Statistics
in Medicine

Research Article

Received 28 August 2014, Accepted 24 May 2015 Published online 21 June 2015 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6558

Optimization of multi-stage dynamic treatment regimes utilizing accumulated data

Xuelin Huang,[a*†] Sangbum Choi,[b] Lu Wang[c] and Peter F. Thall[a]

In medical therapies involving multiple stages, a physician's choice of a subject's treatment at each stage depends on the subject's history of previous treatments and outcomes. The sequence of decisions is known as a dynamic treatment regime or treatment policy. We consider dynamic treatment regimes in settings where each subject's final outcome can be defined as the sum of longitudinally observed values, each corresponding to a stage of the regime. Q-learning, which is a backward induction method, is used to first optimize the last stage treatment then sequentially optimize each previous stage treatment until the first stage treatment is optimized. During this process, model-based expectations of outcomes of late stages are used in the optimization of earlier stages. When the outcome models are misspecified, bias can accumulate from stage to stage and become severe, especially when the number of treatment stages is large. We demonstrate that a modification of standard Q-learning can help reduce the accumulated bias. We provide a computational algorithm, estimators, and closed-form variance formulas. Simulation studies show that the modified Q-learning method has a higher probability of identifying the optimal treatment regime even in settings with misspecified models for outcomes. It is applied to identify optimal treatment regimes in a study for advanced prostate cancer and to estimate and compare the final mean rewards of all the possible discrete two-stage treatment sequences. Copyright © 2015 John Wiley & Sons, Ltd.

Keywords: backward induction; multi-stage treatment; optimal treatment sequence; Q-learning; treatment decision-making

1. Introduction

A dynamic treatment regime is a mathematical formalism for what physicians do routinely when making therapeutic decisions sequentially. The physician chooses a first treatment using diagnostic information, administers it, and observes the patient's response. A second decision is based on the diagnostic information, first treatment, and newly observed response. This process may be continued, using the patient's history up to the current stage for each decision, until either a satisfactory outcome is achieved or no further treatment is considered acceptable. The dynamic treatment regime is the sequence of decision rules embedded in the sequence of alternating observations and treatments.

Methods for evaluating dynamic treatment regimes have been used increasingly for patients undergoing long-term care involving multi-stage therapies. It is challenging to identify optimal decision rules in such multi-stage treatment settings because of the complicated relationships between the alternating sequences of observed outcomes and treatments. The decision at each treatment stage depends on all observed historical data and influences all future outcomes and treatments. In turn, outcomes at each stage are affected by all previous treatments and influence all future treatment decisions. It is well known that simply optimizing the immediate outcome of each stage, which is called a myopic or greedy optimization, may not achieve the best final outcome. Simulation studies in Section 3 demonstrate this point.

Despite these complications, many approaches have been proposed to identify, estimate, or optimize

[a]*Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX 77230, U.S.A.*
[b]*Department of Internal Medicine, The University of Texas Health Science Center at Houston, Houston, TX 77030, U.S.A.*
[c]*Department of Biostatistics, The University of Michigan, Ann Arbor, MI 48109, U.S.A.*
**Correspondence to: Xuelin Huang, Department of Biostatistics - Unit 1411, The University of Texas MD Anderson Cancer Center, 1400 Pressler Street, Houston, TX 77230-1402, U.S.A.*
†*E-mail: xlhuang@mdanderson.org*

dynamic treatment regimes based on observational data. Most methods have their origins in the seminal papers by Robins and colleagues [1–5] and Murphy [6], including inverse probability of treatment weighting (IPTW) and g-estimation for structural nested models. Many variants of IPTW and g-estimation have been proposed [7–16]. They show the importance, when evaluating causal treatment effects, of correcting for bias introduced by physicians' adaptive treatment decisions based on each patient's current health status. Recent applications include estimation of survival for dynamic treatment regimes in a sequentially randomized prostate cancer trial [17] and in a partially randomized trial of chemotherapy for acute leukemia [18]. Clinical trial designs that compare multi-stage treatment strategies by adaptively re-randomizing patients have been proposed [19–22].

We address the multi-stage decision-making problem in settings where the final outcome to be optimized can be expressed as the sum of values observed at each stage. An important application is that where survival time is the final outcome and the cumulative value at each stage is the patient's current survival time. Another application is a study where the cumulative outcome is the amount of time that the patient's health index was within a specific target range. In such situations, a natural approach is to assume a conditional model for the outcome of each stage. However, because of the dependence of each outcome on previous treatments and outcomes, this approach may lead to a very complex model for the final outcome, which in turn makes global optimization of the treatment sequence difficult or intractable.

An alternative approach, called Q-learning [23–30], is to use backward induction [31] to first optimize the last stage treatment, then sequentially optimize the treatment in each previous stage. At each stage, a model is assumed for the cumulative rewards from this stage onward, with all the future stages already optimized. This approach is well suited for global optimization but depends on the correct specification of reward models at all stages. At the optimization at each stage, misspecified models cause bias, and a severe problem is that bias is carried forward from each stage to the optimization of the previous stage. Consequently, even small bias at each stage may produce a large cumulative bias for the optimization of the entire regime.

In this article, we show that a slight modification of standard Q-learning can reduce cumulative bias. At each stage, when computing the rewards from those 'already-optimized' future stage, standard Q-learning uses the maximum of model-based values. The modified Q-learning uses the actual observed rewards plus estimated loss due to sub-optimal actions. If the optimal actions were actually taken for all future stages, then the estimated loss is zero, and the actual observed rewards are used. Consequently, comparing with standard Q-learning, the modified Q-learning is more likely to use the originally observed rewards instead of model-based ones; thus, it is more robust against model misspecification and less likely to carry the bias from stage to stage during the backward reduction. This is demonstrated by simulation studies in Section 3.

When applying the modified Q-learning to a prostate cancer study in Section 4, we provide a by-product that is convenient for the practical use of both standard and modified Q-learning. In many situations where the treatment options are discrete, it is often of interest to estimate the mean rewards of all the treatment sequences. However, Q-learning does not require fully specified reward functions for all possible treatment strategies. This helps achieving simplicity. On the other hand, it fails to provide all the estimates of interest in these situations. Therefore, this seems to be a shortcoming of Q-learning. Nevertheless, in Section 4.3, we provide a trick to use Q-learning (standard or modified) to estimate the mean rewards for all possible multi-stage (discrete) treatment sequences and make inference about the reward differences between any two treatment sequences. This technique of Q-learning application is useful in practice but has not appeared in the literature, to the best of our knowledge.

The article is organized as follows. The modified Q-learning method and its properties are described in Section 2. Its performance for identifying optimal treatment regimes in settings with misspecified reward models is evaluated by simulation in Section 3. We apply the method in Section 4 to analyze data from a prostate cancer trial. A summary and discussion of the modified Q-learning method are presented in Section 5. Mathematical details are provided in Appendices.

2. Backward induction for sequential treatment optimization

We use the framework of potential (counterfactual) outcomes of all possible treatment options for each individual and make the usual assumptions [32] that (i) an individual's potential outcome under the treatment actually received is the observed outcome (consistency), (ii) given the history at each stage, the treatment decision is independent of the potential outcomes (sequential randomization or no unmeasured confounders), and (iii) all treatment strategies being considered have a positive probability

of being observed (positivity). To identify the optimal action at each stage of backward induction, the estimated reward is computed for each possible action assuming that the actions at all future stages will be optimal. This is performed by fitting a parametric model for the counterfactual future reward as a function of actions and current history. The final cumulative reward is the estimate of what the patient's total reward would be if all actions were optimal. In the sequel, we will use the terms 'payoff' and 'reward' interchangeably to mean the same thing.

2.1. Notation and method

For each subject $i = 1, \cdots, n$ and stage $s = 1, \cdots, K$, where n denotes the sample size and K denotes the total number of multiple stages, let $Z_{i,s}$ denote the time-dependent covariates measured at the beginning of s-th stage, $A_{i,s}$, the treatment or action, and $Y_{i,s}$, the observed outcome. Without loss of generality, we assume that larger values of $Y_{i,s}$ are preferable. We denote the corresponding vectors of these variables through s stages by $\bar{Z}_{i,s} = \left(Z_{i,1}, \cdots, Z_{i,s}\right)$, $\bar{A}_{i,s} = \left(A_{i,1}, \cdots, A_{i,s}\right)$, and $\bar{Y}_{i,s} = \left(Y_{i,1}, \cdots, Y_{i,s}\right)$. At stage 1, the subject's history is simply $\bar{H}_{i,1} = Z_{i,1}$. For each subsequent stage $s \geqslant 2$, the history is $\bar{H}_{i,s} = \left(\bar{Z}_{i,s}, \bar{A}_{i,s-1}, \bar{Y}_{i,s-1}\right)$. We denote the optimal action at stage s by $A_{i,s}^{opt}$, and the associated counterfactual outcome that would occur if this optimal action were taken by $Y_{i,s}^{*}$. For all $s < K$, we define the counterfactual cumulative outcome for stages s through K under $\left(A_{i,s}, A_{i,s+1}^{opt}, \cdots, A_{i,K}^{opt}\right)$ as follows:

$$Q_{i,s}^{M} =_{def} Y_{i,s} + \sum_{r=s+1}^{K} Y_{i,r}^{*}, \tag{1}$$

where each $A_{i,j}^{opt}$ is conditional on all the historic information observed prior to stage j, including previous treatments and responses. For simplicity, we have suppressed conditional notation. In words, this equation says that

$$[\text{current + future payoff}] \;=\; [\text{observed stage } s \text{ outcome}] + [\text{best possible future outcomes after stage } s]$$

That is, $Q_{i,s}^{M}$ is the future payoff, starting at stage s, if all actions from $s+1$ to K are optimized, but the actual (possibly suboptimal) action $A_{i,s}$ is taken at stage s. For the final stage, because there are no future actions to optimize, we define $Q_{i,K}^{M} = Y_{i,K}$. In our notation, $Q_{i,s}^{M}$'s are random variables, rather than mean functions as denoted by other authors.

We next define $\Delta_{i,s}$ to be the total future loss from stages s to K if action $A_{i,s}$ is taken instead of $A_{i,s}^{opt}$, while all actions from stage $s+1$ to K are optimal. Thus, if $A_{i,s} = A_{i,s}^{opt}$, then $\Delta_{i,s} = 0$, whereas if $A_{i,s}$ is not optimal, then $\Delta_{i,s} > 0$. This $\Delta_{i,s}$ is essentially Murphy's regret function [6]. Robins defined a similar blip function by comparison with a 'zero' treatment instead of the optimal one [5]. We use the counterfactual in (1) and loss $\Delta_{i,s}$ to define the cumulative future reward to patient i, from stages s to stage K, for taking the optimal action from stage s onward as

$$R_{i,s} \;=_{def}\; Q_{i,s}^{M} + \Delta_{i,s} I\left(A_{i,s} \neq A_{i,s}^{opt}\right) \;=\; Y_{i,s} + \sum_{r=s+1}^{K} Y_{i,r}^{*} \;+\; \Delta_{i,s} I\left(A_{i,s} \neq A_{i,s}^{opt}\right). \tag{2}$$

The basic idea is that the reward $R_{i,s}$ is obtained from $Q_{i,s}^{M}$ by adding back the future loss due to taking a suboptimal action at stage s. For example, if $Y_{i,s}$ is the increment in survival time for stage s, then $R_{i,s}$ is the sum of the stagewise survival times from stage s onward associated with all current and future actions being optimal, given the past treatment and response history.

Given the aforementioned structure for the future reward $R_{i,s}$ at each stage s, the counterfactual cumulative outcome in Equation (1) can be written in the more compact form

$$Q_{i,s}^{M} \;=\; Y_{i,s} + R_{i,s+1}, \quad \text{for} \quad s < K. \tag{3}$$

If all optimal actions and $\Delta_{i,s}$'s were known, one could simply work backward and compute $R_{i,s}$ for each $s = K, K-1, \cdots, 1$, and thus obtain the final payoff $R_{i,1}$. To derive $A_{i,K}^{opt}, \cdots, A_{i,1}^{opt}$ in the steps of the backward induction, we will exploit the decomposition given by Equations (1)–(3) by assuming an

additive parametric regression model for $Q_{i,s}^M$ as a function of $A_{i,s}$ and the most recent history $\bar{H}_{i,s}^M = \left(Z_{i,s-1}, Z_{i,s}, Y_{i,s-1}, A_{i,s-1}\right)^T$. We formulate the regression model in terms of $\bar{H}_{i,s}^M$, rather than the complete history $\bar{H}_{i,s}$, to control the number of parameters so that the method may be applied feasibly. This requires us to make a Markovian assumption. The fitted model will provide estimates of the $\Delta_{i,s}$'s and thus identify the $A_{i,s}^{opt}$ values and thus the optimal payoffs.

Appendix A provides details of the parametric regression model that we assume for $Q_{i,s}^M$. In particular, Appendix D provides an illustration for a special case of three treatment stages with three treatment options each stage. An important aspect of our approach is that we assume a regression model for $Q_{i,s}^M$, rather than assuming simple models for Y_s, $1 \leqslant s \leqslant K$, which could result in a model for $Y = \sum_{s=1}^K Y_s$ that may be too complicated for optimization.

2.2. *Backward induction*

Our method requires the the cumulative causal effect of treatment j versus l at stage s, which we define formally, for $j > l$, as

$$D_{i,s}(j,l) = Q_{i,s}^M\left(A_{i,s} = j, \bar{H}_{i,s}^M\right) - Q_{i,s}^M\left(A_{i,s} = l, \bar{H}_{i,s}^M\right)$$

In words, this cumulative causal effect is

[current + future payoff for action j at stage s] $-$ [current + future payoff for action l at stage s].

Substituting the parameter estimates obtained from the fitted regression model, given in Appendix A, gives the estimated cumulative payoff, $\hat{Q}_{i,s}^M$, estimated causal effects, $\hat{D}_{i,s}(j,l)$, and estimated cumulative future rewards $\hat{R}_{i,s}$. The backward induction is carried out as follows:

Step 1. Start with $s = K$ and set $Q_{i,K}^M = Y_{i,K}$.
Step 2. For the current step s,

Step 2.1 Fit the regression model (15) for $Q_{i,s}^M$ to obtain $\left(\hat{\beta}_s, \hat{\psi}_s^{(2)}, \cdots, \hat{\psi}_s^{(J_s)}\right)$ and thus $\hat{Q}_{i,s}^M$.
Step 2.2 Use the estimated causal effects $\hat{D}_{i,s}(j,l)$ given by (16) to identify the estimated optimal action $\hat{A}_{i,s}^{opt} = \text{argmax}_{1 \leqslant j \leqslant J_s}\{\max_{l \neq j} \hat{D}_{i,s}(j,l)\}$.
Step 2.3 Define the estimated future loss due to taking action $A_{i,s}$ to be $\hat{\Delta}_{i,s} =_{def} \hat{Q}_{i,s}^M\left(\hat{A}_{i,s}^{opt}\right) - \hat{Q}_{i,s}^M(A_{i,s})$.
Step 2.4 By (2), step 2.3 gives the estimate $\hat{R}_{i,s} =_{def} Q_{i,s}^M + \hat{\Delta}_{i,s} I\left(A_{i,s} \neq \hat{A}_{i,s}^{opt}\right)$.

Step 3. If $s > 1$, decrement $s \to s-1$. By (3), set $Q_{i,s}^M = Y_{i,s} + \hat{R}_{i,s+1}$, and go to Step 2.1 If $s = 1$, stop.

At the end of these steps, $\hat{A}_{i,1}^{opt}, \cdots, \hat{A}_{i,K}^{opt}$, the optimal treatments for all subjects at all stages, have been identified, and $\hat{R}_{i,1}$ is the estimated total payoff from taking these estimated optimal actions. With this algorithm, the optimization is global rather than myopic or local.

Asymptotic properties of the estimators are given in Appendix B. The 'Sandwich' formula [33] is used to account for the extra variation because of plugging in an estimator from a late treatment stage into the regression models for an early stage.

2.3. *Comparison with standard Q-learning*

The method described previously is a robust modification of standard Q-learning [23, 28]. For all treatment stages except the last, to estimate counterfactual outcomes under optimal actions, standard Q-learning uses predicted values from previously fitted linear models plus estimated loss due to suboptimal actions. In contrast, our modified Q-learning method uses the values actually observed plus the estimated loss. Let $Q_{i,s}^{std}$ denote the standard Q-learning method's objective function that plays the role of our function $Q_{i,s}^M$. For the first step of the backward induction used in our method, $Q_{i,K}^{std} = Q_{i,K}^M$.

Standard Q-learning does backward induction using the same steps as our backward induction algorithm through step 2.3, but it uses the estimated stage K reward

$$\hat{R}^{std}_{i,K} =_{def} \hat{Q}^{std}_{i,K} + \hat{\Delta}_{i,K} I\left(A_{i,K} \neq \hat{A}^{opt}_{i,K}\right).$$

In contrast, our estimated stage K reward is

$$\hat{R}_{i,K} =_{def} Q^{M}_{i,K} + \hat{\Delta}_{i,K} I\left(A_{i,K} \neq \hat{A}^{opt}_{i,K}\right).$$

The difference is that standard Q-learning uses predicted values $\hat{Q}^{std}_{i,K}$ obtained from a regression model, while our modified Q-learning method uses the observed values $Q^{M}_{i,K}$. This difference is carried to the next stage, $s = K-1$, through the formulations of $\hat{Q}^{std}_{i,K-1} = Y_{K-1} + \hat{R}^{std}_{i,K}$ and $Q^{M}_{i,K-1} = Y_{K-1} + \hat{R}_{i,K}$. Similar differences are accumulated during the iteration of this process in the backward induction steps from $s = K-1$ to $s = 1$.

The modified Q-learning method has the following advantages. First, $Q^{M}_{i,s}$ uses observed outcomes whenever possible for any $s \leqslant K$, whereas $Q^{std}_{i,s}$ uses model-based expectations for any $s < K$. Retaining the original outcomes helps the modified Q-learning rely less on the specification of the models used in (15), and thus improves robustness. This is shown by our simulation study in the next section when the model (15) is misspecified, for example, when some relevant covariates are not included in the data set. The simulations show that $Q^{M}_{i,s}$ has a more robust performance than $Q^{std}_{i,s}$.

The second advantage of the modified Q-learning method follows from the fact that definition (3) ensures $Q^{M}_{i,s} \geqslant \sum_{r=s}^{K} Y_{i,r}$. This means that the predicted reward under optimal treatment regimes for stage $s+1$ onward is always at least the observed reward under the actual regimes, which may be suboptimal. This is a desirable property that does not always hold for standard Q-learning because, in practice, one may observe $Q^{std}_{i,s} < \sum_{j=s}^{K} Y_i$ for some $s < K$. This happens simply because, for some subjects, the predicted rewards under optimal treatment regimes for stage $s+1$ onward are less than their observed actual reward. Furthermore, for $s < K$, if a patient has received the optimal treatment regimes for stage $s+1$ onward, then with the modified Q-learning, because $\Delta_{i,r} = 0$ for all $r \geqslant s+1$, the potential outcome under the treatment sequence $\left\{A_{is}, A^{opt}_{i,s+1}, \cdots, A^{opt}_{i,K}\right\}$ for this patient, is $Q^{M}_{i,s} = \sum_{r=s}^{K} Y_{i,r}$, the observed reward from stage s onward. This is in agreement with the 'consistency assumption'. This assumption, stated at the beginning of Section 2, requires that the assumed counterfactual outcomes under the actual observed actions must be equal to the observed outcomes. It is a very natural assumption and commonly required in causal inference [34]. In contrast, with standard Q-learning, $Q^{std}_{i,s}$ may not equal $\sum_{r=s}^{K} Y_{i,r}$, even if a patient receives the optimal treatment regimes for stage $s+1$ onward. That is, as an estimate of the counterfactual outcome, $Q^{std}_{i,s}$ may violate the consistency assumption on individual basis, although it satisfies this assumption in expectation when the reward models are correctly specified. We will compare the performance of standard and modified Q-learning in the next section by simulation.

3. Simulation studies

The correct specification of reward models is very important for Q-learning [32]. In this section, we use simulations to show that, in some scenarios, when the reward models are misspecified, the modified Q-learning outperforms standard Q-learning. For simplicity, we evaluate two-stage treatment sequences. Sample sizes 50, 100, 200, and 400 are considered. We use three scenarios, each simulation scenario is replicated 1000 times.

3.1. Scenario I

In scenario I, we assume an unobserved variable $V \sim \text{Normal}(0, 2^2)$. For the first treatment stage, we generate covariate $Z_1 \sim \text{Normal}(0,1)$, treatment $A_1 \sim \text{Bernoulli}(0.5)$, and outcome $Y_1 = Z_1(A_1 - 0.5) + V + \varepsilon_1$, with $\varepsilon_1 \sim \text{Normal}(0,1)$. The second stage treatment $A_2 \sim \text{Bernoulli}(0.5)$, and outcome $Y_2 = -2Z_1(A_1 - 0.5) + (A_1 - 0.5)(A_2 - 0.5) - V + \varepsilon_2$, with $\varepsilon_2 \sim \text{Normal}(0,1)$. The final cumulative outcome is $Y = Y_1 + Y_2$. With the observed data of $(Z_1, A_1, Y_1, A_2, Y_2)$ for all subjects, the goal is to find the optimal two-stage treatment regimes that maximizes Y.

In this scenario, both treatments A_1 and A_2 are randomized. The optimal stage 2 treatment is $A_2^{opt} = I(A_1 = 1)$. Then the reward for stage 2 under A_2^{opt} is $R_2 = -2Z_1(A_1 - 0.5) + 0.25 - V + \varepsilon_2$. Recalling that $Q_2^M = Y_2$, and $Q_1^M = Y_1 + R_2 = -Z_1(A_1 - 0.5) + 0.25 + \varepsilon_1 + \varepsilon_2$. We assume the following model to optimize A_1.

$$Q_1^M = \beta_{10} + \beta_{11}Z_1 + A_1(\psi_{10} + \psi_{11}Z_1) + e_1, \tag{4}$$

The true values for the aforementioned parameters are $\beta_1 = (\beta_{10}, \beta_{11})^T = (0.25, 0.5)^T$ and $\psi_1 = (\psi_{10}, \psi_{11})^T = (0, -1)^T$. The optimal stage 1 treatment is $A_1^{opt} = I(Z_1 < 0)$. If we use a myopic strategy to optimize A_1 by maximizing $Y_1 = Z_1(A_1 - 0.5) + V + \varepsilon_1$, we will obtain a wrong solution $A_1^{opt} = I(Z_1 \geqslant 0)$.

To apply the modified Q-learning, we first fit the following model,

$$\begin{aligned} Q_2^M = Y_2 = {} & \beta_{20} + \beta_{21}Z_1 + \beta_{22}A_1 + \beta_{23}Y_1 \\ & + A_2\left(\psi_{20} + \psi_{21}Z_1 + \psi_{22}A_1 + \psi_{23}Y_1\right) + e_2. \end{aligned} \tag{5}$$

From the data generation mechanism, it can be derived that the true values for the aforementioned parameters are $\beta_2 = (\beta_{20}, \beta_{21}, \beta_{22}, \beta_{23})^T = (0.25, 0, -0.5, -0.857)^T$, and $\psi_2 = (\psi_{20}, \psi_{21}, \psi_{22}, \psi_{23})^T = (-0.5, 0, 1, 0)^T$. The coefficient $\beta_{23} = -0.857$ is the effect of V on Y_2 through Y_1, namely, $\beta_{23} = E(Y_1Y_2)/E(Y_1^2)$. Details about this derivation is provided in Appendix C. After fitting the above model to obtain $\hat{\beta}_2$ and $\hat{\psi}_2$, the estimated optimal stage 2 treatment is

$$\hat{A}_2^{opt} = I\left(\hat{\psi}_{20} + \hat{\psi}_{21}Z_1 + \hat{\psi}_{22}A_1 + \hat{\psi}_{23}Y_1 > 0\right). \tag{6}$$

Then let

$$\begin{aligned} \hat{R}_2 &= Y_2 + I\left(A_2 \neq \hat{A}_2^{opt}\right)\left|\hat{\psi}_{20} + \hat{\psi}_{21}Z_1 + \hat{\psi}_{22}A_1 + \hat{\psi}_{23}Y_1\right|, \\ Q_1^M &= Y_1 + \hat{R}_2, \end{aligned} \tag{7}$$

with $|\cdot|$ denotes absolute value. We use the outcome Q_1^M to fit the model in (4). After estimators $\hat{\beta}_1$ and $\hat{\psi}_1$ are obtained, the estimated optimal stage 1 treatment is

$$\hat{A}_1^{opt} = I(\hat{\psi}_{10} + \hat{\psi}_{11}Z_1 > 0). \tag{8}$$

The simulation results for samples of size 200 are given Tables I and II (each first panel from the left, scenario I). In general, the bias is small, and the empirical and asymptotic standard errors (SE and ASE) match well, with coverage probabilities of the 95% confidence intervals all close to nominal. The modified Q-learning correctly identified the optimal stage 1 and stage 2 treatments 91.1% and 88.4% of the time, respectively. Parameter estimations for other sample sizes (n=50, 100, or 400) are also performed well by the modified Q-learning (results not shown).

We apply standard Q-learning to the same data sets. Both standard and modified Q-learning fit the same regression models (5) for treatment stage 2. Naturally, they obtain exactly the same results for stage 2 (Table II) but differ for the stage 1 estimates (Table I). As shown in Equation (7), the outcome used by the modified Q-learning is the actually observed values Y_2 plus the estimated loss due to suboptimal stage 2 actions. In contrast, the outcome used by standard Q-learning for stage 1 is the predicted value $\hat{Y}_2$ from stage 2 by model (5) plus the same estimated loss, as follows.

$$\begin{aligned} \hat{R}_2^{std} &= \hat{Y}_2 + I\left(A_2 \neq \hat{A}_2^{opt}\right)\left|\hat{\psi}_{20} + \hat{\psi}_{21}Z_1 + \hat{\psi}_{22}A_1 + \hat{\psi}_{23}Y_1\right|, \\ Q_1^{std} &= Y_1 + \hat{R}_2^{std}. \end{aligned} \tag{9}$$

Using Q_1^{std} to replace Q_1^M in (4), the same linear regression model in (4) is fit to identify the optimal stage 1 treatments. Note that Q_1^M carries this information from V by using the original data Y_2, but Q_1^{std} discards this information by using the model based value $\hat{Y}_2$. Due to this difference, the two methods obtain different stage 1 estimates (Table I). Standard Q-learning gives biased estimates for β_{11} and ψ_{11}. Consequently, the probability that it correctly identifies the optimal stage 1 treatments is only 38.2%, much lower than that achieved by the modified Q-learning (91.1%). In summary, standard Q-learning uses model-based values in the construction of counterfactual outcomes and is prone to bias introduced by

Table I. Parameter estimates for treatment stage 1.

		Modified Q-Learning				Standard Q-Learning				g-estimation*				Regret Minimization*			
	True	Est	SE	ASE	CP	Est	SE	ASE	CP	Est	SE	ASE	CP	Est	SE	ASE	CP
Scenario I: $\Pr(A_1 = 1) = \Pr(A_2 = 1) = 0.5$																	
β_{10}	0.25	0.282	0.180	0.243	0.988	0.281	0.192	0.202	0.952								
β_{11}	0.5	0.489	0.171	0.200	0.980	−0.068	0.157	0.147	0.039								
ψ_{10}	0	−0.008	0.270	0.341	0.988	−0.004	0.281	0.282	0.953	−0.009	0.269	0.256	0.947	0.009	0.264	0.256	0.928
ψ_{11}	−1	−0.969	0.264	0.263	0.933	0.139	0.178	0.178	0.000	−0.980	0.263	0.249	0.926	−1.012	0.274	0.258	0.922
Scenario II: $\text{logit}\{\Pr(A_1 = 1)\} = Z_1$, $\text{logit}\{\Pr(A_2 = 1)\} = 0.3Y_1$																	
β_{10}	0.25	0.274	0.199	0.263	0.989	0.044	0.205	0.217	0.851								
β_{11}	0.5	0.483	0.193	0.217	0.970	−0.072	0.177	0.159	0.074								
ψ_{10}	0	0.008	0.274	0.369	0.993	0.009	0.291	0.304	0.942	0.008	0.275	0.276	0.945	0.021	0.293	0.282	0.934
ψ_{11}	−1	−0.963	0.288	0.287	0.938	0.147	0.202	0.194	0.000	−0.990	0.329	0.315	0.926	−1.051	0.346	0.347	0.948
Scenario III: $\text{logit}\{\Pr(A_1 = 1)\} = Z_1 + V$, $\text{logit}\{\Pr(A_2 = 1)\} = 0.3Y_1 + V$																	
β_{10}	0.25	0.419	0.194	0.378	0.996	0.289	0.192	0.348	0.999								
β_{11}	0.5	0.556	0.177	0.229	0.977	0.098	0.168	0.179	0.393								
ψ_{10}	0	−0.086	0.253	0.474	0.998	−0.087	0.256	0.429	0.995	−0.056	0.270	0.274	0.945	−0.038	0.284	0.301	0.949
ψ_{11}	−1	−0.982	0.261	0.281	0.950	−0.066	0.206	0.196	0.002	−0.991	0.264	0.284	0.959	−1.016	0.287	0.339	0.968

*g-estimation and regret minimization specify treatment selection probabilities (not shown), but do not need β_{10} or β_{11}. Est, mean of estimates; SE, empirical standard error; ASE, average of standard error estimates; ASE for g-estimation and regret minimization obtained by 200 bootstrap samples. CP, coverage probability of 95% confidence interval.

Table II. Parameter estimates for treatment stage 2.

		Standard or modified Q-Learning*				g-estimation**				Regret Minimization**			
	True	Est	SE	ASE	CP	Est	SE	ASE	CP	Est	SE	ASE	CP
Scenario I: $\Pr(A_1 = 1) = \Pr(A_2 = 1) = 0.5$													
β_{20}	0.25	0.255	0.207	0.210	0.949								
β_{21}	0	0.002	0.170	0.149	0.907								
β_{22}	−0.5	−0.495	0.298	0.297	0.947								
β_{23}	−0.857	−0.858	0.063	0.065	0.960								
ψ_{20}	−0.5	−0.510	0.294	0.297	0.952	−0.510	0.284	0.289	0.952	−0.509	0.282	0.292	0.947
ψ_{21}	0	−0.003	0.240	0.212	0.919	−0.004	0.229	0.227	0.946	−0.014	0.214	0.219	0.955
ψ_{22}	1	0.997	0.413	0.420	0.949	0.997	0.378	0.396	0.959	1.027	0.410	0.420	0.948
ψ_{23}	0	0.004	0.091	0.093	0.952	0.004	0.091	0.094	0.950	−0.002	0.100	0.097	0.932
Scenario II: $\text{logit}\{\Pr(A_1 = 1)\} = Z_1$, $\text{logit}\{\Pr(A_2 = 1)\} = 0.3Y_1$													
β_{20}	0.25	0.019	0.217	0.220	0.823								
β_{21}	0	0.006	0.194	0.165	0.904								
β_{22}	−0.5	−0.501	0.321	0.319	0.942								
β_{23}	−0.857	−0.842	0.070	0.069	0.939								
ψ_{20}	−0.5	−0.517	0.309	0.316	0.965	−0.514	0.310	0.318	0.955	−0.476	0.325	0.324	0.944
ψ_{21}	0	−0.008	0.268	0.230	0.906	−0.004	0.269	0.258	0.933	−0.002	0.266	0.271	0.960
ψ_{22}	1	1.023	0.455	0.456	0.943	1.019	0.453	0.464	0.945	0.996	0.488	0.488	0.938
ψ_{23}	0	−0.005	0.097	0.096	0.954	−0.005	0.107	0.107	0.949	0.001	0.114	0.119	0.957
Scenario III: $\text{logit}\{\Pr(A_1 = 1)\} = Z_1 + V$, $\text{logit}\{\Pr(A_2 = 1)\} = 0.3Y_1 + V$													
β_{20}	0.25	0.639	0.233	0.235	0.617								
β_{21}	0	0.343	0.182	0.157	0.422								
β_{22}	−0.5	−1.31	0.394	0.364	0.391								
β_{23}	−0.857	−0.724	0.0871	0.0873	0.669								
ψ_{20}	−0.5	−1.32	0.401	0.363	0.407	−0.965	0.409	0.399	0.774	−1.089	0.393	0.392	0.667
ψ_{21}	0	−0.539	0.245	0.218	0.315	−0.020	0.327	0.314	0.934	0.046	0.316	0.353	0.966
ψ_{22}	1	1.725	0.548	0.503	0.697	1.020	0.591	0.579	0.932	0.969	0.606	0.625	0.953
ψ_{23}	0	−0.010	0.124	0.123	0.948	−0.005	0.202	0.206	0.953	−0.009	0.222	0.238	0.950

*Standard and modified Q-Learning methods use the same estimating equations and give the same results for the last treatment stage;

**g-estimation and regret minimization specify treatment selection probabilities (not shown) but do not need βs. Est, mean of estimates; SE, empirical standard error; ASE, average of standard error estimates; ASE for g-estimation and regret minimization obtained by 200 bootstrap samples. CP, coverage probability of 95% confidence interval.

model misspecification. As the number of treatment stages increases, the model-based values will be used more times during the backward induction, and this bias problem will become more severe. In contrast, the modified Q-learning achieves robustness against model misspecification by using the original data instead of model-based values whenever possible.

Because the main goal of Q-learning is to correctly identify optimal treatments, we conducted additional simulations for a range of sample sizes and compared performance of modified Q-learning with standard Q-learning. We also compared their performances with the myopic strategy that uses Y_1 to optimize A_1 and Y_2 to optimize A_2. Figure 1 shows these probabilities under a range of sample sizes. The modified Q-learning has larger probabilities than standard Q-learning to correctly identify the optimal stage 1 treatments. Both the modified and standard Q-learning have much better performances than the myopic strategy.

It is also interesting to note that in this setting with misspecified reward models, the optimal treatment selection power of the modified Q-learning increases with sample size, but this is not true for either standard Q-learning or myopic optimization. This shows that in situations of model misspecification, a large sample size cannot remedy a non-robust or incorrectly designed optimization algorithm and may even make things worse. Specifically in this simulation setting, because there are only two treatment options at each stage, a pure random selection of any one of them has a probability of 50% of being correct. The less than 50% power of standard Q-learning and myopic optimization shown in Figure 1 reveal that they are severely biased in such a situation with misspecified reward models. It also explains why their empirical power levels decrease with sample size. This is because their true power levels are less than 50% as $n \to \infty$, and equal to 50% as $n \to 0$ (equivalent to a pure random selection).

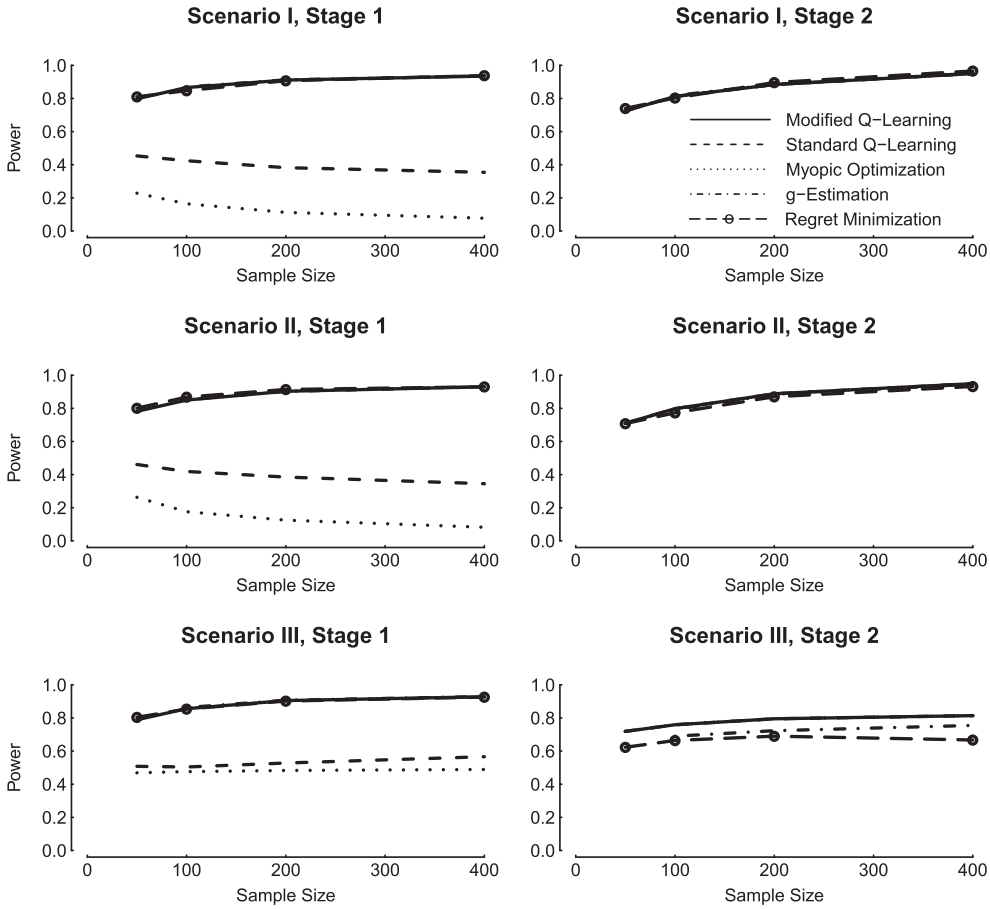

Figure 1. Comparisons between the modified and standard Q-learning, myopic optimization, g-estimation and regret minimization: probability (power) of correctly identifying the optimal stage 1 and stage 2 treatments in scenarios I, II, and III. Note: For stage 1, in all scenarios, the power curves for the modified Q-learning, g-estimation and regret minimization almost overlap with each other. For stage 2, in all scenarios, the power curves for the modified and standard Q-learning and myopic optimization overlap with each other, so that only the solid curve (modified Q-learning) is visible; they almost overlap with those for g-estimation and regret minimization in scenarios I and II and are higher than those for g-estimation and regret minimization in scenario III.

3.2. Scenarios II and III, and other optimization methods

Treatments in scenario I are randomized. We also consider other treatment selection models. In scenario II, treatment assignment probabilities depend on observed covariates and outcomes, namely, $A_1 \sim \text{Bernoulli}(Z_1)$, $A_2 \sim \text{Bernoulli}(0.3Y_1)$. In scenario III, treatment assignments further depend on the unobserved variable V, as in the succeeding texts, $A_1 \sim \text{Bernoulli}(Z_1 + V)$, $A_2 \sim \text{Bernoulli}(0.3Y_1 + V)$. All the other data generation mechanisms remain the same as in scenario I. The same data analyses shown in the previous subsection for scenario I by standard and modified Q-learning, and by the myopic optimization method, are also conducted for scenarios II and III.

As suggested by a reviewer, we also compare the proposed estimator to the estimators by Murphy [6] and Robins [5]. Moodie *et al*. [25] provided a nice description of these estimators together with R functions for implementing them, which are used below with combination of her and our notations.

In our setting of data generation, the g-estimator [5] starts with the following estimation functions. Denote by $\mathfrak{h}_j$ the observed history prior to A_j, $j = 1, 2$. Let $\gamma_1(\mathfrak{h}_1, a_1; \psi) = a_1(\psi_{10} + \psi_{11}z_1)$ and $\gamma_2(\mathfrak{h}_2, a_2; \psi) = a_2(\psi_{20} + \psi_{21}z_1 + \psi_{22}a_1 + \psi_{23}y_1)$ be the blip functions [5] defining the expected differences in outcome, respectively, between treatments $A_1 = a_1$ and $A_1 = 0$, and between $A_2 = a_2$ and $A_2 = 0$. Consequently, this would identify the optimal stage 2 treatment as $d_2^{opt}(\mathfrak{h}_2) = I(\psi_{20} + \psi_{21}z_1 + \psi_{22}a_1 + \psi_{23}y_1 > 0)$, and $d_1^{opt}(\mathfrak{h}_1) = I(\psi_{10} + \psi_{11}z_1 > 0)$. By the data generation described earlier, we have $\psi_1 = (\psi_{10}, \psi_{11})^T = (0, -1)^T$ and $\psi_2 = (\psi_{20}, \psi_{21}, \psi_{22}, \psi_{23})^T = (-0.5, 0, 1, 0)^T$. Denote $S_1(a_1) = a_1(1, z_1)^T$

and $S_2(a_2) = a_2(1, z_1, a_1, y_1)^T$. Define

$$H_{mod,2}(\psi) = Y - \gamma_2(\mathfrak{h}_2, a_2; \psi) \tag{10}$$

$$H_{mod,1}(\psi) = Y - \gamma_1(\mathfrak{h}_1, a_1; \psi) + \{\gamma_2(\mathfrak{h}_2, d_2^{opt}; \psi) - \gamma_2(\mathfrak{h}_2, a_2; \psi)\} \tag{11}$$

Also assume treatment selection models as follows:

$$\text{logit}\{\Pr(A_1 = 1|Z_1)\} = \alpha_{10} + \alpha_{11} Z_1 \,,$$

$$\text{logit}\{\Pr(A_2 = 1|Z_1, A_1, Y_1)\} = \alpha_{20} + \alpha_{21} Z_1 + \alpha_{22} A_1 + \alpha_{23} Y_1 \,,$$

The aforementioned resulting estimator $\hat{\alpha}$ is used in the following estimating equation to estimate ψ and then find out the optimal treatment regimes

$$U_j^{\dagger}(\psi, s, \hat{\alpha}) = \left[H_{mod,j}(\psi) - E\{H_{mod,j}(\psi)|\mathfrak{h}_j\}\right] \times \left[S_j(a_j) - E\{S_j(A_j)|\mathfrak{h}_j\}\right] \,, \qquad j = 1, 2. \tag{12}$$

Murphy [6] defined a regret function as the loss due to taking a suboptimal action. In the above setting, the regret functions for the two stages can be written as $\mu_j(\mathfrak{h}_j, a_j) = \max_a\{\gamma_j(\mathfrak{h}_j, a) - \gamma_j(\mathfrak{h}_j, a_j)\}$ [25]. If an optimal action is taken, that is, $a_j = d_j^{opt}(\mathfrak{h}_j)$, then the regret is zero, namely, $\mu_j(\mathfrak{h}_j, a_j) = 0$. The true regret functions are $\mu_1(\mathfrak{h}_1, a_1) = |z_1|I\{a_1 \neq I(z_1 < 0)\}$ and $\mu_2(\mathfrak{h}_2, a_2) = |a_1 - 0.5|I(a_2 \neq a_1)$. Using the methods proposed by Murphy [6] and also adopted by Moodie *et al*. [25], logistic functions as the following are used to approximate these piecewise linear functions

$$\mu_1^*(z_1, a_1) = (\psi_{10} + \psi_{11} z_1)\left\{\frac{\exp\{30(\psi_{10} + \psi_{11} z_1)\}}{1 + \exp\{30(\psi_{10} + \psi_{11} z_1)\}} - a_1\right\} \tag{13}$$

$$\mu_2^*(z_1, a_1, y_1, a_2) = (\psi_{20} + \psi_{21} z_1 + \psi_{22} a_1 + \psi_{23} y_1)\left\{\frac{\exp\{30(\psi_{20} + \psi_{21} z_1 + \psi_{22} a_1 + \psi_{23} y_1)\}}{1 + \exp\{30(\psi_{20} + \psi_{21} z_1 + \psi_{22} a_1 + \psi_{23} y_1)\}} - a_2\right\}. \tag{14}$$

The true values of the above ψs are the same as that for the g-estimation.

For all the methods, the parameter estimates and their empirical and asymptotic standard errors are reported in Tables I and II. The standard and modified Q-learning and the myopic optimization use more parameters for their outcome models (both βs and ψs), whereas the g-estimation and regret minimization use less parameters for their outcome models (only the ψs). However, they use logistic regression models for treatment selection, for which the estimated parameters are not reported. Similarly, as Moodie et al. [25], the bootstrap method (200 randomly drawn samples from the original data with replacement) are used to compute the asymptotic standard errors for the g-estimation and regret minimization. In the aforementioned tables, standard Q-learning estimators show substantial bias for the stage 1 parameters β_{11} and ψ_{11} in all the three scenarios. The modified Q-learning, g-estimation and regret minimization have only very small bias in scenarios I and II, but some bias in scenario III, in which both the outcome and treatment assignment models are misspecified.

The power levels for all the methods to correctly identify the optimal treatments are depicted in Figure 1. For stage 1, in all three scenarios, the modified Q-learning, g-estimation and regret minimization perform almost the same, whereas the myopic optimization and standard Q-learning have poor performance. For stage 2, the standard and modified Q-learning, and the myopic optimization have the same performance. Comparing with them, the g-estimation and regret minimization have slightly lower power levels in scenario III, but the same power levels in scenarios I and II. In this figure, the case of $n = 50$ for g-estimation is not shown because it involves singular matrices and other computation issues.

Table III. Estimates of Δ_2 in computation of optimal potential stage 2 score $R_2 = Y_2 + \Delta_2$.

Stage 1 treatment	Optimal stage 2 treatment conditional on stage 1 treatment	Actual stage 2 treatment			
		CVD	KA/VE	TEC	TEE
CVD	TEC	NA	0.03	0	0.24
KA/VE	TEC	0.5	NA	0	0.55
TEC	CVD	0	0.175	NA	0.25
TEE	TEC	0.125	0.125	0	NA

4. Application to a prostate cancer study

We applied the modified Q-learning to analyze data from a clinical trial of advanced prostate cancer conducted at MD Anderson Cancer Center from 1998 to 2006 to evaluate multi-stage therapeutic strategies [17, 19]. One hundred and fifty patients with advanced prostate cancer were randomized at enrollment to receive one of four chemotherapy combinations, abbreviated as CVD, KA/VE, TEC, and TEE, during an initial treatment period of 8 to 24 weeks. Thereafter, response-based assignment to the second stage treatment was made. Patients with a favorable response to the initial treatment stayed on the same treatment during the second stage ('respond $\mapsto$ stay'), while patients who did not have a favorable response were randomized among the three remaining treatments ('no response $\mapsto$ switch'). Because 47 patients did not follow this protocol because of severe toxicities or progressive disease or other reasons, Wang *et al.* [17] defined viable dynamic treatment regimes including such discontinuation and accounting for both efficacy and toxicity. This evaluation was based on expert score defined from the bivariate outcomes of efficacy and toxicity in each stage. The scores for the first and second stages were denoted by Y_1 and Y_2, respectively. It was further specified that patients who went off treatment during the first stage received a score of $Y_2 = 0$ for stage 2. We used the modified Q-learning to identify the optimal treatments for the two stages that maximized $Y = Y_1 + Y_2$. The data set included the following covariates: patient age, radiation treatment (yes or no), length of time hormone therapy was received (in months) before registration, location of evidence of disease at enrollment, strata (low or high risk), baseline prostate-specific antigen level, and alkaline phosphatase hemoglobin concentrations.

4.1. Stage 2 estimation

By design, patients with a favorable response in stage 1 had that treatment repeated, and we assumed that they received the optimal A_2. Because patients whose first stage treatment failed were re-randomized, this produced a saturated factorial design with 12 different two-stage treatment sequences. Because of the limited sample size, we fit a model with 12 indicators for the 12 treatment sequences, without including their interactions with patients' characteristics. The fitted model showed that for patients who received TEC in stage 1 and did not have a favorable response, the best stage 2 treatment was CVD. For patients who did not receive TEC in stage 1 and did not have a favorable response, the best stage 2 treatment was TEC. The computation of potential stage 2 scores under the aforementioned optimal stage 2 treatment is shown in Table III. If the stage 1 treatment failed, the score indicated in Table III is added to each patient's actual stage 2 score, Y_2, to obtain a hypothetical optimal score, R_2, which is used in the next step of the analysis. For patients who had a favorable response in stage 1 treatment, we set $R_2 = Y_2$. For patients who went off treatment during the first stage, because they did not receive any stage 2 treatment, they could not be used in the estimation of stage 2 treatment effects. They were still included in the analyses for stage 1 and overall outcomes by assigning $R_2 = 0$. This had an impact on the interpretation of the identified optimal regimes, as shown in the next subsection.

4.2. Stage 1 estimation

After the stage 2 estimation, we defined

$$Q_1^M = Y_1 + R_2.$$

For the four stage 1 treatments, we fit a linear regression model for all main effects and interactions associated with the stage 1 treatment with response Q_1^M. All covariates mentioned at the beginning of

Table IV. Linear regression for the effects of stage 1 treatments on the potential final outcome if their corresponding optimal stage 2 treatments had been received.

	Estimate	SE	p-value
Intercept	1.248	0.1004	< 0.001
Age	−0.0109	0.0053	0.039
KA/VE versus CVD	0.2366	0.1286	0.066
TEC versus CVD	0.2757	0.1294	0.033
TEE versus CVD	0.0475	0.1370	0.729

this Section 4 were considered. Using the Akaike information criterion (AIC) to conduct a stepwise variable selection, we found that age seemed to be the only significant covariate. Interactions between age and treatments were not statistically significant. Age was centered at 65 years, which is roughly the mean. The fitted model given in Tables III and IV shows that the stage 1 treatment may be ranked in the following order: TEC, KA/VE, TEE, and CVD, and they roughly can be put into two groups, {TEC, KA/VE} and {TEE, CVD}, with substantial difference between the groups, but not much difference within either group. Combining these results with those in Table III, which show the optimal stage 2 treatment conditional on stage 1 treatment, we conclude that the optimal treatment sequence (strategy) for these patients is as follows. Start with initial treatment TEC. If a patient achieves a favorable response, then continue to treat with TEC in the second stage. Otherwise, that is, if a patient does not achieve a favorable response to the initial treatment, then treat with CVD in the second stage. We denote this regime by (TEC, CVD). Other regimes are denoted similarly.

The estimates in Table IV are not for stage 1 outcomes only, but rather for the mean final rewards if the stage 2 treatments had been optimized conditional on the stage 1 treatment. For example, compared with CVD, the initial treatment TEC could have improved mean final outcome score by 0.2757 (standard deviation = 0.1294), if all subjects had received their respective optimal stage 2 treatments conditional on their stage 1 treatments. Referring to Table III, the optimal two-stage treatment strategy is (TEC, CVD) for subjects who receive TEC in stage 1, and is (CVD, TEC) for subjects whose stage 1 treatments are CVD. The noted difference of 0.2757 in Table IV between initial treatments TEC and CVD is actually the difference between the two regimes (TEC, CVD) and (CVD, TEC). This difference is statistically significant ($p = 0.033$).

4.3. Estimation for the mean rewards of 16 regimes

Similar to standard Q-learning, the modified Q-learning does not require fully specified reward functions for all possible treatment strategies. For the aforementioned example, combining the results in Tables III and IV, we have estimated the mean rewards of the following four regimes: (TEC, CVD), (KA/VE, TEC), (TEE, TEC), and (CVD, TEC). However, we have not obtained estimates for other regimes, for example, (TEC, TEE). There are 12 such regimes. This might be viewed as an inconvenience for Q-learning or the modified Q-learning. One may try to introduce some extra models to estimate the mean rewards for the other 12 regimes. However, we show in the succeeding discussions that this is unnecessary.

In Table III, our purpose was to identify the optimal regimes, thus we used the optimal stage 2 treatments as references and computed the potential loss Δ_2 due to not taking the optimal stage 2 treatment. When our purpose is to compute mean final rewards for other regimes rather than identifying optimal regimes, we replace those optimal stage 2 treatments in Table III by the treatments for which we intend to estimate and use them as the new references. Then we figure out the new potential loss (or gain) Δ_2' due to not taking the new reference treatments in stage 2 and compute the final reward values for regimes using the new reference treatments in stage 2. We put these Δ_2' values in Table V. For convenience, we copy Table III to the top of Table V. The middle part of Table V shows the new Δ_2' values for estimating the mean final rewards for the following regimes, (CVD, KA/VE), (KA/VE, CVD), (TEC, TEE), and (TEE, CVD), which are referred to as target regimes. With these Δ_2', we define a new final reward $Q_1^{M'} = Y_1 + R_2'$, with $R_2' = Y_2 + \Delta_2'$, and then proceed similarly to use regression models for $Q_1^{M'}$ as we do for Q_1^M. The bottom part of Table V uses another different set of stage 2 reference treatments to estimate the final mean rewards for regimes (CVD, TEE), (KA/VE, TEE), (TEC, KA/VE), and (TEE, KA/VE). Throughout Table V, the stage 2 reference treatments have $\Delta_2' = 0$ (or $\Delta_2 = 0$). If the label for a stage 2 reference treatment is j (hence $\Delta_{2j}' = 0$), then a different stage 2 treatment k has $\Delta_{2k}' = \Delta_{2k} - \Delta_{2j}$. Both

Table V. Estimates for Δ_2 and Δ_2' with different stage 2 treatments as references.

Target regimes	Stage 1 treatment	Reference Stage 2 treatment	Actual stage 2 treatment			
			CVD	KA/VE	TEC	TEE
			Δ_2			
(CVD, TEC)	CVD	TEC	NA	0.03	0	0.24
(KA/VE, TEC)	KA/VE	TEC	0.5	NA	0	0.55
(TEC, CVD)	TEC	CVD	0	0.175	NA	0.25
(TEE, TEC)	TEE	TEC	0.125	0.125	0	NA
			Δ_2'			
(CVD, KA/VE)	CVD	KA/VE	NA	0	0–0.03	0.24–0.03
(KA/VE, CVD)	KA/VE	CVD	0	NA	0–0.5	0.55–0.5
(TEC, TEE)	TEC	TEE	0–0.25	0.175–0.25	NA	0
(TEE, CVD)	TEE	CVD	0	0.125–0.125	0-0.125	NA
			Δ_2'			
(CVD, TEE)	CVD	TEE	NA	0.03–0.24	0–0.24	0
(KA/VE, TEE)	KA/VE	TEE	0.5–0.55	NA	0–0.55	0
(TEC, KA/VE)	TEC	KA/VE	0–0.175	0	NA	0.25–0.175
(TEE, KA/VE)	TEE	KA/VE	0.125–0.125	0	0–0.125	NA

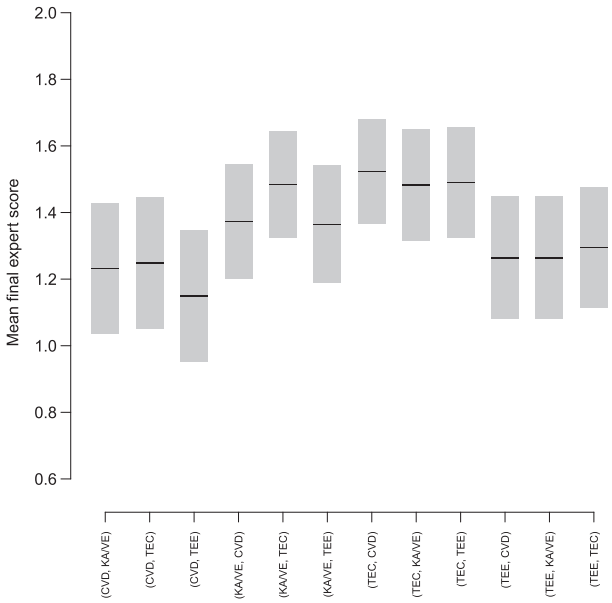

Figure 2. Means and 90% confidence intervals of the final outcome (expert score) for the 12 treatment strategies in the form of (A, B), which means to start with treatment A; if success, stay with A, otherwise switch to treatment B, with A and B each takes one of the four values CVD, KA/VE, TEC, and TEE, and A ≠ B.

Tables III and V are intended to be used for patients who had an unfavorable stage 1 response; consequently, their diagonal elements are not given because of the trial design that only those patients who achieved a successful stage 1 response could receive a stage 2 treatment same as stage 1. The results for 12 possible regimes are shown in Figure 2.

For this particular example, standard Q-learning gives very similar results (not shown). An advantage of both standard and modified Q-learning is that they can identify optimal dynamic treatment regimes for each individual. This can be performed by including interactions between individual level covariates and treatments. For example, if in the aforementioned analysis we include an interaction between patient

age and stage 1 treatment, then we can identify the age-specific optimal treatment regimes. If we include an interaction between stage 1 score and stage 2 treatment in the model in Section 4.1, then such an identified optimal stage 2 treatment will depend on stage 1 score. These are all desirable explorations to maximize benefit for each patient. However, because of the limited sample size, this may not yield stable results and thus is not presented here.

Recall that for patients who went off treatment during stage 1 because of toxicity or progressive disease, and thus did not receive any stage 2 treatment, we set $R_2 = 0$. By doing this, all 150 patients were included in the aforementioned analyses. This practical modification of the original treatment plan is consistent with the idea of 'viable treatment regimes' of Wang *et al*. [17]. For example, a patient received TEC as in stage 1, then went off treatment because of toxicity or progressive disease or other reasons and did not receive any stage 2 treatment, the data from this patient are used in the estimation of final reward for three regimes, namely, (TEC, CVD), (TEC, KA/VE), and (TEC, TEE).

5. Discussion

We have demonstrated a robust modification of Q-learning for optimizing a multi-stage treatment sequence in settings where the payoff is a cumulative outcome, and intermediate values at each stage are available. The modified Q-learning preserves more randomness in the observed outcomes, and thus is more robust against model misspecification, has higher power to identify optimal treatments, and satisfies the consistency assumption. If the treating physician happens to adopt the treatment regime that is optimal for a given patient's condition, the optimal outcome assumed by the modified Q-learning is precisely the observed outcome.

Optimization of a K-stage treatment regime is difficult, because conditioning on the treatment history can result in very complicated models. This is a common problem with all optimization algorithms for multi-stage treatments [5, 6]. We handle this problem by making a Markov assumption. This kind of assumption also was used in others' simulation studies [6]. In reality, this assumption may be violated. The degree of robustness of model results against this assumption is unknown. In such a case, if sample size permits, it is best to explore models without this Markov assumption, that is, include a large number of interaction terms to involve earlier stage history into the reward models. In cancer research, practical values for K are about 2 to 5, corresponding to disease recurrences. In other areas of application where the value of K may be much larger, the advantages of the modified Q-learning, that is, satisfying the consistency assumption and being robust against model misspecifications, may become more prominent.

An attractive feature of both standard and the modified Q-learning is that they do not need model treatment selection probabilities. Most other methods require this additional structure, including the history-adjusted marginal structural models [35] and A-learning [6]. There are very subtle arguments required with the use of modeling treatment selections. It has been argued that small misspecifications in such selection models can accumulate over treatment stages and thus cause severe bias and convergence problems [36]. Therefore, there is an advantage to avoid using such treatment selection models.

Appendix A: regression model for the modified Q-function

For each $s = 1, \cdots, K$, let β_s be a parameter vector of the main effects of $\bar{H}^M_{i,s}$ on $Q^M_{i,s}$. Using $j = 1$ as the reference treatment group, for each $j = 2, \cdots, J_s$, where J_s is the number of treatment options at stage s, let $\psi^{(j)}_s$ be a parameter vector of the interactive effects of action $A_{i,s} = j$ and $\bar{H}^M_{i,s}$ on $Q^M_{i,s}$. Let $\{e_i, i = 1, \cdots, n\}$ be a vector of i.i.d. random errors. Denoting the indicator of an event E by $I(E)$, the regression model for $Q^M_{i,s}$ is

$$Q^M_{i,s} = \beta_s^T \bar{H}^M_{i,s} + \sum_{j=2}^{J_s} I(A_{i,s} = j)\, \psi_s^{(j)T} \bar{H}^M_{i,s} \; + \; e_i, \tag{15}$$

where the main effects are

$$\beta_s^T \bar{H}^M_{i,s} = \beta_{s,0} + \beta_{s,1} Z_{i,s-1} + \beta_{s,2} Z_{i,s} + \beta_{s,3} Y_{i,s-1} + \sum_{l=2}^{J_{s-1}} \beta_{s,4,l} I(A_{i,s-1} = l)$$

and the multiplier of $I(A_{i,s}=j)$ in the sum of the interaction terms is

$$\psi_s^{(j)T}\bar{H}_{i,s}^M=\psi_{s,0}+\psi_{s,1}^{(j)}Z_{i,s-1}+\psi_{s,2}^{(j)}Z_{i,s}+\psi_{s,3}^{(j)}Y_{i,s-1}+\sum_{l=2}^{J_{s-1}}\psi_{s,4,l}^{(j)}I(A_{i,s-1}=l).$$

Thus, each parameter indexed by j is the comparative effect of $A_{i,s}=j$ versus action $A_{i,s}=1$.

Under the model (15), we define the cumulative causal effect of treatment j versus l at stage s, where $j>l$, as

$$\begin{aligned}D_{i,s}(j,l)&=Q_{i,s}^M\left(A_{i,s}=j,\bar{H}_{i,s}^M\right)-Q_{i,s}^M\left(A_{i,s}=l,\bar{H}_{i,s}^M\right)\\&=\begin{cases}\left\{\psi_s^{(j)}-\psi_s^{(l)}\right\}^T\bar{H}_{i,s}^M, & \text{if } l\neq 1\\ \psi_s^{(j)T}\bar{H}_{i,s}^M, & \text{if } l=1,\end{cases}\end{aligned}\tag{16}$$

which depends on the interaction parameters $\psi_s^{(j)}$, $\psi_s^{(l)}$, and recent history $\bar{H}_{i,s}^M$, but not on the main effects β_s.

Given estimates $\hat{\beta}_s$ and $\hat{\psi}_s^{(j)}, j=2,\cdots,J_s$ of the parameters in (15), denote the resulting estimates of $Q_{i,s}^M$ by $\hat{Q}_{i,s}^M$, estimated causal effects by $\hat{D}_{i,s}(j,l)$, and estimated cumulative future rewards by $\hat{R}_{i,s}$.

Appendix B: asymptotic properties

The linear models in (15) are easy to fit. However, due to the use of later stage estimates in models for earlier stages, the covariance formulas for the estimated regression parameters are not straightforward. In the succeeding discussion, we provide closed-form sandwich formula estimators for the covariance matrices. For simplicity, we assume the total number of treatment stages $K=2$ in the following derivation, denote by S_k the matrix formed by $\bar{H}_{i,k}^M, i=1,\cdots,n$, and assume treatment A_k is binary, for $k=1,2$. The results can be generalized to $K>2$.

Rewrite the right-hand side of the regression model in (15) as

$$\mathcal{Q}_k\left(S_k,A_k;\beta_k,\psi_k\right)=\beta_k^TS_{k1}+\left(\psi_k^TS_{k2}\right)A_k,\tag{17}$$

where $S_{k1}\in\mathbb{R}^{p_k}$ and $S_{k2}\in\mathbb{R}^{q_k}$ are sub-vectors of S_k. We allow variable selection so that the model in (15) may not include the full set of variables in $\bar{H}_{i,k}^M$. Denote $\theta_k=\left(\beta_k^T,\psi_k^T\right)^T$, and θ_{k0} its true value, where $\beta_k\in\mathbb{R}^{p_k}$ is the main effect of the current state variables on the outcome, and $\psi_k\in\mathbb{R}^{q_k}$ are the interactions between current state variables and treatment. Note that $Q_{i,2}^M=Y_2$, and $Q_{i,1}^M$ is the potential cumulative outcome given $S_{1,i}$, $A_{1,i}$ and $A_{2,i}^{opt}$. The two-stage backward induction proceeds as follows. Starting with the second stage, we have

$$\hat{\theta}_2=\left(\hat{\beta}_2^T,\hat{\psi}_2^T\right)^T=\arg\min_{\beta_2,\psi_2}\mathbb{P}_n\left\{Q_2^M-\mathcal{Q}_2\left(S_2,A_2;\beta_2,\psi_2\right)\right\}^2=\left[X_2^TX_2\right]^{-1}X_2^T\vec{Y}_2,$$

where $X_2=(S_{21}^T,A_2S_{22}^T)^T$ is the stage 2 design matrix and $\vec{Y}_2=(Y_{21},...,Y_{2n})^T$. Then estimate the second stage individual optimal outcome by $\hat{\mathbf{R}}_2=\left(\hat{R}_{21},\cdots,\hat{R}_{2n}\right)^T$, where

$$\hat{R}_{2i}=Y_{2i}+|\hat{\psi}_2^TS_{22i}|I\left\{A_{2i}\neq I\left(\hat{\psi}_2^TS_{22i}>0\right)\right\}.\tag{18}$$

With this optimized outcome at stage 2, the potential cumulative outcome, given $S_{1,i}$, $A_{1,i}$, and $A_{2,i}^{opt}$, is $\vec{Q}_1^M=(Q_{1,1}^M,\cdots,Q_{n,1}^M)^T$, where $Q_{i,1}^M(\hat{\psi}_2)=Y_{1i}+\hat{R}_{2i}$. After this, we estimate the first stage parameters by

$$\hat{\theta}_1=\left(\hat{\beta}_1^T,\hat{\psi}_1^T\right)^T=\arg\min_{\beta_1,\psi_1}\mathbb{P}_n\left\{Q_1^M(\hat{\psi}_2)-\mathcal{Q}_1\left(S_1,A_1;\beta_1,\psi_1\right)\right\}^2=\left[X_1^TX_1\right]^{-1}X_1^T\vec{Q}_1^M,$$

where $X_1=\left(S_{11}^T,A_1S_{12}^T\right)^T$.

The asymptotic properties of these parameter estimates are presented in the succeeding discussions, under the following technical conditions.

(A1) The true value for θ_2, denoted by $\theta_{20} = \left(\beta_{20}^T, \psi_{20}^T\right)^T$, minimizes

$$\mathcal{P}_0 \left\{Q_2^M - \mathcal{Q}_2(S_2, A_2; \theta_2)\right\}^2,$$

and the true value for θ_1, denoted by $\theta_{10} = \left(\beta_{10}^T, \psi_{10}^T\right)^T$, minimizes

$$\mathcal{P}_0 \left\{Q_1^M(\hat{\psi}_2) - \mathcal{Q}_1(S_1, A_1; \theta_1)\right\}^2,$$

where $\mathcal{P}_0 = \lim_n \mathbb{P}_n$ denotes the true probability measure. We assume that the limit exists and is finite in the aforementioned expressions.

(A2) $\theta_0 = \left(\theta_{10}^T, \theta_{20}^T\right)^T$ is an interior point in a bounded, open, convex subset $\Theta \subset \mathcal{R}^m$, where $m = \sum_k (p_k + q_k)$. For $k = 1, 2$, with probability one, $\mathcal{Q}_k(S_k, A_k; \theta_k)$ is at least twice continuously differentiable with respect to θ_k, and the Hessian matrix, $\mathcal{I}_{k0} = \mathcal{P}_0 \left[\nabla^2_{\theta_k \theta_k} \left\{Q_k^M - \mathcal{Q}_k\left(S_k, A_k; \theta_k\right)\right\}^2\right]$ exists and is positive-definite.

(A3) With probability one, $\Pr\left(\psi_k^T S_{k2} = 0\right) = 0$ for $k = 1, 2$.

Condition A1 says that θ_{10} and θ_{20} are true values that minimize loss function in each step. If Q_k takes the form of the linear model (17), condition A2 is equivalent to non-singularity of the design matrix $X_k = \left(S_{k1}^T, A_k S_{k2}^T\right)^T$ for $k = 1, 2$. From condition A3, we assume there is no possibility of non-regularity. In case of $\Pr(\psi_k^T S_{k2} = 0) > 0$, it has been verified that multi-stage estimation, including standard Q-learning, may be biased and the aforementioned asymptotic properties may be inappropriate, and thus requiring special treatment [13, 37]. Here, we do not consider such complications.

Denote the estimating equation for θ_2 as $\mathbb{P}_n \Psi_2\left(\theta_2; S_2, A_2\right) = 0$, where

$$\Psi_2(\theta_2; S_2, A_2) = \{\mathcal{Q}_2(S_2, A_2; \theta_2)/(\partial\theta_2)\}^T \{Y_2 - \mathcal{Q}_2(S_2, A_2; \theta_2)\}.$$

Theorem 1
Under conditions A1–A3,

$$\sqrt{n}\left(\hat{\theta}_2 - \theta_{20}\right) \sim N\left\{0, V_2(\theta_{20})\right\}, \quad \text{as } n \to \infty,$$

where $V_2(\theta_2) = D_2(\theta_2)^{-1} B_2(\theta_2) \{D_2(\theta_2)^{-1}\}^T$ with $D_2(\theta_2) = -E[\partial\Psi_2(\theta_2; S_2, A_2)/(\partial\theta_2)]$ and $B_2(\theta_2) = E\left[\Psi_2(\theta_2; S_2, A_2)\Psi_2(\theta_2; S_2, A_2)^T\right]$.

Proof
It is a direct application of the 'Sandwich' formula [33], so omitted. □

Because the estimation of θ_1 depends on $\hat{\psi}_2$, let $\Psi_{2,2}$ denote the sub-equation of Ψ_2, and $D_{2,2}$ denote the sub-matrix of D_2, both corresponding to ψ_2 at $\beta_2 = \beta_{20}$. Then, $\hat{\psi}_2 - \psi_{20} \approx D_{2,2}(\psi_{20})^{-1} \mathbb{P}_n[\Psi_{2,2}(\psi_{20}; S_2, A_2)]$, where β_{20} and ψ_{20} are true values of β_2 and ψ_2.

$$\sqrt{n}\left(\hat{\psi}_2 - \psi_{20}\right)^T S_{22} \sim N\left(0, \Sigma_2\right), \ \Sigma_2 = D_{2,2}^{-1} E\left\{\Psi_{2,2}\left(S_{22}^T S_{22}\right)\Psi_{2,2}^T\right\} D_{2,2}^{-1}.$$

The estimating equation for θ_1 is

$$\begin{aligned}
&\mathbb{P}_n \Psi_1(\theta_1; S_1, A_1, \hat{\psi}_2) \\
&= \mathbb{P}_n \left[\left\{\partial\mathcal{Q}_1(S_1, A_1; \theta_1)/(\partial\theta_1)\right\}^T \left\{\vec{Q}_1^M - \mathcal{Q}_1(S_1, A_1; \theta_1)\right\}\right] \\
&= \mathbb{P}_n \Big[\left\{\partial\mathcal{Q}_1(S_1, A_1; \theta_1)/(\partial\theta_1)\right\}^T \\
&\quad \times \left\{Y_1 + Y_2 + |\hat{\psi}_2^T S_{22}| I\left(A_2 \neq I\left(\hat{\psi}_2^T S_{22} > 0\right)\right) - \mathcal{Q}_1(S_1, A_1; \theta_1)\right\}\Big] \\
&= \mathbb{P}_n \Big[\Psi_1(\theta_1; S_1, A_1, \psi_{20}) + \left\{\partial\mathcal{Q}_1(S_1, A_1; \theta_1)/(\partial\theta_1)\right\}^T \\
&\quad \times \left\{|\hat{\psi}_2^T S_{22}| I\left(A_2 \neq I\left(\hat{\psi}_2^T S_{22} > 0\right)\right) - |\psi_{20}^T S_{22}| I\left(A_2 \neq I\left(\psi_{20}^T S_{22} > 0\right)\right)\right\}\Big] \\
&= \mathbb{P}_n \Big[\Psi_1(\theta_1; S_1, A_1, \psi_{20}) + \left\{\partial\mathcal{Q}_1(S_1, A_1; \theta_1)/(\partial\theta_1)\right\}^T \\
&\quad \times \left[I(A_2 = 0)\left\{|\hat{\psi}_2^T S_{22}|_+ - |\psi_{20}^T S_{22}|_+\right\} + I(A_2 = 1)\left\{|\hat{\psi}_2^T S_{22}|_- - |\psi_{20}^T S_{22}|_-\right\}\right].
\end{aligned}$$

Let

$$\bar{S}_2 \equiv \left[S_{21}^T, S_{22}^T \left\{I(A_2 = 0)I\left(\psi_{20}^T S_{22} \geqslant 0\right) + I(A_2 = 1)I\left(\psi_{20}^T S_{22} \leqslant 0\right)\right\}\right]^T.$$

Theorem 2
Under conditions A1–A3,

$$\sqrt{n}\left(\hat{\theta}_1 - \theta_{10}\right) \sim N\left\{0, V_1(\theta_{10}, \theta_{20})\right\}, \quad \text{as } n \to \infty,$$

where $V_1(\theta_{10}, \theta_{20})$ can be estimated by

$$\hat{V}_1(\theta_{10}, \theta_{20}) = D_1(\theta_{10})^{-1}\left[\mathbb{P}_n\left\{\Psi_1\Psi_1^T\right\} + \mathbb{P}_n\left\{X_1\bar{S}_2^T\hat{V}_2(\theta_{20})\bar{S}_2X_1^T\right\}\right]D_1(\theta_{10})^{-1},$$

with $D_1(\theta_1) = -E[\partial\Psi_1(\theta_1; S_1, A_1)/(\partial\theta_1)]$.

Proof
Again, it is a direct application of the 'Sandwich' formula [33], so omitted. □

Appendix C: simulation model

In general, suppose we have random variables $X_1, \cdots, X_p$, and Y, and we would like to do regression of Y on $X_1, \cdots, X_p$ with model $Y = \sum_{i=1}^p \beta_i X_i + \varepsilon$. If $X_1, \cdots, X_p$ are orthogonal to each other, then we have $\beta_i = \frac{E(X_iY)}{E(X_i^2)}$.

In Section 3, we have $Z_1 \sim \text{Normal}(0, 1)$, $A_1 \sim \text{Bernoulli}(0.5)$, $V \sim \text{Normal}(0, 2^2)$, $Y_1 = Z_1(A_1 - 0.5) + V + \varepsilon_1$ with $\varepsilon_1 \sim \text{Normal}(0, 1)$. $A_2 \sim \text{Bernoulli}(0.5)$, $Y_2 = -2Z_1(A_1 - 0.5) + (A_1 - 0.5)(A_2 - 0.5) - V + \varepsilon_2$ with $\varepsilon_2 \sim \text{Normal}(0, 1)$. We use the following model to do regression analysis.

$$\begin{aligned} Q_2^M = Y_2 = {} & \beta_{20} + \beta_{21}Z_1 + \beta_{22}A_1 + \beta_{23}Y_1 \\ & + A_2(\psi_{20} + \psi_{21}Z_1 + \psi_{22}A_1 + \psi_{23}Y_1) + e_2. \end{aligned}$$

To find out the true values for βs in the aforementioned model, we consider regressing Y_2 on the following orthogonal set of random variables $\{1, Z_1, (A_1 - 0.5), Y_1, (A_2 - 0.5), (A_2 - 0.5)Z_1, (A_2 - 0.5)(A_1 - 0.5), (A_2 - 0.5)Y_1\}$. That is to say, we consider the model

$$\begin{aligned} Y_2 = {} & \beta'_{20} + \beta'_{21}Z_1 + \beta'_{22}(A_1 - 0.5) + \beta'_{23}Y_1 \\ & + (A_2 - 0.5)(\psi'_{20} + \psi'_{21}Z_1 + \psi'_{22}(A_1 - 0.5) + \psi'_{23}Y_1) + e_2. \end{aligned}$$

By using the formula $\beta_i = \frac{E(X_iY)}{E(X_i^2)}$ mentioned at the beginning of this Appendix, we obtain

$$\begin{aligned} Y_2 = {} & 0 + 0Z_1 + 0(A_1 - 0.5) - 0.857Y_1 \\ & + (A_2 - 0.5)(0 + 0Z_1 + (A_1 - 0.5) + 0Y_1) + e_2. \end{aligned}$$

For the aforementioned coefficients, the only one that is not straightforward is

$$\beta'_{23} = E(Y_1Y_2)/E(Y_1^2) = \frac{-E(V^2) - 2E[Z_1^2(A_1 - 0.5)^2]}{E(V^2) + E[Z_1^2(A_1 - 0.5)^2] + E(\varepsilon_1^2)} = \frac{-4.5}{5.25} \approx -0.857.$$

Note $e_2 = \varepsilon_2 - 2Z_1(A_1 - 0.5) - V + 0.857Y_1$. It can be verified that e_2 is orthogonal to all explanatory variables on the right-hand side of the above equation, justifying its validity as a residual term.

Appendix D: an illustration for > 2 stages and > 2 treatment options in each stage

Section 2 and Appendix A provide a general description for any number of stages (K), and any number of treatment options in each stage (J_s), for $s = 1, \cdots, K$. As suggested by a referee, for easy understanding, we illustrate through an example in the succeeding discussions. We skip the data generation and pick a particular regression model for each stage. Suppose we observe (Z_s, A_s, Y_s) for each stage $s = 1, 2, 3$, with $A_s = 1, 2$, or 3, indicating three different treatment options for each stage s. Here, $A_s = 1$ and

$A_{s'} = 1$ do not necessarily mean the same medical treatment for stages $s \neq s'$. The modified Q-learning proceeds as follows.

Let $Q_3^M = Y_3$. Fit a linear regression model as below to find out the optimal treatment for stage 3 conditional on current covariates Z_3, previous treatment A_2 and outcome Y_2.

$$Q_3^M = \sum_{j=1}^{3} I(A_3 = j) \left\{ \psi_{30,j} + \psi_{31,j} Z_3 + \sum_{l=1}^{3} \psi_{32,jl} I(A_2 = l) + \psi_{33,j} Y_2 \right\} + e_3.$$

Then the conditional optimal treatment is just the one that maximize the mean reward in stage 3, which is mathematically described as

$$\hat{A}_3^{opt} = \text{argmax}_{j=1,2,3} \left\{ \hat{\psi}_{30,j} + \hat{\psi}_{31,j} Z_3 + \sum_{l=1}^{3} \hat{\psi}_{32,jl} I(A_2 = l) + \hat{\psi}_{33,j} Y_2 \right\}. \tag{19}$$

Then we estimate the potential optimal reward R_3 in stage 3 each individual would have achieved had he/she received his/her conditional optimal treatment as indicated previously. If an individual actually received his/her conditional optimal treatment, the estimated reward $\hat{R}_3$ is set to be the observed Y_3 by our modified Q-learning. Otherwise, $\hat{R}_3$ is set to be Y_3 plus the difference between the rewards for the optimal and actual treatments.

$$\begin{aligned} \hat{R}_3 = Y_3 + I\left(A_3 \neq \hat{A}_3^{opt}\right) \Bigg[& \max_{j=1,2,3} \left\{ \hat{\psi}_{30,j} + \hat{\psi}_{31,j} Z_3 + \sum_{l=1}^{3} \hat{\psi}_{32,jl} I(A_2 = l) + \hat{\psi}_{33,j} Y_2 \right\} \\ & - \sum_{k=1}^{3} I(A_3 = k) \left\{ \hat{\psi}_{30,k} + \hat{\psi}_{31,k} Z_3 + \sum_{l=1}^{3} \hat{\psi}_{32,kl} I(A_2 = l) + \hat{\psi}_{33,k} Y_2 \right\} \Bigg]. \end{aligned}$$

By solving the above equation, the optimal stage3 treatments conditional on historical information are identified. Here, the Markov assumption is used in the above linear regression model so that it depends on A_2 and Y_2, but does not go further to depend on A_1 or Y_1. Without such an assumption, the aforementioned linear regression would have too many predictors and require an extremely large sample size to have a reasonable fit. This Markov assumption is for practical rather than theoretical consideration.

Now consider optimizing treatments for stage 2. Define $Q_2^M = Y_2 + \hat{R}_3$. Similarly, as previously discussed, first fit a linear model as follows:

$$Q_2^M = \sum_{j=1}^{3} I(A_2 = j) \left\{ \psi_{20,j} + \psi_{21,j} Z_2 + \sum_{l=1}^{3} \psi_{22,jl} I(A_1 = l) + \psi_{23,j} Y_1 \right\} + e_2.$$

Then find out the optimal stage 2 treatment conditional on current covariates Z_2, previous treatment A_1 and outcome Y_1, as below.

$$\hat{A}_2^{opt} = \text{argmax}_{j=1,2,3} \left\{ \hat{\psi}_{20,j} + \hat{\psi}_{21,j} Z_2 + \sum_{l=1}^{3} \hat{\psi}_{22,jl} I(A_1 = l) + \hat{\psi}_{23,j} Y_1 \right\}. \tag{20}$$

The potential total reward from stage 2 onwards (i.e., sum of rewards from stages 2 and 3), R_2, had a subject received his optimal stage 2 treatment and his corresponding optimal stage 3 treatment, can be estimated as follows:

$$\begin{aligned} \hat{R}_2 = Y_2 + I(A_2 \neq \hat{A}_2^{opt}) \Bigg[& \max_{j=1,2,3} \left\{ \hat{\psi}_{20,j} + \hat{\psi}_{21,j} Z_2 + \sum_{l=1}^{3} \hat{\psi}_{22,jl} I(A_1 = l) + \hat{\psi}_{23,j} Y_1 \right\} \\ & - \sum_{k=1}^{3} I(A_2 = k) \left\{ \hat{\psi}_{20,k} + \hat{\psi}_{21,k} Z_2 + \sum_{l=1}^{3} \hat{\psi}_{22,kl} I(A_1 = l) + \hat{\psi}_{23,k} Y_1 \right\} \Bigg]. \end{aligned}$$

For stage 1, similarly as previously discussed, define $Q_1^M = Y_1 + \hat{R}_2$. Then fit a linear regression model

$$Q_1^M = \sum_{j=1}^{3} I(A_1 = j)(\psi_{10,j} + \psi_{11,j} Z_1) + e_1.$$

This will give estimates for the optimal treatment stage 1 treatments conditional on Z_1 as follows:

$$\hat{A}_1^{opt} = \text{argmax}_{j=1,2,3}\{\hat{\psi}_{10,j} + \hat{\psi}_{11,j} Z_1\}. \tag{21}$$

Under this optimal stage 1 treatment and corresponding optimal stages 2 and 3 treatments, the total optimal reward is R_1, which can be estimated by

$$\hat{R}_1 = Y_1 + I(A_1 \neq \hat{A}_1^{opt}) \left[\max_{j=1,2,3} \left\{\hat{\psi}_{10,j} + \hat{\psi}_{11,j} Z_1\right\} - \sum_{k=1}^{3} I(A_1 = k) \left\{\hat{\psi}_{10,k} + \hat{\psi}_{11,k} Z_1\right\} \right].$$

The aforementioned procedures are to derive the optimal treatments using a backward induction. After the estimation results are obtained, to apply them in practice, the optimal treatment decision rules are determined as follows. First use (21) to find out the optimal treatment conditional on covariate Z_1. Suppose this gives $A_1 = 1$. After receiving this treatment $A_1 = 1$, the observed stage 1 outcome is Y_1. At the beginning of stage 2, covariate Z_2 is observed. Then at this moment, the optimal stage 2 treatment can be determined by (20) based on Z_2, $A_1 = 1$ and the observed Y_1. Suppose the optimal treatment conditional on these variables is $A_2 = 3$. After receives this treatment $A_2 = 3$, the observed stage 2 outcome is Y_2. At the beginning of stage 3, covariate Z_3 is observed. At this time, the optimal stage 3 treatment is determined by (19) based Z_3, $A_2 = 3$ and the observed Y_2. Suppose the observed stage 3 outcome is Y_3. The above optimal treatment identification method is supposed to maximize $Y = Y_1 + Y_2 + Y_3$.

Acknowledgements

The authors acknowledge the support from the USA National Institutes of Health grants U54 CA096300, U01 CA152958, 5P50 CA100632, R01 CA 83932, and 5P01 CA055164.

References

1. Robins J. A new approach to causal inference in mortality studies with a sustained exposure period: application to control of the healthy worker survivor effect. *Mathematical Modelling* 1986; **7**:9–12.
2. Robins J. The control of confounding by intermediate variables. *Statistics in Medicine* 1989; **8**:679–701.
3. Robins JM. Analytic methods for estimating hiv treatment and cofactor effects. In *Methodological Issues of Aids Mental Health Research*, Ostrow D, Kessler R (eds). Plenum Publishing: New York, 1993; 213–290.
4. Robins JM. Causal inference from complex longitudinal data. In *Latent Variable Modeling and Applications to Causality*, Berkane M (ed.) Springer-Verlag: New York, 1997; 69–117.
5. Robins J. Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium on Biostatistics*, Lin D, Heagerty P (eds). Springer: New York, 2004; 189–326.
6. Murphy SA. Optimal dynamic treatment regimes (with discussion). *Journal of the Royal Statistical Society, Series B* 2003; **65**:331–366.
7. Hernán M, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men. *Epidemiology* 2000; **11**:561–570.
8. Murphy SA, van der Laan MJ, Robins JM, Group CPPR. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association* 2001; **96**:1410–1423.
9. Lunceford JK, Davidian M, Tsiatis AA. Estimation of survival distributions of treatment policies in two-stage randomization designs in clinical trials. *Biometrics* 2002; **58**:48–57.
10. Bang H, Robins J. Doubly robust estimation in missing data and causal inference models. *Biometrics* 2005; **61**:962–973.
11. Petersen M, Sinisi S, van der Laan M. Estimation of direct causal effects. *Epidemiology* 2006; **17.3**:276–284.
12. Robins J, Orellana L, Rotnitzky A. Estimation and extrapolation of optimal treatment and testing strategies. *Statistics in Medicine* 2008; **27**:4678–4721.
13. Moodie EM, Richardson TS. Estimating optimal dyanamic regimes: correcting bias under the null. *Scandinavian Journal of Statistics* 2009; **37**:126–146.
14. Zhao Y, Zeng D, Socinski M, Kosorok M. Reinforcement learning strategies for clinical trials in non-small cell lung cancer. *Biometrics* 2011; **67**:1422–1433.
15. Goldberg Y, Kosorok MR. Q-learning with censored data. *Annals of Statistics* 2012; **40**:529–560.
16. Zhao Y, Zeng D, Rush A, Kosorok M. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association* 2012; **107**:1106–1118.

17. Wang L, Rotnitzky A, Lin X, Millikan RE, Thall PF. Evaluation of viable dynamic treatment regimes in a sequentially randomized trial of advanced prostate cancer. *Journal of American Statistical Association* 2012; **107**:493–508.
18. Wahed A, Thall P. Evaluating joint effects of induction-salvage treatment regimes on overall survival in acute leukemia. *Journal of Royal Statistical Society, Series C* 2013; **62**:67–83.
19. Thall PF, Millikan RE, Sung HG. Evaluating multiple treatment courses in clinical trials. *Statistics in Medicine* 2000; **19**: 1011–1028.
20. Lavori PW. A design for testing clinical strategies: biased adaptive within-subject randomization. *Journal of the Royal Statistical Society, Series A* 2000; **163**:29–38.
21. Lavori PW, Dawson R. Dynamic treatment regimes: practical design considerations. *Statistics in Medicine* 2001; **20**: 1487–1498.
22. Collins LM, Murphy SA, Strecher V. The multiphase optimization strategy (most) and the sequential multiple assignment randomized trial (smart): new methods for more potent e-health interventions. *American Journal of Preventive Medicine* 2007; **32**:S112–S118.
23. Watkins CJCH, Dayan P. Q-learning. *Machine Learning* 1992; **8**:279–292.
24. Murphy SA. A generalization error for q-learning. *Journal of Machine Learning Research* 2005; **6**:1073–1097.
25. Moodie EEM, Richardson TS, Stephens DA. Demystifying optimal dynamic treatment regimes. *Biometrics* 2007; **63**: 447–455.
26. Laber E, Qian M, Lizotte D, Murphy S. Statistical inference in dynamic treatment regimes. *Arxiv preprint* 2010; **arXiv:1006.5831**.
27. Song R, Wang W, Zeng D, Kosorok MR. Penalized Q-learning for dynamic treatment regimes. *arXiv preprint* 2011; **arXiv:1108.5338**.
28. Nahum-Shani I, Qian M, Almirall D, Pelham W, Gnagy B, Fabiano G, Waxmonsky J, Yu J, Murphy S. Q-learning: a data analysis method for developing adaptive interventions. *Psychological Methods* 2012; **17**:478–494.
29. Zhang B, Tsiatis AA, Laber EB, Davidian M. Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika* 2013; **100**:681–691.
30. Schulte PJ, Tsiatis AA, Laber EB, Davidian M. Q-and A-learning methods for estimating optimal dynamic treatments regimes. *Statistical Science* 2014; **29**:640–661.
31. Bellman R. *Dynamic Programming*. Princenton University Press: Princeton, N.J., 1957.
32. Robins J. Robust estimation in sequential ignorable missing data and causal inference models. *Proceedings of the Bayesian Statistical Science Section of the American Statistical Association*, Baltimore, Maryland, 2000, 6–10.
33. White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrika* 1980; **48**:817–838.
34. Henderson R, Ansell P, Alshibani D. Regret-regression for optimal dynamic treatment regimes. *Biometrics* 2010; **66**: 1192–1201.
35. Petersen ML, Deeks SG, Martin JN, van der Laan MJ. History-adjusted marginal structural models to estimate time-varying effect modification. *U.C. Berkeley Division of Biostatistics Working Paper Series* 2005; **Working Paper 199**.
36. Rosthøj S, Fullwood C, Henderson R, Stewart S. Estimation of optimal dynamic anticoagulation regimes from observational data: a regret-based approach. *Statistics in Medicine* 2006; **25**:4197–4215.
37. Chakraborty B, Murphy S, Strecher V. Inference for nonregular parameters in optimal dynamic treatment regimes. *Statistical Methods in Medical Research* 2010; **19**:317–343.