International Statistical Review

*International Statistical Review* (2023) doi: 10.1111/insr.12536

# Optimal Treatment Regimes: A Review and Empirical Comparison

**Zhen Li[1], Jie Chen[2], Eric Laber[3], Fang Liu[4] and Richard Baumgartner[4]**

[1]*Department of Statistics, North Carolina State University, Raleigh, 27607, NC, USA*
[2]*Department of Biometrics, Overland Pharmaceuticals, Dover, 19901, DE, USA*
[3]*Department of Statistical Science, Department of Biostatistics and Bioinformatics, Duke University, Durham, 27708, NC, USA*
[4]*Biostatistics and Research Decision Sciences, Merck & Co., Inc., Kenilworth, NJ 07033, USA*
*E-mail: zhenli861@gmail.com*

## Summary

**A treatment regime is a sequence of decision rules, one per decision point, that maps accumulated patient information to a recommended intervention. An optimal treatment regime maximises expected cumulative utility if applied to select interventions in a population of interest. As a treatment regime seeks to improve the quality of healthcare by individualising treatment, it can be viewed as an approach to formalising precision medicine. Increased interest and investment in precision medicine has led to a surge of methodological research focusing on estimation and evaluation of optimal treatment regimes from observational and/or randomised studies. These methods are becoming commonplace in biomedical research, although guidance about how to choose among existing methods in practice has been somewhat limited. The purpose of this review is to describe some of the most commonly used methods for estimation of an optimal treatment regime, and to compare these estimators in a series of simulation experiments and applications to real data. The results of these simulations along with the theoretical/methodological properties of these estimators are used to form recommendations for applied researchers.**

*Key words*: Adaptive treatment strategies; direct-search estimators; precision medicine; Q-learning.

## 1 Introduction

In clinical practice, treatment decisions are made over the progression of a patient's disease and thus may depend on baseline information as well as evolving treatment and outcome history. A treatment regime is a sequence of decision rules, one per decision point, that maps accumulated patient information to a recommended intervention (Chakraborty & Moodie, 2013; Murphy, 2003; Kosorok & Moodie, 2015; Robins, 2004; Tsiatis *et al.*, 2019). An optimal treatment regime optimises expected cumulative utility when applied to select treatments in the population of interest (for other notions of optimality, see Remark 1). Thus, an optimal treatment regime improves the overall quality of healthcare by personalising intervention decisions to the uniquely evolving health status of each patient. Identification of an optimal treatment regime is closely tied to the problem of subgroup identification, where the goal is to identify patient characteristics that are associated with a favourable response to a given treatment (Ballarini *et al.*, 2018; Lipkovich *et al.*, 2017).

The potential to improve both the quality and efficiency of healthcare has generated tremendous interest in precision medicine. As treatment regimes are the primary mathematical formalisation of precision medicine in biomedical research, there has been a surge of methodological work focused on estimation of optimal treatment regimes from observational and randomised study data (reviews include Chakraborty & Moodie, 2013; Chakraborty & Murphy, 2014; Clifton & Laber, 2020; Kosorok & Laber, 2019; Tsiatis *et al.*, 2019). Within the treatment regimes community, estimators are often classified into one of the following three groups: (i) regression-based, (ii) direct-search, or (iii) model-based planning. As with most coarse categorisations, these groupings are imperfect, and hybrid approaches exist that do not fit cleanly into one of these categories; nevertheless, it is useful to review these broad classes as they (or combinations thereof) cover the vast majority of existing estimators. Regression-based approaches estimate the optimal treatment regime using approximate dynamic programming, implemented through a series of regression models for the conditional mean at each stage, given patient history and treatment. The most widely used regression-based approach is $Q$-/$A$-learning (Murphy, 2003; 2005b; Qian & Murphy, 2011; Robins, 2004; Schulte *et al.*, 2014) along with its many variants (Chakraborty *et al.*, 2010; Ertefaie & Strawderman, 2018; Ertefaie *et al.*, 2021; Goldberg & Kosorok, 2012; Moodie & Richardson, 2010; Moodie *et al.*, 2014; Laber *et al.*, 2014; Leqi & Kennedy, 2022; Song *et al.*, 2015; Zhao *et al.*, 2011; Zhou & Kosorok, 2017). Direct-search methods, also known as value-search methods, construct an estimator of the mean cumulative utility as a function of the treatment regime and then choose the maximiser over a pre-specified class of regimes as the estimated optimal regime. Early work on direct-search estimators includes marginal mean models (Orellana *et al.*, 2010a; 2010b) and classification-based estimators (Rubin & van der Laan, 2012; Zhang *et al.*, 2012; Zhao *et al.*, 2012). These methods have since been extensively studied and generalised (see Huang *et al.*, 2019; Laber & Zhao, 2015; Luckett *et al.*, 2020; Zhang & Zhang, 2018; Zhao *et al.*, 2019; Zhang & van der Schaar, 2020, and references therein). Model-based planning methods estimate the data-generating distribution and then use either estimating equations or Monte Carlo methods to identify an optimal regime. Within the statistics literature, $g$-computation is the primary model-based planning method (Robins, 1986; Robins *et al.*, 1997; Yu & van der Laan, 2002) and underpins fully Bayesian approaches to estimation of an optimal treatment regime (Guan *et al.*, 2020; Lee *et al.*, 2015; Saarela *et al.*, 2015; Thall *et al.*, 2000; Thall *et al.*, 2007; Xu *et al.*, 2016). In the reinforcement learning literature, model-based planning methods are typically used when parsimonious models of the system dynamics are available (see Ghavamzadeh *et al.*, 2015; Sutton & Barto, 2018, for a review and further references).

The large and diverse collection of methodologies to estimate an optimal treatment regime reflects the importance of precision medicine and the increasing maturity of the field. Nevertheless, the field is continuing to evolve and a thorough empirical study designed to evaluate if and when certain methods perform favourably is lacking. In this review, we provide a brief description of a subset of the methods introduced in the preceding paragraph and compare them in a series of simulation experiments and applications. Thus, our intent is to provide an accessible review together with some guidance about how to choose among these methods based on empirical evidence.

In Section 2, we set notation and formalise the notion of an optimal treatment regime. In Section 3, we expound a series of estimators of the optimal treatment regime. In Section 4, we present the results of our simulation experiments. Two real-world data applications are presented in Section 5. A discussion of the simulations and data applications is provided in Section 6, and final conclusions are given in Section 7.

## 2 Setup and Problem Formulation

Suppose that we are planning for the next $K$ treatment decisions in the course of a patient's healthcare. We formalise such a treatment plan as a treatment regime: a sequence of $K$ decision rules, one per treatment decision, each mapping up-to-date patient information to a recommended treatment (Chakraborty & Moodie, 2013; Murphy, 2003; Robins, 2004; Tsiatis *et al.*, 2019). We assume that the available data are composed of $n$ patient trajectories, each an independent copy of $\{S_1, A_1, S_2, A_2, \ldots, S_K, A_K, Y\}$, where $S_1 \in \mathcal{S}_1$ denotes baseline patient information; $A_k \in \mathcal{A}_k$ denotes the treatment assigned at time point $k = 1, \ldots, K$; $S_k \in \mathcal{S}_k$ denotes interim patient information collected during the course of treatment $A_{k-1}$ for $k = 2, \ldots, K$; and $Y \in \mathbb{R}$ denotes the outcome of interest, measured after the final treatment $A_K$, which is coded so that higher values correspond to more desirable health outcomes. The set of available treatments, $\mathcal{A}_k$, $k = 1, \ldots, K$, is typically finite, although in some applications, for example, dose finding or treatment timing, this might be modelled as a continuum (Chen *et al.*, 2016; Guo & Yuan, 2017; Laber & Zhao, 2015; Nie *et al.*, 2020; Schulz & Moodie, 2020; Rich *et al.*, 2014; Zhu *et al.*, 2020). Denote patient history by $\overline{S}^k = (S_1, \ldots, S_k)$ and treatment history by $\overline{A}_k = (A_1, \ldots, A_k)$, so that $H_k = (\overline{S}^k, \overline{A}_{k-1}) \in \mathcal{H}_k$ is the information available to the decision maker at time $k = 1, \ldots, K$, where we define $H_1 \equiv S_1$ and $\mathcal{H}_1 \equiv \mathcal{S}_1$ for convenience. A treatment regime $d = (d_1, \ldots, d_K) \in \mathcal{D} = \mathcal{D}_1 \times \ldots \times \mathcal{D}_K$ is a sequence of decision rules such that $d_k : \mathcal{H}_k \to \mathcal{A}_k$ is a mapping from available patient information to a recommended treatment. Thus, under $d$, a patient with history $H_k = h_k$ at time $k$ will be recommended the intervention $d_k(h_k)$. In general, the observed data are not collected under an optimal treatment regime (lest we could forgo the entire enterprise), and thus, we need additional structure that will allow for identification of the optimal regime in terms of the data-generating model (see Luckett *et al.*, 2017; Wallace *et al.*, 2018, for a discussion of modelling clinician behaviour in observational data). We state these assumptions using the potential outcomes framework (Rubin, 1978; Robins, 1986).

Define $S_k^*(\bar{a}_{k-1})$ to be the potential covariate information under treatment sequence $\bar{a}_{k-1}$ at time $k = 2, \ldots, K$ and let $Y^*(\bar{a}_K)$ be the potential outcome under treatment sequence $\bar{a}_K$. The potential history at stage $k$ under $\bar{a}_{k-1}$ is, therefore, $H_k^*(\bar{a}_{k-1}) = \left\{\overline{S}^{k*}(\bar{a}_{k-1}), \bar{a}_{k-1}\right\}$. The set of all potential outcomes is

$$\mathcal{W} = \left\{Y^*(\bar{a}_K), S_k^*(\bar{a}_{k-1}), k = 1, \ldots, K : \bar{a}_K \in \mathcal{A}_1 \times \ldots \times \mathcal{A}_K\right\},$$

where we have defined $S_1^*(a_0) \equiv S_1$ for convenience. For any treatment regime $d = (d_1, \ldots, d_K)$, the potential outcome under $d$ is thus

$$Y^*(d) = \sum_{\bar{a}_K} Y^*(\bar{a}_K) \prod_{k=1}^{K} 1_{d_k\left\{H_k^*(\bar{a}_{k-1})\right\} = a_k},$$

where $1(\cdot)$ is the indicator function. For any regime, $d \in \mathcal{D}$, define $V(d) = \mathbb{E}Y^*(d)$ to be the mean outcome under $d$. The function $V$, termed the value function, is the primary measure of the performance of a regime. We say a regime, $d^{\text{opt}} \in \mathcal{D}$ is optimal if $V(d^{\text{opt}}) \geq V(d)$ for all $d \in \mathcal{D}$. This definition makes plain that our focus is on estimating an optimal regime within a given (pre-specified) class of regimes. This class may be chosen to be parsimonious, for example, linear regimes or those representable as trees or lists (Laber & Zhao, 2015; Lakkaraju & Rudin, 2017; Qian & Murphy, 2011; Sies & Van Mechelen, 2017; Tao *et al.*, 2018; Zhu *et al.*, 2017; Zhang *et al.*, 2018; Zhang *et al.*, 2015), or may be a large and flexible class, for

example, a Reproducing Kernel Hilbert Space generated by a universal kernel or other machine learning methods (Luedtke & van der Laan, 2016b; Liu *et al.*, 2018; Zhao *et al.*, 2011; Zhao *et al.*, 2012). To characterise the optimal regime in terms of the data-generating distribution, we make the following assumptions: (C1) consistency, $Y = Y^*(\overline{A}_K)$ and $H_k = H_k^*(\overline{A}_{k-1})$, (C2) positivity, $P(A_k = a_k|H_k) > 0$ with probability one for all $a_k \in \mathcal{A}_k$ for $k = 1, \ldots, K$, and (C3) sequential ignorability, $A_k \perp \mathcal{W}|H_k$ for $k = 1, \ldots, K$. We further assume that there is no interference between units and that there are not multiple versions of treatment. These assumptions are standard and have been extensively studied in the literature. For additional discussion see Tsiatis *et al.* (2019).

Throughout, we assume for simplicity that all interventions in $\mathcal{A}_k$ are allowable (feasible) for all patients at stage $k$; in general, this may not hold as the set of allowable treatments may depend on a patient's health status, for example, this set may depend on the presence of co-morbid conditions, patient preference, cost, presence/absence of a biomarker, etc. The methods presented here all generalise immediately to this more general setting (see Tsiatis *et al.*, 2019; van der Laan & Petersen, 2007).

**Remark** *This review, in keeping with vast majority the precision medicine literature, defines an optimal regime as maximising the marginal mean outcome in the target population. However, the mean is not always the best measure of performance. One concern with using the mean is that it is sensitive to extreme observations, so that optimal regime might be heavily influenced by rare events. Another is that it may not align with the clinical goals in a particular application; for example, if the goal is to protect against poor outcomes, one might instead prefer to maximise a lower quantile of the outcome distribution. (Linn* et al.*, 2015; 2017) and Wang* et al. *(2018) consider estimating a regime which maximises the marginal median of the potential outcome distribution, that is,* $d_m^{\text{opt}} = \text{argmax}_D \text{Median}\{Y^*(d)\}$ *(see also Kallus* et al.*, 2019). Recently, Leqi & Kennedy (2022) noted that* $d_m^{\text{opt}}$ *has the rather undesirable property that* $d_m^*(x)$ *can depend on the outcome distribution for patients with* $X \neq x$*. As an alternative, they propose to estimate the regime which maximises the average conditional median* $\mathbb{E}\{\text{Median}\{Y^*(d)|X\}\}$.
*Other criterion used to define an optimal regime include maximising efficacy subject to a constraint on side-effects (Linn* et al.*, 2015; Laber* et al.*, 2018; Wang* et al.*, 2018) or resources (Caniglia* et al.*, 2021; Luedtke & van der Laan, 2016c); maximising choice (Ertefaie* et al.*, 2016; Fard & Pineau, 2011; Laber* et al.*, 2014; Lizotte & Laber, 2016; Wu, 2016); and maximising patient (latent) utility (Butler* et al.*, 2018; Luckett* et al.*, 2020).*

## 3 Methods for Optimal Treatment Regies

### *3.1 Model-based Planning*

The first approach to estimating an optimal treatment regime that we consider involves modelling the complete trajectory distribution under any allowable treatment regime. From such a model, one can derive the marginal mean outcome under any regime $d$, that is, the value function, $V(d) = \mathbb{E}Y^*(d)$, and consequently $d^{\text{opt}} = \text{argmax}_{d \in \mathcal{D}} V(d)$. Using the $g$-computation formula (Chakraborty & Moodie, 2013; Robins, 1986; Tsiatis *et al.*, 2019), the marginal density of $Y^*(d)$ is obtained by first factoring the joint trajectory distribution under $d$ and integrating out over the history. In particular, the density of $Y^*(d)$ at $y$ is

$$f_{Y^*(d)}(y) = \begin{aligned} &\int f_{Y|H_K,A_K}(y|h_K, a_K)\delta_{d_K(h_K)}(a_K)f_{H_K|H_{K-1},A_{K-1}}(h_K|h_{K-1}, a_{K-1}) \\ &\times \ldots \times f_{H_2|H_1,A_1}(h_2|h_1, a_1)\delta_{d_1(h_1)}(a_1)f_{H_1}(h_1)d\mu(h_K, a_K), \end{aligned} \quad (1)$$

where $\delta_u(\cdot)$ is the point mass at $u$; $\mu$ is a measure on $\mathcal{H}_K \times \mathcal{A}_K$; $f_{Y|H_K,A_K}(y|h_K, a_K)$ is the density of $Y$ given $H_K = h_K$ and $A_K = a_K$; $f_{H_k|H_{k-1},A_{k-1}}(h_k|h_{k-1}, a_{k-1})$ is the density of $H_k$ given $H_{k-1} = h_{k-1}$ and $A_{k-1} = a_{k-1}$ for $k = 2, \ldots, K$; and $f_{H_1}(h_1)$ is the density of $H_1$. Each of these densities is identifiable in the observed data and can be estimated using standard methods, for example, maximum likelihood. Plugging these estimators into the integrand in (1) yields an estimator $\widehat{f}_{Y^*(d),n}$ of $f_{Y^*(d)}$. The estimated optimal regime is thus $\widehat{d}_n = \text{argmax}_{d \in \mathcal{D}} \int y \widehat{f}_{Y^*(d),n}(y)dy$. In many realistic applications, it is not possible to obtain $\widehat{f}_{Y^*(d),n}$ in closed form as the integral in (1) is intractable. Instead, one must use Monte Carlo methods whereby one simulates trajectories under any candidate regime $d$ and then averages the simulated outcomes to form an estimator of the associated marginal mean outcome. If $\mathcal{D}$ is complex, one may need to employ sophisticated stochastic optimisation methods to construct an estimator (Gosavi *et al.*, 2015); we will not discuss such computational details here.

Implementation of the $g$-computation formula requires positing models for each of the conditional densities in (1); thus, this approach is best suited to situations where there is either sufficient data to support flexible estimators of these densities or strong underlying clinical theory to inform parsimonious parametric models. When high-quality models can be constructed for these densities, $g$-computation can have good empirical performance (see Chakraborty & Moodie, 2013; Sutton & Barto, 2018; and; Tsiatis *et al.*, 2019 for textbook-level reviews; for recent applications see; Laber *et al.*, 2018; Guan *et al.*, 2020; and; Josefsson & Daniels, 2020).

### 3.1.1 *Model-based planning implementation and extensions*

For the purpose of illustration we describe a simple implementation of model-based planning using location-scale models (Carroll & Ruppert, 1988) for the requisite conditional distributions and a non-parametric model for the distribution of the first stage history. The extension to more complex models is straightforward. We assume that $\mathcal{H}_k \subset \mathbb{R}^{p_k}$ for $k = 1, \ldots, K$. From (1), we must estimate the densities $f_{H_1}, f_{H_2|H_1,A_1}, \ldots, f_{H_K|H_{K-1},A_{K-1}}$, and $f_{Y|H_K,A_K}$. As an estimator of $f_{H_1}$ we take the empirical distribution $\widehat{f}_{H_1} = n^{-1}\Sigma_{i=1}^n \delta_{H_{1,i}}$. For $k = 2, \ldots, K$, we posit models of the form

$$H_k = \mu_k(H_{k-1}, A_{k-1}; \theta_{k,0}) + \Sigma_k^{1/2}(H_{k-1}, A_{k-1}; \theta_{k,1})\epsilon_k,$$

where $\mu_k : \mathcal{H}_{k-1} \times \mathcal{A}_{k-1} \to \mathcal{H}_k$ and $\Sigma_k : \mathcal{H}_{k-1} \times \mathcal{A}_{k-1} \to \mathbb{R}^{p_k \times p_k}$ are smooth functions indexed by the parameter vectors $\theta_{k,0} \in \Theta_{k,0} \subseteq \mathbb{R}^{p_{k,0}}$ and $\theta_{k,1} \in \Theta_{k,1} \subseteq \mathbb{R}^{p_{k,1}}$, and $\epsilon_k \in \mathbb{R}^{p_k}$ is an independent additive error with mean zero and identity covariance. One can estimate $\theta_{k,0}$ using (nonlinear) least squares, that is,

$$\widehat{\theta}_{k,0,n} = \arg\min_{\theta_{k,0}} \sum_{i=1}^n ||H_{k,i} - \mu_k(H_{k-1,i}, A_{k-1,i}; \theta_{k,0})||_2^2,$$

where $||\cdot||_2$ denotes the Euclidean norm. Similarly, one can estimate $\theta_{k,1}$ using

$$\widehat{\theta}_{k,1,n} = \arg\min_{\theta_{k,1}} \sum_{i=1}^n ||\Sigma(H_{k-1,i}, A_{k-1,i}; \theta_{k,1}) - \left\{H_{k,i} - \mu(H_{k-1,i}, A_{k-1,i}; \widehat{\theta}_{k,0,n})\right\}\left\{H_{k,i} - \mu(H_{k-1,i}, A_{k-1,i}; \widehat{\theta}_{k,0,n})\right\}^{\mathsf{T}}||_F^2,$$

where $||\cdot||_F$ denotes the Frobenius norm. Define the standardised empirical residuals $\widehat{\epsilon}_{k,i,n} = \Sigma_k^{-1/2}(H_{k-1,i}, A_{k-1,i}; \widehat{\theta}_{k,1,n})\left\{H_{k,i} - \mu_k(H_{k-1,i}, A_{k-1,i}; \widehat{\theta}_{k,0,n})\right\}$ and define $\widehat{f}_{\epsilon,k,n}$ to be the empirical distribution of the standardised empirical residuals, that is, $\widehat{f}_{\epsilon,k,n} = n^{-1}\Sigma_{i=1}^n \delta_{\widehat{\epsilon}_{i,k,n}}$. A draw, say $\tilde{H}_k$, from the estimated conditional distribution of $H_k$ given $H_{k-1} = h_{k-1}$ and $A_{k-1} = a_{k-1}$ is obtained by drawing $\tilde{\epsilon} \sim \widehat{f}_{\epsilon,k,n}$ and then setting $\tilde{H}_k = \mu(h_{k-1}, a_{k-1}; \widehat{\theta}_{k,0,n}) + \Sigma_k^{1/2}(h_{k-1}, a_{k-1}; \widehat{\theta}_{k,1,n})\tilde{\epsilon}$. Note that, in some settings, one might prefer to model the standardised residuals using a Gaussian copula or other suitable multivariate model.

For the conditional distribution of the outcome $Y$ given $H_K$ and $A_K$ we posit a location-scale model of the form

$$Y = \mu_Y(H_K, A_K; \theta_{Y,0}) + \sigma_Y(H_K, A_K; \theta_{Y,1})\eta_Y,$$

where $\mu_Y : \mathcal{H}_K \times \mathcal{A}_K \to \mathbb{R}$ and $\sigma_Y : \mathcal{H}_K \times \mathcal{A}_K \to \mathbb{R}_+$ are smooth functions indexed by $\theta_{Y,0} \in \Theta_{Y,0} \subseteq \mathbb{R}^{p_{Y,0}}$ and $\theta_{Y,1} \in \Theta_{Y,1} \subseteq \mathbb{R}^{p_{Y,1}}$, and $\eta_Y \in \mathbb{R}$ is an independent additive error with mean zero and unit variance. Estimators $\widehat{\theta}_{Y,0,n}$ and $\widehat{\theta}_{Y,1,n}$ of $\theta_{Y,0}$ and $\theta_{Y,1}$ are constructed using (nonlinear) least squares as described above, that is,

$$\widehat{\theta}_{Y,0,n} = \arg\min_{\theta_{Y,0}} \sum_{i=1}^n \left\{Y_i - \mu_Y(H_{K,i}, A_{K,i}; \theta_{Y,0})\right\}^2$$

and

$$\widehat{\theta}_{Y,1,n} = \arg\min_{\theta_{Y,1}} \sum_{i=1}^n \left[\left\{Y_i - \mu_Y(H_{K,i}, A_{K,i}; \widehat{\theta}_{Y,0,n})\right\}^2 - \sigma_Y^2(H_{K,i}, A_{K,i}; \theta_{Y,1})\right]^2.$$

For each $i = 1, \ldots, n$, define the standardised empirical residuals

$$\widehat{\eta}_{Y,i,n} = \sigma_Y^{-1}(H_K, A_K; \widehat{\theta}_{Y,1,n})\left\{Y_i - \mu_Y(H_{K,i}, A_{K,i}; \widehat{\theta}_{Y,0,n})\right\}.$$

Let $\widehat{f}_{\eta,Y,n}$ denote the empirical distribution of these standardised residuals or some other suitable estimator of the distribution of $\eta_Y$. A draw from the estimated conditional distribution of $Y$ given $H_K = h_K$ and $A_K = a_K$, say $\tilde{Y}$, is thus obtained by drawing $\tilde{\eta}_Y$ from $\widehat{f}_{\eta,Y,n}$ and setting $\tilde{Y} = \mu_Y(h_K, a_K; \widehat{\theta}_{Y,0,n}) + \sigma_Y(h_K, a_K; \widehat{\theta}_{Y,1,n})\tilde{\eta}_Y$.

Given estimated densities for the histories and outcome, we can now use Monte Carlo sampling with the *g*-computation formula to construct an estimator $\widehat{V}_n(d)$ for any regime $d$. Let $d$ be fixed, and let $B$ be large positive integer. An estimator $\widehat{V}_n(d)$ is constructed as follows:

(a) Generate $H_1^{(1)}, \ldots, H_1^{(B)} \sim_{i.i.d} \widehat{f}_{H_1}$;
(b) Set $A_1^{(b)} = d_1(H_1^{(b)})$ for $b = 1, \ldots, B$;
(c) For $k = 2, \ldots, K$;
  (i) Generate $\epsilon_k^{(1)}, \ldots, \epsilon_k^{(B)} \sim_{i.i.d} \widehat{f}_{\epsilon,k,n}$;
  (ii) Set $H_k^{(b)} = \mu_k(H_{k-1}^{(b)}, A_{k-1}^{(b)}; \widehat{\theta}_{k,0,n}) + \Sigma_k^{1/2}(H_{k-1}^{(b)}, A_{k-1}^{(b)}; \widehat{\theta}_{k,1,n})\epsilon_k^{(b)}$;
  (iii) Set $A_k^{(b)} = d_k(H_k^{(b)})$ for $b = 1, \ldots, B$;
(d) Generate $\eta_Y^{(1)}, \ldots, \eta_Y^{(B)} \sim_{i.i.d} \widehat{f}_{\eta,Y,n}$;
(e) Set $Y^{(b)} = \mu_Y(H_K^{(b)}, A_K^{(b)}; \widehat{\theta}_{Y,0,n}) + \sigma_Y(H_K, A_K; \widehat{\theta}_{Y,1,n})\eta_Y^{(b)}$ for $b = 1, \ldots, B$;
(f) Set $\widehat{V}_n(d) = B^{-1}\Sigma_{b=1}^B Y^{(b)}$.

Finally, given a class of regimes $\mathcal{D}$ one can compute $\widehat{d}_n = \text{argmax}_{d \in \mathcal{D}} \widehat{V}_n(d)$.

In two-stage decision problems, the use of mean-variance models can be combined with $Q$-learning (described below) to avoid estimating multivariate conditional densities for the histories (Laber *et al.*, 2014; Linn *et al.*, 2017). However, the use of location-scale models, while convenient, is not necessary. One alternative is to use non-parametric Bayesian models, which have shown strong empirical performance in a number of complex domains (Xu *et al.*, 2016; Laber *et al.*, 2018; Guan *et al.*, 2020). For a survey of model-based planning approaches beyond biomedical applications, see Polydoros & Nalpantidis (2017), Wang *et al.* (2019), and references therein.

### 3.2 Regression-based Methods

The underlying generative model is not needed to identify the optimal treatment regime. In this and the following subsections, we express the optimal regime using estimating equations that require modelling only part of the generative model. These approaches are desirable in settings where there is insufficient data or underlying theory to inform a high-quality model for patient trajectories or if one wants to reduce bias by imposing less structure on the data-generating model.

#### 3.2.1 *Q-learning*

$Q$-/$A$-learning are the most commonly used methods to estimate an optimal treatment regime in biomedical applications. We describe $Q$-learning first as it is easy to implement, flexible, extensible, and robust to some forms of model misspecification (Schulte *et al.*, 2014). $Q$-learning is a regression-based approximate dynamic programming method that is derived from the Bellman optimality equations (Murphy, 2005b; Sutton & Barto, 2018; Clifton & Laber, 2020). Define the $Q$-function at stage $K$ as

$$Q_K(h_K, a_K) = \mathbb{E}(Y | H_K = h_K, A_K = a_K), \tag{2}$$

so that $Q_K(h_K, a_K)$ is the mean outcome for a patient who presents with history $H_K = h_K$ and is assigned intervention $A_K = a_K$ at stage $K$. The mean outcome for a patient presenting with history $H_K$ and who is assigned treatment according to decision rule $d_K$ is thus $Q_K\{h_K, d_K(h_K)\}$, and we define the optimal decision rule in $\mathcal{D}_K$ at stage $K$ to be $d_K^{\text{opt}} = \text{argmax}_{d_K \in \mathcal{D}_K} \mathbb{E} Q_K\{H_K; d_K(H_K)\}$.[1] For $k = K - 1, \ldots, 1$ define the $Q$-function at stage $k$ recursively as follows

$$Q_k(h_k, a_k) = \mathbb{E}\left[Q_{k+1}\left\{H_{k+1}, d_{k+1}^{\text{opt}}(H_{k+1})\right\} \middle| H_k = h_k, A_k = a_k\right], \tag{3}$$

$$d_k^{\text{opt}} = \arg\max_{d_k \in \mathcal{D}_k} \mathbb{E} Q_k\{H_k, d_k(H_k)\}. \tag{4}$$

Thus, the optimal regime in $\mathcal{D}$ can be characterised through the $Q$-functions, which are identifiable in terms of the data-generating model. Furthermore, Equations (2) and (3) immediately suggest an estimator of $d^{\text{opt}}$. First, one can construct an estimator $\widehat{Q}_{K,n}$ of $Q_K$ by estimating the regression of $Y$ on $H_K$ and $A_K$ and subsequently computing the plug-in estimator

$$\begin{aligned} \widehat{d}_{K,n}^{Q} &= \arg\max_{d_k \in \mathcal{D}_k} \mathbb{P}_n \widehat{Q}_{K,n}\{H_K, d_K(H_K)\} \\ &= \arg\max_{d_K \in \mathcal{D}_K} \sum_{i=1}^{n} \widehat{Q}_{K,n}\left\{H_{K,i}, d_K(H_{K,i})\right\}, \end{aligned} \tag{5}$$

where $\mathbb{P}_n$ denotes the empirical expectation operator. Using (3), one can construct an estimator $\widehat{Q}_{k,n}$ of $Q_k$ for $k = K-1, \ldots, 1$ recursively by fitting the regression of $\widehat{Q}_{k+1,n}\left\{H_{k+1}, \widehat{d}^{Q}_{k+1,n}(H_{k+1})\right\}$ on $H_k$ and $A_k$ and subsequently computing

$$\widehat{d}^{Q}_{k,n} = \arg\max_{d_k \in \mathcal{D}_k} \sum_{i=1}^{n} \widehat{Q}_{k,n}\{H_{k,i}, d_k(H_{k,i})\}. \tag{6}$$

The estimator (6) allows for arbitrary regression models for the $Q$-functions, for example, these could be simple parametric models or more flexible machine learning models, which can be chosen independently from the class of regimes $\mathcal{D}$ (however, one should choose the regression model to be sufficiently expressive to capture the complexities in $\mathcal{D}$, see Zhang *et al.*, 2012; Zhang *et al.*, 2015; Zhang *et al.*, 2018; Zhang & Zhang, 2018).

When the set of regimes is unrestricted, for example, if $\mathcal{D}_k$ contains all measurable maps from $\mathcal{H}_k$ into $\mathcal{A}_k$ for $k = 1, \ldots, K$, then it can be shown that $d^{\text{opt}}_K(h_K) = \text{argmax}_{a_K} Q_K(h_K, a_K)$ is the optimal decision rule at the last stage and, recursively for $k = K-1, \ldots, 1$, $d_k(h_k) = \text{argmax}_{a_k} Q_k(h_k, a_k)$ is the optimal rule at stage $k$, where

$$\begin{aligned} Q_k(h_k, a_k) = &\ \mathbb{E}\left[Q_{k+1}\left\{H_{k+1}, d^{\text{opt}}_{k+1}(h_{k+1})\right\} \middle| H_k = h_k, A_k = a_k\right] \\ = &\ \mathbb{E}\left\{\max_{a_{k+1}} Q_{k+1}(H_{k+1}, a_{k+1}) \middle| H_k = h_k, A_k = a_k\right\}. \end{aligned}$$

This expression for the optimal regime suggests an alternative estimation procedure in which $\mathcal{D}$ is *implicitly* defined by the class of models posited for the $Q$-functions. That is, one first constructs an estimator $\widehat{Q}_{K,n}$ of $Q_K$ by estimating the regression of $Y$ on $H_K$ and $A_K$ as described above. The estimated optimal decision rule at stage $K$ is then taken to be $\widehat{d}_{K,n}(h_K) = \text{argmax}_{a_K} \widehat{Q}_{K,n}(h_K, a_K)$. For $k = K-1, \ldots, 1$, one constructs an estimator $\widehat{Q}_{k,n}$ of $Q_k$ by estimating the regression of $\max_{a_{k+1}} \widehat{Q}_{k+1,n}(H_{k+1}, a_{k+1})$ on $H_k$ and $A_k$ and subsequently define $\widehat{d}_{k,n}(h_k) = \text{argmax}_{a_k} \widehat{Q}_{k,n}(h_k, a_k)$. When estimation of the $Q$-functions is carried out this way, it is plain to see that the models for the $Q$-functions and the class of regimes under consideration are conflated. On one hand, this is natural as the $Q$-functions determine the optimal (unrestricted) regime and thus should determine the class of possible regimes. On the other hand, to generate new clinical insights or to inform actual clinical decision making it is often necessary that the estimated regimes be interpretable to domain experts. Thus, if one wants a parsimonious regime when taking this approach, one would need to posit a parsimonious set of $Q$-functions, which may raise concerns about model misspecification. These concerns have, at times, caused some confusion about $Q$-learning forcing an unpleasant choice between model misspecification (via parsimonious models) or unintelligible decision rules (via flexible models). However, as we have seen, these concerns are largely unjustified as one can specify a flexible class of models for the $Q$-functions while enforcing an interpretable or parsimonious class of regimes through specification of $\mathcal{D}$ (see Zhang *et al.*, 2012; Taylor *et al.*, 2015; Zhang & Zhang, 2018; Tsiatis *et al.*, 2019).

### 3.2.2 *Q-learning implementation and extensions*

One of the most appealing aspects of $Q$-learning when the class of regimes is unrestricted (and regression-based procedures in general) is that it can be implemented easily using standard statistical software. All that is required for implementation is the ability to fit regression models and to maximise the fitted regression models over a class of decision rules. Of course, the

difficulty associated with each of these steps depends on the complexity of the models and the class of decision rules. To fix ideas, we first describe linear $Q$-learning with a class of linear decision rules induced by the form of the $Q$-functions. Throughout, we assume that $\mathcal{A}_k = \{0, 1, \ldots, L\}$.

For each $k = 1, \ldots, K$, let $\psi_k : \mathcal{H}_k \to \mathbb{R}^{p_k}$ denote a known feature mapping, for example, this may contain interaction terms or nonlinear basis expansions. We posit models of the form $Q_k(h_k, a_k; \beta_k) = \beta_{0,k}^{\mathrm{T}}\psi_k(h_k) + \sum_{\ell=1}^{L} \beta_{\ell,k}^{\mathrm{T}}\psi_k(h_k)1_{A_k=\ell}$, which is indexed by $\beta_k = (\beta_{k,0}^{\mathrm{T}}, \beta_{k,1}^{\mathrm{T}}, \ldots, \beta_{k,L}^{\mathrm{T}})^{\mathrm{T}}$. We consider decision rules of the form $d_k(h_k; \gamma) = \mathrm{argmax}_{\ell \in \{0,1,\ldots,L\}}\psi_k(h_k)^{\mathrm{T}}\gamma_\ell$, indexed by the parameters $\gamma = \{\gamma_0^{\mathrm{T}}, \gamma_1^{\mathrm{T}}, \ldots, \gamma_L^{\mathrm{T}}\}^{\mathrm{T}}$. Let $\mathcal{D}_k$ denote this class of decision rules and suppose that $Q_k(h_k, a_k) = Q_k(h_k, a_k; \beta_k^*)$ for some $\beta_k^*$. It can be shown that $d_k^{\mathrm{opt}} = \mathrm{argmax}_{d_k \in \mathcal{D}_k}\mathbb{E}Q_k\{H_k, d_k(H_k)\}$ is given by $d_k^{\mathrm{opt}}(h_k) = d_k(h_k; \gamma_k^{\mathrm{opt}})$, where $\gamma_k^{\mathrm{opt}} = (\beta_{k,0}^{*\mathrm{T}}, \beta_{k,0}^{*\mathrm{T}} + \beta_{k,1}^{*\mathrm{T}}, \ldots, \beta_{k,0}^{*\mathrm{T}} + \beta_{k,L}^{*\mathrm{T}})^{\mathrm{T}}$. The estimated optimal regime can thus be constructed using a series of regressions as follows. Define $\widehat{\beta}_{K,n}$ to be the least squares estimator

$$\widehat{\beta}_{K,n} = \arg\min_{\beta_K} \sum_{i=1}^{n} \left\{Y_i - Q_K(H_{K,i}, A_{K,i}; \beta_K)\right\}^2,$$

and subsequently define $\widehat{\gamma}_{K,n} = (\widehat{\beta}_{K,0,n}^{\mathrm{T}}, \widehat{\beta}_{K,1,n}^{\mathrm{T}} + \widehat{\beta}_{K,0,n}^{\mathrm{T}}, \ldots, \widehat{\beta}_{K,L,n}^{\mathrm{T}} + \widehat{\beta}_{K,0,n}^{\mathrm{T}})^{\mathrm{T}}$ so that $\widehat{d}_{K,n}(h_K) = \mathrm{argmax}_\ell \widehat{\gamma}_{K,n,\ell}^{\mathrm{T}}\psi_K(h_K)$. The remaining estimated optimal decision rules $\widehat{d}_k$ are constructed recursively for $k = K - 1, K - 2, \ldots, 1$ via

$$\widehat{\beta}_{k,n} = \arg\min_{\beta_k} \sum_{i=1}^{n} \left[\widehat{Q}_{k+1,n}\left\{H_{k+1,i}, \widehat{d}_{k+1,n}(H_{k+1,i})\right\} - Q_k(H_{k,i}, A_{k,i}; \beta_k)\right]^2,$$

and $\widehat{\gamma}_{k,n} = (\widehat{\beta}_{k,0,n}^{\mathrm{T}}, \widehat{\beta}_{k,1,n}^{\mathrm{T}} + \widehat{\beta}_{k,0,n}^{\mathrm{T}}, \ldots, \widehat{\beta}_{k,L,n}^{\mathrm{T}} + \widehat{\beta}_{k,0,n}^{\mathrm{T}})^{\mathrm{T}}$ so that the estimated optimal decision rule is $\widehat{d}_{k,n}(h_k) = \arg\max_\ell \widehat{\gamma}_{k,n,\ell}^{\mathrm{T}}\psi_k(h_k)$.

We now consider a more complex example in which the $Q$-functions are estimated using a non-parametric class and the class of decision rules is linear. We show that identifying the optimal decision rule in this context can be recast as a mixed integer program. Thus, it requires the use of specialised optimisation libraries and may become prohibitively expensive in large problems or when trying to optimise over all stages jointly (rather than using the one-stage-at-a-time heuristic considered here). In these cases, one can use stochastic search algorithms or weighted classification methods, which we also discuss briefly.

Let $\mathcal{Q}_k$, $k = 1, \ldots, K$ denote a class of models for $Q_k$, for example, these could be neural networks, random forest, super learners (Polley & Van Der Laan, 2010), or linear models with nonlinear basis functions. Consider linear decision rules as before, for example, $d_k(h_k; \gamma_k) = \mathrm{argmax}_{\ell \in \{0,\ldots,L\}}\psi_k(h_k)^{\mathrm{T}}\gamma_{k,\ell}$ indexed by $\gamma_k = \left(\gamma_{k,0}^{\mathrm{T}}, \gamma_{k,1}^{\mathrm{T}}, \ldots, \gamma_{k,L}^{\mathrm{T}}\right)^{\mathrm{T}}$. We describe the procedure at stage $K$ as the procedure for stages $k = K - 1, K - 2, \ldots, 1$ are identical with the outcome $Y$ replaced by $\widehat{Q}_{k+1,n}\left\{H_{k+1}, \widehat{d}_{k+1,n}(H_{k+1})\right\}$. At stage $K$, compute

$$\widehat{Q}_{K,n} = \arg\min_{Q_K \in \mathcal{Q}_K} \mathbb{P}_n\{Y - Q_K(H_K, A_K)\}^2,$$

and for any decision rule $d_K(h_K; \gamma_K)$ define the value

$$\widehat{V}_{K,n}(\gamma_K) \quad = \quad \mathbb{P}_n \widehat{Q}_{K,n}\{H_K, d_K(H_K; \gamma_K)\}.$$

It follows that maximising $\widehat{V}_{K,n}(\gamma_K)$ over $\gamma_K$ is equivalent to solving the following optimisation problem

$$\begin{aligned} \max_{\gamma_K,\, Z_{i,\ell},\, i=1,\ldots,n,\, \ell=0,\ldots,L} \quad & \sum_{i=1}^{n}\sum_{\ell=0}^{L} Z_{i,\ell}\widehat{Q}_{K,n}(H_{K,i}, \ell) \\ \text{such that} \quad & \sum_{\ell=0}^{L} Z_{i,\ell} = 1, \text{ for all } i = 1, \ldots, n \\ & ||\gamma_{K,\ell}|| = 1, \text{ for all } \ell = 0, \ldots, L, \\ & \sum_{\ell=0}^{L} Z_{i,\ell}\phi_K(H_{K,i})^{\mathrm{T}}\gamma_{K,\ell} \geq \phi_K(H_{K,i})^{\mathrm{T}}\gamma_{K,j} \\ & \qquad \text{for all } j = 0, \ldots, L,\ i = 1, \ldots, n \\ & Z_{i,\ell} \in \{0,1\},\ \gamma_{K,\ell} \in \mathbb{R}^{\dim\psi_k(h_k)} \\ & \qquad \text{for all } \ell = 0, \ldots, L,\ i = 1, \ldots, n, \end{aligned}$$

which can be seen to be a mixed integer program. Such problems can be solved to machine precision using specialised optimisation libraries such as CPLEX or XPRESS (Anand *et al.*, 2017; Rudin & Ertekin, 2018), but run-times can be excessive when the dimension is large (for example, see Laber & Murphy, 2011). In cases where computation is not feasible, one can use stochastic search methods, for example, simulated annealing, genetic algorithms, or simultaneous perturbation (Spall, 2005).

Another approach to computing $\text{argmax}_{\gamma_K}\widehat{V}_{K,n}(\gamma_K)$ is to recast it as a weighted classification problem and then to use existing classification algorithms to approximate a solution (Zhang *et al.*, 2012; Zhang *et al.*, 2013; Zhao *et al.*, 2012; Zhang & Zhang, 2018). Let $d_K(h_K; \gamma_K)$ denote a decision rule indexed by (possibly infinite-dimensional) parameter $\gamma_K \in \Gamma_K$, for example, the class $\{d_K(h_K; \gamma_K) : \gamma \in \Gamma_K\}$ may be the class of linear decision rules as discussed above or it may be the set of rules representable as trees or decision lists. Write

$$\begin{aligned} \arg\max_{\gamma_K} \widehat{V}_{K,n}(\gamma_K) \quad & = \arg\max_{\gamma_K} \mathbb{P}_n \widehat{Q}_{K,n}\{H_K, d_K(H_K; \gamma_K)\} \\ & = \arg\min_{\gamma_K} \mathbb{P}_n \left[\max_{\ell} \widehat{Q}_{K,n}(H_K, \ell) - \widehat{Q}_{K,n}\{H_K, d_K(H_K; \gamma_K)\}\right], \end{aligned} \tag{7}$$

where we have used the fact that maximisation is equivalent to minimising the negative of the same objective and the fact that the term $\max_\ell \widehat{Q}_{K,n}(H_K, \ell)$ does not affect the argmin as it does not depend on $\gamma_K$. It can be seen that each term in (7) is non-negative and is equal to zero when $d_K(H_K; \gamma_K) \in \text{argmax}_\ell \widehat{Q}_{K,n}(H_K, \ell)$. Thus, one can view the minimisation problem in (7) is as a weighted classification problem in which the 'correct class' at history $H_K = h_K$ is given by $\text{argmax}_\ell \widehat{Q}_{K,n}(H_K, \ell)$ (with an element chosen arbitrarily if the argmax is not a singleton) and the cost associated with predicting class $a$ at $H_K = h_K$ is equal to $\max_\ell \widehat{Q}_{K,n}(h_K, \ell) - \widehat{Q}_{K,n}(h_K, a)$. With this perspective, one can create a training set of input-label-cost triples

$$\left\{\left[H_{K,i}, \arg\max_{\ell} \widehat{Q}_{K,n}(H_{K,i}, \ell), \left\{\max_{\ell} \widehat{Q}_{K,n}(H_{K,i}, \ell) - \widehat{Q}_{K,n}(H_{K,i}, a)\right\}_{a=0}^{L}\right]\right\}_{i=1}^{n},$$

that is, the input is $H_K$, the correct label is $\text{argmax}_\ell \widehat{Q}_{K,n}(H_K, \ell)$, and the associated vector of costs is $\left\{\max_\ell \widehat{Q}_{K,n}(H_K, \ell) - \widehat{Q}_{K,n}(H_K, a)\right\}_{a=0}^{L}$. This data can then be fed into any multi-class classification algorithm that accommodates sample weights (see Zadrozny

*et al.*, Zadrozny *et al.*; Abe *et al.*, 2004, and references therein for methods that can be used to make any classification algorithm cost-sensitive).

**Remark** *Note that the classification dataset we constructed was only at the observed histories* $\{H_{K,i}\}_{i=1}^{n}$*; however, one could generate an arbitrary number of histories from any surrogate distribution, in particular, one might generate histories from the estimated density of* $H_K$ *computed using g-computation under the estimated optimal strategy. The reasons one might do this are twofold. First, the empirical distribution of* $H_K$ *is under the data-generating model and not the counterfactual distribution induced by the optimal decision rules at times* $k = 1, \ldots, K-1$*; consequently, optimising* $\widehat{V}_n(\gamma_K)$ *over* $\gamma_K$ *needs not lead to a global optimiser unless the class of rules indexed by* $\gamma_K$ *includes the globally optimal rule* $h_K \mapsto \text{argmax}_\ell Q_K(h_K, \ell)$*. Using the estimated distribution could restore global consistency if the g-computation model is correctly specified. Second, one can generate many more than* $n$ *data points thereby potentially reducing variance (Breiman & Shang, 1996).*

*Note that using inverse probability weighting (IPW) or augmented IPW (AIPW) will re-weight the observed histories to ensure the expectation is taken with respect to the correct counterfactual distribution; however, because these estimators average over histories that are consistent at previous stages, they can quickly run out of data in multi-stage problems.*

### 3.2.3 Other methods

We have reviewed a select subset of estimators that we felt illustrated key ideas in regression-based estimation of an optimal treatment regime. In this section, we catalogue several important estimators which were omitted from our detailed discussion along with references for further study.

*A-learning.*

*A*-learning (Murphy, 2003; Schulte *et al.*, 2014) is a doubly-robust alternative to *Q*-learning. For simplicity, we consider $\mathcal{A}_k = \{0, 1\}$ and suppose that the set of regimes is unrestricted. In this setting, *A*-learning targets the so-called contrast function

$$C_k(h_k) = Q_k(h_k, 1) - Q_k(h_k, 0), \quad k = 1, \ldots, K. \tag{8}$$

If $C_k(h_k)$ and either (but not necessarily both) $Q_k(h_k, 0)$ or $P(a_k|h_k)$ are correctly specified, it can be shown that the *A*-learning estimator $\widehat{C}_{k,n}(h_k)$ of $C_k(h_i)$ is consistent. The estimated optimal decision rule is $\widehat{d}^{A}_{k,n}(h_k) = 1\{\widehat{C}_{k,n}(h_k) > 0\}$, $k = 1, \ldots, K$. A description of the *A*-learning estimator and discussion of its merits relative to *Q*-learning can be found in Schulte *et al.* (2014).

*Targeted maximum likelihood.*

van der Laan & Luedtke (2014) proposed a regression method to estimate the blip function by super learning, which is an ensemble machine learning approach. In their setting, the decision rule $d_k$ depends on $H_k$ only through $(X_k, A_{k-1})$, where $X_k$ is a summary function of $H_k$. This approach can consider a more interpretable class of regimes, and it is sometimes more computationally efficient to estimate the optimal treatment regime within the restricted class. For convenience, we consider $\mathcal{A} = \{0, 1\}$ and the problem of a single decision point, that is, $K = 1$. In this case, the blip function is defined as

$$B(X_1) = \mathbb{E}\{\mathbb{E}(Y|A_1 = 1, S_1) - \mathbb{E}(Y|A_1 = 0, S_1)|X_1\}.$$

Note that the blip function is equal to the contrast function (8) if $X_1 = H_1 \equiv S_1$. The optimal decision rule is $d_1^{\text{opt}}(h_1) = 1\{B(x_1) > 0\}$, where $x_1$ is a summary function of $h_1$. The authors proposed a loss function for the blip function based on IPW and performed regression by applying super learners (see Section 3 of van der Laan & Luedtke, 2014). Under this approach, the estimated optimal decision rule is $\widehat{d}^{B}_{1,n}(h_1) = 1\{\widehat{B}_n(x_1) > 0\}$, where $\widehat{B}_n$ is the estimator of the blip function. The method is shown to possess the double-robustness property, that is, if either $Q_k(h_k)$ or $P(a_k|h_k)$ is correctly specified, the proposed loss function is shown to be valid in that the true blip function is the minimiser of the loss function.

*Robust Q-learning.*

Ertefaie *et al.* (2021) proposed a variant of Q-learning, termed robust Q-learning, which posits a parametric model for the contrasts $C_k(h_k)$ and non-parametric models for the main effects $\mu_k(h_k) = \{Q_k(h_k, 1) + Q_k(h_k, 0)\}/2$. Because non-parametric models are used for the main effects, robust Q-learning is less prone to misspecification than fully-parametric Q-learning. Using sample splitting, Ertefaie *et al.* (2021) obtain asymptotic normality for the coefficients indexing the contrast functions even if the estimated main effects converge at sub-parameteric rates. Robust Q-learning is an example of double-machine learning which has become an extremely active area of research in semi-parametric methods as of late (see Chernozhukov *et al.*, 2018; Semenova & Chernozhukov, 2020; Kennedy, 2022, and references therein).

**Remark** *In this review, we have primarily focused on settings in which one had available a parsimonious representation of the history. However, in some settings, one may wish to tailor treatment using genetic, functional, imaging, or other high-dimensional data (McKeague & Qian, 2011; 2014; Ciarleglio* et al.*, 2016; Laber & Staicu, 2018; Ciarleglio* et al.*, 2018). In such settings, one can apply penalised variants of many of the methods discussed in this review by simply adding a sparse penalty, for example, a lasso- or elastic net-type penalty to the estimation procedure. Examples include penalised Q-learning (Qian & Murphy, 2011; Zhu* et al.*, 2019), A-learning (Shi* et al.*, 2018), and value-search (Song* et al.*, 2015). One interesting exception is the penalised Q-learning method of Song* et al. *(2015) which shrinks the treatment contrast in Q-learning rather than the coefficients themselves which induces a kind of fused-lasso effect (see also Chakraborty* et al.*, 2010; Moodie & Richardson, 2010). In our simulation study, we examine Q-learning with a lasso penalty.*

### 3.3 Value-Search

The $Q$-function representation given in Section 3.2 characterises the optimal regime recursively through the so-called optimal $Q$-functions (the qualifier 'optimal' reflects the fact that the $Q$-functions are maximised at each step). However, with only slight modification, this recursion can be used to identify the map $d \mapsto V(d)$ and subsequently to construct an estimator $d \mapsto \widehat{V}_n(d)$ from which we obtain an estimated optimal regime $\widehat{d}_n = \text{argmax}_{d \in \mathcal{D}} \widehat{V}_n(d)$. Estimated optimal regimes constructed by maximising an estimator $\widehat{V}_n$ of $V$ over a class of regimes are known as value-search or policy-search estimators (Sutton & Barto, 2018; Tsiatis *et al.*, 2019). The form of a value-search estimator is appealing in that it mimics the optimality criterion that defines an optimal regime. Furthermore, given a consistent estimator $\widehat{V}_n$ of $V$, using a value-search estimator directly targets the best regime in the class $\mathcal{D}$ regardless of whether this class contains the optimal regime within the class of all measurable maps from history to actions (see Qian & Murphy, 2011, for further discussion).

In the remainder of this section, we describe how to use $Q$-functions to obtain an estimator $\widehat{V}_n^Q(d)$ of any regime. We then describe an alternative formulation using inverse probability weighting (IPW), which does not require models for the $Q$-functions. Next, we discuss augmented inverse probability weighting (AIPW), which combines these estimators and enjoys the double-robustness property. Finally, we introduce some other common value-search approaches such as $C$-learning and DR-IPCW. We focus on two-stage problems $K = 2$ and binary treatments coded so that $\mathcal{A}_k = \{0, 1\}$ for $k = 1, 2$; the fully general case is described in Tsiatis *et al.* (2019).

Let $d = (d_1, d_2)$ be an arbitrary regime. As in Section 3.2, define the terminal (second stage) $Q$-function as $Q_2(h_2, a_2) = \mathbb{E}(Y|H_2 = h_2, A_2 = a_2)$. Rather than maximising, we simply plug-in the second stage decision rule to obtain the first stage $Q$-function under second stage decision rule $d_2$ as follows

$$Q_1(h_1, a_1; d_2) = \mathbb{E}[Q_2\{H_2, d_2(H_2)\}|H_1 = h_1, A_1 = a_1].$$

Thus, $Q_1(h_1, a_1; d_2)$ is the expected outcome for a patient presenting with $H_1 = h_1$, assigned treatment $A_1 = a_1$ and then treated with $d_2$ at the second stage. The value of $d$ is, therefore, $V^Q(d) = \mathbb{E}Q_1\{H_1, d_1(H_1); d_2\}$. An estimator $\widehat{V}_n^Q$ of $V^Q$ is obtained by (i) constructing an estimator $\widehat{Q}_{2,n}$ of $Q_2$ by regressing $Y$ on $H_2$ and $A_2$; (ii) constructing an estimator $\widehat{Q}_{1,n}(\cdot, \cdot; d_2)$ of $Q_1(\cdot, \cdot; d_2)$ by regressing $\widehat{Q}_{2,n}\{H_2, d_2(H_2)\}$ on $H_1$ and $A_1$; and (iii) computing $\widehat{V}_n^Q(d) = \mathbb{P}_n\widehat{Q}_{1,n}\{H_1, d_1(H_1); d_2\}$. As mentioned previously, one is free to select regression models that are deemed appropriate based on domain science and/or exploratory analyses.

While flexible models can be used for the $Q$-functions, choosing such models may not always be straightforward and numerous authors have expressed concerns about misspecification (e.g. Zhao *et al.*, 2012). An alternative approach, which we now review, is to use AIPW, which is consistent as long as either the treatment assignment mechanism (the propensity scores) or the $Q$-functions are correctly specified. Furthermore, such estimators enjoy a number of important theoretical advantages when flexible models are used for the propensities or the $Q$-functions (Chernozhukov *et al.*, 2017; Zhang & Zhang, 2018).

For each $a_k$ and history $h_k$, define the propensity score $P(a_k|h_k) = P(A_k = a_k|H_k = h_k)$ and let $\widehat{P}_n(a_k|h_k)$ denote the estimated propensities, for example, these could be constructed using multinomial logistic regression. The inverse probability weighted estimator (IPWE) of the $V(d)$ is

$$\widehat{V}_n^{\text{IPWE}}(d) = \mathbb{P}_n\left\{\frac{Y1_{A_1=d_1(H_1)}1_{A_2=d_2(H_2)}}{\widehat{P}_n(A_1|H_1)\widehat{P}_n(A_2|H_2)}\right\},$$

which can be seen to be the Horvitz-Thompson estimator of the mean outcome under $d$ (Horvitz & Thompson, 1952). One could construct an estimator of the optimal treatment regime via $\widehat{d}_n^{\text{IPWE}} = \text{argmax}_{d\in\mathcal{D}}\widehat{V}_n^{\text{IPWE}}(d)$. However, the IPWE is highly unstable as it only uses a small fraction of the observed data, for example, if data were collected in a two-stage binary treatment sequential multiple assignment randomised trial (SMART Murphy, 2005a) then only a quarter of the observed data would be used in the IPWE (in expectation). For this reason, it is common to augment the IPWE with a regression term to gain efficiency. The Augmented IPWE (AIPWE) of $V(d)$ is given by

$$\widehat{V}_n^{\text{AIPWE}}(d) = \widehat{V}_n^{\text{IPWE}}(d) - \mathbb{P}_n\left[\frac{1_{A_1=d_1(H_1)} - \widehat{P}_n(A_1|H_1)}{\widehat{P}_n(A_1|H_1)}\,\widehat{Q}_{1,n}\{H_1, d_1(H_1); d_2\}\right]$$
$$-\mathbb{P}_n\left[\frac{1_{A_1=d_1(H_1)}\left\{1_{A_2=d_2(H_2)} - \widehat{P}_n(A_2|H_2)\right\}}{\widehat{P}_n(A_1|H_1)\widehat{P}_n(A_2|H_2)}\,\widehat{Q}_{2,n}\{H_2, d_2(H_2)\}\right],$$

where $\widehat{Q}_{1,n}$ and $\widehat{Q}_{2,n}$ are the estimated $Q$-functions as described previously. The estimated optimal regime is given by $\widehat{d}_n^{\text{AIPWE}} = \text{argmax}_{d\in\mathcal{D}}\widehat{V}_n^{\text{AIPWE}}(d)$. The AIPWE is doubly-robust in that it is consistent if either the propensity scores or $Q$-functions are correctly specified and furthermore attains the semi-parametric efficiency bound when both are correct (for reviews, see Tsiatis (2007), Molenberghs *et al.* (2014), & Tsiatis *et al.* (2019) and for some interesting recent developments see Luedtke *et al.* (2017) and references therein).

$C$-learning (Zhang & Zhang, 2018) is another value-search method, which can be viewed as a weighted classification problem considering $\mathcal{A}_k = \{0, 1\}$. The loss function is defined as a 0-1 loss multiplied by a weight of the contrast function, and the optimal decision rule minimises the weighted loss:

$$d_k^{\text{opt}} = \arg\min_{d_k\in\mathcal{D}_k} \mathbb{E}\left[|C_k(H_k)|1_{d_k(H_k)\neq 1\{C_k(H_k)>0\}}\right],$$

where $C_k$ is the contrast function given in (8), $k = 1, 2$. Estimation for $d_k^{\text{opt}}$ reduces to constructing an estimator $\widehat{C}_{k,n}$ of $C_k$, which can be carried out using regression as shown in Zhang *et al.* (2012). The estimated optimal decision rule at stage $k$ is then given by

$$\widehat{d}_{k,n}^{C} = \arg\min_{d_k\in\mathcal{D}_k} n^{-1}\sum_{i=1}^{n}\left[|\widehat{C}_{k,n}(H_{k,i})|1_{d_k(H_{k,i})\neq 1\{\widehat{C}_{k,n}(H_{k,i})>0\}}\right].$$

The method of double robust inverse probability of treatment and censoring weighting (DR-IPCW) is an alternative value-search method (van der Laan & Luedtke, 2014). The authors proposed a loss function for $d \in \mathcal{D}$ with expectation equaling to $-\mathbb{E}Y^*(d)$ if either $Q_k(h_k, a_k)$ or $P(a_k|h_k)$ is correctly specified. Thus, estimating the optimal decision rule $d^{\text{opt}}$ is equivalent to minimising the loss function over $\mathcal{D}$.

Other direct optimisation methods include targeted learning (van der Laan & Rose, 2011; 2018), decision lists (Zhang *et al.*, 2018), outcome weighted learning (Zhao *et al.*, 2015), quantile learning (Wang *et al.*, 2018) and marginal mean models (Orellana *et al.*, 2010a; 2010b). Because the majority of the above methods directly estimate the expected outcome across a class of regimes, an estimator of $\mathbb{E}Y^*(d^{\text{opt}})$ can also be obtained by applying these methods.

### 3.3.1 *Value-search implementation and extensions*

Implementation details in the estimation of the $Q$-functions has already been discussed in Section 3.2.2.; thus, the remaining considerations are (i) estimation of the propensity scores and (ii) optimisation over the class of candidate regimes. Estimation of the propensity scores can be carried out using any classification algorithm that produces probability estimates. Logistic regression is often used in practice although one could use kernel-methods, neural networks, or boosting (possibly with Platt scaling, etc.) There is currently a rich line of research on the use of sample-splitting to obtain $\sqrt{n}$ -consistent and asymptotically normal estimators when flexible (e.g. machine learning) models are used to estimate the propensity score and the $Q$-functions

(Naimi & Kennedy, 2017; Chernozhukov *et al.*, 2017; Chernozhukov *et al.*, 2018; Semenova & Chernozhukov, 2020); however, the development of these methods is ongoing and rather technical so we do not include them in the present review.

Given an estimator $\widehat{V}_n(d)$, for example, constructed using (A)IPWE or $Q$-learning, one must then construct an approximate maximiser of $d \mapsto \widehat{V}_n(d)$ over a class of regimes $d \in \mathcal{D}$. In cases where $\mathcal{D}$ comprises parametric regimes, that is, $\mathcal{D} = \left\{ d = (d_1(\cdot\,; \beta_1), d_2(\cdot\,; \beta_2) : \beta = (\beta_1^{\mathrm{T}}, \beta_2^{\mathrm{T}})^{\mathrm{T}} \in \mathcal{B} \right\}$, where $\mathcal{B}$ is a compact subset of Euclidean space, one can use stochastic search methods to approximate a solution. Let $d_\beta = \{d_1(\cdot\,; \beta_1), d_2(\cdot\,; \beta_2)\}$, and define $\widehat{V}_n(\beta) = \widehat{V}_n(d_\beta)$. Let $\widehat{\beta}_n^{(0)}$ denote a starting value, for example, one might construct a warm-start $\widehat{\beta}_n^{(0)}$ using the *ad hoc* iterative $Q$-learning procedure described in the preceding section. A simple stochastic gradient descent algorithm uses updates of the form

$$\widehat{\beta}_n^{(k+1)} = \widehat{\beta}_n^{(k)} + \lambda^{(k)} \left\{ \widehat{V}_n(\widehat{\beta}_n^{(k)} + Z^{(k)}) - \widehat{V}_n(\widehat{\beta}_n^{(k)}) \right\} Z^{(k)},$$

for $k \geq 1$, where $Z^{(1)}$, $Z^{(2)}$, …, are independent random vectors with mean zero and identity covariance matrix and $\lambda^{(1)}$, $\lambda^{(2)}$, …, are (possibly data-dependent) tuning parameters that govern the learning rate (Spall, 2005; Bottou, 2010).

If the class $\mathcal{D}$ is comprised of trees or lists, one can use greedy heuristics analogous to those used in classification and regression trees (Loh, 2014; Laber & Zhao, 2015; Zhang *et al.*, 2018). For kernel-based decision rules, (Zhao *et al.*, 2015) use convex surrogates for $\widehat{V}_n$ to approximate iterative (stage-wise) and simultaneous (joint) optimal decision rules. For more complex classes of treatment regimes, one can use simulated annealing, genetic algorithms, or other heuristics to approximate a solution (Zhang *et al.*, 2013).

## 4 Simulation Studies

In this section, we investigate the performance of some of the methods described in the preceding section using Monte Carlo simulations. We explore their performance in estimating the optimal treatment regime as well as their performance in estimating the expected outcome under the optimal treatment regime. The simulation setup is similar to Schulte *et al.* (2014) and van der Laan & Luedtke (2014); additional details are provided in the next section.

For the estimation of the optimal treatment regime, we consider the following methods: (1) $Q$-learning with multiple linear regression (QMR), (2) penalised QMR, (3) $Q$-learning with random forests (QRF, Taylor *et al.*, 2015), (4) $A$-learning, (5) targeting the blip function (van der Laan & Luedtke, 2014), (6) targeting the regret function (Murphy, 2003), (7) DR-IPCW (van der Laan & Luedtke, 2014), (8) $C$-learning (Zhang & Zhang, 2018), (9) backward outcome weighted learning (BOWL) (Zhao *et al.*, 2015), (10) decision lists (Zhang *et al.*, 2018) and (11) model-based planning. In terms of the three broad categories we introduced previously: (1)–(6) are regression-based methods; (7)–(10) are direct/value-search methods; and (11) is (of course) a model-based planning method. We measure the performance of these estimators using V-efficiency as defined in Schulte *et al.* (2014); that is,

$$R(\widehat{d}_n) = \frac{\mathbb{E}\{\mathbb{E}Y^*(\widehat{d}_n)\}}{\mathbb{E}Y^*(d^{\mathrm{opt}})},$$

where $\widehat{d}_n$ denotes the estimated optimal treatment regime constructed from data on $n$ patients. V-efficiency captures the extent to which $\widehat{d}_n$ achieves the value of the true optimal regime. Higher values of $R(\widehat{d}_n)$ indicate better performance. When presenting results, we scale the V-efficiency by 100 to obtain the percentage of $\mathbb{E}Y^*(d^{\text{opt}})$ obtained by $\widehat{d}_n$ on average. In addition to V-efficiency, we also calculate the expected potential outcome $\mathbb{E}Y^*(\widehat{d}_n)$ and the average agreement with the optimal treatment, that is, $T(\widehat{d}_n) = K^{-1}\mathbb{E}\,\Sigma_{k=1}^{K} 1_{\widehat{d}_{k,\,n}(H_k)=d_k^{\text{opt}}(H_k)}$, for each method.

We next evaluate several methods for estimating $V(d^{\text{opt}})$; this quantity is of interest when one wishes to assess the value of applying a personalised treatment strategy in a given domain (Laber *et al.*, 2016; Rose *et al.*, 2019). Thus, one must construct an estimator $\widehat{d}_n$ *and* evaluate it via $\widehat{V}_n(\widehat{d}_n)$. To estimate the expected outcome under the estimated optimal treatment regime, we consider (1) $Q$-learning with multiple linear regression (QMR), (2) $Q$-learning with random forests (QRF), (3) $A$-learning, (4) targeting the regret function, (5) targeted learning (TMLE) (van der Laan & Rubin, 2006), and (6) inverse probability treatment weighting (IPTW) (van der Laan & Rose, 2018). For TMLE and IPTW, we first use DR-IPCW to estimate the optimal treatment regime (because of its superior performance in estimating the optimal treatment regime) and then apply these two methods to estimate the expected outcome under the estimated optimal treatment regime. For the measure of performance, we consider the mean squared error $\text{MSE} = \mathbb{E}\left\{\widehat{V}_n(\widehat{d}_n) - V(d^{\text{opt}})\right\}^2$. Lower MSE corresponds to better performance. In Appendix D1, we also show results on *regime evaluation* in which we compare estimators $\widehat{V}_n(d)$ of $V(d)$ for fixed regimes $d$.

Most of the methods considered here require that the $Q$-function $Q_k(h_k, a_k)$ be estimated. With the exceptions of QMR and model-based planning, we estimate $Q_k(h_k, a_k)$ using a random forest (we use the default settings in the `randomForest` package in R). For QMR, we estimate $Q_k(h_k, a_k)$ using linear regression on the main effects, that is, $h_k$, $a_k$, and the first-order interaction terms between treatment and history. We considere $L_2$ penalisation terms in penalised QMR. For the method of model-based planning, we estimate the $Q$- and $V$-functions using Monte Carlo methods (see details in Section 3). We assume that the transition and conditional outcome density follow normal/binomial distributions with linear mean models and constant covariance-matrices which are estimated using maximum likelihood (additional details about the working models are given in Appendix B1). Thus, model-based planning coincides with QMR when $K = 1$.

Under the scenarios for which the propensity is unknown, we estimate $P(a_k|h_k)$ using logistic regression on the main terms of $h_k$. Note that not all methods require the estimation of $P(a_k|h_k)$. For methods such as $C$-learning and BOWL, for which the class of regimes $\mathcal{D}_k$ must be computationally tractable (e.g. parametric), we consider $\mathcal{D}_k$ to be a class of linear regimes. For all others, the class of regimes is unrestricted, that is, $\mathcal{D}_k$ contains all measurable maps from $\mathcal{H}_k$ into $\mathcal{A}_k$, $k = 1, \ldots, K$. We use $Q$-learning with correctly specified models as the gold standard.

### 4.1 Data Generating Models

We consider the scenarios where continuous and binary outcomes are measured after $K = 1, 2, 4, 8$ decision points. For each scenario, we take the number of patients to be $n = 250, 500, 1000$. The available treatments at each decision point are binary. We perform 500 Monte Carlo runs for each scenario.

#### 4.1.1 Continuous outcome

First, we consider a continuous outcome in $K = 1, 2, 4, 8$ decision points. For $K = 1, 2$, we use the data-generating models shown below, which are similar to those used in Schulte et al. (2014, Section 6). The data-generating models for $K = 4, 8$ are more cumbersome and therefore relegated to Appendix A1.

In the case of one decision point, suppose that we observe i.i.d. data $(S_i, A_i, Y_i),\ i = 1, \ldots, n$ with

$$\begin{aligned} S &\sim N\left\{\begin{pmatrix}0\\3\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right\} \\ A|S &= (s_1, s_2) \sim \text{Bernoulli}\left\{\text{expit}(\phi_{10}^0 + \phi_{11}^0 s_1 + \phi_{12}^0 s_1^2)\right\}, \\ Y|S &= (s_1, s_2), A = a \sim N\left\{\beta_{10}^0 + \beta_{11}^0 s_1 + \beta_{12}^0 s_1^2 + \beta_{13}^0 s_2 + a(\psi_{10}^0 + \psi_{11}^0 s_1),\ 1\right\}, \end{aligned}$$

where $\boldsymbol{\phi}_1^0 = (0, -2, 1)$, $\boldsymbol{\beta}_1^0 = (1, 1, 5, 1)$, $\boldsymbol{\psi}_1^0 = (1, 0.5)$ and $\text{expit(x)} = e^x/(1 + e^x)$.

In the case of two decision points, suppose we observe i.i.d. data $(S_{1i}, A_{1i}, S_{2i}, A_{2i}, Y_i),\ i = 1, \ldots, n$ with

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

where $m(s_1, s_2, a_1, a_2) = \beta_{20}^0 + \beta_{21}^0 s_1 + \beta_{22}^0 a_1 + \beta_{23}^0 s_1 a_1 + \beta_{24}^0 s_2 + \beta_{25}^0 s_2^2 + a_2(\psi_{20}^0 + \psi_{21}^0 a_1 + \psi_{22}^0 s_2)$, $\boldsymbol{\phi}_1^0 = (0.3, 0.5)$, $\boldsymbol{\delta}_1^0 = (0, 0.5, -0.75, 0.25)$, $\boldsymbol{\phi}_2^0 = (0, 0.5, 0.1, -1, -0.1, 1)$, $\boldsymbol{\beta}_2^0 = (3, 0, 0.1, -0.5, -0.5, 5)$, and $\boldsymbol{\psi}_2^0 = (1, 0.25, 0.5)$.

The true value of the expected outcome under the optimal treatment regime ($\psi$) for one and two decision points are 10.009 and 16.222, respectively.

#### 4.1.2 Binary outcome

Here, we consider a binary outcome in either one or two decision points using data-generating models similar to those used in van der Laan & Luedtke (2014).

In the case of one decision point, suppose that we observe i.i.d. data $(S_i, A_i, Y_i),\ i = 1, \ldots, n$ with

$$\begin{aligned} &S \sim N(\mathbf{0}, I_4), A|S = (s_1, s_2, s_3, s_4) \sim \text{Bernoulli}(0.5),\ Y|S = (s_1, s_2, s_3, s_4), \\ &A = a \sim \text{Bernoulli}\left[\frac{1}{2}\text{expit}(1 - s_1^2 + 3s_2^2 + 5s_3^2 - 4.45) + \frac{1}{2}\text{expit}\{-0.5 - s_3 + 2s_1 s_2 + a(3|s_2| - 1.5)\}\right]. \end{aligned}$$

In the case of two decision points, suppose that we observe i.i.d. data $(S_{1i}, A_{1i}, S_{2i}, A_{2i}, Y_i),\ i = 1, \ldots, n$ with

$$\begin{aligned} &S_{11}, S_{12} \overset{iid}{\sim} \text{Unif}(-1, 1), A_1|S_{11}, S_{12} \sim \text{Bernoulli}(0.5),\ U_1, U_2|A_1, S_{11}, S_{12} \overset{iid}{\sim} \text{Unif}(-1, 1), \\ &S_{21}|S_{11}, S_{12}, A_1 \sim U_1(1, 25A_1 + 0.25), S_{22}|S_{11}, S_{12}, A_1, \\ &S_{21} \sim U_2(1, 25A_1 + 0.25), A_2|S_{11}, S_{12}, A_1, S_{21}, S_{22} \sim \text{Bernoulli}(0.5), \\ &Y|S_{11}, S_{12}, A_1, S_{21}, S_{22}, A_2 \sim \text{Bernoulli}\{0.4 + 0.069b(S_{11}, S_{12}, A_1, S_{21}, S_{22}, A_2)\}, \end{aligned}$$

where $b(S_{11}, S_{12}, A_1, S_{21}, S_{22}, A_2) = 0.5A_1[-0.8 - 3\{\text{sign}(S_{11}) + S_{11}\} - S_{12}^2] + A_1\{-0.35 + (S_{21} - 0.5)^2\} + 0.08A_1A_2$.

The true value of the expected outcome under the optimal treatment regime ($\psi$) for one and two decision points are 0.561 and 0.487, respectively. For the regression-based methods, we employ the logistic regression to model the binary outcome and for the model-based planning, we assume the conditional outcome follows a binomial distribution with a mean which is linear in the history.

### 4.2 Simulation Results

Table 1 displays the values of $R(\widehat{d}_n)$ for each method when the outcome is continuous. In this case, we provide results when the propensity is either known or unknown, thus reflecting data from randomised and observational studies. It can be seen that the non-parametric methods, that is, Blip, DR-IPCW and QRF, typically have better predictive performance compared with (semi-)parametric methods (QMR, Planning, *A*-learning, *C*-learning, Regret, BOWL, Decision lists), and their performance is close to the gold standard. The performance of BOWL is superior when $K = 1$ but degrades significantly when $K = 2, 4, 8$. Because *Q*-learning and decision lists do not require estimating the propensity score $P(a_k|h_k)$, their results do not depend on whether or not the propensity score is known. It is interesting to note that a few methods (e.g. DR-IPCW for $K = 1, 2$) perform better when the propensity score is unknown (such behaviour is anticipated by known results in semi-parametric theory Tsiatis, 2007). Table 2 and 3 show the expected potential outcomes and the percentage of selecting the optimal treatment under various methods, which are consistent with the observation of $R(\widehat{d}_n)$.

Table 4 displays the MSE of the estimates for the expected outcome under the optimal treatment regime when the outcome is continuous. We see that non-parametric methods like *Q*-learning with random forest (QRF) always outperform (semi-)parametric methods. Moreover, the methods perform better when propensity scores $P(a_k|h_k)$ are known.

For binary outcomes, we assume only known propensity scores $P(a_k|h_k) = 0.5$. Results for $R(\widehat{d}_n)$, $\mathbb{E}Y^*(\widehat{d}_n)$, $T(\widehat{d}_n)$ and MSE are presented in Tables 5–8, respectively. Similar to the results for continuous outcomes, we can see that the predictive performance of the non-parametric methods is most often superior to that of the (semi-)parametric methods in terms of $R(\widehat{d}_n)$, $\mathbb{E}Y^*(\widehat{d}_n)$, $T(\widehat{d}_n)$ and MSE.

Combining Tables 1–8, we summarise our findings in Table 9, where the methods are ranked based on their performance. We see that non-parametric methods like Blip, DR-IPCW and QRF have higher rankings compared with (semi-)parametric methods when estimating the optimal treatment regime for both when the propensity scores are known and unknown. For the estimated outcome under the optimal regime, QRF consistently outperforms the other methods.

## 5 Real Data Applications

### 5.1 HIV Data Analysis

In this section, we apply several dynamic treatment regime methods to the ACTG175 data set, which is available through the R package *speff2trial* (Juraska & Juraska, 2010). The data set consists of 2139 patients infected with HIV who were randomised to one of four treatment arms: (i) zidovudine (AZT) monotherapy, (ii) AZT+didanosine (ddI), (iii) AZT+zalcitabine (ddC) and (iv) ddI monotherapy. The outcome measure is the CD4 T-cell count at 96 ± 5 weeks as CD4 count represents a vital signal for disease progression in HIV-infected patients. A basic

Table 1. *The value of $R(\widehat{d}_n)$ (standard error) under the settings in Section 4.1.1 for observational (Obs.) data type with unknown propensity scores and experimental (Exp.) data type with known propensity scores.*

| Data type | Method | $R(\widehat{d}_n)$ ($n = 250$) | | $R(\widehat{d}_n)$ ($n = 500$) | | $R(\widehat{d}_n)$ ($n = 1000$) | |
|---|---|---|---|---|---|---|---|
| | | $K = 1$ | $K = 2$ | $K = 1$ | $K = 2$ | $K = 1$ | $K = 2$ |
| Obs. | QMR | 97.0 (0.1) | 94.0 (0.2) | 97.0 (0.1) | 94.5 (0.1) | 97.1 (0.1) | 95.1 (2.2) |
| | Penalised QMR | 97.0 (0.1) | 94.0 (0.2) | 96.9 (0.1) | 94.6 (0.1) | 96.8 (0.1) | 95.1 (0.1) |
| | QRF | 98.9 (0.1) | 96.2 (0.3) | 99.1 (0.1) | 97.0 (0.2) | 99.1 (0.1) | 97.8 (0.1) |
| | Planning | 97.0 (0.1) | 94.0 (0.2) | 97.0 (0.1) | 94.6 (0.1) | 97.1 (0.1) | 95.2 (0.1) |
| | *A*-learning | 98.4 (0.1) | 96.1 (0.3) | 98.2 (0.1) | 97.1 (0.2) | 98.2 (0.1) | 98.2 (0.1) |
| | *C*-learning | 99.6 (0.1) | 87.5 (0.2) | 99.7 (0.1) | 88.2 (0.1) | **99.8 (0.1)** | 88.8 (0.1) |
| | BOWL | **99.9 (0.1)** | 83.9 (0.1) | **99.8 (0.1)** | 83.8 (0.1) | **99.9 (0.1)** | 84.0 (0.1) |
| | Regret | 94.2 (0.3) | 85.4 (0.2) | 94.0 (0.2) | 85.7 (0.2) | 93.6 (0.1) | 86.0 (0.1) |
| | Blip | **99.9 (0.1)** | **97.1 (0.3)** | **99.9 (0.1)** | **98.0 (0.2)** | **99.9 (0.1)** | **99.0 (0.1)** |
| | DR-IPCW | **99.9 (0.1)** | **97.2 (0.3)** | **99.9 (0.1)** | **98.2 (0.1)** | **99.9 (0.1)** | **98.9 (0.1)** |
| | Decision list | 93.5 (0.1) | 85.8 (0.3) | 93.5 (0.1) | 85.2 (0.1) | 93.4 (0.1) | 84.5 (0.1) |
| Exp. | QMR | 97.0 (0.1) | 94.0 (0.2) | 97.0 (0.1) | 94.5 (0.1) | 97.1 (0.1) | 95.1 (0.1) |
| | Penalised QMR | 97.0 (0.1) | 94.0 (0.2) | 96.9 (0.1) | 94.6 (0.1) | 96.8 (0.1) | 95.1 (0.1) |
| | QRF | 98.9 (0.1) | **96.2 (0.3)** | 99.1 (0.1) | **97.0 (0.2)** | 99.1 (0.1) | 97.8 (0.1) |
| | Planning | 97.0 (0.1) | 94.0 (0.2) | 97.0 (0.1) | 94.6 (0.1) | 97.1 (0.1) | 95.2 (0.1) |
| | *A*-learning | 97.4 (0.1) | 92.8 (0.4) | 97.9 (0.1) | 94.2 (0.2) | 98.5 (0.1) | 95.5 (0.1) |
| | *C*-learning | 98.9 (0.1) | 87.0 (0.2) | 99.1 (0.1) | 87.3 (0.1) | 99.4 (0.1) | 87.8 (0.1) |
| | BOWL | **99.9 (0.1)** | 83.6 (0.1) | **99.8 (0.1)** | 83.6 (0.1) | **99.9 (0.1)** | 83.6 (0.1) |
| | Regret | 95.0 (0.3) | 92.7 (0.4) | 94.9 (0.2) | 93.0 (0.2) | 94.9 (0.1) | 92.7 (0.2) |
| | Blip | 98.5 (0.1) | **95.8 (0.3)** | 99.2 (0.1) | **97.3 (0.2)** | 99.3 (0.1) | **98.1 (0.1)** |
| | DR-IPCW | 98.5 (0.1) | **96.3 (0.3)** | 99.3 (0.1) | **97.4 (0.2)** | 99.5 (0.1) | **98.1 (0.1)** |
| | Decision list | 93.5 (0.1) | 85.8 (0.3) | 93.5 (0.1) | 85.2 (0.2) | 93.4 (0.1) | 84.5 (0.1) |
| | *Q*-learning (corr) | 99.8 (0.1) | 96.4 (0.3) | 99.9 (0.1) | 97.6 (0.1) | 99.9 (0.1) | 98.4 (0.1) |
| | | $K = 4$ | $K = 8$ | $K = 4$ | $K = 8$ | $K = 4$ | $K = 8$ |
| Obs. | QMR | 80.8 (3.1) | 85.0 (2.2) | 84.5 (2.1) | 83.3 (1.4) | 82.5 (1.4) | 80.5 (1.1) |
| | Penalised QMR | 79.7 (3.0) | 82.6 (2.0) | 84.1 (2.1) | 82.3 (1.4) | 82.3 (1.4) | 79.8 (1.1) |
| | QRF | **89.1 (2.9)** | **93.4 (1.9)** | **94.9 (1.9)** | **94.1 (1.1)** | **93.1 (1.3)** | **98.0 (0.9)** |
| | Planning | 76.2 (3.1) | 75.4 (1.9) | 75.7 (2.2) | 71.4 (1.4) | 70.6 (1.5) | 72.4 (1.0) |
| | *A*-learning | 73.3 (3.2) | 63.5 (2.5) | 81.3 (2.2) | 63.0 (1.8) | 80.0 (1.4) | 63.1 (1.3) |
| | *C*-learning | 83.7 (3.1) | 85.8 (1.9) | 88.9 (2.1) | 88.7 (1.3) | 85.3 (1.5) | 90.2 (0.9) |
| | BOWL | 63.6 (2.0) | 58.3 (1.2) | 63.0 (1.4) | 58.7 (0.8) | 63.3 (1.0) | 58.8 (0.6) |
| | Regret | 62.7 (3.2) | 67.3 (2.1) | 66.8 (2.2) | 65.4 (1.4) | 64.8 (1.5) | 63.9 (1.1) |
| | Blip | 82.3 (0.2) | 80.9 (0.2) | 85.7 (0.1) | 83.8 (0.1) | 90.9 (0.1) | 86.7 (0.1) |
| | DR-IPCW | 82.4 (0.2) | 81.0 (0.2) | 85.9 (0.1) | 83.8 (0.1) | 91.0 (0.1) | 86.6 (0.1) |
| | Decision list | 74.8 (2.8) | 79.6 (1.9) | 71.6 (1.9) | 78.7 (1.5) | 69.8 (1.4) | 75.2 (1.1) |
| Exp. | QMR | 80.8 (3.1) | 85.0 (2.2) | 84.5 (2.1) | 83.3 (1.4) | 82.5 (1.4) | 80.5 (1.1) |
| | Penalised QMR | 79.7 (3.0) | 82.6 (2.0) | 84.1 (2.1) | 82.3 (1.4) | 82.3 (1.4) | 79.8 (1.1) |
| | QRF | **89.1 (2.9)** | **93.4 (1.9)** | **94.9 (1.9)** | **94.1 (1.1)** | **93.1 (1.3)** | **98.0 (0.9)** |
| | Planning | 76.2 (3.1) | 75.4 (1.9) | 75.7 (2.2) | 71.4 (1.4) | 70.6 (1.5) | 72.4 (1.0) |
| | *A*-learning | 72.3 (3.3) | 65.2 (2.5) | 79.2 (2.2) | 63.4 (1.8) | 80.3 (1.5) | 64.0 (1.3) |
| | *C*-learning | 85.1 (3.0) | 84.8 (1.9) | 85.5 (2.1) | 86.6 (1.4) | **91.9 (1.5)** | 89.5 (1.0) |
| | BOWL | 63.6 (2.2) | 55.9 (1.4) | 63.2 (1.5) | 55.4 (1.0) | 65.0 (1.1) | 56.2 (0.7) |
| | Regret | 62.7 (3.2) | 67.3 (2.1) | 66.8 (2.2) | 65.4 (1.4) | 64.8 (1.5) | 63.9 (1.1) |
| | Blip | 82.4 (0.3) | 81.7 (0.2) | 85.9 (0.1) | 83.8 (0.1) | 91.0 (0.1) | 86.6 (0.1) |
| | DR-IPCW | 82.5 (0.3) | 81.8 (0.2) | 85.8 (0.1) | 83.9 (0.1) | 91.1 (0.1) | 86.5 (0.1) |
| | Decision list | 74.8 (2.8) | 79.6 (1.9) | 71.6 (1.9) | 78.7 (1.5) | 69.8 (1.4) | 75.2 (1.1) |
| | *Q*-learning (corr) | 99.6 (2.8) | 99.9 (1.8) | 99.9 (1.9) | 99.3 (1.2) | 99.6 (1.3) | 97.8 (0.9) |

'corr' represents 'correctly specified model'. Standard errors of the estimators are shown in parenthesis. The boldfaced-values represent the best results among different methods.

Table 2. *The potential outcomes under the settings in Section 4.1.1 for observational (Obs.) data type with unknown propensity scores and experimental (Exp.) data type with known propensity scores.*

| Data type | Method | $\mathbb{E}Y^*(\widehat{d}_n)$ ($n = 250$) | | $\mathbb{E}Y^*(\widehat{d}_n)$ ($n = 500$) | | $\mathbb{E}Y^*(\widehat{d}_n)$ ($n = 1000$) | |
|---|---|---|---|---|---|---|---|
| | | $K = 1$ | $K = 2$ | $K = 1$ | $K = 2$ | $K = 1$ | $K = 2$ |
| Obs. | QMR | 9.7 (0.2) | 15.2 (0.5) | 9.7 (0.1) | 15.3 (0.5) | 9.7 (0.1) | 15.4 (0.4) |
| | Penalised QMR | 9.7 (0.2) | 15.2 (0.5) | 9.7 (0.1) | 15.3 (0.5) | 9.7 (0.1) | 15.4 (0.4) |
| | QRF | 9.9 (0.1) | 15.6 (0.7) | 9.9 (0.1) | 15.7 (0.5) | 9.9 (0.1) | 15.9 (0.5) |
| | Planning | 9.7 (0.2) | 15.3 (0.5) | 9.7 (0.1) | 15.3 (0.4) | 9.7 (0.1) | 15.4 (0.3) |
| | *A*-learning | 9.8 (0.1) | 15.6 (0.7) | 9.8 (0.1) | 15.7 (0.6) | 9.8 (0.1) | 15.9 (0.5) |
| | *C*-learning | 10.0 (0.1) | 14.2 (0.4) | 10.0 (0.1) | 14.3 (0.4) | **10.0 (0.1)** | 14.4 (0.4) |
| | BOWL | **10.0 (0.1)** | 13.6 (0.3) | **10.0 (0.1)** | 13.6 (0.3) | **10.0 (0.1)** | 13.6 (0.3) |
| | Regret | 9.4 (0.4) | 13.8 (0.6) | 9.4 (0.4) | 13.9 (0.6) | 9.4 (0.4) | 14.0 (0.6) |
| | Blip | **10.0 (0.1)** | **15.8 (0.6)** | **10.0 (0.1)** | **15.9 (0.6)** | **10.0 (0.1)** | **16.1 (0.4)** |
| | DR-IPCW | **10.0 (0.1)** | **15.8 (0.7)** | **10.0 (0.1)** | **15.9 (0.5)** | **10.0 (0.1)** | **16.1 (0.4)** |
| | Decision list | 9.4 (0.1) | 13.9 (0.6) | 9.4 (0.1) | 13.8 (0.5) | 9.4 (0.1) | 13.7 (0.4) |
| Exp. | QMR | 9.7 (0.2) | 15.2 (0.5) | 9.7 (0.1) | 15.3 (0.5) | 9.7 (0.1) | 15.4 (0.4) |
| | Penalised QMR | 9.7 (0.2) | 15.2 (0.5) | 9.7 (0.1) | 15.3 (0.5) | 9.7 (0.1) | 15.4 (0.4) |
| | QRF | 9.9 (0.1) | **15.6 (0.7)** | 9.9 (0.1) | **15.7 (0.5)** | 9.9 (0.1) | 15.9 (0.5) |
| | Planning | 9.7 (0.2) | 15.3 (0.5) | 9.7 (0.1) | 15.3 (0.4) | 9.7 (0.1) | 15.4 (0.3) |
| | *A*-learning | 9.8 (0.2) | 15.1 (0.9) | 9.8 (0.2) | 15.3 (0.9) | 9.9 (0.2) | 15.5 (0.7) |
| | *C*-learning | 9.9 (0.2) | 14.1 (0.4) | 9.9 (0.2) | 14.2 (0.4) | 9.9 (0.2) | 14.2 (0.4) |
| | BOWL | **10.0 (0.1)** | 13.6 (0.2) | **10.0 (0.1)** | 13.6 (0.2) | **10.0 (0.1)** | 13.6 (0.3) |
| | Regret | 9.5 (0.4) | 15.0 (0.9) | 9.5 (0.4) | 15.1 (0.9) | 9.5 (0.4) | 15.0 (0.9) |
| | Blip | 9.9 (0.2) | **15.5 (0.7)** | 9.9 (0.2) | **15.8 (0.6)** | 9.9 (0.2) | **15.9 (0.5)** |
| | DR-IPCW | 9.9 (0.2) | **15.6 (0.7)** | 9.9 (0.2) | **15.8 (0.6)** | **10.0 (0.1)** | **15.9 (0.5)** |
| | Decision list | 9.4 (0.1) | 13.9 (0.6) | 9.4 (0.1) | 13.8 (0.5) | 9.4 (0.1) | 13.7 (0.4) |
| | *Q*-learning (corr) | 10.0 (0.1) | 15.6 (0.7) | 10.0 (0.1) | 15.8 (0.5) | 10.0 (0.1) | 16.0 (0.4) |
| | | $K = 4$ | $K = 8$ | $K = 4$ | $K = 8$ | $K = 4$ | $K = 8$ |
| Obs. | QMR | 4.2 (2.5) | 8.7 (3.5) | 4.4 (2.4) | 8.5 (3.3) | 4.3 (2.3) | 8.2 (3.5) |
| | Penalised QMR | 4.1 (2.5) | 8.4 (3.3) | 4.3 (2.4) | 8.4 (3.2) | 4.2 (2.3) | 8.1 (3.4) |
| | QRF | **4.6 (2.4)** | **9.5 (3.0)** | **4.9 (2.2)** | **9.6 (2.6)** | **4.8 (2.2)** | **10.0 (2.9)** |
| | Planning | 3.9 (2.5) | 7.7 (3.1) | 3.9 (2.6) | 7.3 (3.2) | 3.6 (2.4) | 7.4 (3.3) |
| | *A*-learning | 3.8 (2.6) | 6.5 (4.1) | 4.2 (2.5) | 6.4 (4.1) | 4.1 (2.3) | 6.4 (4.3) |
| | *C*-learning | 4.3 (2.5) | 8.8 (3.0) | 4.6 (2.4) | 9.1 (3.1) | 4.4 (2.5) | 9.2 (3.0) |
| | BOWL | 3.3 (1.6) | 6.0 (2.0) | 3.2 (1.6) | 6.0 (1.9) | 3.3 (1.7) | 6.0 (2.1) |
| | Regret | 3.2 (2.6) | 6.9 (3.4) | 3.4 (2.5) | 6.7 (3.3) | 3.3 (2.5) | 6.5 (3.5) |
| | Blip | 4.2 (0.2) | 8.3 (0.3) | 4.4 (0.2) | 8.6 (0.2) | 4.7 (0.1) | 8.9 (0.2) |
| | DR-IPCW | 4.2 (0.2) | 8.3 (0.2) | 4.4 (0.2) | 8.6 (0.2) | 4.7 (0.1) | 8.9 (0.2) |
| | Decision list | 3.9 (2.3) | 8.1 (3.1) | 3.7 (2.2) | 8.0 (3.3) | 3.6 (2.4) | 7.7 (3.6) |
| Exp. | QMR | 4.2 (2.5) | 8.7 (3.5) | 4.4 (2.4) | 8.5 (3.3) | 4.3 (2.3) | 8.2 (3.5) |
| | Penalised QMR | 4.1 (2.5) | 8.4 (3.3) | 4.3 (2.4) | 8.4 (3.2) | 4.2 (2.3) | 8.1 (3.4) |
| | QRF | **4.6 (2.4)** | **9.5 (3.0)** | **4.9 (2.2)** | **9.6 (2.6)** | **4.8 (2.2)** | **10.0 (2.9)** |
| | Planning | 3.9 (2.5) | 7.7 (3.1) | 3.9 (2.6) | 7.3 (3.2) | 3.6 (2.4) | 7.4 (3.3) |
| | *A*-learning | 3.7 (2.7) | 6.7 (4.0) | 4.1 (2.6) | 6.5 (4.1) | 4.1 (2.4) | 6.5 (4.1) |
| | *C*-learning | 4.4 (2.4) | 8.7 (3.0) | 4.4 (2.4) | 8.8 (3.2) | **4.7 (2.4)** | 9.1 (3.2) |
| | BOWL | 3.3 (1.8) | 5.7 (2.3) | 3.3 (1.7) | 5.6 (2.4) | 3.3 (1.8) | 5.7 (2.4) |
| | Regret | 3.2 (2.6) | 6.9 (3.4) | 3.4 (2.5) | 6.7 (3.3) | 3.3 (2.5) | 6.5 (3.5) |
| | Blip | 4.2 (0.2) | 8.3 (0.3) | 4.4 (0.2) | 8.6 (0.2) | 4.7 (0.1) | 8.8 (0.2) |
| | DR-IPCW | 4.2 (0.2) | 8.3 (0.2) | 4.4 (0.2) | 8.6 (0.2) | 4.7 (0.1) | 8.8 (0.2) |
| | Decision list | 3.9 (2.3) | 8.1 (3.1) | 3.7 (2.2) | 8.0 (3.3) | 3.6 (2.4) | 7.7 (3.6) |
| | *Q*-learning (corr) | 5.1 (2.3) | 10.3 (2.8) | 5.3 (2.2) | 10.1 (2.7) | 5.1 (2.2) | 10.0 (2.9) |

'corr' represents 'correctly specified model'. Standard errors of the potential outcomes are shown in parenthesis. The boldfaced-values represent the best results among different methods.

conclusion from the study is that for patients who had taken AZT before entering the trial, treatment with ddI or AZT+ddI were better than continuing to take AZT alone (see Hammer *et al.*, 1996, for more information about the study). In this analysis we expand on this

Table 3. *The percentage of correct decisions (standard error) under the settings in Section 4.1.1 for observational (Obs.) data type with unknown propensity scores and experimental (Exp.) data type with known propensity scores.*

| Data type | Method | $T(\widehat{d}_n)$ ($n = 250$) | | $T(\widehat{d}_n)$ ($n = 500$) | | $T(\widehat{d}_n)$ ($n = 1000$) | |
|---|---|---|---|---|---|---|---|
| | | $K = 1$ | $K = 2$ | $K = 1$ | $K = 2$ | $K = 1$ | $K = 2$ |
| Obs. | QMR | 80.1 (0.5) | 68.6 (0.4) | 80.6 (0.3) | 69.3 (0.3) | 80.5 (0.1) | 70.2 (0.1) |
| | Penalised QMR | 79.9 (0.5) | 68.7 (0.4) | 80.0 (0.3) | 69.5 (0.2) | 80.1 (0.1) | 70.1 (0.1) |
| | QRF | 92.1 (0.2) | 80.5 (0.9) | 92.8 (0.1) | 83.2 (0.6) | 92.8 (0.1) | 85.4 (0.4) |
| | Planning | 80.1 (0.5) | 68.5 (0.4) | 80.6 (0.3) | 69.3 (0.2) | 80.5 (0.1) | 70.1 (0.1) |
| | *A*-learning | 70.2 (0.6) | 77.4 (1.0) | 68.4 (0.3) | 80.9 (0.7) | 67.8 (0.1) | 84.7 (0.4) |
| | *C*-learning | 95.6 (0.2) | 53.5 (0.7) | 96.6 (0.1) | 56.1 (0.5) | **97.1 (0.1)** | 59.1 (0.3) |
| | BOWL | **97.6 (0.1)** | 33.2 (0.3) | **97.6 (0.1)** | 33.0 (0.2) | **97.6 (0.1)** | 33.7 (0.2) |
| | Regret | 50.8 (2.3) | 37.5 (0.6) | 50.0 (1.5) | 38.3 (0.4) | 47.7 (1.4) | 38.1 (0.2) |
| | Blip | **97.6 (0.1)** | **84.5 (0.8)** | **97.5 (0.1)** | **88.2 (0.5)** | **97.7 (0.1)** | **91.4 (0.3)** |
| | DR-IPCW | **97.6 (0.1)** | **84.2 (0.9)** | **97.5 (0.1)** | **88.7 (0.5)** | **97.7 (0.1)** | **91.7 (0.3)** |
| | Decision list | 27.0 (0.2) | 31.8 (0.5) | 27.0 (0.1) | 30.4 (0.3) | 27.0 (0.1) | 28.7 (0.1) |
| Exp. | QMR | 80.1 (0.5) | 68.6 (0.4) | 80.6 (0.3) | 69.3 (0.3) | 80.5 (0.1) | 70.2 (0.1) |
| | Penalised QMR | 79.9 (0.5) | 68.7 (0.4) | 80.0 (0.3) | 69.5 (0.2) | 80.1 (0.1) | 70.1 (0.1) |
| | QRF | 92.1 (0.2) | **80.5 (0.9)** | 92.8 (0.1) | **83.2 (0.6)** | 92.8 (0.1) | 85.4 (0.4) |
| | Planning | 80.1 (0.5) | 68.5 (0.4) | 80.6 (0.3) | 69.3 (0.2) | 80.5 (0.1) | 70.1 (0.1) |
| | *A*-learning | 72.4 (1.0) | 61.6 (1.3) | 76.6 (0.7) | 67.1 (0.9) | 81.0 (0.4) | 72.1 (0.6) |
| | *C*-learning | 89.4 (0.9) | 47.8 (0.8) | 91.5 (0.6) | 50.8 (0.5) | 93.3 (0.3) | 53.4 (0.3) |
| | BOWL | **97.6 (0.1)** | 32.5 (0.3) | **97.6 (0.1)** | 32.5 (0.2) | **97.6 (0.1)** | 32.9 (0.1) |
| | Regret | 51.5 (2.3) | 70.8 (1.1) | 50.9 (1.7) | 72.0 (0.8) | 50.6 (1.2) | 71.1 (0.5) |
| | Blip | 86.8 (0.9) | 76.7 (1.0) | 91.2 (0.6) | **83.8 (0.6)** | 93.6 (0.4) | **87.2 (0.4)** |
| | DR-IPCW | 87.6 (0.7) | 77.6 (1.0) | 92.2 (0.4) | **83.7 (0.9)** | 94.8 (0.2) | **87.8 (0.4)** |
| | Decision list | 27.0 (0.2) | 31.8 (0.5) | 27.0 (0.1) | 30.4 (0.3) | 27.0 (0.1) | 28.7 (0.1) |
| | *Q*-learning (corr) | 95.7 (0.3) | 79.0 (1.0) | 97.3 (0.1) | 83.5 (0.6) | 98.1 (0.1) | 87.3 (0.4) |
| | | $K = 4$ | $K = 8$ | $K = 4$ | $K = 8$ | $K = 4$ | $K = 8$ |
| Obs. | QMR | 66.9 (0.3) | 70.6 (0.4) | 67.7 (0.2) | 69.3 (0.2) | 68.1 (0.1) | 67.3 (0.1) |
| | Penalised QMR | 65.4 (0.3) | 64.1 (0.4) | 66.8 (0.2) | 65.4 (0.2) | 67.7 (0.1) | 65.3 (0.1) |
| | QRF | **85.1 (0.3)** | **86.2 (0.2)** | **90.0 (0.1)** | **90.7 (0.1)** | **93.3 (0.1)** | **93.9 (0.1)** |
| | Planning | 60.4 (0.1) | 57.8 (0.1) | 60.6 (0.1) | 57.9 (0.1) | 60.7 (0.1) | 57.9 (0.1) |
| | *A*-learning | 61.9 (0.6) | 50.2 (0.4) | 65.7 (0.3) | 52.3 (0.3) | 67.1 (0.2) | 54.5 (0.2) |
| | *C*-learning | 73.6 (0.3) | 72.0 (0.3) | 76.5 (0.2) | 76.3 (0.2) | 78.7 (0.1) | 79.5 (0.1) |
| | BOWL | 46.1 (0.1) | 36.9 (0.2) | 46.3 (0.1) | 37.9 (0.1) | 46.5 (0.1) | 38.6 (0.1) |
| | Regret | 49.8 (0.7) | 50.1 (0.6) | 49.8 (0.5) | 50.1 (0.4) | 49.7 (0.3) | 50.1 (0.3) |
| | Blip | 72.0 (0.4) | 66.8 (0.3) | 76.5 (0.2) | 70.7 (0.2) | 84.2 (0.2) | 75.0 (0.2) |
| | DR-IPCW | 71.8 (0.4) | 67.3 (0.3) | 77.5 (0.2) | 70.9 (0.2) | 83.7 (0.2) | 74.6 (0.2) |
| | Decision list | 58.8 (0.3) | 57.1 (0.4) | 57.2 (0.3) | 58.5 (0.3) | 55.2 (0.2) | 59.5 (0.2) |
| Exp. | QMR | 66.9 (0.3) | 70.6 (0.4) | 67.7 (0.2) | 69.3 (0.2) | 68.1 (0.1) | 67.3 (0.1) |
| | Penalised QMR | 65.4 (0.3) | 64.1 (0.4) | 66.8 (0.2) | 65.4 (0.2) | 67.7 (0.1) | 65.3 (0.1) |
| | QRF | **85.1 (0.3)** | **86.2 (0.2)** | **90.0 (0.1)** | **90.7 (0.1)** | **93.3 (0.1)** | **93.9 (0.1)** |
| | Planning | 60.4 (0.1) | 57.8 (0.1) | 60.6 (0.1) | 57.9 (0.1) | 60.7 (0.1) | 57.9 (0.1) |
| | *A*-learning | 59.8 (0.6) | 50.5 (0.4) | 64.3 (0.3) | 52.0 (0.3) | 67.4 (0.2) | 53.9 (0.2) |
| | *C*-learning | 73.5 (0.3) | 71.8 (0.3) | 76.6 (0.2) | 76.3 (0.2) | 78.7 (0.1) | 79.6 (0.1) |
| | BOWL | 36.5 (0.3) | 32.5 (0.3) | 37.3 (0.2) | 32.5 (0.2) | 38.1 (0.2) | 32.9 (0.2) |
| | Regret | 49.8 (0.7) | 50.1 (0.6) | 49.8 (0.5) | 50.1 (0.4) | 49.7 (0.3) | 50.1 (0.3) |
| | Blip | 72.1 (0.4) | 67.8 (0.3) | 76.9 (0.2) | 70.4 (0.2) | 84.5 (0.2) | 74.7 (0.2) |
| | DR-IPCW | 71.7 (0.4) | 68.2 (0.3) | 77.4 (0.2) | 71.0 (0.2) | 84.1 (0.2) | 75.2 (0.2) |
| | Decision list | 58.8 (0.3) | 57.1 (0.4) | 57.2 (0.3) | 58.5 (0.3) | 55.2 (0.2) | 59.5 (0.2) |
| | *Q*-learning (corr) | 100 (0) | 100 (0) | 100 (0) | 100 (0) | 100 (0) | 100 (0) |

'corr' represents 'correctly specified model'. Standard errors of the estimators are shown in parentheses. The boldfaced-values represent the best results among different methods.

conclusion seeking to determine the optimal assignment of these two treatments for the subset of patients that has previously taken AZT. Specifically, we select $n = 532$ patients who had taken AZT before the study and received AZT+ddI or ddI monotherapy in the study and for

Table 4. *The MSE under the simulation settings in Section 4.1.1 for observational (Obs.) data type with unknown propensity scores and experimental (Exp.) data type with known propensity scores.*

| Data type | Method | MSE ($n = 250$) | | MSE ($n = 500$) | | MSE ($n = 1000$) | |
|---|---|---|---|---|---|---|---|
| | | $K = 1$ | $K = 2$ | $K = 1$ | $K = 2$ | $K = 1$ | $K = 2$ |
| Obs. | QMR | 1.93 (0.08) | 41.75 (2.28) | 1.51 (0.04) | 33.64 (1.49) | 1.53 (0.04) | 28.92 (0.93) |
| | QRF | **0.28 (0.02)** | **1.58 (0.12)** | **0.15 (0.01)** | **0.80 (0.06)** | **0.08 (0.01)** | **0.45 (0.03)** |
| | *A*-learning | 18.77 (0.85) | 66.61 (5.67) | 16.50 (0.49) | 38.75 (2.64) | 16.48 (0.35) | 27.46 (1.71) |
| | Regret | 2.80 (0.18) | 11.03 (0.53) | 2.61 (0.15) | 10.57 (0.40) | 2.41 (0.14) | 10.68 (0.38) |
| | TMLE | 1.64 (0.18) | 9.19 (0.39) | 0.91 (0.03) | 5.19 (0.22) | 0.80 (0.02) | 4.25 (0.14) |
| | IPTW | 1.93 (0.20) | 10.37 (0.48) | 1.15 (0.04) | 5.43 (0.23) | 1.12 (0.03) | 4.46 (0.15) |
| Exp. | QMR | 1.93 (0.08) | 41.75 (2.28) | 1.51 (0.04) | 33.64 (1.49) | 1.53 (0.04) | 28.92 (0.93) |
| | QRF | **0.28 (0.02)** | **1.58 (0.12)** | **0.15 (0.01)** | **0.80 (0.06)** | **0.08 (0.01)** | **0.45 (0.03)** |
| | *A*-learning | 2.09 (0.28) | 62.28 (5.46) | 1.18 (0.14) | 22.14 (2.40) | 0.51 (0.05) | 9.06 (0.72) |
| | Regret | 25.71 (0.27) | 51.54 (1.14) | 25.94 (0.20) | 56.39 (1.18) | 26.38 (0.15) | 31.48 (0.34) |
| | TMLE | 1.44 (0.07) | 8.71 (0.48) | 0.85 (0.04) | 5.11 (0.21) | 0.73 (0.02) | 4.35 (0.18) |
| | IPTW | 1.87 (0.10) | 9.91 (0.59) | 1.10 (0.05) | 6.03 (0.28) | 0.98 (0.03) | 5.66 (0.35) |
| | *Q*-learning (corr) | 0.22 (0.01) | 2.92 (0.19) | 2.61 (0.15) | 1.51 (0.09) | 2.41 (0.14) | 0.69 (0.04) |

'corr' represents 'correctly specified model'. The standard errors are shown in parentheses. The boldfaced-values indicate the best results among different methods.

Table 5. *The value of $R(\widehat{d}_n)$ (standard error) under the settings in Section 4.1.2.*

| Method | $R(\widehat{d}_n)$ ($n = 250$) | | $R(\widehat{d}_n)$ ($n = 500$) | | $R(\widehat{d}_n)$ ($n = 1000$) | |
|---|---|---|---|---|---|---|
| | $K = 1$ | $K = 2$ | $K = 1$ | $K = 2$ | $K = 1$ | $K = 2$ |
| QMR | 85.1 (0.2) | 93.2 (0.3) | 85.8 (0.1) | 95.4 (0.1) | 86.4 (0.1) | 96.8 (0.1) |
| Penalised QMR | 80.0 (0.2) | 92.5 (0.3) | 79.4 (0.1) | 94.7 (0.1) | 78.7 (0.1) | 96.2 (0.1) |
| QRF | **92.1 (0.2)** | 89.9 (0.3) | **94.2 (0.1)** | 93.2 (0.1) | **96.0 (0.1)** | 95.7 (0.1) |
| Planning | 85.1 (0.2) | 93.2 (0.3) | 85.8 (0.1) | 95.4 (0.1) | 86.4 (0.1) | 96.8 (0.1) |
| *A*-learning | 85.0 (0.2) | 88.9 (0.4) | 85.7 (0.1) | 90.8 (0.2) | 86.3 (0.1) | 92.6 (0.1) |
| *C*-learning | 87.4 (0.2) | 92.4 (0.3) | 88.6 (0.1) | 94.6 (0.1) | 89.8 (0.1) | 96.3 (0.1) |
| BOWL | 79.5 (0.2) | **94.9 (0.4)** | 78.9 (0.1) | **96.9 (0.1)** | 78.3 (0.1) | **97.8 (0.1)** |
| Regret | 82.6 (0.2) | 82.1 (0.1) | 82.7 (0.1) | 82.1 (0.1) | 82.7 (0.1) | 82.2 (0.1) |
| Blip | 88.2 (0.2) | 90.7 (0.4) | 91.2 (0.1) | 93.3 (0.2) | 95.3 (0.1) | 96.0 (0.1) |
| DR-IPCW | 88.2 (0.2) | 91.2 (0.4) | 91.3 (0.1) | 94.1 (0.2) | 95.4 (0.1) | 96.7 (0.1) |
| Decision list | 84.2 (0.2) | 80.5 (0.1) | 85.1 (0.1) | 80.4 (0.1) | 84.8 (0.1) | 80.5 (0.1) |

The boldfaced-values indicate the best results among different methods.

whom full CD4 and the selected covariates are available. This problem can be formulated as a one-decision-point problem for which the treatment indicator $A_i$ is set to be 1 if patient $i$ is assigned to AZT+ddI therapy and 0 if the patient is assigned to ddI monotherapy. Because the trial is randomised with equal randomisation for both therapies, the propensity score equals 0.5.

For illustrative purposes, we consider only two covariates in the estimation of the optimal treatment regimes: the baseline body weight $S_1$ and the baseline CD4 T-cell count $S_2$. We include the body weight in our analysis because it has been observed that body weight has a significant role on AZT pharmacokinetic profile. Burger *et al.* (1994) reported that AZT clearance

Table 6. *The potential outcomes under the settings in Section 4.1.2.*

| | $R(\widehat{d}_n)$ $(n = 250)$ | | $R(\widehat{d}_n)$ $(n = 500)$ | | $R(\widehat{d}_n)$ $(n = 1000)$ | |
|---|---|---|---|---|---|---|
| Method | $K = 1$ | $K = 2$ | $K = 1$ | $K = 2$ | $K = 1$ | $K = 2$ |
| QMR | 0.48 (0.01) | 0.45 (0.02) | 0.48 (0.02) | 0.46 (0.01) | 0.48 (0.02) | 0.47 (0.01) |
| Penalised QMR | 0.45 (0.02) | 0.45 (0.02) | 0.45 (0.02) | 0.46 (0.02) | 0.44 (0.02) | 0.47 (0.01) |
| QRF | **0.52 (0.01)** | 0.44 (0.02) | **0.53 (0.01)** | 0.45 (0.02) | **0.54 (0.01)** | 0.47 (0.01) |
| Planning | 0.48 (0.01) | 0.45 (0.02) | 0.48 (0.02) | 0.46 (0.01) | 0.48 (0.02) | 0.47 (0.01) |
| *A*-learning | 0.48 (0.02) | 0.43 (0.03) | 0.48 (0.02) | 0.44 (0.02) | 0.48 (0.02) | 0.45 (0.02) |
| *C*-learning | 0.49 (0.02) | 0.45 (0.02) | 0.50 (0.01) | 0.46 (0.02) | 0.50 (0.01) | 0.47 (0.01) |
| BOWL | 0.45 (0.01) | **0.46 (0.03)** | 0.44 (0.01) | **0.47 (0.01)** | 0.44 (0.01) | **0.48 (0.01)** |
| Regret | 0.46 (0.02) | 0.40 (0.01) | 0.46 (0.02) | 0.40 (0.01) | 0.46 (0.02) | 0.40 (0.01) |
| Blip | 0.49 (0.02) | 0.44 (0.03) | 0.51 (0.02) | 0.45 (0.02) | 0.53 (0.01) | 0.47 (0.01) |
| DR-IPCW | 0.49 (0.02) | 0.45 (0.03) | 0.51 (0.02) | 0.45 (0.02) | 0.53 (0.01) | 0.48 (0.01) |
| Decision list | 0.47 (0.02) | 0.39 (0.01) | 0.48 (0.01) | 0.39 (0.01) | 0.48 (0.01) | 0.39 (0.01) |

Table 7. *The percentage of correct decisions (standard error) under the settings in Section 4.1.2.*

| | $T(\widehat{d}_n)$ $(n = 250)$ | | $T(\widehat{d}_n)$ $(n = 500)$ | | $T(\widehat{d}_n)$ $(n = 1000)$ | |
|---|---|---|---|---|---|---|
| Method | $K = 1$ | $K = 2$ | $K = 1$ | $K = 2$ | $K = 1$ | $K = 2$ |
| QMR | 54.2 (0.3) | 69.8 (0.5) | 55.4 (0.2) | 73.9 (0.3) | 56.4 (0.1) | 77.1 (0.2) |
| Penalised QMR | 49.5 (1.3) | 68.1 (0.5) | 49.5 (0.8) | 71.3 (0.3) | 50.5 (0.4) | 74.0 (0.2) |
| QRF | **68.6 (0.3)** | 64.7 (0.5) | **73.3 (0.2)** | 69.6 (0.3) | **77.5 (0.1)** | 74.2 (0.1) |
| Planning | 54.2 (0.3) | 69.5 (0.5) | 55.4 (0.2) | 73.8 (0.3) | 56.4 (0.1) | 77.0 (0.2) |
| *A*-learning | 53.9 (0.3) | 62.1 (0.6) | 55.1 (0.2) | 64.9 (0.3) | 56.2 (0.1) | 67.5 (0.2) |
| *C*-learning | 57.1 (0.3) | 69.9 (0.5) | 59.3 (0.2) | 73.5 (0.3) | 61.6 (0.1) | 77.0 (0.2) |
| BOWL | 43.8 (0.2) | **75.4 (0.6)** | 42.6 (2.6) | **78.8 (0.2)** | 41.4 (0.1) | **80.7 (0.1)** |
| Regret | 50.0 (0.3) | 59.2 (0.1) | 49.8 (0.2) | 59.2 (0.5) | 50.0 (0.1) | 59.1 (0.1) |
| Blip | 60.8 (0.4) | 65.2 (0.6) | 67.2 (0.3) | 69.3 (0.3) | 77.9 (0.2) | 74.4 (0.2) |
| DR-IPCW | 60.8 (0.4) | 65.4 (0.6) | 67.2 (0.3) | 69.3 (0.3) | 77.9 (0.2) | 74.6 (0.2) |
| Decision list | 51.0 (0.2) | 50.1 (0.1) | 51.4 (0.1) | 50.0 (0.1) | 51.1 (0.1) | 50.1 (0.1) |

is significantly lower in patients with a lower body weight, which indicates a qualitative interaction of body weight with AZT.[2]

We apply the methods discussed previously in Section 4 to estimate the optimal treatment regime. Moreover, we provide IPW estimators given by van der Laan & Rose (2018) for the mean outcome and associated 95% confidence intervals under each estimated policy. For training and testing purposes, we randomly split the data set into a training set with 300 patients and a testing set with 232 patients. Then we estimate the optimal treatment regime using the training set and give IPW estimators and confidence intervals for the mean outcome under the estimated optimal treatment regime, for which the results are provided in Table 10. It can be seen that the estimators of the mean outcome under the optimal treatment regime estimated by QMR and *A*-learning are the largest; however, the confidence intervals do not show a significant difference among all the methods. This result implies a larger sample size is needed to test which methods outperform others. Because we posit the same parametric model for QMR and *A*-learning, the estimators are the same for these two methods.

Recall that non-parametric methods such as QRF, Blip and DR-IPCW do not provide explicit forms for the estimated optimal treatment regime, whereas the parametric methods do. The interpretability of decision rules obtained from parametric methods are seen by many researchers as an advantage over the black-box decision rules generated by non-parametric methods, especially in the field of precision medicine. Although explicit forms cannot generally be extracted

Table 8. *The MSE under the settings in Section 4.1.2. The standard errors are shown in parentheses.*

| | MSE ($n = 250$) | | MSE ($n = 500$) | | MSE ($n = 1000$) | |
|---|---|---|---|---|---|---|
| Method | $K = 1$ | $K = 2$ | $K = 1$ | $K = 2$ | $K = 1$ | $K = 2$ |
| QMR | **0.003** | 0.004 | 0.004 | 0.001 | 0.004 | 0.0006 |
| | **(0.0002)** | (0.0002) | (0.0001) | (0.0001) | (0.0001) | (0.0001) |
| QRF | **0.003** | **0.001** | **0.002** | **0.0006** | **0.002** | **0.0004** |
| | **(0.0001)** | **(0.0001)** | **(0.0001)** | **(0.0001)** | **(0.0001)** | **(0.0001)** |
| *A*-learning | **0.003** | 0.038 | 0.004 | 0.014 (0.001) | 0.005 | 0.005 |
| | **(0.0002)** | (0.0002) | (0.0001) | | (0.0001) | (0.0003) |
| Regret | 0.011 | 0.012 | 0.013 | 0.012 | 0.014 | 0.010 |
| | (0.0003) | (0.0005) | (0.0002) | (0.0004) | (0.0002) | (0.0004) |
| TMLE | 0.013 | 0.040 | 0.006 | 0.017 (0.001) | 0.007 | 0.003 |
| | (0.0005) | (0.002) | (0.0003) | | (0.0001) | (0.0001) |
| IPTW | 0.018 | 0.042 | 0.008 | 0.018 (0.001) | 0.009 | 0.003 |
| | (0.0006) | (0.002) | (0.0003) | | (0.0001) | (0.0002) |

The bold-faced-values indicate the best results among different methods.

Table 9. *Average ranks of the methods in Tables A1–A8 based on $R(\widehat{d}_n)$ and MSE for observational (Obs.) data type with unknown propensity scores and experimental (Exp.) data type with known propensity scores.*

| $R(\widehat{d}_n)$ | | | MSE | | |
|---|---|---|---|---|---|
| Method | Obs. | Exp. | Method | Obs. | Exp. |
| QMR | 4.8 | 5.8 | QMR | 3.2 | 4.7 |
| QRF | 4.1 | 3.7 | QRF | 1.0 | 1.0 |
| Planning | 5.1 | 6.0 | *A*-learning | 4.5 | 4.2 |
| *A*-learning | 6.1 | 5.3 | Regret | 4.6 | 5.8 |
| *C*-learning | 4.5 | 6.0 | TMLE | 3.2 | 2.2 |
| BOWL | 5.5 | 5.3 | IPTW | 4.1 | 3.2 |
| Regret | 8.8 | 8.5 | | | |
| Blip | 2.5 | 2.0 | | | |
| DR-IPCW | 2.7 | 2.0 | | | |
| DL | 8.9 | 9.2 | | | |

Here, we only consider the cases when $K = 1, 2$.

Table 10. *The estimated optimal treatment regime $(\widehat{d}^{\text{opt}})$, mean expected outcomes under the estimated optimal treatment regime $(\widehat{\mathbb{E}_{d^{\text{opt}}}Y})$, and associated 95% confidence interval (95% CI) on $\widehat{\mathbb{E}_{d^{\text{opt}}}Y}$ for the HIV data.*

| Method | $\widehat{d}^{\text{opt}}$ | $\widehat{\mathbb{E}_{d^{\text{opt}}}Y}$ | 95% CI |
|---|---|---|---|
| QMR | $I(24.0 + 1.7S_1 - 0.4S_2 > 0)$ | 367.9 | 309.0, 426.7 |
| QRF | NA* | 347.3 | 290.3, 404.3 |
| *A*-learning | $I(24.0 + 1.7S_1 - 0.4S_2 > 0)$ | 367.9 | 309.0, 426.7 |
| *C*-learning | $I(-57.5 + 0.9S_1 - 0.003S_2 > 0)$ | 347.1 | 291.7, 402.5 |
| Blip | NA* | 354.9 | 299.6, 410.8 |
| DR-IPCW | NA* | 354.8 | 300.6, 409.1 |
| BOWL | 1 | 360.7 | 306.7, 414.8 |
| Decision list | 1 | 360.7 | 306.7, 414.8 |

*denotes treatment regimes without explicit forms.

from non-parametric methods, we illustrate how these rules depend on the selected covariates in Figure 1, which displays the original and optimal treatment assignment using QRF for the 232 patients in the testing set.

Combining Table 10 and Figure 1, we can see that for both parametric and non-parametric methods, the estimated optimal treatment regime tends to recommend AZT+ddI combination therapy for patients with large body weight and small baseline CD4 count; otherwise, ddI monotherapy is recommended. The level of agreement across methods can be seen in Table 11 and Figure 1 where we compare the recommended treatment for each patient of the testing set. Although the recommended treatment is not always consistent across the methods, the treatment recommended by a parametric method may be preferred because of interpretability of the results due to known roles of body weight and baseline CD4 count in disease progression, which can be easily communicated with physicians in medical practice. However, when domain knowledge about disease progression is limited, non-parametric methods can be desirable to provide guidance for the treating physician in choosing the optimal treatment in order to achieve better prediction accuracy.

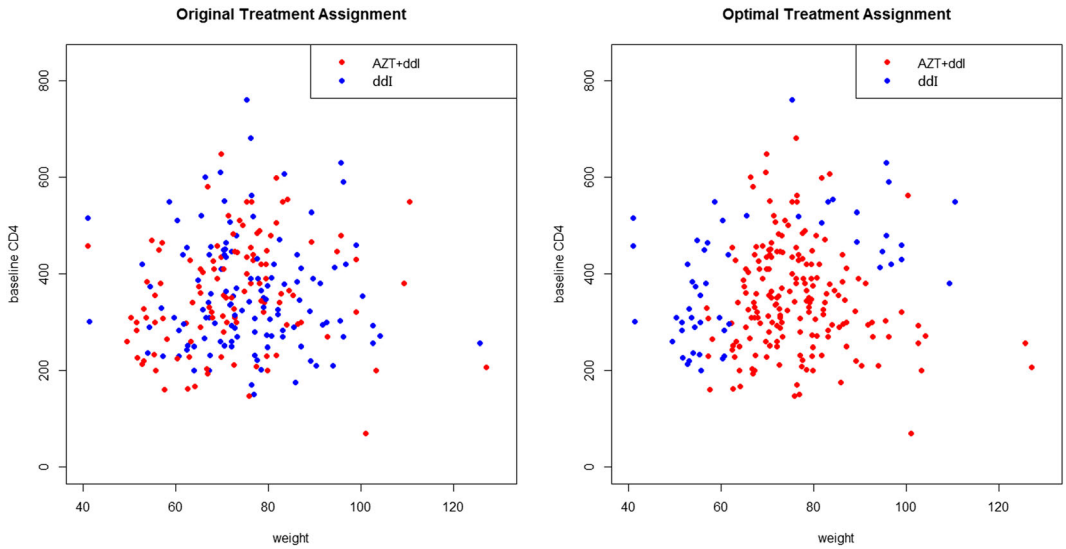


**FIGURE 1.** *Original and optimal treatment assignment using QRF for 232 patients in the testing set of HIV data*

Table 11. *The overlap rates for the optimal treatment assignment given by different methods for 232 patients in the testing set of the HIV data.*

| Method | QMR | QRF | *A*-learning | *C*-learning | Blip | DR-IPCW | BOWL | Decision list |
|---|---|---|---|---|---|---|---|---|
| QMR | 1 | 0.65 | 1 | 0.66 | 0.72 | 0.57 | 0.71 | 0.71 |
| QRF | | 1 | 0.65 | 0.92 | 0.78 | 0.63 | 0.76 | 0.76 |
| *A*-learning | | | 1 | 0.66 | 0.72 | 0.57 | 0.71 | 0.71 |
| *C*-learning | | | | 1 | 0.71 | 0.56 | 0.82 | 0.82 |
| Blip | | | | | 1 | 0.85 | 0.73 | 0.73 |
| DR-IPCW | | | | | | 1 | 0.66 | 0.66 |
| BOWL | | | | | | | 1 | 1 |
| Decision list | | | | | | | | 1 |

### 5.2 Melanoma Data Analysis

Melanoma is the fifth most common malignant tumour among men and the sixth most common malignant tumour among women, with a total of 6850 deaths in the US in 2020 (ACS, 2020). It is currently the deadliest form of skin cancer with a median survival time of approximately 24 months when diagnosed and treated in the earlier stages (Song *et al.*, 2015). Typically, treatment includes a sequence of therapies (such as chemotherapies and immunotherapies) that are selected based on the patient's baseline and time-evolving covariates (including disease progression). Here, we apply some of the methods described in Section 3 to a real-world dataset comprising 7730 patients diagnosed with melanoma. We seek to estimate an optimal treatment regime that maximises the mean survival rate at two years after a patient starts a first-line treatment.

The dataset was derived from the Flatiron health database. Each patient received the first-line of therapy during January 2011 and May 2017 and continued to receive multiple lines of treatment over multiple years. In this analysis, we consider only five of the major therapies available at the time and encode them as A, B, C, D, and E (to maintain confidentiality, the names of the treatment are not disclosed). Based on domain knowledge and a common variable selection method (SIS-Lasso) (Fan & Lv, 2008), we select the following baseline and time-dependent covariates: age, gender, Tstage (tumour size), Mstage (metastatic status), Nstage (lymph nodes), Gstage (tumour grade), BRAF (a human gene that encodes a protein called B-Raf that helps to control cell growth) and protein in serum or plasma. The subset of patients used in this analysis includes the 781 patients that received one of the selected treatment and for whom complete data are available.

To formalise this problem in the context of dynamic treatment regime, we consider 8 decision points over 2 years (3-month interval) and 6 treatment options at each decision point: A, B, C, D, E, and no-treatment. Because there are multiple treatment options available at each decision point, we only consider estimating the optimal treatment regime using *Q*-learning with multiple regression (QMR), *Q*-learning with random forest (QRF) and compare these to a method based on Nelson-Aalen estimators commonly used in the survival analysis literature (Shen *et al.*, 2017). We find that the optimal dynamic treatment regime estimated using the method in Shen *et al.* (2017) selects 'no-treatment' for the majority of decision points, which does not have clinical value. Although the precise reason for this unsatisfactory result is not clear, it might be due to the misspecified models for system dynamics or the class of candidate regimes might be unsuitable. The estimated mean survival rate under the optimal treatment regime for QMR and QRF are 0.726 and 0.596, respectively, with the empirical survival rate 0.410. According to the simulation performance and domain knowledge, the estimate for QMR is biased upwards while that for QRF is relatively accurate.

To study which treatment is optimal for patients at different decision points, we separate the data sets into a training set with size of 500 and a testing set with size of 281. We estimate the optimal treatment regime with the training set and compute the optimal treatment for patients in the testing set. Table 12 shows the frequency (as a proportion) with which each treatment option was recommended by QRF and QMR at all available decision points for the 281 patients in the testing set. We see that these two methods most often select treatment A as the optimal treatment, whereas treatment E, the most frequently selected treatment in the data, is less likely to be recommended. Figure 2 shows the original and optimal treatment assignments applying QRF and QMR for the 281 patients in the testing set. The regime estimated under QMR switches treatment more frequently than that of QRF. However, the QMR regime is more interpretable in that the influence of prior treatments and patient covariates on the decision rule can

Table 12. *The proportion of selecting different treatments under the estimated optimal treatment regime at all decision points combined for the 281 patients in the testing set of melanoma data.*

| Treatment | QRF | QMR | Empirical |
|---|---|---|---|
| A | **0.509** | *0.318* | 0.166 |
| B | 0.135 | **0.327** | 0.168 |
| C | *0.178* | 0.159 | *0.240* |
| D | 0.069 | 0.176 | 0.173 |
| E | 0.108 | 0.020 | **0.253** |

The boldfaced and italic values represent the highest and second highest proportions. The last column shows the original proportion in the observed data.

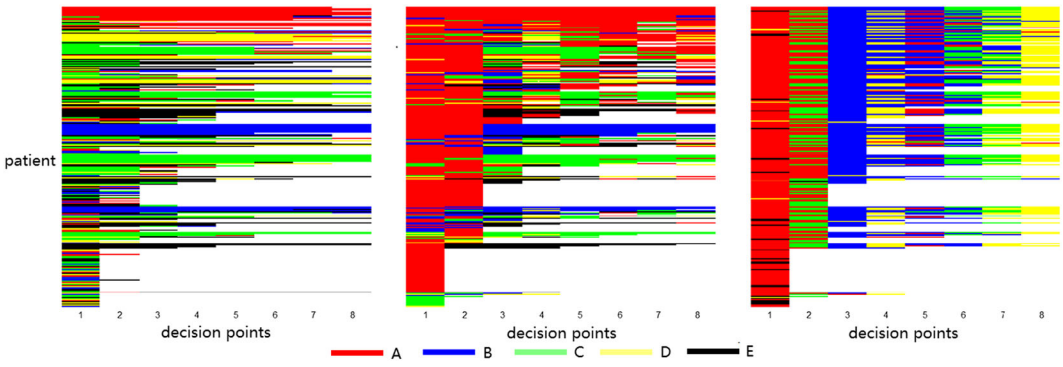


**FIGURE 2.** *Original (left) and optimal treatment assignment using QRF (middle) and QMR (right) for 281 patients in the testing set of the melanoma data. Red, blue, green, yellow, black, white represent treatments A, B, C, D, E, and no-treatment, respectively*

be easily identifiable, whereas QRF yields a black-box decision making mechanism that lacks clear clinical interpretation.

Unlike the HIV example in Section 5.1 where there is only one decision point and the treatment is recommended based on baseline body weight and CD4 count, in this example there are 8 decision points and the treatment at each decision point is recommended based on prior history (e.g. prior treatments received and other evolving covariates) of individual patients, leading to the complexity of decision-making process. Therefore, interpretability of the results is highly desirable in order to communicate with the treating physician for the choice of right therapy during the treatment. In this sense, QMR is always preferred.

## 6 Summary and Appraisal of the Methods

In this review, we introduced and evaluated a number of the most widely used estimators of an optimal treatment regime and the expected outcome under the optimal treatment regime. As might have been expected, no single method was uniformly best, we discuss here the pros and cons of each method.

### *6.1 Predictive Performance*

In settings with strong prior knowledge that can be used to inform a high-quality system dynamics mode, model-based planning can offer excellent predictive performance while still being

meaningful (interpretable) to domain experts. However, when such information is not available one is often faced with a potentially difficult trade-off between performance and interpetability/parsimony. Our simulation results showed that simple parametric models can be sensitive to misspecification while non-parametric models (esp., targeting the blip function, DR-IPCW, and random forests) provided good performance across the range of examples considered. While this point is often emphasised by proponents of non-parametric methods, it should be noted that the model diagnostics and interactive model-building that are carried out in practice are likely to catch severe misspecification thus weakening these criticisms. Furthermore, as estimation of an optimal regime is often carried out as part of exploratory and hypothesis-generating analyses, it is critical that the estimated optimal regime be interpretable to clinical and intervention scientists.

### *6.2 Model Constraints*

Even if the true optimal regime is a complex function of patient covariates, there is often interest in estimating an optimal regime within a pre-specified class of regimes, for example, the class or linear regimes or those representable as trees etc. (Zhang *et al.*, 2012; Zhang *et al.*, 2013; Zhang *et al.*, 2018). In some cases one can estimate a regime within the class by restricting the form of the *Q*-function or contrast although this may risk misspecification. A potentially more robust solution is to use direct-search methods such as *C*-learning or model-based planning. Such methods are also amenable to constraints on cost, risk of harm/adverse events, and local availability as these constraints can be implemented through constraints on the class of regimes or by augmenting the value function (Linn *et al.*, 2015; Luedtke & van der Laan, 2016c; Laber *et al.*, 2018; Wang *et al.*, 2018).

### *6.3 Interpretability*

As we have noted previously, interpetability is critical in when optimal dynamic treatment regimes are being used to generate new clinical insights and guide subsequent research. Clinicians are often unwilling to inform clinical decision making using an unintelligble black box. Furthermore, there is strong empirical and theoretical evidence to suggest the existence of effective but parsimonious treatment regimes (Zhang *et al.*, 2015; Wang *et al.*, 2018; Rudin, 2019). However, in online learning, problems where decision rules are being used to select patient interventions in real-time performance (subject to safety constraints) may be paramount. In such settings, the use of more flexible models may be justified.

### *6.4 Other Considerations*

In an era of increasingly large and complex data types issues such as extensibility and computational scalability are becoming important factors in selection of a method for estimating an optimal intervention strategy. *Q*-learning is among the most versatile methods in that it applies in virtually any setting in which regression can be applied; for example, recent examples include the use of functional predictors (McKeague & Qian, 2014; Laber & Staicu, 2018; Ciarleglio *et al.*, 2018; Dziak *et al.*, 2019), and ordinal treatments (Chen *et al.*, 2018).

Another concern is the efficiency and robustness of accompanying inference procedures. It is well-established that the value function is a non-smooth functional of the data-generating model and thus confidence intervals and tests for the value function are non-regular and may not perform well without adaptation. (Robins, 2004; Chakraborty *et al.*, 2010; Moodie & Richardson, 2010; Laber *et al.*, 2014; Chakraborty *et al.*, 2014; Song *et al.*, 2015; Luedtke &

Van Der Laan, 2016a). Valid inference procedures for many parametric approaches exist, for example, $A$-/$Q$-learning or BOWL with parametric decision rules, but work on inference for non-parametric/machine learning methods is an ongoing and active area of research (see Chernozhukov *et al.*, 2017; Naimi & Kennedy, 2017; Shi *et al.*, 2020, and references therein).

## 7 Conclusion

In this review we sought to provide an accessible introduction to a variety of commonly used methods for estimation of an optimal treatment regime along with an empirical evaluation of these methods. The results of our empirical experiments illustrated trade-offs between interpretability and performance and showed that there is no method that is likely to work best in all settings. Estimation and inference for optimal treatment regimes remains an active and exciting area of research. We hope that this review might serve as a bridge for researchers to engage with this important area.

## Endnotes

[1]The expectation here is taken over the data-generating distribution. If the class $\mathcal{D}_K$ contains the function $d_K^{\text{opt}}(h_K) = \text{argmax}_{a_K} Q_K(h_K, a_K)$ the data-generating distribution does not matter in the sense that $d_K^{\text{opt}}$ maximises $d_K \mapsto \mathbb{E}Q_K\{H_K, d_K(H_K)\}$ over any distribution on $H_K$. Otherwise, the argmax over $\mathcal{D}_K$ may depend on this distribution and one may wish to consider globally optimising over all of $d$ as described later in the next section. A similar argument applies for each $k = 1, \ldots, K - 1$. Optimising sequentially need not find the globally optimal regime, or even the optimal regime in $\mathcal{D}$, unless $d_k^{\text{opt}}$ is in $\mathcal{D}_k$ for all $k$. See Tsiatis *et al.* (2019) for additional discussion.

[2]In medicine, drug clearance is a pharmacokinetic measurement of the rate at which the active drug is removed from the body and drug clearance is correlated with the time course of a drug's action.

## Acknowledgement

The authors would like to thank Dr. Shannon Holloway for her helpful suggestions in editing this paper.

## References

ACS (2020). Cancer Facts & Figures 2020. https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2020/cancer-facts-and-figures-2020.pdf

Abe, N., Langford, J. & Zadrozny, B. (2004). An iterative method for multi-class cost-sensitive learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 3–11. https://doi.org/10.1145/1014052.1014056

Anand, R., Aggarwal, D. & Kumar, V. (2017). A comparative analysis of optimization solvers. *J. Stat. Manag. Syst.*, **20**(4), 623–635. https://doi.org/10.1080/09720510.2017.1395182

Ballarini, N.M., Rosenkranz, G.K., Jaki, T., König, F. & Posch, M. (2018). Subgroup identification in clinical trials via the predicted individual treatment effect. *PloS One*, **13**(10), e0205971. https://doi.org/10.1371/journal.pone.0205971

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of compstat'2010*, Springer, pp. 177–186.

Breiman, L. & Shang, N. (1996). Born again trees. *University of California, Berkeley, Berkeley, CA, Technical Report*, **1**(2), 4.

Burger, D.M., Meenhorst, P.L., Mulder, J.W., Koks, C.H.W., Bult, A. & Beijnen, J.H. (1994). Pharmacokinetics of zidovudine and metabolites in a patient with HIV-associated nephropathy and severe renal impairment. *Drug Investigation*, **7**(5), 282–287. https://doi.org/10.1007/BF03257420

Butler, E.L., Laber, E.B., Davis, S.M. & Kosorok, M.R. (2018). Incorporating patient preferences into estimation of optimal individualized treatment rules. *Biometrics*, **74**(1), 18–26.

Caniglia, E.C., Murray, E.J., Hernán, M.A. & Shahn, Z. (2021). Estimating optimal dynamic treatment strategies under resource constraints using dynamic marginal structural models. *Stat. Med.*, **40**(23), 4996–5005.

Carroll, R.J. & Ruppert, D. (1988). *Transformation and weighting in regression*, Vol. **30**. CRC Press, One Penn Plaza: New York, NY.

Chakraborty, B., Laber, E.B. & Zhao, Y.-Q. (2014). Inference about the expected performance of a data-driven dynamic treatment regime. *Clinical Trials*, **11**(4), 408–417.

Chakraborty, B. & Moodie, E. (2013). *Statistical methods for dynamic treatment regimes*. Springer: New York, NY.

Chakraborty, B. & Murphy, S.A. (2014). Dynamic treatment regimes. *Ann. Rev. Stat. Appl.*, **1**, 447–464. https://doi.org/10.1146/annurev-statistics-022513-115553

Chakraborty, B., Murphy, S. & Strecher, V. (2010). Inference for non-regular parameters in optimal dynamic treatment regimes. *Stat. Methods Med. Res.*, **19**(3), 317–343. https://doi.org/10.1177/0962280209105013

Chen, J., Fu, H., He, X., Kosorok, M.R. & Liu, Y. (2018). Estimating individualized treatment rules for ordinal treatments. *Biometrics*, **74**(3), 924–933. https://doi.org/10.1111/biom.12865

Chen, G., Zeng, D. & Kosorok, M.R. (2016). Personalized dose finding using outcome weighted learning. *J. Am. Stat. Assoc.*, **111**(516), 1509–1521. https://doi.org/10.1080/01621459.2016.1148611

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C. & Newey, W. (2017). Double/debiased/neyman machine learning of treatment effects. *Am. Econ. Rev.*, **107**(5), 261–65. https://doi.org/10.1257/aer.p20171038

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econ. J.*, **21**, C1–C68. https://doi.org/10.1111/ectj.12097

Ciarleglio, A., Petkova, E., Ogden, T. & Tarpey, T. (2018). Constructing treatment decision rules based on scalar and functional predictors when moderators of treatment effect are unknown. *J. Royal Stat. Soc. Ser. C, Appl. Stat.*, **67**(5), 1331.

Ciarleglio, A., Petkova, E., Tarpey, T. & Ogden, R.T. (2016). Flexible functional regression methods for estimating individualized treatment rules. *Stat.*, **5**(1), 185–199.

Clifton, J. & Laber, E. (2020). Q-learning: Theory and applications. *Ann. Rev. Stat. Appl.*, **7**, 279–301. https://doi.org/10.1146/annurev-statistics-031219-041220

Dziak, J.J., Coffman, D.L., Reimherr, M., Petrovich, J., Li, R., Shiffman, S. & Shiyko, M.P. (2019). Scalar-on-function regression for predicting distal outcomes from intensively gathered longitudinal data: Interpretability for applied scientists. *Stat. Surv.*, **13**, 150.

Ertefaie, A., McKay, J.R., Oslin, D. & Strawderman, R.L. (2021). Robust q-learning. *J. Am. Stat. Assoc.*, **116**(533), 368–381.

Ertefaie, A. & Strawderman, R.L. (2018). Constructing dynamic treatment regimes over indefinite time horizons. *Biometrika*, **105**(4), 963–977. https://doi.org/10.1093/biomet/asy043

Ertefaie, A., Wu, T., Lynch, K.G. & Nahum-Shani, I. (2016). Identifying a set that contains the best dynamic treatment regimes. *Biostatistics*, **17**(1), 135–148. https://doi.org/10.1093/biostatistics/kxv025

Fan, J. & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. Royal Stat. Soc.: Ser. B (Stat. Methodol.)*, **70**(5), 849–911. https://doi.org/10.1111/j.1467-9868.2008.00674.x

Fard, M.M. & Pineau, J. (2011). Non-deterministic policies in Markovian decision processes. *J. Artif. Intell. Res.*, **40**, 1–24. https://doi.org/10.1613/jair.3175

Ghavamzadeh, M., Mannor, S., Pineau, J., Tamar, A. et al. (2015). Bayesian reinforcement learning: A survey. *Foundat. Trends® Mach. Learn.*, **8**(5-6), 359–483. https://arxiv.org/abs/1609.04436

Goldberg, Y. & Kosorok, M.R. (2012). Q-learning with censored data. *Ann. Stat.*, **40**(1), 529. https://doi.org/10.1214/12-AOS968

Gosavi, A. et al. (2015). *Simulation-based optimization*. Springer.

Guan, Q., Reich, B.J., Laber, E.B. & Bandyopadhyay, D. (2020). Bayesian nonparametric policy search with application to periodontal recall intervals. *J. Am. Stat. Assoc.*, **115**(531), 1066–1078. https://doi.org/10.1080/01621459.2019.1660169

Guo, B. & Yuan, Y. (2017). Bayesian phase I/II biomarker-based dose finding for precision medicine with molecularly targeted agents. *J. Am. Stat. Assoc.*, **112**(518), 508–520. https://doi.org/10.1080/01621459.2016.1228534

Hammer, S.M., Katzenstein, D.A., Hughes, M.D., Gundacker, H., Schooley, R.T., Haubrich, R.H., Henry, W.K., Lederman, M.M., Phair, J.P. & Niu, M. (1996). A trial comparing nucleoside monotherapy with combination

therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *New England J. Med.*, **335**(15), 1081–1090. https://www.nejm.org/doi/full/10.1056/NEJM199610103351501
Horvitz, D.G. & Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *J. Am. Stat. Assoc.*, **47**(260), 663–685.
Huang, X., Goldberg, Y. & Xu, J. (2019). Multicategory individualized treatment regime using outcome weighted learning. *Biometrics*, **75**(4), 1216–1227. https://doi.org/10.1111/biom.13084
Josefsson, M. & Daniels, M.J. (2020). Bayesian semi-parametric G-computation for causal inference in a cohort study with non-ignorable dropout and death. https://doi.org/10.1111/rssc.12464
Juraska, M. & Juraska, M.M. (2010). Package speff2trial. R software.
Kallus, N., Mao, X. & Uehara, M. (2019). Localized debiased machine learning: Efficient estimation of quantile treatment effects, conditional value at risk, and beyond.
Kennedy, E.H. (2022). Semiparametric doubly robust targeted double machine learning: a review. arXiv preprint arXiv:2203.06469.
Kosorok, M.R. & Laber, E.B. (2019). Precision medicine. *Ann. Rev. Stat. Appl.*, **6**, 263–286. https://doi.org/10.1146/annurev-statistics-030718-105251
Kosorok, M.R. & Moodie, E.E.M. (2015). *Adaptive TreatmentStrategies in Practice: Planning Trials and Analyzing Data for Personalized Medicine*, Vol. **21**. SIAM. https://doi.org/10.1137/1.9781611974188
Laber, E.B., Linn, K.A. & Stefanski, L.A. (2014). Interactive model building for Q-learning. *Biometrika*, **101**(4), 831–847. https://doi.org/10.1093/biomet/asu043
Laber, E.B., Lizotte, D.J. & Ferguson, B. (2014). Set-valued dynamic treatment regimes for competing outcomes. *Biometrics*, **70**(1), 53–61. https://doi.org/10.1111/biom.12132
Laber, E.B., Meyer, N.J., Reich, B.J., Pacifici, K., Collazo, J.A. & Drake, J.M. (2018). Optimal treatment allocations in space and time for on-line control of an emerging infectious disease. *J. Royal Stat. Soc.: Ser. C (Appl. Stat.)*, **67**(4), 743–789. https://doi.org/10.1111/rssc.12266
Laber, E.B. & Murphy, S.A. (2011). Adaptive confidence intervals for the test error in classification. *J. Am. Stat. Assoc.*, **106**(495), 904–913. https://doi.org/10.1198/jasa.2010.tm10053
Laber, E.B. & Staicu, A.-M. (2018). Functional feature construction for individualized treatment regimes. *J. Am. Stat. Assoc.*, **113**(523), 1219–1227.
Laber, E.B., Wu, F., Munera, C., Lipkovich, I., Colucci, S. & Ripa, S. (2018). Identifying optimal dosage regimes under safety constraints: An application to long term opioid treatment of chronic pain. *Stat. Med.*, **37**(9), 1407–1418. https://doi.org/10.1002/sim.7566
Laber, E.B. & Zhao, Y.Q. (2015). Tree-based methods for individualized treatment regimes. *Biometrika*, **102**(3), 501–514. https://doi.org/10.1093/biomet/asv028
Laber, E.B., Zhao, Y.-Q., Regh, T., Davidian, M., Tsiatis, A., Stanford, J.B., Zeng, D., Song, R. & Kosorok, M.R. (2016). Using pilot data to size a two-arm randomized trial to find a nearly optimal personalized treatment strategy. *Stat. Med.*, **35**(8), 1245–1256.
Lakkaraju, H. & Rudin, C. (2017). Learning cost-effective and interpretable treatment regimes. In *Artificial intelligence and statistics*, pp. 166–175. http://proceedings.mlr.press/v54/lakkaraju17a/lakkaraju17a.pdf
Lee, J., Thall, P.F., Ji, Y. & Müller, P. (2015). Bayesian dose-finding in two treatment cycles based on the joint utility of efficacy and toxicity. *J. Am. Stat. Assoc.*, **110**(510), 711–722. https://doi.org/10.1080/01621459.2014.926815
Leqi, L. & Kennedy, E.H. (2022). Median optimal treatment regimes. arXiv preprint arXiv:2103.01802.
Linn, K.A., Laber, E.B. & Stefanski, L.A. (2015). Chapter 15: Estimation of dynamic treatment regimes for complex outcomes: Balancing benefits and risks. In *Adaptive treatment strategies in practice: Planning trials and analyzing data for personalized medicine*, pp. 249–262. SIAM.
Linn, K.A., Laber, E.B. & Stefanski, L.A. (2017). Interactive Q-learning for quantiles. *J. Am. Stat. Assoc.*, **112**(518), 638–649. https://doi.org/10.1080/01621459.2016.1155993
Lipkovich, I., Dmitrienko, A. & B D'Agostino Sr, R. (2017). Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Stat. Med.*, **36**(1), 136–196. https://doi.org/10.1002/sim.7064
Liu, Y., Wang, Y., Kosorok, M.R., Zhao, Y. & Zeng, D. (2018). Augmented outcome-weighted learning for estimating optimal dynamic treatment regimens. *Stat. Med.*, **37**(26), 3776–3788. https://doi.org/10.1002/sim.7844
Lizotte, D.J. & Laber, E.B. (2016). Multi-objective markov decision processes for data-driven decision support. *The J. Mach. Learn. Res.*, **17**(1), 7378–7405.
Loh, W.-Y. (2014). Fifty years of classification and regression trees. *Int. Stat. Rev.*, **82**(3), 329–348. https://doi.org/10.1111/insr.12016
Luckett, D.J., Laber, E.B., Kahkoska, A.R., Maahs, D.M., Mayer-Davis, E. & Kosorok, M.R. (2020). Estimating dynamic treatment regimes in mobile health using v-learning. *J. Am. Stat. Assoc.*, **115**(530), 692–706. https://doi.org/10.1080/01621459.2018.1537919

Luckett, D.J., Laber, E.B. & Kosorok, M.R. (2017). Estimation and optimization of composite outcomes. https://arxiv.org/pdf/1711.10581.pdf

Luedtke, A.R., Sofrygin, O., van der Laan, M.J. & Carone, M. (2017). Sequential double robustness in right-censored longitudinal models. https://arxiv.org/pdf/1705.02459.pdf

Luedtke, A.R. & Van Der Laan, M.J. (2016a). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Ann. Stat.*, **44**(2), 713.

Luedtke, A.R. & van der Laan, M.J. (2016b). Super-learning of an optimal dynamic treatment rule. *The Int. J. Biostat.*, **12**(1), 305–332. https://doi.org/10.1515/ijb-2015-0052

Luedtke, A.R. & van der Laanc, M.J. (2016c). Optimal individualized treatments in resource-limited settings. *The Int. J. Biostat.*, **12**(1), 283–303.

McKeague, I.W. & Qian, M. (2011). Sparse functional linear regression with applications to personalized medicine. In *Recent advances in functional data analysis and related topics*, Springer, pp. 213–218.

McKeague, I.W. & Qian, M. (2014). Estimation of treatment policies based on functional predictors. *Stat. Sinica*, **24**(3), 1461.

Molenberghs, G., Fitzmaurice, G., Kenward, M.G., Tsiatis, A. & Verbeke, G. (2014). *Handbook of missing data methodology.* CRC Press.

Moodie, E.ricaE.M., Dean, N. & Sun, Y.R. (2014). Q-learning: Flexible learning about useful utilities. *Stat. Biosci.*, **6**(2), 223–243. https://doi.org/10.1007/s12561-013-9103-z

Moodie, E.E.M. & Richardson, T.S. (2010). Estimating optimal dynamic regimes: Correcting bias under the null. *Scandinavian J. Stat.*, **37**(1), 126–146. https://doi.org/10.1111/j.1467-9469.2009.00661.x

Murphy, S.A. (2003). Optimal dynamic treatment regimes. *J. Royal Stat. Soc.: Ser. B (Stat. Methodol.) Methodology)*, **65**(2), 331–355. https://doi.org/10.1111/1467-9868.00389

Murphy, S.A. (2005a). An experimental design for the development of adaptive treatment strategies. *Stat. Med.*, **24**(10), 1455–1481. https://doi.org/10.1002/sim.2022

Murphy, S.A. (2005b). A generalization error for Q-learning. *J. Mach. Learn. Res.*, **6**(Jul), 1073–1097. https://www.jmlr.org/papers/volume6/murphy05a/murphy05a.pdf

Naimi, A.I. & Kennedy, E.H. (2017). Nonparametric double robustness. arXiv preprint arXiv:1711.07137.

Nie, X., Brunskill, E. & Wager, S. (2020). Learning when-to-treat policies. *J. Am. Stat. Assoc.*, **2020**, 1–18. https://doi.org/10.1080/01621459.2020.1831925

Orellana, L., Rotnitzky, A. & Robins, J.M. (2010a). Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, Part I: main content. *The Int. J. Biostat.*, **6**(2). https://doi.org/10.2202/1557-4679.1200

Orellana, L., Rotnitzky, A. & Robins, J.M. (2010b). Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, Part II: proofs of results. *The Int. J. Biostat.*, **6**(2). https://doi.org/10.2202/1557-4679.1242

Polley, E.C. & Van Der Laan, M.J. (2010). Super learner in prediction. In *Targeted learning: Causal inference for observational and experimental data*, Eds. van der Laan, M.J. & Rose, S. http://biostats.bepress.com/ucbbiostat/paper266

Polydoros, A.S. & Nalpantidis, L. (2017). Survey of model-based reinforcement learning: Applications on robotics. *J. Intell. Robotic Syst.*, **86**(2), 153–173. https://doi.org/10.1007/s10846-017-0468-y

Qian, M. & Murphy, S.A. (2011). Performance guarantees for individualized treatment rules. *An. Stat.*, **39**(2), 1180. https://doi.org/10.1214/10-AOS864

Rich, B., Moodie, E.E.M. & Stephens, D.A. (2014). Simulating sequential multiple assignment randomized trials to generate optimal personalized warfarin dosing strategies. *Clinical Trials*, **11**(4), 435–444. https://doi.org/10.1177/1740774513517063

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period application to control of the healthy worker survivor effect. *Math. Model.*, **7**(9-12), 1393–1512. https://doi.org/10.1016/0270-0255(86)90088-6

Robins, J.M. (2004). Optimal structural nested models for optimal sequential decisions. In *Proceedings of the second seattle symposium in biostatistics*, pp. 189–326, Springer.

Robins, J.M., Wasserman, L., Geiger, D. & Shenoy, P. (1997). Estimation of Effects of Sequential Treatments by Reparameterizing Directed Acyclic Graphs, Proceedings of the thirteenth conference on uncertainty in artificial intelligence, pp. 409–420. https://arxiv.org/ftp/arxiv/papers/1302/1302.1566.pdf

Rose, E.J., Laber, E.B., Davidian, M., Tsiatis, A.A., Zhao, Y.-Q. & Kosorok, M.R. (2019). Sample size calculations for smarts. arXiv preprint arXiv:1906.06646.

Rubin, D.B. (1978). Bayesian inference for causal effects: The role of randomization. *The Ann. Stat.*, **1978**, 34–58. https://www.jstor.org/stable/2958688

Rubin, D.B. & van der Laan, M.J. (2012). Statistical issues and limitations in personalized medicine research with clinical trials. *The Int. J. Biostat.*, **8**(1), 18. https://doi.org/10.1515/1557-4679.1423
Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Mach. Intell.*, **1**(5), 206–215.
Rudin, C. & Ertekin, S. (2018). Learning customized and optimized lists of rules with mathematical programming. *Math. Program. Comput.*, **10**(4), 659–702.
Saarela, O., Arjas, E., Stephens, D.A. & Moodie, E.E.M. (2015). Predictive Bayesian inference and dynamic treatment regimes. *Biometric. J.*, **57**(6), 941–958. https://doi.org/10.1002/bimj.201400153
Schulte, P.J., Tsiatis, A.A., Laber, E.B. & Davidian, M. (2014). Q-and A-learning methods for estimating optimal dynamic treatment regimes. *Stat. Sci.*, **29**(4), 640. https://doi.org/10.1214/13-STS450
Schulz, J. & Moodie, E.E.M. (2020). Doubly robust estimation of optimal dosing strategies. *J. Am. Stat. Assoc.*, **2020**, 1–13. https://doi.org/10.1080/01621459.2020.1753521
Semenova, V. & Chernozhukov, V. (2020). Debiased machine learning of conditional average treatment effects and other causal functions. *The Econ. J.*, **2020**, utaa027. https://doi.org/10.1093/ectj/utaa027
Shen, J., Wang, L. & Taylor, J.M.G. (2017). Estimation of the optimal regime in treatment of prostate cancer recurrence from observational data using flexible weighting models. *Biometrics*, **73**(2), 635–645. https://doi.org/10.1111/biom.12621
Shi, C., Fan, A., Song, R. & Lu, W. (2018). High-dimensional a-learning for optimal dynamic treatment regimes. *Ann. Stat.*, **46**(3), 925.
Shi, C., Lu, W. & Song, R. (2020). Breaking the curse of nonregularity with subagging—inference of the mean outcome under optimal treatment regimes. *J. Mach. Learn. Res.*, **21**(176), 1–67.
Sies, A. & Van Mechelen, I. (2017). Comparing four methods for estimating tree-based treatment regimes. *The Int. J. Biostat.*, **13**(1). https://doi.org/10.1515/ijb-2016-0068
Song, R., Kosorok, M., Zeng, D., Zhao, Y., Laber, E. & Yuan, M. (2015). On sparse representation for optimal individualized treatment selection with penalized outcome weighted learning. *Stat*, **4**(1), 59–68.
Song, R., Wang, W., Zeng, D. & Kosorok, M.R. (2015). Penalized q-learning for dynamic treatment regimens. *Stat. Sinica*, **25**(3), 901. https://doi.org/10.5705/ss.2012.364
Song, X., Zhao, Z., Barber, B., Farr, A.M., Ivanov, B. & Novich, M. (2015). Overall survival in patients with metastatic melanoma. *Current Med. Res. Opinion*, **31**(5), 987–991. https://doi.org/10.1185/03007995.2015.1021904
Spall, J.C. (2005). *Introduction to stochastic search and optimization: estimation, simulation, and control*, Vol. **65**. John Wiley & Sons.
Sutton, R.S. & Barto, A.G. (2018). *Introduction to reinforcement learning*, Vol. **135**. MIT press: Cambridge.
Tao, Y., Wang, L. & Almirall, D. (2018). Tree-based reinforcement learning for estimating optimal dynamic treatment regimes. *The Ann. Appl. Stat.*, **12**(3), 1914. https://doi.org/10.1214/18-AOAS1137
Taylor, J.eremyM.G., Cheng, W. & Foster, J.C. (2015). Reader reaction to "A robust method for estimating optimal treatment regimes" by Zhang et al. (2012). *Biometrics*, **71**(1), 267–273. https://doi.org/10.1111/biom.12228
Thall, P.F., Millikan, R.E. & Sung, H.-G. (2000). Evaluating multiple treatment courses in clinical trials. *Stat. Med.*, **19**(8), 1011–1028. https://doi.org/10.1002/(SICI)1097-0258(20000430)19:8<1011::AID-SIM414>3.0.CO;2-M
Thall, P.F., Wooten, L.H., Logothetis, C.J., Millikan, R.E. & Tannir, N.M. (2007). Bayesian and frequentist two-stage treatment strategies based on sequential failure times subject to interval censoring. *Stat. Med.*, **26**(26), 4687–4702. https://doi.org/10.1002/sim.2894
Tsiatis, A. (2007). *Semiparametric theory and missing data*. Springer Science & Business Media.
Tsiatis, A.A., Davidian, M., Holloway, S. & Laber, E.B. (2019). *Dynamic treatment regimes: Statistical methods for precision medicine*. CRC Press: New York, NY.
van der Laan, M.J. & Luedtke, A.R. (2014). Targeted learning of an optimal dynamic treatment, and statistical inference for its mean outcome. http://biostats.bepress.com/ucbbiostat/paper329
van der Laan, M.J. & Petersen, M.L. (2007). Causal effect models for realistic individualized treatment and intention to treat rules. *The Int. J. Biostat.*, **3**(1). https://doi.org/10.2202/1557-4679.1022
van der Laan, M.J. & Rose, S. (2011). *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media.
van der Laan, M.J. & Rose, S. (2018). *Targeted learning in data science: causal inference for complex longitudinal studies*. Springer.
van der Laan, M.J. & Rubin, D. (2006). Targeted maximum likelihood learning. *The Int. J. Biostat.*, **2**(1). https://doi.org/10.2202/1557-4679.1043
Wallace, M.P., Moodie, E.ricaE.M. & Stephens, D.A. (2018). Reward ignorant modeling of dynamic treatment regimes. *Biomet. J.*, **60**(5), 991–1002. https://doi.org/10.1002/bimj.201700322
Wang, T., Bao, X., Clavera, I., Hoang, J., Wen, Y., Langlois, E., Zhang, S., Zhang, G., Abbeel, P. & Ba, J. (2019). Benchmarking model-based reinforcement learning. arXiv preprint arXiv:1907.02057.

Wang, Y., Fu, H. & Zeng, D. (2018). Learning optimal personalized treatment rules in consideration of benefit and risk: with an application to treating type 2 diabetes patients with insulin therapies. *J. Am. Stat. Assoc.*, **113**(521), 1–13. https://doi.org/10.1080/01621459.2017.1303386

Wang, L., Zhou, Y., Song, R. & Sherwood, B. (2018). Quantile-optimal treatment regimes. *J. Am. Stat. Assoc.*, **113**(523), 1243–1254. https://doi.org/10.1080/01621459.2017.1330204

Wu, T. (2016). Set Valued Dynamic Treatment Regimes. Ph.D. Thesis. https://deepblue.lib.umich.edu/bitstream/handle/2027.42/133462/wutiansh_1.pdf?sequence=1&isAllowed=y

Xu, Y., Müller, P., Wahed, A.S. & Thall, P.F. (2016). Bayesian nonparametric estimation for dynamic treatment regimes with sequential transition times. *J. Am. Stat. Assoc.*, **111**(515), 921–950. https://doi.org/10.1080/01621459.2015.1086353

Yu, Z. & van der Laan, M.J. (2002). Construction of counterfactuals and the G-computation formula. https://biostats.bepress.com/ucbbiostat/paper122/

Zhang, Y., Laber, E.B., Davidian, M. & Tsiatis, A.A. (2018). Interpretable dynamic treatment regimes. *J. Am. Stat. Assoc.*, **113**(524), 1541–1549. https://doi.org/10.1080/01621459.2017.1345743

Zhang, Y., Laber, E.B., Tsiatis, A. & Davidian, M. (2015). Using decision lists to construct interpretable and parsimonious treatment regimes. *Biometrics*, **71**(4), 895–904. https://doi.org/10.1111/biom.12354

Zhang, B., Tsiatis, A.A., Davidian, M., Zhang, M. & Laber, E. (2012). Estimating optimal treatment regimes from a classification perspective. *Stat.*, **1**(1), 103–114. https://doi.org/10.1002/sta.411

Zhang, B., Tsiatis, A.A., Laber, E.B. & Davidian, M. (2012). A robust method for estimating optimal treatment regimes. *Biometrics*, **68**(4), 1010–1018. https://doi.org/10.1111/j.1541-0420.2012.01763.x

Zhang, B., Tsiatis, A.A., Laber, E.B. & Davidian, M. (2013). Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika*, **100**(3), 681–694. https://doi.org/10.1093/biomet/ast014

Zhang, Y. & van der Schaar, M. (2020). Gradient regularized v-learning for dynamic treatment regimes. *Adv. Neural Inf. Process. Syst.*, **33**.

Zhang, B. & Zhang, M. (2018). C-learning: A new classification framework to estimate optimal dynamic treatment regimes. *Biometrics*, **74**(3), 891–899. https://doi.org/10.1111/biom.12836

Zhao, Y.-Q., Laber, E.B., Ning, Y., Saha, S. & Sands, B.E. (2019). Efficient augmentation and relaxation learning for individualized treatment rules using observational data. *J. Mach. Learn. Res.*, **20**(48), 1–23. https://www.jmlr.org/papers/volume20/18-191/18-191.pdf

Zhao, Y.-Q., Zeng, D., Laber, E.B. & Kosorok, M.R. (2015). New statistical learning methods for estimating optimal dynamic treatment regimes. *J. Am. Stat. Assoc.*, **110**(510), 583–598. https://doi.org/10.1080/01621459.2014.937488

Zhao, Y., Zeng, D., Rush, A.J. & Kosorok, M.R. (2012). Estimating individualized treatment rules using outcome weighted learning. *J. Am. Stat. Assoc.*, **107**(499), 1106–1118. https://doi.org/10.1080/01621459.2012.695674

Zhao, Y., Zeng, D., Socinski, M.A. & Kosorok, M.R. (2011). Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. *Biometrics*, **67**(4), 1422–1433. https://doi.org/10.1111/j.1541-0420.2011.01572.x

Zhou, X. & Kosorok, M.R. (2017). Causal nearest neighbor rules for optimal treatment regimes. https://arxiv.org/abs/1711.08451

Zhou, X., Mayer-Hamblett, N., Khan, U. & Kosorok, M.R. (2017). Residual weighted learning for estimating individualized treatment rules. *J. Am. Stat. Assoc.*, **112**(517), 169–187. https://doi.org/10.1080/01621459.2015.1093947

Zhu, L., Lu, W., Kosorok, M.R. & Song, R. (2020). Kernel assisted learning for personalized dose finding. In *Proceedings of the 26th acm sigkdd international conference on knowledge discovery & data mining*, pp. 56–65. https://doi.org/10.1145/3394486.3403048

Zadrozny, B., Langford, J. & Abe, N. (2017). Cost-sensitive learning by cost-proportionate example weighting. In *Third IEEE international conference on data mining*, pp. 435–442. IEEE. https://doi.org/10.1109/ICDM.2003.1250950

Zhu, W., Zeng, D. & Song, R. (2019). Proper inference for value function in high-dimensional q-learning for dynamic treatment regimes. *J. Am. Stat. Assoc.*, **114**(527), 1404–1417.

Zhu, R., Zhao, Y.-Q., Chen, G., Ma, S. & Zhao, H. (2017). Greedy outcome weighted tree learning of optimal personalized treatment rules. *Biometrics*, **73**(2), 391–400. https://doi.org/10.1111/biom.12593

## Appendix A: Data Generating Models for $K = 4, 8$ in Section 4.1.1

When $K = 4, 8$, suppose we observe i.i.d. data $(\{S_{ki}, A_{ki}\}_{k=1}^{K}, Y_i)$, $i = 1, \ldots, n$ with

$S_1 \sim N(0, 1)$,
$A_1|S_1 = s_1 \sim \text{Bernoulli}(0.5)$,
$for\ k = 2, \ldots, K$,
$S_k|\{S_m = s_m, A_m = a_m\}_{m=1}^{k-1} \sim N\left[1 + a_{k-1}\left\{1_{0 \le s_{k-1} \le K/2} - 1_{s_{k-1} < 0} - 1_{s_{k-1} > K/2}\right\} + s_{k-1}, 1\right]$,
$A_k|\{S_m = s_m, A_m = a_m\}_{m=1}^{k-1}, S_k = s_k \sim \text{Bernoulli}(0.5)$,
$Y_K|\{S_m = s_m, A_m = a_m\}_{m=1}^{K} \sim N\left[1 + a_K\left\{1_{0 \le s_K \le K/2} - 1_{s_K < 0} - 1_{s_K > K/2}\right\} + s_K, 1\right]$ .

## Appendix B: Working Models for QMR and Model-based Planning in Section 4

For QMR, we estimate $Q_k(h_k, a_k)$ using linear regression on the main effects, that is, $h_k$, $a_k$, and the first-order interaction terms between treatment and history. To be specific, for the continuous outcome scenario, we assume $Q_k(h_k, a_k) = \phi_k(h_k, a_k)^{\mathrm{T}}\beta_k$, where $\phi_k(h_k, a_k) = \left(\{s_i\}_{i=1}^{k}, \{a_i\}_{i=1}^{k}, \{a_i s_1, \ldots, a_i s_i\}_{i=1}^{k}, \{a_i a_1, \ldots, a_i a_{i-1}\}_{i=1}^{k}\right)$, $k = 1, \ldots, K$. We estimate the linear parameters $\beta_k$ using OLS or penalisation approaches. For the binary outcome scenario, we assume $\text{logit}\{Q_k(h_k, a_k)\} = \phi_k(h_k, a_k)^{\mathrm{T}}\beta_k$, where $\text{logit}(x) = \log\left(\frac{x}{1-x}\right)$. We estimate the linear parameters $\beta_k$ using (penalised) logistic regression.

For the method of model-based planning, we estimate the $Q$- and $V$-functions using Monte Carlo methods (see details in Section 3). We assume that the transition and conditional outcome density follow normal/binomial distributions with linear mean models and constant covariance-matrices which are estimated using maximum likelihood. To be specific, for continuous outcome $Y_K$, we assume $Y_K|H_K, A_K \sim N\left[\phi_K(H_K, A_K)^{\mathrm{T}}\beta_K, \sigma_K^2\right]$. For continuous state $S_k$, $k = 2, \ldots, K$, we assume $S_k|H_{k-1}, A_{k-1} \sim N\left[\phi_{k-1}(H_{k-1}, A_{k-1})^{\mathrm{T}}\beta_{k-1}, \sigma_{k-1}^2\right]$. We estimate $\beta_k$ using OLS or penalisation approaches. For binary outcome $Y_K$, we assume $Y_K|H_K, A_K \sim \text{Bernoulli}\left[\text{expit}\{\phi_K(H_K, A_K)^{\mathrm{T}}\beta_K\}\right]$. For binary state $S_k$, $k = 2, \ldots, K$, we assume $S_k|H_{k-1}, A_{k-1} \sim \text{Bernoulli}\left[\text{expit}\{\phi_{k-1}(H_{k-1}, A_{k-1})^{\mathrm{T}}\beta_{k-1}\}\right]$. We estimate $\beta_k$ using (penalised) logistic regression.

## Appendix C: A brief review of Targeted Maximum Likelihood Estimator (TMLE)

We consider the TMLE of $V(d)$; for a general introduction to TMLE see (van der Laan & Rose, 2011). Define $V_k^d(h_k) \triangleq \mathbb{E}_d(Y|H_k = h_k)$ and a submodel $V_k^{d, \epsilon_k, g}(h_k)$ such that $\text{logit}\{V_k^{d, \epsilon_k, g}(h_k)\} = \text{logit}\{V_k^d(h_k, a_k)\} + \epsilon_k \prod_{i=1}^{k-1} \frac{I\{A_i = d_i(h_i)\}}{g_i(h_i)}$, where $g$ denotes the propensity score at each stage and is assumed to be known here. To solve TMLE, a loss function for $V_k^d(h_k)$ is introduced such that the gradient of the loss function at $\epsilon_k = 0$ equals the efficient influence curve of the probability distribution as following:

$$L_{V_{k+1}^d}(V_k^d)(h_{k+1}) = -\left[V_{k+1}^d(h_{k+1}) \log V_k^d(h_k) + \{1 - V_{k+1}^d(h_{k+1})\}\log\{1 - V_k^d(h_k)\}\right].$$

The TMLE is solved iteratively from $k = K$ to $k = 1$. Initially at stage $K$, we obtain estimators $V^d_{K,n}$ for $V^d_K$ with super learning. At stage $k = K - 1, \ldots, 1$, we compute the maximum likelihood estimate

$$\epsilon_{k,n} = \arg\min_{\epsilon_k} \mathbb{P}_n L_{V^d_{k+1,n}}(V^{d,\epsilon_k,g}_k)(h_{k+1})$$

and set $V^d_{k,n} = V^{d,\epsilon_{k,n},g}_k$. The final estimator for $V(d)$ is $V^d_{1,n}(h_1)$.

## Appendix D: Comparisons Across Different Regime Evaluation Methods

In Section 4, we compared the performance across various methods in estimating the expected outcome under the estimated optimal treatment regime, that is, $\widehat{V}_n(\widehat{d}_n)$. In other settings, one may be interested in the performance of value estimators in the context of regime evaluation, that is, the quality of an estimator $\widehat{V}_n(d)$ of $V(d)$, where $d$ is fixed. In Table D1 we compare TMLE and IPTW for their performance on estimating the expected outcome under three fixed treatment regimes $d^{\text{opt}}$, $d^{\text{const}}$ and $d^{\text{random}}$. $d^{\text{opt}}$ is the true optimal treatment regime; $d^{\text{const}}$ is a constant treatment regime in which we assume $d_k(h_k) \equiv 1$; and $d^{\text{random}}$ is a completely randomised treatment regime with propensity score 0.5. Note that while $Q$-learning, $A$-learning and some other methods can estimate $V(d^{\text{opt}})$, they cannot estimate $V(d)$ given a fixed regime $d$ and thus we do not consider them here.

Table D1. *Comparisons across different regime evaluation methods under the data-generating settings in Section 4.1.1.*

| Trt. regime | Method | MSE ($n = 250$) | | MSE ($n = 500$) | | MSE ($n = 1000$) | |
|---|---|---|---|---|---|---|---|
| | | $K = 1$ | $K = 2$ | $K = 1$ | $K = 2$ | $K = 1$ | $K = 2$ |
| $d^{\text{opt}}$ | TMLE | 0.88 (0.31) | 6.36 (0.25) | 0.23 (0.01) | 5.79 (0.19) | 0.13 (0.01) | 5.12 (0.13) |
| | IPTW | 1.22 (0.32) | 4.48 (0.21) | 0.51 (0.03) | 3.47 (0.15) | 0.42 (0.02) | 2.75 (0.10) |
| $d^{\text{const}}$ | TMLE | 1.73 (0.32) | 4.59 (0.29) | 0.95 (0.04) | 3.55 (0.17) | 0.75 (0.02) | 3.29 (0.13) |
| | IPTW | 1.97 (0.32) | 4.80 (0.30) | 1.20 (0.04) | 3.73 (0.18) | 1.07 (0.03) | 3.49 (0.14) |
| $d^{\text{random}}$ | TMLE | 0.81 (0.17) | 11.27 (0.43) | 0.45 (0.02) | 10.78 (0.30) | 0.41 (0.02) | 10.52 (0.24) |
| | IPTW | 0.94 (0.23) | 10.69 (0.42) | 0.34 (0.02) | 15.64 (0.43) | 0.22 (0.01) | 25.06 (0.45) |

The propensity score is unknown here and is estimated using super learning. The MSE (standard error) is calculated for three treatment regimes $d^{\text{opt}}$, $d^{\text{const}}$ and $d^{\text{random}}$.

## Appendix E: Optimal Treatment Assignment Given Across Various Methods in 5.1

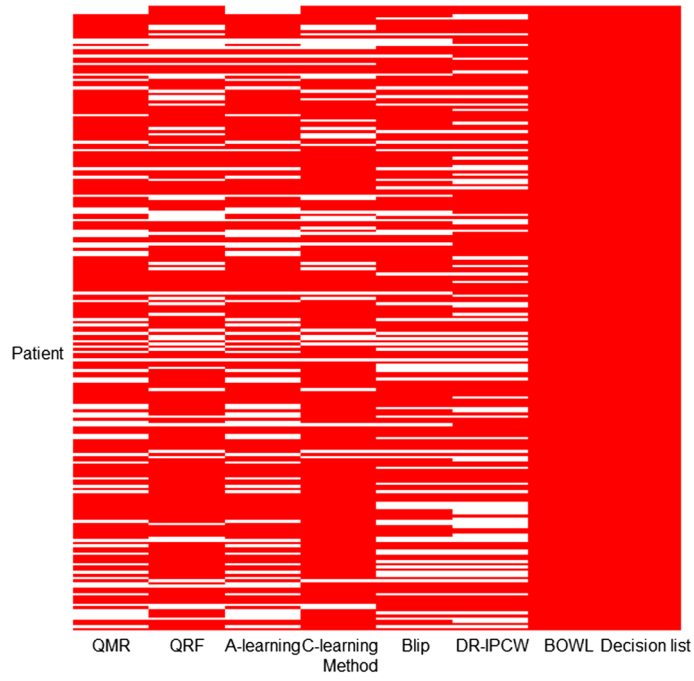


**FIGURE E1.** *The optimal treatment assignment given by different methods for 232 patients in the testing set of the HIV data. The red and white areas represent the treatment of AZT+ddI and ddI, respectively*

[*Received April 2021; accepted December 2022*]