{L}_{\text{obs}})$, where $\mathbf{L}_{\text{obs}} = \mathbf{L}$ when Δ is 1 and $\mathbf{L}_{\text{obs}} = \mathbf{V}$ when $\Delta = 0$. In realistic epidemiologic studies it would not be unusual for the sample size n to be between 500 and 2000 and yet for $\mathbf{V}$ to be 50–100 dimensional.

Suppose that interest lies in estimating the unconditional mean μ of Y based on n i.i.d. copies of $\mathbf{O}_i$ $(i = 1, \ldots, n)$. If, as we assume, Y is MAR and the probability of observing complete data is always positive, that is, $P(\Delta = 1 \mid Y, \mathbf{V}) = P(\Delta = 1 \mid \mathbf{V}) \equiv \pi(\mathbf{V}) > 0$ with probability 1, then we can represent the mean $\mu = E(Y) = E\{E(Y \mid \mathbf{V})\}$ of Y in terms of the distribution of the observed data as either $E\{E(Y \mid \Delta = 1, \mathbf{V})\}$ or $E\{\Delta Y/\pi(\mathbf{V})\}$. The second representation of Y suggests (i) fitting a model for the "propensity score" (PS) $\pi(\mathbf{V})$ based on a parametric model $\pi(\mathbf{V};\ \boldsymbol{\alpha})$, such as the linear logistic regression model logit $\{\pi(\mathbf{V};\ \boldsymbol{\alpha})\} = \boldsymbol{\alpha}'\mathbf{V}$, where logit $(x) = \log\ \{x/(1 - x)\}$, and (ii) then estimating μ with the Horvitz–Thompson (HT) estimator $\hat{\mu}_{\text{HT}} = n^{-1}\sum_i \Delta_i Y_i/\pi(\mathbf{V}_i;\ \hat{\boldsymbol{\alpha}})$, where $\hat{\boldsymbol{\alpha}}$ is the maximum likelihood estimator (MLE) of $\boldsymbol{\alpha}$ (Horvitz and Thompson, 1952; Rosenbaum, 1987). Note that because $\mathbf{V}$ is very high-dimensional, it may not be feasible to estimate $\pi(\mathbf{V})$ nonparametrically using smoothing techniques. Rather we must specify a dimension reducing parametric model for $\pi(\mathbf{V}\,; \boldsymbol{\alpha})$.

The first representation of Y suggests (i) fitting a model $\Psi\{s(\mathbf{V}\,;\boldsymbol{\beta})\}$ for $E(Y \mid \Delta = 1,\ \mathbf{V})$ with Ψ^{-1} a known link function and $s(\mathbf{V};\boldsymbol{\beta})$ a known regression function of an unknown finite-dimensional parameter $\boldsymbol{\beta}$, and (ii) then estimating μ by the outcome regression (OR) estimator $\hat{\mu}_{\text{OR}} = n^{-1}\sum_i \Psi\{s(\mathbf{V}_i;\tilde{\boldsymbol{\beta}})\}$, the sample average over all subjects of the predicted values $\Psi\{s(\mathbf{V}_i;\tilde{\boldsymbol{\beta}})\}$ of the Y_i. Here $\tilde{\boldsymbol{\beta}}$ solves the "normal equations" $\mathbf{0} = \sum_{i=1}^{n} \Delta_i \partial s(\mathbf{V}_i;\boldsymbol{\beta})/\partial\boldsymbol{\beta}'[Y_i - \Psi\{s(\mathbf{V}_i;\boldsymbol{\beta})\}]$, where $\mathbf{0}$ is a vector of all zeros of an appropriate dimension. Note that if Ψ^{-1} is the canonical link function of a generalized linear model (GLM), these equations are precisely the likelihood (score) equations for the model and the resulting estimator is the MLE. This estimator is often referred to as the iteratively reweighted least squares (IRLS) estimator because an IRLS algorithm is often used to solve the score equations. For example, if Y were dichotomous, we choose $\Psi^{-1}\ (x) = \ln\{x/(1 - x)\}$ to be the logit link, $[\Psi(x) = e^x/(1 + e^x)]$, and might choose $s(\mathbf{V};\boldsymbol{\beta})$ to be the linear function $s(\mathbf{V};\boldsymbol{\beta}) = \boldsymbol{\beta}'\mathbf{V}$. Then $\tilde{\boldsymbol{\beta}}$ is the MLE of $\boldsymbol{\beta}$ in this linear logistic model among subjects on whom Y was observed.

There has been considerable debate as to which approach to estimating the mean of Y is to be preferred as the approach based on the HT estimator $\hat{\mu}_{\text{HT}}$ is inconsistent if the model for Δ (i.e., the PS model) is misspecified while the approach based on the OR estimator $\hat{\mu}_{\text{OR}}$ is inconsistent if the OR model $s(\mathbf{V};\boldsymbol{\beta})$ is misspecified. This controversy could be resolved if an estimator were available that was guaranteed to be consistent for μ whenever at least one of the two models was correct. We refer to such an estimator as DR as it can protect against misspecification of either the OR model or the PS model, although not against simultaneous misspecification of both. Because with observational data one can never be sure that either model is correct, the best that can be hoped for is to find a DR estimator.

Scharfstein et al. (1999, p. 1140–1141) showed that to obtain a DR estimator in this setting it suffices to model $E(Y \mid \Delta = 1, \mathbf{V})$ as $e(\mathbf{V};\boldsymbol{\beta},\phi) = \Psi\{s(\mathbf{V};\boldsymbol{\beta}) + \phi\pi^{-1}(\mathbf{V};\hat{\boldsymbol{\alpha}})\}$, which adds the covariate $\pi^{-1}(\mathbf{V};\hat{\boldsymbol{\alpha}})$ to the OR model $\Psi\{s(\mathbf{V};\boldsymbol{\beta})\}$. Then the estimator

$$
\begin{aligned}
\hat{\mu}_{dr} &= n^{-1}\sum_i e(\mathbf{V}_i;\hat{\boldsymbol{\beta}},\hat{\phi}) \\
&= n^{-1}\sum_i \Psi\{s(\mathbf{V}_i;\hat{\boldsymbol{\beta}}) + \hat{\phi}\pi^{-1}(\mathbf{V}_i;\hat{\boldsymbol{\alpha}})\}
\end{aligned}
$$

is DR in the sense that $\hat{\mu}_{dr}$ is consistent asymptotically normal (CAN) if either the model $e(\mathbf{V};\boldsymbol{\beta},\ \phi)$ (with $\pi^{-1}(\mathbf{V};\hat{\boldsymbol{\alpha}})$ replaced by its probability limit) for $E(Y \mid \Delta = 1,\ \mathbf{V})$ or the PS model $\pi(\mathbf{V};\alpha)$ is correct. Here $(\hat{\boldsymbol{\beta}},\hat{\phi})$ jointly solve $\mathbf{0} = \sum_i \Delta_i \partial e(\mathbf{V}_i;\boldsymbol{\beta},\phi)/\partial(\boldsymbol{\beta}',\phi)\{Y_i - e(\mathbf{V}_i;\boldsymbol{\beta},\phi)\}$. Thus if $Y$is dichotomous, Ψ^{-1} is the logit link and $s(\mathbf{V};\boldsymbol{\beta}) = \boldsymbol{\beta}'\mathbf{V}$,$(\hat{\boldsymbol{\beta}}',\hat{\phi})'$ are the MLEs among subjects with $\Delta = 1$ in the logistic regression model with covariates $\mathbf{V}$ and $1/\pi(\mathbf{V};\hat{\boldsymbol{\alpha}})$. It is clear that the $\hat{\mu}_{dr}$ is CAN when the model $e(\mathbf{V};\boldsymbol{\beta},\ \phi)$ is correct.

To see why it is DR, consider the estimator $\tilde{\mu}_{AIPW}$ solving $0 = \hat{U}(\mu)$ where $\hat{U}(\mu)$ can be written in either of the following algebraically equivalent forms:

$$
\begin{aligned}
\hat{U}(\mu) &= \sum_i \pi^{-1}(\mathbf{V};\hat{\boldsymbol{\alpha}})\Delta_i(Y_i - \mu) \\
&\quad - \Big\{\pi^{-1}(\mathbf{V};\hat{\boldsymbol{\alpha}})\Delta_i - 1\Big\}\{e(\mathbf{V}_i;\hat{\boldsymbol{\beta}},\hat{\phi}) - \mu\} \\
&= \sum_i \pi^{-1}(\mathbf{V};\hat{\boldsymbol{\alpha}})\Delta_i\{Y_i - e(\mathbf{V}_i;\hat{\boldsymbol{\beta}},\hat{\phi})\} \\
&\quad + \sum_i \{e(\mathbf{V}_i;\hat{\boldsymbol{\beta}},\hat{\phi}) - \mu\}.
\end{aligned}
$$

Because it solves an AIPW estimating equation, $\tilde{\mu}_{AIPW}$ is obviously CAN if the model $\pi(\mathbf{V};\alpha)$ is correct. We now show that $\tilde{\mu}_{AIPW} = \hat{\mu}_{dr}$ (i.e., $\hat{\mu}_{dr}$ is simply a regression representation of $\tilde{\mu}_{AIPW}$), which proves the double robustness of $\tilde{\mu}_{AIPW} = \hat{\mu}_{dr}$. It follows from the second representation of $\hat{U}(\mu)$, that $\tilde{\mu}_{AIPW} = \hat{\mu}_{dr}$ provided $\sum_i \pi^{-1}(V;\hat{\boldsymbol{\alpha}})\Delta_i\{Y_i - e(V_i;\hat{\boldsymbol{\beta}},\hat{\phi})\} = 0$. But this equality is immediately seen from the normal equations satisfied by $(\hat{\boldsymbol{\beta}},\hat{\phi})$. The key to the double robustness property of $\hat{U}(\mu)$ is that it estimates the orthogonal estimating function $U_{orth}\,(\mu)$ obtained by replacing the estimates $\pi^{-1}(V;\hat{\boldsymbol{\alpha}})$ and $e(V;\hat{\boldsymbol{\beta}},\hat{\phi})$ in $\hat{U}(\mu)$ with the true functions $\pi^{-1}\;(V)$ and $E(Y\mid\mathbf{V})$. An AIPW estimating equation is said to be orthogonal if it is uncorrelated with the set $\{\sum_i\{\Delta_i - \pi(\mathbf{V}_i)\}h(V_i); h(V_i)$ arbitrary$\}$ of scores of any PS model.

One could wonder about the actual advantage of using DR estimators as, in practice, all models including the OR and PS models are misspecified and thus even the DR estimator of μ may be considerably biased. In our opinion, a DR estimator has the following advantage that argues for its routine use: if either the model for the OR or the model for the PS is nearly correct, then the bias of a DR estimator of μ will be small. Thus, the DR estimator $\hat{\mu}_{dr}$, in contrast with both the OR estimator $\hat{\mu}_{\mathrm{OR}}$ and the HT estimator $\hat{\mu}_{\mathrm{HT}}$, gives the analyst two chances to get nearly correct inference about the mean of Y. Of course, there can be an efficiency cost to using a DR estimator rather than the OR estimator of μ. However, we will see in the simulation study reported later in this section that the use of DR estimators may provide major improvements in robustness while incurring strikingly little efficiency loss.

A further advantage of DR estimation is that comparison of the three estimators $\hat{\mu}_{dr}, \hat{\mu}_{\mathrm{HT}}$, and $\hat{\mu}_{\mathrm{OR}}$ with one another serves as a useful goodness of fit test (Robins and Rotnitzky, 2001). To formalize here, let $\hat{\tau}^2_{dr-\mathrm{HT}}$ and $\hat{\tau}^2_{dr-\mathrm{OR}}$ be the empirical variance of $(\hat{\mu}_{dr} - \hat{\mu}_{\mathrm{HT}})$ and $(\hat{\mu}_{dr} - \hat{\mu}_{\mathrm{OR}})$, respectively, calculated from a large number of nonparametric bootstrap replications of the study data. Then the tests with rejection regions $|(\hat{\mu}_{dr} - \hat{\mu}_{\mathrm{HT}})/\hat{\tau}_{dr-\mathrm{HT}}| > 1.96$ and $|(\hat{\mu}_{dr} - \hat{\mu}_{\mathrm{OR}})/\hat{\tau}_{dr-\mathrm{OR}}| > 1.96$ are valid large sample 0.05 level tests of the null hypotheses that the PS model and the OR model, respectively, are correctly specified. However, the tests are not consistent. That is, there exist laws under which the PS and OR models are incorrect but the estimators $\hat{\mu}_{dr}$ and $\hat{\mu}_{\mathrm{HT}}$ converge in probability to a common value μ^* that differs from the true parameter μ, resulting in misleading inference. The same holds true with $\hat{\mu}_{\mathrm{OR}}$ replacing $\hat{\mu}_{\mathrm{HT}}$. However, although logically possible, such inconsistency may be uncommon in practice.

One possible theoretical objection to $\hat{\mu}_{dr}$ is that when the PS is either known or correctly modeled, $\hat{\mu}_{dr}$ can be less efficient than $\hat{\mu}_{\mathrm{HT}}$ if the model for $E(Y\mid\Delta = 1,\, V)$ is badly misspecified. Robins (2002, Appendix 4) has developed an alternative DR estimator, referred to as $\hat{\mu}_{IPCW}$, that, as noted by Robins, Rotnitzky, and Bonetti (2001), is always guaranteed to be at least as efficient as $\hat{\mu}_{\mathrm{HT}}$ when the PS is either known or correctly modeled. However, $\hat{\mu}_{IPCW}$ is more difficult than $\hat{\mu}_{dr}$ to compute with standard software. Furthermore, in practice, it would be rare for the model for $E(Y\mid\Delta = 1,\, V)$ to be so badly misspecified that $\hat{\mu}_{dr}$ was seriously inefficient.

2.2 *A Treatment-Effect Model*

In this subsection we show the DR estimator of the mean of Y can be generalized to provide an estimator of the average causal effect of a binary treatment from observational data under the assumption of no unmeasured confounders. Consider an observational study with i.i.d. data $\{\mathbf{O}_i = (\Delta_i,\, Y_i,\, \mathbf{V}_i)\ ;\ i = 1,\ldots,n\}$ on n study subjects, where Δ is the indicator of the dichotomous treatment, Y is the outcome, and $\mathbf{V}$ is a high-dimensional vector of pretreatment confounding variables.

We assume ignorable treatment assignment, that is, $Y(\delta)\amalg\Delta\mid\mathbf{V}$, where $Y(\delta)$ is the counterfactual outcome at treatment level $\delta(\delta\in\{0,\,1\})$, and $A\amalg B\mid C$ denotes independence between A and B conditional on C. We often refer to the assumption of ignorable treatment assignment as the assumption of no unmeasured confounders. Under the assumption of no unmeasured confounders, the average treatment effect $\mu\equiv E\{Y(1)\} - E\{Y(0)\}$ can be written in two different ways as a function of the joint distribution of the observed data. Specifically, $\mu\equiv E\{E(Y\mid\Delta = 1,\,\mathbf{V}) - E(Y\mid\Delta = 0,\,\mathbf{V})\}$ and $\mu = E\{\Delta Y/\pi(\mathbf{V})\} - E[(1-\Delta)Y/\{1-\pi(\mathbf{V})\}]$. Thus given a parametric OR model $\Psi\{s(\Delta,\,\mathbf{V};\boldsymbol{\beta})\}$ for $E(Y\mid\Delta,\,\mathbf{V})$, we could estimate μ by $\hat{\mu}_{\mathrm{OR}} = n^{-1}\sum_i[\Psi\{s(1,\mathbf{V}_i;\tilde{\boldsymbol{\beta}})\} - \Psi\{s(0,\mathbf{V}_i;\tilde{\boldsymbol{\beta}})\}]$, the difference in the treatment-specific OR estimators of $E\{Y(\delta)\}$. Here $\tilde{\boldsymbol{\beta}}$ solves

$$\mathbf{0} = \sum_{i=1}^{n}\partial s\left(\Delta_i,\mathbf{V}_i;\boldsymbol{\beta}\right)/\partial\boldsymbol{\beta}'[Y_i - \Psi\{s\left(\Delta_i,\mathbf{V}_i;\boldsymbol{\beta}\right)\}],$$

which reduce to the ordinary least squares (OLS) normal equations when, for example, $\Psi(\cdot)$ is the identity link. A simple choice for $s(\Delta,\,\mathbf{V};\boldsymbol{\beta})$ would be $\boldsymbol{\beta}'\;(\Delta,\,\mathbf{V}')'$ in the absence of the interactions between the treatment and covariates.

Alternatively, we can estimate μ by $\hat{\mu}_{\mathrm{HT}} = n^{-1}[\sum_i\Delta_iY_i/\pi(\mathbf{V}_i;\hat{\boldsymbol{\alpha}}) - \sum_i(1-\Delta_i)Y_i/\{1-\pi(\mathbf{V}_i;\hat{\boldsymbol{\alpha}})\}]$, the difference in the treatment-arm-specific HT estimators of $E\{Y(\delta)\}$, where $\hat{\boldsymbol{\alpha}}$ is as in the previous subsection. Now $\hat{\mu}_{\mathrm{HT}}$ is inconsistent if the PS (i.e., treatment) model is misspecified, while $\hat{\mu}_{\mathrm{OR}}$ is inconsistent if the OR model is misspecified. Scharfstein et al. (1999, p. 1141) also showed that to obtain a DR estimator, we can model $E(Y\mid\Delta,\,\mathbf{V})$ by

$$\begin{aligned} e(\Delta,\mathbf{V};\boldsymbol{\beta},\phi_1,\phi_2) = \Psi[s(\Delta,\mathbf{V};\boldsymbol{\beta}) + \phi_1\Delta\pi^{-1}(\mathbf{V};\hat{\boldsymbol{\alpha}})\\ + \phi_2(1-\Delta)\{1-\pi(\mathbf{V};\hat{\boldsymbol{\alpha}})\}^{-1}],\end{aligned}$$

which adds the covariates $\Delta\pi^{-1}(\mathbf{V};\hat{\boldsymbol{\alpha}})$ and $(1-\Delta)\{1-\pi(\mathbf{V};\hat{\boldsymbol{\alpha}})\}^{-1}$ to the original OR model. In fact, an alternative DR estimator $\hat{\mu}_{dr}$, that is more efficient than Scharfstein et al.'s when only the OR model $\Psi\{s(\Delta,\,\mathbf{V};\,\boldsymbol{\beta})\}$ is correct, is to impose $\phi_1 = \phi_2$ in the previous model. That is, to obtain $\hat{\mu}_{dr}$ we fit the model

$$e(\Delta,\mathbf{V};\boldsymbol{\beta},\phi) = \Psi[s(\Delta,\mathbf{V};\boldsymbol{\beta}) + \phi\{f(\Delta\mid V;\hat{\boldsymbol{\alpha}})\}^{-1}],$$

where $f(\Delta\mid V;\boldsymbol{\alpha}) = \Delta\pi(V;\hat{\boldsymbol{\alpha}}) + (1-\Delta)\{1-\pi(V;\hat{\boldsymbol{\alpha}})\}$ is a subject's estimated probability of getting the treatment they actually received.

The estimator $\hat{\mu}_{dr} = n^{-1}\sum_i\{e(1,\mathbf{V}_i;\hat{\boldsymbol{\beta}},\hat{\phi}) - e(0,\mathbf{V}_i;\hat{\boldsymbol{\beta}},\hat{\phi})\}$ is DR in the sense that $\hat{\mu}_{dr}$ is CAN if either the model $e(\Delta,\,\mathbf{V};\boldsymbol{\beta},\,\phi)$ for $E(Y\mid\Delta,\,\mathbf{V})$ or the PS model $\pi(\mathbf{V};\alpha)$ is correct. Here, $(\hat{\boldsymbol{\beta}},\hat{\phi})$ jointly solve $\mathbf{0} = \sum_i\partial e(\Delta_i,\mathbf{V}_i;\,\boldsymbol{\beta},\phi)/\partial(\boldsymbol{\beta}',\phi)\{Y_i - e(\Delta_i,\mathbf{V}_i;\,\boldsymbol{\beta},\phi)\}$.

The estimator $\hat{\mu}_{dr}$ solves a long-standing open problem in the estimation of treatment effects: what function (or functions) of the PS needs to be added to a model $\Psi\{s(\Delta,\,\mathbf{V};\boldsymbol{\beta})\}$ for $E(Y\mid\Delta,\,\mathbf{V})$ in order to ensure consistent estimation of the average treatment effect when the PS is modeled correctly but the OR model $\Psi\{s(\Delta,\, \mathbf{V}; \boldsymbol{\beta})\}$ is incorrect. We see that we must add to the regression the inverse probability of treatment weighted (IPTW) covariate $1/f(\Delta \,|\, V; \hat{\boldsymbol{\alpha}})$, which is the (estimated) inverse of the PS for treated subjects ($\Delta = 1$) and the inverse of "1 minus the PS" for untreated subjects ($\Delta = 0$). Other choices can result in inconsistent estimation of the average treatment effect.

2.3 *A Simulation Study*

Numerical studies were performed to compare the finite sample behavior of the standard estimators and the proposed DR estimator. In our numerical experiments we assumed a linear regression model for $E(Y \,|\, \Delta,\, V)$ (i.e., an identity link function), but a nonlinear model with other canonical link functions could have been used. Simulation results are summarized in terms of the bias, variance, and interquartile range of the estimates. The precise definitions of the estimators and all the models employed for data generation are summarized in Table 1 and in the footnote of Table 2, and are omitted from the main text. In all simulations, the sample size was 500 and 1000 simulations were conducted.

Turn first to the missing data model. Recall the full data are $\mathbf{L} = (\mathbf{V}',\, Y)'$. We took $\mathbf{V} = (V_1,\, V_2,\, V_3)'$ to be a vector of always observed baseline variables. We generated V_k ($k = 1,\, 2,\, 3$) independently from a standard normal distribution and then Y from a normal distribution with mean of $s(\mathbf{V}; \boldsymbol{\beta})$ and a unit variance (see Table 1A). The parameter values chosen in Table 1A imply the marginal mean μ of Y is 1. The missingness indicator Δ was generated from the logistic regression model logit $\{\pi(\mathbf{V}; \boldsymbol{\alpha})\}$ also given in Table 1A. To investigate the robustness to misspecification, we also considered false models for both the missingness

Table 1
Simulation scenarios

	A. Nonlongitudinal model
True	$s(\mathbf{V}; \boldsymbol{\beta}) = \boldsymbol{\beta} \cdot [1,\, V_1^2,\, V_2,\, V_2V_3\,]',\, \boldsymbol{\beta} = [0,\, 1,\, 2.5,\, 3].$
	$\text{logit}\{\pi(\mathbf{V}; \boldsymbol{\alpha})\} = \boldsymbol{\alpha} \cdot [1,\, I_1,\, I_2,\, I_3,\, I_1I_2\,]',\, \boldsymbol{\alpha} = [-1,\, 1,\, 0,\, 0,\, -1].$
False	$s(\mathbf{V}; \boldsymbol{\beta}) = \boldsymbol{\beta} \cdot [1,\, V_1,\, V_2^2\,]'.$
	$\text{logit}\{\pi(\mathbf{V}; \boldsymbol{\alpha})\} = \boldsymbol{\alpha} \cdot [I_1,\, I_3\,]'.$
	B. Treatment-effect model
True	$s(\Delta,\, \mathbf{V}; \boldsymbol{\beta}) = \boldsymbol{\beta} \cdot [1,\, V_1^2\Delta,\, V_2\Delta,\, V_2V_3\,(1-\Delta),\, V_3\,(1-\Delta)]',\, \boldsymbol{\beta} = [0,\, 2,\, 3,\, 2,\, -4].$
	$\text{logit}\{\pi(\mathbf{V}; \boldsymbol{\alpha})\} = \boldsymbol{\alpha} \cdot [1,\, I_1,\, I_2,\, I_3,\, I_1I_2\,]',\, \boldsymbol{\alpha} = [-3,\, 2.5,\, 3,\, 1,\, -3].$
False	$s(\Delta,\, \mathbf{V}; \boldsymbol{\beta}) = \boldsymbol{\beta} \cdot [1,\, V_1\Delta,\, V_2^2\,(1-\Delta)]'.$
	$\text{logit}\{\pi(\mathbf{V}; \boldsymbol{\alpha})\} = \boldsymbol{\alpha} \cdot [I_3,\, V_4\,]'.$
	C. Longitudinal data model
True	$s_1\,(\mathbf{L}_1;\, \boldsymbol{\beta}) = \boldsymbol{\beta} \cdot [1,\, V_{11},\, V_{11}V_{13}\,]',\, \boldsymbol{\beta} = [0,\, 3,\, -2].$
	$s_2(\overline{\boldsymbol{L}}_2; \boldsymbol{\beta}) = \boldsymbol{\beta} \cdot [1, V_{11}^2, V_{12}, V_2^2, V_{12}V_2]', \boldsymbol{\beta} = [0, -3, 3, 1, -2].$
	$\text{logit}\{\lambda(1 \,|\, \mathbf{L}_1;\, \boldsymbol{\alpha})\} = \boldsymbol{\alpha} \cdot [1,\, I_{11},\, I_{12},\, I_{13},\, I_{11}I_{12}\,]',\, \boldsymbol{\alpha} = [-1,\, 1,\, 1,\, -1,\, -1].$
	$\text{logit}\{\lambda(2 \,|\, \overline{\mathbf{L}}_2; \boldsymbol{\alpha})\} = \boldsymbol{\alpha} \cdot [1, I_{11}, I_{12}, I_{13}, I_{11}I_{12}, I_2, I_2I_{13}]', \boldsymbol{\alpha} = [0, 1, 1, 0, -1, 0, -2].$
False	$s_1\,(\mathbf{L}_1;\, \boldsymbol{\beta}) = \boldsymbol{\beta} \cdot [1,\, V_{11},\, V_{12}\,]'.$
	$s_2(\overline{\mathbf{L}}_2; \boldsymbol{\beta}) = \boldsymbol{\beta} \cdot [1, V_{11}, V_{12}^2, V_{13}^2, V_2]'.$
	$\text{logit}\{\lambda(1 \,|\, \mathbf{L}_1;\, \boldsymbol{\alpha})\} = \boldsymbol{\alpha} \cdot [1,\, I_{12},\, I_{13}\,]'.$
	$\text{logit}\{\lambda(2 \,|\, \overline{\mathbf{L}}_2; \boldsymbol{\alpha})\} = \boldsymbol{\alpha} \cdot [1, I_2]'.$
	D. MSM
True	$s_1\,(\mathbf{L}_1,\, a_1 = 1;\, \boldsymbol{\beta}) = \boldsymbol{\beta} \cdot [1,\, V_{11},\, V_{11}V_{13}\,]',\, \boldsymbol{\beta} = [0,\, 3,\, -2].$
	$s_1\,(\mathbf{L}_1,\, a_1 = 0;\, \boldsymbol{\beta}) = \boldsymbol{\beta} \cdot [1,\, V_{11},\, V_{12}\,]',\, \boldsymbol{\beta} = [0,\, -1,\, 3].$
	$s_2(\overline{\mathbf{L}}_2, \overline{\mathbf{a}}_2 = (1,1); \boldsymbol{\beta}) = \boldsymbol{\beta} \cdot [1, V_{11}^2, V_{12}, V_2^2, V_{12}V_2]', \boldsymbol{\beta} = [0, -3, 3, 1, -2].$
	$s_2(\overline{\mathbf{L}}_2, \overline{\mathbf{a}}_2 = (1,0); \boldsymbol{\beta}) = \boldsymbol{\beta} \cdot [1, V_{11}^2, V_{12}, V_2^2]', \boldsymbol{\beta} = [0, 5, -2, 1].$
	$s_2(\overline{\mathbf{L}}_2, \overline{\mathbf{a}}_2 = (0,1); \boldsymbol{\beta}) = \boldsymbol{\beta} \cdot [1, V_{11}, V_{12}, V_2]', \boldsymbol{\beta} = [0, -1, 3, 1].$
	$s_2(\overline{\mathbf{L}}_2, \overline{\mathbf{a}}_2 = (0,0); \boldsymbol{\beta}) = \boldsymbol{\beta} \cdot [1, V_{11}, V_{11}V_2]', \boldsymbol{\beta} = [0, 2, 1].$
	$\text{logit}\{P(A_1 = 1 \,|\, \mathbf{L}_1;\, \boldsymbol{\alpha})\} = \boldsymbol{\alpha} \cdot [1,\, I_{11},\, I_{12},\, I_{13},\, I_{11}I_{12}\,]',\, \boldsymbol{\alpha} = [-1,\, 1,\, 1,\, -1,\, -1].$
	$\text{logit}\{P(A_2 = 1 \,|\, \overline{\mathbf{L}}_2, a_1 = 1; \boldsymbol{\alpha})\} = \boldsymbol{\alpha} \cdot [1, I_{11}, I_{12}, I_{13}, I_{11}I_{12}, I_2, I_2I_{13}]',$
	$\boldsymbol{\alpha} = [0{,}1,\, -1{,}0{,}0,\, -0.4,\, -0.3].$
	$\text{logit}\{P(A_2 = 1 \,|\, \overline{\mathbf{L}}_2, a_1 = 0; \boldsymbol{\alpha})\} = \boldsymbol{\alpha} \cdot [1, I_{11}, I_{12}, I_{13}, I_2]', \boldsymbol{\alpha} = [0, 2, 1, -1, -1].$
False	$s_2(\overline{\mathbf{L}}_2, \overline{\mathbf{a}}_2 = (1,1); \boldsymbol{\beta}) = \boldsymbol{\beta} \cdot [1, V_{11}^2]'.$
	$s_2(\overline{\mathbf{L}}_2, \overline{\mathbf{a}}_2 = (1,0); \boldsymbol{\beta}) = \boldsymbol{\beta} \cdot [1, V_{11}^2]'.$
	$s_2(\overline{\mathbf{L}}_2, \overline{\mathbf{a}}_2 = (0,1); \boldsymbol{\beta}) = \boldsymbol{\beta} \cdot [1, V_{11}, V_{12}, V_2]'.$
	$s_2(\overline{\mathbf{L}}_2, \overline{\mathbf{a}}_2 = (0,0); \boldsymbol{\beta}) = \boldsymbol{\beta} \cdot [1, V_{11}]'.$
	$\text{logit}\{P(A_1 = 1 \,|\, \mathbf{L}_1;\, \boldsymbol{\alpha})\} = \boldsymbol{\alpha} \cdot [1,\, I_{11},\, I_{12}\,]'.$
	$\text{logit}\{P(A_2 = 1 \,|\, \overline{\mathbf{L}}_2, a_1 = 1; \boldsymbol{\alpha})\} = \boldsymbol{\alpha} \cdot [1, I_{11}, I_{12}, I_{13}, I_{11}I_{12}]'.$
	$\text{logit}\{P(A_2 = 1 \,|\, \overline{\mathbf{L}}_2, a_1 = 0; \boldsymbol{\alpha})\} = \boldsymbol{\alpha} \cdot [1, I_{11}, I_2, I_{13}I_2]'.$

For simple notation, we let $I_l = I(V_l > 0)$ and $\text{logit}(r) \equiv \log\{r/(1-r)\}$. $\bar{\mathbf{a}}_2$ denotes $(a_1,\, a_2)$. $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are parameter vectors of appropriate dimensions.

Table 2
Simulation result: estimating $\mu = E(Y)$ (upper) and treatment effect (lower)

Estimator	Bias[1]	Variance	Bias[2]	IQR
$\hat{\mu}_{\text{HT}}$	−0.01	0.11	0.00	0.43
$\hat{\mu}_{\text{HT.fal}}$	−0.31	0.12	−0.29	0.44
$\hat{\mu}_{\text{OR}}$	0.00	0.04	−0.00	0.28
$\hat{\mu}_{\text{OR.fal}}$	−0.36	0.12	−0.35	0.44
$\hat{\mu}_{dr}$	0.00	0.04	−0.00	0.28
$\hat{\mu}_{dr.\text{ofal}}$	−0.02	0.11	0.00	0.45
$\hat{\mu}_{dr.\text{pfal}}$	0.00	0.04	−0.00	0.28
$\hat{\mu}_{dr.o\oplus\text{pfal}}$	−0.32	0.12	−0.30	0.45
$\hat{\mu}_{\text{HT}}$	−0.01	0.21	−0.00	0.59
$\hat{\mu}_{\text{HT.fal}}$	0.86	0.15	0.87	0.52
$\hat{\mu}_{\text{OR}}$	0.00	0.07	0.01	0.35
$\hat{\mu}_{\text{OR.fal}}$	−1.56	0.07	−1.56	0.34
$\hat{\mu}_{dr}$	0.00	0.09	0.01	0.41
$\hat{\mu}_{dr.\text{ofal}}$	−0.09	0.28	−0.06	0.63
$\hat{\mu}_{dr.\text{pfal}}$	0.00	0.08	0.01	0.39
$\hat{\mu}_{dr.o\oplus\text{pfal}}$	0.92	0.15	0.93	0.54

True parameter μ is 1 for mean parameter (upper) and 2 for treatment effect (lower). Bias[1], Bias[2], and Variance denote bias in mean, bias in median, and variance of the estimates from 1000 simulations, respectively. IQR denotes the interquartile range, that is, upper quartile (75%)–lower quartile (25%). Each simulation is based on the sample size of 500.

Description of estimators

- $\hat{\mu}_{\text{HT}}$ is the Horvitz–Thompson estimator with the correct model for π.
- $\hat{\mu}_{\text{HT.fal}}$ is the Horvitz–Thompson estimator with the false model for π.
- $\hat{\mu}_{\text{OR}}$ is the OLS estimator using the correct model for s.
- $\hat{\mu}_{\text{OR.fal}}$ is the OLS estimator using the false model for s.
- $\hat{\mu}_{dr}$ is $\hat{\mu}_{dr}$ using the correct models for π and s.
- $\hat{\mu}_{dr.\text{ofal}}$ is $\hat{\mu}_{dr}$ using the correct model for π and the false model for s.
- $\hat{\mu}_{dr.\text{pfal}}$ is $\hat{\mu}_{dr}$ using the false model for π and the correct model for s.
- $\hat{\mu}_{dr.o\oplus\text{pfal}}$ is $\hat{\mu}_{dr}$ using the false models for π and s.

Note that π denotes missingness or treatment allocation probability, and s represents the OR model for complete data.

mechanism and OR. We implemented the estimators $\hat{\mu}_{\text{HT}}, \hat{\mu}_{\text{OR}}$, and $\hat{\mu}_{dr}$ of Section 2.1 and results are reported in the upper half of Table 2.

Turn next to the treatment-effect model based on data $\mathbf{O} = \{\Delta,\ Y,\ \mathbf{V}' = (V_1,\ V_2,\ V_3)\}$. The parameter of interest is the average treatment effect. V_k ($k = 1, 2, 3$) were generated as above and Y from $N(s(\Delta, \mathbf{V}; \boldsymbol{\beta}), 1)$ where $s(\Delta, \mathbf{V}; \boldsymbol{\beta})$ is given in the same table. The parameter values used in Table 1B imply an average treatment effect μ of 2. The treatment indicator Δ was generated from the logistic model logit $\{\pi(\mathbf{V}; \boldsymbol{\alpha})\}$ provided in Table 1B. To investigate the impact of model misspecifications, we also generated the data from false models for the treatment mechanism as well as the OR model. In our false model for the treatment mechanism, we regressed the treatment indicator on two covariates: the first was one of the four covariates that actually determined treatment and the second was a noise variable independent of these four. Results are presented in the lower half of Table 2.

Reading from Table 2, we observe that in both the missing data and treatment-effect models, as expected, $\hat{\mu}_{\text{HT}}$ was virtually unbiased if we adopted a correct model for $\pi(\mathbf{V})$ but was badly biased otherwise; similarly $\hat{\mu}_{\text{OR}}$ was unbiased under a correct OR model but badly biased otherwise. In contrast, $\hat{\mu}_{dr}$ was virtually unbiased when either (or both) the PS or OR model was correct, although, as anticipated, $\hat{\mu}_{dr}$ was considerably biased when both were incorrect. Consider next the variance and interquartile range of the estimators. Because $\hat{\mu}_{\text{OR}}$ is the MLE of μ in both the missing data and treatment-effect models it should have minimum variance among all consistent estimators. What is remarkable in Table 2 is that whenever the OR model was correctly specified (so that $\hat{\mu}_{\text{OR}}$ was consistent), $\hat{\mu}_{dr}$ was nearly as efficient as the MLE $\hat{\mu}_{\text{OR}}$. Thus a very small price is paid in terms of efficiency loss by using $\hat{\mu}_{dr}$ in place of $\hat{\mu}_{\text{OR}}$, and yet, when the PS model was correct, huge benefits were obtained in terms of robustness against misspecification of the OR model.

It follows from the theory of semiparametric efficiency bounds that, when the PS model is correct, $\hat{\mu}_{dr}$ based on a correct model for the OR is asymptotically more efficient than $\hat{\mu}_{dr}$ based on an incorrect model for the OR (Scharfstein et al., 1999). These theoretical results are born out here; indeed we see that $\hat{\mu}_{dr}$ based on an incorrect model for the OR may have variance two to four times that of $\hat{\mu}_{dr}$ based on a correct model for the OR.

3. Longitudinal Models with Monotone Missing Data

Next we turn to longitudinal missing data models. We let $\mathbf{L} = \bar{\mathbf{L}}_{K+1} = (\mathbf{L}_1', \ldots, \mathbf{L}_{K+1}')'$ represent the full data obtained at times $m = 1, \ldots, K+1$. Let C be the censoring time such that if $C = m$, then $\mathbf{L}_{\text{obs}} = \bar{\mathbf{L}}_m \equiv (\mathbf{L}_1', \ldots, \mathbf{L}_m')'$ is observed and $\underline{\mathbf{L}}_{m+1} \equiv (\mathbf{L}_{m+1}', \ldots, \mathbf{L}_{K+1}')'$ is missing. That is, we observe n i.i.d. copies of $\boldsymbol{O} = (C, \bar{\mathbf{L}}_C)$. The sample space for C is $\{1, \ldots, K+1\}$, implying that $\mathbf{L}_1$ is an always observed baseline variable. We assume that the data are MAR, which implies that $\lambda(m \mid \mathbf{L}) = \lambda(m \mid \bar{\mathbf{L}}_m)$ for $m = 1, \ldots, K$ where $\lambda(m \mid \cdot) = P(C = m \mid C \geq m; \cdot)$ is the discrete hazard of censoring, that is, censoring at time m depends on the full data $\mathbf{L} = \bar{\mathbf{L}}_{K+1}$ only through the observed past $\bar{\mathbf{L}}_m$. In addition, we assume that $\lambda(m \mid \bar{\mathbf{L}}_m) < 1 - \sigma$ with probability one for all m and positive σ.

Suppose the parameter of interest μ is the mean of $Y = L_{K+1}$, which we will assume to be univariate for simplicity. Again we can represent μ as a function of the distribution of the observed data in two different ways. The first representation, analogous to the HT inverse probability-weighted (IPW) representation of Section 2.1, is $E(Y) = E(\Delta Y / \bar{\pi}_{K+1})$ where $\bar{\pi}_m = \prod_{j=1}^{m}\{1 - \lambda(j \mid \bar{\mathbf{L}}_j)\}$ is the probability of not being censored at any time less than or equal to $m - 1$, and now $\Delta = 1$ if a subject stays uncensored through the end of the study, that is, $\Delta = I(C = K + 1)$.

The second representation is most easily defined recursively. Let $H_{K+1} = Y$, then $H_K = E(H_{K+1} \mid C \geq K+1, \bar{\mathbf{L}}_K), \ldots, H_{m-1} = E(H_m \mid C \geq m, \bar{\mathbf{L}}_{m-1}), \ldots, H_1 = E(H_2 \mid C \geq 2, \bar{\mathbf{L}}_1)$. Finally $\mu = E(Y) = E(H_1)$ where H_1 is a function of the always observed $\mathbf{L}_1$. It follows that if we can specify a correct model for $\lambda(j \mid \bar{\mathbf{L}}_j)$ then we can obtain a consistent estimator of μ as the sample average of $\Delta Y/\hat{\bar{\pi}}_{K+1}$, where $\hat{\bar{\pi}}_{K+1}$ is the estimated value of $\bar{\pi}_{K+1}$ under the parametric model. Alternatively, we could correctly specify parametric regression models for $E(H_m \mid C \geq m, \bar{\mathbf{L}}_{m-1})$ for each m, and then estimate μ by the sample average of the estimated H_1's obtained from the recursive regression models for the H_m. However, the first approach will be inconsistent if the models for $\lambda(j \mid \bar{\mathbf{L}}_j)$ are misspecified, whereas the second approach will be inconsistent if the models for $E(H_m \mid C \geq m, \bar{\mathbf{L}}_{m-1})$ are misspecified. Thus it would be useful to derive DR estimators that are CAN if either the PS models for the missingness mechanism or the sequential OR models are correctly specified.

Let us introduce some theoretical background. Our goal in this section is to make inference about the finite-dimensional, say p, parameter $\boldsymbol{\mu}$ in a semiparametric or nonparametric model with likelihood $f(\mathbf{L}; \boldsymbol{\mu}, \ \boldsymbol{\theta})$ where $\boldsymbol{\mu} \in \mathcal{R}^p$ (i.e., Euclidean space) and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is an infinite-dimensional nuisance parameter. We assume that, in the absence of missing data, we would estimate $\boldsymbol{\mu}$ by solving a p-dimensional unbiased estimating function $\mathbf{0} = \sum_i d(\mathbf{L}_i; \boldsymbol{\mu})$ for some $d \in \mathcal{D} = \{d(\mathbf{L}; \boldsymbol{\mu}); E_{\boldsymbol{\mu},\boldsymbol{\theta}}\{d(\mathbf{L}; \boldsymbol{\mu})\} = \mathbf{0}$ for all $\boldsymbol{\theta}\}$. For instance, when $\boldsymbol{\mu}$ is the marginal mean of $\mathbf{L}_{K+1}$ as above, then $d(\mathbf{L}; \boldsymbol{\mu}) = (\mathbf{L}_{K+1} - \boldsymbol{\mu})$ and the model $f(\mathbf{L}; \boldsymbol{\mu}, \ \boldsymbol{\theta})$ for $\mathbf{L}$ is nonparametric in the sense that we allow the full data $\mathbf{L}$ to have an arbitrary unknown distribution restricted only by $\mathbf{L}_{K+1}$ having a finite expectation. If we were interested in the regression coefficients $E(\mathbf{L}_{K+1}\mathbf{L}_1^{'}) \ \mathrm{cov}(\mathbf{L}_1)^{-1}$ of the population least squares regression of $\mathbf{L}_{K+1}$ on $\mathbf{L}_1$, then $d(\mathbf{L}; \boldsymbol{\mu})$ could be taken to be $\mathbf{L}'_1$ $(\mathbf{L}_{K+1} - \mathbf{L}_1\boldsymbol{\mu})$ (i.e., the OLS normal equations). Henceforth, to simplify notation, we will go back to one-dimensional μ and L_{K+1}.

In the presence of missing data, even when we assume MAR, we will in general not be able to estimate μ without making further modeling assumptions due to the curse of dimensionality. Formally, we are assuming that when there are missing data, the curse of dimensionality appropriate information bound for μ is zero (Robins and Ritov, 1997). One approach to reducing the dimension is to assume a parametric submodel $f(\mathbf{L}; \ \mu, \ \boldsymbol{\beta})$ for the distribution of the full data, where $\boldsymbol{\beta} \in \mathcal{B} \subset \boldsymbol{\Theta}$ with a finite-dimensional space $\mathcal{B}$ and estimate the parameters by maximum likelihood, using the EM algorithm. An alternative approach is to specify a parametric model for $\lambda(m \mid \bar{\mathbf{L}}_m; \boldsymbol{\alpha})$ for the censoring hazard $\lambda(m \mid \bar{\mathbf{L}}_m)$ and estimate μ with the inverse probability of censoring estimators of Robins, Rotnitzky, and Zhao (1995). The first approach will be inconsistent if the model $f(\mathbf{L}; \mu, \boldsymbol{\beta})$ is incorrect and the second will be inconsistent if the model $\lambda(m \mid \bar{\mathbf{L}}_m; \boldsymbol{\alpha})$ is incorrect.

It is possible to construct an estimator of μ based on the full data estimating function $d(\mathbf{L}; \mu)$, that is, CAN in the semiparametric *union* model that assumes that (i) the data are MAR, (ii) the semi- or nonparametric model $f(\mathbf{L}; \mu, \ \boldsymbol{\theta})$ is true, and (iii) at least one (but not necessarily both) of a lower dimensional model $f(\mathbf{L}; \mu, \ \boldsymbol{\beta})$ for $\boldsymbol{\beta} \in \mathcal{B}$ or a parametric model $\lambda(m \mid \bar{\mathbf{L}}_m; \boldsymbol{\alpha})$ is correct (Scharfstein et al., 1999; Robins, 2000). Any such estimator is referred to as DR. Note (ii) will always hold if $f(\mathbf{L}; \mu, \ \boldsymbol{\theta})$ is a nonparametric model.

But, as we now show we can do even better. Specifically, we do not need to specify a parametric model $f(\mathbf{L}; \mu, \ \boldsymbol{\beta})$ for the entire joint distribution of $\mathbf{L}$. Rather, to be DR, it suffices to specify parametric models $\Psi\{s_m(\bar{\mathbf{L}}_m; \boldsymbol{\beta}_m)\}$ for the regression functions $H_m(\mu) \equiv E\{d(\mathbf{L}; \mu) \mid \bar{\mathbf{L}}_m\}$ for $m = K, \ldots, 2$ and then to estimate the regression parameters $\boldsymbol{\beta}_m$ from the observed data. This latter task we will carry out recursively for $m = K, \ldots, 2$, based on the observations that (i) by definition, $E\{H_m(\mu) \mid \bar{\mathbf{L}}_{m-1}\} = H_{m-1}(\mu)$, (ii) under the MAR assumption, $E\{H_m(\mu) \mid \bar{\mathbf{L}}_{m-1}\} = E\{H_m(\mu) \mid \bar{\mathbf{L}}_{m-1}, C \geq m\}$, and (iii) $H_m \ (\mu)$ is a function of $\bar{\mathbf{L}}_m$, which is observed whenever $C \geq m$. Robins (2000) proved that the estimator $\hat{\mu}_{dr}$ constructed in the following algorithm is CAN for μ under the union model that differs from the above union model by replacing "a lower dimensional model $f(\mathbf{L}; \mu, \ \boldsymbol{\beta})$ for $\boldsymbol{\beta} \in \mathcal{B}$" with "a parametric model $\Psi\{s_m(\bar{\mathbf{L}}_m; \boldsymbol{\beta}_m)\}$ for $E\{H_{m+1}(\mu) \mid \bar{\mathbf{L}}_m\}(m = K, \ldots, 2)$." In what follows we describe how to compute the DR estimator $\hat{\mu}_{dr}$.

1. Compute the MLE $\hat{\boldsymbol{\alpha}}$ of $\boldsymbol{\alpha}$ from the observed data.
2. Select a particular d from $\mathcal{D}$. (The choice can only affect efficiency.)
3. Set $\hat{H}_{K+1}(\mu) = d(\mathbf{L}; \mu)$.
4. Recursively, for $m = K + 1, \ldots, 2$,
   a: For subjects with $C \geq m$, specify and fit by IRLS a parametric regression model $e_{m-1}(\bar{\mathbf{L}}_{m-1}; \boldsymbol{\beta}_{m-1}, \phi_{m-1}) = \Psi\{s_{m-1}(\bar{\mathbf{L}}_{m-1}; \boldsymbol{\beta}_{m-1}) + \phi_{m-1}\bar{\pi}_{m-1}^{-1}(\hat{\boldsymbol{\alpha}})\}$ for the conditional expectation $E\{\hat{H}_m(\mu) \mid C \geq m, \bar{\mathbf{L}}_{m-1}\}$, where $s_{m-1}(\bar{\mathbf{L}}_{m-1}; \boldsymbol{\beta}_{m-1})$ is a known function with unknown parameter $\boldsymbol{\beta}_{m-1}$, Ψ^{-1} is the canonical link function of a given GLM, and $\bar{\pi}_m(\hat{\boldsymbol{\alpha}}) = \prod_{j=1}^{m}\{1 - \lambda(j \mid \bar{\mathbf{L}}_j; \hat{\boldsymbol{\alpha}})\}$. Note that $\boldsymbol{\beta}_{m-1} \equiv \boldsymbol{\beta}_{m-1} \ (\mu)$ and $\phi_{m-1} \equiv \phi_{m-1} \ (\mu)$ depend on μ.
   b: For subjects with $C \geq m - 1$, let $\hat{H}_{m-1}(\mu) = \Psi\{s_{m-1}(\bar{\mathbf{L}}_{m-1}; \hat{\boldsymbol{\beta}}_{m-1}) + \hat{\phi}_{m-1}\bar{\pi}_{m-1}^{-1}(\hat{\boldsymbol{\alpha}})\}$ be the predicted value from IRLS fit of the model where $\hat{\phi}_{m-1}$ and $\hat{\boldsymbol{\beta}}_{m-1}$ are the (joint) IRLS estimators. This means that $(\hat{\boldsymbol{\beta}}'_{m-1}, \hat{\phi}_{m-1})'$ satisfies $\mathbf{0} = \tilde{E}[I(C \geq m)$ $[\hat{H}_m(\mu) - \Psi\{s_{m-1}(\bar{\mathbf{L}}_{m-1}; \hat{\boldsymbol{\beta}}_{m-1}) + \hat{\phi}_{m-1}\bar{\pi}_{m-1}^{-1}(\hat{\boldsymbol{\alpha}})\}]$ $\{\partial s_{m-1}(\bar{\mathbf{L}}_{m-1}; \hat{\boldsymbol{\beta}}_{m-1})/\partial\boldsymbol{\beta}'_{m-1}, \bar{\pi}_{m-1}^{-1}(\hat{\boldsymbol{\alpha}})\}]$ where $\tilde{E}(V) = n^{-1}\sum_{i=1}^{n} V_i$.
5. Finally $\hat{\mu}_{dr}$ solves $0 = \sum_i \hat{H}_{1i}(\mu)$.

Remark. Depending on their functional forms, it is possible that the parametric models $\Psi\{s_m(\bar{\mathbf{L}}_m; \boldsymbol{\beta}_m), m = K, \ldots, 2\}$ are mutually incompatible in the sense that no joint distribution satisfies all $K - 1$ simultaneously, so, by definition, they must be misspecified. Even if such is the case, we do not regard this as a practical drawback, because each of the $K - 1$ models $\Psi\{s_m(\bar{\mathbf{L}}_m; \boldsymbol{\beta}_m)\}$ may still have small (approximation) bias for its estimand $H_m(\mu)$. After all, even for parametric models that are mutually compatible, the models are practically (although not logically) certain to be misspecified. Thus, the most that can be hoped for is an estimator $\hat{\mu}_{dr}$ of μ with small bias if either the models for censoring hazards $\lambda(m \mid \bar{\mathbf{L}}_m)$ or the models for the full data regression functions $H_m(\mu)$ have small (approximation) bias.

Finally, even if the model $\Psi\{s_m(\bar{\mathbf{L}}_m; \boldsymbol{\beta}_m)\}$ for $H_m(\mu)$ is misspecified, $\hat{\mu}_{dr}$ remains DR if the larger model $e_m(\bar{\mathbf{L}}_m; \boldsymbol{\beta}_m, \phi_m)$ is correct (with $\bar{\pi}_{m-1}^{-1}(\hat{\boldsymbol{\alpha}})$ replaced by its probability limit). When both the parametric models $e_m(\bar{\mathbf{L}}_m; \boldsymbol{\beta}_m, \phi_m)$ and $\lambda(m \mid \bar{\mathbf{L}}_m; \boldsymbol{\alpha})$ are correct and the model

Table 3
Simulation result: estimating $\mu = E(Y)$ in the longitudinal missing data model

Estimator	Bias[1]	Variance	Bias[2]	IQR
$\hat{\mu}_{\mathrm{HT}}$	0.02	10.31	−0.34	4.07
$\hat{\mu}_{\mathrm{HT.fal}}$	−3.31	5.10	−3.48	2.98
$\hat{\mu}_{\mathrm{OR}}$	0.02	1.73	−0.04	1.77
$\hat{\mu}_{\mathrm{OR.fal}}$	−4.82	3.34	−4.99	2.41
$\hat{\mu}_{dr}$	0.02	1.74	−0.02	1.76
$\hat{\mu}_{dr.\mathrm{ofal}}$	−0.20	9.79	−0.56	3.94
$\hat{\mu}_{dr.\mathrm{pfal}}$	0.01	1.74	−0.03	1.74
$\hat{\mu}_{dr.o\oplus\mathrm{pfal}}$	−2.24	7.82	−2.51	3.56

The true mean $\mu = E(Y)$ is 11. Bias[1], Bias[2], and Variance denote bias in mean, bias in median, and variance of the estimates from 1000 simulations, respectively. IQR denotes the interquartile range, that is, upper quartile (75%)–lower quartile (25%). Each simulation is based on the sample size of 500. See Table 2 footnote.

$f(\mathbf{L}; \mu, \boldsymbol{\theta})$ does not restrict the distribution of $\mathbf{L}$ (i.e., it is nonparametric), Robins (2000) shows our estimator $\hat{\mu}_{dr}$ will attain the semiparametric variance bound for the union model.

3.1 *A Simulation Study*

Let $\mathbf{L} = (\mathbf{L}_1', L_2, L_3)'$ represent the full data with $\mathbf{L}_1 = (V_{11}, V_{12}, V_{13})'$ and $L_3 = Y$. So the censoring variable C takes a value in $\{1, 2, 3\}$. V_{1i} $(i = 1, 2, 3)$ were generated independently from a standard normal, L_2 from $N(s_1(\mathbf{L}_1; \boldsymbol{\beta}), 1)$, and Y from $N(s_2(\bar{\mathbf{L}}_2; \boldsymbol{\beta}), 1)$ as presented in Table 1C. We are interested in estimating $\mu = E(Y) = 11$. Ignorable MAR data were created according to the missingness probabilities of $\lambda(1 \mid \mathbf{L}_1; \boldsymbol{\alpha})$ and $\lambda(2 \mid \bar{\mathbf{L}}_2; \boldsymbol{\alpha})$. Under this data configuration, L_2 and L_3 are missing for approximately 33% and 70% of subjects, respectively. Data were additionally generated from false models to explore how each estimator behaves under misspecification. $\hat{\mu}_{dr}$ was constructed based on the sequential regression analysis described above; first we regressed $L_3 = Y$ on $\bar{\mathbf{L}}_2 = (\mathbf{L}_1, L_2)$ and the estimated inverse PS $\pi_2^{-1}(\hat{\boldsymbol{\alpha}})$ jointly among those who had $C = 3$, and computed the corresponding predicted values for all subjects with $C \geq 2$. Next, these predicted values were regressed on $\mathbf{L}_1$ and $\pi_1^{-1}(\hat{\boldsymbol{\alpha}})$. Hence, the new predicted values were obtained as a function of $\mathbf{L}_1$ only, which is never missing. The average of this quantity is the final estimate. As evident in Table 3, the performance of the estimators under comparison is in agreement with the results predicted by the theory of double robustness.

4. MSM for Causal Inference

In this section, let the temporally ordered observed data be $\mathbf{O} = (\mathbf{L}_1, A_1, \mathbf{L}_2, A_2, \ldots, \mathbf{L}_K, A_K, \mathbf{L}_{K+1})$ where A_k is a treatment given at time k and $\mathbf{L}_k$ are other variables measured just prior to treatment. For easier presentation, we assume that each of A_m, L_{K+1}, and μ are all one-dimensional. Associated with each treatment history $\bar{\mathbf{a}} = (a_1, \ldots, a_K)$, there is a counterfactual random variable $\mathbf{L}_{\bar{\mathrm{a}}} = \bar{\mathbf{L}}_{\bar{\mathrm{a}},K+1}$ recording a subject's response history if treatment regime $\bar{\mathbf{a}}$ was followed. We link the counterfactual data to the observed data through the consistency assumption $\bar{\mathbf{L}}_{\bar{\mathrm{a}},m} = \bar{\mathbf{L}}_m$ if $\bar{\mathbf{A}}_{m-1} = \bar{\mathbf{a}}_{m-1}$ which states that the observed and counterfactual response through m will be equal if the observed and counterfactual treatments agree through $m - 1$. That is to say, the future cannot determine the past. We impose the assumption of sequential ignorability (i.e., no unmeasured confounders) that for all $\bar{\mathbf{a}}$ and m

$$L_{\bar{\mathrm{a}}} \coprod A_m \mid \bar{\mathbf{L}}_m, \bar{\mathbf{A}}_{m-1} = \bar{\mathbf{a}}_{m-1}, \tag{1}$$

which implies that sufficient covariates have been recorded in the $\mathbf{L}_m$ so that, as in a sequential randomized trial, the treatment A_m is independent of the counterfactuals given the observed past. Further we assume that, for all A_m in the support of A_m,

$$\text{if } f(\bar{\mathbf{A}}_{m-1}, \bar{\mathbf{L}}_m) > 0 \quad \text{then} \quad f(a_m \mid \bar{\mathbf{A}}_{m-1}, \bar{\mathbf{L}}_m) > 0, \tag{2}$$

which says that there is a positive probability that, in the observed study, any regime $\bar{\mathbf{a}}$ may be followed by a given subject.

We shall consider inference concerning the parameter μ of the marginal structural mean model (MSMM) $E(L_{\bar{\mathrm{a}},K+1}) = g(\bar{\mathbf{a}}; \mu)$ with $g(\cdot; \cdot)$ a known function. The parameter μ quantifies the effect of the regime $\bar{\mathbf{a}}$ on the mean of L_{K+1}. The MSMM is a semiparametric model characterized by the restriction that $E\{d(L_{\bar{\mathrm{a}}}, \bar{\mathbf{a}}; \mu)\} = 0$ for $d \in \mathcal{D}$, with $\mathcal{D} = \{d(L_{\bar{\mathrm{a}}}, \bar{\mathbf{a}}; \mu) = d^*(\bar{\mathbf{a}})\{L_{\bar{\mathrm{a}},K+1} - g(\bar{\mathbf{a}}; \mu)\}; d^*(\cdot)\ \textit{arbitrary}\}$. If the assumption that $E\{d(L_{\bar{\mathrm{a}}}, \bar{\mathbf{a}}; \mu)\} = 0$ does not restrict the distribution of the $L_{\bar{\mathrm{a}}}$, we say our MSMM is saturated (i.e., the observed data model is nonparametric). Under sequential ignorability, an MSMM induces a semiparametric model for the observed data with likelihood $f(\mathbf{O}; \mu, \boldsymbol{\theta}, \boldsymbol{\rho}) = \prod_{m=1}^{K+1} f(L_m \mid \bar{\mathbf{L}}_{m-1}, \bar{\mathbf{A}}_{m-1}; \mu, \boldsymbol{\theta}) \times \prod_{m=1}^{K} f(A_m \mid \bar{\mathbf{A}}_{m-1}, \bar{\mathbf{L}}_m; \boldsymbol{\rho})$, where $f(L_m \mid \bar{\mathbf{L}}_{m-1}, \bar{\mathbf{A}}_{m-1}; \mu, \boldsymbol{\theta})$ and $f(A_m \mid \bar{\mathbf{A}}_{m-1}, \bar{\mathbf{L}}_m; \boldsymbol{\rho})$ are densities with respect to some dominating measures ν_l and ν_a, respectively, where $(\mu, \boldsymbol{\theta})$ and $\boldsymbol{\rho}$ are variation independent, and $\boldsymbol{\theta}$ and $\boldsymbol{\rho}$ are (often infinite-dimensional) nuisance parameters. Robins (2000) notes that, by sequential ignorability, the observed data model is characterized by the restriction for all functions $d \in \mathcal{D}$

$$E\{d\left(\bar{\mathbf{L}}_{K+1}, \bar{\mathbf{A}}_K; \mu\right)/\bar{\pi}_K\} = 0, \tag{3}$$

where now $\bar{\pi}_m = \prod_{j=1}^{m} f(A_j \mid \bar{\mathbf{L}}_j, \bar{\mathbf{A}}_{j-1})$.

In order to reduce dimensionality, we could specify parametric submodels $f(\mathbf{l}_{\bar{\mathrm{a}},m} \mid \bar{\mathbf{l}}_{\bar{\mathrm{a}},m-1}; \mu, \boldsymbol{\beta})$ where $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are finite-dimensional parameters. Robins (2000) shows that, by sequential ignorability, $f(\mathbf{l}_{\bar{\mathrm{a}},m} \mid \bar{\mathbf{l}}_{\bar{\mathrm{a}},m-1}) = f(\mathbf{l}_m \mid \bar{\mathbf{l}}_{m-1}, \bar{\mathbf{a}}_{m-1})$ so in terms of the observables we are modeling $f(\mathbf{l}_m \mid \bar{\mathbf{l}}_{m-1}, \bar{\mathbf{a}}_{m-1})$ by the model $f(\mathbf{l}_m \mid \bar{\mathbf{l}}_{m-1}, \bar{\mathbf{a}}_{m-1}; \mu, \boldsymbol{\beta})$. We could then estimate $(\mu, \boldsymbol{\beta})$ by maximum likelihood since the MLE does not depend on the treatment mechanism $f(a_m \mid \bar{\mathbf{l}}_m, \bar{\mathbf{a}}_{m-1})$. An alternative approach is to specify a parametric model $f(a_m \mid \bar{\mathbf{l}}_m, \bar{\mathbf{a}}_{m-1}; \boldsymbol{\alpha})$ for $f(a_m \mid \bar{\mathbf{l}}_m, \bar{\mathbf{a}}_{m-1})$ and estimate μ with the inverse probability of treatment estimators of Hernán, Brumback, and Robins (2001), as these estimators do not require models for $f(\mathbf{l}_m \mid \bar{\mathbf{l}}_{m-1}, \bar{\mathbf{a}}_{m-1})$ (Robins et al., 1995). The first approach will be inconsistent if the model $f(\mathbf{l}_m \mid \bar{\mathbf{l}}_{m-1}, \bar{\mathbf{a}}_{m-1}; \mu, \boldsymbol{\beta})$ is incorrect and the second will be inconsistent if the model $f(a_m \mid \bar{\mathbf{l}}_m, \bar{\mathbf{a}}_{m-1}; \boldsymbol{\alpha})$ is incorrect.

Robins (2000) constructs a CAN estimator of μ in the semiparametric *union* model that assumes (1), (2), and the MSMM model are true (so (3) holds), and at least one of the two finite-dimensional submodels indexed by $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ is correct as well.

Again we can do better. Specifically, we need not parametrically model all of the joint law of $L_{\bar{\mathrm{a}}}$ through the models $f(\mathbf{l}_{\bar{\mathrm{a}},m} \mid \bar{\mathbf{l}}_{\bar{\mathrm{a}},m-1})$. Rather we need only specify parametric models for the counterfactual regression functions $m = K, \ldots, 2, 1$:

$$h_m(\bar{\mathbf{l}}_m, \bar{\mathbf{a}}_m; \mu) \equiv \int \cdots \int E\{d(L_{\bar{\mathrm{a}}}, \bar{\mathbf{a}}; \mu) \,|\, \bar{\mathbf{L}}_{\bar{a},m} = \bar{\mathbf{l}}_m\} \prod_{j=m+1}^{K} d\nu(a_j),$$

where ν is counting measure if A_j is discrete and Lebesgue measure if A_j is continuous, and $\boldsymbol{\beta} \equiv \boldsymbol{\beta}\ (\mu)$. Note the integral over the measure of the a_j $(j > m)$ is required to make $h_m(\bar{\mathbf{l}}_m, \bar{\mathbf{a}}_m; \mu)$ a function $\bar{\mathbf{a}}$ only through $\bar{\mathbf{a}}_m$. Robins (2000) shows that if we let $t_{K+1}(\bar{\mathbf{l}}_{K+1}, \bar{\mathbf{a}}_K; \mu) = d(\bar{\mathbf{l}}_{K+1}, \bar{\mathbf{a}}_K; \mu)$ and $t_{m+1}(\bar{\mathbf{l}}_{m+1}, \bar{\mathbf{a}}_m; \mu) = \int h_{m+1}(\bar{\mathbf{l}}_{m+1}, \bar{\mathbf{a}}_{m+1}; \mu) d\nu(a_{m+1}) < \infty$ for $m = K, \ldots, 1$, then, under sequential ignorability, $H_m(\mu) = h_m(\bar{\mathbf{L}}_m, \bar{\mathbf{A}}_m; \mu)$ equals $E\{T_{m+1}(\mu) \,|\, \bar{\mathbf{L}}_m, \bar{\mathbf{A}}_m\}$ where $T_{m+1}(\mu) = t_{m+1}(\bar{\mathbf{L}}_{m+1}, \bar{\mathbf{A}}_m; \mu)$. Thus we can fit a model for the counterfactual regression $h_m(\bar{\mathbf{l}}_m, \bar{\mathbf{a}}_m; \mu)$ by fitting a model for the observed data regression $E\{T_{m+1}(\mu) \,|\, \bar{\mathbf{L}}_m, \bar{\mathbf{A}}_m\}$.

Robins (2000) proves that the estimator constructed in the following algorithm is CAN for μ under the union model that differs from the above union model by replacing the model $f(\mathbf{l}_{\bar{\mathrm{a}},m} \mid \bar{\mathbf{l}}_{\bar{\mathrm{a}},m-1}; \mu)$ with parametric models for $h_m(\bar{\mathbf{L}}_m, \bar{\mathbf{A}}_m; \mu) = E\{T_{m+1}(\mu) \,|\, \bar{\mathbf{L}}_m, \bar{\mathbf{A}}_m\}$. When all our parametric submodels are correct both for treatment and for the $h_m(\bar{\mathbf{L}}_m, \bar{\mathbf{A}}_m; \mu)$ and our MSMM is saturated, the DR estimator $\hat{\mu}_{dr}$ will attain the semiparametric variance bound for the union model.

1. Compute the MLE $\hat{\boldsymbol{\alpha}}$ of $\boldsymbol{\alpha}$ from the observed data.
2. Select a particular d from $\mathcal{D}$. (The choice can only affect efficiency.)
3. Set $\hat{T}_{K+1}(\mu) = d(\bar{\mathbf{L}}_{K+1}, \bar{\mathbf{A}}_K; \mu)$.
4. Recursively, for $m = K + 1, \ldots, 2$,
   a: Specify and fit by IRLS a parametric regression model $h_{m-1}(\bar{\mathbf{L}}_{m-1}, \bar{\mathbf{A}}_{m-1}; \boldsymbol{\beta}_{m-1}, \phi_{m-1}) = \Psi\{s_{m-1}(\bar{\mathbf{L}}_{m-1}, \bar{\mathbf{A}}_{m-1}; \boldsymbol{\beta}_{m-1}) + \phi_{m-1}\bar{\pi}_{m-1}^{-1}(\hat{\boldsymbol{\alpha}})\}$ for the conditional expectation $E\{\hat{T}_m(\mu) \mid \bar{\mathbf{A}}_{m-1}, \bar{\mathbf{L}}_{m-1}\}$, where $s_{m-1}(\bar{\mathbf{L}}_{m-1}, \bar{\mathbf{A}}_{m-1}; \boldsymbol{\beta}_{m-1})$ is a known function with the unknown parameter $\boldsymbol{\beta}_{m-1}$, Ψ is the canonical link function of a given GLM, and $\bar{\pi}_m(\hat{\boldsymbol{\alpha}}) = \prod_{j=1}^{m} f(A_j \mid \bar{\mathbf{L}}_j, \bar{\mathbf{A}}_{j-1}; \hat{\boldsymbol{\alpha}})$. Implicitly, $\boldsymbol{\beta}_{m-1} \equiv \boldsymbol{\beta}_{m-1}\ (\mu)$ and $\phi_{m-1} \equiv \phi_{m-1}\ (\mu)$ depend on μ.
   b: Let $\hat{H}_{m-1}(\mu) \equiv \hat{h}_{m-1}(\bar{\mathbf{L}}_{m-1}, \bar{\mathbf{A}}_{m-1}; \mu) = \Psi\{s_{m-1}(\bar{\mathbf{L}}_{m-1}, \bar{\mathbf{A}}_{m-1}; \hat{\boldsymbol{\beta}}_{m-1}) + \hat{\phi}_{m-1}\bar{\pi}_{m-1}^{-1}(\hat{\boldsymbol{\alpha}})\}$ be the predicted value from IRLS fit of the model. This implies that $(\hat{\boldsymbol{\beta}}'_{m-1}, \hat{\phi}_{m-1})'$ is a solution of $\mathbf{0} = \tilde{E}[[\hat{T}_m(\mu) - \Psi\{s_{m-1}(\bar{\mathbf{L}}_{m-1}, \bar{\mathbf{A}}_{m-1}; \hat{\boldsymbol{\beta}}_{m-1}) + \hat{\phi}_{m-1}\bar{\pi}_{m-1}^{-1}(\hat{\boldsymbol{\alpha}})\}]\{\partial s(\bar{\mathbf{L}}_{m-1}; \hat{\boldsymbol{\beta}}_{m-1})/\partial\boldsymbol{\beta}'_{m-1}, \bar{\pi}_{m-1}^{-1}(\hat{\boldsymbol{\alpha}})\}]$ where $\tilde{E}(V) = n^{-1}\sum_{i=1}^{n} V_i$.
   c: Here, for $m = K, \ldots, 1$, we have recursively defined $\hat{T}_m(\mu) = \int \hat{h}_m(\bar{\mathbf{L}}_m, \bar{\mathbf{A}}_m; \mu) d\nu_a(A_m)$.
5. Finally $\hat{\mu}_{dr}$ solves $0 = \sum_i \hat{T}_{1i}(\mu)$.

In the Appendix, we will show that this sequential regression estimator $\hat{\mu}_{dr}$ is indeed an AIPW estimator.

4.1 *A Simulation Study*

As a last simulation, we show how to implement the above algorithm and illustrate the finite sample efficiency and robustness of $\hat{\mu}_{dr}$. The following longitudinal data $\boldsymbol{O} = \{\mathbf{L}_1 = (V_{11}, V_{12}, V_{13}), A_1, L_2, A_2, L_3 = Y\}$ were generated as follows. We supposed that there exists a counterfactual outcome $Y_{\bar{\mathrm{a}}}$ associated with treatment history $\bar{\mathbf{a}} = (a_1, a_2)$. Let V_{1i} $(i = 1, 2, 3)$ be distributed as independent $N(0, 1)$ and treatment A_1 was assigned according to the probability mass functions $P(A_1 = 1 \,|\, \mathbf{L}_1; \boldsymbol{\alpha})$. Then generate $L_{a_1,2} \,|\, \mathbf{L}_1$ from $N(s_1(\mathbf{L}_1, A_1 = a_1; \boldsymbol{\beta}), 1)$. Next A_2 was assigned according to the probability mass function $P(A_2 = 1 \,|\, A_1 = a_1, \bar{\mathbf{L}}_2; \boldsymbol{\alpha})$. Finally, the counterfactual outcome $Y_{\bar{\mathrm{a}}} \,|\, (L_{\bar{\mathrm{a}},2}, \mathbf{L}_1)$ was distributed as $N(s_2(\bar{\mathbf{L}}_2, \bar{\mathbf{A}}_2 = \bar{\mathbf{a}}_2; \boldsymbol{\beta}), 1)$ (see Table 1D).

Under the underlying correct models, Robins's (1986) G-computation algorithm gives $E(Y_{1,1}) = 11$, $E(Y_{1,0}) = 19$, $E(Y_{0,1}) = 0$, and $E(Y_{0,0}) = -1$ where, for example, $Y_{1,1}$ is $Y_{\bar{\mathrm{a}}}$ with $\bar{\mathbf{a}} = (1, 1)$. Thus, equivalently, $E(Y_{\bar{\mathrm{a}}}) = E(Y_{a_1,a_2}) = -1 + 20a_1 + a_2 - 9a_1a_2$ so all main effects and interactions of the MSM are nonzero. To construct a biologically interesting nonsaturated model, we considered a transformation $Y^*_{\bar{\mathrm{a}}} = Y_{\bar{\mathrm{a}}} + 9a_1a_2 + 19a_2$ of $Y_{\bar{\mathrm{a}}}$. Then $Y^*_{\bar{\mathrm{a}}}$ satisfies $E(Y^*_{\bar{\mathrm{a}}}) = -1 + 20\ cum(\bar{\mathbf{a}}_2)$, where $cum(\bar{\mathbf{a}}_2) = a_1 + a_2$. Such a model is typical in occupational health studies where it is often hypothesized that the exposure effect only depends on cumulative exposure. Finally we took the observed outcome $Y = L_3$ to be $Y^*_{\bar{\mathbf{A}}}$, that is, $Y^*_{\bar{\mathrm{a}}}$ with $\bar{\mathbf{a}}$ evaluated at the observed treatment $\bar{\mathbf{A}} = (A_1, A_2)$. It follows that $\boldsymbol{\mu} \equiv (\mu_0, \mu_1)' = (-1, 20)'$ is the true parameter to be estimated, where μ_0 is the intercept and μ_1 is the slope.

A naive OLS estimator, $\hat{\boldsymbol{\mu}}_{\text{assoc}}$ was obtained by regressing Y^* on $cum(\bar{\mathbf{A}}_2)$. The OLS estimator will converge to a value that differs from the causal parameter $\boldsymbol{\mu}$ of the MSM in the presence of confounding by $(\mathbf{L}_1, L_2)$. The simple HT-like IPW estimator, $\hat{\boldsymbol{\mu}}_{\text{HT}}$, of the MSM parameter is defined as the solution of $\mathbf{0} = \sum_i d(\bar{\mathbf{L}}_{3i}, \bar{\mathbf{A}}_{2i}; \boldsymbol{\mu})/\bar{\pi}_{2i}(\hat{\boldsymbol{\alpha}})$, where $d(\bar{\mathbf{L}}_3, \bar{\mathbf{A}}_2; \boldsymbol{\mu}) = \{Y^* - \mu_0 - \mu_1 cum(\bar{\mathbf{A}}_2)\}f(A_1) f(A_2 \,|\, A_1)\{1, cum(\bar{\mathbf{A}}_2)\}'$ and $\bar{\pi}_2(\boldsymbol{\alpha}) = \prod_{j=1}^{2} f(A_j \,|\, \bar{\mathbf{L}}_j, \bar{\mathbf{A}}_{j-1}; \boldsymbol{\alpha})$, and $\hat{\boldsymbol{\alpha}}$ is estimated by maximum likelihood. Both correct and incorrect models $f(A_j \,|\, \bar{\mathbf{L}}_j, \bar{\mathbf{A}}_{j-1}; \boldsymbol{\alpha})$ were tried. This estimator is consistent under the correct PS (i.e., treatment) effect model but neither robust to its misspecification nor efficient. The estimator $\hat{\boldsymbol{\mu}}_{\text{OR}}$ simply replaced the unknown conditional probabilities in the G-computation formula with their estimates based on the fit of parametric models for $Y = Y^*_{\bar{\mathbf{A}}}$ given $(A_2, L_2, A_1, \mathbf{L}_1)$, L_2 given $(A_1, \mathbf{L}_1)$, and $\mathbf{L}_1$ using the OR models in Table 1D. Both correct and incorrect models were tried. The estimator is efficient under correct specification but inconsistent otherwise. Finally, the DR estimator was computed using the above algorithm, with the required regressions based on both correct and incorrect models.

Simulation results are summarized in Table 4. $\hat{\boldsymbol{\mu}}_{\text{assoc}}$, which failed to account for confounding, was severely biased. Again $\hat{\boldsymbol{\mu}}_{\text{OR}}$ was consistent and had the smallest variance under the correct regression models but was substantially biased under their misspecification. Similarly, $\hat{\boldsymbol{\mu}}_{\text{HT}}$ was considerably biased with the incorrect model for the probability of treatment. Our $\hat{\boldsymbol{\mu}}_{dr}$ performed best, being unbiased if either the

Table 4
Simulation result: estimating regression parameters in the MSM: intercept (upper) and slope (lower)

Estimator	Bias[1]	Variance	Bias[2]	IQR
$\hat{\mu}_{\text{assoc}}$	−1.24	0.92	−1.22	1.24
$\hat{\mu}_{\text{HT}}$	−0.02	0.73	−0.00	1.09
$\hat{\mu}_{\text{HT.fal}}$	−0.31	1.11	−1.26	1.30
$\hat{\mu}_{\text{OR}}$	0.00	0.05	0.00	0.30
$\hat{\mu}_{\text{OR.fal}}$	−1.66	0.89	−1.60	1.24
$\hat{\mu}_{dr}$	0.00	0.06	0.01	0.32
$\hat{\mu}_{dr.\text{ofal}}$	0.10	1.33	0.12	1.40
$\hat{\mu}_{dr.\text{pfal}}$	0.00	0.05	0.00	0.30
$\hat{\mu}_{dr.o\oplus\text{pfal}}$	−1.04	1.01	−0.99	1.28
$\hat{\mu}_{\text{assoc}}$	1.97	2.52	1.85	2.09
$\hat{\mu}_{\text{HT}}$	0.03	1.81	−0.10	1.66
$\hat{\mu}_{\text{HT.fal}}$	1.89	3.31	1.74	2.18
$\hat{\mu}_{\text{OR}}$	0.01	0.49	−0.04	0.90
$\hat{\mu}_{\text{OR.fal}}$	1.35	1.61	1.24	1.70
$\hat{\mu}_{dr}$	0.01	0.50	−0.03	0.85
$\hat{\mu}_{dr.\text{ofal}}$	−0.04	1.80	−0.13	1.69
$\hat{\mu}_{dr.\text{pfal}}$	0.00	0.49	−0.04	0.90
$\hat{\mu}_{dr.o\oplus\text{pfal}}$	1.41	1.74	1.29	1.73

The true intercept and slope parameters are −1 and 20, respectively, in a simple linear regression model. $\hat{\mu}_{\text{assoc}}$ is the OLS estimator from the simple linear regression model ignoring relevant confounders. Bias[1], Bias[2], and Variance denote bias in mean, bias in median, and variance for the estimates from 1000 simulations, respectively. IQR denotes the interquartile range, that is, upper quartile (75%)–lower quartile (25%). Each simulation is based on the sample size of 500. See Table 2 footnote.

counterfactual OR model or the PS model was correct and being nearly as efficient as $\hat{\boldsymbol{\mu}}_{\text{OR}}$ when both were correct.

5. An Example: The ENRICHD Study

We applied the method for time-independent treatment effects presented in Section 2.2 to the recently conducted ENRICHD (Enhancing Recovery in Coronary Heart Disease) trial (2003). The trial protocol randomized postmyocardial infarction patients suffering from depression or social isolation to a cognitive behavior therapy program or to usual care. Primary endpoints were time to reinfarction or death. In both arms an antidepressant(s) was allowed when prescribed by a physician. Thus antidepressant therapy was a nonrandomized concomitant treatment. Here, we analyze the effect of postrandomization antidepressant drug therapy on a secondary endpoint, the Beck Depression Inventory (BDI) measured at 6 months from randomization exclusively for statistical illustration. In the analysis we coded postrandomization antidepressant drug therapy as 1 if an antidepressant were prescribed any time in the first 6 months and 0 otherwise. We restricted the analysis to those who were depressed at baseline with nonmissing BDI scores. We adjusted only for baseline variables.

We report a naive crude estimate of the treatment effect equal to difference in mean BDI among antidepressant users ($N = 206$) and nonusers ($N = 1126$). We also report the HT, OR, and DR estimators of Section 2.2.

To select our OR model, we used the following algorithm. We considered as the potential regressors in a linear regression model, main effects of antidepressant use, treatment arm, baseline BDI, and 23 remaining baseline characteristics (as shown in Table 1 of ENRICHD, 2003) and their two-way interactions with BDI score at baseline, antidepressant use, and treatment arm. We used backward elimination to simplify our multivariate model. At each step, the factor with the largest p-value was dropped one at a time until all factors are significant with a cutpoint of p-value = 0.06. In this process, the main effect terms corresponding to each significant interaction were retained. Our final OR model included the following factors: main effect terms for antidepressant use, treatment arm, baseline BDI, age, education level, perceived stress score (PSS), perceived social support scale (PSSS), comorbidity index, vasodilator use, diabetes, cerebrovascular disease, and interaction terms for antidepressant by education, antidepressant by BDI score, BDI by comorbidity, BDI by age, BDI by diabetes, treatment by age, and treatment by PSS.

A completely analogous algorithm was used to build our final PS model but with logistic replacing linear regression. The final PS model included BDI at randomization, treatment arm, age, race, comorbidity score, creatinine, and an interaction of BDI and creatinine.

Remark. We chose this particular model selection methodology not because we believe it to be optimal, but rather because we believe it approximates current practice. Indeed the issue of how to select an optimal PS model is difficult. Specifically, one cannot simply choose a very large model that includes all main effects and all possible interactions to many orders. This reflects the fact that the issue is not only bias but variance. Specifically, in order for the HT estimator to be CAN or for our DR estimator to be CAN when the OR model is misspecified, the estimated PS must converge to the true score at rate $n^{1/4}$ or better. But the rate of convergence depends both on the degree of model misspecification (approximation bias) and on the variance of the estimated PS. To control the variance, the number of parameters in the PS model can increase no faster than the square root of the sample size n. Indeed the question of how to optimally choose a PS model that optimally trades off bias with variance is beyond the scope of this article.

Results are summarized in Table 5. Standard errors and the corresponding 95% confidence intervals were obtained from 1000 nonparametric bootstrap samples. The estimates $\hat{\mu}_{\text{HT}}$, $\hat{\mu}_{\text{OR}}$, and $\hat{\mu}_{dr}$ varied between 2.40–2.76, a maximum

Table 5
The ENRICHD study data analysis: estimating the effect of antidepressants on BDI

Estimator	Mean (SE)	95% CI
$\hat{\mu}_{naive}$	3.32 (0.73)	(1.91, 4.71)
$\hat{\mu}_{\text{HT}}$	2.40 (0.71)	(0.96, 3.81)
$\hat{\mu}_{\text{OR}}$	2.64 (0.61)	(1.47, 3.90)
$\hat{\mu}_{dr}$	2.76 (0.68)	(1.38, 4.12)

BDI stands for Beck Depression Inventory. SE and CI denote standard error and confidence interval, respectively. Treatment effect is defined by the difference in mean BDI score between the treated group and the untreated group. $\hat{\mu}_{naive}$ is computed as a (unweighted) sample average as observed. SE and CI are estimated from 1000 bootstrap samples. See Table 2 footnote.

difference of approximately 1/2 a standard deviation. In contrast, the unadjusted crude "naive estimate" was 3.32. The values 1.08 and 1.5 of the test statistics $|(\hat{\mu}_{dr} - \hat{\mu}_{\mathrm{HT}})/\hat{\tau}_{dr-HT}|$ and $|(\hat{\mu}_{dr} - \hat{\mu}_{\mathrm{OR}})/\hat{\tau}_{dr-\mathrm{OR}}|$ offer no evidence against the null hypotheses that the PS and OR models were correctly specified. The most parsimonious summary of the evidence appears to be that OR and PS models are nearly correct and thus, in this data set, $\hat{\mu}_{dr}, \hat{\mu}_{\mathrm{HT}}$, and $\hat{\mu}_{\mathrm{OR}}$ may have nearly fully corrected for confounding by the measured baseline variables. In contrast the naive crude estimator is biased upward due to uncontrolled confounding by the measured variables.

The 95% confidence intervals for μ constructed from $\hat{\mu}_{dr}, \hat{\mu}_{\mathrm{HT}}$, and $\hat{\mu}_{\mathrm{OR}}$ based on the bootstrap standard errors exclude the null value of 0, suggesting an adverse effect of antidepressants on BDI. However, randomized trials of antidepressant therapy in patients with coronary heart disease have previously shown a beneficial effect of these drugs on BDI. The most likely explanation for the discrepancy between our findings and these previous findings is time-dependent confounding by depressive symptoms and BDI scores. For example, subjects in the cognitive behavior therapy arm were given a repeat BDI test at 5 weeks postrandomization. If the repeat test showed less than 50% reduction in score from baseline, the subject was referred to a psychiatrist for consideration of antidepressant therapy. Thus BDI test at 5 weeks is a confounder as it predicts both antidepressant treatment and the study endpoint, BDI score at 6 months. At time of randomization, 4.8% of the usual care arm and 9.1% of the intervention arm were placed on antidepressants. By 6 months, the cumulative rates of antidepressant use had increased to 13.4% in the usual care and to 20.5% in the intervention arm. Since in our analysis we only adjusted for baseline variables, we did not eliminate confounding caused by time-varying determinants (such as the BDI score at 5 weeks and clinical symptoms of depression) of postrandomization antidepressant therapy.

6. Discussion

In this article we have considered both the theoretical and, via simulation, the practical advantages of DR estimators in four different epidemiologic settings. A DR estimator offers the analyst two chances to make nearly correct inference about the parameter of interest, a crucial property not shared by standard IPW estimators or standard likelihood-based estimators. Although a DR estimator will be less efficient than an MLE when the likelihood model is correct, nonetheless, in our opinion, the additional robustness of the DR estimator to misspecification argues for its routine use. Furthermore, in our simulation studies, we have seen that the use of DR estimators may incur surprisingly little efficiency loss compared to MLEs when both are consistent, and yet provide major improvements in robustness when the likelihood model is incorrect.

Although the DR estimators are attractive and exist in the four models we studied in this article, in many models they do not exist and even when they do, their construction may not be obvious. Robins and Rotnitzky (2001) characterized necessary and sufficient conditions for the existence of DR estimators in a number of models including various nonignorable missing data models and the semiparametric regression model.

Acknowledgements

The authors thank Professor Butch Tsiatis for reading an earlier draft of this article, and Professor Diane Catellier and the other ENRICHD Investigators for providing the permission to use the ENRICHD data. The comments from two reviewers greatly improved this article.

References

The ENRICHD Investigators. (2003). Effects of treating depression and low perceived social support on clinical events after myocardial infarction: The Enhancing Recovery in Coronary Heart Disease Patients (ENRICHD) Randomized Trial. *Journal of the American Medical Association* **289,** 3106–3116.

Hernán, M., Brumback, B., and Robins, J. M. (2001). Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *Journal of the American Statistical Association* **96,** 440–448.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47,** 663–685.

Lipsitz, S. R., Ibrahim, J. G., and Zhao, L. P. (1999). Weighted estimating equation for missing covariate data with properties similar to maximum likelihood. *Journal of the American Statistical Association* **94,** 1147–1160.

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine* **23,** 2937–2960.

Neugebauer, R. and Van der Laan, M. J. (2005). Why prefer double robust estimates? *Journal of Statistical Planning and Inference* **129,** 405–426.

Robins, J. M. (1986). A new approach to causal inference in mortality studies with sustained exposure periods—Application to control of the healthy worker survivor effect. *Mathematical Modelling* **7,** 1393–1512.

Robins, J. M. (1999). Marginal structural models versus structural nested models as tools for causal inference. In *Statistical Models in Epidemiology: The Environment and Clinical Trials*, M. E. Halloran and D. Berry (eds), 95–134. New York: Springer-Verlag.

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

Robins, J. M. (2002). Commentary on "Using inverse weighting and predictive inference to estimate the effects of time-varying treatments on the discrete-time hazard, by Dawson and Lavori. *Statistics in Medicine*, **21,** 1663–1680.

Robins, J. M. and Ritov, Y. (1997). A curse of dimensionality appropriate (CODA) asymptotic theory for semiparametric models. *Statistics in Medicine* **16,** 285–319.

Robins, J. M. and Rotnitzky, A. (2001). Comment on the Bickel and Kwon article, "On double robustness." *Statistica Sinica* **11,** 920–936.

Robins, J. M., Rotnitzky, A., and Zhao L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89,** 846–866.

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

Robins, J. M., Rotnitzky, A., and Van der Laan, M. J. (2000). Comment on the Murphy and Van der Vaart article, "On profile likelihood." *Journal of the American Statistical Association* **95,** 431–435.

Robins, J. M., Rotnitzky, A., and Bonetti, M. (2001). Discussion of the Frangakis and Rubin article, "Addressing an idiosyncrasy in estimating survival curves using double sampling in the presence of self-selected right censoring." *Biometrics* **57,** 343–347.

Rosenbaum, P. (1987). Model-based direct adjustment. *Journal of the American Statistical Association* **82,** 387–394.

Rotnitzky, A., Robins, J. M., and Scharfstein, D. O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association* **93,** 1321–1339.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association* **94,** 1096–1120 (with Rejoinder, 1135–1146).

Van der Laan, M. J. and Robins, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality.* New York: Springer-Verlag.

Received February 2004. *Revised November* 2004.
Accepted February 2005.

Appendix

Equivalence between Sequential Regression Estimators and AIPW Estimators

We will show that the sequential regression estimator $\hat{\mu}_{dr}$ in Section 4 is precisely an orthogonal AIPW estimator. Define, following Robins (1999), the AIPW estimating function $U_{AIPW}(\mu) = d(\bar{\mathbf{L}}_{K+1}, \bar{\mathbf{A}}_K; \mu)/\bar{\pi}_K(\hat{\boldsymbol{\alpha}}) - \sum_{m=1}^{K} [c(m, \bar{\mathbf{L}}_m, \bar{\mathbf{A}}_m; \mu) - E_{\hat{\boldsymbol{\alpha}}}\{c(m, \bar{\mathbf{L}}_m, \bar{\mathbf{A}}_m; \mu) \mid \bar{\mathbf{L}}_m, \bar{\mathbf{A}}_{m-1}\}]$, where the choice $c(m, \bar{\mathbf{L}}_m, \bar{\mathbf{A}}_m; \mu) = \hat{H}_m(\mu)/\bar{\pi}_m(\hat{\boldsymbol{\alpha}})$ makes $U_{AIPW}(\mu)$ orthogonal. We will show that $\sum_i U_{AIPW,i}(\mu) = \sum_i \hat{T}_{1i}(\mu)$ for all μ in the proposition below.

It follows from Robins (1999) that, because $\hat{\mu}_{dr}$ is an AIPW estimator, it is regular asymptotically linear when the model $f(a_m \mid \bar{\mathbf{L}}_m, \bar{\mathbf{a}}_{m-1}; \boldsymbol{\alpha})$ is correct. The key step in showing that $\hat{\mu}_{dr}$ is regular asymptotically linear when the model for $h_m(\bar{\mathbf{l}}_m, \bar{\mathbf{a}}_m; \mu)$ is correct is that $\hat{\phi}_m$ converges to 0 for each m. It then immediately follows that $n^{-1}\sum_i \hat{T}_{1i}(\mu)$ converges to $E\{d(\bar{\mathbf{L}}_{K+1}, \bar{\mathbf{A}}_K; \mu)/\bar{\pi}_K\}$. Efficiency results from the fact that, if the MSMM is saturated, then at laws where both parametric models are true, the tangent space (i.e., the closed linear span of scores for correctly specified regular parametric submodels) for the union model is all random variables with finite variance. This implies that all regular estimators have the same efficient influence function.

Moreover, the monotone missing data model in Section 3 is actually a special case of the MSMM model of Section 4. To see why, we show a correspondence between the two models by recoding our monotone missing data model via the following: define $A_m = 1$ if $C > m$ and $A_m = 0$ otherwise. Then we can write $\Delta d(\mathbf{L}; \mu)$ as $I(\bar{\mathbf{A}}_K = \mathbf{1})d(\mathbf{L}; \mu)$ where $\mathbf{1}$ is the vector with all components equal to 1. Then define $d(\bar{\mathbf{L}}_{K+1}, \bar{\mathbf{A}}_K; \mu)$ to be $I(\bar{\mathbf{A}}_K = \mathbf{1})d(\mathbf{L}; \mu)$ and the correspondence is complete. In this special case, $\hat{H}_m(\mu) = \hat{T}_m(\mu)$.

Proposition A.1: $\sum_i U_{AIPW,i}(\mu) = \sum_i \hat{T}_{1i}(\mu)$.

Proof: Let us set $d(\bar{\mathbf{L}}_{K+1}, \bar{\mathbf{A}}_K; \mu) = \hat{T}_{K+1}(\mu)$ and $c(m, \bar{\mathbf{L}}_m, \bar{\mathbf{A}}_m; \mu) = \hat{H}_m(\mu)\bar{\pi}_m^{-1}(\hat{\boldsymbol{\alpha}})$. The relationships of $E_{\hat{\boldsymbol{\alpha}}}\{c(m, \bar{\mathbf{L}}_m, \bar{\mathbf{A}}_m; \mu) \mid \bar{\mathbf{L}}_m, \bar{\mathbf{A}}_{m-1}\} = \hat{T}_m(\mu)\bar{\pi}_{m-1}^{-1}(\hat{\boldsymbol{\alpha}})$ and $\bar{\pi}_0(\hat{\boldsymbol{\alpha}}) = 1$ lead to

$$\begin{aligned}
&\sum_{i=1}^{n} U_{AIPTW,i}(\mu) \\
&= \sum_{i=1}^{n}\left[\frac{d\left(\bar{\mathbf{L}}_{K+1i}, \bar{\mathbf{A}}_{Ki}; \mu\right)}{\bar{\pi}_{Ki}(\hat{\boldsymbol{\alpha}})} - \sum_{m=1}^{K}\left[c(m, \bar{\mathbf{L}}_{mi}, \bar{\mathbf{A}}_{mi}; \mu)\right.\right. \\
&\qquad\left.\left. - E_{\hat{\boldsymbol{\alpha}}}\{c(m, \bar{\mathbf{L}}_{mi}, \bar{\mathbf{A}}_{mi}; \mu) \mid \bar{\mathbf{L}}_{mi}, \bar{\mathbf{A}}_{m-1i}\}\right]\right] \\
&= \sum_{i=1}^{n}\left[\frac{d(\bar{\mathbf{L}}_{K+1i}, \bar{\mathbf{A}}_{Ki}; \mu)}{\bar{\pi}_{Ki}(\hat{\boldsymbol{\alpha}})} - \sum_{m=1}^{K}\left\{\frac{\hat{H}_{mi}(\mu)}{\bar{\pi}_{mi}(\hat{\boldsymbol{\alpha}})} - \frac{\hat{T}_{mi}(\mu)}{\bar{\pi}_{m-1i}(\hat{\boldsymbol{\alpha}})}\right\}\right] \\
&= \sum_{i=1}^{n}\left[\frac{d(\bar{\mathbf{L}}_{K+1i}, \bar{\mathbf{A}}_{Ki}; \mu)}{\bar{\pi}_{Ki}(\hat{\boldsymbol{\alpha}})} - \sum_{m=1}^{K}\left\{\frac{\hat{T}_{m+1i}(\mu)}{\bar{\pi}_{mi}(\hat{\boldsymbol{\alpha}})} - \frac{\hat{T}_{mi}(\mu)}{\bar{\pi}_{m-1i}(\hat{\boldsymbol{\alpha}})}\right\}\right] \\
&= \sum_{i=1}^{n} \hat{T}_{1i}(\mu),
\end{aligned}$$

because the sample averages of $\hat{H}_m(\mu)\bar{\pi}_m^{-1}(\hat{\boldsymbol{\alpha}})$ and $\hat{T}_{m+1}(\mu)\bar{\pi}_m^{-1}(\hat{\boldsymbol{\alpha}})$ are equal for $m = 1, \ldots, K$. This is guaranteed by including the term $\phi_{m-1}\bar{\pi}_{m-1}^{-1}(\hat{\boldsymbol{\alpha}})$ in the GLM in Step 4 in Section 4.