Published in final edited form as:
Stat Sci. 2014 November ; 29(4): 640–661. doi:10.1214/13-STS450.

Q- and A-learning Methods for Estimating Optimal Dynamic Treatment Regimes

Phillip J. Schulte,
Biostatistician, Duke Clinical Research Institute, Durham, North Carolina 27701, USA (phillip.schulte@duke.edu).

Anastasios A. Tsiatis,
Gertrude M. Cox Distinguished Professor, Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695-8203, USA (tsiatis@ncsu.edu).

Eric B. Laber, and
Assistant Professor, Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695-8203, USA (eblaber@ncsu.edu).

Marie Davidian
William Neal Reynolds Professor, Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695-8203, USA (davidian@ncsu.edu).

Abstract

In clinical practice, physicians make a series of treatment decisions over the course of a patient's disease based on his/her baseline and evolving characteristics. A dynamic treatment regime is a set of sequential decision rules that operationalizes this process. Each rule corresponds to a decision point and dictates the next treatment action based on the accrued information. Using existing data, a key goal is estimating the optimal regime, that, if followed by the patient population, would yield the most favorable outcome on average. *Q*- and *A*-learning are two main approaches for this purpose. We provide a detailed account of these methods, study their performance, and illustrate them using data from a depression study.

Key words and phrases

Advantage learning; bias-variance tradeoff; model misspecification; personalized medicine; potential outcomes; sequential decision making

1. INTRODUCTION

An area of current interest is personalized medicine, which involves making treatment decisions for an individual patient using all information available on the patient, including

Supplementary Material
Supplement A: Supplement to "Q- and A-learning methods for estimating optimal dynamic treatment regimes" (doi: COMPLETED BY TYPESETTER). Due to space constraints technical details and further results are given in the supplementary document Schulte et al. (2012).

genetic, physiologic, demographic, and other clinical variables, to achieve the "best" outcome for the patient given this information. In treating a patient with an ongoing disease or disorder, a clinician makes a series of decisions based on the patient's evolving status. A dynamic treatment regime is a list of sequential decision rules formalizing this process. Each rule corresponds to a key decision point in the disease/disorder progression and takes as input the information on the patient to that point and outputs the treatment that s/he should receive from among the available options. A key step toward personalized medicine is thus finding the optimal dynamic treatment regime, that which, if followed by the entire patient population, would yield the most favorable outcome on average.

The statistical problem is to estimate the optimal regime based on data from a clinical trial or observational study. *Q*-learning (*Q* denoting "quality," Watkins, 1989; Watkins and Dayan, 1992; Nahum-Shani et al., 2010) and advantage learning (*A*-learning, Murphy, 2003; Robins, 2004; Blatt, Murphy and Zhu, 2004) are two main approaches for this purpose and are related to reinforcement learning methods for sequential decision-making in computer science. *Q*-learning is based roughly on posited regression models for the outcome of interest given patient information at each decision point and is implemented through a backwards recursive fitting procedure that is related to the dynamic programming algorithm (Bather, 2000), a standard approach for deducing optimal sequential decisions. *A*-learning involves the same recursive strategy, but requires only posited models for the part of the outcome regression representing contrasts among treatments and for the probability of observed treatment assignment given patient information at each decision point. As discussed later, this may make *A*-learning more robust to model misspecification than *Q*-learning for consistent estimation of the optimal treatment regime.

Examples of the use of *Q*- and *A*-learning and alternative methods to deduce optimal strategies for treatment of substance abuse, psychiatric disorders, cancer, and HIV infection and for dose adjustment in response to evolving patient status have been presented (Rosthøj et al., 2006; Murphy et al., 2007a,b; Zhao, Kosorok and Zeng, 2009; Henderson, Ansell and Alshibani, 2010). Relevant work includes Thall, Millikan and Sung (2000), Thall, Sung and Etsey (2002), Robins (2004), Moodie, Richardson and Stephens (2007), Thall et al. (2007), van der Laan and Petersen (2007), Robins, Orellana and Rotnitzky (2008), Almirall, Ten Have and Murphy (2010), Orellana, Rotnitzky and Robins (2010), Zhang et al. (2012a,b), Zhao et al. (2012), Zhang et al. (2013) and Zhao et al. (2013).

The objective of this article is to provide readers interested in an introduction to estimation of optimal dynamic treatment regimes with a self-contained, detailed description of an appropriate statistical framework in which to define formally an optimal regime, of some of the operational and philosophical considerations involved, and of *Q*- and *A*-learning methods. Section 2 introduces the statistical framework, and Sections 3 and 4 discuss the form of the optimal regime. We describe and contrast *Q*- and *A*-learning in Section 5 and present systematic empirical studies of their relative performance and the effects of misspecification of the postulated models involved in Section 6. The methods are demonstrated using data from the Sequenced Treatment Alternatives to Relieve Depression (STAR*D, Rush et al., 2004) study in Section 7.

2. FRAMEWORK AND ASSUMPTIONS

Consider the setting of K prespecified, ordered decision points, indexed by $k = 1,\ldots, K$, which may be times or events in the disease or disorder process that necessitate a treatment decision, where, at each point, a set of treatment options is available. Assume that there is a final outcome Y of interest for which large values are preferred. The outcome may be ascertained following the Kth decision, as with CD4 T-cell count at a prespecified follow-up time in HIV infection (Moodie et al., 2007); or may be a function of information accrued over the entire sequence of decisions, as in Henderson et al. (2010), where the outcome is the overall proportion of time a measure of blood clotting speed is kept within a target range in dosing of anticoagulant agents.

In order to define an optimal treatment regime and discuss its estimation based on data from an observational study or clinical trial, we define a suitable conceptual framework. For simplicity, our presentation is heuristic. Imagine that there is a superpopulation of patients, denoted by Ω, where one may view an element $\omega \in \Omega$ as a patient from this population. We assume that patients in the population have been treated according to routine clinical practice for the disease or disorder prior to the first treatment decision. Consequently, immediately prior to this first decision, patient ω would present to the decision-maker with a set of baseline information (covariates) denoted by the random variable S_1, discussed further below. Thus, $S_1(\omega)$ is the value of his/her information immediately prior to decision 1, taking values s_1, say, in a set $\mathcal{S}_1$. Assume that, at each decision point $k = 1,\ldots, K$, there is a finite set of all possible treatment options $\mathcal{A}_k$, with elements a_k. We do not consider the case of continuous treatment and henceforth restrict attention to a finite set of options. Denote by $\bar{a}_k = (a_1, \ldots, a_k)$ a possible treatment history that could be administered through decision k, taking values in $\bar{\mathcal{A}}_k = \mathcal{A}_1 \times \ldots \times \mathcal{A}_k$, the set of all possible treatment histories $\bar{a}_K$ through all K decisions.

We then define the potential outcomes (Robins, 1986)

$$W^* = \left\{ S_2^*(a_1), S_3^*(\bar{a}_2), \ldots, S_k^*(\bar{a}_{k-1}), \ldots, S_K^*(\bar{a}_{K-1}), Y^*(\bar{a}_K) \text{for all} \bar{a}_K \in \bar{\mathcal{A}}_K \right\}. \quad (1)$$

In (1), $S_k^*(\bar{a}_{k-1})(\omega)$ denotes the value of covariate information that would arise between decisions $k - 1$ and k for a patient $\omega \in \Omega$ in the hypothetical situation that s/he were to have received previously treatment history $\bar{a}_{k-1}$, taking values s_k in a set $\mathcal{S}_k$, $k = 2, \ldots, K$. Similarly, $Y^*(\bar{a}_K)(\omega)$ is the hypothetical outcome that would result for ω were s/he to have been administered the full set of K treatments in $\bar{a}_K$. This notation implies that, for random variables such as $S_k^*(\bar{a}_{k-1}), \bar{a}_{k-1}$ is an index representing prior treatment history. Write $\bar{S}_k^*(\bar{a}_{k-1}) = \{S_1, S_2^*(a_1), \ldots, S_k^*(\bar{a}_{k-1})\}, k=1, \ldots, K$, where $\bar{S}_k^*(\bar{a}_{k-1})(\omega)$ takes values $\bar{s}_k$ in $\bar{\mathcal{S}}_k = \mathcal{S}_1 \times \ldots \times \mathcal{S}_k$; this definition includes the baseline covariate S_1 and is taken equal to S_1 when $k = 1$. The elements of the $\bar{S}_k^*(\bar{a}_{k-1}) \text{and} Y^*(\bar{a}_K)$ may be discrete or continuous; in what follows, for simplicity, we take these random variables to be discrete, but the results hold more generally.

A dynamic treatment regime $d = (d_1, \ldots, d_K)$ is a set of rules that forms an algorithm for treating a patient over time; it is "dynamic" because treatment is determined based on a patient's previous history. At the kth decision point, the kth rule $d_k(s_k, \bar{a}_{k-1})$, say, takes as input the patient's realized covariate and treatment history prior to the kth treatment decision and outputs a value $a_k \in \Psi_k(s_k, \bar{a}_{k-1}) \subseteq \mathcal{A}_k$; for $k = 1$, there is no prior treatment (a_0 is null), and we write $d_1(s_1)$ and $\Psi_1(s_1)$. Here, $\Psi_k(s_k, \bar{a}_{k-1})$ is a specified set of possible treatment options for a patient with realized history $(s_k, \bar{a}_{k-1})$, discussed further below. Accordingly, although we suppress this in the notation for brevity, the definition of a dynamic treatment regime we now present depends on the specified $\Psi_k(s_k, \bar{a}_{k-1})$, $k = 1, \ldots, K$. Because $d_k(s_k, \bar{a}_{k-1}) \in \Psi_k(s_k, \bar{a}_{k-1}) \subseteq \mathcal{A}_k$, d_k need only map a subset of $\mathcal{S}_k \times \mathcal{A}_{k-1}$ to $\mathcal{A}_k$. We define these subsets recursively as

$$\Gamma_k = \Big\{ (\bar{s}_k, \bar{a}_{k-1}) \in \overline{\mathcal{S}}_k \times \overline{\mathcal{A}}_{k-1} \text{satisfying(i)} a_j \in \Psi_j(\bar{s}_j, \bar{a}_{j-1}), j = 1, \ldots, k-1, \text{and(ii)} \Pr\big\{ \overline{S}_k^*(\bar{a}_{k-1}) = \bar{s}_k \big\} > 0 \Big\}, k = 1, \ldots, K, \tag{2}$$

determined by $\Psi = (\Psi_1, \ldots, \Psi_K)$. The Γ_k contain all realizations of covariate and treatment history consistent with having followed such Ψ-specific regimes to decision k. Define the class $\mathcal{D}$ of (Ψ-specific) dynamic treatment regimes to be the set of all d for which d_k, $k = 1, \ldots, K$, is a mapping from Γ_k into $\mathcal{A}_k$ satisfying $d_k(s_k, \bar{a}_{k-1}) \in \Psi_k(s_k, \bar{a}_{k-1})$ for every $(s_k, \bar{a}_{k-1}) \in \Gamma_k$.

Specification of the $\Psi_k(s_k, \bar{a}_{k-1})$, $k = 1, \ldots, K$, is dictated by the scientific setting and objectives. Some treatment options may be unethical or impossible for patients with certain histories, making it natural to restrict the set of possible options for such patients. In the context of public health policy, the focus may be on regimes involving only treatment options that are less costly or widely available unless a patient's condition is especially serious, as reflected in his/her covariate information. In what follows, we assume that a particular fixed set Ψ is specified, and by an optimal regime we mean an optimal regime within the class of corresponding Ψ-specific regimes.

An optimal regime should represent the "best" way to intervene to treat patients in Ω. To formalize, for any $d \in \mathcal{D}$, writing $\bar{d}_k = (d_1, \ldots, d_k)$, $k = 1, \ldots, K$, $\bar{d}_K = d$, define the potential outcomes associated with d as $\{S_2^*(d_1), \ldots, S_k^*(\bar{d}_{k-1}), \ldots, S_K^*(\bar{d}_{K-1}), Y^*(d)\}$ such that, for any $\omega \in \Omega$, with $S_1(\omega) = s_1$,

$$\begin{aligned} &d_1(s_1) = u_1, S_2^*(d_1)(\omega) = S_2^*(u_1)(\omega) = s_2, d_2(\bar{s}_2, u_1) = u_2, \ldots, d_{K-1}(\bar{s}_{K-1}, \bar{u}_{K-2}) = u_{K-1}, \\ &S_K^*(\bar{d}_{K-1})(\omega) = S_K^*(\bar{u}_{K-1})(\omega) = s_K, d_K(\bar{s}_K, \bar{u}_{K-1}) = u_K, Y^*(d)(\omega) = Y^*(\bar{u}_K)(\omega) = y. \end{aligned} \tag{3}$$

The index $\bar{d}_{k-1}$ emphasizes that $S_k^*(\bar{d}_{k-1})(\omega)$ represents the covariate information that would arise between decisions $k - 1$ and k were patient ω to receive the treatments sequentially dictated by the first $k - 1$ rules in d. Similarly, $Y^*(d)(\omega)$ is the final outcome that ω would experience if s/he were to receive the K treatments dictated by d.

With these definitions, the expected outcome in the population if all patients with initial state $S_1 = s_1$ were to follow regime d is $\mathrm{E}\{Y^*(d)|S_1 = s_1\}$. An optimal regime, $d^{\mathrm{opt}} \in \mathscr{D}$, say, satisfies

$$\mathrm{E}\{Y^*(d)\,\big|S_1{=}s_1\} \leq \mathrm{E}\{Y^*(d^{\mathrm{opt}})\,\big|\,S_1{=}s_1\} \text{for all} d \in \mathscr{D} \text{and all} s_1 \in \mathscr{S}_1. \quad (4)$$

Because (4) is true for any fixed s_1, in fact $\mathrm{E}\{Y^*(d)\} \leq \mathrm{E}\{Y^*(d^{(1)\mathrm{opt}})\}$ for any $d \in \mathscr{D}$. In Section 3, we give the form of d^{opt} satisfying (4).

Alternative specifications of Ψ may lead to different classes of regimes across which the optimal regime may differ. We emphasize that the definition (4) is predicated on the particular set Ψ, and hence class $\mathscr{D}$, of interest. In principle, the class $\mathscr{D}$ of interest is conceived based on scientific or policy objectives without reference to data available from a particular study.

Of course, potential outcomes for a given patient for all $d \in \mathscr{D}$ are not observed. Thus, the goal is to estimate d^{opt} in (4) using data from a study carried out on a random sample of n patients from Ω that record baseline and evolving covariate information and treatments actually received. Denote these available data as independent and identically distributed (i.i.d.) time-ordered random variables $(S_{1i}, A_{1i}, \ldots, S_{Ki}, A_{Ki}, Y_i)$, $i = 1, \ldots, n$, on Ω. Here, S_1 is as before; S_k, $k = 2, \ldots, K$, is covariate information recorded between decisions $k - 1$ and k, taking values $s_k \in \mathscr{S}_k$; A_k, $k = 1, \ldots, K$, is the recorded, observed treatment assignment, taking values $a_k \in \mathscr{A}_k$; and Y is the observed outcome, taking values $y \in \mathscr{Y}$. As above, define $\bar{S}_k = (S_1, \ldots. S_k)$ and $\bar{A}_k = (A_1, \ldots, A_k)$, $k = 1, \ldots, K$, taking values $\bar{s}_k \in \bar{\mathscr{S}}_k$ and $\bar{a}_k \in \bar{\mathscr{A}}_k$.

The available data may arise from an observational study involving n participants randomly sampled from the population; here, treatment assignment takes place according to routine clinical practice in the population. Alternatively, the data may arise from an intervention study. A clinical trial design that has been advocated for collecting data suitable for estimating optimal treatment regimes is that of a so-called sequential multiple-assignment randomized trial (SMART, Lavori and Dawson, 2000; Murphy, 2005). In a SMART involving K pre-specified decision points, each participant is randomized at each decision point to one of a set of treatment options, where, at the kth decision, the randomization probabilities may depend on past realized information $\bar{s}_k, \bar{a}_{k-1}$.

In order to use the observed data from either type of study to estimate an optimal regime, several assumptions are required. As is standard, we make the consistency assumption (e.g., Robins, 1994) that the covariates and outcomes observed in the study are those that potentially would be seen under the treatments actually received; that is, $S_k = S_k^*(\bar{A}_{k-1}), k = 2, \ldots, K$, and $Y = Y^*(\bar{A}_K)$. We also make the stable unit treatment value assumption (Rubin, 1978), which ensures that a patient's covariates and outcome are unaffected by how treatments are allocated to her/him and other patients. The critical assumption of no unmeasured confounders, also referred to as the sequential randomization assumption (Robins, 1994), must be satisfied. A strong version of this assumption states that A_k is conditionally independent of W^* in (1) given $\{\bar{S}_k, \bar{A}_{k-1}\}$, $k = 1, \ldots, K$, where A_0 is null,

written $A_k \perp\!\!\!\perp W^* | \bar{S}_k, \bar{A}_{k-1}$. In a SMART, this assumption is satisfied by design; in an observational study, it is unverifiable from the observed data. The strong version is sufficient for identification of the distribution of not only $Y^*(\bar{a}_K)$ but of the joint distribution of $Y^*(\bar{a}_K)$ and $\overline{S}^*_K(\overline{a}_{K-1})$ and allows the results of Section 4 to hold. Although in the population patients and their providers may make decisions based only on past covariate information available to them, the issue is whether or not all of the information that is related to treatment assignment and future covariates and outcome is recorded in the S_k; see Robins (2004, Sections 2–3) for discussion and a relaxation of the version of the sequential randomization assumption given here. We assume henceforth that these assumptions hold.

Whether or not it is possible to estimate d^{opt} from the available data is predicated on the treatment options in $\Psi_k(\bar{s}_k, \bar{a}_{k-1})$, $k = 1,\ldots, K$, being represented in the data. For a prospectively-designed SMART, ordinarily, Ψ defining the class $\mathscr{D}$ of interest would dictate the design. At decision k, subjects would be randomized to the options in $\Psi_k(\bar{s}_k, \bar{a}_{k-1})$, satisfying this condition. If the data are from an observational study, all treatment options in $\Psi_k(\bar{s}_k, \bar{a}_{k-1})$ at each decision k must have been assigned to some patients. That is, if we define recursively

$$\begin{aligned}\Gamma_1^{\max} = \{ & s_1 \in \mathscr{S}_1 : \Pr(S_1 = s_1) > 0\},\ \Psi_1^{\max}(s_1) = \{a_1 \in \mathscr{A}_1 : \Pr(A_1 = a_1 \mid S_1 = s_1) > 0 \text{ for all } s_1 \in \Gamma_1^{\max}\},\ \Gamma_k^{\max} \\ = & [(\overline{s}_k, \overline{a}_{k-1}) \in \overline{\mathscr{S}}_k \times \overline{\mathscr{A}}_{k-1} \quad \text{satisfying (i) } a_j \in \Psi_j^{\max}(\overline{s}_j, \overline{a}_{j-1}), j = 1, \ldots, k-1, \text{ and (ii)} \\ & \Pr\{\overline{S}_k^*(\overline{a}_{k-1}) = \overline{s}_k\} > 0],\ \Psi_k^{\max}(\overline{s}_k, \overline{a}_{k-1}) = \{a_k \in \mathscr{A}_k : \Pr(A_k = a_k \mid \overline{S}_k = \overline{s}_k, \overline{A}_{k-1} = \overline{a}_{k-1}) > 0 \text{ for all } (\overline{s}_k, \overline{a}_{k-1}) \in \Gamma_k^{\max}\},\ k = 2, \ldots, K,\end{aligned}$$

we must have $\Psi_k(\overline{s}_k, \overline{a}_{k-1}) \subseteq \Psi_k^{\max}(\overline{s}_k, \overline{a}_{k-1}), k = 1, \ldots, K$. The class of regimes dictated by $\Psi^{\max} = (\Psi_1^{\max}, \ldots, \Psi_K^{\max})$ is the largest that can be considered based on the data, sometimes referred to as the class of "feasible regimes" (Robins, 2004). If this inclusion condition does not hold for all $k = 1, \ldots, K$, d^{opt} cannot be estimated from the data, and the class of regimes $\mathscr{D}$ of interest must be reevaluated or another data source found.

3. OPTIMAL TREATMENT REGIMES

Q- and *A*-learning are two approaches to estimating d^{opt} satisfying (4) under the foregoing framework. Both involve recursive fitting algorithms; the main distinguishing feature is the form of the underlying models. To appreciate the rationale, one must understand how d^{opt} is determined via dynamic programming, also known as backward induction. We demonstrate the formulation of d^{opt} in terms of the potential outcomes and then show how d^{opt} may be expressed in terms of the observed data under assumptions including those in Section 2. We sometimes highlight dependence on specific elements of quantities such as $\bar{a}_k$, writing, for example, $\bar{a}_k$ as $(\bar{a}_{k-1}, a_k)$.

At the Kth decision point, for any $s_K \in \mathcal{S}_K$, $\bar{a}_{K-1} \in \mathcal{A}_{K-1}$ for which $(s_K, \bar{a}_{K-1}) \in \Gamma_K$, define

$$d_K^{(1)\text{opt}}(\bar{s}_K, \bar{a}_{K-1}) = \arg \max_{a_K \in \Psi_K(\bar{s}_K, \bar{a}_{K-1})} \mathrm{E}\{Y^*(\bar{a}_{K-1}, a_K) | \bar{S}_K^*(\bar{a}_{K-1}) = \bar{s}_K\}, \quad (5)$$

$$V_K^{(1)}(\bar{s}_K, \bar{a}_{K-1}) = \max_{a_K \in \Psi_K(\bar{s}_K, \bar{a}_{K-1})} \mathrm{E}\{Y^*(\bar{a}_{K-1}, a_K) | \bar{S}_K^*(\bar{a}_{K-1}) = \bar{s}_K\}. \quad (6)$$

For $k = K - 1, \ldots, 1$ and any $s_k \in \mathcal{S}_k$, $\bar{a}_{k-1} \in \mathcal{A}_{k-1}$ for which $(s_k, \bar{a}_{k-1}) \in \Gamma_k$, which clearly holds if $(s_K, \bar{a}_{K-1}) \in \Gamma_K$, let

$$d_k^{(1)\text{opt}}(\bar{s}_k, \bar{a}_{k-1}) = \arg \max_{a_k \in \Psi_k(\bar{s}_k, \bar{a}_{k-1})} \mathrm{E}[V_{k+1}^{(1)}\{\bar{s}_k, S_{k+1}^*(\bar{a}_{k-1}, a_k), \bar{a}_{k-1}, a_k\} | \bar{S}_k^*(\bar{a}_{k-1}) = \bar{s}_k], \quad (7)$$

$$V_k^{(1)}(\bar{s}_k, \bar{a}_{k-1}) = \max_{a_k \in \Psi_k(\bar{s}_k, \bar{a}_{k-1})} \mathrm{E}[V_{k+1}^{(1)}\{\bar{s}_k, S_{k+1}^*(\bar{a}_{k-1}, a_k), \bar{a}_{k-1}, a_k\} | \bar{S}_k^*(\bar{a}_{k-1}) = \bar{s}_k]; \quad (8)$$

thus, for

$$s_1 \in \mathcal{S}_1, d_1^{(1)\text{opt}}(s_1) = \arg \max_{a_1 \in \Psi_1(s_1)} \mathrm{E}[V_2^{(1)}\{s_1, S_2^*(a_1), a_1\} \Big| S_1 = s_1], V_1^{(1)}(s_1) = \max_{a_1 \in \Psi_1(s_1)} \mathrm{E}[V_2^{(1)}\{s_1, S_2^*(a_1), a_1\} \Big| S_1 = s_1]$$

. Conditional expectations are well-defined by (2)(ii).

Clearly, $d^{(1)\text{opt}} = (d_1^{(1)\text{opt}}, \ldots, d_K^{(1)\text{opt}})$ is a treatment regime, as it comprises a set of rules that uses patient information to assign treatment from among the options in Ψ. The superscript (1) indicates that $d^{(1)\text{opt}}$ provides K rules for a patient presenting prior to decision 1 with baseline information $S_1 = s_1$; Section 4 considers optimal treatment of patients presenting at subsequent decisions after receiving possibly sub-optimal treatment at prior decisions. Note that $d^{(1)\text{opt}}$ is defined in a backward iterative fashion. At decision K, (5) gives the treatment that maximizes the expected potential final outcome given the prior potential information, and (6) is the maximum achieved. At decisions $k = K - 1, \ldots, 1$, (7) gives the treatment that maximizes the expected outcome that would be achieved if subsequent optimal rules already defined were followed henceforth. In Section A.1 of the supplemental article [Schulte et al. (2012)], we show that $d^{(1)\text{opt}}$ defined in (5)–(8) is an optimal treatment regime in the sense of satisfying (4).

The foregoing developments express optimal regimes in terms of the distribution of potential outcomes. If an optimal regime is to be identifiable, it must be possible under the assumptions in Section 2 to express $d^{(1)\text{opt}}$ in terms of the distribution of the observed data. To this end, define

$$Q_K(\bar{s}_K, \bar{a}_K) = \mathrm{E}(Y | \bar{S}_K = \bar{s}_K, \bar{A}_K = \bar{a}_K), \quad (9)$$

$$d_K^{\text{opt}}(\bar{s}_K, \bar{a}_{K-1}) = \arg \max_{a_K \in \Psi_K(\bar{s}_K, \bar{a}_{K-1})} Q_K(\bar{s}_K, \bar{a}_{K-1}, a_K), \quad (10)$$

$$V_K(\overline{s}_K, \overline{a}_{K-1}) = \max_{a_K \in \Psi_K(\overline{s}_K, \overline{a}_{K-1})} Q_K(\overline{s}_K, \overline{a}_{K-1}, a_K), \quad (11)$$

and for $k = K - 1, \ldots, 1$, define

$$Q_k(\overline{s}_k, \overline{a}_k) = \mathrm{E}\{V_{k+1}(\overline{s}_k, S_{k+1}, \overline{a}_k) | \overline{S}_k = \overline{s}_k, \overline{A}_k = \overline{a}_k\} \quad (12)$$

$$d_k^{\mathrm{opt}}(\overline{s}_k, \overline{a}_{k-1}) = \arg \max_{a_k \in \Psi_k(\overline{s}_k, \overline{a}_{k-1})} Q_k(\overline{s}_k, \overline{a}_{k-1}, a_k), \quad (13)$$

$$V_k(\overline{s}_k, \overline{a}_{k-1}) = \max_{a_k \in \Psi_k(\overline{s}_k, \overline{a}_{k-1})} Q_k(\overline{s}_k, \overline{a}_{k-1}, a_k). \quad (14)$$

.

The expressions in (9)–(14) are well-defined under assumptions we discuss next. In (9) and (12), $Q_k(s_k, \bar{a}_k)$ are referred to as "Q-functions," viewed as measuring the "quality" associated with using treatment a_k at decision k given the history up to that decision and then following the optimal regime thereafter. The "value functions" $V_k(s_k, \bar{a}_{k-1})$ in (11) and (14) reflect the "value" of a patient's history $s_k, \bar{a}_{k-1}$ assuming that optimal decisions are made in the future. We emphasize that the $d_k^{\mathrm{opt}}, k = 1, \ldots, K$, defined (9)–(14) may not be optimal unless the sequential randomization, consistency, and positivity assumptions hold.

As in Section 2, the treatment options in Ψ must be represented in the data, i.e., $\Psi_k(\overline{s}_k, \overline{a}_{k-1}) \subseteq \Psi_k^{\max}(\overline{s}_k, \overline{a}_{k-1}), k = 1, \ldots, K$, in order to estimate an optimal regime. Formally, this implies that

$$\mathrm{pr}(A_k = a_k | \overline{S}_k = \overline{s}_k, \overline{A}_{k-1} = \overline{a}_{k-1}) > 0 \text{ if } (\overline{s}_k, \overline{a}_{k-1}) \in \Gamma_k \text{ and } a_k \in \Psi_k(\overline{s}_k, \overline{a}_{k-1}) \quad (15)$$

for all $k = 1, \ldots, K$. In Section A.2 of the supplemental article [Schulte et al. (2012)], under the consistency and sequential randomization assumptions and the positivity assumption (15), we show that, for any $(s_k, \bar{a}_{k-1}) \in \Gamma_k$ and $a_k \in \Psi_k(s_k, \bar{a}_{k-1})$, $k = 1, \ldots, K$,

$$\mathrm{pr}(\overline{S}_k = \overline{s}_k, \overline{A}_k = \overline{a}_k) > 0, \quad (16)$$

$$\mathrm{pr}(S_{k+1} = s_{k+1} \left| \overline{S}_k = \overline{s}_k, \overline{A}_k = \overline{a}_k \right.) = \mathrm{pr}\{S_{k+1}^*(\overline{a}_k) = s_{k+1} \left| \overline{S}_k = \overline{s}_k, \overline{A}_{k-1} = \overline{a}_{k-1} \right.\} \quad (17)$$

$$= \mathrm{pr}\{S_{k+1}^*(\overline{a}_k) = s_{k+1} | \overline{S}_j = \overline{s}_j, \overline{A}_{j-1} = \overline{a}_{j-1}, S_{j+1}^*(\overline{a}_j) = s_{j+1}, \ldots, S_k^*(\overline{a}_{k-1}) = s_k\}, \quad (18)$$

for $j = 1, \ldots, k$, where (18) with $j = k$ is the same as the right-hand side of (17), $S_{K+1} = Y$ and $S_{K+1}^*(\overline{a}_K) = Y^*(\overline{a}_K)$, and when $j = 1$ the conditioning events do not involve treatment. By (16), the quantities in (9)–(14) are well-defined. Under (17)–(18), the conditional distributions of the observed data involved in (9)–(14) are the same as the conditional distributions of the potential outcomes involved in (5)–(8). It follows that

$$d_k^{(1)\mathrm{opt}}(\bar{s}_k, \bar{a}_{k-1}) = d_k^{\mathrm{opt}}(\bar{s}_k, \bar{a}_{k-1}), V_k^{(1)}(\bar{s}_k, \bar{a}_{k-1}) = V_k(\bar{s}_k, \bar{a}_{k-1}), \quad (19)$$

for $(s_k, \bar{a}_{k-1}) \in \Gamma_k$, $k = 1, \ldots, K$. The equivalence in (19) shows that, under the consistency, sequential randomization, and positivity assumptions, an optimal treatment regime in the (Ψ-specific) class of interest $\mathscr{D}$ may be obtained using the distribution of the observed data.

There may not be a unique d^{opt}. At any decision k, if there is more than one possible option a_k maximizing the Q-function, then any rule d_k^{opt} yielding one of these a_k defines an optimal regime.

4. OPTIMAL "MIDSTREAM" TREATMENT REGIME

In Section 3, we define a (Ψ-specific) optimal treatment regime starting at decision point 1 and elucidate conditions under which it may be estimated using data from a clinical or observational study collected through all K decisions on a sample from the patient population. The goal is to estimate the optimal regime and implement it in new such patients presenting at the first decision.

In routine clinical practice, however, a new patient may be encountered subsequent to decision point 1. For definiteness, suppose a new patient presents "midstream," immediately prior to the ℓth decision point, $\ell = 2, \ldots, K$. A natural question is how to treat this patient optimally henceforth. For such a patient, the first $\ell - 1$ treatment decisions presumably have been made according to routine practice, and s/he has a realized past history that may be viewed as realizations of random variables $(S_1^{(P)}, A_1^{(P)}, \ldots, S_{\ell-1}^{(P)}, A_{\ell-1}^{(P)}, S_\ell^{(P)})$. Here, $A_k^{(P)}, k = 1, \ldots, \ell - 1$, represent the treatments received by such a patient according to the treatment assignment mechanism governing routine practice; and $S_k^{(P)}, k = 1, \ldots, \ell - 1$, denote covariate information collected up to the ℓth decision. Write $\overline{A}_k^{(P)} = (A_1^{(P)}, \ldots, A_k^{(P)}), k = 1, \ldots, \ell - 1 \text{ and } \overline{S}_k^{(P)} = (S_1^{(P)}, \ldots, S_k^{(P)}), k = 1, \ldots, \ell$.

As $\mathscr{A}_k$ denotes the set of all possible treatment options at decision k, $\overline{A}_{\ell-1}^{(P)}$ takes on values $\bar{a}_{\ell-1} \in \bar{\mathscr{A}}_{\ell-1}$. To define Ψ-specific regimes starting at decision ℓ, at the least, $S_k^{(P)}$ must contain the same information as S_k in the data, $k = 1, \ldots, \ell$. Because the available data dictate the covariate information incorporated in the class of regimes $\mathscr{D}$, if $S_k^{(P)}$ contains additional information, it cannot be used in the context of such regimes. We thus take $S_k^{(P)}$ and S_k to contain the same information, stated formally as the consistency assumption $S_k^{(P)} = S_k^*(\overline{A}_{k-1}^{(P)}), k = 1, \ldots, \ell$. Moreover, we can only consider treating new patients with realized histories $(s_\ell, \bar{a}_{\ell-1})$ that are contained in Γ_ℓ; that is, that could have resulted from following a Ψ-specific regime through decision $\mp - 1$. If the data arise from a SMART including only a subset of the treatments employed in practice, this may not hold.

We thus desire rules $d_k^{(\ell)}(\overline{s}_k, \overline{a}_{k-1}), k=\ell, \ell+1, \ldots, K$, say, that dictate how to treat such midstream patients presenting with realized past history $(\overline{S}_\ell^{(P)}, \overline{A}_{\ell-1}^{(P)})=(\overline{s}_\ell, \overline{a}_{\ell-1})$. In the following, we regard $(s_\ell^-, \bar{a}_{\ell-1})$ as fixed, corresponding to the particular new patient. Let $\Gamma_k^{(\ell)}$ be all elements of Γ_k with $(s_\ell^-, \bar{a}_{\ell-1})$ fixed at the values for the given new patient. Write $d^{(\ell)}=(d_\ell^{(\ell)}, d_{\ell+1}^{(\ell)}, \ldots, d_K^{(\ell)})$ to denote regimes starting at the ℓth decision point, and define the class $\mathscr{D}^{(\ell)}$ of all such regimes to be the set of all $d^{(\ell)}$ for which $d_k^{(\ell)}(\overline{s}_k, \overline{a}_{k-1})=a_k$ for $(\overline{s}_k, \overline{a}_{k-1}) \in \Gamma_k^{(\ell)}$ and $a_k \in \Psi_k(s_k^-, \bar{a}_{k-1})$ for $k = \ell,\ldots, K$. Then, by analogy to (4), we seek $d^{(\ell)\text{opt}}$ satisfying

$$\mathrm{E}\{Y^*(\overline{a}_{\ell-1}, d^{(\ell)}) \,\big|\, \overline{S}_\ell^{(P)}=\overline{s}_\ell, \overline{A}_{\ell-1}^{(P)}=\overline{a}_{\ell-1}\} \le \mathrm{E}\{Y^*(\overline{a}_{\ell-1}, d^{(\ell)\text{opt}}) \,\big|\, \overline{S}_\ell^{(P)}=\overline{s}_\ell, \overline{A}_{\ell-1}^{(P)}=\overline{a}_{\ell-1}\} \quad (20)$$

for all $d^{(\ell)} \in \mathscr{D}^{(\ell)}$ and $s_\ell^- \in \mathscr{S}_\ell^-$, $\bar{a}_{\ell-1} \in \mathscr{A}_{\ell-1}^-$ for which $\mathrm{pr}(\overline{S}_\ell^{(P)}=\overline{s}_\ell, \overline{A}_{\ell-1}^{(P)}=\overline{a}_{\ell-1})>0$. Viewing this as a problem of making $K - \ell+1$ decisions at decision points $\ell, \ell+1,\ldots, K$, with initial state $\overline{S}_\ell^{(P)}=\overline{s}_\ell, \overline{A}_{\ell-1}^{(P)}=\overline{a}_{\ell-1}$, by an argument analogous to that in Section A.1 of the supplemental article [Schulte et al. (2012)] for $\ell = 1$ and initial state $S_1 = s_1$ letting $\mathscr{V}_{\ell,k}=\{\overline{S}_\ell^{(P)}=\overline{s}_\ell, \overline{A}_{\ell-1}^{(P)}=\overline{a}_{\ell-1}, S_{\ell+1}^*(\overline{a}_\ell)=s_{\ell+1}, \ldots, S_k^*(\overline{a}_{k-1})=s_k\}$, it may be shown that $d^{(\ell)\text{opt}}$ satisfying (20) is given by

$$d_K^{(\ell)\text{opt}}(\overline{s}_K, \overline{a}_{K-1})=\arg\max_{a_K \in \Psi_K(\overline{s}_K, \overline{a}_{K-1})} \mathrm{E}\{Y^*(\overline{a}_{K-1}, a_K) \,|\, \mathscr{V}_{\ell,K}\}, \quad (21)$$

$$V_K^{(\ell)}(\overline{s}_K, \overline{a}_{K-1})=\max_{a_K \in \Psi_K(\overline{s}_K, \overline{a}_{K-1})} \mathrm{E}\{Y^*(\overline{a}_{K-1}, a_K) \,|\, \mathscr{V}_{\ell,K}\} \quad (22)$$

for any $s_K^- \in \mathscr{S}_K^-$, $\bar{a}_{K-1} \in \mathscr{A}_{K-1}^-$ for which $(\overline{s}_K, \overline{a}_{K-1}) \in \Gamma_K^{(\ell)}$; and, for $k = K - 1,\ldots \ell$,

$$d_k^{(\ell)\text{opt}}(\overline{s}_k, \overline{a}_{k-1})=\arg\max_{a_k \in \Psi_k(\overline{s}_k, \overline{a}_{k-1})} \mathrm{E}[V_{k+1}^{(\ell)}\{\overline{s}_k, S_{k+1}^*(\overline{a}_{k-1}, a_k), \overline{a}_{k-1}, a_k\} \,|\, \mathscr{V}_{\ell,k}], \quad (23)$$

$$V_k^{(\ell)}(\overline{s}_k, \overline{a}_{k-1})=\max_{a_k \in \Psi_k(\overline{s}_k, \overline{a}_{k-1})} \mathrm{E}[V_{k+1}^{(\ell)}\{\overline{s}_k, S_{k+1}^*(\overline{a}_{k-1}, a_k), \overline{a}_{k-1}, a_k \,|\, \mathscr{V}_{\ell,k}] \quad (24)$$

for any $s_k^- \in \mathscr{S}_k^-$, $\bar{a}_{k-1} \in \mathscr{A}_{k-1}^-$ for which $(\overline{s}_k, \overline{a}_{k-1}) \in \Gamma_k^{(\ell)}$, so that

$$d_\ell^{(\ell)\text{opt}}(\overline{s}_\ell, \overline{a}_{\ell-1})=\arg\max_{a_\ell \in \Psi_\ell(\overline{s}_\ell, \overline{a}_{\ell-1})} \mathrm{E}[V_{\ell+1}^{(\ell)}\{\overline{s}_\ell, S_{\ell+1}^*(\overline{a}_{\ell-1}, a_\ell), \overline{a}_{\ell-1}, a_\ell\} \,|\, \overline{S}_\ell^{(P)}=\overline{s}_\ell, \overline{A}_{\ell-1}^{(P)}=\overline{a}_{\ell-1}].$$

Comparison of (5)–(8) to (21)–(24) shows that the ℓth to Kth rules of the optimal regime $d^{(1)\text{opt}}$ that would be followed by a patient presenting at the first decision are not necessarily the same as those of the optimal regime $d^{(\ell)\text{opt}}$ that would be followed by a patient presenting at the ℓth decision. In particular, noting that the conditioning sets in (5)–(8) are $\mathscr{V}_{1,K}$ and $\mathscr{V}_{1,k}$, the rules are ℓ-dependent through dependence of the conditioning sets $\mathscr{V}_{\ell,k}$, ℓ

= 1, …, K, $k = \ell,\ldots, K$, on ℓ. However, we now demonstrate that these rules coincide under certain conditions.

Make the consistency, sequential randomization, and positivity (15) assumptions on the available data required to show (19) in Section 3, along with the consistency assumption on the $S_k^{(P)}$ above and the sequential randomization assumption $A_k^{(P)} \perp\!\!\!\perp W^* \big| \overline{S}_k^{(P)}, \overline{A}_{k-1}^{(P)}, k=1,\ldots,\ell-1$, which ensures that the $\overline{S}_k^{(P)}$ include all information related to treatment assignment and future covariates and outcome up to decision ℓ. Note that (21)–(24) are expressed in terms of the conditional distributions $\mathrm{pr}\{S_{k+1}^*(\overline{a}_k)=s_{k+1} \big| \overline{S}_\ell^{(P)}=\overline{s}_\ell, \overline{A}_{\ell-1}^{(P)}=\overline{a}_{\ell-1}, S_{\ell+1}^*(\overline{a}_\ell)=s_\ell,\ldots, S_k^*(\overline{a}_{k-1})=s_k\}, k=\ell,\ldots,K$. We can then use (18) with $j = \ell$ to deduce that these conditional distributions can be written equivalently as $\mathrm{pr}\{S_{k+1}^*(\overline{a}_k)=s_{k+1} \big| \overline{S}_k^*(\overline{a}_{k-1})=\overline{s}_k\}, k=\ell,\ldots,K$, so solely in terms of the distribution of the potential outcomes. By (17) and (18) with $j = 1$, this can be written as $\mathrm{pr}(S_{k+1} = s_{k+1} \big| \bar{S}_k = \bar{s}_k, \bar{A}_k = \bar{a}_k)$. This shows that (21)–(24) can be reexpressed in terms of the observed data, so that, for $(\bar{s}_k, \bar{a}_{k-1}) \in \Gamma_k$ for $\ell = 1,\ldots, K$ and $k = \ell,\ldots, K$,

$$d_k^{(\ell)\mathrm{opt}}(\overline{s}_k, \overline{a}_{k-1}) = d_k^{\mathrm{opt}}(\overline{s}_k, \overline{a}_{k-1}), V_k^{(\ell)}(\overline{s}_k, \overline{a}_{k-1}) = V_k(\overline{s}_k, \overline{a}_{k-1}). \quad (25)$$

Note that (25) subsumes (19) when $\ell = 1$. The equivalence in (25) not only demonstrates that an optimal treatment regime can be obtained using the distribution of the observed data but also that the corresponding rules dictating treatment do not depend on ℓ under these assumptions. Thus, the single set of rules $d^{\mathrm{opt}} = (d_1^{\mathrm{opt}}, \ldots, d_K^{\mathrm{opt}})$ defined in (10) and (13) is relevant regardless of when a patient presents. That is, treatment at the ℓth decision point for a patient who presents at decision 1 and has followed the rules in d^{opt} to that point would be determined by d_ℓ^{opt} evaluated at his/her history up to that point, as would treatment for a subject presenting for the first time immediately prior to decision ℓ. See Robins (2004, pages 305–306) for more discussion.

5. Q- AND A-LEARNING

5.1 Q-Learning

From (10), (13) and (19), an optimal (Ψ-specific) regime d^{opt} may be represented in terms of the Q-functions (9), (12). Thus, estimation of d^{opt} based on i.i.d. data $(S_{1i}, A_{1i},\ldots, S_{Ki}, A_{Ki}, Y_i)$, $i = 1,\ldots, n$, may be accomplished via direct modeling and fitting of the Q-functions. This is the approach underlying Q-learning. Specifically, one may posit models $Q_k(\bar{s}_k, \bar{a}_k; \xi_k)$, say, for $k = K, K-1,\ldots, 1$, each depending on a finite-dimensional parameter ξ_k. The models may be linear or nonlinear in ξ_k and include main effects and interactions in the elements of $\bar{s}_k$ and $\bar{a}_k$.

Estimators $\hat{\xi}_k$ may be obtained in a backward iterative fashion for $k = K, K-1,\ldots, 1$ by solving suitable estimating equations [e.g., ordinary (OLS) or weighted (WLS) least squares]. Assuming the latter, for $k = K$, letting $\tilde{V}_{(K+1)i} = Y_i$ one would first solve

$$\sum_{i=1}^{n}\frac{\partial Q_K(\overline{S}_{Ki},\overline{A}_{Ki};\xi_K)}{\partial\xi_K}\Sigma_K^{-1}(\overline{S}_{Ki},\overline{A}_{Ki})\{\tilde{V}_{(K+1)i}-Q_K(\overline{S}_{Ki},\overline{A}_{Ki};\xi_K)\}=0 \quad (26)$$

in ξ_K to obtain $\hat{\xi}_K$, where $\Sigma_K(s_K, \bar{a}_K)$ is a working variance model. Substituting the model $Q_K(s_K, \bar{a}_K; \xi_K)$ in (10) and accordingly writing $d_K^{\text{opt}}(\overline{s}_K, \overline{a}_{K-1};\xi_K)$, substituting $\hat{\xi}_K$ for ξ_K yields an estimator for the optimal treatment choice at decision K for a patient with past history $S_K = s_K$, $\bar{A}_{K-1} = \bar{a}_{K-1}$. With $\hat{\xi}_K$ in hand, one would form for each i, based on (11), $\tilde{V}_{Ki} = \max_{a_K\in\Psi_K(S_{Ki},\bar{A}_{(K-1)i})} Q_K(S_{Ki},\bar{A}_{(K-1)i},a_K;\hat{\xi}_K)$. To obtain $\hat{\xi}_{K-1}$, setting $k = K - 1$, based on (12) and letting $\Sigma_k(s_k, \bar{a}_k)$ be a working variance model, one would then solve for ξ_k

$$\sum_{i=1}^{n}\frac{\partial Q_k(\overline{S}_{ki},\overline{A}_{ki};\xi_k)}{\partial\xi_k}\Sigma_k^{-1}(\overline{S}_{ki},\overline{A}_{ki})\{\tilde{V}_{(k+1)i}-Q_k(\overline{S}_{ki},\overline{A}_{ki};\xi_k)\}=0. \quad (27)$$

The corresponding $d_{K-1}^{\text{opt}}(\overline{s}_{K-1},\overline{a}_{K-2};\hat{\xi}_{K-1})$ yields an estimator for the optimal treatment choice at decision $K - 1$ for a patient with past history $S_{K-1} = s_{K-1}$, $\bar{A}_{K-2} = \bar{a}_{K-2}$, assuming s/he will take the optimal treatment at decision K. One would continue this process in the obvious fashion for $k = K - 2,\ldots, 1$, forming $\tilde{V}_{ki} = \max_{a_k\in\Psi_k(S_{ki},\bar{A}_{(k-1)i})} Q_k(S_{ki},\bar{A}_{(k-1)i},a_k;\hat{\xi}_k)$, and solving equations of form (27) to obtain $\hat{\xi}_k$ and corresponding $d_k^{\text{opt}}(\overline{s}_k,\overline{a}_{k-1};\hat{\xi}_k)$.

We may now summarize the estimated optimal regime as $\hat{d}_Q^{\text{opt}}=(\hat{d}_{Q,1}^{\text{opt}},\ldots,\hat{d}_{Q,K}^{\text{opt}})$, where

$$\hat{d}_{Q,1}^{\text{opt}}(s_1)=d_1^{\text{opt}}(s_1;\hat{\xi}_1),\ \hat{d}_{Q,k}^{\text{opt}}(\overline{s}_k,\overline{a}_{k-1})=d_k^{\text{opt}}(\overline{s}_k,\overline{a}_{k-1};\hat{\xi}_k),\ k=2,\ldots,K. \quad (28)$$

It is important to recognize that, even under the sequential randomization assumption, the estimated regime (28) may not be a consistent estimator for the true optimal regime unless all the models for the Q-functions are correctly specified.

We illustrate the approach for $K = 2$, where at each decision there are two possible treatment options coded as 0 and 1; i.e., $\Psi_1(s_1) = \mathcal{A}_1 = \{0,1\}$ for all s_1 and $\Psi_2(s_2, a_1) = \mathcal{A}_2 = \{0,1\}$ for all s_2 and $a_1 \in \{0,1\}$. Let $\mathcal{H}_1=(1,s_1^T)^T$ and $\mathcal{H}_2=(1,s_1^T,a_1,s_2^T)^T$. As in many modeling contexts, it is standard to adopt linear models for the Q-functions; accordingly, consider the models

$$Q_1(s_1,a_1;\xi_1)=\mathcal{H}_1^T\beta_1+a_1(\mathcal{H}_1^T\psi_1),\ Q_2(\overline{s}_2,\overline{a}_2;\xi_2)=\mathcal{H}_2^T\beta_2+a_2(\mathcal{H}_2^T\psi_2), \quad (29)$$

where $\xi_k=(\beta_k^T,\psi_k^T)^T$, $k=1,2$. In (29), $Q_2(s_2, \bar{a}_2; \xi_2)$ is a model for $E(Y|S_2 = s_2, \bar{A}_2 = \bar{a}_2)$, a standard regression problem involving observable data, whereas $Q_1(s_1, a_1;\xi_1)$ is a model for the conditional expectation of $V_2(s_2, a_1 = \max_{a_2\in\{0,1\}} E(Y|S_2 = s_2, A_1 = a_1, A_2 = a_2)$ given $S_1 = s_1$ and $A_1 = a_1$, which is an approximation to a complex true relationship; see Section 5.3. Under (29), $V_2(\overline{s}_2,a_1;\xi_2)=\max_{a_2\in\{0,1\}}Q_2(\overline{s}_2,a_1,a_2;\xi_2)=\mathcal{H}_2^T\beta_2+(\mathcal{H}_2^T\psi_2)I(\mathcal{H}_2^T\psi_2>0)$ and $V_1(s_1;\xi_1)=\max_{a_1\in\{0,1\}}Q_1(s_1,a_1;\xi_1)=\mathcal{H}_1^T\beta_1+(\mathcal{H}_1^T\psi_1)I(\mathcal{H}_1^T\psi_1>0)$. Substituting the

Q-functions in (29) in (10) and (13) then yields $d_1^{\text{opt}}(s_1;\xi_1)=I(\mathscr{H}_1^T\psi_1>0)$ and $d_2^{\text{opt}}(\bar{s}_2,a_1;\xi_2)=I(\mathscr{H}_2^T\psi_2>0)$.

We have presented (26) and (27) in the conventional WLS form, with leading term in the summand $\partial/\partial\xi_k Q_k(\overline{S}_{ki},\overline{A}_{ki};\xi_k)\Sigma_k^{-1}(\overline{S}_{ki},\overline{A}_{ki})$; taking Σ_k to be a constant yields OLS. At the *K*th decision, with responses Y_i, standard theory implies that this is the optimal leading term when $\text{var}(Y|\bar{S}_K = s_K, \bar{A}_K = a_K) = \Sigma_K(s_K, \bar{a}_K)$, yielding the (asymptotically) efficient estimator for ξ_K. For $k < K$, with "responses" $\tilde{V}_{(k+1)i}$, this theory may no longer apply; however, deriving the optimal leading term involves considerable complication. Accordingly, it is standard to fit the posited models $Q_k(s_k, \bar{a}_k; \xi_k)$ via OLS or WLS; some authors define *Q*-learning as using OLS (Chakraborty, Murphy and Strecher, 2010). The choice may be dictated by apparent relevance of the homoscedasticity assumption on the $\tilde{V}_{(k+1)i}$, $k = K, K-1, \ldots, 1$, and whether or not linear models are sufficient to approximate the relationships may also be evaluated, but see Section 5.3.

5.2 A-Learning

Advantage learning (*A*-learning, Blatt et al., 2004) is a term used to describe a class of alternative methods to *Q*-learning predicated on the fact that the entire *Q*-function need not be specified to estimate the optimal regime. For simplicity, we consider here only the case of two treatment options coded as 0 and 1 at each decision; i.e., $\Psi_k(s_k, \bar{a}_{k-1}) = \mathscr{A}_k = \{0,1\}$, $k = 1,\ldots, K$.

To fix ideas, consider (29). Note that $d_1^{\text{opt}}(s_1;\xi_1)$ implied by (29) depends only on $\mathscr{H}_1^T\psi_1=Q_1(s_1,1;\xi_1)-Q_1(s_1,0;\xi_1)$; likewise, $d_2^{\text{opt}}(\bar{s}_2,a_1;\xi_2)$ depends only on $\mathscr{H}_2^T\psi_2=Q_2(\bar{s}_2,a_1,1;\xi_2)-Q_2(\bar{s}_2,a_1,0;\xi_2)$. This reflects the general result that, for purposes of deducing the optimal regime, for each $k = 1,\ldots, K$, it suffices to know the contrast function $C_k(s_k, \bar{a}_{k-1}) = Q_k(s_k, \bar{a}_{k-1}, 1) - Q_k(s_k, \bar{a}_{k-1}, 0)$. This can be appreciated by noting that any arbitrary $Q_k(s_k, \bar{a}_k)$ may be written as $h_k(s_k, \bar{a}_{k-1}) + a_k C_k(s_k, \bar{a}_{k-1})$, where $h_k(s_k, \bar{a}_{k-1}) = Q_k(s_k, \bar{a}_{k-1}, 0)$, so that $Q_k(s_k, \bar{a}_{k-1}, a_k)$ is maximized by taking $a_k = I\{C_k(s_k, \bar{a}_{k-1}) > 0\}$; and the maximum itself is the expression $h_k(s_k, \bar{a}_{k-1}) + C_k(s_k, \bar{a}_{k-1})I\{C_k(s_k, \bar{a}_{k-1}) > 0\}$. In the case of two treatment options we consider here, the contrast function is also referred to as the optimal-blip-to-zero function (Robins, 2004; Moodie et al., 2007). Murphy (2003) considers the expression $C_k(S_k, \bar{A}_{k-1})[I\{C_k(S_k, \bar{A}_{k-1}) > 0\} - A_k]$, referred to as the advantage or regret function, as it represents the "advantage" in response incurred if the optimal treatment at the *k*th decision were given relative to that actually received (or, equivalently, the "regret" incurred by not using the optimal treatment). See Robins (2004) and Moodie et al. (2007) for discussion of the relationship between regrets and optimal blip functions in this and settings other than binary treatment options.

We discuss here an *A*-learning method based on explicit modeling of the contrast functions, which we refer to as contrast-based *A*-learning. This approach is implemented via recursive solution of certain estimating equations given below developed by Robins (2004), often referred to as g-estimation. See Moodie et al. (2007) and the supplementary material to Zhang et al. (2013) for details. Contrast-based *A*-learning is distinguished from the regret-

based *A*-learning methods of Murphy (2003) and Blatt et al. (2004), which rely on direct modeling of the regret functions and are implemented using a different estimating equation formulation called Iterative Minimization for Optimal Regimes by Moodie et al. (2007).

All of these methods are alternatives to *Q*-learning, which involves modeling the full *Q*-functions. For $k = K - 1,\ldots, 1$, the *Q*-functions involve possibly complex relationships, raising concern over the consequences of model misspecification for estimation of the optimal regime. As identifying the optimal regime depends only on correct specification of the contrast or regret functions, *A*-learning methods may be less sensitive to mismodeling; see Sections 5.3 and 6.

Although we consider these methods only in the case of binary treatment options here, they may be extended to more than two treatments at the expense of complicating the formulation; see Robins (2004) and Moodie et al. (2007).

Contrast-based *A*-learning proceeds as follows. Posit models $C_k(s_k, \bar{a}_{k-1}; \psi_k)$, $k = 1,\ldots, K$, for the contrast functions, depending on parameters ψ_k. Consider decision *K*. Let $\pi_K(s_K, \bar{a}_{K-1}) = \text{pr}(A_K = 1|S_K = s_K, \bar{A}_{K-1} = \bar{a}_{K-1})$ be the propensity of receiving treatment 1 in the observed data as a function of past history and $\tilde{V}_{(K+1)i} = Y_i$. Robins (2004) showed that all consistent and asymptotically normal estimators for ψ_K are solutions to estimating equations of the form

$$\sum_{i=1}^{n}\lambda_K(\overline{S}_{Ki}, \overline{A}_{(K-1)i})\{A_{Ki}-\pi_K(\overline{S}_{Ki}, \overline{A}_{(K-1)i})\}\times\{\tilde{V}_{(K+1)i}-A_{Ki}C_K(\overline{S}_{Ki}, \overline{A}_{(K-1)i};\psi_K)-\theta_K(\overline{S}_{Ki}, \overline{A}_{(K-1)i})\}=0 \quad (30)$$

for arbitrary functions $\lambda_K(s_K, \bar{a}_{K-1})$ of the same dimension as ψ_K and arbitrary functions $\theta_K(s_K, \bar{a}_{K-1})$. Assuming that the model $C_K(s_K, \bar{a}_{K-1}; \psi_K)$ is correct, if $\text{var}(Y|S_K = s_k, \bar{A}_{K-1} = a_{k-1})$ is constant, the optimal choices of these functions are given by $\lambda_K(s_K, \bar{a}_{K-1}; \psi_K) = \partial/\partial\psi_K C_K(s_K, \bar{a}_{K-1}; \psi_K)$ and $\theta_K(s_{Ki}, \bar{a}_{(K-1)i}) = h_K(s_K, \bar{a}_{K-1})$; otherwise, if the variance is not constant, the optimal λ_K is complex (Robins, 2004).

To implement estimation of ψ_K via (30), one may adopt parametric models for these functions. Although *A*-learning obviates the need to specify fully the *Q*-functions, one may posit models for the optimal θ_K, $h_K(s_K, \bar{a}_{K-1}; \beta_K)$, say. Moreover, unless the data are from a SMART study, in which case the propensities $\pi_K(s_K, \bar{a}_{K-1})$ are known, these may be modeled as $\pi_K(s_K, \bar{a}_{K-1}; \varphi_K)$ (e.g., by a logistic regression). These models are only adjuncts to estimating ψ_K; as long as at least one of these models is correctly specified, (30) will yield a consistent estimator for ψ_K, the so-called double robustness property. In contrast, *Q*-learning requires correct specification of all *Q*-functions; see Section 5.3 and Section A.5 of the supplemental article [Schulte et al. (2012).]

Substituting these models in (30), one solves (30) jointly in $(\psi_K^T, \beta_K^T, \phi_K^T)^T$ with

$$\sum_{i=1}^{n}\frac{\partial h_K(\overline{S}_K, \overline{A}_{K-1};\beta_K)}{\partial\beta_K}\{\tilde{V}_{(K+1)i}-A_{Ki}C_K(\overline{S}_{Ki}, \overline{A}_{(K-1)i};\psi_K)-h_K(\overline{S}_{Ki}, \overline{A}_{(K-1)i};\beta_K)\}=0$$

and the usual binary regression likelihood score equations in φ_K. We then have $d_K^{\text{opt}}(\bar{s}_K, \bar{a}_{K-1}; \psi_K) = I\{C_K(\bar{s}_k, \bar{a}_{K-1}; \psi_K) > 0\}$; as in *Q*-learning, substituting $\hat{\psi_K}$ yields an estimator for the optimal treatment choice at decision *K* for a patient with past history $\bar{S}_K = \bar{s}_K$, $\bar{A}_{K-1} = \bar{a}_{K-1}$.

With $\hat{\psi_K}$ in hand, the contrast-based *A*-learning algorithm proceeds in a backward iterative fashion to yield $\hat{\psi_k}$, $k = K - 1, \ldots, 1$. At the *k*th decision, given models $h_k(\bar{s}_k, \bar{a}_{k-1}; \beta_k)$ and $\pi_k(\bar{s}_k, \bar{a}_{k-1}; \varphi_k)$, one solves jointly in $(\psi_K^T, \beta_K^T, \phi_K^T)^T$ a system of estimating equations analogous to those above. The *k*th set of equations is based on "optimal responses" $\tilde{V}_{(k+1)i}$, where, for each *i*, $\tilde{V}_{ki}$ estimates $V_k(\bar{S}_{ki}, \bar{A}_{(k-1),i})$. It may be shown (see Section A.3 of the supplemental article [Schulte et al. (2012)]) that $E(V_{k+1}(\bar{S}_{k+1}, \bar{A}_k) + C_k(\bar{S}_k, \bar{A}_{k-1})[I\{C_k(\bar{S}_k, \bar{A}_{k-1}) > 0\} - A_k] \mid \bar{S}_k, \bar{A}_{k-1}) = V_k(\bar{S}_k, \bar{A}_{k-1})$. Accordingly, define recursively $\tilde{V}_{ki} = \tilde{V}_{(k+1)i} + C_k(\bar{S}_{ki}, \bar{A}_{(k-1)i}; \hat{\psi_k})[I(C_k(\bar{S}_{ki}, \bar{A}_{(k-1)i}; \hat{\psi_k}) > 0\} - A_{ki}]$, $k = K, K-1, \ldots 1$, $\tilde{V}_{(K+1)i} = Y_i$. The equations at the *k*th decision are then

$$\sum_{i=1}^{n} \lambda_k(\bar{S}_{ki}, \bar{A}_{(k-1)i}; \psi_k)\{A_{ki} - \pi_k(\bar{S}_{ki}, \bar{A}_{(k-1)i}; \phi_k)\} \times \{\tilde{V}_{(k+1)i} - A_{ki} C_k(\bar{S}_{ki}, \bar{A}_{(k-1)i}; \psi_k) - h_k(\bar{S}_{ki}, \bar{A}_{(k-1)i}; \beta_k)\} = 0,$$
$$\sum_{i=1}^{n} \frac{\partial h_k(\bar{S}_k, \bar{A}_{k-1}; \beta_k)}{\partial \beta_k} \{\tilde{V}_{(k+1)i} - A_{ki} C_k(\bar{S}_{ki}, \bar{A}_{(k-1)i}; \psi_k) - h_k(\bar{S}_{ki}, \bar{A}_{(k-1)i}; \beta_k)\} = 0, \quad (31)$$

for a given specification $\lambda_k(\bar{s}_k, \bar{a}_{k-1}; \psi_k)$, solved jointly with the maximum likelihood score equations for binary regression in φ_k. It follows that $d_k^{\text{opt}}(\bar{s}_k, \bar{a}_{k-1}; \hat{\psi}_k) = I\{C_k(\bar{s}_k, \bar{a}_{k-1}; \hat{\psi}_k) > 0\}$. As above, the optimal λ_k is complex (Robins, 2004); taking $\lambda_k(\bar{s}_k, \bar{a}_{k-1}; \psi_k) = \partial/\partial\psi_k \, C_k(\bar{s}_k, \bar{a}_{k-1}; \psi_k)$ is reasonable for practical implementation.

Summarizing, the estimated optimal regime $\hat{d}_A^{\text{opt}} = (\hat{d}_{A,1}^{\text{opt}}, \ldots, \hat{d}_{A,K}^{\text{opt}})$ is

$$\hat{d}_{A,1}^{\text{opt}}(s_1) = d_1^{\text{opt}}(s_1; \hat{\psi}_1), \hat{d}_{A,k}^{\text{opt}}(\bar{s}_k, \bar{a}_{k-1}) = d_k^{\text{opt}}(\bar{s}_k, a_{k-1}; \hat{\psi}_k), k = 2, \ldots, K, \quad (32)$$

How well $\hat{d}_A^{\text{opt}}$ estimates d^{opt} and hence $d^{(1)\text{opt}}$ depends on how close the posited $C_k(\bar{s}_k, \bar{a}_{k-1}; \psi_k)$ are to the true contrast functions as well as correct specification of the functions h_k or π_k.

Henceforth, for brevity, we suppress the descriptor "contrast-based" and refer to the foregoing approach simply as *A*-learning.

5.3 Comparison and Practical Considerations

When $K = 1$, the *Q*-function is a model for $E(Y|S_1 = s_1, A_1 = a_1)$. If in *Q*-learning this model and the variance model Σ_1 in (26) are correctly specified, then, as above, the form of (26) is optimal for estimating ξ_1. Accordingly, even if $C_1(s_1; \psi_1)$ and $h_1(s_1; \beta_1)$ are correctly modeled, (31) with $K = 1$ is generally not of this optimal form for any choice $\lambda_1(s_1; \psi_1)$, and hence *A*-learning will yield relatively inefficient inference on ψ_1 and the optimal regime. However, if in *Q*-learning the *Q*-function is mismodeled, but in *A*-learning $C_1(s_1; \psi_1)$ and

$\pi_1(s_1;\varphi_1)$ are both correctly specified, then *A*-learning will still yield consistent inference on ψ_1 and hence the optimal regime, whereas inference on ξ_1 and the optimal regime via *Q*-learning may be inconsistent. We assess the trade-off between consistency and efficiency in this case in Section 6. For $K > 1$, owing to the complications involved in specifying optimal estimating equations for *Q*- and *A*-learning, relative performance is not readily apparent; we investigate empirically in Section 6.

In special cases, *Q*- and *A*-learning lead to identical estimators for the *Q*-function (Chakraborty et al., 2010). For example, this holds if the propensities for treatment are constant, as would be the case under pure randomization at each decision point, and certain linear models are used for $C_1(s_1;\psi_1)$ and $h_1(s_1;\beta_1)$; Section A.4 of the supplemental article [Schulte et al. (2012)] demonstrates when $K = 1$ and $\text{pr}(A_1 = 1|S_1 = s_1)$ does not depend on s_1. See Robins (2004, page 1999) and Rosenblum and van der Laan (2009) for further discussion.

As we have emphasized, for *Q*-learning, while modeling the *Q*-function at decision *K* is a standard regression problem with response *Y*, for decisions $k = K - 1,\ldots, 1$, this involves modeling the estimated value function, which at decision *k* depends on relationships for future decisions $k + 1,\ldots, K$. Ideally, the sequence of posited models $Q_k(s_k, \bar{a}_k; \xi_k)$ should respect this constraint. However, this may be difficult to achieve with standard regression models. To illustrate, consider (29), and assume S_1, S_2 are scalar, where the conditional distribution of S_2 given $S_1 = s_1$, $A_1 = a_1$ is Normal $(\mathcal{K}_1^T\gamma, \sigma^2)$, say, $\mathcal{K}_1 = (1, s_1, a_1)^T$. Recall that $V_2(\bar{s}_2, a_1;\xi_2)=\mathcal{H}_2^T\beta_2+(\mathcal{H}_2^T\psi_2)I(\mathcal{H}_2^T\psi_2>0)$, where $\mathcal{H}_2^T\beta_2=\mathcal{K}_1^T\beta_{21}+s_2\beta_{22}$ and $\mathcal{H}_2^T\psi_2=\mathcal{K}_1^T\psi_{21}+s_2\psi_{22}$. Then, if the model Q_2 in (29) were correct, from (12), ideally, $Q_1(s_1, a_1) = \text{E}\{V_2(s_1, S_2, a_1; \xi_2)|S_1 = s_1, A_1 = a_1\}$. Letting $\phi(\cdot)$ and $\Phi(\cdot)$ be the standard normal density and cumulative distribution function, respectively, it may be shown (see Section A.5 of the supplemental article [Schulte et al. (2012)]) that

$$\begin{aligned} Q_1(s_1,a_1) &= \text{E}\{V_2(s_1,S_2,a_1;\xi_2)|S_1 \\ &= s_1, A_1 \\ &= a_1\} = \mathcal{K}_1^T(\beta_{21} \\ &+ \gamma\beta_{22}) \\ &+ (\mathcal{K}_1^T\psi_{21})\{1 \\ &- \Phi(\eta)\} + \psi_{22}\{\sigma\varphi(\eta) \\ &+ (\mathcal{K}_1^T\gamma)\{1 \\ &- \Phi(\eta)\}\}, \eta \\ &= -\mathcal{K}_1^T(\psi_{21}/\psi_{22}+\gamma)/\sigma, \end{aligned} \tag{33}$$

taking $\psi_{22} > 0$. The true $Q_1(s_1, a_1)$ in (33) is clearly highly nonlinear and likely poorly approximated by the posited linear model $Q_1(s_1, a_1; \xi_1)$ in (29). For larger *K*, this incompatibility between true and assumed models would propagate from $K - 1,\ldots, 1$. Thus, while using linear models for the *Q*-functions is popular in practice, the potential for such mismodeling should be recognized.

An approach that may mitigate the risk of mismodeling is to employ flexible models for the *Q*-functions; Zhao, Kosorok and Zeng (2009) use support vector regression models. Developments in statistical learning suggest a large collection of powerful regression methods that might be used. Many of these methods must be tuned in order to balance bias and variance, a natural approach to which is to minimize the cross-validated mean squared error of the *Q*-functions at each decision point. An obvious downside is that the final model may be difficult to interpret, and clinicians may not be willing to use "black box" rules. One compromise is to fit a simple, interpretable model, such as a decision tree, to the fitted values of the complex model in order explore the factors driving the recommended treatment decisions. This simple model can then be checked against scientific theory. If it appears sensible, then clinicians may be willing to use predictions from the complex model. For discussion, see Craven and Shavlik (1996).

A-learning represents a middle ground between *Q*-learning and these approaches in that it allows for flexible modeling of the functions $h_k(s_k, \bar{a}_{k-1})$ while maintaining simple parametric models for the contrast functions $C_k(s_k, \bar{a}_{k-1})$. Thus, the resulting decision rule, which depends only on the contrast function, remains interpretable, while the model for the response is allowed to be nonlinear. This is also appealing in that it may be reasonable to expect, based on the underlying science, that the relationship between patient history and outcome is complex while the optimal rule for treatment assignment is dependent, in a simple fashion, on a small number of variables. The flexibility allowed by a semi-parametric model also has its drawbacks. Techniques for formal model building, critique, and diagnosis are well understood for linear models but much less so for semi-parametric models. Consequently, *Q*-learning based on building a series of linear models may be more appealing to an analyst interested in formal diagnostics.

A-learning may have certain advantages for making inference under the null hypothesis of no effect of any treatment regime in $\mathscr{D}$ on outcome. For example, in a SMART, the propensities are specified by design, and under the null, the contrast functions are identically zero and hence correctly specified. Thus, *A*-learning will yield consistent estimators for the parameters defining the contrast function. See Robins (2004) and the references in Section 8.

6. SIMULATION STUDIES

We examine the finite sample performance of *Q*- and *A*-learning on a suite of simple test examples via Monte Carlo simulation. We emphasize that the methods are straightforward to implement in more complex settings than those here. To illustrate trade-offs between the methods, we begin with correctly specified models and systematically introduce misspecification of the *Q*-function, the propensity model, and both. We focus here on situations where the contrast function is correctly specified to gain insight into impact of other model components. Scenarios with a misspecified contrast model can be constructed to include or exclude the target d^{opt}, precluding generalizable conclusions. See Section A.9 of the supplemental article [Schulte et al. (2012)], Zhang et al. (2012a,b), and Zhang et al. (2013) for simulations involving misspecified contrast functions and Robins (2004, Section 9) for discussion.

In all scenarios, 10,000 Monte Carlo replications were used, and, for each generated data set, $\hat{d}_Q^{\text{opt}}$ and $\hat{d}_A^{\text{opt}}$ in (28) and (32) were obtained using the *Q*- and *A*-learning procedures in Sections 5.1 and 5.2. For simplicity, we consider one ($K = 1$) and two ($K = 2$) decision problems, where, at each decision point, there are two treatment options coded as 0 and 1. In all cases, we used *Q*-functions of the form $Q_1(s_1, a_1;\xi_1) = h_1(s_1; \beta_1) + a_1C_1(s_1; \psi_1)$ and $Q_2(s_2^-, \bar{a}_2; \xi_2) = h_2(s_2^-, \bar{a}_1; \beta_2) + a_2,C_2(s_2^-, a_1; \psi_2)$ to represent both true and assumed working models. With the contrast functions correctly specified, ψ_k, $k = 1,2$, dictate the optimal regime. Thus, as one measure of performance, we focus on relative efficiency of the estimators of components of ψ_k as reflected by the ratio of Monte Carlo mean squared errors (MSEs) given by MSE of *A*-learning/MSE of *Q*-learning, so that values greater than 1 favor *Q*-learning. Recognizing that $E(Y^*(d^{\text{opt}}))$ is the benchmark achievable outcome on average, as a second measure, we consider the extent to which the estimated regimes $\hat{d}_Q^{\text{opt}}$ and $\hat{d}_A^{\text{opt}}$ achieve $E(Y^*(d^{\text{opt}}))$ if followed by the population. Specifically, for regime d indexed by ψ_1 ($K = 1$) or $(\psi_1^T, \psi_2^T)^T$ ($K=2$), let $H(d) = E\{Y^*(d)\}$, a function of these parameters. Then $H(d^{\text{opt}}) = E\{Y^*(d^{\text{opt}})\}$ is this function evaluated at the true parameter values, and $H(d^{\hat{\text{opt}}})$ is this function evaluated the estimated parameter values for a given data set, where $d^{\hat{\text{opt}}}$ is $\hat{d}_Q^{\text{opt}}$ or $\hat{d}_A^{\text{opt}}$. Define $R(d^{\hat{\text{opt}}}) = E\{H(d^{\hat{\text{opt}}})\}/H(d^{\text{opt}})$, where the expectation in the numerator is with respect to the distribution of the estimated parameters in $d^{\hat{\text{opt}}}$. We refer to $R(d^{\hat{\text{opt}}})$ as the v-efficiency of $d^{\hat{\text{opt}}}$, as it reflects the extent to which $d^{\hat{\text{opt}}}$ achieves the "value" of the true optimal regime. In Section A.6 of the supplemental article [Schulte et al. (2012); we discuss calculation of $R(d^{\hat{\text{opt}}})$.

6.1 One Decision Point

In this and the next section, $n = 200$. Here, the observed data are (S_{1i},A_{1i},Y_i), $i = 1,\ldots, n$. With $\text{expit}(x) = e^x/(1 + e^x)$, we used the class of generative models

$$\begin{aligned} S_1 \sim \text{Normal}(0,1), A_1 \,\big|\, S_1 = s_1 \sim \text{Bernoulli}\{\text{expit}(\phi_{10}^0 + \phi_{11}^0 s_1 + \phi_{12}^0 s_1^2)\}, \\ Y \,\big|\, S_1 = s_1, A_1 = a_1 \sim \text{Normal}\{\beta_{10}^0 + \beta_{11}^0 s_1 + \beta_{12}^0 s_1^2 + a_1(\psi_{10}^0 + \psi_{11}^0 s_1), 9\}, \end{aligned} \tag{34}$$

indexed by $\phi^0 = (\phi_{10}^0, \phi_{11}^0, \phi_{12}^0)^T, \beta^0 = (\beta_{10}^0, \beta_{11}^0, \beta_{12}^0)^T, \psi^0 = (\psi_{10}^0, \psi_{11}^0)^T$, so that $d^{\text{opt}} = d_1^{\text{opt}}, d_1^{\text{opt}}(s_1) = I(\psi_{10}^0 + \psi_{11}^0 s_1 > 0)$. For *A*-learning, we assumed models $h_1(s_1; \beta_1) = \beta_{10} + \beta_{11}s_1$, $C_1(s_1;\psi_1) = \psi_{10} + \psi_{11}s_1$, and $\pi_1(s_1;\varphi_1) = \text{expit}(\varphi_{10} + \varphi_{11}s_1)$, and for *Q*-learning we used $Q_1(s_1, a_1;\xi_1) = h_1(s_1;\beta_1) + a_1C_1(s_1;\psi_1)$. These models involve correctly specified contrast functions and are nested within the true models, with $h_1(s_1; \beta_1)$, and hence the *Q*-function, correctly specified when $\beta_{12}^0 = 0$. The propensity model $\pi_1(s_1; \varphi_1)$ is correctly specified when $\phi_{12}^0 = 0$. To study the effects of misspecification, we varied β_{12}^0 and ϕ_{12}^0 while keeping the others fixed, considering parameter settings of the form $\phi^0 = (0, -2, \phi_{12}^0)^T, \beta^0 = (1, 1, \beta_{12}^0)^T, \psi^0 = (1, 0.5)^T$.

Correctly specified models—As noted in Section 5.3, when all working models are correctly specified, *Q*-learning is more efficient than *A*-learning, which for (34) occurs when

$\beta_{12}^0 = \phi_{12}^0 = 0$. Here, the efficiency of *Q*-learning relative to *A*-learning is 1.06 for estimating ψ_{10}^0 and 2.74 for ψ_{11}^0. Thus, *Q*-learning is a modest 6% more efficient in estimating ψ_{10}^0 but a dramatic 174% more efficient in estimating ψ_{11}^0. Interestingly, the v-efficiency of the decision rules produced by the methods is similar, with $R(\hat{d}_Q^{\text{opt}}) = 0.97 \text{ and } R(\hat{d}_A^{\text{opt}}) = 0.95$, so that inefficiency in estimation of ψ_1 via *A*-learning does not translate into a regime of poorer quality than that found by *Q*-learning.

Misspecified propensity model—Under (34), this situation corresponds to $\beta_{12}^0 = 0$ and nonzero ϕ_{12}^0. An appeal of *A*-learning is the double robustness property noted in Section 5.2, which implies that ψ_1 is estimated consistently when the propensity model is misspecified provided that the *Q*-function is correct. In contrast, *Q*-learning does not depend on the propensity model, so its performance is unaffected. Figure 1 shows the relative efficiency in estimating $\psi_{10}^0 \text{ and } \psi_{11}^0$ and the efficiency of $\hat{d}_Q^{\text{opt}} \text{ and } \hat{d}_A^{\text{opt}}$ as ϕ_{12}^0 varies from −1 to 1. The leftmost panel shows that there is minimal efficiency gain by using *Q*-learning instead of *A*-learning in estimation of ψ_{10}^0. From the center panel, *Q*-learning yields substantial gains over *A*-learning for estimating ψ_{11}^0. Interestingly, the gain is largest when $\phi_{12}^0 = 0$, which corresponds to a correctly specified propensity model. Letting $\pi^0(s_1; \phi_1^0)$ be the true propensity, $\phi_1^0 = (\phi_{10}^0, \phi_{11}^0, \phi_{12}^0)^T$, a possible explanation for this seemingly contradictory result in this scenario is that, as $\left|\phi_{12}^0\right|$ gets larger, $\text{logit}\{\pi^0(S_1; \phi_1^0\} = \phi_{10}^0 + \phi_{11}^0 s_1 + \phi_{12}^0 s_1^2$ becomes more profoundly quadratic. Consequently, the estimator for φ_{11} in the posited model $\pi_1(s_1; \varphi_1) = \text{expit}(\varphi_{10} + \varphi_{11} s_1)$ approaches zero, so that the estimated posited propensity approaches a constant. Because *Q*- and *A*-learning are algebraically equivalent under constant propensity here, substituting an estimated propensity that is nearly constant leads to an estimator very similar to that from *Q*-learning. Consequently, empirical efficiency gains decrease as $\left|\phi_{12}^0\right| \to \infty$. The right panel of Figure 1 shows a small gain in v-efficiency of $\hat{d}_Q^{\text{opt}}$ over $\hat{d}_A^{\text{opt}}$; both achieve good performance.

See Section A.9 of the supplemental article [Schulte et al. (2012)] for evidence demonstrating this behavior of the propensity score and for further summaries reflecting the relative efficiency of the estimated regimes in all scenarios in this and the next section.

Misspecified Q-function—This scenario examines the second aspect of *A*-learning's double-robustness, characterized in (34) by $\phi_{12}^0 = 0$ and nonzero β_{12}^0. Here, *A*-learning leads to consistent estimation while *Q*-learning need not. The left panel of Figure 2 shows that the gain in efficiency using *A*-learning is minimal in estimating ψ_{10}^0. The center panel illustrates the bias-variance trade-off associated with *Q*- versus *A*-learning. For β_{12}^0 far from zero, bias in the misspecified *Q*-function dominates the variance, and *A*-learning enjoys smaller MSE while, for small values of β_{12}^0, variance dominates bias, and *Q*-learning is more efficient. The

right panel shows that large bias in the *Q*-function can lead to meaningful loss (~10%) in v-efficiency of $\hat{d}_Q^{\text{opt}}$ relative to $\hat{d}_A^{\text{opt}}$.

Both propensity model and Q-function misspecified—In our class of generative models (34), this corresponds to nonzero values of both $\beta_{12}^0 \text{and} \phi_{12}^0$. Rather than vary both values, (e.g., over a grid), we varied one and chose the other so that it is "equivalently misspecified." In particular, for a given value of ϕ_{12}^0, we selected $\beta_{12}^0 = \beta_{12}^0(\phi_{12}^0)$ so that the *t*-statistic associated with testing $\phi_{12}^0 = 0$ in the logistic propensity model and the *t*-statistic associated with testing $\beta_{12}^0 = 0$ in the linear *Q*-function would be approximately equal in distribution. Consequently, across data sets, an analyst would be equally likely to detect either form of misspecification. Details of this construction are given in Section A.7 of the supplemental article [Schulte et al. (2012)].

As in the preceding scenario, Figure 3 illustrates the bias-variance trade-off associated with *Q*- and *A*-learning. For large misspecification, *A*-learning provides a large enough reduction in bias to yield lower MSE; for small misspecification, *Q*-learning incurs some bias but reduces the variance enough to yield lower MSE. From the right panel of the figure, bias seems to translate into a larger loss in v-efficiency of the estimators of d^{opt} than variance.

6.2 Two Decision Points

For $K = 2$, the observed data available to estimate $d^{\text{opt}} = (d_1^{\text{opt}}, d_2^{\text{opt}})$ are $(S_{1i}, A_{1i}, S_{2i}, A_{2i}, Y_i)$, $i = 1, \ldots, n$. For these scenarios, we used a class of true generative data models that differs from those of Chakraborty et al. (2010), Song et al. (2010), and Laber et al. (2010) only in that S_2 is continuous instead of binary; as the model at the first stage is saturated, this allows correct specification of the *Q*-function at decision 1. The generative model is

$$
\begin{aligned}
&S_1 \sim \text{Bernoulli}(0.5), A_1 \,|\, S_1 = s_1 \sim \text{Bernoulli}\{\text{expit}(\phi_{10}^0 + \phi_{11}^0 s_1)\},\\
&S_2 | S_1 = s_1, A_1 = a_1 \sim \text{Normal}(\delta_{10}^0 + \delta_{11}^0 s_1 + \delta_{12}^0 a_1 + \delta_{13}^0 s_1 a_1, 2),\\
&A_2 | S_1 = s_1, S_2 = s_2, A_1 = a_1 \sim \text{Bernoulli}\{\text{expit}(\phi_{20}^0 + \phi_{21}^0 s_1 + \phi_{22}^0 a_1 + \phi_{23}^0 s_2 + \phi_{24}^0 a_1 s_2 + \phi_{25}^0 s_2^2)\},\\
&Y S_1 = s_1, S_2 = s_2, A_1 = a_1, A_2 = a_2 \sim \text{Normal}\{m(s_1, s_2, a_1, a_2), 10\},\\
&m(s_1, s_2, a_1, a_2) = \beta_{20}^0 + \beta_{21}^0 s_1 + \beta_{22}^0 a_1 + \beta_{23}^0 s_1 a_1 + \beta_{24}^0 s_2 + \beta_{25}^0 s_2^2 + a_2(\psi_{20}^0 + \psi_{21}^0 a_1 + \psi_{22}^0 s_2).
\end{aligned}
$$

The model is indexed by

$$
\begin{aligned}
\phi_1^0 &= (\phi_{10}^0, \phi_{11}^0)^T, \delta_1^0\\
&= (\delta_{10}^0, \delta_{11}^0, \delta_{12}^0, \delta_{13}^0)^T, \phi_2^0\\
&= (\phi_{20}^0, \phi_{21}^0, \phi_{22}^0, \phi_{23}^0, \phi_{24}^0, \phi_{25}^0)^T, \beta_2^0\\
&= (\beta_{20}^0, \beta_{21}^0, \beta_{22}^0, \beta_{23}^0, \beta_{24}^0, \beta_{25}^0)^T, \text{and} \psi_2^0\\
&= (\psi_{20}^0, \psi_{21}^0, \psi_{22}^0)^T
\end{aligned}
$$

, with true $h_2^0(s_1, s_2, a_1) = \beta_{20}^0 + \beta_{21}^0 s_1 + \beta_{22}^0 a_1 + \beta_{23}^0 s_1 a_1 + \beta_{24}^0 s_2 + \beta_{25}^0 s_2^2$ and contrast function $C_2^0(s_1, s_2, a_1) = \psi_{20}^0 + \psi_{21}^0 a_1 + \psi_{22}^0 s_2$, say. Because A_1 and S_1 are binary, the true functions $h_1^0(s_1) = \beta_{10}^0 + \beta_{11}^0 s_1 \text{and} C_1^0(s_1) = \psi_{10}^0 + \psi_{11}^0 s_1$ are linear in $s_1; \beta_{10}^0, \beta_{11}^0, \psi_{10}^0, \text{and} \psi_{10}^0$ are derived

in terms of parameters indexing the generative model in Section A.8 of the supplemental article [Schulte et al. (2012)]. Thus, the true optimal regime has

$$d_1^{\text{opt}}(s_1)=I(\psi_{10}^0+\psi_{11}^0 s_1>0)\text{and}d_2^{\text{opt}}(s_1,s_2,a_1)=I(\psi_{20}^0+\psi_{21}^0 a_1+\psi_{22}^0 s_2>0).$$

We assumed working models for *A*-learning of the form $h_1(s_1;\beta_1) = \beta_{10} + \beta_{11}s_1$, $C_1(s_1;\psi_1) = \psi_{10} + \psi_{11}s_1$, $\pi_1(s_1;\varphi_1) = \text{expit}(\varphi_{10} + \varphi_{11}s_1)$, $h_2(s_1, s_2, a_1;\beta_2) = \beta_{20} + \beta_{21}s_1 + \beta_{22}a_1 + \beta_{23}s_1a_1 + \beta_{24}s_2$, $C_2(s_1,s_2, a_1;\psi_2) = \psi_{20} + \psi_{21}a_1 + \psi_{22}s_2$, and $\pi_2(s_1,s_2, a_1;\varphi_2) = \text{expit}(\varphi_{20} + \varphi_{21}s_1 + \varphi_{22}a_1 + \varphi_{23}s_2 + \varphi_{24}a_1s_2)$; and, similarly, *Q*-functions $Q_1(s_1, a_1;\xi_1) = h_1(s_1;\beta_1) + a_1C_1(s_1;\psi_1)$ and $Q_2(s_1,s_2, a_1, a_2;\xi_2) = h_2(s_1,s_2, a_1;\beta_2) + a_2C_2(s_1,s_2, a_1;\psi_2)$ for *Q*-learning, so that the contrast functions are correctly specified in each case. Comparison of the working and generative models shows that the former are correctly specified when $\phi_{25}^0\text{and}\beta_{25}^0$ are both zero and are misspecified otherwise. Thus, we systematically varied these parameters to study the effects of misspecification, leaving all other parameter values fixed, taking

$$\begin{aligned}\phi_1^0&=(0.3,-0.5)^T,\delta_1^0\\&=(0,0.5,-0.75,0.25)^T,\phi_2^0\\&=(0,0.5,0.1,-1,-0.1,\phi_{25}^0)^T,\beta_2^0\\&=(3,0,0.1,-0.5,-0.5,\beta_{25}^0)^T,\text{and}\psi_2^0\\&=(1,0.25,0.5)^T\end{aligned}.$$

Correctly specified models—This occurs when $\phi_{25}^0=\beta_{25}^0=0$. As discussed previously, *Q*-learning is efficient when the models are correctly specified. Efficiencies of *Q*- learning relative to *A*-learning for estimating $\psi_{10}^0,\psi_{11}^0,\psi_{20}^0,\psi_{21}^0,\text{and}\psi_{22}^0$ are 1.07, 1.03, 1.19, 1.44, and 1.98, respectively. Hence, *Q*-learning is markedly more efficient in estimating the second stage parameters but only modestly so for first stage parameters. More efficient estimators of the parameters do not translate into greater v-efficiency of the estimated regimes in this scenario, as $R(\hat{d}_Q^{\text{opt}})=0.96\text{and}R(\hat{d}_A^{\text{opt}})=0.96$.

Misspecified propensity model—The propensity model at the second stage is misspecified when ϕ_{25}^0 is nonzero. To isolate the effects of such misspecification, we set $\beta_{25}^0=0$ and varied ϕ_{25}^0 between −1 and 1. From Figure 4, *Q*-learning is more efficient than *A*-learning for estimation of all parameters in ψ_1 and ψ_2, and, as in the one decision case, the efficiency gain is largest when $\phi_{25}^0=0$, corresponding to a correctly specified propensity model. From the lower right panel, there appears to be little difference in v-efficiency of $\hat{d}_Q^{\text{opt}}\text{and}\hat{d}_A^{\text{opt}}$.

Misspecified Q-function—Under our class of generative models, the *Q*-function is misspecified when β_{25}^0 is nonzero. We set $\phi_{25}^0=0$ to focus on the effects of such misspecification. Figure 5 shows that, for the first stage parameters $\psi_{10}^0\text{and}\psi_{11}^0$, there is little difference in efficiency between *Q*- and *A*-learning. The upper panels illustrate varying degrees of the bias-variance trade-off between the methods. In particular, in estimating ψ_{22}^0, a small amount of misspecification leads to significant bias, and hence *A*-learning produces

a much more accurate estimator, while, for ψ_{20}^0 the bias-variance trade-off is present but attenuated, and there is little difference between *Q*- and *A*-learning. In estimation of ψ_{21}^0, variance appears to dominate bias, and *Q*-learning is preferred for the chosen range of β_{25}^0 values. From the lower right panel, relative efficiency for estimating ψ_{22}^0 weakly tracks the relative efficiencies of the estimated regimes $\hat{d}_Q^{\text{opt}}\text{and}\hat{d}_A^{\text{opt}}$, suggesting that the efficiency gain for *A*-learning in estimating ψ_{22}^0 leads to improved estimation of d^{opt}.

Both the propensity model and Q-function misspecified—This scenario corresponds to nonzero values of $\beta_{25}^0\text{and}\phi_{25}^0$. Analogous to the one decision case, we chose pairs $(\beta_{25}^0, \phi_{25}^0)$ that are "equivalently misspecified;" see Section A.7 of the supplemental article [Schulte et al. (2012)]. From Figure 6, there is no general trend in efficiency of estimation across parameters that might recommend one method over the other. Furthermore, from the lower right panel, there is little difference in v-efficiency of the estimated regimes. One should not expect to draw broad conclusions, as neither *Q*- nor *A*-learning need be consistent here. Interestingly, despite misspecification of both models, $\hat{d}_Q^{\text{opt}}\text{and}\hat{d}_A^{\text{opt}}$ still enjoy high v-efficiency in this scenario.

6.3 Moodie, Richardson, and Stephens Scenario

The foregoing simulation scenarios deliberately involve simple models for the *Q*-functions in order to allow straightforward interpretation. To investigate the relative performance of the methods in a more challenging setting, we generated data from a scenario similar to that in Moodie et al. (2007) in which the true contrast functions are simple yet the *Q*-functions are complex.

The data generating process used mimics a study in which HIV-infected patients are randomized to receive antiretroviral therapy (coded as 1) or not (coded as 0) at baseline and again at six months, where the randomization probabilities depend on baseline and six month CD4 counts. Specifically, we generated baseline CD4 count S_1 ~ Normal(450, 150^2), and baseline treatment A_1 was then assigned according to $A_1|S_1=s_1\sim\text{Bernoulli}\{\text{expit}(\phi_{10}^0+\phi_{11}^0 s_1)\}$. We generated six month CD4 count S_2, distributed conditional on $S_1 = s_1$, $A_1 = a_1$ as Normal($1.25s_1$,60^2). Treatment A_2 was then generated according to $A_2|S_1=s_1, A_1=a_1, S_2=s_2\sim\text{Bernoulli}\{\text{expit}(\phi_{20}^0+\phi_{21}^0 s_2)\}$. In contrast to the scenario in Moodie et al. (2007), this allows all possible treatment combinations. The outcome *Y* is CD4 count at one year; following Moodie et al. (2007), *Y* was generated as $Y=Y^{\text{opt}}-\mu_1^0(S_1,A_1)-\mu_2^0(S_1,S_2,A_1,A_2)$, where $Y^{\text{opt}}|S_1 = s_1, A_1 = a_1, S_2 = s_2, A_2 = a_2$ ~ Normal(400 + $1.6s_1$, 60^2). Here, $\mu_1^0(S_1,A_1)\text{and}\mu_2^0(S_1,S_2,A_1,A_2)$ are the true advantage (regret) functions; we took $C_1^0(s_1)=\psi_{10}^0+\psi_{11}^0 s_1\text{and}C_2^0(s_1,s_2,a_1)=\psi_{20}^0+\psi_{21}^0 s_2$ to be the true contrast functions, so that, from Section 5.2,

$$\mu_1^0(S_1,A_1)=(\psi_{10}^0+\psi_{11}^0 S_1)\{I(\psi_{10}^0+\psi_{11}^0 S_1>0)-A_1\}, \quad (35)$$

$$\mu_2^0(S_1, S_2, A_1, A_2) = (\psi_{20}^0 + \psi_{21}^0 S_2)\{I(\psi_{20}^0 + \psi_{21}^0 S_2 > 0) - A_2\}. \quad (36)$$

It follows that the optimal treatment regime $d^{\text{opt}} = (d_1^{\text{opt}}, d_2^{\text{opt}})$ has $d_1^{\text{opt}}(s_1) = I(\psi_{10}^0 + \psi_{11}^0 s_1 > 0)$ and $d_2^{\text{opt}}(\bar{s}_2, a_1) = I(\psi_{20}^0 + \psi_{21}^0 s_1 > 0)$. While the true contrast functions are linear in $\psi_k^0, k = 1, 2$, the true implied $h_1^0(s_1) \text{and} h_2^0(s_1, a_1, s_2)$ are nonsmooth and possibly complex.

Following Moodie et al. (2007), for *A*-learning, we assumed working models $h_1(s_1;\beta_1) = \beta_{10} + \beta_{11}s_1$, $C_1(s_1;\psi_1) = \psi_{10} + \psi_{11}s_1$, $h_2(s_1, s_2, a_1; \beta_2) = \beta_{20} + \beta_{21}s_1 + \beta_{22}a_1 + \beta_{23}s_1a_1 + \beta_{24}s_2$, and $C_2(s_1, s_2, a_1; \psi_2) = \psi_{20} + \psi_{21}s_2$, and propensity models $\pi_1(s_1;\varphi_1) = \text{expit}(\varphi_{10} + \varphi_{11}s_1)$ and $\pi_2(s_1, s_2, a_1; \varphi_2) = \text{expit}(\varphi_{20} + \varphi_{21}s_2)$. For *Q*-learning, we analogously assumed *Q*-functions $Q_1(s_1, a_1; \xi_1) = h_1(s_1; \beta_1) + a_1C_1(s_1; \psi_1)$ and $Q_2(s_1, s_2, a_1, a_2; \xi_2) = h_2(s_1, s_1, a_1; \beta_2) + a_2C_2(s_1, s_2, a_1;\psi_2)$. Note that the contrast functions in each case are correctly specified, as are the propensity models; however, the *Q*-functions are misspecified, as the linear models $h_1(s_1; \beta_1)$ and $h_2(s_1, s_1, a_1; \beta_2)$ are poor approximations to the complex forms of the true $h_1^0(s_1) \text{and} h_2^0(s_1, s_2, a_1)$.

We report results for $n = 1000$ with

$$\begin{aligned}\phi_1^0 &= (\phi_{10}^0, \phi_{11}^0)^T \\ &= (2.0, -0.006)^T, \phi_2^0 \\ &= (\phi_{20}^0, \phi_{21}^0)^T \\ &= (0.8, -0.004)^T, \psi_1^0 \\ &= (\psi_{10}^0, \psi_{11}^0)^T \\ &= (250, -1.0)^T, \text{and} \psi_2^0 \\ &= (\psi_{20}^0, \psi_{21}^0)^T = (720, -2.0)^T\end{aligned}$$

in Table 1. Because the *Q*-functions are misspecified, the *Q*-learning estimators for $\psi_1^0 \text{and} \psi_2^0$ are biased, while those obtained via *A*-learning are consistent owing to the double robustness property. This leads to the dramatic relative inefficiency of *Q*-learning reflected by the MSE ratios. Under the assumed models, the estimated optimal regime for *Q*-learning dictates that, at baseline, therapy be given to patients with baseline CD4 count less than 199.7, while that estimated using *A*-learning gives treatment to those with baseline CD4 count less than 249.1, almost perfectly achieving the true optimal CD4 threshold of 250. Under the data generative process, using the baseline decision rule estimated via *Q*-learning may result in as many as 4.4% of patients who would receive therapy at baseline under the true optimal regime being assigned no treatment. Similarly, at the second decision, the estimated optimal regimes obtained by *Q*- and *A*-learning dictate that therapy be given to patients with six month CD4 count less than 320.2 and 360.1, respectively. Again, *A*-learning yields an estimated threshold almost identical to the optimal value of 360. Although that obtained via *Q*-learning is lower, 4.3% of patients who should receive therapy at six months would not if the estimated six month rule from *Q*-learning were followed by the population.

By Section A.6 of the supplemental article [Schulte et al. (2012)], $H(d^{\text{opt}}) = 1120$, whereas $\mathrm{E}\{H(\hat{d}_Q^{\text{opt}})\} \approx 1117.1$ (estimated standard error 1.3) and $\mathrm{E}\{H(\hat{d}_A^{\text{opt}})\} \approx 1119.9 (0.3)$, so that $R(\hat{d}_Q^{\text{opt}}) \text{and} R(\hat{d}_A^{\text{opt}})$ are virtually equal to one. Thus, although *Q*-learning yields poor estimation of parameters in the contrast functions, loss in v-efficiency of the estimated optimal regime is negligible. A possible explanation is as follows. For (35) and (36), some patients near the true treatment decision boundary would have $C_k^0(\overline{S}_k, \overline{A}_{k-1}), k=1, 2$, close to zero. Thus, even if a regime improperly assigns treatment to these patients, they would experience only a small loss in outcome and hence have little effect on the overall average. For other patients for whom the true contrast is not close to zero, improper assignment could result in considerable degradation of outcome. Because the proportion of patients receiving improper assignment is small in this scenario, the effect of these latter patients on the overall expected outcome is not substantial, leading to the relatively good expected outcome under the estimated *Q*-learning regime.

7. APPLICATION TO STAR*D

Sequenced Treatment Alternatives to Relieve Depression (STAR*D) was a randomized clinical trial enrolling 4041 patients designed to compare treatment options for patients with major depressive disorder. The trial involved four levels, where each level consisted of a 12 week period of treatment, with scheduled clinic visits at weeks 0, 2, 4, 6, 9, 12. Severity of depression at any visit was assessed using clinician-rated and self-reported versions of the Quick Inventory of Depressive Symptomatology (QIDS) score (Rush et al., 2003), for which higher values correspond to higher severity. At the end of each level, patients deemed to have an adequate clinical response to that level's treatment did not move on to future levels, where adequate response was defined by 12-week clinician-rated QIDS score ≤ 5 (remission) or showing a 50% or greater decrease from the baseline score at the beginning of level 1 (successful reduction). During level 1, all patients were treated with citalopram. Patients continuing to level 2 due to inadequate response, conferring with their physicians, expressed preference to (i) switch or (ii) augment citalopram and within that preference were randomized to one of several options: (i) switch: sertraline, bupropion, venlafaxine, or cognitive therapy, or (ii) augment: citalopram plus one of either bupropion, buspirone, or cognitive therapy. Patients randomized to cognitive therapy (alone or augmented with citalopram) were eligible, in the case of inadequate response, to move to a supplementary level 2A and be randomized to switch to bupropion or venlafaxine. All patients without adequate response at level 2 (or 2A) continued to level 3 and, depending on preference to (i) switch or (ii) augment, were randomized within that preference to (i) switch: mirtazepine or nortriptyline or (ii) augment with either: lithium or triiodothyronine. Patients without adequate response continued to level 4, requiring a switch to tranylcypromine or mirtazepine combined with venlafaxine (determined by preference). Thus, although the study involved randomization, it is observational with respect to the treatment options switch or augment. For a complete description see Rush et al. (2004); see Section A.10 of the supplemental article [Schulte et al. (2012)] for a schematic of the design.

To demonstrate formulation of this problem within the framework of Sections 2 and 3, we take level 2A to be part of level 2 and consider only levels 2 and 3, calling them stages

(decision points) 1 and 2, respectively ($K = 2$). Some patients in stage 1 without adequate response dropped out of the study without continuing to stage 2. Hence, we analyze complete case data, excluding dropouts, from 795 patients entering stage 1; 330 of these subsequently continued to stage 2. Let A_k, $k = 1, 2$, be the treatment at stage k, taking values 0 (augment) or 1 (switch); both options are feasible for all eligible subjects. Let S_{10} denote baseline (study entry) QIDS score and S_{11} denote the most recent QIDS score at the beginning of stage 1, respectively, so that $S_1 = (S_{10}, S_{11})^T$ is information available immediately prior to the first decision. Similarly, let S_2 be the information available immediately prior to stage 2; here, S_2 is the most recent QIDS score at the end of stage 1/ beginning of stage 2. Finally, let T be QIDS score at the end of stage 2. Because some patients exhibited adequate response at the end of stage 1 and did not progress to stage 2, we define the outcome of interest to be $-S_2$ (negative QIDS score at the end of stage 1) for patients not moving to stage 2 and $-(S_2 + T)/2$ (average of negative QIDS scores at the end of stages 1 and 2) otherwise. Thus, writing $L_0 = \max(5, S_{10}/2)$, $Y = -S_2 I(S_2 \le L_0) - (S_2 + T) I(S_2 > L_0)/2$, the cumulative average negative QIDS score. Thus, this demonstrates the case where outcome is a function of accrued information over the sequence of decisions.

From (9), $Q_2(s_2^-, \bar{a}_2) = E(Y|S_2^- = s_2^-, \bar{A}_2 = \bar{a}_2) = -s_2\{I\{s_2 \le l_0) + I(s_2 > l_0)/2\} + E(-T|S_2^- = s_2^-, \bar{A}_2 = \bar{a}_2, S_2 > l_0) I(s_2 > l_0)/2$, so that $V_2(s_2^-, a_1) = -s_2 I(s_2 \le l_0) + \{-s_2 + U_2(s_2^-, a_1)\} I(s_2 > l_0)/2$, where $U_2(s_2^-, a_1) = \max_{a2} E(-T|S_2^- = s_2^-, \bar{A}_1 = \bar{a}_1, A_2 = a_2, S_2 > l_0)$. Thus, from (12),

$$Q_1(s_1, a_1) = E[-S_2 I(S_2 \le l_0) + \{-S_2 + U_2(\bar{s}_2, a_1)\} I(S_2 > l_0)/2 \big| S_1 = s_a, A_1 = a_1].$$

We describe implementation for Q-learning. At the second decision point, we must posit a model for $Q_2(s_2^-, \bar{a}_2)$. From the form of $Q_2(s_2^-, \bar{a}_2)$, we need only specify a model for $E(-T|S_2^- = s_2^-, \bar{A}_2 = \bar{a}_2, S_2 > l_0)$; given the form of the conditioning set, this may be carried out using only the data from patients moving to stage 2. Based on exploratory analysis, defining s_{22} to be the slope of QIDS score over stage 1 based on s_{11} and s_2, we took this model to be of the form $\beta_{20} + \beta_{21} s_2 + \beta_{22} s_{22} + \psi_{20} a_2$, so that the posited Q-function is

$$Q_2(\bar{s}_2, \bar{a}_2; \xi_2) = -s_2\{I(s_2 \le l_0) + I(s_2 > l_0)/2\} + I(s_2 > l_0)(\beta_{20} + \beta_{21} s_2 + \beta_{22} s_{22} + \psi_{20} a_2)/2, \quad (37)$$

$\xi_2 = (\beta_{20}, \beta_{21}, \beta_{22}, \psi_{20})^T$. Under (37), $V_2(s_2^-, a_1; \xi_2) = -s_2\{I(s_2 \le l_0) + I(s_2 > l_0)/2\} + I(s_2 > l_0)\{\beta_{20} + \beta_{21} s_2 + \beta_{22} s_{22} + \psi_{20} I(\psi_{20} > 0)\}/2$, and the "responses" $\tilde{V}_{2,i}$ for use in (27) may then be formed by substituting the estimate for ξ_2. Based on exploratory analysis, we took the posited Q-function at the first stage to be $Q_1(s_1, a_1; \xi_1) = \beta_{10} + \beta_{11} s_{11} + \beta_{12} s_{12} + a_1 (\psi_{10} + \psi_{11} s_{12})$, where s_{12} is the slope of QIDS score prior to stage 1 based on s_{10} and s_{11}; and $\xi_1 = (\beta_{10}, \beta_{11}, \beta_{12}, \psi_{10}, \psi_{11})^T$. For A-learning, we posited models for the functions $h_k(s_k^-, \bar{a}_{k-1})$ and $C_k(s_k^-, \bar{a}_{k-1})$, $k = 1,2$, in the obvious way analogous to those above, and we took the propensity models to be of the form $\pi_2(s_2^-, a_1; \varphi_2) = \text{expit}(\varphi_{20} + \varphi_{21} s_2 + \varphi_{22} s_{22} + \varphi_{23} a_1)$ and $\pi_1(s_1; \varphi_1) = \text{expit}(\varphi_{10} + \varphi_{11} s_{11} + \varphi_{12} s_{12})$. Section A.11 of the supplemental article [Schulte et al. (2012)] presents model diagnostics.

The results are given in Table 2. To describe implementation, we consider interactions significant based on a test at level $\alpha = 0.10$. At the first stage, Q-learning suggests a

treatment switch for those with QIDS slope prior to stage 1 greater than −1.09 (obtained by solving $1.11 + 1.02S_{12} = 0$); *A*-learning assigns a treatment switch for those with this QIDS slope greater than −1.66. At stage 2, the results suggest that all patients should switch and not augment their existing treatments.

8. DISCUSSION

We have provided a self-contained account of *Q*- and *A*-learning methods for estimating optimal dynamic treatment regimes, including a detailed discussion of the underlying statistical framework in which these methods may be formalized and of their relative merits. Our discussion of *A*-learning is limited to the case of two treatment options at each decision. Our simulation studies suggest that, while *A*-learning may be inefficient relative to *Q*-learning in estimating parameters that define the optimal regime when the *Q*-functions required for the latter are correctly specified, *A*-learning may offer robustness to such misspecification. Nonetheless, *Q*-learning may have practical advantages in that it involves modeling tasks familiar to most data analysts, allowing the use of standard diagnostic tools. On the other hand, *A*-learning may be preferred in settings where it is expected that the form of the decision rules defining the optimal regime is not overly complex. However, *A*-learning increases in complexity with more than two treatment options at each stage, which may limit its appeal. Interestingly, in the simulation scenarios we consider, inefficiency and bias in estimation of parameters defining the optimal regime does not necessarily translate into large degradation of average performance of the estimated regime for either method.

Although our simple simulation studies provide some insight into the relative merits of these methods, there remain many unresolved issues in estimation of optimal treatment regimes. Approaches to address the challenges of high-dimensional information and large numbers of decision points are required. Existing methods for model selection focusing on minimization of prediction error may not be best for developing models optimal for decision-making. When *K* is very large, the number of parameters in the models required for *Q*- and *A*-learning becomes unwieldy. The analyst may wish to postulate models in which parameters are shared across decision points; see Robins (2004), Robins et al. (2008), Orellana et al. (2010) and Chakraborty and Moodie (2012).

In our development, we have invoked a strong version of the sequential randomization assumption to simplify supporting arguments. Richardson and Robins (2013) allow identification of potential outcomes under possibly weaker assumptions via graphical representations. These authors also extend the notion of a dynamic treatment regime.

Formal inference procedures for evaluating the uncertainty associated with estimation of the optimal regime are challenging due to the nonsmooth nature of decision rules, which in turn leads to nonregularity of the parameter estimators; see Robins (2004), Chakraborty et al. (2010), Laber et al. (2010), Moodie and Richardson (2010), Song et al. (2010), and Laber and Murphy (2011).

We have discussed sequential decision-making in the context of personalized medicine, but many other applications exist where, at one or more times in an evolving process, an action

must be taken from among a set of plausible actions. Indeed, *Q*-learning was originally proposed in the computer science literature with these more general problems in mind; see Shortreed et al. (2010).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

This work was supported by NIH grants R37 AI031789, R01 CA051962, R01 CA085848, P01 CA142538, and T32 HL079896.

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.

Blatt, D.; Murphy, SA.; Zhu, J. Technical Report 04-63. The Methodology Center, Pennsylvania State University; 2004. A-learning for approximate planning.

Estimating optimal dynamic treatment regimes with shared decision rules across stages: An extension of Q-learning. 2012 Unpublished manuscript.

Chakraborty B, Murphy SA, Strecher V. Inference for non-regular parameters in optimal dynamic treatment regimes. Statistical Methods in Medical Research. 2010; 19:317–343. [PubMed: 19608604]

Craven, MW.; Shavlik, JW. In Advances in Neural Information Processing Systems. Vol. ume 8. Denver, CO: MIT Press; 1996. Extracting tree-structured representations of trained networks; p. 24-30.

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

Lavori PW, Dawson R. A design for testing clinical strategies: biased adaptive within-subject randomization. Journal of the Royal Statistical Society, Series A. 2000; 163:29–38.

Laber EB, Murphy SA. Adaptive confidence intervals for the test error in classification. J. Amer. Statist. Assoc. 2011; 106:904–913.

Laber EB, Qian M, Lizotte DJ, Murphy SA. Statistical inference in dynamic treatment regimes. 2010 Pre-print, arXiv:1006.5831v1.

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

Moodie EEM, Richardson TS. Estimating optimal dynamic regimes: correcting bias under the null. Scand. J. Statist. 2010; 37:126–146.

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. Stat. Med. 2005; 24:1455–1481. [PubMed: 15586395]

Murphy SA, Lynch KG, Oslin D, McKay JR, Ten Have T. Developing adaptive treatment strategies in substance abuse research. Drug Alcohol Depend. 2007a; 88S:S24–S30. [PubMed: 17056207]

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

Nahum-Shani I, Qian M, Almirall D, Pelham WE, Gnagy B, Fabiano G, Waxmonsky J, Yu J, Murphy SA. Q-Learning: A data analysis method for constructing adaptive interventions. Technical report. 2010

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, DOI: 10.2202/1557-4679.1200.

Richardson TS, Robins JM. Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. 2013 Available at http://www.csss.washington.edu/Papers/.

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. Correcting for non-compliance in randomized trials using structural nested mean models. Comm. Statist. - Theory Meth. 1994; 23:2379–2412.

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. Stat. Med. 2008; 27:4678–4721. [PubMed: 18646286]

Rosenblum M, van der Laan MJ. Using regression models to analyze randomized trials: Asymptotically valid hypothesis tests despite incorrectly specified models. Biometrics. 2009; 65:937–945. [PubMed: 19210739]

Rosthøj S, Fullwood C, Henderson R, Stewart S. Estimation of optimal dynamic anticoagulation regimes from observational data: A regret-based approach. Stat. Med. 2006; 25:4197–4215. [PubMed: 16981226]

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

Rush AJ, Fava M, Wisniewski SR, Lavori PW, Trivedi MH, Sackeim HA, Thase ME, Nierenberg AA, Quitkin FM, Kashner TM, Kupfer DJ, Rosenbaum JF, Alpert J, Stewart JW, McGrath PJ, Biggs MM, Shores-Wilson K, Lebowitz BD, Ritz L, Niederehe G. Sequenced Treatment Alternatives to Relieve Depression (STAR*D): rationale and design. Control. Clin. Trials. 2004; 25:119–142. [PubMed: 15061154]

Rush AJ, Trivedi MH, Ibrahim HM, Carmody TJ, Arnow B, Klein DN, Markowitz JC, Ninan PT, Kornstein S, Manber R, Thase ME, Kocsis JH, Keller MB. The 16-item quick inventory of depressive symptomatology (qids), clinician rating (qids-c), and self-report (qids-sr): a psychometric evaluation in patients with chronic major depression. Biological Psychiatry. 2003; 54:573–583. [PubMed: 12946886]

Schulte PJ, Tsiatis AA, Laber EB, Davidian M. Supplement to “Q- and A-learning Methods for Estimating Optimal Dynamic Treatment Regimes”. 2012

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. 2010; 11:109–136.

Song R, Wang W, Zeng D, Kosorok MR. Penalized q-learning for dynamic treatment regimes. 2010 Pre-Print, arXiv:1108.5338v1.

Thall PF, Millikan RE, Sung H. Evaluating multiple treatment courses in clinical trials. Stat. Med. 2000; 19:1011–1028. [PubMed: 10790677]

Thall PF, Sung H, Etsey E. Selecting therapeutic strategies based on efficacy and death in multicourse clinical trials. J Amer. Statist. Assoc. 2002; 97:29–39.

Thall PF, Wooten LH, Logothetis CJ, Millikan RE, Tannir NM. Bayesian and frequentist two-stage treatment strategies based on sequential failure times subject to interval censoring. Stat. Med. 2007; 26:4687–4702. [PubMed: 17427204]

van der Laan MJ, Petersen ML. Causal effect models for realistic individualized treatment and intention to treat rules. Int. J. Biostat. 2007; 3:3.

Watkins, CJGH. Ph.D. Thesis. Cambridge, UK: King's College; 1989. Learning from Delayed Rewards.

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

Zhang B, Tsiatis AA, Davidian M, Zhang M, Laber EB. Estimating optimal treatment regimes from a classification perspective. Stat. 2012; 1:103–114. [PubMed: 23645940]

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

Zhang B, Tsiatis AA, Laber EB, Davidian M. Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. Biometrika. 2013 in press. doi: 10.1093/biomet/ast014.

Zhao Y, Kosorok MR, Zeng D. Reinforcement learning design for cancer clinical trials. Stat. Med. 2009; 28:3294–3315. [PubMed: 19750510]

Zhao Y, Zeng D, Rush AJ, Kosorok MR. Estimating individualized treatment rules using outcome weighted learning. J Amer. Statist. Assoc. 2012; 107:1106–1118.

Zhao Y, Zeng D, Laber EB, Kosorok MR. New statistical learning methods for estimating optimal dynamic treatment regimes. 2013 Unpublished manuscript.

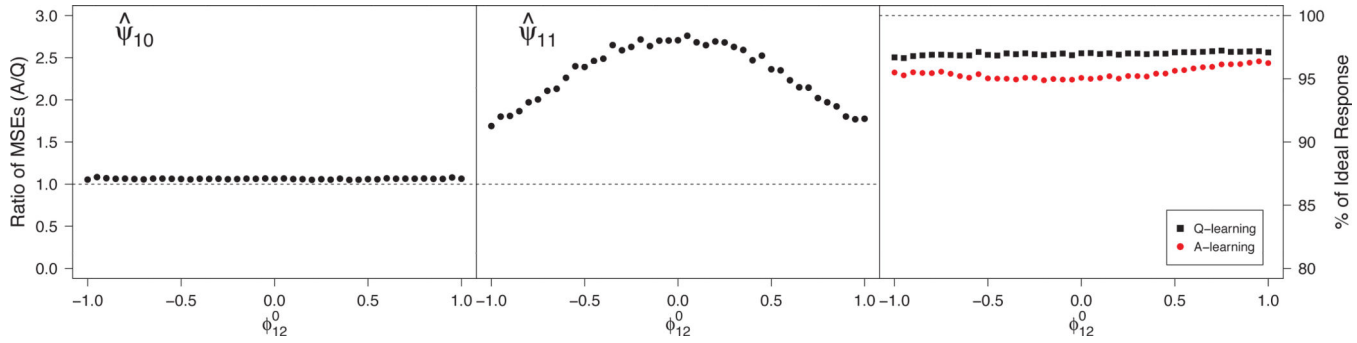

Fig 1.

Monte Carlo MSE ratios for estimators of components of ψ_1 (left and center panels) and efficiencies $R(\hat{d}_Q^{\text{opt}})$ and $R(\hat{d}_A^{\text{opt}})$ for estimating the true d^{opt} (right panel) under misspecification of the propensity model. MSE ratios > 1 favor Q-learning

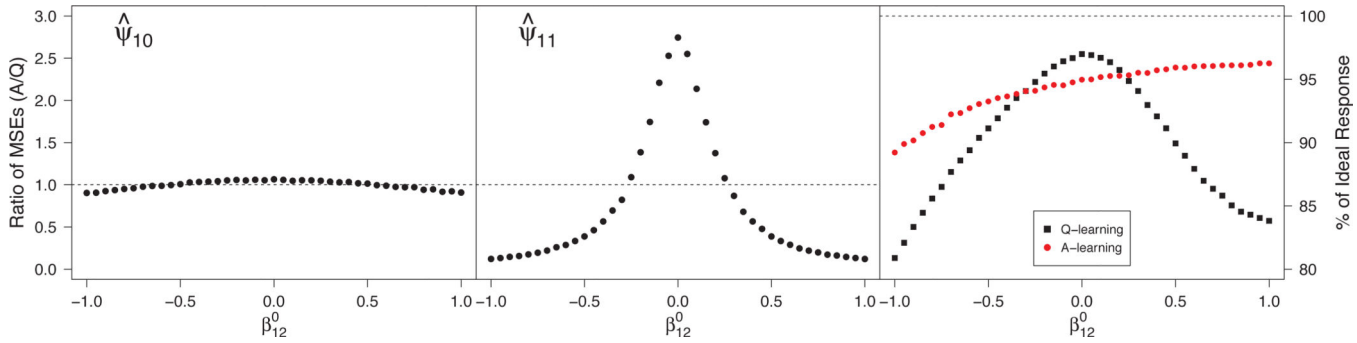

Fig 2.

Monte Carlo MSE ratios for estimators of components of ψ_1 (left and center panels) and efficiencies $R(\hat{d}_Q^{\text{opt}})$ and $R(\hat{d}_A^{\text{opt}})$ for estimating the true d^{opt} (right panel) under misspecification of the Q-function. MSE ratios > 1 favor Q-learning

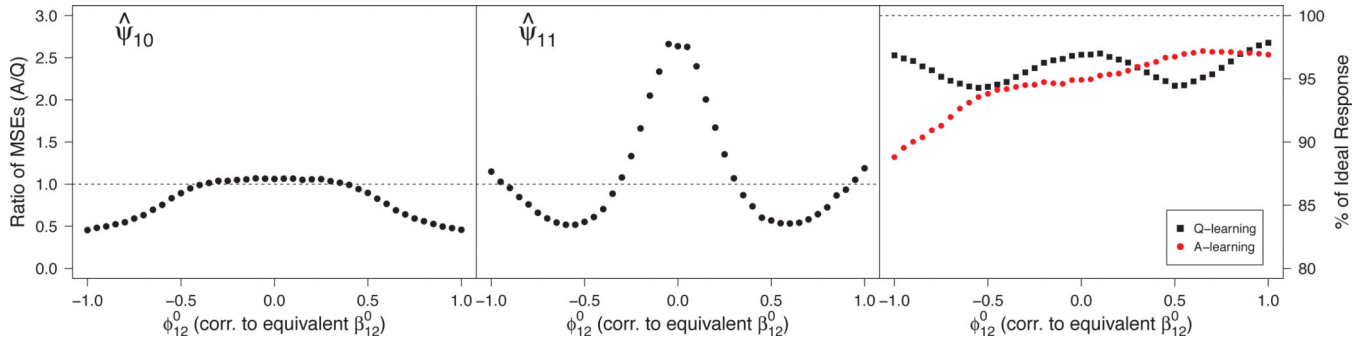

Fig 3.

Monte Carlo MSE ratios for estimators of components of ψ_1 (left and center panels) and efficiencies $R(\hat{d}_Q^{\text{opt}})$ and $R(\hat{d}_A^{\text{opt}})$ for estimating the true d^{opt} (right panel) under misspecification of both the propensity model and the Q-function. MSE ratios > 1 favor Q-learning

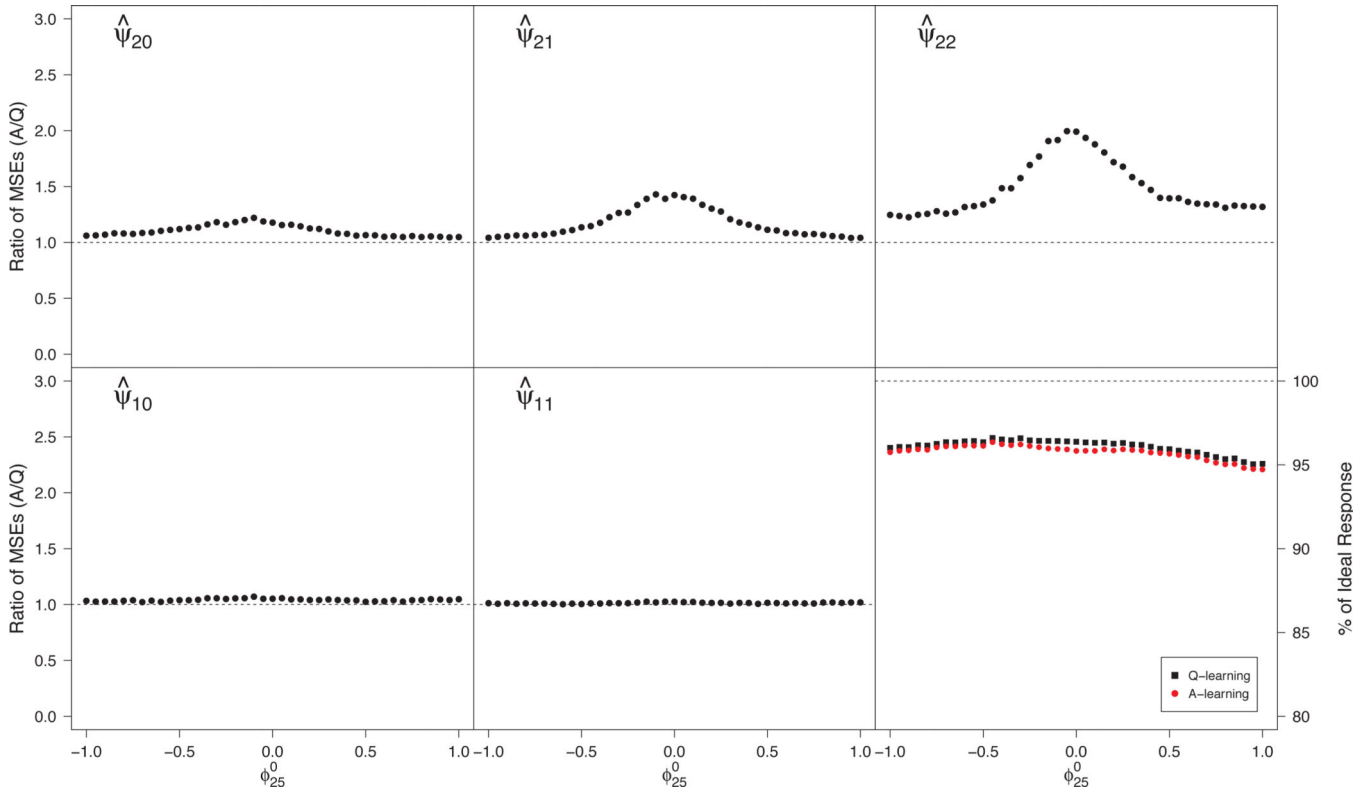

Fig 4.

Monte Carlo MSE ratios for estimators of components of ψ_2 and ψ_1 (upper row and lower row left and center panels) and efficiencies $R(\hat{d}_Q^{\text{opt}})\text{and}R(\hat{d}_A^{\text{opt}})$ for estimating the true d^{opt} (lower right panel) under misspecification of the propensity model. MSE ratios > 1 favor Q-learning

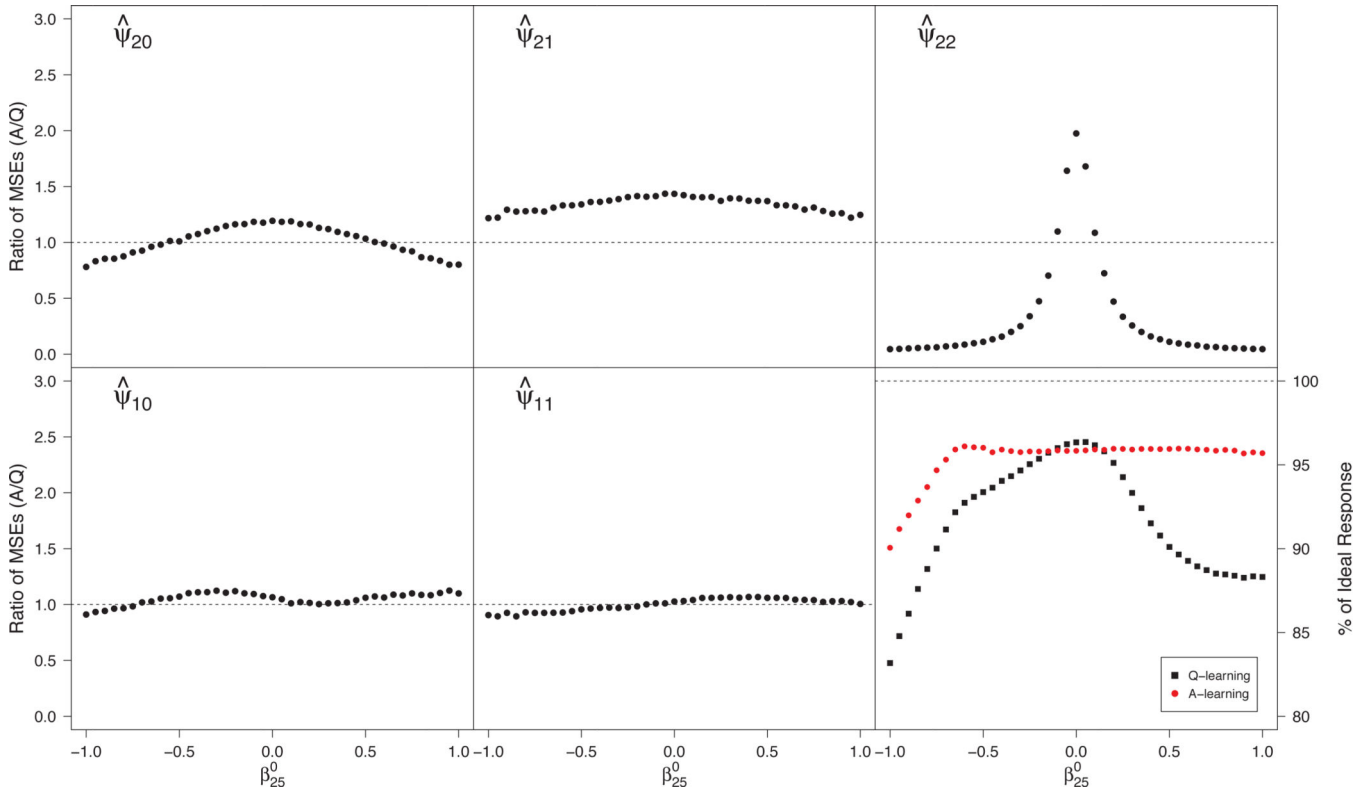

Fig 5.

Monte Carlo MSE ratios for estimators of components ofψ_2 and ψ_1 (upper row and lower row left and center panels) and efficiencies $R(\hat{d}_Q^{\text{opt}})\text{and}R(\hat{d}_A^{\text{opt}})$ for estimating the true d^{opt} (lower right panel) under misspecification of the Q-functions. MSE ratios > 1 favor Q-learning

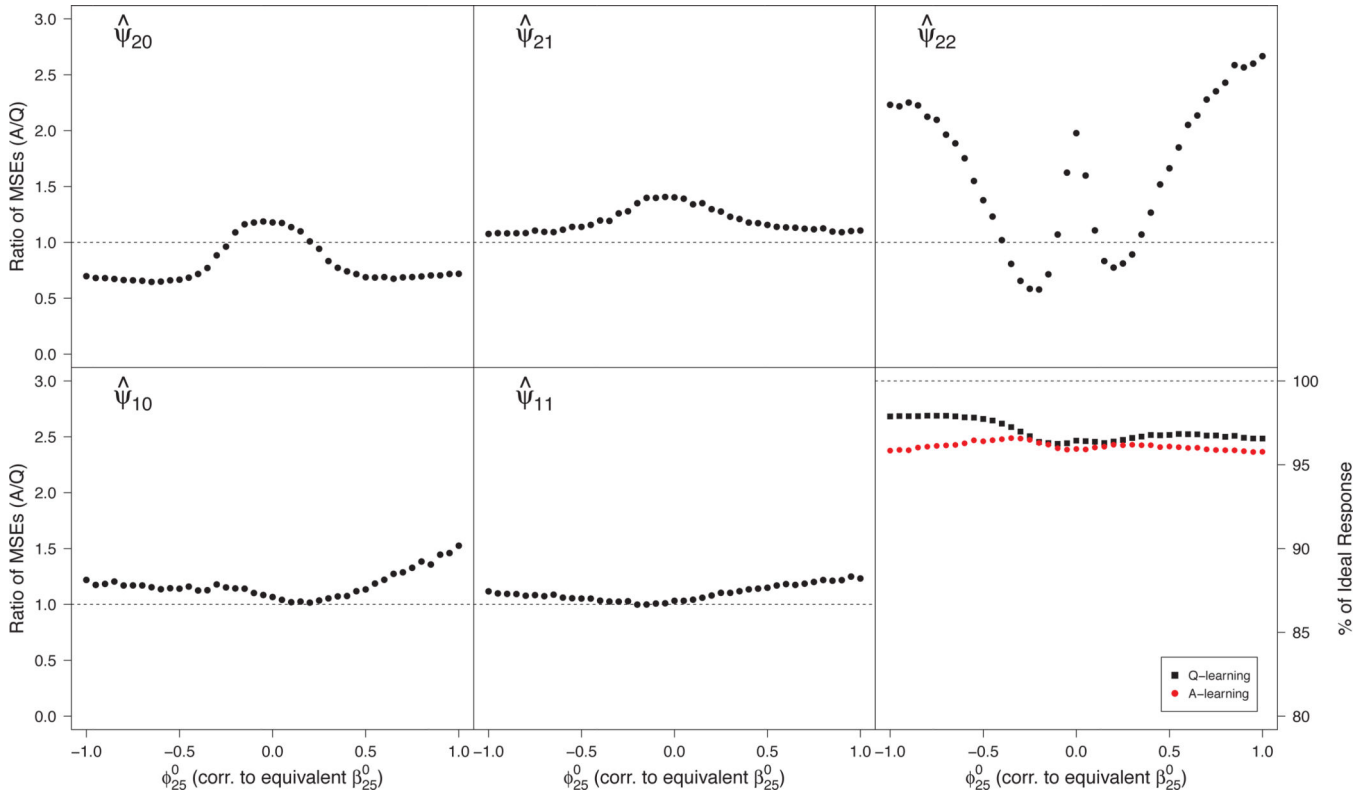

Fig 6.

Monte Carlo MSE ratios for estimators of components of ψ_2 and ψ_1 (upper row and lower row left and center panels) and efficiencies $R(\hat{d}_Q^{\text{opt}})\text{and}R(\hat{d}_A^{\text{opt}})$ for estimating the true d^{opt} (lower right panel) under misspecification of both the propensity models and Q-functions. MSE ratios > 1 favor Q-learning

Table 1

Monte Carlo average (standard deviation) of estimates obtained via Q- and A-learning and ratio of Monte Carlo MSE for the Moodie and Richardson scenario; MSE ratios > 1 favor Q-learning

Parameter (true value)	*Q*-learning	*A*-learning	MSE ratio
$\psi_{10}^0 = 250$	154.8 (23.2)	249.1 (18.7)	0.036
$\psi_{11}^0 = -1.0$	−0.775 (0.052)	−0.998 (0.041)	0.032
$\psi_{20}^0 = 720$	507.3 (49.2)	720.3 (48.4)	0.050
$\psi_{21}^0 = -2.0$	−1.584 (0.092)	−2.001 (0.085)	0.040

Table 2

STAR*D data analysis results

	Q-learning			A-learning		
Parameter	**Estimate**	**95% CI**	**p-value**	**Estimate**	**95% CI**	**p-value**
Stage 2						
β_{20}	−1.46	(−3.47 , 0.55)		−1.47	(−3.49 , 0.54)	
β_{21}	−0.75	(−0.88 , −0.61)	*	−0.75	(−0.88 , −0.61)	*
β_{22}	1.17	(0.52 , 1.81)	*	1.17	(0.52 , 1.81)	*
ψ_{20}	1.10	(0.02 , 2.19)	*	1.12	(0.03 , 2.22)	*
Stage 1						
β_{10}	−0.62	(−1.94 , 0.70)		−0.30	(−1.69 , 1.09)	
β_{11}	−0.54	(−0.62 , −0.45)	*	−0.55	(−0.64 , −0.46)	*
β_{12}	−0.08	(−0.60 , 0.45)		0.10	(−0.46 , 0.66)	
Ψ_{10}	1.11	(0.28 , 1.94)	*	0.73	(−0.18 , 1.65)	
ψ_{11}	1.02	(−0.08 , 2.11)	*	0.44	(−0.83 , 1.72)	

Asterisks indicate evidence at level of significance 0.05 (0.10) that the main effect (treatment contrast) parameter is non-zero