Received: 21 June 2019 | Revised: 28 April 2020 | Accepted: 30 April 2020

DOI: 10.1002/sim.8678

RESEARCH ARTICLE

Statistics in Medicine WILEY

Ascertaining properties of weighting in the estimation of optimal treatment regimes under monotone missingness

Lin Dong[1] | **Eric Laber**[1] | **Yair Goldberg**[2] | **Rui Song**[1] | **Shu Yang**[1]

[1]Department of Statistics, North Carolina State University, Raleigh, North Carolina

[2]Department of Statistics, Technion Israel Institute of Technology, Haifa, Israel

Correspondence
Eric Laber, Department of Statistics, North Carolina State University, Raleigh, NC, 27695, USA.
Email: eblaber@ncsu.edu

Funding information
Division of Mathematical Sciences, Grant/Award Number: DMS-1555141; National Cancer Institute, Grant/Award Number: P01 CA142538; National Institute of Diabetes and Digestive and Kidney Diseases, Grant/Award Number: R01-DK-108073

Dynamic treatment regimes operationalize precision medicine as a sequence of decision rules, one per stage of clinical intervention, that map up-to-date patient information to a recommended intervention. An optimal treatment regime maximizes the mean utility when applied to the population of interest. Methods for estimating an optimal treatment regime assume the data to be fully observed, which rarely occurs in practice. A common approach is to first use multiple imputation and then pool the estimators across imputed datasets. However, this approach requires estimating the joint distribution of patient trajectories, which can be high-dimensional, especially when there are multiple stages of intervention. We examine the application of inverse probability weighted estimating equations as an alternative to multiple imputation in the context of monotonic missingness. This approach applies to a broad class of estimators of an optimal treatment regime including both Q-learning and a generalization of outcome weighted learning. We establish consistency under mild regularity conditions and demonstrate its advantages in finite samples using a series of simulation experiments and an application to a schizophrenia study.

KEYWORDS

augmented inverse probability weighting, dynamic treatment regimes, monotonic coarseness, outcome weighted learning, Q-learning

1 | INTRODUCTION

Dynamic treatment regimes operationalize clinical decision making as a sequence of decision rules, one per stage of intervention, that map current patient information to a recommended intervention.[1,2] An optimal treatment regime maximizes the mean utility if applied to select interventions in the patient population of interest.[3-5] Optimal treatment regimes have been estimated across a wide range of application areas including anticoagulation,[6-8] cancer,[9,10] mental disorders,[11-13] and HIV.[14-16] In these and nearly all other biomedical application areas, the observed data are subject to missingness, which can include missing measurements, treatments, and outcomes.[3,17]

There is a large body of literature on estimation of optimal treatment regimes using complete data. This body of research includes: approximate dynamic programming methods like Q- and A-learning[1,2,18-21] and its many variants;[22-31] direct-search methods including outcome weighted learning;[13,32-42] and model-based planning via g-computation.[10,43-47] Because these methods require complete data, it is often necessary to employ methods to address missing data. A common approach is to apply multiple imputation to complete the data, compute a given estimator of an optimal regime on each of the imputed data sets, and then aggregate these estimators, for example, by averaging or voting.[48-53] Although estimation

of an optimal treatment regime is often but one part of a suite of secondary analyses, the requirement to develop a complete dataset is convenient as it can be used for a variety of other analyses.

Despite its appealing features, multiple imputation can be problematic with complex longitudinal data arising in the context of sequential decision problems because the data can be high-dimensional and subjects are often missing large segments of data. Furthermore, multiple imputation requires estimating the joint distribution of patient trajectories and constructing a high-quality imputation model in this context is difficult. If a misspecified model is used to impute large amounts of missing data, the estimators may be biased and inferences inaccurate.[54] We examine a class of augmented inverse probability weighted estimators of the optimal treatment regime that applies to approximate dynamic programming and to direct-search methods when the data have a monotone missingness pattern. The application of standard arguments of semiparametric efficiency theory establishes a double robustness property for this class of estimators.[55] We show that augmented weighting performs favorably as compared to multiple imputation and to simple inverse probability weighting in simulation examples. These results suggest that investigators should give serious consideration to using weighting methods as an alternative to multiple imputation in the context of estimating optimal treatment regimes in practice. This work fills the gap in using weighting estimator in the context of monotonic missingness in the dynamic treatment regime literature. Moreover, a unified estimating equation framework is given that can be generalized to a broader class of estimation methods in this setting.

The remainder of this article is organized as follows. In Section 2, we set notation and define an optimal treatment regime using potential outcomes. In Section 3, we describe a class of estimators of an optimal treatment regimes for use with complete data; this class includes Q-learning and outcome weighted learning as special cases. In Section 4, we derive an augmented inverse probability weighted estimator for the proposed class of estimators that applies under a monotone missingness pattern. In Section 5, we present simulation examples and an application to data from a sequential multiple assignment randomized trial on schizophrenia. A brief discussion of future work is given in Section 6.

2 | SETUP AND NOTATION

We consider longitudinal data arising from an observational study or a sequential multiple assignment randomized trial.[56-58] The complete data are assumed to be of the form $\{(\mathbf{X}_{1i}, A_{1i}, \mathbf{X}_{2i}, A_{2i}, \dots, \mathbf{X}_{Ti}, A_{Ti}, Y_i)\}_{i=1}^{n}$, which comprise n independent replicates of $(\mathbf{X}_1, A_1, \mathbf{X}_2, A_2, \dots, \mathbf{X}_T, A_T, Y)$, where: T is the number of treatment stages, $\mathbf{X}_1 \in \mathbb{R}^{p_1}$ is baseline patient information, and $\mathbf{X}_t \in \mathbb{R}^{p_t}$ is information collected during stage $(t-1)$ for $t = 2, \dots, T$, $A_t \in \mathcal{A}_t = \{-1, 1\}$ is the treatment assigned during stage $t = 1, \dots, T$, and $Y \in \mathcal{Y} = \mathbb{R}$ is the terminal outcome coded so that higher values are better. The restriction to binary treatments is not necessary for approximate dynamic programming methods; however, most variants of outcome weighted learning require binary treatments[59,60] so we impose this restriction to allow for a simple and unified notation.

Define $\mathbf{H}_1 = \mathbf{X}_1$, and recursively define $\mathbf{H}_t = (\mathbf{H}_{t-1}, A_{t-1}, \mathbf{X}_t)$ for $t = 2, \dots, T$. Thus, $\mathbf{H}_t$ is the information available to the decision maker at time $t = 1, \dots, T$. Let $\mathcal{H}_t$ denote the support of $\mathbf{H}_t$, and for each $\mathbf{h}_t \in \mathcal{H}_t$, define $\Psi_t(\mathbf{h}_t) \subseteq \mathcal{A}_t$ to be the set of allowable treatments for a patient presenting with history $\mathbf{H}_t = \mathbf{h}_t$ at time t. A treatment regime in this context is a sequence of maps, $\boldsymbol{\pi} = (\pi_1, \dots, \pi_T)$, with $\pi_t : \mathcal{H}_t \to \mathcal{A}_t$ and $\pi_t(\mathbf{h}_t) \in \Psi_t(\mathbf{h}_t)$ for all $\mathbf{h}_t \in \mathcal{H}_t$ and $t = 1, \dots, T$. Under $\boldsymbol{\pi}$, a patient with history $\mathbf{H}_t = \mathbf{h}_t$ at time t would be recommended to receive treatment $\pi_t(\mathbf{h}_t)$. Let Π denote the space of all feasible regimes. An optimal treatment regime, say $\boldsymbol{\pi}^{\text{opt}} \in \Pi$, maximizes the mean outcome if applied to the population of interest. We formalize this definition using potential outcomes.[61,62]

For each $t = 1, \dots, T$, define $\bar{\mathbf{a}}_t = (a_1, \dots, a_t)$. Let $\mathbf{H}_t^*(\bar{\mathbf{a}}_{t-1})$ denote the potential history under treatment sequence $\bar{\mathbf{a}}_{t-1}$, and let $Y^*(\bar{\mathbf{a}}_T)$ denote the potential outcome under treatment sequence $\bar{\mathbf{a}}_T$. Therefore, the set of all potential outcomes is

$$\mathcal{W}^* = \{\mathbf{H}_t^*(\bar{\mathbf{a}}_{t-1}), Y^*(\bar{\mathbf{a}}_T) \ : \ a_t \in \Psi_t\{\mathbf{H}_t^*(\bar{\mathbf{a}}_{t-1})\}, t = 1, \dots, T\},$$

where we have defined $\mathbf{H}_1^*(\bar{\mathbf{a}}_0) \equiv \mathbf{H}_1$. Let $1_{(\cdot)}$ be the indicator function. The potential outcome under a regime $\boldsymbol{\pi} \in \Pi$ is

$$Y^*(\boldsymbol{\pi}) = \sum_{\bar{\mathbf{a}}_T} Y^*(\bar{\mathbf{a}}_T) \prod_{v=1}^{T} 1_{[\pi_v\{\mathbf{H}_v^*(\bar{\mathbf{a}}_{v-1})\} = a_v]}.$$

Define the value of regime $\boldsymbol{\pi}$ to be $V(\boldsymbol{\pi}) = \mathbb{E}Y^*(\boldsymbol{\pi})$, that is, the marginal mean outcome if all subjects were assigned treatment according to $\boldsymbol{\pi}$. The optimal regime, $\boldsymbol{\pi}^{\text{opt}} \in \Pi$, satisfies $V(\boldsymbol{\pi}^{\text{opt}}) \geq V(\boldsymbol{\pi})$ for all $\boldsymbol{\pi} \in \Pi$. To identify $\boldsymbol{\pi}^{\text{opt}}$ in terms of the data-generating model, we make the following assumptions: (i) consistency, $\mathbf{H}_t = \mathbf{H}_t^*(\overline{\mathbf{A}}_{t-1})$ for $t = 2, \ldots, T$ and $Y = Y^*(\overline{\mathbf{A}}_T)$, (ii) strong ignorability, $A_t \perp \mathcal{W}^* | \mathbf{H}_t$ for $t = 1, \ldots, T$, and (iii) positivity, $P(A_t = a_t | \mathbf{H}_t = \mathbf{h}_t) > 0$ for all $a_t \in \Psi_t(\mathbf{h}_t)$ and $t = 1, \ldots, T$. These assumptions are standard in the dynamic treatment regimes literature.[2,3,21,63] Hereafter, we implicitly assume that these conditions hold.

3 | ESTIMATION WITH COMPLETE DATA

In this section, we review estimation of an optimal treatment regime when the data are completely observed. We consider a class of estimators that are representable as solutions to a set of estimating equations. This class is quite broad and includes most of the estimators commonly used in practice. To illustrate this point, we show in the Appendix that Q-learning and a generalization of outcome weighted learning belong to this class.

We consider treatment regimes of the form $\boldsymbol{\pi}_{\boldsymbol{\beta}} = \{\pi_1(\,\cdot\,; \boldsymbol{\beta}_1), \ldots, \pi_T(\,\cdot\,; \boldsymbol{\beta}_T)\}$ in which the decision rules composing the regime are indexed by parameters $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}}, \ldots, \boldsymbol{\beta}_T^{\mathrm{T}})^{\mathrm{T}} \in \mathcal{B}$, where $\mathcal{B}$ is a normed linear space with norm $\|\cdot\|_{\mathcal{B}}$. For example, one might consider linear decision rules of the form $\pi_t(\mathbf{h}_t; \boldsymbol{\beta}_t) = \text{sign}(\boldsymbol{\beta}_t^{\mathrm{T}} \mathbf{h}_{t,0})$, where $\mathbf{h}_{t,0}$ is a feature vector constructed from $\mathbf{h}_t$ and $\text{sign}(u)$ is 1 if u is positive and -1 otherwise. We do not exclude the case in which $\boldsymbol{\beta}_t$ includes nuisance parameters so that $\pi_t(\cdot\,:\,\boldsymbol{\beta}_t)$ depends only on a subvector of $\boldsymbol{\beta}_t$; however, we do not make any special distinction for such nuisance parameters as it is not important for our purposes. We assume that an estimator $\hat{\boldsymbol{\beta}}_n = (\hat{\boldsymbol{\beta}}_{1,n}^{\mathrm{T}}, \ldots, \hat{\boldsymbol{\beta}}_{T,n}^{\mathrm{T}})^{\mathrm{T}}$ of $\boldsymbol{\beta}$ is constructed by solving the estimating equation

$$\mathbb{P}_n \mathbf{m}_n(\mathbf{H}_T, A_T, Y; \boldsymbol{\beta}) = 0 \tag{1}$$

over $\boldsymbol{\beta} \in \mathcal{B}$, where $\mathbb{P}_n$ denotes the empirical measure, and $\mathbf{m}_n : \mathcal{H}_T \times \mathcal{A}_T \times \mathcal{Y} \to \mathbb{R}^J$. The dependence of $\mathbf{m}_n$ on n is to allow for regularization or other factors that may vary with the sample size.[29,34,37,64]

Many estimators of an optimal treatment regime are based on backwards recursion. For these estimators, some components of $\mathbf{m}_n$ will depend on the partial history $\mathbf{S}_t \triangleq (\mathbf{H}_t^{\mathrm{T}}, A_t)^{\mathrm{T}}$ for $t = 1, \ldots, T$ rather than the complete data $\mathbf{S}_{T+1} \triangleq (\mathbf{H}_T^{\mathrm{T}}, A_T, Y)^{\mathrm{T}}$. For $t = 1, \ldots, T$ define $\mathcal{J}_t$ to be the indices of $\mathbf{m}_n$ such that $m_{n,j}$ depends on $\mathbf{S}_{T+1}$ only through $\mathbf{S}_t$, that is,

$$\mathcal{J}_t = \{1 \leq j \leq J \;:\; m_{n,j}(\mathbf{h}_T, a_T, y; \boldsymbol{\beta}) = \tilde{m}_{n,j}(\mathbf{h}_t, a_t; \boldsymbol{\beta}) \text{ for some } \tilde{m}_{n,j} : \mathcal{H}_t \times \mathcal{A}_t \to \mathbb{R}\},$$

and $\mathcal{J}_{T+1}$ are the indices that rely on the complete data. Under this representation, the estimation equation in Equation (1) can be equivalently expressed as

$$\mathbb{P}_n \tilde{m}_{n,j}(\mathbf{S}_t; \boldsymbol{\beta}) = 0 \quad \text{for all} \quad j \in \mathcal{J}_t, \quad t = 1, \ldots, T+1. \tag{2}$$

We will exploit this representation to use more of the observed data in constructing weighted complete case estimators in Section 4.

Let $\boldsymbol{\beta}_n^*$ denote the population analog of $\hat{\boldsymbol{\beta}}_n$, that is, the solution to Equation (2) with $\mathbb{P}_n$ replaced by P. We say that $\hat{\boldsymbol{\beta}}_n$ is consistent if $\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n^*\|_{\mathcal{B}}$ converges to zero in probability. Because our objective is not to propose new estimators of an optimal treatment regime, we will assume that the estimating equation has been suitably constructed to ensure consistency under the data-generating model in the complete data case and avoid stating specific conditions under which such consistency holds. Giving such general conditions would be cumbersome; for example, the conditions under which Q-learning with linear models is consistent are quite different from those under which kernel-based outcome weighted learning is consistent.

4 | ESTIMATION WITH INCOMPLETE DATA

4.1 | Missingness mechanism

We assume that baseline covariate information and initial treatment assignment, $(\boldsymbol{X}_1, A_1)$, are always observed. This assumption generally holds in practice because patients who do not receive an initial treatment assignment are often

unenrolled from the study and excluded from subsequent analyses. We further assume that the missingness pattern is nearly monotone; that is, any item missingness that violates this monotone pattern is sparse, and the missingness pattern has been made monotone through artificial censoring or single imputation. Because patient dropout is the primary cause for missing data in longitudinal studies, for example, SMARTS, this assumption is common in the literature.[17]

Let $C \in \{1, \dots, T+1\}$ denote the dropout time so that $C = t$ if the patient dropped out after assignment of A_t for $t = 1, \dots, T$ and $C = T+1$ if the patient's trajectory is fully observed, that is, if $C = t$ for $t = 1, \dots, T$, we observe $(\boldsymbol{X}_1, A_1, \dots, \boldsymbol{X}_t, A_t) = (\mathbf{H}_t, A_t)$; if $C = T+1$, we observe complete case $(\boldsymbol{X}_1, A_1, \dots, \boldsymbol{X}_T, A_T, Y) = (\mathbf{H}_T, A_T, Y)$. We further assume that the data are missing at random[65,66] so that $1_{C=t} \perp (\mathbf{X}_{t+1}, A_{t+1}, \dots, X_T, A_T, Y) | \mathbf{H}_t, A_t$ for all $t = 1, \dots, T$.

The simplest strategy for adapting the estimating equations presented in Section 3 for missing data is through inverse probability weighting of complete cases. Note that "complete case" for a term $m_{n,j}(\mathbf{S}_t; \boldsymbol{\beta})$, where $j \in \mathcal{J}_t$, is the one in which $\mathbf{S}_t$ is observed and not necessarily the one for which the complete trajectory, $\mathbf{S}_{T+1}$, is observed.

4.2 | Inverse probability weighted estimating equations

Inverse probability weighted complete case (IPWCC) estimators re-weight the terms in the estimating equation for an optimal regime by their respective probabilities of being observed.[55,67] Define the discrete hazard of dropout at time $t = 1, \dots, T$ to be $\lambda_t(\mathbf{s}_t) = P(C = t | C \geq t, \mathbf{S}_t = \mathbf{s}_t)$. Thus, $\lambda_t(\mathbf{s}_t)$ is the probability of dropping out at stage t for a patient with covariate and treatment history $\mathbf{S}_t = \mathbf{s}_t$. The survivor function at time t is thus $K_t(\mathbf{s}_t) = P(C > t | \mathbf{S}_t = \mathbf{s}_t) = \prod_{v=1}^{t} \{1 - \lambda_v(\mathbf{s}_v)\}$. Under the MAR assumption, we can model the hazards using a binary regression model. For concreteness, we use a logistic regression model so that

$$\lambda_t(\mathbf{s}_t; \boldsymbol{\psi}_t) = \text{expit}\{g_t(\mathbf{s}_t; \boldsymbol{\psi}_t)\},$$

where $\text{expit}(u) \triangleq \exp(u)/\{1 + \exp(u)\}$, $\boldsymbol{\psi}_t \in \mathcal{K}_t \subseteq \mathbb{R}^{q_t}$ is a vector of parameters, and $g_t(\mathbf{s}_t; \boldsymbol{\psi}_t)$ is continuously differentiable in $\boldsymbol{\psi}_t$ for all $\mathbf{s}_t$. Define $\boldsymbol{\psi} \triangleq (\boldsymbol{\psi}_t^{\mathrm{T}}, \dots, \boldsymbol{\psi}_T^{\mathrm{T}})^{\mathrm{T}} \in \mathcal{K} \subseteq \mathbb{R}^q$, where $q = q_1 + \dots + q_T$. Define $\varsigma_t(c, \mathbf{s}_t; \boldsymbol{\psi}_t) \triangleq 1_{c=t} \nabla_{\boldsymbol{\psi}_t} g_t(\mathbf{s}_t; \boldsymbol{\psi}_t) - 1_{c \geq t} \nabla_{\boldsymbol{\psi}_t} g_t(\mathbf{s}_t; \boldsymbol{\psi}_t) \text{expit}\{g_t(\mathbf{s}_t; \boldsymbol{\psi}_t)\}$ to be the score function of the posited logistic regression model, and let $\hat{\boldsymbol{\psi}}_{t,n}$ be a solution to $\mathbb{P}_n \varsigma_t(C, \mathbf{S}_t; \boldsymbol{\psi}_t) = 0$. Let $\overline{\boldsymbol{\psi}}_t = (\boldsymbol{\psi}_1^{\mathrm{T}}, \boldsymbol{\psi}_2^{\mathrm{T}}, \dots, \boldsymbol{\psi}_t^{\mathrm{T}})^{\mathrm{T}}$ so that $\hat{\overline{\boldsymbol{\psi}}}_{t,n} = (\hat{\boldsymbol{\psi}}_{1,n}^{\mathrm{T}}, \hat{\boldsymbol{\psi}}_{2,n}^{\mathrm{T}}, \dots, \hat{\boldsymbol{\psi}}_{t,n}^{\mathrm{T}})^{\mathrm{T}}$. The estimated survivor function is $K_t(\mathbf{s}_t; \hat{\overline{\boldsymbol{\psi}}}_{t,n}) = \prod_{v=1}^{t} \{1 - \lambda_v(\mathbf{s}_v; \hat{\boldsymbol{\psi}}_{v,n})\}$.

Define the complete case weights at level $t = 2, \dots, T+1$ and $j \in \mathcal{J}_t$ under parameters $\boldsymbol{\psi}$ to be $w_j^{\text{cc}}(c, \mathbf{s}_{t-1}; \overline{\boldsymbol{\psi}}_{t-1}) = 1_{c > (t-1)} / K_{t-1}(\mathbf{s}_{t-1}; \overline{\boldsymbol{\psi}}_{t-1})$. The IPWCC estimator of an optimal treatment regime based on Equation (2) solves

$$\mathbb{P}_n w_j^{\text{cc}}(C, \mathbf{S}_{t-1}; \hat{\overline{\boldsymbol{\psi}}}_{t-1,n}) \check{m}_{n,j}(\mathbf{S}_t; \boldsymbol{\beta}) = 0, \quad j \in \mathcal{J}_t, \quad t = 1, \dots, T+1,$$

with the understanding that $w_j^{\text{cc}}(c, \mathbf{s}_0; \overline{\boldsymbol{\psi}}_0) \equiv 1$ for $j \in \mathcal{J}_1$. The preceding equations along with those for $\boldsymbol{\psi}$ could be expressed as a single stacked estimating equation by concatenating the score equation for the logistic regression models onto $\mathbf{m}_n$. Let $\mathbb{P}_n \check{\mathbf{m}}_n(\mathbf{S}_{T+1}; \boldsymbol{\beta}, \boldsymbol{\psi}) = 0$ denote this joint estimating equation. It follows from the derivations given in the Appendix that both the Q-learning and outcome weighted learning estimators can thus be constructed under a monotone missingness pattern using IPWCC by means of the preceding estimating equation. The following result can be used to establish consistency of the IPWCC estimator when combined with standard conditions for Z-estimators, for example, the estimating equation has a unique isolated minimizer.[68] The proofs are relegated to Appendix.

Theorem 1. *Assume that the survivor function is correctly specified so that $K_t(\mathbf{s}_t) = K_t(\mathbf{s}_t; \overline{\boldsymbol{\psi}}_t^*)$ for all t and $\mathbf{s}_t$, for some $\boldsymbol{\psi}^* \in \mathcal{K}$ and that the $\hat{\boldsymbol{\psi}}_n \to \boldsymbol{\psi}^*$ in probability. If $\boldsymbol{\beta}_n^*$ satisfies $\|P\mathbf{m}_n(\mathbf{S}_{T+1}; \boldsymbol{\beta}_n^*)\| = o(1)$, then $\|P\check{\mathbf{m}}_n(\mathbf{S}_{T+1}; \boldsymbol{\beta}_n^*, \hat{\boldsymbol{\psi}}_n)\| = o_p(1)$.*

4.3 | Augmented inverse probability weighted estimating equations

Using semiparametric efficiency theory for monotone coarsening, of which monotone missingness is a special case, we derive an augmented inverse probability weighted complete case (AIPWCC) estimator.[55,69] Define for each $t = 1, \dots, T$, the conditional mean $\mathbf{d}_{n,t}(\mathbf{s}_t) \triangleq \mathbb{E}\{\mathbf{m}_n(\mathbf{S}_{T+1}; \boldsymbol{\beta}^*) | \mathbf{S}_t = \mathbf{s}_t\}$ for which we posit parametric models $\mathbf{d}_{n,t}(\mathbf{s}_t; \boldsymbol{\alpha}_t)$ indexed by $\boldsymbol{\alpha}_t \in \mathbb{R}^{\nu_t}$. Let $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^{\mathrm{T}}, \dots, \boldsymbol{\alpha}_T^{\mathrm{T}})^{\mathrm{T}}$. An estimator of $\mathbf{d}_{n,t}$ can thus be obtained by regressing $\mathbf{m}_n(\mathbf{S}_{T+1}; \hat{\boldsymbol{\beta}}_n)$ on $d_{n,t}(\mathbf{S}_t; \boldsymbol{\alpha}_t)$ restricted to patients with $C = T+1$, which gives $\hat{\boldsymbol{\alpha}}_{t,n}$ and subsequently $\hat{\mathbf{d}}_{n,t}(\mathbf{s}_t) = d_{n,t}(\mathbf{s}_t, \hat{\boldsymbol{\alpha}}_{t,n})$ for each $t = 1, \dots, T$. Let $\lambda_t(\mathbf{s}_t; \boldsymbol{\psi}_t)$ and

$K_t(\mathbf{s}_t;\overline{\boldsymbol{\psi}}_t)$ be as defined in the previous section. For each $t=2,\ldots,T+1, j\in\mathcal{J}_t$, and $r=1,\ldots,t-1$, define the augmentation weights

$$w_{r,j}^{\text{aug}}(c,\mathbf{s}_r;\overline{\boldsymbol{\psi}}_r)=\frac{1_{c=r}-\lambda_r(\mathbf{s}_r;\boldsymbol{\psi}_r)}{K_r(\mathbf{s}_r;\overline{\boldsymbol{\psi}}_r)}.$$

The AIPWCC estimating equations are

$$\mathbb{P}_n\left\{w_j^{\text{cc}}(C,\mathbf{S}_{t-1};\hat{\overline{\boldsymbol{\psi}}}_{t-1,n})\tilde{m}_{n,j}(\mathbf{S}_t;\boldsymbol{\beta})+\sum_{r=1}^{t-1}w_{r,j}^{\text{aug}}(C,\mathbf{S}_r;\hat{\overline{\boldsymbol{\psi}}}_{r,n})d_{n,r,j}(\mathbf{S}_r;\hat{\boldsymbol{\alpha}}_{r,n})\right\}=0,\quad \text{for all}\quad j\in\mathcal{J}_t,\ t=1,\ \ldots\ ,T+1.$$

To obtain a single set of estimating equations, one could concatenate the estimating equations for $\boldsymbol{\psi}$ and $\boldsymbol{\alpha}$ to those given above. Let $\mathbb{P}_n\check{\mathbf{m}}_n(\mathbf{S}_{T+1};\boldsymbol{\beta},\boldsymbol{\psi},\boldsymbol{\alpha})=0$ denote the joint estimating equations. Explicit forms of these estimating equations for both Q-learning and outcome weighted learning are provided in the Supplemental Material.

Theorem 2. *Assume that the hazard functions for dropout are correctly specified so that $\lambda_t(\mathbf{s}_t)=\lambda_t(\mathbf{s}_t;\boldsymbol{\psi}_t^*)$ for all $\mathbf{s}_t$ for some $\boldsymbol{\psi}_t^*\in\mathcal{K}_t$ and $\hat{\boldsymbol{\psi}}_n\to\boldsymbol{\psi}^*$ in probability or that the regression functions are correctly specified so that $\mathbf{d}_{n,t}(\mathbf{s}_t)=\mathbf{d}_{n,t}(\mathbf{s}_t;\boldsymbol{\alpha}_t^*)$ for some $\boldsymbol{\alpha}_t^*\in\mathcal{A}_t$ and $\hat{\boldsymbol{\alpha}}_n\to\boldsymbol{\alpha}^*$ in probability. If $\boldsymbol{\beta}_n^*$ satisfies $\|P\mathbf{m}_n(\mathbf{S}_{T+1};\boldsymbol{\beta}_n^*)\|=o(1)$, then $\|P\check{\mathbf{m}}_n(\mathbf{S}_{T+1};\boldsymbol{\beta}_n^*,\hat{\boldsymbol{\psi}}_n,\hat{\boldsymbol{\alpha}}_n)\|=o_p(1)$. Thus, the AIPWCC estimator is doubly robust.*

Remark 1. When the dimension of the trajectory space is high, the solutions to the estimating equations may be unstable. In these cases, regularization may be necessary to stabilize the solution and to reduce over-fitting..[70,71] In our simulation experiments, we regularize the estimating equation using an adaptive ridge penalty; that is, given a preliminary estimator $\hat{\boldsymbol{\beta}}_n^0$ of $\boldsymbol{\beta}^*$ based on the unpenalized estimating equation, we compute $\hat{\boldsymbol{\beta}}_n^{\lambda_n}$ as the solution to $\mathbb{P}_n\check{\mathbf{m}}_n(\mathbf{S}_{T+1};\boldsymbol{\beta},\hat{\boldsymbol{\psi}}_n,\hat{\boldsymbol{\alpha}}_n)+\lambda_n\boldsymbol{\beta}/|\hat{\boldsymbol{\beta}}_n|=0$, where the division is taken elementwise and $\lambda_n\geq 0$ is a tuning parameter.

5 | SIMULATION AND DATA APPLICATION

5.1 | Simulation studies

We compare the performances of IPWCC, AIPWCC, and multiple imputation (MI) in terms of the value of the estimated optimal regime; these methods are applied with monotone missing data and the estimation is done using either Q-learning or outcome weighted learning. Data are simulated to mimic a two-stage SMART with binary treatments at each stage. The complete data are generated as follows:

$$\begin{aligned}
&\boldsymbol{X}_1=(X_{11},\ \ldots\ ,X_{1p})^{\mathrm{T}},X_{1k}\sim\text{Bernoulli}(0.5),k=1,\ \ldots\ ,p;\\
&A_1\sim\text{Uniform}\{-1,1\};\\
&\boldsymbol{X}_2|\boldsymbol{X}_1=\boldsymbol{x}_1,A_1=a_1\sim\mathcal{N}_p\{(\boldsymbol{\Gamma}_0+\boldsymbol{\Gamma}_1a_1)\boldsymbol{x}_1,\tau^2\boldsymbol{I}_p\};\\
&A_2\sim\text{Uniform}\{-1,1\};\\
&Y|\boldsymbol{X}_1=\boldsymbol{x}_1,A_1=a_1,\boldsymbol{X}_2=\boldsymbol{x}_2,A_2=a_2\sim\mathcal{N}\{\mu_Y(\boldsymbol{x}_1,a_1,\boldsymbol{x}_2,a_2),\sigma_Y^2\},
\end{aligned}\tag{3}$$

where $\mu_Y(\boldsymbol{x}_1,a_1,\boldsymbol{x}_2,a_2)=\gamma_{20}+\gamma_{21}a_1+\boldsymbol{\gamma}_{22}^{\mathrm{T}}\boldsymbol{x}_1a_1+\boldsymbol{\gamma}_{23}^{\mathrm{T}}\boldsymbol{x}_2+(\phi_{20}+\phi_{21}a_1+\boldsymbol{\phi}_{22}^{\mathrm{T}}\boldsymbol{x}_2)a_2$. Thus, the model is indexed by the matrices $\boldsymbol{\Gamma}_0,\boldsymbol{\Gamma}_1\in\mathbb{R}^{p\times p}$, coefficients $\gamma_{20},\gamma_{21},\boldsymbol{\gamma}_{22},\boldsymbol{\gamma}_{23},\phi_{20},\phi_{21},\boldsymbol{\phi}_{22}$, and variance components $\tau^2,\sigma_Y^2>0$.

For the missingness mechanism, we consider hazard functions of the form

$$P(C=t|C\geq t,\mathbf{H}_t,A_t)=\lambda_t(\mathbf{H}_t,A_t)=\text{expit}\{(1,X_{t1},A_t\cdot X_{t2})^{\mathrm{T}}\boldsymbol{\psi}_t\},\quad t=1,2.$$

We vary the parameters $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$ to obtain 35% and 65% missingness. The actual parameter values are provided in the Supplemental Materials. All simulation experiments use training sets of size $n=1000$ and 500 Monte Carlo replications.

In our implementation of Q-learning, we use linear models of the form set $Q_t(\mathbf{H}_t,A_t;\boldsymbol{\beta}_t)=\boldsymbol{\beta}_t^{\mathrm{T}}\mathbf{B}_{t,0}$, where $\mathbf{B}_{1,0}=(1,\mathbf{X}_1^{\mathrm{T}},A_1,\mathbf{X}_1^{\mathrm{T}}A_1)^{\mathrm{T}}$ and $\mathbf{B}_{2,0}=(1,\mathbf{X}_1^{\mathrm{T}},A_1,\mathbf{X}_1^{\mathrm{T}}A_1,A_2,A_1A_2,\mathbf{X}_2^{\mathrm{T}}A_2)^{\mathrm{T}}$. For outcome weighted learning, we use linear decision

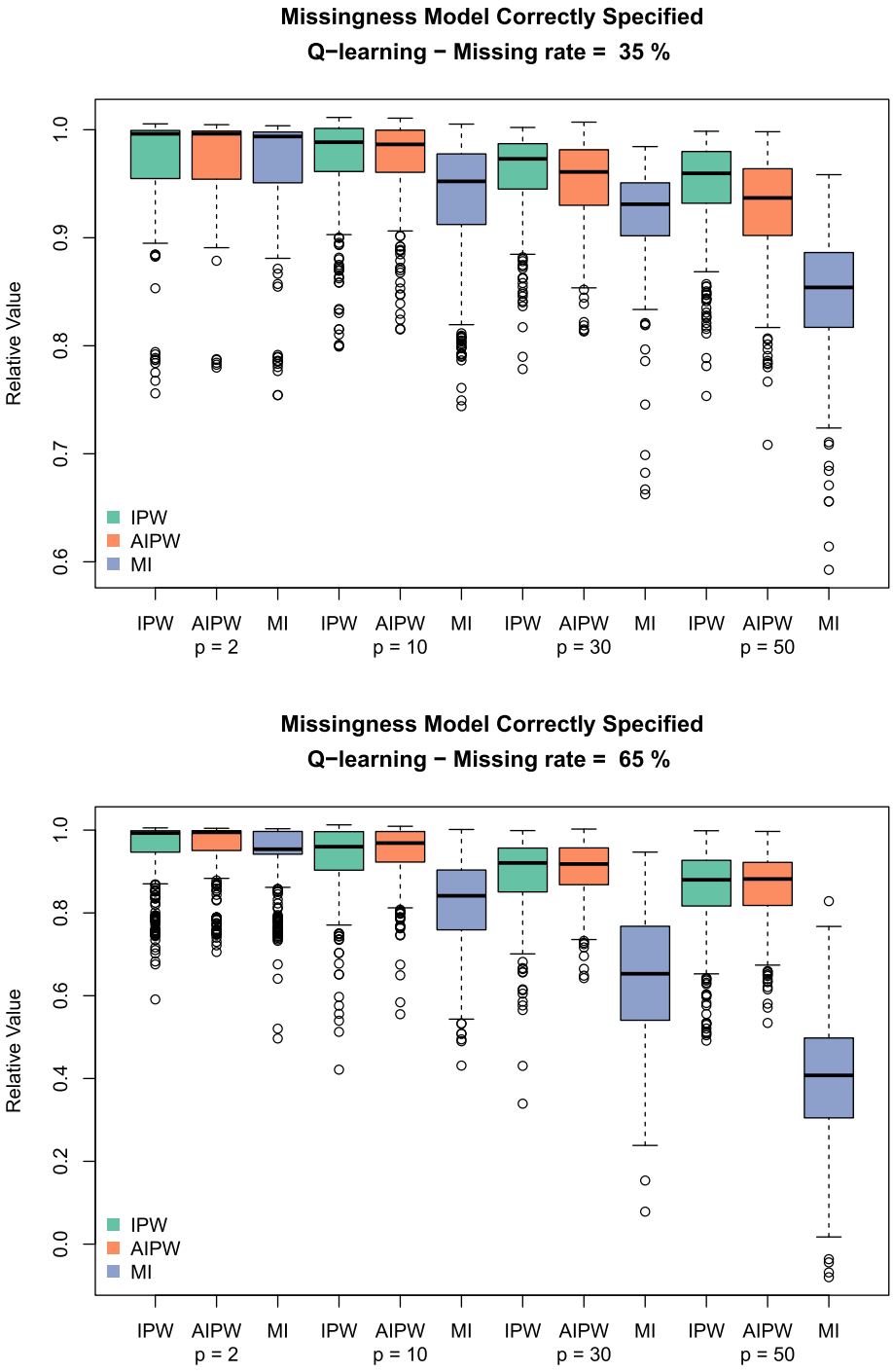

FIGURE 1 Relative value of *Q-learning* with IPWCC estimator, AIPWCC estimator, and MI when the missingness model is correctly specified. The length of the corresponding vertical bar is the Monte Carlo standard deviation of the relative value estimates [Colour figure can be viewed at wileyonlinelibrary.com]

rules of the form $f_t(\mathbf{H}_t;\boldsymbol{\eta}_t)=\boldsymbol{\eta}_t^{\mathrm{T}}\mathbf{H}_{t,0}$, where $\mathbf{H}_{1,0}=(1,\mathbf{X}_1^{\mathrm{T}})^{\mathrm{T}}$ and $\mathbf{H}_{2,0}=(1,\mathbf{X}_1^{\mathrm{T}},A_1,\mathbf{X}_1^{\mathrm{T}}A_1,X_2^{\mathrm{T}})$. These rules are estimated using logistic loss, also known as the entropy loss, as the convex surrogate.[72,73]

To illustrate the double-robustness property of the AIPWCC estimator, we consider both correctly and incorrectly specified models for the hazard functions. In the correctly specified case, we fit a logistic regression model at each stage with the correct features, that is, $(1,X_{t1},A_tX_{t2})$ for $t=1,2$. For the incorrectly specified model, we fit a logistic regression model with features $(1,X_1,X_{11}X_{12})$ at stage 1 and $(1,X_{21}^2,X_{22}^2)$ at stage 2. The conditional mean model is not correctly specified throughout.

We evaluate the performance of an estimated optimal regime, $\hat{\boldsymbol{\pi}}$, in terms of its relative value, which is defined as

$$\mathrm{RV}(\hat{\boldsymbol{\pi}})=\frac{V(\hat{\boldsymbol{\pi}})-V(\boldsymbol{\pi}_0)}{V(\boldsymbol{\pi}^{\mathrm{opt}})-V(\boldsymbol{\pi}_0)},$$

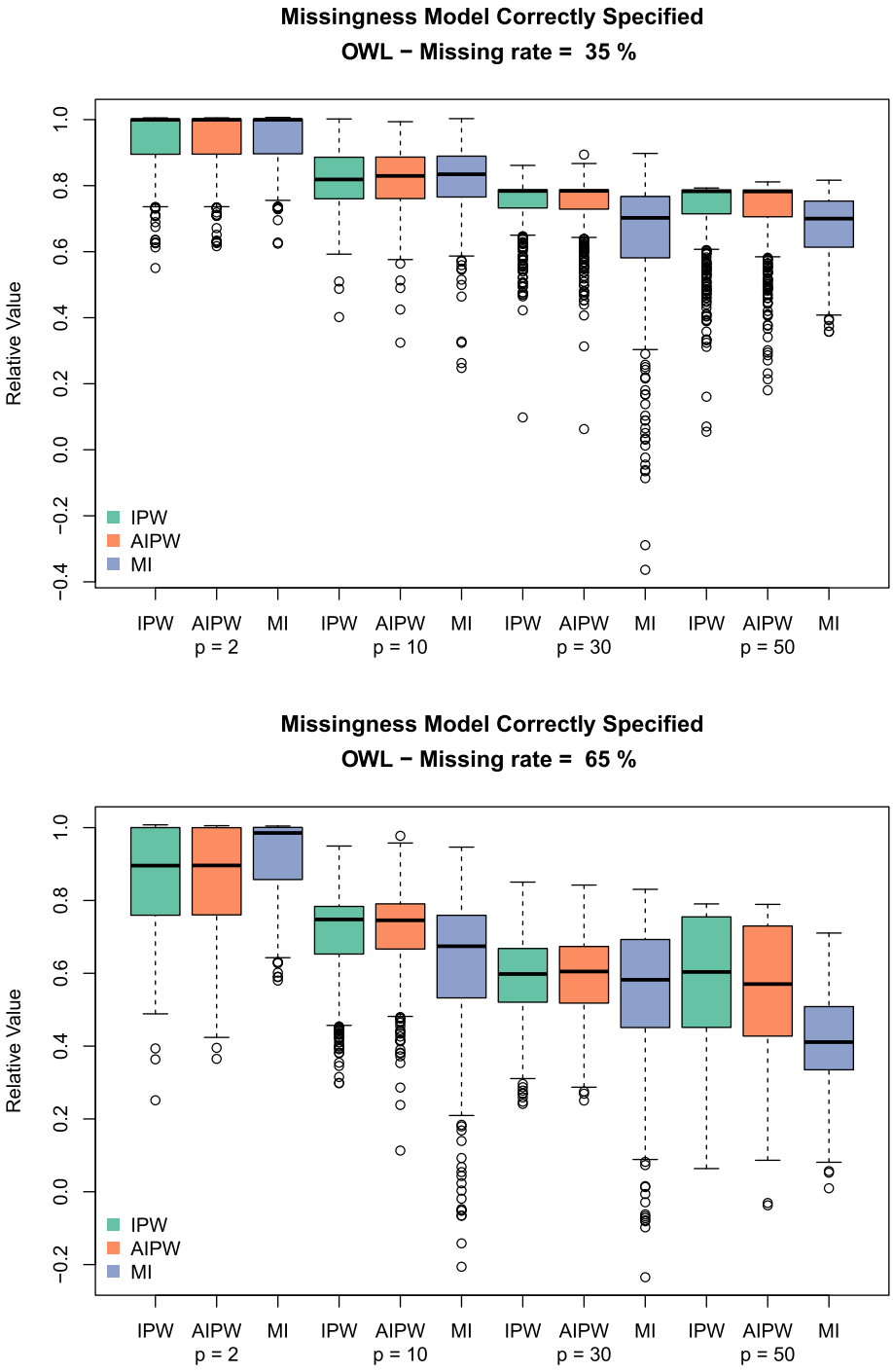

FIGURE 2 Relative value of *outcome weighted learning* with IPWCC estimator, AIPWCC estimator, and MI when the missingness model is correctly specified. The length of the corresponding vertical bar is the Monte Carlo standard deviation of the relative value estimates [Colour figure can be viewed at wileyonlinelibrary.com]

where $\boldsymbol{\pi}_0$ is the stochastic policy that assigns treatments at each stage using a fair-sided coin flip. The reason for using the relative value, for instance, instead of the raw value, is to allow for comparison across a range of generative models, for example, different problem dimensions. The requisite values are estimated using Monte Carlo methods with 40 000 simulated patients.[21]

We approximate the roots of the estimating equations using R package `nleqslv` with multiple starts. We implement MI using R package `MICE` with default settings and 10 imputed datasets.[74] The conditional expectation model in the AIPWCC estimator is fitted using ridge regression and tuned using the 5-fold cross-validation estimator of the value under the optimal regime.

The results for the correctly specified hazard models with Q-learning and outcome weighted learning are displayed in Figures 1 and 2, respectively. It can be seen that both the IPWCC and AIPWCC estimators generally outperform MI

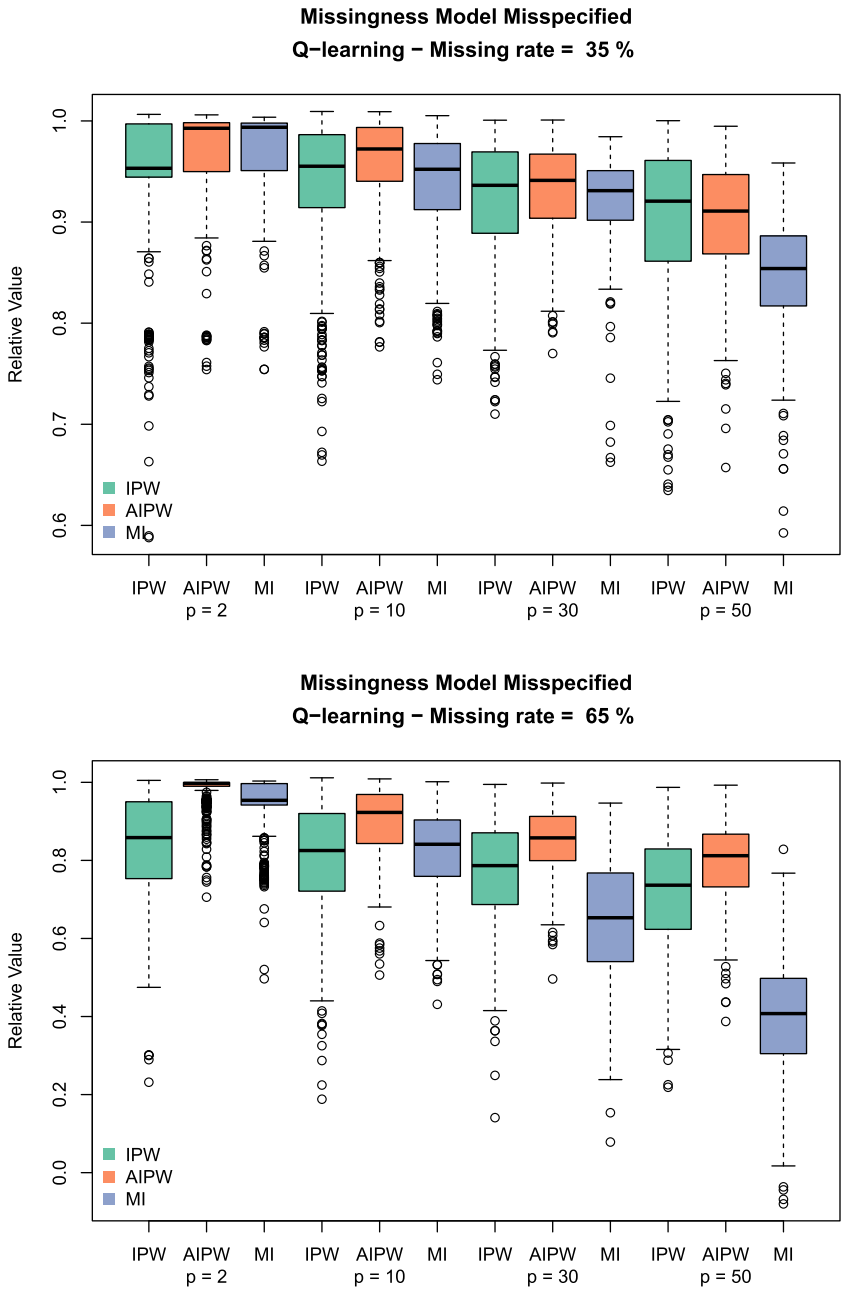

FIGURE 3 Relative value of *Q-learning* with IPWCC estimator, AIPWCC estimator, and MI when the missingness model is misspecified. The length of the corresponding vertical bar is the Monte Carlo standard deviation of the relative value estimates [Colour figure can be viewed at wileyonlinelibrary.com]

when the dimension of the covariates at each stage, p, is large. The results for the incorrectly specified hazard models estimated using Q-learning and outcome weighted learning are displayed in Figures 3 and 4, respectively. The results are qualitatively similar though the robustness of the AIPWCC shows improved performance relative to the IPWCC in some scenarios.

5.2 | CATIE trial analysis

We use data from the CATIE schizophrenia study[75] to illustrate the proposed methods. The CATIE study is a SMART, which enrolled 1460 schizophrenia patients. This dataset was chosen in part because it was used as an illustrative case study with MI by others.[17,76-78]

As done elsewhere in the literature, we compare two treatments of primary clinical interest at each stage: Perphenazine (coded -1) and Olanzapine (coded 1). The dataset consisting of 506 patients receiving these treatments , 46% of whom followed the entire course (ie, they are complete cases), 34% dropped out after stage 1, and 20% dropped out after

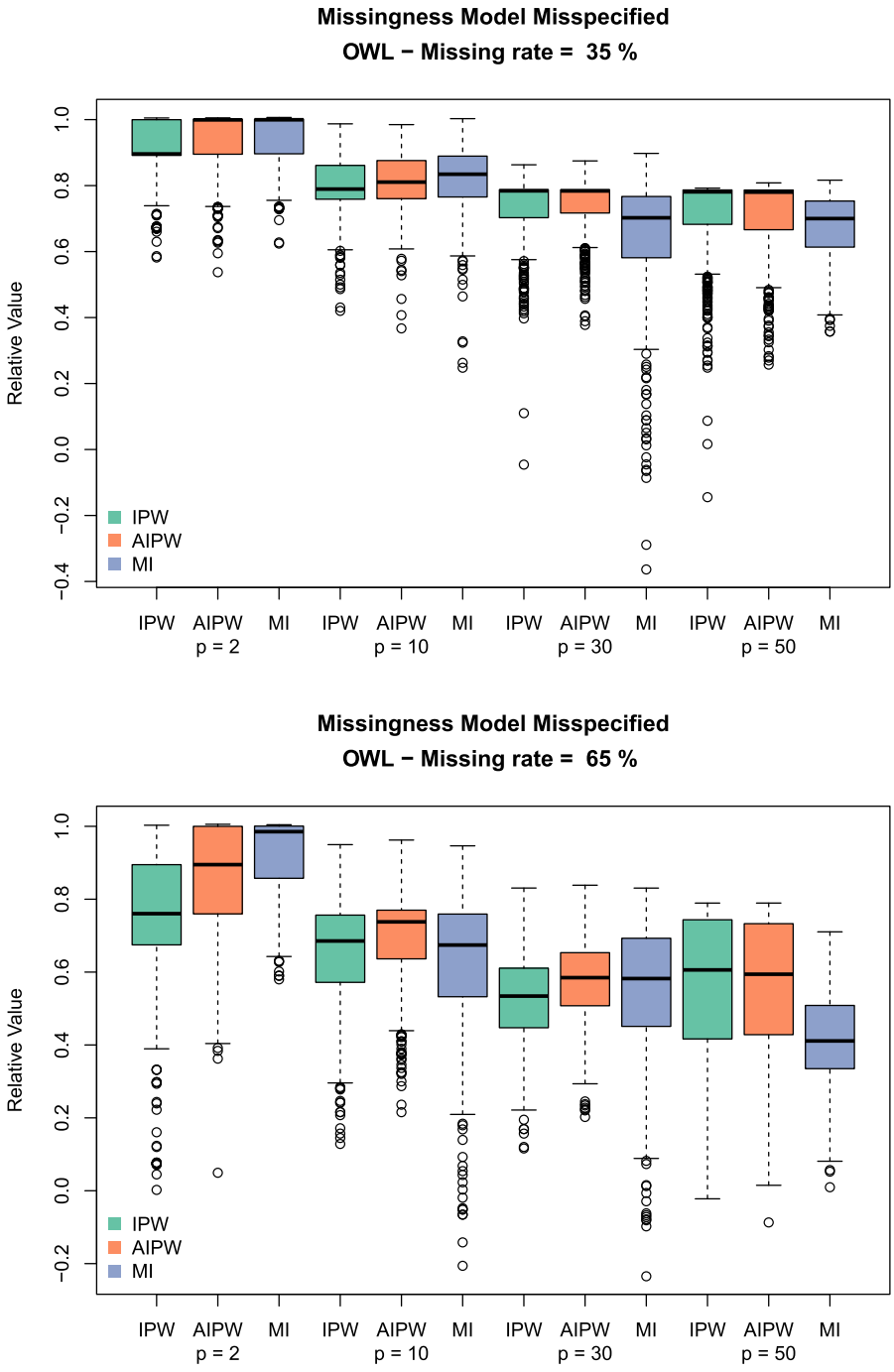

FIGURE 4 Relative value of *outcome weighted learning* with IPWCC estimator, AIPWCC estimator, and MI when the missingness model is misspecified. The length of the corresponding vertical bar is the Monte Carlo standard deviation of the relative value estimates [Colour figure can be viewed at wileyonlinelibrary.com]

stage 2. The missingness pattern is shown in Figure 5. Item missingness was sparse (less than 2% of the observed data) and singly imputed using mean imputation.

The positive and negative syndrome scale (PANSS) score is the standard medical scale for measuring symptom severity in schizophrenia. This score is a time-varying variable and was measured at each stage: baseline (PANSS0); stage 1 (PANSS1); and stage 2 (PANSS2). A higher PANSS score is associated with more severe symptoms, so we use 100-PANSS2 as the final outcome to match our convention of higher values representing better clinical outcomes. We include four baseline covariates in our analyses: PANNS0, baseline PANSS; EXACER, an indicator that the patient has been recently hospitalized; SEX; and TD, an indicator that the patient has Tardive Dyskinesia, a serious movement disorder associated with some antipscyhotic medications. In addition, we include PANNS1, first stage PANSS, in our second stage models. As in the simulation study, we used linear models for the Q-functions of Q-learning and linear decision rules for outcome weighted learning.

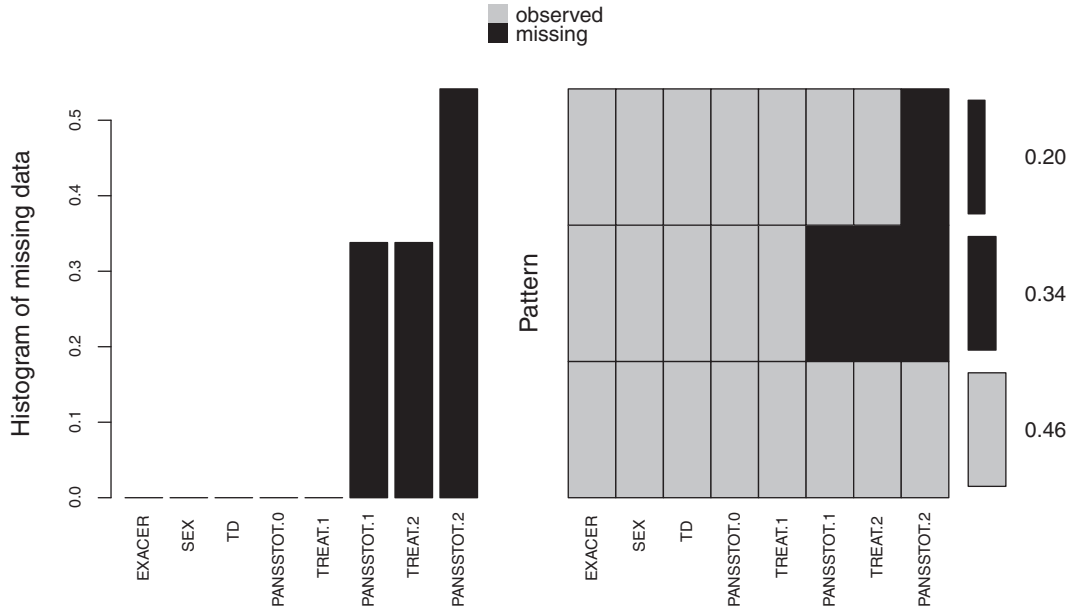

FIGURE 5 Missiningness pattern of CATIE study after cleaning

TABLE 1 Cross-validated value estimates of the optimal regimes estimated using different methods for CATIE study

	Q-MI	Q-IPW	Q-AIPW	OWL-MI	OWL-IPW	OWL-AIPW
$\hat{V}(\hat{\pi}^{\text{opt}})$	35.406	40.684	41.147	48.220	49.366	48.164

We compared the following six approaches: Q-learning with MI (Q-MI), Q-learning with IPWCC (Q-IPW), Q-learning with AIPWCC (Q-AIPW), outcome weighted learning with MI (OWL-MI), outcome weighted learning with IPWCC (OWL-IPW), and outcome weighted learning with AIPWCC (OWL-AIPW). The cross-validated value for each approach is reported in Table 1. With the CATIE data, outcome weighted learning generally performed favorably to Q-learning. In terms of adjustment for missing data, MI appears to be worse than AIWPCC/IPWCC with Q-learning but about the same with outcome weighted learning.

6 | DISCUSSION

Missing data are essentially unavoidable with SMARTS and other longitudinal study designs commonly used to estimate optimal treatment regimes. Multiple Imputation has been shown to be an effective tool for accommodating missing data in such studies. However, in some settings, imputation can involve modeling complex processes, which may be prone to misspecification and high variance. We examined the use of inverse probability weighted methods and showed such methods are consistent for a broad class of estimators of an optimal treatment regime. Furthermore, in empirical experiments, these methods outperformed imputation with the gap in performance widening with the increasing trajectory dimension as well as with the increasing amount of missingness. Thus, we recommend that such weighting methods be given serious consideration by researchers estimating optimal treatment regimes from randomized or observational studies. Alternatively, it may be beneficial to forgo choosing between imputation and weighting and instead combine them.[17] We leave such a hybrid approach to future work.

ORCID

Eric Laber https://orcid.org/0000-0003-2640-7696

REFERENCES

1. Murphy SA. Optimal dynamic treatment regimes. *J Royal Stat Soc Ser B (Stat Methodol)*. 2003;65(2):331-355.
2. Robins JM. Optimal structural nested models for optimal sequential decisions. Paper presented at: Proceedings of the 2nd seattle Symposium in Biostatistics; 2004:189-326; Springer.
3. Kosorok MR, Moodie EE. *Adaptive Treatment Strategies in Practice: Planning Trials and Analyzing Data for Personalized Medicine*. Vol 21. Philadelphia, PA: SIAM; 2015.
4. Linn KA, Laber EB, Stefanski LA. Interactive Q-learning for quantiles. *J Am Stat Assoc*. 2017;112(518):638-649.
5. Wang L, Zhou Y, Song R, Sherwood B. Quantile-optimal treatment regimes. *J Am Stat Assoc*. 2018;113(523):1243-1254.
6. Henderson R, Ansell P, Alshibani D. Regret-regression for optimal dynamic treatment regimes. *Biometrics*. 2010;66(4):1192-1201.
7. Barrett JK, Henderson R, Rosthøj S. Doubly robust estimation of optimal dynamic treatment regimes. *Stat Biosci*. 2014;6(2):244-260.
8. Rich B, Moodie EE, Stephens DA. Simulating sequential multiple assignment randomized trials to generate optimal personalized warfarin dosing strategies. *Clin Trials*. 2014;11(4):435-444.
9. Wang L, Rotnitzky A, Lin X, Millikan RE, Thall PF. Evaluation of viable dynamic treatment regimes in a sequentially randomized trial of advanced prostate cancer. *J Am Stat Assoc*. 2012;107(498):493-508.
10. Xu Y, Thall PF, Hua W, Andersson BS. Bayesian non-parametric survival regression for optimizing precision dosing of intravenous busulfan in allogeneic stem cell transplantation. *J Royal Stat Soc Ser C (Appl Stat)*. 2019;68(3):809-828.
11. Nahum-Shani I, Qian M, Almirall D, et al. Q-learning: a data analysis method for constructing adaptive interventions. *Psychol Methods*. 2012;17(4):478.
12. Laber EB, Lizotte DJ, Qian M, Pelham WE, Murphy SA. Dynamic treatment regimes: technical challenges and applications. *Electr J Stat*. 2014;8(1):1225.
13. Zhang Y, Laber EB, Davidian M, Tsiatis AA. Estimation of optimal treatment regimes using lists. *J Am Stat Assoc*. 2017; Just-Accepted. 113(524):1541–1549.
14. van der Laan MJ, Petersen ML. Causal effect models for realistic individualized treatment and intention to treat rules. *Int J Biostat*. 2007;3(1):1–52.
15. Petersen ML, Porter KE, Gruber S, Wang Y, van der Laan MJ. Diagnosing and responding to violations in the positivity assumption. *Stat Methods Med Res*. 2012;21(1):31-54.
16. Young JG, Cain LE, Robins JM, O'Reilly EJ, Hernán MA. Comparative effectiveness of dynamic treatment regimes: an application of the parametric g-formula. *Stat Biosci*. 2011;3(1):119.
17. Shortreed SM, Laber E, Scott Stroup T, Pineau J, Murphy SA. A multiple imputation strategy for sequential multiple assignment randomized trials. *Stat Med*. 2014;33(24):4202-4214.
18. Blatt D, Murphy S, Zhu J. *A-Learning for Approximate Planning*. Ann Arbor, MI: University of Michigan; 2004.
19. Murphy SA. A generalization error for Q-learning. *J Mach Learn Res*. 2005;6(July):1073-1097.
20. Moodie EE, Richardson TS, Stephens DA. Demystifying optimal dynamic treatment regimes. *Biometrics*. 2007;63(2):447-455.
21. Schulte PJ, Tsiatis AA, Laber EB, Davidian M. Q-and A-learning methods for estimating optimal dynamic treatment regimes. *Stat Sci*. 2014;29(4):640.
22. Zhao Y, Kosorok MR, Zeng D. Reinforcement learning design for cancer clinical trials. *Stat Med*. 2009;28(26):3294-3315.
23. Goldberg Y, Kosorok MR. Q-learning with censored data. *Ann Stat*. 2012;40(1):529.
24. Lu W, Zhang HH, Zeng D. Variable selection for optimal treatment decision. *Stat Methods Med Res*. 2013;22(5):493-504.
25. Moodie EE, Dean N, Sun YR. Q-learning: flexible learning about useful utilities. *Stat Biosci*. 2014;6(2):223-243.
26. Tian L, Alizadeh AA, Gentles AJ, Tibshirani R. A simple method for estimating interactions between a treatment and a large number of covariates. *J Am Stat Assoc*. 2014;109(508):1517-1532.
27. Laber EB, Linn KA, Stefanski LA. Interactive model building for Q-learning. *Biometrika*. 2014;101(4):831-847.
28. Zhou X, Kosorok MR. Causal nearest neighbor rules for optimal treatment regimes; 2017. arXiv preprint arXiv:171108451.
29. Jeng XJ, Lu W, Peng H, et al. High-dimensional inference for personalized treatment decision. *Electr J Stat*. 2018;12(1):2074-2089.
30. Shi C, Fan A, Song R, Lu W. High-dimensional *A*-learning for optimal dynamic treatment regimes. *Ann Stat*. 2018;46(3):925-957.
31. Kosorok MR, Laber EB. Precision medicine. annual review of statistics and its application; 2019; In press.
32. Orellana L, Rotnitzky A, Robins JM. Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, Part I: main content. *Int J Biostat*. 2010;6(2):1–46.
33. Zhang B, Tsiatis AA, Laber EB, Davidian M. A robust method for estimating optimal treatment regimes. *Biometrics*. 2012;68(4):1010-1018.
34. Zhao Y, Zeng D, Rush AJ, Kosorok MR. Estimating individualized treatment rules using outcome weighted learning. *J Am Stat Assoc*. 2012;107(499):1106-1118.
35. Zhang B, Tsiatis AA, Laber EB, Davidian M. Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika*. 2013;100(3):681-694.
36. Zhao YQ, Zeng D, Laber EB, Song R, Yuan M, Kosorok MR. Doubly robust learning for estimating individualized treatment with censored data. *Biometrika*. 2014;102(1):151-168.
37. Zhao YQ, Zeng D, Laber EB, Kosorok MR. New statistical learning methods for estimating optimal dynamic treatment regimes. *J Am Stat Assoc*. 2015;110(510):583-598.
38. Zhou X, Mayer-Hamblett N, Khan U, Kosorok MR. Residual weighted learning for estimating individualized treatment rules. *J Am Stat Assoc*. 2017;112(517):169-187.
39. Athey S, Wager S. Efficient policy learning; 2017. arXiv preprint arXiv:170202896.

40. Zhang B, Zhang M. C-learning: a new classification framework to estimate optimal dynamic treatment regimes. *Biometrics*. 2018;74(3):891-899.
41. Liu Y, Wang Y, Kosorok MR, Zhao Y, Zeng D. Augmented outcome-weighted learning for estimating optimal dynamic treatment regimens. *Stat Med*. 2018;37(26):3776–3788.
42. Luckett DJ, Laber EB, Kahkoska AR, Maahs DM, Mayer-Davis E, Kosorok MR. Estimating dynamic treatment regimes in mobile health using V-learning. *J Am Stat Assoc*. 2018;115(530):692–706.
43. Robins JM. *Causal Inference from Complex Longitudinal Data Latent Variable Modeling and Applications to Causality*. New York, NY: Springer; 1997:69-117.
44. Yu Z, van der Laan MJ. Construction of counterfactuals and the G-computation formula; 2002.
45. Xu Y, Müller P, Wahed AS, Thall PF. Bayesian nonparametric estimation for dynamic treatment regimes with sequential transition times. *J Am Stat Assoc*. 2016;111(515):921-950.
46. Guan Q, Reich BJ, Laber EB, Bandyopadhyay D. Bayesian nonparametric policy search with application to periodontal recall intervals; 2018. arXiv preprint arXiv:181004338.
47. Laber EB, Meyer NJ, Reich BJ, Pacifici K, Collazo JA, Drake JM. Optimal treatment allocations in space and time for on-line control of an emerging infectious disease. *J Royal Stat Soc Ser C (Appl Stat)*. 2018;67(4):743-789.
48. Almirall D, DiStefano C, Chang YC, et al. Longitudinal effects of adaptive interventions with a speech-generating device in minimally verbal children with ASD. *J Clin Child Adolesc Psychol*. 2016;45(4):442-456.
49. Lu X, Nahum-Shani I, Kasari C, et al. Comparing dynamic treatment regimes using repeated-measures outcomes: modeling considerations in SMART studies. *Stat Med*. 2016;35(10):1595-1615.
50. Ertefaie A, Shortreed S, Chakraborty B. Q-learning residual analysis: application to the effectiveness of sequences of antipsychotic medications for patients with schizophrenia. *Stat Med*. 2016;35(13):2221-2234.
51. Nahum-Shani I, Ertefaie A, Lu X, et al. A SMART data analysis method for constructing adaptive treatment strategies for substance use disorders. *Addiction*. 2017;112(5):901-909.
52. Kilbourne AM, Smith SN, Choi SY, et al. Adaptive School-based Implementation of CBT (ASIC): clustered-SMART for building an optimized adaptive implementation intervention to improve uptake of mental health interventions in schools. *Implement Sci*. 2018;13(1):119.
53. Kidwell KM, Seewald NJ, Tran Q, Kasari C, Almirall D. Design and analysis considerations for comparing dynamic treatment regimens with binary outcomes from sequential multiple assignment randomized trials. *J Appl Stat*. 2018;45(9):1628-1651.
54. Seaman SR, White IR, Copas AJ, Li L. Combining multiple imputation and inverse-probability weighting. *Biometrics*. 2012;68(1):129-137.
55. Tsiatis A. *Semiparametric Theory and Missing Data*. New York, NY: Springer Science & Business Media; 2007.
56. Lavori PW, Dawson R. Dynamic treatment regimes: practical design considerations. *Clin Trials*. 2004;1(1):9-20.
57. Murphy SA. An experimental design for the development of adaptive treatment strategies. *Stat Med*. 2005;24(10):1455-1481.
58. Kidwell KM. SMART designs in cancer research: Past, present, and future. *Clin Trials*. 2014;11(4):445-456.
59. Laber E, Zhao Y. Tree-based methods for individualized treatment regimes. *Biometrika*. 2015;102(3):501-514.
60. Chen G, Zeng D, Kosorok MR. Personalized dose finding using outcome weighted learning. *J Am Stat Assoc*. 2016;111(516):1509-1521.
61. Rubin DB. Bayesian inference for causal effects: The role of randomization. *Ann Stat*. 1978;6:34-58.
62. Splawa-Neyman J, Dabrowska DM, Speed T. On the application of probability theory to agricultural experiments. essay on principles. Section 9. *Stat Sci*. 1990;5:465-472.
63. Chakraborty B, Moodie E. *Statistical Methods for Dynamic Treatment Regimes*. New York, NY: Springer; 2013.
64. Qian M, Murphy SA. Performance guarantees for individualized treatment rules. *Ann Stat*. 2011;39(2):1180.
65. Rubin DB. Inference and missing data. *Biometrika*. 1976;63(3):581-592.
66. Little RJ, Rubin DB. *Statistical Analysis with Missing Data*. Vol 333. Hoboken, NJ: John Wiley & Sons; 2014.
67. Tsiatis AA, Kenward MG, Fitzmaurice G, Verbeke G, Molenberghs G. *Handbook of Missing Data Methodology*. Boca Raton, FL: Chapman & Hall/CRC; 2014.
68. Kosorok MR. *Introduction to Empirical Processes and Semiparametric Inference*. New York, NY: Springer Science & Business Media; 2007.
69. Robins JM, Rotnitzky A. Semiparametric efficiency in multivariate regression models with missing data. *J Am Stat Assoc*. 1995;90(429):122-129.
70. Fu WJ. Penalized estimating equations. *Biometrics*. 2003;59(1):126-132.
71. Johnson BA, Lin D, Zeng D. Penalized estimating functions and variable selection in semiparametric regression models. *J Am Stat Assoc*. 2008;103:672-680.
72. Zhao YQ, Laber EB, Ning Y, Saha S, Sands B. Efficient augmentation and relaxation learning for individualized treatment rules using observational data. *J Mach Learn Res*. 2019;48:1–23.
73. Jiang B, Song R, Li J, Zeng D. Entropy learning for dynamic treatment regimes. *Stat Sinica*. 2019;29:1633–1655.
74. Buuren S, Groothuis-Oudshoorn K. Mice: multivariate imputation by chained equations in R. *J Stat Softw*. 2011;45.
75. Stroup TS, McEvoy JP, Swartz MS, et al. The national institute of mental health clinical antipsychotic trials of intervention effectiveness (CATIE) project: schizophrenia trial design and protocol development. *Schizophr Bull*. 2003;29(1):15.
76. Shortreed SM, Laber E, Lizotte DJ, Stroup TS, Pineau J, Murphy SA. Informing sequential clinical decision-making through reinforcement learning: an empirical study. *Mach Learn*. 2011;84(1-2):109-136.
77. Shortreed SM, Moodie EE. Estimating the optimal dynamic antipsychotic treatment regime: evidence from the sequential multiple-assignment randomized clinical antipsychotic trials of intervention and effectiveness schizophrenia study. *J Royal Stat Soc Ser C (Appl Stat)*. 2012;61(4):577-599.

78. Laber EB, Lizotte DJ, Ferguson B. Set-valued dynamic treatment regimes for competing outcomes. *Biometrics*. 2014;70(1):53-61.
79. Busoniu L, Babuska R, De Schutter B, Ernst D. *Reinforcement Learning and Dynamic Programming Using Function Approximators*. Boca Raton, FL: CRC Press; 2010.
80. Geramifard A, Walsh TJ, Tellex S, et al. A tutorial on linear function approximators for dynamic programming and reinforcement learning. *Found Trends Mach Learn*. 2013;6(4):375-451.
81. Sutton RS, Barto AG. Reinforcement Learning: An Introduction: MIT Press; 2018.
82. Bellman R. *Dynamic Programming*. Princeton, NJ: Princeton University Press; 1957:151.
83. Zhang B, Tsiatis AA, Davidian M, Zhang M, Laber E. Estimating optimal treatment regimes from a classification perspective. *Stat*. 2012;1(1):103-114.
84. Cristianini N, Shawe-Taylor J. *An Introduction to Support Vector Machines and other Kernel-based Learning Methods*. Cambridge, MA: Cambridge University Press; 2000.
85. Moguerza JM, Muñoz A. Support vector machines with applications. *Stat Sci*. 2006;21(3):322-336.
86. Berlinet A, Thomas-Agnan C. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. New York, NY: Springer Science & Business Media; 2011.
87. Nosedal-Sanchez A, Storlie CB, Lee TC, Christensen R. Reproducing kernel Hilbert spaces for penalized regression: a tutorial. *Am Stat*. 2012;66(1):50-60.
88. Rubin DB, van der Laan MJ. Statistical issues and limitations in personalized medicine research with clinical trials. *Int J Biostat*. 2012;8.
89. Qi Z, Liu Y. D-learning to estimate optimal individual treatment rules. *Electron J Stat*. 2018a;12(2):3601-3638.
90. Qi Z, Liu D, Fu H, Liu Y. Multi-armed angle-based direct learning for estimating optimal individualized treatment rules with various outcomes. *J Am Stat Assoc*. 2018b;115:1-35.
91. Davidian M, Tsiatis AA, Holloway S, Laber EB. *Introduction to Treatment Regimes*. Boca Raton, FL: Chapman Hall (forthcoming; 2019.
92. Taylor JM, Cheng W, Foster JC. Reader reaction to "A robust method for estimating optimal treatment regimes" by Zhang et al.(2012). *Biometrics*. 2015;71(1):267-273.
93. Wolsey LA, Nemhauser GL. *Integer and Combinatorial Optimization*. Hoboken, NJ: John Wiley & Sons; 2014.
94. Boyd S, Vandenberghe L. *Convex Optimization*. Cambridge, MA: Cambridge University Press; 2004.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Dong L, Laber E, Goldberg Y, Song R, Yang S. Ascertaining properties of weighting in the estimation of optimal treatment regimes under monotone missingness. *Statistics in Medicine*. 2020;39:3503–3520. https://doi.org/10.1002/sim.8678

APPENDIX

Q-learning with complete data

Q-learning is an approximate dynamic programming algorithm that has been applied to estimate optimal treatment regimes in a variety of biomedical and engineering applications.[19,21,79-81] The basis for *Q*-learning is the dynamic programming characterization of an optimal regime. Define $Q_T(\mathbf{h}_T, a_T) = \mathbb{E}(Y|\mathbf{H}_T = \mathbf{h}_T, A_T = a_T)$, and recursively for $t = T-1, T-2, \ldots, 1$ define $Q_t(\mathbf{h}_t, a_t) = \mathbb{E}\{\max_{a_{t+1}\in\Psi_{t+1}(\mathbf{H}_{t+1})} Q_{t+1}(\mathbf{H}_{t+1}, a_{t+1})|\mathbf{H}_t = \mathbf{h}_t, A_t = a_t\}$. It follows from dynamic programming that the optimal regime satisfies $\pi_t^{\text{opt}}(\mathbf{h}_t) = \arg\max_{a_t\in\Psi_t(\mathbf{h}_t)} Q_t(\mathbf{h}_t, a_t)$.[82] Let $Q_t(\mathbf{h}_t, a_t; \boldsymbol{\beta}_t)$ denote a posited class of models for $Q_t(\mathbf{h}_t, a_t)$ indexed by $\boldsymbol{\beta}_t \in \mathcal{B}_t$ for $t = 1, \ldots, T$. The induced class of treatment regimes is thus of the form $\pi_t(\mathbf{h}_t; \boldsymbol{\beta}_t) = \arg\max_{a_t\in\Psi_t(\mathbf{h}_t)} Q_t(\mathbf{h}_t, a_t; \boldsymbol{\beta}_t)$.[83] If $\mathcal{B}_t = \mathbb{R}^{p_t}$ and $Q(\mathbf{h}_t, a_t; \boldsymbol{\beta}_t)$ is differentiable in $\boldsymbol{\beta}_t$ for all $\mathbf{h}_t, a_t \in \mathcal{H}_t \times \mathcal{A}_t$, then $\hat{\boldsymbol{\beta}}_{T,n}$ solves

$$\mathbb{P}_n\{Y - Q_T(\mathbf{H}_T, A_T; \boldsymbol{\beta}_T)\}\nabla_{\boldsymbol{\beta}_T} Q_T(\mathbf{H}_T, A_T; \boldsymbol{\beta}_T) = 0, \tag{A1}$$

and for $t = T-1, T-2, \ldots, 1$ the estimators $\hat{\boldsymbol{\beta}}_{t,n}$ solve

$$\mathbb{P}_n\{\max_{a_{t+1}\in\Psi_t(\mathbf{H}_{t+1})} Q_{t+1}(\mathbf{H}_{t+1}, a_{t+1}; \hat{\boldsymbol{\beta}}_{t+1,n}) - Q_t(\mathbf{H}_t, A_t; \boldsymbol{\beta}_t)\}\nabla_{\boldsymbol{\beta}_t} Q_t(\mathbf{H}_t, A_t; \boldsymbol{\beta}_t) = 0. \tag{A2}$$

Thus, the estimator $\hat{\boldsymbol{\beta}}_n$ is obtained by solving $\mathbb{P}_n \mathbf{m}_n(\boldsymbol{\beta}) = 0$, where $\mathbf{m}_n$ is constructed by stacking Equations (A1) and (A2) for $T-1, \ldots, 1$ so that $\hat{\boldsymbol{\beta}}_n$ is a root of

$$\mathbb{P}_n \begin{bmatrix} \{Y - Q_T(\mathbf{H}_T, A_T; \boldsymbol{\beta}_T)\} \nabla_{\boldsymbol{\beta}_T} Q_T(\mathbf{H}_T, A_T; \boldsymbol{\beta}_T) \\ \{\max_{a_T \in \Psi_T(\mathbf{H}_T)} Q_T(\mathbf{H}_T, a_T; \boldsymbol{\beta}_T) - Q_{T-1}(\mathbf{H}_{T-1}, A_{T-1}; \boldsymbol{\beta}_{T-1})\} \nabla_{\boldsymbol{\beta}_{T-1}} Q_{T-1}(\mathbf{H}_{T-1}, A_{T-1}; \boldsymbol{\beta}_{T-1}) \\ \vdots \\ \{\max_{a_2 \in \Psi_2(\mathbf{H}_2)} Q_2(\mathbf{H}_2, a_2; \boldsymbol{\beta}_2) - Q_1(\mathbf{H}_1, A_1; \boldsymbol{\beta}_1)\} \nabla_{\boldsymbol{\beta}_1} Q_1(\mathbf{H}_1, A_1; \boldsymbol{\beta}_1) \end{bmatrix}.$$

The estimated optimal decision at stage t is thus $\hat{\pi}_{n,t}(\mathbf{h}_t) = \arg\max_{a_t \in \Psi_t(\mathbf{h}_t)} Q_t(\mathbf{h}_t, a_t; \hat{\boldsymbol{\beta}}_{t,n})$. Similar expressions can be obtained for nonparametric variants of Q-learning.[13,22,25] For the purpose of illustration, we briefly describe kernel ridge regression for Q-learning. Suppose that $\mathcal{H}_t \subseteq \mathbb{R}^{p_t}$ for all $t = 1, \ldots, T$. For each t, let $K_t : \mathbb{R}^{p_t} \times \mathbb{R}^{p_t} \to \mathbb{R}$ be symmetric and positive definite and write $\mathbb{H}_t$ to denote the corresponding reproducing kernel Hilbert space with norm $\|\cdot\|_{\mathbb{H}_t}$.[84-87] To approximate Q_T within $\mathbb{H}_T$, one solves for each $a \in \{-1, 1\}$

$$\hat{Q}_{T,n}(\cdot, a) = \arg\min_{f_a \in \mathbb{H}_T} \mathbb{P}_n 1_{A_T = a} \{Y - f_a(\mathbf{H}_T)\}^2 + \lambda_{T,a,n} \|f_a\|^2_{\mathbb{H}_T}, \tag{A3}$$

where $\lambda_{T,a,n} \geq 0$ is a tuning parameter. For each $a \in \{-1, 1\}$ define $I_{T,a} = \{i \ : \ A_{T,i} = a\}$ to be the subset of patients to receive treatment a at time T and define $\mathbb{Z}^{\mathrm{T}}_{T,a}(\mathbf{h}_T) = \{K_T(\mathbf{H}_{T,i}, \mathbf{h}_T)\}_{i \in I_{T,a}}$. It follows that $\hat{Q}_{T,n}(\mathbf{h}_T, a_T) = \mathbb{Z}^{\mathrm{T}}_{T,a_T}(\mathbf{h}_T) \hat{\boldsymbol{\beta}}_{T,a,n}$, where $\hat{\boldsymbol{\beta}}_{T,a,n}$ is a solution of $\mathbb{P}_n 1_{A_T = a_T} \{Y - (1 + \tilde{\lambda}_{T,a,n}) \mathbb{Z}^{\mathrm{T}}_{T,a_T}(\mathbf{H}_T) \boldsymbol{\beta}_{T,a}\} \mathbb{Z}_{T,a_T}(\mathbf{H}_T) = 0$. Note that in the previous estimating equation, we have replaced $\lambda_{T,a,n}$ by $\tilde{\lambda}_{T,a,n}$ to reflect in re-writing the estimator the penalty has been scaled by the number of subjects receiving treatment a_T. Constructing $\mathbb{Z}_{t,a_t}(\mathbf{h}_t)$ analogously for $t = T-1, T-2, \ldots, 1$, one can construct estimators $\hat{Q}_{t,n}(\mathbf{h}_t, a_t) = \mathbb{Z}^{\mathrm{T}}_{t,a_t}(\mathbf{h}_t) \hat{\boldsymbol{\beta}}_{t,a,n}$ where $\hat{\boldsymbol{\beta}}_{t,a,n}$ is a solution of

$$\mathbb{P}_n 1_{A_t = a_t} \{\max_{a_{t+1} \in \Psi_{t+1}(\mathbf{H}_{T+1})} \hat{Q}_{t+1,n}(\mathbf{H}_{t+1}, a_{t+1}) - (1 + \tilde{\lambda}_{t,a,n}) \mathbb{Z}^{\mathrm{T}}_{t,a_t} \boldsymbol{\beta}_{t,a}\} \mathbb{Z}_{t,a_t}(\mathbf{H}_t) = 0.$$

The preceding estimating equations can be stacked to obtain a single estimating equation; see Zhang et al[13] for additional details.

Outcome weighted learning with complete data

Direct-search methods, also known as value-search methods, estimate an optimal treatment regime by maximizing an estimator of the marginal mean outcome over a pre-specified class of regimes.[32,33,35,40,88] Outcome weighted learning (OWL) comprises a subclass of these methods, which rely on the use of a convex surrogate to carry out the proposed maximization (see below for details). Outcome weighted learning was introduced for single-stage decisions by Zhao et al[34] and has since been generalized to multistage decisions[37] and undergone a number of other modifications and refinements.[36,38,41,60,89,90]

We consider a variant of outcome weighted learning that uses a convex relaxation of the augmented inverse probability weighted estimator of the marginal mean outcome.[35,41,72] We use a backwards recursive procedure[37] to extend the single-stage procedure proposed by Zhao et al.[72] to multiple-stages; while this is our not main methodological contribution, it may of be interest in its own right.[40,91]

We describe the estimator as a sequence of models fit, each indexed by their own parameters, before stacking the models and parameters into a single estimating equation. To ease bookkeeping and clarify development, we begin by using $\boldsymbol{\eta} = (\boldsymbol{\eta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\eta}_T^{\mathrm{T}})^{\mathrm{T}}$ solely to index the decision rules and use $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\theta}_T^{\mathrm{T}})^{\mathrm{T}}$ to index nuisance models; we later pool these into a single collection of parameters, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\beta}_T^{\mathrm{T}})^{\mathrm{T}}$, to match the notation used in our general framework. For simplicity, we assume linear decision rules of the form $\pi_t(\mathbf{h}_t; \boldsymbol{\eta}_t) = \mathrm{sign}(\boldsymbol{\eta}_t^{\mathrm{T}} \mathbf{h}_{t,0})$, where $\mathbf{h}_{t,0}$ is known feature vector constructed from $\mathbf{h}_t$. Furthermore, let $Q_T(\mathbf{h}_T, a_T; \boldsymbol{\theta}_T)$ be a posited model for $Q_T(\mathbf{h}_T, a_T) = \mathbb{E}(Y | \mathbf{H}_T = \mathbf{h}_T, A_T = a_T)$. For each $\boldsymbol{\eta}_T$, define $V_T(\mathbf{h}_T, \boldsymbol{\eta}_T) = Q_T\{\mathbf{h}_T, \pi_T(\mathbf{h}_T; \boldsymbol{\eta}_T)\}$ with corresponding posited model $V_T(\mathbf{h}_T, \boldsymbol{\eta}_T, \boldsymbol{\theta}_T) = Q_T\{\mathbf{h}_t, \pi_T(\mathbf{h}_T; \boldsymbol{\eta}_T); \boldsymbol{\theta}_T\}$. Define

$$Q_{T-1}(\mathbf{h}_{T-1}, a_{T-1}, \boldsymbol{\eta}_T, \boldsymbol{\theta}_T) = \mathbb{E}\{V_T(\mathbf{H}_T, \boldsymbol{\eta}_T, \boldsymbol{\theta}_T) | \mathbf{H}_{T-1} = \mathbf{h}_{T-1}, A_{T-1} = a_{T-1}\},$$

thus $Q_{T-1}(\mathbf{h}_{T-1}, a_{T-1}, \boldsymbol{\eta}_T, \boldsymbol{\theta}_T)$ is the mean outcome for a patient presenting with $\mathbf{H}_{T-1} = \mathbf{h}_{T-1}$, treated with $A_{T-1} = a_{t-1}$, and subsequently treated according to $\pi_T(\cdot; \boldsymbol{\eta}_T)$ assuming that the model $Q_T(\mathbf{h}_T, a_T; \boldsymbol{\theta}_T)$ is correct.

Let $Q_{T-1}(\mathbf{h}_{T-1}, a_{T-1}, \boldsymbol{\eta}_T, \boldsymbol{\theta}_T; \boldsymbol{\theta}_{T-1})$ be a posited model for $Q_{T-1}(\mathbf{h}_{T-1}, a_{T-1}, \boldsymbol{\eta}_T, \boldsymbol{\theta}_T)$. Using an underbar to denote future, for example, $\underline{\boldsymbol{\eta}}_t = (\boldsymbol{\eta}_t^{\mathrm{T}}, \dots, \boldsymbol{\eta}_T^{\mathrm{T}})^{\mathrm{T}}$, define

$$V_{T-1}(\mathbf{h}_{T-1}, \underline{\boldsymbol{\eta}}_{T-1}, \underline{\boldsymbol{\theta}}_{T-1}) = Q_{T-1}\{\mathbf{h}_{T-1}, \pi_{T-1}(\mathbf{h}_{T-1}; \boldsymbol{\eta}_{T-1}), \boldsymbol{\eta}_T, \boldsymbol{\theta}_T; \boldsymbol{\theta}_{T-1}\},$$

to be the expected outcome for a patient presenting with $\mathbf{H}_{T-1} = \mathbf{h}_{T-1}$ and treated according to $\pi_{T-1}(\cdot; \boldsymbol{\eta}_{T-1})$ and $\pi_T(\cdot; \boldsymbol{\eta}_T)$ at times $T-1$ and T under the models indexed by $\underline{\boldsymbol{\theta}}_{T-1}$. Recursively, for $t = T-2, \dots, 1$ define

$$Q_t(\mathbf{h}_t, a_t, \underline{\boldsymbol{\eta}}_{t+1}, \underline{\boldsymbol{\theta}}_{t+1}) = \mathbb{E}\{V_{t+1}(\mathbf{H}_{t+1}, \underline{\boldsymbol{\eta}}_{t+1}, \underline{\boldsymbol{\theta}}_{t+1}) | \mathbf{H}_t = \mathbf{h}_t, A_t = a_t\},$$

and let $Q_t(\mathbf{h}_t, a_t, \underline{\boldsymbol{\eta}}_{t+1}, \underline{\boldsymbol{\theta}}_{t+1}; \boldsymbol{\theta}_t)$ denote a posited model. Subsequently, define

$$V_t(\mathbf{h}_t, \underline{\boldsymbol{\eta}}_t, \underline{\boldsymbol{\theta}}_t) = Q_t\{\mathbf{h}_t, \pi_t(\mathbf{h}_t; \boldsymbol{\eta}_t), \underline{\boldsymbol{\eta}}_{t+1}, \underline{\boldsymbol{\theta}}_{t+1}; \boldsymbol{\theta}_t\}.$$

Thus, for any regime $\boldsymbol{\pi}_{\boldsymbol{\eta}}$ indexed by $\boldsymbol{\eta} = \underline{\boldsymbol{\eta}}_1$ the marginal mean outcome under the models indexed by $\boldsymbol{\theta} = \underline{\boldsymbol{\theta}}_1$ is $V(\boldsymbol{\pi}_{\boldsymbol{\eta}}) = \mathbb{E}V_1(\mathbf{H}_1, \boldsymbol{\eta}, \boldsymbol{\theta})$. This construction is the basis for Q-learning with policy-search;[13,92] however, we will not be using the Q-functions in this way. Our purpose in deriving them is to include them as augmentation terms in a doubly robust estimator of the incremental (stagewise) regret.

We assume that $Q_t(\mathbf{h}_t, a_t, \underline{\boldsymbol{\eta}}_{t+1}, \underline{\boldsymbol{\theta}}_{t+1}; \boldsymbol{\theta}_t)$ is continuously differentiable in $\boldsymbol{\theta}_t$ for all $\mathbf{h}_t, a_t, \underline{\boldsymbol{\eta}}_{t+1}, \underline{\boldsymbol{\theta}}_{t+1}$. For each $\boldsymbol{\eta}$, let $\hat{\boldsymbol{\theta}}_n(\boldsymbol{\eta}) = \{\hat{\boldsymbol{\theta}}_{1,n}(\underline{\boldsymbol{\eta}}_2), \hat{\boldsymbol{\theta}}_{2,n}(\underline{\boldsymbol{\eta}}_3), \dots, \hat{\boldsymbol{\theta}}_{T-1,n}(\boldsymbol{\eta}_T), \hat{\boldsymbol{\theta}}_{T,n}\}$ be a root of the estimating equation

$$\mathbb{P}_n \begin{bmatrix} \{Y - Q_T(\mathbf{H}_T, A_T; \boldsymbol{\theta}_T)\} \nabla_{\boldsymbol{\theta}_T} Q_T(\mathbf{H}_T, A_T; \boldsymbol{\theta}_T) \\ \{V_T(\mathbf{H}_T, \boldsymbol{\eta}_T, \boldsymbol{\theta}_T) - Q_{T-1}(\mathbf{H}_{T-1}, A_{T-1}, \boldsymbol{\eta}_T, \boldsymbol{\theta}_T; \boldsymbol{\theta}_{T-1})\} \nabla_{\boldsymbol{\theta}_{T-1}} Q_{T-1}(\mathbf{H}_{T-1}, A_{T-1}, \boldsymbol{\eta}_T, \boldsymbol{\theta}_T; \boldsymbol{\theta}_{T-1}) \\ \vdots \\ \{V_{t+1}(\mathbf{H}_{t+1}, \underline{\boldsymbol{\eta}}_{t+1}, \underline{\boldsymbol{\theta}}_{t+1}) - Q_t(\mathbf{H}_t, A_t, \underline{\boldsymbol{\eta}}_{t+1}, \underline{\boldsymbol{\theta}}_{t+1}; \boldsymbol{\theta}_t)\} \nabla_{\boldsymbol{\theta}_t} Q_t(\mathbf{H}_t, A_t, \underline{\boldsymbol{\eta}}_{t+1}, \underline{\boldsymbol{\theta}}_{t+1}; \boldsymbol{\theta}_t) \\ \vdots \\ \{V_2(\mathbf{H}_2, \underline{\boldsymbol{\eta}}_2, \underline{\boldsymbol{\theta}}_2) - Q_1(\mathbf{H}_1, A_1, \underline{\boldsymbol{\eta}}_2, \underline{\boldsymbol{\theta}}_2; \boldsymbol{\theta}_1)\} \nabla_{\boldsymbol{\theta}_1} Q_1(\mathbf{H}_1, A_1, \underline{\boldsymbol{\eta}}_2, \underline{\boldsymbol{\theta}}_2; \boldsymbol{\theta}_1) \end{bmatrix}, \tag{A4}$$

where we note that $\boldsymbol{\eta}_1$ does not affect $\hat{\boldsymbol{\theta}}_n(\boldsymbol{\eta})$; however, to estimate the performance of $\boldsymbol{\pi}_{\boldsymbol{\eta}}$ with the estimated Q-functions one would use $\mathbb{P}_n Q_1\{\mathbf{H}_1, \pi_1(\mathbf{H}_1; \boldsymbol{\eta}_1), \underline{\boldsymbol{\eta}}_2, \underline{\hat{\boldsymbol{\theta}}}_{2,n}(\underline{\boldsymbol{\eta}}_3); \hat{\boldsymbol{\theta}}_{1,n}(\boldsymbol{\eta}_2)\}$, which can be seen to depend on all of $\boldsymbol{\eta}$.

The second set of estimating equations are based on a backwards recursive representation of an augmented inverse probability weighted estimator of the incremental regret. As noted previously, the estimated Q-functions derived above serve as augmentation terms. Define $\Delta_T(\mathbf{H}_T, A_T; \boldsymbol{\theta}_T) = Q_T(\mathbf{H}_T, A_T; \boldsymbol{\theta}_T) - Q_T(\mathbf{H}_T, -A_T; \boldsymbol{\theta}_T)$ and $\hat{\Delta}_{T,n}(\mathbf{H}_T, A_T) = \Delta_T(\mathbf{H}_T, A_T; \hat{\boldsymbol{\theta}}_{T,n})$. Define the estimated incremental regret at stage T as

$$\begin{aligned} J_{T,n}(\boldsymbol{\eta}_T; \boldsymbol{\theta}_T) &= \mathbb{P}_n \mathbf{1}_{\mathrm{sign}\{W_T(\mathbf{H}_T, A_T, \boldsymbol{\theta}_T)\} A_T \neq \pi_T(\mathbf{H}_T; \boldsymbol{\eta}_T)} |W_T(\mathbf{H}_T, A_T, Y, \boldsymbol{\theta}_T)| \\ &= \mathbb{P}_n \mathbf{1}_{\mathrm{sign}\{W_T(\mathbf{H}_T, A_T, Y, \boldsymbol{\theta}_T)\} A_T \boldsymbol{\eta}_T^{\mathrm{T}} \mathbf{H}_{T,0} < 0} |W_T(\mathbf{H}_T, A_T, Y, \boldsymbol{\theta}_T)|, \end{aligned} \tag{A5}$$

where

$$W_T(\mathbf{H}_T, A_T, Y, \boldsymbol{\theta}_T) = \frac{\{Y - Q_T(\mathbf{H}_T, -A_T; \boldsymbol{\theta}_T) - \{1 - P(A_T|\mathbf{H}_T)\}\Delta_T(\mathbf{H}_T, A_T; \boldsymbol{\theta}_T)\}}{P(A_T|\mathbf{H}_T)}.$$

Define $\hat{J}_{T,n}(\boldsymbol{\eta}_T) = J_{T,n}(\boldsymbol{\eta}_T; \hat{\boldsymbol{\theta}}_{T,n})$. It can be shown that $\hat{J}_{T,n}(\boldsymbol{\eta}_T)$ is, up to an additive constant that does not depend on $\boldsymbol{\eta}_T$, a doubly robust estimator for the difference between: (i) the marginal mean outcome for a patient receiving treatment as per protocol (ie, under the data-generating model) for the first $T-1$ time points, followed by treatment under an optimal regime and (ii) the marginal mean outcome for a patient receiving treatment per protocol for the first $T-1$ time points followed by treatment under $\pi_T(\cdot; \boldsymbol{\eta}_T)$.[91] Thus, $\hat{J}_{T,n}(\boldsymbol{\eta}_T)$ is a measure of the loss incurred by treatment of patients at time T,

who had theretofore been treated per protocol, with $\pi_T(\cdot;\boldsymbol{\eta}_T)$ rather than an optimal treatment. Because of the indicator function, direct minimization of Equation (A5) over $\boldsymbol{\eta}_T$ is a mixed integer program,[93] which is NP-hard in general and thus typically requires the use of either specialized software or the use of heuristics.[35,40] A distinguishing feature of outcome weighted learning is the relaxation of Equation (A5) by replacing the indicator function with a convex function. Given a convex function $L:\mathbb{R}\to\mathbb{R}$, called a convex surrogate, backwards recursive outcome weighted learning uses the objective

$$J^L_{T,n}(\boldsymbol{\eta}_T;\boldsymbol{\theta}_T)=\mathbb{P}_n L[\text{sign}\{W_T(\mathbf{H}_T,A_T,Y,\boldsymbol{\theta}_T)\}A_T\boldsymbol{\eta}_T^{\mathrm{T}}\mathbf{H}_{T,0}]|W_T(\mathbf{H}_T,A_T,Y,\boldsymbol{\theta}_T)|,$$

which can be seen to be a convex function of $\boldsymbol{\eta}_T$ for each $\boldsymbol{\theta}_T$. Define $\hat{J}^L_{T,n}(\boldsymbol{\eta}_T)=J^L_{T,n}(\boldsymbol{\eta}_T;\hat{\boldsymbol{\theta}}_{T,n})$. We define $\hat{\boldsymbol{\eta}}_{T,n}$ to be the solution to $\nabla_{\boldsymbol{\eta}_T}\hat{J}^L_{T,n}(\boldsymbol{\eta}_T)=0$, where the gradient can be replaced with a sub-gradient if L is not differentiable.[94] Thus, the estimated optimal rule at stage T is $\hat{\pi}_T(\cdot;\hat{\boldsymbol{\eta}}_{T,n})$.

Estimators of the optimal decision rules at stages $t=T-1,T-2,\ldots,1$ solve analogous estimating equations, which are defined recursively as follows. For $t=T-1$ define

$$\Delta_{T-1}(\mathbf{H}_{T-1},A_{T-1};\boldsymbol{\eta}_T,\underline{\boldsymbol{\theta}}_{T-1})=Q_{T-1}\{\mathbf{H}_{T-1},A_{T-1},\boldsymbol{\eta}_T,\boldsymbol{\theta}_T;\boldsymbol{\theta}_{T-1}\}-Q_{T-1}\{\mathbf{H}_{T-1},-A_{T-1},\boldsymbol{\eta}_T,\boldsymbol{\theta}_T;\boldsymbol{\theta}_{T-1}\},$$

and define $\hat{\Delta}_{T-1,n}(\mathbf{H}_{T-1},A_{T-1})=\Delta_{T-1}\{\mathbf{H}_{T-1},A_{T-1};\hat{\underline{\boldsymbol{\theta}}}_{T-1,n}(\hat{\boldsymbol{\eta}}_{T,n})\}$, where $\hat{\underline{\boldsymbol{\theta}}}_{T-1,n}(\hat{\boldsymbol{\eta}}_{T,n})=\{\hat{\boldsymbol{\theta}}_{T-1,n}(\hat{\boldsymbol{\eta}}_{T,n}),\hat{\boldsymbol{\theta}}_T\}$. Subsequently, define the relaxed loss function

$$J^L_{T-1,n}(\underline{\boldsymbol{\eta}}_{T-1};\underline{\boldsymbol{\theta}}_{T-1})=\mathbb{P}_n L[\text{sign}\{W_{T-1}(\mathbf{H}_T,A_T,Y,\underline{\boldsymbol{\theta}}_{T-1},\boldsymbol{\eta}_T)\}A_{T-1}\boldsymbol{\eta}_{T-1}^{\mathrm{T}}\mathbf{H}_{T-1,0}]|W_{T-1}(\mathbf{H}_T,A_T,Y,\underline{\boldsymbol{\theta}}_{T-1},\boldsymbol{\eta}_T)|,$$

where the weights are given by

$$\begin{aligned}W_{T-1}\{\mathbf{H}_T,A_T,Y,\underline{\boldsymbol{\theta}}_{T-1},\boldsymbol{\eta}_T\}=&\frac{1_{A_T=\pi_T(\mathbf{H}_T;\boldsymbol{\eta}_T)}Y}{P(A_T|\mathbf{H}_T)P(A_{T-1}|\mathbf{H}_{T-1})}\\&-\frac{Q_{T-1}\{\mathbf{H}_{T-1},-A_{T-1};\boldsymbol{\eta}_T,\underline{\boldsymbol{\theta}}_{T-1}\}}{P(A_{T-1}|\mathbf{H}_{T-1})}-\frac{\{1-P(A_{T-1}|\mathbf{H}_{T-1})\}\Delta_{T-1,n}(\mathbf{H}_{T-1},A_{T-1};\boldsymbol{\eta}_T,\underline{\boldsymbol{\theta}}_{T-1})}{P(A_{T-1}|\mathbf{H}_{T-1})}\\&-\frac{1_{A_T=\pi_T(\mathbf{H}_T;\boldsymbol{\eta}_T)}}{P(A_T|\mathbf{H}_T)P(A_{T-1}|\mathbf{H}_{T-1})}\left\{\frac{1_{A_T=\pi_T(\mathbf{H}_T;\boldsymbol{\eta}_T)}-P(A_T|\mathbf{H}_T)}{P(A_T|\mathbf{H}_T)}\right\}Q_{T,n}\{\mathbf{H}_T,\pi_T(\mathbf{H}_T;\boldsymbol{\eta}_T);\boldsymbol{\theta}_T\}.\end{aligned}$$

Define $\hat{J}^L_{T-1,n}(\boldsymbol{\eta}_{T-1})=J_{T-1}\{\boldsymbol{\eta}_{T-1},\hat{\boldsymbol{\eta}}_{T,n},\hat{\underline{\boldsymbol{\theta}}}_{T-1,n}(\hat{\boldsymbol{\eta}}_{T,n})\}$, and let $\hat{\boldsymbol{\eta}}_{T-1,n}$ be a solution to $\nabla_{\boldsymbol{\eta}_{T-1}}\hat{J}^L_{T-1,n}(\boldsymbol{\eta}_{T-1})=0$. For a generic $t<T-1$, define

$$\Delta_t(\mathbf{H}_t,A_t;\underline{\boldsymbol{\theta}}_t,\underline{\boldsymbol{\eta}}_{t+1})=Q_t(\mathbf{H}_t,A_t,\underline{\boldsymbol{\eta}}_{t+1},\underline{\boldsymbol{\theta}}_{t+1};\boldsymbol{\theta}_t)-Q_t(\mathbf{H}_t,-A_t,\underline{\boldsymbol{\eta}}_{t+1},\underline{\boldsymbol{\theta}}_{t+1};\boldsymbol{\theta}_t),$$

and define $\hat{\Delta}_{t,n}(\mathbf{H}_t,A_t)=\Delta\{\mathbf{H}_t,A_t;\hat{\underline{\boldsymbol{\theta}}}_{t,n}(\hat{\underline{\boldsymbol{\eta}}}_{t+1,n}),\hat{\underline{\boldsymbol{\eta}}}_{t+1,n}\}$, where $\hat{\underline{\boldsymbol{\theta}}}_{t,n}(\hat{\underline{\boldsymbol{\eta}}}_{t+1,n})=\{\hat{\boldsymbol{\theta}}_{t,n}(\hat{\underline{\boldsymbol{\eta}}}_{t+1,n}),\hat{\boldsymbol{\theta}}_{t+1,n}(\hat{\underline{\boldsymbol{\eta}}}_{t+2,n}),\ldots,\hat{\boldsymbol{\theta}}_{T,n}\}$. Subsequently, define

$$J^L_t(\underline{\boldsymbol{\eta}}_t,\underline{\boldsymbol{\theta}}_t)=\mathbb{P}_n L[\text{sign}\{W_t(\mathbf{H}_T,A_T,Y,\underline{\boldsymbol{\eta}}_{t+1},\underline{\boldsymbol{\theta}}_t)\}A_t\boldsymbol{\eta}_t^{\mathrm{T}}\mathbf{H}_{t,0}]|W_t(\mathbf{H}_T,A_T,Y,\underline{\boldsymbol{\eta}}_{t+1},\underline{\boldsymbol{\theta}}_t)|,$$

where the weights are given by

$$\begin{aligned}W_t(\mathbf{H}_T,A_T,Y,\underline{\boldsymbol{\eta}}_{t+1},\underline{\boldsymbol{\theta}}_t)=&\frac{Y\prod_{s=t+1}^T 1_{A_s=\pi_s(\mathbf{H}_s;\boldsymbol{\eta}_s)}}{\prod_{s=t}^T P(A_s|\mathbf{H}_s)}\\&-\frac{Q_t(\mathbf{H}_t,-A_t,\underline{\boldsymbol{\eta}}_{t+1},\underline{\boldsymbol{\theta}}_{t+1};\boldsymbol{\theta}_t)}{P(A_t|\mathbf{H}_t)}-\frac{\{1-P(A_t|\mathbf{H}_t)\}\Delta_{t,n}(\mathbf{H}_t,A_t,\underline{\boldsymbol{\eta}}_{t+1},\boldsymbol{\theta}_t)}{P(A_t|\mathbf{H}_t)}\\&-\sum_{r=t+1}^T\left[\frac{\prod_{s=t+1}^{r-1}1_{A_s=\pi(\mathbf{H}_s;\boldsymbol{\eta}_s)}}{\prod_{s=t}^{r-1}P(A_s|\mathbf{H}_s)}\left\{\frac{1_{A_r=\pi_r(\mathbf{H}_t;\boldsymbol{\eta}_r)}-P(A_r|\mathbf{H}_r)}{P(A_r|\mathbf{H}_r)}\right\}Q_r\{\mathbf{H}_r,\pi_r(\mathbf{H}_r;\boldsymbol{\eta}_r);\underline{\boldsymbol{\eta}}_{r+1},\underline{\boldsymbol{\theta}}_{r+1};\boldsymbol{\theta}_r\}\right].\end{aligned}$$

Define $\hat{J}^L_{t,n}(\boldsymbol{\eta}_t)=J^t\{\boldsymbol{\eta}_t,\hat{\underline{\boldsymbol{\eta}}}_{t+1,n},\hat{\underline{\boldsymbol{\theta}}}_{t,n}(\hat{\underline{\boldsymbol{\eta}}}_{t+1,n})\}$, and let $\hat{\boldsymbol{\eta}}_{t,n}$ be a solution to $\nabla_{\boldsymbol{\eta}_t}\hat{J}^L_{t,n}(\boldsymbol{\eta}_t)=0$.

The joint estimating equations for $\boldsymbol{\beta} = (\boldsymbol{\eta}^{\mathrm{T}}, \boldsymbol{\theta}^{\mathrm{T}})^{\mathrm{T}}$ are obtained by stacking Equation (A4) and the estimating equations $\nabla_{\boldsymbol{\eta}_t} J_t(\underline{\boldsymbol{\eta}}_t, \underline{\boldsymbol{\theta}}_t) = 0$ for $t = 1, \ldots, T$.

Proof of Theorem 1

By stacking the estimating equations for $\boldsymbol{\psi}$ and IPWCC estimating equations for $\boldsymbol{\beta}$, we construct a Z-estimator $(\hat{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\beta}}_n)$ that solves

$$\mathbb{P}_n \begin{bmatrix} 1_{C=T}\nabla_{\boldsymbol{\psi}_T} g_T(\mathbf{S}_T; \boldsymbol{\psi}_T) - 1_{C\geq T}\nabla_{\boldsymbol{\psi}_T} g_T(\mathbf{S}_T; \boldsymbol{\psi}_T)\mathrm{expit}\{g_T(\mathbf{S}_T; \boldsymbol{\psi}_T)\} \\ \vdots \\ 1_{C=1}\nabla_{\boldsymbol{\psi}_1} g_1(\mathbf{S}_1; \boldsymbol{\psi}_1) - 1_{C\geq 1}\nabla_{\boldsymbol{\psi}_1} g_1(\mathbf{S}_1; \boldsymbol{\psi}_1)\mathrm{expit}\{g_1(\mathbf{S}_1; \boldsymbol{\psi}_1)\} \\ \mathbf{w}^{\mathrm{cc}}(C, \mathbf{S}_T; \boldsymbol{\psi}) \cdot \mathbf{m}_n(\mathbf{H}_T, A_T, Y; \boldsymbol{\beta}) \end{bmatrix} = \mathbb{P}_n \check{\mathbf{m}}_n(\mathbf{S}_T, Y; \boldsymbol{\psi}, \boldsymbol{\beta}) = 0,$$

where $\mathbf{w}^{\mathrm{cc}}(C, \mathbf{S}_T; \boldsymbol{\psi})$ and $\mathbf{m}_n(\mathbf{H}_T, A_T, Y; \boldsymbol{\beta})$ are constructed by aggregating $w_j^{\mathrm{cc}}(C, \mathbf{S}_{t-1}; \overline{\boldsymbol{\psi}}_{t-1})$ and $\tilde{m}_{n,j}(\mathbf{S}_t; \boldsymbol{\beta})$ in the order of $j \in \mathcal{J}_{T+1} \cup \mathcal{J}_T \cup \ldots \cup \mathcal{J}_1$ respectively; the $\cdot$ notation is the elementwise product operator for vectors.

Assume that the following standard regularity conditions for Z-estimator hold.

1. $P\check{\mathbf{m}}_n(\mathbf{S}_T, Y; \boldsymbol{\psi}, \boldsymbol{\beta})$ exists for all $\boldsymbol{\psi}, \boldsymbol{\beta} \in \mathcal{K} \times \mathcal{B}$; there exists $\boldsymbol{\psi}^*, \boldsymbol{\beta}_n^* \in \mathcal{K} \times \mathcal{B}$ such that $\|P\check{\mathbf{m}}_n(\mathbf{S}_T, Y; \boldsymbol{\psi}^*, \boldsymbol{\beta}_n^*)\| = o(1)$, and $\|P\check{\mathbf{m}}_n(\mathbf{S}_T, Y; \boldsymbol{\psi}, \boldsymbol{\beta})\| \neq o(1)$ for $\boldsymbol{\psi} \neq \boldsymbol{\psi}^*, \boldsymbol{\beta} \neq \boldsymbol{\beta}_n^*$.
2. Each function in $\check{\mathbf{m}}_n(\cdot)$ is continuous in $\mathcal{K} \times \mathcal{B}$ and bounded by an integrable function of the data that does not depend on $(\boldsymbol{\psi}, \boldsymbol{\beta})$.

We show the consistency of $\hat{\boldsymbol{\beta}}_n$ given that $\lambda_t(\mathbf{s}_t) = \lambda_t(\mathbf{s}_t; \boldsymbol{\psi}_t^*)$. Assuming the correctness of the hazard models, it follows that $K_t(\mathbf{s}_t) = K_t(\mathbf{s}_t; \overline{\boldsymbol{\psi}}_t^*)$ for all t and $\mathbf{s}_t$, for some $\boldsymbol{\psi}^* \in \mathcal{K}$ and that the $\hat{\boldsymbol{\psi}}_n \to \boldsymbol{\psi}^*$ in probability. We only need to show

$$\mathbb{E}\{w_j^{\mathrm{cc}}(C, \mathbf{S}_{t-1}; \overline{\boldsymbol{\psi}}_{t-1}^*)\tilde{m}_{n,j}(\mathbf{S}_t; \boldsymbol{\beta})\} = \mathbb{E}\{\tilde{m}_{n,j}(\mathbf{S}_t; \boldsymbol{\beta})\}, j \in \mathcal{J}_t, t = 1, \ldots T+1.$$

Take the double expectation with the inner expectation condition on $\mathbf{S}_t$ and by MAR assumption, we have

$$\begin{aligned} \mathbb{E}\left\{ \frac{1_{C>t-1}}{K_{t-1}(\mathbf{S}_{t-1})} \tilde{m}_{n,j}(\mathbf{S}_t; \boldsymbol{\beta}) \right\} &= \mathbb{E}\left[\mathbb{E}\left\{ \frac{1_{C>t-1}}{K_{t-1}(\mathbf{S}_{t-1})} \tilde{m}_{n,j}(\mathbf{S}_t; \boldsymbol{\beta}) \middle| \mathbf{S}_t \right\} \right] \\ &= \mathbb{E}\left[\tilde{m}_{n,j}(\mathbf{S}_t; \boldsymbol{\beta}) \frac{\mathbb{E}(1_{C>t-1}|\mathbf{S}_t)}{K_{t-1}(\mathbf{S}_{t-1})} \right] \\ &= \mathbb{E}\{\tilde{m}_{n,j}(\mathbf{S}_t; \boldsymbol{\beta})\} \end{aligned}$$

Therefore, if $\boldsymbol{\beta}_n^*$ satisfies $\|P\mathbf{m}_n(\mathbf{S}_{T+1}; \boldsymbol{\beta}_n^*)\| = o(1)$, we have $\|P\check{\mathbf{m}}_n(\mathbf{S}_{T+1}; \boldsymbol{\beta}_n^*, \hat{\boldsymbol{\psi}}_n)\| = o_p(1)$.

Proof of Theorem 2

We prove Theorem 2 by showing the consistency in either one of two scenarios: (i) the missingness model is correctly specified, that is, $\lambda_t(\mathbf{s}_t) = \lambda_t(\mathbf{s}_t; \boldsymbol{\psi}_t^*)$ for all $\mathbf{s}_t$ for some $\boldsymbol{\psi}_t^* \in \mathcal{K}_t$ and $\hat{\boldsymbol{\psi}}_n \to \boldsymbol{\psi}^*$ in probability; (ii) the conditional expectation is correctly specified, that is, $\mathbf{d}_{n,t}(\mathbf{s}_t) = \mathbf{d}_{n,t}(\mathbf{s}_t; \boldsymbol{\alpha}^*)$ for some $\boldsymbol{\alpha}^* \in \mathcal{A}$.

By stacking together the AIPWCC estimation equation for $\boldsymbol{\beta}$, estimating equation for $\boldsymbol{\psi}$ and $\boldsymbol{\alpha}$, we form a unified Z-estimator representation $\mathbb{P}_n\check{\mathbf{m}}_n(\mathbf{S}_{T+1}; \boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{\alpha}) = 0$. Under same regularity conditions as in proof of Theorem 1 for general Z-estimator, it is sufficient to show for $t = 1, \ldots, T+1$, one has

$$\mathbb{E}\left\{ w_j^{\mathrm{cc}}(C, \mathbf{S}_{t-1}; \hat{\overline{\boldsymbol{\psi}}}_{t-1,n})\tilde{m}_{n,j}(\mathbf{S}_t; \boldsymbol{\beta}) + \sum_{r=1}^{t-1} w_{r,j}^{\mathrm{aug}}(C, \mathbf{S}_r; \hat{\overline{\boldsymbol{\psi}}}_{r,n}) d_{n,r,j}(\mathbf{S}_r; \hat{\boldsymbol{\alpha}}_{r,n}) \right\} = \mathbb{E}\{\tilde{m}_{n,j}(\mathbf{S}_t; \boldsymbol{\beta})\} \quad \text{for all} \quad j \in \mathcal{J}_t.s \tag{A6}$$

By lemma 10.4 of Tsiatis,[55] we have

$$1 - \frac{1_{C>t}}{K_t(\mathbf{S}_t)} = \sum_{r=1}^{t} \left[\frac{1_{C=r} - \lambda_r(\mathbf{S}_r, \psi_r)1_{C\geq r}}{K_r(\mathbf{S}_r, \overline{\boldsymbol{\psi}}_r)} \right], \quad \text{for all} \quad t = 1, \ldots, T. \tag{A7}$$

Therefore

$$\frac{1_{C>t-1}}{K_{t-1}(\mathbf{S}_{t-1},\overline{\boldsymbol{\psi}}_{t-1})}\tilde{m}_{n,j}(\mathbf{S}_t;\boldsymbol{\beta})=\tilde{m}_{n,j}(\mathbf{S}_t;\boldsymbol{\beta})+\left\{\frac{1_{C>t-1}}{K_{t-1}(\mathbf{S}_{t-1};\overline{\boldsymbol{\psi}}_{t-1})}-1\right\}\tilde{m}_{n,j}(\mathbf{S}_t;\boldsymbol{\beta}),$$

and the AIPWCC estimation equation on the left of Equation (A6) can be rearranged as

$$\tilde{m}_{n,j}(\mathbf{S}_t;\boldsymbol{\beta})-\sum_{r=1}^{t-1}\frac{1_{C=r}-\lambda_r(\mathbf{S}_r;\hat{\boldsymbol{\psi}}_{r,n})1_{C\geq r}}{K_r(\mathbf{S}_r;\hat{\overline{\boldsymbol{\psi}}}_{r,n})}\{\tilde{m}_{n,j}(\mathbf{S}_t;\boldsymbol{\beta})-d_{n,r,j}(\mathbf{S}_r;\hat{\boldsymbol{\alpha}}_{r,n})\}.$$

As we assume that the first term $\|P\tilde{m}_{n,j}(\mathbf{S}_t;\boldsymbol{\beta})\|=o(1)$, now it is sufficient to show that

$$\mathbb{E}\left[\frac{1_{C=r}-\lambda_r(\mathbf{S}_r;\hat{\boldsymbol{\psi}}_{r,n})1_{C\geq r}}{K_r(\mathbf{S}_r;\hat{\overline{\boldsymbol{\psi}}}_{r,n})}\{\tilde{m}_{n,j}(\mathbf{S}_t;\boldsymbol{\beta})-d_{n,r,j}(\mathbf{S}_r;\hat{\boldsymbol{\alpha}}_{r,n})\}\right]=0, \tag{A8}$$

for $r=1,\ldots,t-1$ for a specific $t=1,\ldots,T+1$.

We first prove for scenario 1 where the hazards models are correctly specified so that $\hat{\boldsymbol{\psi}}_n\to\boldsymbol{\psi}^*$. Define a sequence of random vectors $G_r=(\mathbf{S}_t^{\mathrm{T}},1_{C=1},\ldots,1_{C=r-1})^{\mathrm{T}}$ for $r=1,\ldots t-1$. By the law of iterated expectations, we take conditional expectation on G_r. Noting that the only unknown random variable is $1_{C=r}$ given G_r, we have

$$\mathbb{E}\left[\frac{\mathbb{E}(1_{C=r}|G_r)-\lambda_r(\mathbf{S}_r)1_{C\geq r}}{K_r(\mathbf{S}_r)}\{\tilde{m}_{n,j}(\mathbf{S}_t;\boldsymbol{\beta})-d_{n,r,j}(\mathbf{S}_r;\hat{\boldsymbol{\alpha}}_{r,n})\}\right].$$

Because of MAR assumption, we have

$$\mathbb{E}\{1_{C=r}|G_r\}=P(C=r|C\geq r,\mathbf{S}_t)1_{C\geq r}=P(C=r|C\geq r,\mathbf{S}_r)1_{C\geq r}=\lambda_r(\mathbf{S}_r)1_{C\geq r}.$$

Thus, we prove that Equation (A8) holds.

Now we consider scenario 2—the conditional expectation models are correctly specified so that $\mathbf{d}_{n,t}(\mathbf{s}_t)=\mathbf{d}_{n,t}(\mathbf{s}_t;\boldsymbol{\alpha}_t^*)$. Rewrite the left-hand side of Equation (A8) as

$$\mathbb{E}\left[\frac{1_{C=r}}{K_r(\mathbf{S}_r;\hat{\overline{\boldsymbol{\psi}}}_{r,n})}\{\tilde{m}_{n,j}(\mathbf{S}_t;\boldsymbol{\beta})-\mathbf{d}_{n,t}(\mathbf{S}_t)\}\right]-\mathbb{E}\left[\frac{\lambda_r(\mathbf{S}_r,\hat{\boldsymbol{\psi}}_r)1_{C\geq r}}{K_r(\mathbf{S}_r;\hat{\overline{\boldsymbol{\psi}}}_r)}\{\tilde{m}_{n,j}(\mathbf{S}_t;\boldsymbol{\beta})-\mathbf{d}_{n,t}(\mathbf{S}_t)\}\right].$$

Applying the law of iterated expectation again, we take conditional expectation of the first term in the above equation on $(1_{C=r},\mathbf{S}_r^{\mathrm{T}})^{\mathrm{T}}$. This leads to

$$\mathbb{E}\left\{\frac{1_{C=r}}{K_r(\mathbf{S}_r;\overline{\boldsymbol{\psi}}_r)}[\mathbb{E}\{\tilde{m}_{n,j}(\mathbf{S}_t;\boldsymbol{\beta})|1_{C=r},\mathbf{S}_r\}-\mathbb{E}\{\tilde{m}_{n,j}(\mathbf{S}_t;\boldsymbol{\beta})|\mathbf{S}_r\}]\right\}.$$

By MAR, we have $P(\mathbf{S}_t|1_{C=r},\mathbf{S}_r\}=P(\mathbf{S}_t|\mathbf{S}_r\}$, that is, $\mathbb{E}\{\tilde{m}_{n,j}(\mathbf{S}_t;\boldsymbol{\beta})|1_{C=r},\mathbf{S}_r\}=\mathbb{E}\{\tilde{m}_{n,j}(\mathbf{S}_t;\boldsymbol{\beta})|\mathbf{S}_r\}$. This implies the first term in is zero. Similarly, we take conditional expectation of the second term on $(1_{C\geq r},\mathbf{S}_r^{\mathrm{T}})^{\mathrm{T}}$ and obtain

$$\mathbb{E}\left\{\frac{\lambda_r(\mathbf{S}_r,\psi_r)1_{C\geq r}}{K_r(\mathbf{S}_r,\overline{\boldsymbol{\psi}}_r)}[\mathbb{E}\{\tilde{m}_{n,j}(\mathbf{S}_t;\boldsymbol{\beta})|1_{C\geq r},\mathbf{S}_r\}-\mathbb{E}\{\tilde{m}_{n,j}(\mathbf{S}_t;\boldsymbol{\beta})|\mathbf{S}_r\}]\right\}.$$

By MAR again, we have $P(\mathbf{S}_t|1_{C\geq r},\mathbf{S}_r)=P(\mathbf{S}_t|\mathbf{S}_r)$, which implies the second term in is zero.

Therefore, we show that Equation (A8) holds under the two scenarios and the doubly robust conclusion follows.