Published in final edited form as:

Biometrika. 2013 ; 100(3): . doi:10.1093/biomet/ast014.

Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions

Baqun Zhang,
Department of Preventive Medicine, 680 N. Lakeshore Drive, Suite 1400 Northwestern University, Chicago, Illinois, 60611 U.S.A

Anastasios A. Tsiatis,
Department of Statistics, North Carolina State University, Raleigh, North Carolina, 27695-8203, U.S.A

Eric B. Laber, and
Department of Statistics, North Carolina State University, Raleigh, North Carolina, 27695-8203, U.S.A

Marie Davidian
Department of Statistics, North Carolina State University, Raleigh, North Carolina, 27695-8203, U.S.A

Baqun Zhang: baqun.zhang@northwestern.edu; Anastasios A. Tsiatis: tsiatis@ncsu.edu; Eric B. Laber: eblaber@ncsu.edu; Marie Davidian: davidian@ncsu.edu

Summary

A dynamic treatment regime is a list of sequential decision rules for assigning treatment based on a patient's history. Q- and A-learning are two main approaches for estimating the optimal regime, i.e., that yielding the most beneficial outcome in the patient population, using data from a clinical trial or observational study. Q-learning requires postulated regression models for the outcome, while A-learning involves models for that part of the outcome regression representing treatment contrasts and for treatment assignment. We propose an alternative to Q- and A-learning that maximizes a doubly robust augmented inverse probability weighted estimator for population mean outcome over a restricted class of regimes. Simulations demonstrate the method's performance and robustness to model misspecification, which is a key concern.

Keywords

A-learning; Double robustness; Outcome regression; Propensity score; Q-learning

1. Introduction

Treatment of patients with chronic disease involves a series of decisions, where the clinician determines the next treatment to be administered based on all information available to that point. A dynamic treatment regime is a set of sequential decision rules, each corresponding to a decision point in the treatment process. Each rule inputs the available information and

Supplementary Material
Supplementary material available at *Biometrika* online includes technical arguments, more details on the estimators studied, and additional simulation results.

outputs the treatment to be given from among the possible options. The optimal regime is that yielding the most favorable outcome on average if followed by the patient population.

Q- and A-learning are two main approaches for estimating the optimal dynamic treatment regime using data from a clinical trial or observational study. Q-learning (Watkins & Dayan, 1992) involves postulating at each decision point regression models for outcome as a function of patient information to that point. In A-learning (Robins, 2004; Murphy, 2003), models are posited only for the part of the regression involving contrasts among treatments and for treatment assignment at each decision point. Both are implemented through a backward recursive fitting procedure based on a dynamic programming algorithm (Bather, 2000). Under certain assumptions and correct specification of these models, Q- and A-learning lead to consistent estimation of the optimal regime. See Rosthøj et al. (2006), Murphy et al. (2007), Zhao et al. (2009) and Henderson et al. (2010) for applications; related methods are discussed by Robins (2004), Moodie et al. (2007), Robins et al. (2008), Almirall et al. (2010) and Orellana et al. (2010).

A concern with both Q- and A-learning is the effect of model misspecification on the quality of the estimated optimal regime. If one attempts to circumvent this difficulty by using flexible nonparametric regression techniques (Zhao et al., 2009), the estimated optimal rules may be complicated functions of possibly high-dimensional patient information that are difficult to interpret or implement and hence are unappealing to clinicians wary of black box approaches.

Given these drawbacks, we focus on a restricted class of treatment regimes indexed by a finite number of parameters, where the form of regimes in the class may be derived from posited regression models or prespecified on the grounds of interpretability or cost to depend on key subsets of patient information. Zhang et al. (2012) proposed an approach for estimating the optimal regime within such a restricted class for a single treatment decision based on maximizing directly a doubly robust augmented inverse probability weighted estimator for the population mean outcome over all regimes in the class, assuming that larger outcomes are preferred. Via the double robustness property, the estimated optimal regimes enjoy protection against model misspecification and comparable or superior performance than do competing methods. With judicious choice of the augmentation term, increased efficiency of estimation of the mean outcome is achieved, which translates into more precise estimators for the optimal regime.

We adapt this approach to two or more decision points. This is considerably more complex than for one decision and is based on casting the problem as one of monotone coarsening (Tsiatis, 2006, Chapter 7). We focus for simplicity on the case of two treatment options at each decision point, though the methods extend to a finite number of options. The methods lead to estimated optimal regimes achieving comparable performance to those derived via Q- or A-learning under correctly specified models and have the added benefit of protection against misspecification.

2. Framework

Assume there are K prespecified, ordered decision points and an outcome of interest, a function of information collected across all K decisions or ascertained after the Kth decision, with larger values preferred. At each decision $k = 1, \dots, K$, there are two k-specific treatment options coded as 0,1 in the set of options $\mathscr{A}_k$; write a_k to denote an element of $\mathscr{A}_k$. Denote a possible treatment history up to and including decision k as $\bar{a}_k = (a_1, \dots, a_k) \in \mathscr{A}_1 \times \cdots \times \mathscr{A}_k = \bar{\mathscr{A}}_k$.

We consider a potential outcomes framework. For a randomly chosen patient, let X_1 denote baseline covariates recorded prior to the first decision, and let $X_k^*(\bar{a}_{k-1})$ be the covariate information that would accrue between decisions ($k - 1$) and k were s/he to receive treatment history $\bar{a}_{k-1}$ ($k = 2, \ldots, K$), taking values $x_k \in \mathscr{X}_k$. Let $Y^*(\bar{a}_K)$ be the outcome that would result were s/he to receive full treatment history $\bar{a}_K$. Then define the potential outcomes (Robins, 1986) as

$$W=\{X_1, X_2^*(a_1), \ldots, X_K^*(\bar{a}_{K-1}), Y^*(\bar{a}_K)\text{for all}\bar{a}_K \in \bar{\mathscr{A}}_K\}.$$

For convenience later, we include X_1, which is always observed and hence is not strictly a potential outcome, in W, and write $\bar{\mathrm{X}}_k^*(\bar{a}_{k-1})=\{X_1, X_2^*(a_1)\ldots, X_k^*(\bar{a}_{k-1})\}$ and $\bar{x}_k = (x_1, \ldots, x_k)$ for $k = 1, \ldots, K$, where then $\bar{x}_k \in \bar{\mathscr{X}}_k = \mathscr{X}_1 \times \cdots \times \mathscr{X}_k$.

A dynamic treatment regime $g = (g_1, \ldots, g_K)$ is an ordered set of decision rules, where $g_k(\bar{x}_k, \bar{a}_{k-1})$ corresponding to the kth decision takes as input a patient's realized covariate and treatment history up to decision k and outputs a treatment option $a_k \in \Phi_k(\bar{x}_k, \bar{a}_{k-1}) \subseteq \mathscr{A}_k$. In general, $\Phi_k(\bar{x}_k, \bar{a}_{k-1})$ is the set of feasible options at decision k for a patient with realized history $(\bar{x}_k, \bar{a}_{k-1})$, allowing that some options in $\mathscr{A}_k$ may not be possible for patients with certain histories; here, $\Phi_k(\bar{x}_k, \bar{a}_{k-1}) \subseteq \{0, 1\}$. Thus, a feasible treatment regime must satisfy $g_k(\bar{x}_k, \bar{a}_{k-1}) \in \Phi_k(\bar{x}_k, \bar{a}_{k-1})$ ($k = 1, \ldots, K$). Denote the class of all feasible regimes by $\mathscr{G}$

For $g \in \mathscr{G}$ writing $\bar{g}_k = (g_1, \ldots, g_k)$ for $k = 1, \ldots, K$ and $\bar{g}_K = g$, define the potential outcomes associated with g to be $W_g=\{X_1, X_2^*(g_1), \ldots, X_K^*(\bar{g}_{K-1}), Y^*(g)\}$, where $X_k^*(\bar{g}_{k-1})$ is the covariate information that would be seen between decisions $k - 1$ and k were a patient to receive the treatments dictated sequentially by the first $k - 1$ rules in g, and $Y^*(g)$ is the outcome if s/he were to receive the K treatments determined by g. Thus, W_g is an element of W.

Define an optimal treatment regime $g^{\text{opt}}=(g_1^{\text{opt}}, \ldots, g_K^{\text{opt}}) \in \mathscr{G}$ as satisfying

$$E\{Y^*(g^{\text{opt}})\} \geq E\{Y^*(g)\}, g \in \mathscr{G}. \quad (1)$$

That is, g^{opt} is a regime that maximizes expected outcome were all patients in the population to follow it. The optimal regime g^{opt} may be determined via dynamic programming, also referred to as backward induction. At the Kth decision point, for any $\bar{x}_K \in \bar{\mathscr{X}}_K$, $\bar{a}_{K-1} \in \bar{\mathscr{A}}_{K-1}$, define

$$g_K^{\text{opt}}(\bar{\mathrm{x}}_K, \bar{a}_{K-1})=\arg\max_{a_K \in \Phi_K(\bar{\mathrm{x}}_K, \bar{a}_{K-1})} E\{Y^*(\bar{a}_{K-1}, a_K)|\bar{\mathrm{X}}_K^*(\bar{a}_{K-1})=\bar{\mathrm{x}}_K\}, \quad (2)$$

$$V_K(\bar{\mathrm{x}}_K, \bar{a}_{K-1})=\max_{a_K \in \Phi_K(\bar{\mathrm{x}}_K, \bar{a}_{K-1})} E\{Y^*(\bar{a}_{K-1}, a_K)|\bar{\mathrm{X}}_K^*(\bar{a}_{K-1})=\bar{\mathrm{x}}_K\}. \quad (3)$$

For $k = K - 1, \ldots, 2$ and any $\bar{x}_k \in \bar{\mathscr{X}}_k$, $\bar{a}_{k-1} \in \bar{\mathscr{A}}_{k-1}$, define

$$\begin{aligned} g_k^{\text{opt}}(\bar{\mathrm{x}}_k, \bar{a}_{k-1})&=\arg\max_{a_k \in \Phi_k(\bar{\mathrm{x}}_k, \bar{a}_{k-1})} E[V_{k+1}\{\bar{\mathrm{x}}_k, X_{k+1}^*(\bar{a}_{k-1}, a_k), \bar{a}_{k-1}, a_k\}|\bar{\mathrm{X}}_k^*(\bar{a}_{k-1})=\bar{\mathrm{x}}_k], \\ V_k(\bar{\mathrm{x}}_k, \bar{a}_{k-1})&=\max_{a_k \in \Phi_k(\bar{\mathrm{x}}_k, \bar{a}_{k-1})} E[V_{k+1}\{\bar{\mathrm{x}}_k, X_{k+1}^*(\bar{a}_{k-1}, a_k), \bar{a}_{k-1}, a_k\}|\bar{\mathrm{X}}_k^*(\bar{a}_{k-1})=\bar{\mathrm{x}}_k]. \end{aligned}$$

For $k = 1$, $x_1 \in \mathscr{X}_1$, $g_1^{\text{opt}}(x_1)=\arg\ \max_{a_1\in\Phi_1(x_1)}E[V_2\{x_1,X_2^*(a_1),a_1\}|X_1=x_1]$ and $V_1(x_1)=\max_{a_1\in\Phi_1(x_1)}E[V_2\{x_1,X_2^*(a_1),a_1\}|X_1=x_1]$. Thus, g_K^{opt} yields the treatment option at decision K that maximizes the expected potential outcome given prior covariate and treatment history. At decisions $k = K - 1, \ldots, 1$, g_k^{opt} dictates the option that maximizes the expected potential outcome that would be achieved if the optimal rules were followed in the future. An argument that g^{opt}, so defined, satisfies (1) is given in an unpublished report by Schulte et al. (2013) available from the last author.

This definition of an optimal regime is intuitively given in terms of potential outcomes. In practice, with the exception of X_1, W cannot be observed for any patient. Rather, a patient is observed to experience only a single treatment history. Let A_k be the observed treatment received at decision k and let $\bar{A}_k = (A_1, \ldots, A_k)$ be observed treatment history up to decision k. Let X_k be the covariate information observed between decisions $k - 1$ and k under the observed treatment history $\bar{A}_{k-1}$ ($k = 2, \ldots, K$), with history $\bar{X}_k = (X_1, \ldots, X_k)$ for $k = 1, \ldots, K$ to decision k. Let Y be the observed outcome under $\bar{A}_K$. The observed data on a patient are $(\bar{X}_K, \bar{A}_K, Y)$, and the data available from a clinical trial or observational study involving n subjects are independent and identically distributed $(\bar{X}_{Ki}, \bar{A}_{Ki}, Y_i)$ for $i = 1, \ldots, n$.

Under the following standard assumptions, g^{opt} may equivalently be expressed in terms of the observed data. The consistency assumption states that $X_k=X_k^*(\bar{A}_{k-1})=\Sigma_{\bar{a}_{k-1}\in\bar{\mathscr{A}}_{k-1}}X_{k-1}^*(\bar{a}_{k-1})I(\bar{A}_{k-1}=\bar{a}_{k-1})$ for $k = 2, \ldots, K$, and $Y = Y^*(\bar{A}_K) = \sum_{\bar{a}_K \in \bar{\mathscr{A}}_K} Y^*(\bar{a}_K)I(\bar{A}_K = \bar{a}_K)$; that is, a patient's observed covariates and outcome are the same as the potential ones s/he would exhibit under the treatment history actually received. The stable unit treatment value assumption (Rubin, 1978), implies that a patient's covariates and outcome are not influenced by treatments received by other patients. A version of the sequential randomization assumption (Robins, 2004) states that W is independent of A_k conditional on $(\bar{X}_k, \bar{A}_{k-1})$. This is satisfied by default for data from a sequentially randomized clinical trial (Murphy, 2005), but is not verifiable from data from an observational study. It is reasonable to believe that decisions made in an observational study are based on a patient's covariate and treatment history; however, all such information associated with treatment assignment and outcome must be recorded in the $\bar{X}_k$ to validate the assumption.

Under these assumptions, from §1 of the Supplementary Material, $p_{Y^*(\bar{a}_K)|\bar{\mathrm{X}}_K^*(\bar{a}_{K-1})}(y|\bar{\mathrm{x}}_K)=p_{Y|\bar{\mathrm{X}}_K,\bar{A}_K}(y|\bar{\mathrm{x}}_K,\bar{a}_K)$, so that $E\{Y^*(\bar{a}_K)|\bar{\mathrm{X}}_K^*(\bar{a}_{K-1})=\bar{\mathrm{x}}_K\}=E(Y|\bar{\mathrm{X}}_K=\bar{\mathrm{x}}_K,\bar{A}_K=\bar{a}_K)$. Thus, letting $Q_K(\bar{x}_K, \bar{a}_K) = E(Y| \bar{X}_K = \bar{x}_K, \bar{A}_K = \bar{a}_K)$, (2) and (3) become

$$g_K^{\text{opt}}(\bar{\mathrm{x}}_K,\bar{a}_{K-1})=\arg\max_{a_K\in\Phi_K(\bar{\mathrm{x}}_K,\bar{a}_{K-1})}Q_K(\bar{\mathrm{x}}_K,\bar{a}_{K-1},a_K),\ V_K(\bar{\mathrm{x}}_K,\bar{a}_{K-1})=\max_{a_K\in\Phi_K(\bar{\mathrm{x}}_K,\bar{a}_{K-1})}Q_K(\bar{\mathrm{x}}_K,\bar{a}_{K-1},a_K).$$

Using $p_{X_k^*(\bar{a}_{k-1})|\bar{\mathrm{X}}_{k-1}^*(\bar{a}_{k-2})}(\bar{\mathrm{x}}_k|\bar{\mathrm{x}}_{k-1})=p_{X_k|\bar{\mathrm{X}}_{k-1},\bar{A}_{k-1}}(x_k|\bar{\mathrm{x}}_{k-1},\bar{a}_{k-1})$, for $k = K, \ldots, 2$,

$$Q_k(\bar{\mathrm{x}}_k,\bar{a}_k)=E\{V_{k+1}(\bar{\mathrm{x}}_k,X_{k+1},\bar{a}_k)|\bar{\mathrm{X}}_k=\bar{\mathrm{x}}_k,\bar{A}_k=\bar{a}_k\}(k=K-1,\ldots,1),$$
$$g_k^{\text{opt}}(\bar{\mathrm{x}}_k,\bar{a}_{k-1})=\arg\max_{a_k\in\Phi_k(\bar{\mathrm{x}}_k,\bar{a}_{k-1})}Q_k(\bar{\mathrm{x}}_k,\bar{a}_{k-1},a_k)(k=K-1,\ldots,2),$$
$$V_k(\bar{\mathrm{x}}_k,\bar{a}_{k-1})=\max_{a_k\in\Phi_k(\bar{\mathrm{x}}_k,\bar{a}_{k-1})}Q_k(\bar{\mathrm{x}}_k,\bar{a}_{k-1},a_k)(k=K-1,\ldots,2),$$

and $g_1^{\text{opt}}(x_1)=\arg\max_{a_1\in\Phi_1(x_1)}Q_1(x_1,a_1)$, $V_1(x_1) = \max_{a_1\in\Phi_1(x_1)} Q_1(x_1, a_1)$. The $Q_k(\bar{x}_k, \bar{a}_k)$ and $V_k(\bar{x}_k, \bar{a}_{k-1})$ are referred to as Q-functions and value functions and are derived from the distribution of the observed data.

3. Q- and A- learning

Q-learning is based on the developments in §2. Linear or nonlinear models $Q_k(\bar{x}_k, \bar{a}_k; \beta_k)$ in a finite-dimensional parameter β_k may be posited and estimators $\hat{\beta}_k$ obtained via a backward iterative process for $k = K, \ldots, 1$ by solving least squares estimating equations; see §2 of the Supplementary Material. The estimated optimal regime is $\hat{g}_Q^{\text{opt}}=(\hat{g}_{Q,1}^{\text{opt}},\ldots,\hat{g}_{Q,K}^{\text{opt}})$, where $\hat{g}_{Q,1}^{\text{opt}}(x_1)=g_{Q,1}^{\text{opt}}(x_1;\hat{\beta}_1)=\arg\max_{a_1\in\Phi_1(x_1)}Q_1(x_1,a_1;\hat{\beta}_1)$, and $\hat{g}_{Q,k}^{\text{opt}}(\bar{x}_k,\bar{a}_{k-1})=g_{Q,k}^{\text{opt}}(\bar{x}_k,\bar{a}_{k-1};\hat{\beta}_k)=\arg\max_{a_k\in\Phi_k(\bar{x}_k,\bar{a}_{k-1})}Q_k(\bar{x}_k,\bar{a}_{k-1},a_k;\hat{\beta}_k)$ for $k = 2, \ldots, K$. Unless all models are correctly specified, $\hat{g}_Q^{\text{opt}}$ may not be a good estimator for g^{opt}.

The A-learning method we consider is a version of g-estimation (Robins, 2004); see §2 of the Supplementary Material. Write $Q_k(\bar{x}_k, \bar{a}_k)$ as $h_k(\bar{x}_k, \bar{a}_{k-1}) + a_kC_k(\bar{x}_k, \bar{a}_{k-1})$, $h_k(\bar{x}_k, \bar{a}_{k-1}) = Q_k(\bar{x}_k, \bar{a}_{k-1}, 0)$ and $C_k(\bar{x}_k, \bar{a}_{k-1}) = Q_k(\bar{x}_k, \bar{a}_{k-1}, 1) - Q_k(\bar{x}_k, \bar{a}_{k-1}, 0)$. We refer to $C_k(\bar{x}_k, \bar{a}_{k-1})$ as the Q-contrast function; with two treatment options, $A_kC_k(\bar{x}_k, \bar{a}_{k-1})$ is the optimal-blip-to-zero function of Robins (2004). Posit models $C_k(\bar{x}_k, \bar{a}_{k-1}; \psi_k)$ and $C_1(x_1; \psi_1)$, depending on parameters ψ_k; and models $h_k(\bar{x}_k, \bar{a}_{k-1}; \alpha_k)$ and $h_1(x_1; \alpha_1)$, with parameters α_k for $k = K, \ldots, 2$. Let $\pi_k(\bar{x}_k, \bar{a}_{k-1}) = \text{pr}(A_k = 1 \mid \bar{X}_k = \bar{x}_k, \bar{A}_{k-1} = \bar{a}_{k-1})$ and $\pi_1(x_1) = \text{pr}(A_1 = 1 \mid X_1 = x_1)$ be the propensities for treatment, which are unknown unless the data are from a sequentially randomized trial, and specify models $\pi_k(\bar{x}_k, \bar{a}_{k-1}; \gamma_k)$, $k = K, \ldots, 2$, and $\pi_1(x_1; \gamma_1)$, e.g., logistic regression models. Estimators $\hat{\psi}_k$ may be found iteratively for $k = K, \ldots, 1$ by solving for ψ_k and α_k estimating equations given in §2 of the Supplementary Material, substituting the maximum likelihood estimators $\hat{\gamma}_k$. As $Q_k(\bar{x}_k, \bar{a}_k)$ is maximized by $a_k = I\{C_k(\bar{x}_k, \bar{a}_{k-1}) > 0\}$, the estimated optimal regime is $\hat{g}_A^{\text{opt}}=(\hat{g}_{A,1}^{\text{opt}},\ldots,\hat{g}_{A,K}^{\text{opt}})$, where $\hat{g}_{A,1}^{\text{opt}}(x_1)=g_{A,1}^{\text{opt}}(x_1;\hat{\psi}_1)=I\{C_1(x_1;\hat{\psi}_1)>0\}$ and $\hat{g}_{A,k}^{\text{opt}}(\bar{x}_k,\bar{a}_{k-1})=g_{A,k}^{\text{opt}}(\bar{x}_k,\bar{a}_{k-1};\hat{\psi}_k)=I\{C_k(\bar{x}_k,\bar{a}_{k-1};\hat{\psi}_k)>0\}$, for $k = 2, \ldots, K$. If the contrast and propensity models are correctly specified, then $\hat{\psi}_k$ will be consistent for ψ_k even if $h_k(\bar{x}_k, \bar{a}_{k-1}; \alpha_k)$ for $k = K, \ldots, 2$, and $h_1(x_1; \alpha_1)$ are misspecified, and $\hat{g}_A^{\text{opt}}$ will consistently estimate g^{opt}. Thus, the quality of $\hat{g}_A^{\text{opt}}$ depends on how close the $C_k(\bar{x}_k, \bar{a}_{k-1}; \psi_k)$ are to the true contrast functions.

As discussed in §2 of the Supplementary Material, the efficient version of A-learning is so complex as to be infeasible to implement. The implementation of A-learning we use in the empirical studies of §5 is likely as close to efficient as could be hoped in practice.

See the unpublished report of Schulte et al. (2013) for a detailed account of both methods.

4. Proposed robust method

Q- and A-learning are predicated on the postulated models for the Q-functions and Q-contrast functions, respectively, so the resulting estimated regime may be far from g^{opt} if these models are misspecified. We propose an alternative approach that may be robust to such misspecification, based on directly estimating the optimal regime in a specified class of regimes.

Models $Q_k(\bar{x}_k, \bar{a}_k; \beta_k)$ or $C_k(\bar{x}_k, \bar{a}_{k-1}; \psi_k)$, whether correct or not, define classes of regimes $\mathscr{G}_\beta$, $\beta=(\beta_1^{\mathrm{T}}, \ldots, \beta_K^{\mathrm{T}})^{\mathrm{T}}$, or $\mathscr{G}_\psi$, indexed analogously by ψ, whose elements may often be simplified. For example, with $K = 2$, if $C_2(\bar{x}_2, a_1; \psi_2) = \psi_{02} + \psi_{12} x_2$ and $C_1(x_1; \psi_1) = \psi_{01} + \psi_{11} x_1$, the corresponding regimes $g_\psi = (g_{\psi 1}, g_{\psi 2})$ take $g_{\psi 1}(x_1) = I(\psi_{01} + \psi_{11} x_1 > 0)$ and $g_{\psi 2}(\bar{x}_2, a_1) = I(\psi_{02} + \psi_{12} x_2 > 0)$. If prior knowledge suggests that treatment 1 would benefit patients with smaller values of X_1 or X_2, then all reasonable regimes should have $\psi_{11} < 0$ and $\psi_{12} < 0$, and elements of $\mathscr{G}_\psi$ may be expressed in terms of $\eta_1 = -\psi_{01}/\psi_{11}$ and $\eta_2 = -\psi_{02}/\psi_{12}$ as $g_\eta = (g_{\eta 1}, g_{\eta 2})$, $g_{\eta 1}(x_1) = I(\eta_1 > x_1)$ and $g_{\eta 2}(\bar{x}_2, a_1) = I(\eta_2 > x_2)$, $\eta = (\eta_1, \eta_2)^{\mathrm{T}}$.

This suggests considering a class $\mathscr{G}_\eta$, with elements $g_\eta = (g_{\eta 1}, \ldots, g_{\eta K})$, indexed by $\eta=(\eta_1^{\mathrm{T}}, \ldots, \eta_K^{\mathrm{T}})^{\mathrm{T}}$, of form $\{g_{\eta 1}(x_1), \ldots, g_{\eta K}(\bar{x}_K, \bar{a}_{K-1})\}$. If $\mathscr{G}_\eta$ is derived from models $Q_k(\bar{x}_k, \bar{a}_k; \beta_k)$ or $C_k(\bar{x}_k, \bar{a}_{k-1}; \psi_k)$, then $\eta = \eta(\beta)$ or $\eta = \eta(\psi)$ is a many-to-one function of β or ψ, and $g^{\mathrm{opt}} \in \mathscr{G}_\eta$ if these models are correct. Here, estimating $\eta^{\mathrm{opt}} = \arg\max_\eta E\{Y^*(g_\eta)\}$ defining the regime g_η^{opt}, say, will yield an estimator for g^{opt}. If these models are misspecified, $\eta(\hat{\beta})$ or $\eta(\hat{\psi})$ may not converge in probability to η^{opt}, and resulting regimes may be far from optimal. If instead the form of elements of $\mathscr{G}_\eta$ is chosen directly based on interpretability or cost, independently of such models, $\mathscr{G}_\eta$ may or may not contain g^{opt}, but g_η^{opt} is still of interest as the optimal regime among those deemed realistic in practice.

We propose an approach to estimation of g_η^{opt} in a given class $\mathscr{G}_\eta$ by developing an estimator for $E\{Y^*(g_\eta)\}$ that is robust to model misspecification and maximizing it in η. We cast the problem as one of monotone coarsening. Following Tsiatis (2006, §7.1), for fixed η, let $\bar{g}_{\eta k} = (g_{\eta 1}, \ldots, g_{\eta k})$, for $k = 1, \ldots, K$, and let $\bar{g}_{\eta K} = g_\eta$. Identify full data as the potential outcomes $W_{g_\eta}=\{X_1, X_2^*(g_{\eta 1}), \ldots, X_K^*(\bar{g}_{\eta_{K-1}}), Y^*(g_\eta)\}$, and let $\bar{X}_k^*(\bar{g}_{\eta_{k-1}})=\{X_1, X_2^*(g_{\eta 1}), \ldots, X_k^*(\bar{g}_{\eta_{k-1}})\}$. Let $\mathscr{C}_\eta$ be a discrete coarsening variable taking values $1, \ldots, K, \infty$ corresponding to $K + 1$ levels of coarsening, reflecting the extent to which the observed treatments received are consistent with those dictated by g_η. In the general coarsened data set up, when $\mathscr{C}_\eta = k$, we observe $G_k(W_{g_\eta})$, a many-to-one function of W_{g_η}; when $\mathscr{C}_\eta = \infty$, we observe $G_\infty(W_{g_\eta}) = W_{g_\eta}$, the full data. Here, under the consistency assumption, this is as follows. If $A_1 \neq g_{\eta 1}(X_1)$, then $\mathscr{C}_\eta = 1$; that is, $I(\mathscr{C}_\eta = 1) = I\{A_1 \neq g_{\eta 1}(X_1)\}$, and we observe $G_{\mathscr{C}_\eta}(W_{g_\eta}) = G_1(W_{g_\eta}) = X_1$. None of the observed treatments are consistent with following g_η, so $X_2, \ldots, X_K, Y$ are not consistent with g_η. If $A_1 = g_{\eta 1}(X_1)$ and $A_2 \neq g_{\eta 2}\{\bar{X}_2, g_{\eta 1}(X_1)\}$, then $\mathscr{C}_\eta = 2$, $I(\mathscr{C}_\eta = 2) = I\{A_1 = g_{\eta 1}(X_1)\} I[A_2 \neq g_{\eta 2}\{\bar{X}_2, g_{\eta 1}(X_1)\}]$, and $G_{\mathscr{C}_\eta}(W_{g_\eta})=G_2(W_{g_\eta})=\bar{X}_2^*(g_{\eta_1})=\bar{X}_2$. Only the treatment at decision 1 and the ensuing X_2 are consistent with g_η. Likewise, $I(\mathscr{C}_\eta = 3) = I\{A_1 = g_{\eta 1}(X_1)\} I[A_2 = g_{\eta 2}\{\bar{X}_2, g_{\eta 1}(X_1)\}] I\{A_3 \neq g_{\eta 3}(\bar{X}_3)\}$, where $g_{\eta 3}(\bar{X}_3)$ is shorthand for $g_{\eta 3}[\bar{X}_3, g_{\eta 1}(X_1), g_{\eta 2}\{\bar{X}_2, g_{\eta 1}(X_1)\}] = g_{\eta 3}\{\bar{X}_3, \bar{g}_{\eta 2}(\bar{X}_2)\}$ and $\bar{g}_{\eta 2}(\bar{X}_2) = [g_{\eta 1}(X_1), g_{\eta 2}\{\bar{X}_2, g_{\eta 1}(X_1)\}]$ and similarly for general k; and $G_{\mathscr{C}_\eta}(W_{g_\eta})=G_3(W_{g_\eta})=\bar{X}_3^*(\bar{g}_{\eta_2})=\bar{X}_3$. Continuing in this fashion, $I(\mathscr{C}_\eta = K) = I[\bar{A}_{K-1} = \bar{g}_{\eta K-1}\{\bar{X}_{K-1}, \bar{g}_{\eta K-2}(\bar{X}_{K-2})\}] I[A_K \neq g_{\eta K}\{\bar{X}_K, \bar{g}_{\eta K-1}(\bar{X}_{K-1})\}]$, and $G_{\mathscr{C}_\eta}(W_{g_\eta})=G_K(W_{g_\eta})=\bar{X}_K^*(\bar{g}_{\eta_{K-1}})=\bar{X}_K$. Finally, if $\bar{A}_K = \bar{g}_{\eta K}\{\bar{X}_K, \bar{g}_{\eta K-1}(\bar{x}_{K-1})\}$, $G_{\mathscr{C}_\eta}(W_{g_\eta}) = G_\infty(W_{g_\eta}) = W_{g_\eta} = (X_1, \ldots, X_K, Y)$. Here, the observed data are consistent with having followed all K rules in g_η. The coarsening is monotone in that $G_k(W_{g_\eta})$ is a coarsened version of $G_{k'}(W_{g_\eta})$, $k' > k$, and $G_k(W_{g_\eta})$ is a many-to-one function of $G_{k+1}(W_{g_\eta})$.

Coarsened data are said to be coarsened at random if, for each k, the probability that the data are coarsened at level k, given the full data, depends only on the coarsened data, so only on data that are observed at level k (Tsiatis, 2006, §7.1). Under the consistency and sequential randomization assumptions, it may be shown using results in §3 of the Supplementary

Material that the coarsening here is at random. Define the coarsening discrete hazard pr($\mathscr{C}_\eta = k \mid \mathscr{C}_\eta \geq k, W_{g_\eta}$) to be the probability that the observed treatments cease to be consistent with g_η at decision k, given they are consistent prior to k and all potential outcomes. Under coarsening at random, this hazard is a function only of the coarsened data, that is, the data observed through decision k, which we write as pr($\mathscr{C}_\eta = k \mid \mathscr{C}_\eta \geq k, W_{g_\eta}$) = $\lambda_{\eta,k}\{G_k(W_{g_\eta})\}$. Then, from above, for $k = 1$, $\lambda_{\eta,1}\{G_1(W_{g_\eta})\} = \lambda_{\eta,1}(X_1) = \text{pr}\{A_1 \neq g_{\eta 1}(X_1) \mid X_1\}$, which can be expressed in terms of the propensity for treatment at decision 1 as $\pi_1(X_1)^{1-g_{\eta 1}(X_1)}\{1 - \pi_1(X_1)\}^{g_{\eta 1}(X_1)}$. Similarly, for $k = 2, \ldots, K$,

$$
\begin{aligned}
\lambda_{\eta,k}\{G_k(W_{g_\eta})\} = \lambda_{\eta,k}(\bar{X}_k) = \text{pr}\{A_k \neq g_{\eta_k}(\bar{X}_k, \bar{A}_{k-1}) | \bar{X}_k, \bar{A}_{k-1} \\
= \bar{g}_{\eta_{k-1}}(\bar{X}_{k-1})\} \\
= \pi_k\{\bar{X}_k, \bar{g}_{\eta_{k-1}}(\bar{X}_{k-1})\}^{1-g_{\eta_k}\{\bar{X}_k, \bar{g}_{\eta_{k-1}}(\bar{X}_{k-1})\}} \\
\times [1 - \pi_k\{\bar{X}_k, \bar{g}_{\eta_{k-1}}(\bar{X}_{k-1})\}]^{g_{\eta_k}\{\bar{X}_k, \bar{g}_{\eta_{k-1}}(\bar{X}_{k-1})\}}.
\end{aligned}
$$

We may then express the probabilities of being consistent with g_η through at least the kth decision, so having $\mathscr{C}_\eta > k$, given all potential outcomes, in terms of the discrete hazards. Under coarsening at random, these probabilities depend only on the observed data through decision k. That is, pr($\mathscr{C}_\eta > k \mid W_{g_\eta}$) = $K_{\eta,k}\{G_k(W_{g_\eta})\} = K_{\eta,k}(\bar{X}_k)$, where $K_{\eta,k}(\bar{X}_k) = \prod_{k'=1}^{k}\{1 - \lambda_{\eta,k'}(\bar{X}_{k'})\}$ (Tsiatis, 2006, §8.1).

We now use these developments to deduce the form of estimators for $E\{Y^*(g_\eta)\}$. From the theory of Robins et al. (1994) for general monotonely coarsened data, under coarsening at random, if the coarsening mechanism is correctly specified, which corresponds here to correct specification of the $\lambda_{\eta,k}(\bar{X}_k)$, and hence of the propensity models, all regular, asymptotically linear, consistent estimators (Tsiatis, 2006, Chapter 3) for $E\{Y^*(g_\eta)\}$ for fixed η have the form

$$
\sum_{i=1}^{n} \frac{I(\mathscr{C}_{\eta,i} = \infty)}{K_{\eta,K}(\bar{X}_{Ki})} Y_i + \sum_{k=1}^{K} \frac{I(\mathscr{C}_{\eta,i} = k) - \lambda_{\eta,k}(\bar{X}_{ki}) I(\mathscr{C}_{\eta,i} \geq k)}{K_{\eta,k}(\bar{X}_{ki})} L_k(\bar{X}_{ki}), \quad (4)
$$

where $L_k(\bar{X}_k)$ are arbitrary functions of $\bar{X}_k$. The optimal choice leading to (4) with smallest asymptotic variance is $L_{\eta,k}^{\text{opt}}(\bar{x}_k) = E\{Y^*(g_\eta) | \bar{X}_k^*(\bar{g}_{\eta_{k-1}}) = \bar{x}_k\}$. The right hand term in (4) augments the first, itself a consistent estimator for $E\{Y^*(g_\eta)\}$ when the $\lambda_{\eta,k}(\bar{X}_k)$ are correctly specified, to gain efficiency. As in Tsiatis (2006, §10.3), (4) is doubly robust in that it is a consistent estimator for $E\{Y^*(g_\eta)\}$ if either the $\lambda_{\eta,k}(\bar{X}_{ki})$ are correctly specified or if the $L_k(\bar{X}_{ki})$ are equal to $L_{\eta,k}^{\text{opt}}(\bar{X}_{ki}) (k = 1, \ldots, K)$; see §4 of the Supplementary Material.

To implement (4), one must specify $\lambda_{\eta,k}(\bar{X}_{ki})$ and $L_k(\bar{X}_{ki})$. The first follow from specifying $\pi_1(x_1) = \text{pr}(A_1 = 1 \mid X_1 = x_1)$, $\pi_k(\bar{x}_k, \bar{a}_{k-1}) = \text{pr}(A_k = 1 \mid \bar{X}_k = \bar{x}_k, \bar{A}_{k-1} = \bar{a}_{k-1})$ for $k = K, \ldots, 2$. If these are unknown, as in A-learning, posit models $\pi_1(x_1; \gamma_1)$, $\pi_k(\bar{x}_k, \bar{a}_{k-1}; \gamma_k)$ for $k = 2, \ldots, K$, and estimate γ_k by $\hat{\gamma}_k$ ($k = 1, \ldots, K$). With $\gamma = (\gamma_1^{\mathrm{T}}, \ldots, \gamma_K^{\mathrm{T}})^{\mathrm{T}}$ and $\hat{\gamma}^{\mathrm{T}} = (\hat{\gamma}_1^{\mathrm{T}}, \ldots, \hat{\gamma}_K^{\mathrm{T}})^{\mathrm{T}}$, this implies that $\lambda_{\eta,1}(X_1; \gamma_1) = \pi_1(X_1; \gamma_1)^{1-g_{\eta 1}(X_1)}\{1 - \pi_1(X_1; \gamma_1)\}^{g_{\eta 1}(X_1)}$,

$$
\lambda_{\eta,k}(\bar{X}_k; \gamma_k) = \pi_k\{\bar{X}_k, \bar{g}_{\eta_{k-1}}(\bar{X}_{k-1}); \gamma_k\}^{1-g_{\eta_k}\{\bar{X}_k, \bar{g}_{\eta_{k-1}}(\bar{X}_{k-1})\}} \times [1 - \pi_k\{\bar{X}_k, \bar{g}_{\eta_{k-1}}(\bar{X}_{k-1}); \gamma_k\}]^{g_{\eta_k}\{\bar{X}_k, \bar{g}_{\eta_{k-1}}(\bar{X}_{k-1})\}}
$$

and $K_{\eta,k}(\bar{\mathrm{X}}_k;\gamma)=\prod_{k'=1}^{k}\{1-\lambda_{\eta,k'}(\bar{\mathrm{X}}_{k'};\gamma_{k'})\}$, and suggests substituting $\lambda_{\eta,k}(\bar{X}_k;\hat{\gamma}_k)$ and $K_{\eta,k}(\bar{X}_k;\hat{\gamma})$ in (4).

Several options exist for specification of the $L_k(\bar{X}_k)$. The simplest is to take $L_k(\bar{X}_k)\equiv 0$, yielding the inverse probability weighted estimator

$$\mathrm{IPWE}(\eta)=\sum_{i=1}^{n}\frac{I(\mathscr{C}_{\eta,i}=\infty)}{K_{\eta,K}(\bar{\mathrm{X}}_{Ki};\hat{\gamma})}Y_i, \qquad (5)$$

which is consistent for $E\{Y^*(g_\eta)\}$ if $\eta_1(X_1;\gamma_k)$ and $\pi_k(\bar{X}_k,\bar{\mathrm{A}}_{k-1};\gamma_k)$ ($k = 2, \ldots, K$), and hence $K_{\eta,K}(\bar{X}_K;\gamma)$, are correctly specified, but otherwise may be inconsistent. The corresponding estimator for g_η^{opt} is found by estimating η^{opt} by $\hat{\eta}_{\mathrm{IPWE}}^{\mathrm{opt}}$, say, maximizing (5) in η. As (5) is based on data only from subjects whose entire treatment history is consistent with g_η, it is relatively less efficient than estimators that use all the data, discussed next.

To take greatest advantage of the potential for improved efficiency through the augmentation term in (4), an obvious approach is to posit and fit parametric models approximating the conditional expectations $L_{\eta,k}^{\mathrm{opt}}(\bar{\mathrm{x}}_k)=E\{Y^*(g_\eta)|\bar{\mathrm{X}}_k^*(\bar{g}_{\eta_{k-1}})=\bar{\mathrm{x}}_k\}$, and substitute these into (4) along with $\lambda_{\eta,k}(\bar{X}_k;\hat{\gamma}_k)$ and $K_{\eta,k}(\bar{X}_k;\hat{\gamma})$. To this end, let $\mu_{\eta_K}(\bar{x}_K,\bar{\mathrm{a}}_K) = E(Y\,|\,\bar{X}_K=\bar{x}_K, \bar{\mathrm{A}}_K=\bar{\mathrm{a}}_K)$ and $f_{\eta_K}(\bar{x}_K,\bar{\mathrm{a}}_{K-1})=\mu_{\eta_K}\{\bar{x}_K,\bar{\mathrm{a}}_{K-1}, g_{\eta_K}(\bar{x}_K,\bar{\mathrm{a}}_{K-1})\}$. Then define iteratively, for $k = K-1, \ldots, 2$, the quantities $\mu_{\eta_k}(\bar{x}_k,\bar{\mathrm{a}}_k)=E\{f_{\eta_{k+1}}(\bar{x}_k,X_{k+1},\bar{\mathrm{a}}_k)\,|\,\bar{X}_k=\bar{x}_k, \bar{\mathrm{A}}_k=\bar{\mathrm{a}}_k\}$ and $f_{\eta_k}(\bar{x}_k,\bar{\mathrm{a}}_{k-1})=\mu_{\eta_k}\{\bar{x}_k,\bar{\mathrm{a}}_{k-1},g_{\eta_k}(\bar{x}_k,\bar{\mathrm{a}}_{k-1})\}$; for $k = 1$, $\mu_{\eta_1}(x_1,a_1)=E\{f_{\eta_2}(x_1,X_2,a_1)\,|\,X_1=x_1, A_1=a_1\}$, $f_{\eta_1}(x_1)=\mu_{\eta_1}\{x_1,g_{\eta_1}(x_1)\}$. In §5 of the Supplementary Material, we demonstrate that $L_{\eta,k}^{\mathrm{opt}}(\bar{\mathrm{X}}_k)=\mu_{\eta_k}\{\bar{\mathrm{X}}_k,\bar{g}_{\eta_k}(\bar{\mathrm{X}}_k)\}$.

This suggests specifying η-dependent models $\mu_{\eta_k}(\bar{x}_k,\bar{\mathrm{a}}_k;\xi_k)$ depending on parameters ξ_k, $k = 1, \ldots, K$. For fixed η, estimators $\hat{\xi}_k$ for ξ_k may be found iteratively by solving in ξ_k

$$\sum_{i=1}^{n}\frac{\partial\mu_{\eta_k}(\bar{\mathrm{X}}_{ki},\bar{A}_{ki};\xi_k)}{\partial\xi_k}\{\tilde{\mathrm{f}}_{(k+1)i}-\mu_{\eta_k}(\bar{\mathrm{X}}_{ki},\bar{A}_{ki};\xi_k)\}=0(k=1,\ldots K),$$

where $\partial/\partial\xi_k\{\mu_{\eta_k}(\bar{X}_{ki},\bar{\mathrm{A}}_{ki};\xi_k)\}$ is the vector of partial derivatives of $\mu_{\eta_k}(\bar{X}_{ki},\bar{\mathrm{A}}_{ki};\xi_k)$ with respect to elements of ξ_k, $\tilde{f}_{(K+1)i}=Y_i$ and $\tilde{f}_{ki}=\mu_{\eta_k}[\bar{X}_{ki},\bar{\mathrm{A}}_{(k-1)i},g_{\eta_k}\{\bar{X}_{ki},\bar{\mathrm{A}}_{(k-1)i}\};\hat{\xi}_k]$ ($k = K, \ldots, 2$). The fitted $\mu_{\eta_k}\{\bar{X}_k,\bar{g}_{\eta_k}(\bar{X}_k);\hat{\xi}_k\}$ may then be used to approximate $L_{\eta,k}^{\mathrm{opt}}(\bar{\mathrm{x}}_k)$ in (4). While these models almost certainly are not correct, as specification of a compatible sequence of models for $k = 1, \ldots, K$ is a significant challenge, they may be reasonable approximations to the true conditional expectations. Thus, define

$$\mathrm{DR}(\eta)=\sum_{i=1}^{n}\frac{I(\mathscr{C}_{\eta,i}=\infty)}{K_{\eta,K}(\bar{\mathrm{X}}_{Ki};\hat{\gamma})}Y_i+\sum_{k=1}^{K}\frac{I(\mathscr{C}_{\eta,i}=k)-\lambda_{\eta,k}(\bar{\mathrm{X}}_{ki};\hat{\gamma}_k)I(\mathscr{C}_{\eta,i}\geq k)}{K_{\eta,k}(\bar{\mathrm{X}}_{ki};\hat{\gamma})}\mu_{\eta_k}\{\bar{\mathrm{X}}_{ki},\bar{g}_{\eta_k}(\bar{\mathrm{X}}_{ki});\hat{\xi}_k\}, \qquad (6)$$

which, by virtue of the double robustness property, will be consistent for $E\{Y^*(g_\eta)\}$ if either $\pi_1(x_1;\gamma_k)$ and $\pi_k(\bar{x}_k,\bar{\mathrm{a}}_{k-1};\gamma_k)$ ($k = K, \ldots, 2$), are correctly specified, or the the $\mu_{\eta_k}(\bar{x}_k,\bar{\mathrm{a}}_k;\xi_k)$ are. If all of these models were correct, then (6) would achieve optimal efficiency. As for (5), estimation of g_η^{opt} follows by maximizing (6) in η to obtain $\hat{\eta}_{\mathrm{DR}}^{\mathrm{opt}}$.

A computational challenge is that the models $\mu_{\eta_k}(\bar{x}_k,\bar{\mathrm{a}}_{k-1};\xi_k)$ must be refitted for each value of η encountered in the optimization algorithm used to carry out the maximization. A practical alternative when regimes in $\mathscr{G}_\eta$ are derived from models is to substitute for $L_k(\bar{X}_{k,i})$

in (4) fitted Q-functions $Q_k\{\bar{X}_k, \bar{g}_{\eta k}(\bar{X}_k); \hat{\beta}_k\}$ for $k = K, \ldots, 1$ obtained from Q-learning; holding $\hat{\beta}_k$ fixed, these depend on η only through $\bar{g}_{\eta k}(\bar{X}_k)$. While these are not strictly models for $E\{Y^*(g_\eta)|\bar{\mathrm{X}}_k^*(\bar{g}_{\eta_k})\}$, the hope is that they will be close enough to achieve near optimal efficiency gains over (5). Thus, estimate g_η^{opt} by maximizing in η to obtain $\hat{\eta}_{\mathrm{AIPWE}}^{\mathrm{opt}}$

$$\mathrm{AIPWE}(\eta)=\sum_{i=1}^{n}\frac{I(\mathscr{C}_{\eta,i}=\infty)}{K_{\eta,K}(\bar{\mathrm{X}}_{Ki};\hat{\gamma})}Y_i+\sum_{k=1}^{K}\frac{I(\mathscr{C}_{\eta,i}=k)-\lambda_{\eta,k}(\bar{\mathrm{X}}_{ki};\hat{\gamma}_k)I(\mathscr{C}_{\eta,i}\geq k)}{K_{\eta,k}(\bar{\mathrm{X}}_{ki};\hat{\gamma})}Q_k\{\bar{\mathrm{X}}_{ki},\bar{g}_{\eta_k}(\bar{\mathrm{X}}_{ki});\hat{\beta}_k\}. \quad (7)$$

See §6 of the Supplementary Material for a similar proposal when $\mathscr{G}_\eta$ is determined directly.

Standard errors for these estimators for $E\{Y^*(g_\eta^{\mathrm{opt}})\}$ may be obtained via the sandwich technique (Stefanski & Boos, 2002) based on the argument in Zhang et al. (2012, Equation (4)).

5. Simulation studies

We have carried out several simulation studies to evaluate the performance of the proposed methods, each involving 1000 Monte Carlo data sets.

The first simulation adopts the scenario in Moodie et al. (2007) of a study in which HIV-infected patients are randomized to initiate antiretroviral therapy or not, coded as 1 or 0, at baseline and again at six months to determine the optimal regime for therapy initiation. We generated baseline CD4 count $X_1 \sim N(450, 150)$, where $N(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2; baseline treatment A_1 as Bernoulli with success probability $\mathrm{pr}(A_1 = 1 \mid X_1) = \mathrm{expit}(2 - 0.006X_1)$, where $\mathrm{expit}(u) = e^u/(1 + e^u)$; six-month CD4 count X_2, conditional on (X_1,A_1), as $N(1.25X_1, 60)$; and treatment at six months A_2 as Bernoulli with $\mathrm{pr}(A_2 = 1 \mid \hat{X}_2, A_1) = A_1 + (1 - A_1)\mathrm{expit}(0.8 - 0.004X_2)$. Here, patients with $A_1 = 1$ continue on therapy with certainty. The outcome Y, one-year CD4 count, conditional on $(\bar{X}_2, \bar{A}_2)$, was normal with mean $400 + 1.6X_1 - |250 - X_1|\{A_1 - I(250 - X_1 > 0)\}^2 - (1 - A_1)|720 - 2X_2|\{A_2 - I(720 - 2X_2 > 0)\}^2$ and variance 60^2. The true Q-contrast functions are thus $C_2(x_1, x_2, a_1) = (1 - a_1)(720 - 2x_2)$, $C_1(x_1) = 250 - x_1$, the optimal treatment regime $g^{\mathrm{opt}}=(g_1^{\mathrm{opt}}, g_2^{\mathrm{opt}})$ has $g_1^{\mathrm{opt}}(x_1)=I(250-x_1>0), g_2^{\mathrm{opt}}(\bar{\mathrm{x}}_2, a_1)=I\{a_1+(1-a_1)(720-2x_2)>0\}=I\{a_1+(1-a_1)(360-x_2)>0\}$ and $E\{Y^*(g^{\mathrm{opt}})\} = 1120$.

For A-learning, we took

$$h_2(\bar{\mathrm{x}}_2, a_1;\alpha_2)=\alpha_{20}+\alpha_{21}x_1+\alpha_{22}a_1+\alpha_{23}a_1x_1+\alpha_{24}(1-a_1)x_2,\ C_2(\bar{\mathrm{x}}_2, a_1;\psi_2)=(1-a_1)(\psi_{20}+\psi_{21}x_2),$$

$h_1(x_1, \alpha_1) = \alpha_{10} + \alpha_{11}x_1$, and $C_1(x_1; \psi_1) = \psi_{10} + \psi_{11}x_1$; and, analogously, for Q-learning,

$$Q_2(\bar{\mathrm{x}}_2, \bar{a}_2;\beta_2)=\beta_{20}+\beta_{21}x_1+a_1(\beta_{22}+\beta_{23}x_1)+\beta_{24}(1-a_1)x_2+a_2(1-a_1)(\beta_{25}+\beta_{26}x_2),$$
$$Q_1(x_1, a_1;\beta_1)=\beta_{10}+\beta_{11}x_1+a_1(\beta_{12}+\beta_{13}x_1),$$

so the Q-contrast functions are correct, but the Q-functions are misspecified. Here, $C_2(\bar{x}_2, 1; \psi_2) = 0$, respecting that $\Phi_2(\bar{x}_2, 1) = \{1\}$. We used correct propensity models $\pi_2(\bar{x}_2, a_1 = 0; \gamma_2) = \mathrm{expit}(\gamma_{20} + \gamma_{21}x_2)$, $\pi_1(x_1; \gamma_1) = \mathrm{expit}(\gamma_{10} + \gamma_{11}x_1)$ and incorrect models $\pi_2(\bar{x}_2, a_1 = 0; \gamma_2) = \gamma_2$, $\pi_1(x_1; \gamma_1) = \gamma_1$.

For maximizing ipwe(η) in (5), dr(η) in (6), and aipwe(η) in (7) to obtain $\hat{\eta}^{\text{opt}}_{\text{IPWE}}$, $\hat{\eta}^{\text{opt}}_{\text{DR}}$, and $\hat{\eta}^{\text{opt}}_{\text{AIPWE}}$, we considered the class of regimes $\mathscr{G}_\eta$ with elements $g_\eta = (g_{\eta 1}, g_{\eta 2})$,

$$g_{\eta_2}(\bar{\text{x}}_2, a_1){=}I\{a_1{+}(1-a_1)(\eta_{20}{+}\eta_{21}x_2){>}0\}, g_{\eta_1}(x_1){=}I(\eta_{10}{+}\eta_{11}x_1{>}0),$$

so that $\eta_2 = (\eta_{20}, \eta_{21})^T$, $\eta_1 = (\eta_{10}, \eta_{11})^T$, $\eta{=}(\eta_1^T, \eta_2^T)^T$ and $\eta^{\text{opt}} = (250, -1, 360, -1)^T$. Clearly, $g^{\text{opt}} \in \mathscr{G}_\eta$. We used the same propensity models, and, for (7), Q-function models as above; for (6), we posited $\mu_{\eta 2}(\bar{x}_2, \bar{a}_2; \xi_2) = \xi_{20} + \xi_{21}x_1 + a_1(\xi_{22} + \xi_{23}x_1) + \xi_{24}(1 - a_1)x_2 + a_2(1 - a_1)(\xi_{25} + \xi_{26}x_2)$ and $\mu_{\eta 1}(x_1, a_1; \xi_1) = \xi_{10} + \xi_{11}x_1 + a_1(\xi_{12} + \xi_{13}x_1)$ for each η. To achieve a unique representation, we fixed $(\eta_{21}, \eta_{11}) = (-1, -1)$ and determined η_{20}, η_{10} via a grid search; because ipwe(η), dr(η) and aipwe(η) are step functions of η with jumps at (x_{1i}, x_{2j}) $(i, j = 1, \ldots, n)$, we maximized in η over all (x_{1i}, x_{2j}).

The second scenario is the same as the first except that the models for the Q-contrast functions are misspecified. Specifically, the generative distribution of Y given $(\bar{X}_2, \bar{A}_2)$ is now normal with mean $400 + 1.6X_1 - |250 - 0.6X_1|\{A_1 - I(250 - X_1 > 0)\}^2 - (1 - A_1)|720 - 1.4X_2|\{A_2 - I(720 - 2X_2 > 0)\}^2$ and variance 60^2, so that, from the discussion below (2) of Moodie et al. (2007), the implied true contrast functions are no longer of the form above, but all posited models were taken to be the same as in the first simulation.

Tables 1 and 2 show the results. For Q- and A-learning, we report $\eta(\hat{\beta})$ and $\eta(\hat{\psi})$. The column $\hat{E}$ $(\hat{\eta}^{\text{opt}})$ shows for each estimator the Monte Carlo average and standard deviation of the estimated values of $E\{Y^*(g^{\text{opt}}_\eta)\}$ reflecting performance for estimating the true achievable mean outcome under the true optimal regime, while $E(\hat{\eta}^{\text{opt}})$ reflects performance of the estimated optimal regime itself. For each Monte Carlo data set, this is the true mean outcome that would be achieved if the estimated optimal regime were followed by the population was determined by simulation, and the values reported are the Monte Carlo average and standard deviation of these simulated quantities. When compared to the true $E\{Y^*(g^{\text{opt}}_\eta)\}{=}1120$, these measure the extent to which the estimated optimal regimes approach the performance of the true optimal regime.

For the first simulation, from Table 1, because the Q-functions are misspecified, the Q-learning estimators for η_{10} and η_{20} are biased, while those from A-learning based on postulated Q-contrast functions that include the truth are consistent when the propensity model is correct. When the propensity model is incorrect, Q-learning is unaffected; however, A-learning yields biased estimators for η_{10} and η_{20} identical to those from Q-learning, as linear models are used for $C_2(\bar{x}_2, a_1; \psi_2)$, $C_1(x_1; \psi_1)$, $h_2(\bar{x}_2, a_1; \alpha_2)$ and $h_1(x_1; \alpha_1)$ (Chakraborty et al., 2010). Although Q-learning results in poor estimation of η_{10} and η_{20}, efficiency loss for estimating the optimal regime is negligible, as the proportion of benefit the estimated regime achieves if used in the entire population relative to the true optimal regime is virtually one. A possible explanation is that patients near the true decision boundary have $C_2(\bar{X}_2, a_1)$, $C_1(X_1)$ close to zero, and few patients would receive treatment 1 according to the true decision rule for the first time point. This also follows from the fact that for regime $g{=}(0, g_2^{\text{opt}})$, the corresponding expectation is 1114. When the propensity model is correct, the estimators based on dr(η) and aipwe(η) yield estimated regimes comparable to those found by A-learning in terms of true mean outcome achieved, despite yielding relatively inefficient estimators for η_{10} and η_{20} A-learning, perhaps for the same reason as above. When the propensity model is incorrect, the dr(η) and aipwe(η) estimators yield estimated regimes that are still close to the optimal. The ipwe(η) estimator show relatively poorer performance, especially when the propensity score model is incorrect, which is not unexpected; this estimator only uses information from patients whose treatment histories are consistent with following g_η and hence is inefficient.

In the second simulation, the values of $|C_2(\bar{X}_2, A_1)|$ and $|C_1(X_1)|$ for patients near the true decision boundary are larger than in the first simulation, and the posited Q-contrast functions are no longer correct. From Table 2, the A- and Q- learning estimators perform similarly, both yielding estimated regimes far from optimal. Those based on dr(η) and aipwe(η) are almost identical to g^{opt} on average and perform almost identically to the true optimal regime, regardless of whether or not the propensity model is correct. Again, the estimator based on ipwe(η) in (5) performs poorly. Evidently, augmentation even using incorrect models leads to considerable gains over ipwe(η) regardless of whether or not the propensity model is correct.

The third scenario involved $K = 3$ decision points. To achieve average numbers of patients consistent with the regime comparable to those in the $K = 2$ cases, we took $n = 1000$. We generated X_1, A_1, X_2 as in the previous two scenarios; A_2 as Bernoulli with $\text{pr}(A_2 = 1 \mid \bar{X}_2, A_1) = \text{expit}(0.8 - 0.004X_2)$; twelve-month CD4 count X_3, conditional on $(\bar{X}_2, \bar{A}_2)$, as $N(0.8X_2, 60)$; treatment at twelve months A_3 as Bernoulli with $\text{pr}(A_3 = 1 \mid \bar{X}_3, \bar{A}_2) = \text{expit}(1 - 0.004X_3)$; and the outcome Y, 18-month CD4 count, conditional on $(\bar{X}_3, \bar{A}_3)$, as normal with mean $400 + 1.6X_1 - |500 - 1.4X_1|\{A_1 - I(500 - 2X_1 > 0)\}^2 - |720 - 1.4X_2|\{A_2 - I(720 - 2X_2 > 0)\}^2 - |600 - 1.4X_3|\{A_3 - I(600 - 2X_3 > 0)\}^2$ and variance 60^2. The optimal treatment regime $g^{\text{opt}}=(g_1^{\text{opt}}, g_2^{\text{opt}}, g_3^{\text{opt}})$ has $g_1^{\text{opt}}(x_1)=I(250-x_1>0), g_2^{\text{opt}}(\bar{x}_2, a_1)=I(360-x_2>0), g_3^{\text{opt}}(\bar{x}_3, \bar{a}_2)=I(300-x_3>0)$ and $E\{Y^*(g^{\text{opt}})\} = 1120$.

For A-learning, we took

$$\begin{aligned}
h_3(\bar{x}_3, \bar{a}_2;\alpha_3)&=\alpha_{30}+\alpha_{31}x_1+a_1(\alpha_{32}+\alpha_{33}x_1)+\alpha_{34}x_2+a_2(\alpha_{35}+\alpha_{36}x_2)+\alpha_{37}x_3,\\
C_3(\bar{x}_3, \bar{a}_2;\psi_2)&=\psi_{30}+\psi_{31}x_3, h_2(\bar{x}_2, a_1;\alpha_2)=\alpha_{20}+\alpha_{21}x_1+a_1(\alpha_{22}+\alpha_{23}x_1)+\alpha_{24}x_2,\\
C_2(\bar{x}_2, a_1;\psi_2)&=\psi_{20}+\psi_{21}x_2, h_1(x_1;\alpha_1)=\alpha_{10}+\alpha_{11}x_1, C_1(x_1;\psi_1)=\psi_{10}+\psi_{11}x_1,
\end{aligned}$$

and for Q-learning

$$\begin{aligned}
Q_3(\bar{x}_3, \bar{a}_3;\beta_3)&=\beta_{30}+\beta_{31}x_1+a_1(\beta_{32}+\beta_{33}x_1)+\beta_{34}x_2+a_2(\beta_{35}+\beta_{36}x_2)+\alpha_{37}x_3+a_3(\beta_{38}+\beta_{39}x_3),\\
Q_2(\bar{x}_2, \bar{a}_2;\beta_2)&=\beta_{20}+\beta_{21}x_1+a_1(\beta_{22}+\beta_{23}x_1)+\beta_{24}x_2+a_2(\beta_{25}+\beta_{26}x_2),\\
Q_1(x_1, a_1;\beta_1)&=\beta_{10}+\beta_{11}x_1+a_1(\beta_{12}+\beta_{13}x_1);
\end{aligned}$$

thus, both Q- and Q-contrast functions are misspecified. We used correct propensity models $\pi_3(\bar{x}_3, \bar{a}_2; \gamma_3) = \text{expit}(\gamma_{30} + \gamma_{31}x_3)$, $\pi_2(\bar{x}_2, a_1; \gamma_2) = \text{expit}(\gamma_{20} + \gamma_{21}x_2)$, $\pi_1(x_1; \gamma_1) = \text{expit}(\gamma_{10} + \gamma_{11}x_1)$ and incorrect models $\pi_3(\bar{x}_3, \bar{a}_2; \gamma_3) = \gamma_3$, $\pi_2(\bar{x}_2, a_1; \gamma_2) = \gamma_2$, $\pi_1(x_1; \gamma_1) = \gamma_1$.

For the three proposed estimators, we took the class of regimes $\mathscr{G}_\eta$ to have elements $g_\eta = (g_{\eta 1}, g_{\eta 2}, g_{\eta 3})$, $g_{\eta 3}(\bar{x}_3, \bar{a}_2) = I(\eta_{30} + \eta_{31}x_3 > 0)$, $g_{\eta 2}(\bar{x}_2, a_1) = I(\eta_{20} + \eta_{21}x_2 > 0)$, $g_{\eta 1}(x_1) = (\eta_{10} + \eta_{11}x_1 > 0)$, so $\eta_3 = (\eta_{30}, \eta_{31})^{\mathrm{T}}$ $\eta_2 = (\eta_{20}, \eta_{21})^{\mathrm{T}}$, $\eta_1 = (\eta_{10}, \eta_{11})^{\mathrm{T}}$, $\eta=(\eta_1^{\mathrm{T}}, \eta_2^{\mathrm{T}}, \eta_3^{\mathrm{T}})^{\mathrm{T}}$ and $\eta^{\text{opt}} = (250, -1, 360, -1, 300, -1)^{\mathrm{T}}$, so $g^{\text{opt}} \in \mathscr{G}_\eta$. We used the same propensity models, and, for (7), Q-function models as above, and fixed $(\eta_{31}, \eta_{21}, \eta_{11}) = (-1, -1, -1)$. To carry out the maximizations, we used a genetic algorithm discussed by Goldberg (1989), implemented in the `rgenoud` package in R (Mebane & Sekhon, 2011); see §7 of the Supplementary Material for details.

Table 3 show the results. Q-learning performs poorly, as expected. When the propensity model is correctly specified, results for A-learning and the proposed methods are similar to those in the second scenario, with the estimated regimes based on dr(η) and aipwe(η) achieving near-optimal performance and associated reliable inference on the true achievable mean outcome $E\{Y^*(g_\eta)\}$. When the propensity models are misspecified, the situation is similar for these estimators in terms of performance; however, inference on $E\{Y^*(g_\eta)\}$ is markedly degraded. In both cases, performance of the estimator based on ipwe(η) is quite poor. Intuitively, as the number of decisions K increases, it is not unexpected that all methods can suffer from diminished performance. Research is needed on the design of sequentially randomized trials to ensure adequate sample size for reliable inference on multi-decision regimes.

In §8 of the Supplementary Material, we present results of a more complex scenario; the qualitative conclusions are similar.

All simulations here, and others we have conducted, suggest that Q- and A-learning can yield biased estimators for parameters defining the optimal regime if the Q-functions or Q-contrast functions are misspecified. Under these conditions, the resulting estimated optimal regimes can perform poorly in terms of achieving the expected potential outcome of the true optimal regime. In contrast, the proposed approach using (6) or (7) exhibits robustness to misspecification of either one of the outcome regression or propensity score models. Under these circumstances, the estimators of regime parameters are relatively unbiased, and the expected potential outcome under the estimated optimal regime approaches that of the true optimal regime. Moreover, the proposed methods lead to reliable estimation of the expected potential outcome under the true regime, with coverage probabilities close to the nominal level. Even when both outcome regression and propensity models are misspecified, the proposed methods can yield estimated optimal regimes that do not show substantial degradation of performance in terms of achieved expected potential outcome relative to the true optimal regime. In this case, inference on the expected outcome under the true optimal regime can be compromised, although, interestingly, the methods performed well in this regard under these conditions in the second simulation scenario. Collectively, our results suggest that the proposed methods are attractive alternatives to Q- and A-learning owing to their robustness to such model misspecification. As the estimator based on aipwe(η) is much less computationally intensive than dr(η) and performs similarly, we recommend it for practical use.

In §9 of the Supplementary Material, we report on application of the methods to a study to compare treatment options in patients with nonpsychotic major depressive disorder.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported by grants from the US National Institutes of Health.

References

Almirall D, Ten Have T, Murphy SA. Structural nested mean models for assessing time-varying effect moderation. Biometrics. 2010; 66:131–139. [PubMed: 19397586]

Bather, J. Decision Theory: an Introduction to Dynamic Programming and Sequential Decisions. Chichester: Wiley; 2000.

Chakraborty B, Murphy SA, Strecher V. Inference for non-regular parameters in optimal dynamic treatment regimes. Statist. Meth. Med. Res. 2010; 19:317–343.

Goldberg, DE. Genetic Algorithms in Search, Optimization, and Machine Learning. Reading, MA: Addison-Wesley; 1989.

Henderson R, Ansell P, Alshibani D. Regret-regression for optimal dynamic treatment regimes. Biometrics. 2010; 66:1192–1201. [PubMed: 20002404]

Mebane WR, Sekhon JS. Genetic optimization using derivatives: the rgenoud package for R. J. Statist. Soft. 2011; 42:1–26.

Moodie EEM, Richardson TS, Stephens DA. Demystifying optimal dynamic treatment regimes. Biometrics. 2007; 63:447–455. [PubMed: 17688497]

Murphy SA. Optimal dynamic treatment regimes (with discussion). J. Royal Statist. Soc., Ser. B. 2003; 58:331–366.

Murphy SA. An experimental design for the development of adaptive treatment strategies. Statist. Med. 2005; 24:1455–1481.

Murphy SA, Oslin DW, Rush AJ, Zhu J. Methodological challenges in constructing effective treatment sequences for chronic psychiatric disorders. Neuropsychopharmacology. 2007; 32:257–262. [PubMed: 17091129]

Orellana L, Rotnitzky A, Robins J. Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, part I: Main content. Int. J. Biostatist. 2010; 6(Issue 2) Article 8.

Robins JM. A new approach to causal inference in mortality studies with sustained exposure periods: Applications to control of the healthy worker survivor effect. Math. Model. 1986; 7:1393–1512.

Robins, JM. Optimal structured nested models for optimal sequential decisions. In: Lin, DY.; Heagerty, PJ., editors. Proceedings of the Second Seattle Symposium on Biostatistics. New York: Springer; 2004. p. 189-326.

Robins J, Orellana L, Rotnitzky A. Estimation and extrapolation of optimal treatment and testing strategies. Statist. Med. 2008; 27:4678–4721.

Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. J. Am. Statist. Assoc. 1994; 89:846–866.

Rosthøj S, Fullwood C, Henderson R, Stewart S. Estimation of optimal dynamic anticoagulation regimes from observational data: A regret-based approach. Statist. Med. 2006; 25:4197–4215.

Rubin DB. Bayesian inference for causal effects: The role of randomization. Ann. Statist. 1978; 6:34–58.

Stefanski LA, Boos DD. The calculus of M-estimation. Amer. Statist. 2002; 56:29–38.

Tsiatis, AA. Semiparametric Theory and Missing Data. New York: Springer; 2006.

Watkins CJCH, Dayan P. Q-learning. Mach. Learn. 1992; 8:279–292.

Zhang B, Tsiatis AA, Laber EB, Davidian M. A robust method for estimating optimal treatment regimes. Biometrics. 2012; 68:1010–1018. [PubMed: 22550953]

Zhao Y, Kosorok MR, Zeng D. Reinforcement learning design for cancer clinical trials. Statist. Med. 2009; 28:3294–3315.

Table 1

Results for the first simulation scenario, Q-contrast functions correct, 1000 Monte Carlo data sets, n = 500. For the true optimal regime $g^{\mathrm{opt}} = g_\eta^{\mathrm{opt}} \in \mathscr{G}_\eta$, $\eta^{\mathrm{opt}} = (250, -1, 360, -1)^{\mathrm{T}}$ and $E\{Y^*(g_\eta^{\mathrm{opt}})\} = 1120$

Estimator	$\hat\eta_{10}$	$\hat\eta_{20}$	$\hat{E}(\hat\eta^{\mathrm{opt}})$	SE	Cov.	$E(\hat\eta^{\mathrm{opt}})$
Q-learning	228 (17)	322 (25)	1117 (12)	–	–	1119 (1)
		Propensity score correct				
A-learning	245 (18)	359 (20)	1121 (11)	–	–	1120 (1)
AIPWE (7)	210 (73)	363 (33)	1125 (12)	12	93.1	1118 (2)
DR (6)	211 (73)	363 (34)	1125 (12)	12	93.1	1118 (2)
IPWE (5)	268 (72)	397 (83)	1183 (24)	34	59.2	1105 (18)
		Propensity score correct				
A-learning	228 (17)	322 (25)	1117 (12)	–	–	1119 (1)
AIPWE (7)	259 (51)	390 (47)	1123 (12)	12	93.8	1116 (4)
DR (6)	262 (48)	386 (45)	1123 (12)	12	94.3	1116 (4)
IPWE (5)	349 (49)	471 (63)	1554 (56)	64	0.0	1075 (22)

AIPWE, DR, and IPWE, estimators based on maximizing aipwe(η) dr(η), and ipwe(η), respectively; $\hat\eta_{10}$, $\hat\eta_{20}$, Monte Carlo average estimates (standard deviation); $\hat{E}(\hat\eta^{\mathrm{opt}})$, Monte Carlo average (standard deviation) of estimated $E\{Y^*(g_\eta^{\mathrm{opt}})\}$; SE, Monte Carlo average of sandwich standard errors; Cov., coverage of associated 95% Wald-type confidence intervals for $E(\eta^{\mathrm{opt}})$; $E(\hat\eta^{\mathrm{opt}})$, Monte Carlo average (standard deviation) of values $E\{Y^*(\hat{g}_\eta^{\mathrm{opt}})\}$ obtained using 10^6 Monte Carlo simulations for each data set.

Table 2

Results for the second simulation scenario, Q-contrast functions incorrect, 1000 Monte Carlo data sets, n = 500. For the true optimal regime $g^{\text{opt}}=g_\eta^{\text{opt}} \in \mathscr{G}_\eta$, $\eta^{\text{opt}} = (250, -1, 360, -1)^{\text{T}}$ and $E\{Y^*(g_\eta^{\text{opt}})\}=1120$. All quantities are as in Table 1

Estimator	$\hat{\eta}_{10}$	$\hat{\eta}_{20}$	$\hat{E}(\hat{\eta}^{\text{opt}})$	**SE**	**Cov.**	$E(\hat{\eta}^{\text{opt}})$
Q-learning	381 (33)	386 (45)	1104 (12)	–	–	1088 (6)
			Propensity score correct			
A-learning	364 (29)	453 (26)	1115 (12)	–	–	1087 (3)
AIPWE (7)	250 (21)	359 (9)	1120 (12)	12	94.7	1118 (3)
DR (6)	250 (23)	360 (13)	1121 (11)	12	96.3	1118 (3)
IPWE (5)	305 (67)	432 (86)	1182 (27)	38	70.1	1096 (12)
			Propensity score incorrect			
A-learning	381 (33)	386 (45)	1104 (12)	–	–	1088 (6)
AIPWE (7)	255 (24)	363 (28)	1116 (12)	12	93.5	1118 (6)
DR (6)	255 (25)	364 (28)	1116 (12)	12	93.3	1117 (7)
IPWE (5)	361 (47)	480 (69)	1571 (59)	67	0.0	1086 (5)

Table 3

Results for the third simulation scenario, K = 3, Q-contrast functions incorrect, 1000 Monte Carlo data sets, n = 1000. For the true optimal regime $g^{\mathrm{opt}}=g_{\eta}^{\mathrm{opt}} \in \mathscr{G}_{\eta}$, $\eta^{\mathrm{opt}} = (250, -1, 360, -1, 300, -1)^{\mathrm{T}}$ and $E\{Y^{*}(g_{\eta}^{\mathrm{opt}})\}=1120$. All quantities are as in Table 1

Estimator	$\hat{\eta}_{10}$	$\hat{\eta}_{20}$	$\hat{\eta}_{30}$	$\hat{E}(\hat{\eta}^{\mathrm{opt}})$	**SE**	**Cov.**	$E(\hat{\eta}^{\mathrm{opt}})$
Q-learning	179 (58)	412.9 (28)	341 (33)	1058 (13)	–	–	1086 (9)
Propensity score correct							
A-learning	319 (12)	462 (11)	387 (12)	1108 (12)	–	–	1071 (3)
AIPWE (7)	263 (41)	362 (14)	300 (7)	1121 (10)	10	94.6	1116 (5)
DR (6)	263 (37)	361 (11)	300 (8)	1121 (10)	10	94.2	1117 (5)
IPWE (5)	399 (132)	618 (138)	450 (132)	1297 (63)	103	56.2	1008 (75)
Propensity score incorrect							
A-learning	179 (58)	413 (28)	341 (33)	1041 (12)	–	–	1086 (9)
AIPWE (7)	360 (48)	371 (39)	310 (30)	1200 (26)	27	9.0	1104 (10)
DR (6)	386 (35)	362 (26)	314 (39)	1208 (26)	27	4.5	1102 (9)
IPWE (5)	412 (42)	521 (60)	415 (58)	2459 (148)	167	0.0	1055 (14)