Q-Learning Based Methods for Dynamic Treatment Regimes

Xinyi Li, Nikki L. B. Freeman, and Lily Wang

Abstract Precision medicine seeks to find the optimal treatments tailored to the individual characteristics of each patient. Dynamic treatment regimes consists of a sequence of personalized treatment decisions that formalizes the process of decision-making that translated the patients' information into the recommended treatment. Q-learning is a popular approach to estimate the optimal treatment regime, which is closely related to the regression-based analysis in statistics, and reinforcement learning methods. In this book chapter, we provide a comprehensive introduction of Q-learning based methods for the estimation of dynamic treatment regimes. We start with the potential outcome framework that lay the ground for Q-learning, followed by the introduction of reinforcement learning and its application in precision medicine study. We then delve into Q-learning based methods in dynamic treatment regime with finite time horizon, including both single-decision setting and multi-stage decision setting, and infinite time horizon, respectively. To concretize the concepts discussed, we present a simple example of Q-learning implementation for the two-stage setting using the R statistical programming language.

Key words: precision medicine, reinforcement learning, causal inference, Markov decision process, data-driven decision science

Xinyi Li
Clemson University, School of Mathematical and Statistical Sciences, O-329 Martin Hall, Clemson, SC 29634, e-mail: lixinyi@clemson.edu

Nikki L. B. Freeman
University of North Carolina at Chapel Hill, Department of Biostatistics, 135 Dauer Drive, Chapel Hill, NC 27599, e-mail: nlbf@live.unc.edu

Lily Wang
George Mason University, Department of Statistics, 4400 University Drive, MS 4A7, Fairfax, VA 22030, e-mail: lwang41@gmu.edu

1 Introduction

The goal of precision medicine is to "match the right treatment to the right patient at the right time" [1, 7]. Unlike more traditional approaches used to generate evidence for health and healthcare decision-making, the precision medicine paradigm focuses on learning from data the best way to tailor treatment or a sequence of treatments to individuals based on their specific characteristics. This is often referred to as "leveraging heterogeneity". The idea is that by learning how to optimally tailor treatments to individuals, individual outcomes and thereby population outcomes can be improved over those from one-size-fits-all or non-data-driven treatment strategies.

In precision medicine, the formal structure for matching patients to treatments is the dynamic treatment regime (DTR), which is also known as adaptive treatment strategies [17, 23], adaptive interventions [8, 24], treatment policies [20, 42], or individualized treatment rules [28]. DTRs are sets of treatment rules, one for each key decision point, that map from patient features to an available treatment. The features may include disease and treatment history, clinical, prognostic, and/or genomic characteristics, behaviors, and response to past treatments. In the language of DTRs, the goal of precision medicine then is to learn optimal DTRs, decision rules that, if followed, would favorably optimize the expected value of a target outcome over the patient population of interest.

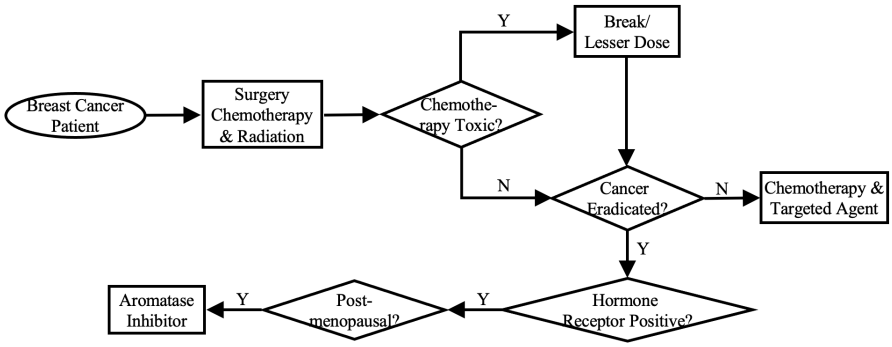

Fig. 1 Schematic depiction of DTR for breast cancer patients reproduced from [14]. A patient diagnosed with breast cancer may get the surgery, followed by chemotherapy and radiation at the first stage. The physician will modify this to take a break from chemotherapy and radiation or give a lesser dose if these treatment actions are too toxic to the patient. If the cancer is eradicated by the surgery or these treatment actions, and if the patient is hormone receptor-positive and post-menopausal, an aromatase inhibitor will be prescribed. If the cancer has not been eradicated yet or there is evidence of cancer, the physician may suggest additional chemotherapy and a targeted agent.

In practice, DTRs can be widely applied to the study of chronic diseases or disorders, and are particularly well-poised to support decision-making in disease settings that may require a sequence of medical interventions; examples include cancer [39, 47], attention deficit hyperactivity disorder (ADHD; [27]), human immunodefi-

ciency virus (HIV) infection/acquired immune deficiency syndrome (AIDS) [4], and diabetes [19]. The schematic in Figure 1, which is used to describe the possible medical interventions for the treatment of breast cancer in [14], can help us understand DTRs more concretely. As shown in Figure 1, a patient diagnosed with breast cancer may initially be treated with surgery followed by chemotherapy and radiation. Depending on the toxicity of chemotherapy for the individual patient, the provider will modify the treatment strategy: a break from or reduced dose of chemotherapy if chemotherapy is too toxic for the patient, augmented chemotherapy with a targeted agent if the chemotherapy is not too toxic, and the cancer is not yet eradicated, or an additional aromatase inhibitor if the patient's cancer is eradicated and they are hormone receptor-positive and post-menopausal. Patients whose cancer is eradicated after chemotherapy and are not hormone receptor-positive, or are hormone receptor-positive and not post-menopausal, are surveilled for recurrence but not given any further active therapies.

To illustrate how a DTR might support clinical decision-making, let us reconsider the case of a 60-year-old, hormone receptor-positive and post-menopausal woman with stage II breast cancer using the DTR depicted in Figure 1. For this woman, the treatment recommendations are as follows [14]:

> Following surgery, treat with chemotherapy for six cycles. If there is no evidence of cancer and the lymph nodes are negative, treat with an aromatase inhibitor for five years. If there is evidence of cancer following chemotherapy, continue chemotherapy for another six cycles. If the patient experiences a grade III or higher toxicity on the prescribed chemotherapy, switch to another chemotherapy.

We note that the DTR provides guidance on how to tailor treatments based on the woman's health and disease status (post-menopausal, lymph node involvement, cancer persistence or eradication, hormone receptor-positive) and response to previous treatment (degree of chemotherapy toxicity) – in other words, how to tailor treatment to this particular patient.

The example DTR outlined in Figure 1 can also give us an idea of the data and statistical elements needed to learn such a DTR. We observe that in the example, there are multiple decision points at which treatment decisions need to be made and possible treatments for each decision point. These include the decision whether to continue and/or augment chemotherapy and the decision whether to prescribe an aromatase inhibitor. The recommendations rendered by the DTR at each of these decision points, in turn, depend on individual patient features and responses to previous therapy. For example, the decision to continue and/or augment chemotherapy depends on whether the chemotherapy is too toxic for the patient and whether or not the patient's cancer has responded to the previous treatment of surgery in combination with chemotherapy and radiation. Similarly, the decision to prescribe an aromatase inhibitor for patients whose cancer has been eradicated depends on the patient's cancer characteristics, specifically, whether the cancer is hormone receptor-positive. While not explicitly illustrated in Figure 1, the goal of the DTR is to optimize patient outcomes such as progression-free survival or recurrence-free survival. Thus, at a minimum, the statistical framework for learning DTRs will include decision points, patient features, treatments, and outcomes. Additional features may be

included in model-specific settings or problem attributes, but in general, the main elements remain the same.

Data collected from randomized controlled trials (RCTs) and sequentially multiple adaptive randomized trials (SMARTs), as well as observational data, may be used to estimate DTRs. Generally, by design, data from RCTs are suitable for learning DTRs with a single decision point, and data from SMARTs are suitable for learning DTRs with multiple decision points. For each, the decision points correspond to the randomization points, and the possible treatments correspond to the treatments being randomized at the randomization points. Point exposure or multiple exposure observational data, too, may be used, although additional assumptions are needed to identify key statistical estimands related to DTRs when using observational data. For additional discussion on trial and study designs that generate data appropriate for DTR estimation, see [16].

Various statistical analysis methods have been developed in support of data-driven decision-making [15, 16, 40], including outcome weighted learning (OWL) [45, 46] and related methods, inverse probability weighted (IPW), and doubly robust augmented inverse probability weighted (AIPWE) estimators [43, 44], among others (see [15, 16, 40]). Among the methods for learning optimal DTRs, Q-learning is one of the widely used approaches and is our focus in this chapter. Q-learning, where "Q" stands for "quality," is a strategy for learning optimal DTRs by positing models for and estimating Q-functions. As we will see, Q-functions are conditional expectations, which in turn are closely related to the widely adopted regression-based methods in statistics.

We proceed as follows: In Section 2, we lay out the precision medicine framework, providing the notation and language of precision medicine used throughout this chapter. This framework sets the statistical goal of learning optimal DTRs. Section 3 formalizes Q-learning in precision medicine and contrasts it with reinforcement learning, providing a conceptual overview and highlighting the distinctions. Sections 4 and 5 introduce Q-Learning for optimal DTRs in the finite and infinite time horizon settings, respectively. In Section 6, we provide a simple example with code to illustrate how to implement Q-learning for the finite horizon setting in R.

2 Introduction to the Precision Medicine Framework

The idea of "matching the right treatment to the right person at the right time" can be formalized concisely as the goal of learning an optimal DTR. Put otherwise, the goal is to learn a sequence of functions, one for each key decision point, that maps from observed patient features to a recommended treatment. An optimal DTR is a DTR that, if followed, would optimize outcomes on average in the target population. In this section, we develop this idea formally.

2.1 Notation and potential outcomes

We begin by discussing causal inference and the potential outcomes or counterfactual framework. The concept of potential outcomes was first proposed by [25] for causal effects in the context of completely randomized experiments, and was then extended to a general framework for causation in [34] in both observational and randomized studies. In the standard causal inference setting, or as one might describe as the single timepoint decision setting, a potential outcome is defined as the outcome a subject would have had under a particular treatment. For each subject, there is a potential outcome associated with each possible treatment. At most, only one potential outcome is realized for each individual, and the not-realized potential outcomes are thus counter to fact. Formally, we let $\mathscr{A}$ denote the set of treatment options and $a \in \mathscr{A}$ denote a particular treatment option. Then, we denote $Y^*(a)$ as the potential outcome that would be observed if a subject were assigned treatment $a \in \mathscr{A}$, and $\{Y^*(a)\}_{a\in\mathscr{A}}$ is the set of all potential outcomes. A more detailed overview of the potential outcomes framework and causal inference can be found in [12] and [26].

In the precision medicine setting, however, we are often interested in clinical decision-making over time rather than a single timepoint and thus need a more general notion of potential outcomes that can accommodate potential outcomes under sequences of treatments. We let $t = 1,\ldots,T$ denote the sequence of timepoints at which a treatment decision is made and $T < \infty$ the total number of decision points. Later we will consider the case with infinitely many decision points (Section 5). Potential outcomes in the multiple decision point (multiple stage) setting are slightly more complicated than those in the single decision point (single stage) setting. Most obviously, instead of considering a single treatment assigned at a single decision point, we will need to consider entire sequences of treatments. To denote a sequence of treatments, we will use an overline, for example, $\overline{\boldsymbol{a}}_t = (a_1,\ldots,a_t)$. Additionally, we define $\overline{\boldsymbol{a}} \equiv \overline{\boldsymbol{a}}_T = (a_1,\ldots,a_T)$. Because the treatments that are available at each decision point may differ, the set of possible treatments available at time t is denoted by $\mathscr{A}_t$ and we denote the set of possible treatments that could be realized by the sequence $\overline{\boldsymbol{a}}_t$ as $\mathscr{A}_1 \times \cdots \times \mathscr{A}_t$. Letting Y denote the outcome of interest measured after the T-th treatment decision, we let $Y^*(\overline{\boldsymbol{a}})$ denote the potential outcome we would observe if an individual received a particular sequence of treatments $\overline{\boldsymbol{a}} \in \mathscr{A}_1 \times \cdots \times \mathscr{A}_T$.

In addition to treatments and final outcomes, precision medicine often involves incorporating patient features that evolve over time as well as proximal outcomes that are collected in close proximity to when a treatment is administered. An example of proximal outcomes can be found in mobile health (mHealth) studies [18]. In mHealth interventions, mobile devices can deliver treatments/interventions to individuals during daily lives, and those outcomes collected in the near time the treatment is taken are considered proximal outcomes, such as stress or physical activity like step counts. Because the treatments applied at time t affect both proximal outcomes and interim measurements, when extending the potential outcomes framework to multiple treatments over time, we will need to consider both potential proximal outcomes and potential interim measurements. We let $A_t \in \mathscr{A}_t$ denote the assigned treatment, Y_t denote a proximal outcome measured after the treatment

at stage t, and let $Y_t^*(\overline{\boldsymbol{a}}_t)$ be the potential proximal outcome we would observe if $\overline{\boldsymbol{a}}_t \in \mathscr{A}_1 \times \cdots \times \mathscr{A}_t$ had been assigned for $t = 1, \ldots, T-1$. Further, we let $\mathbf{X}_1 \in \mathscr{X}_1$ denote baseline patient information at time $t = 1$, and for $t = 2, \ldots, T$, $\mathbf{X}_t \in \mathscr{X}_t$ stand for interim patient information collected after the $(t-1)$-th decision and before the t-th decision is to be made. Baseline patient information, also referred to as baseline covariates or baseline features, might include co-morbid conditions, biomarkers, disease status, and other patient characteristics observed at the time of the first decision. Interim patient information might include patient features, or covariates, that evolve over time and possibly in response to prior treatments. We let $\mathbf{X}_t^*(\overline{\boldsymbol{a}}_{t-1})$ denote the potential interim measurements of patient information at time t we would observe if treatment sequence $\overline{\boldsymbol{a}}_{t-1} \in \mathscr{A}_1 \times \cdots \times \mathscr{A}_{t-1}$ had been assigned.

Thus far, we have considered potential outcomes for sequences of treatments rather than treatment rules which are the central target of precision medicine. To extend from sequences of treatments to treatment rules, the history notation is convenient. Define $\mathbf{H}_t$ as the history data at time t, so $\mathbf{H}_1 = \mathbf{X}_1$ and $\mathbf{H}_t = (\mathbf{H}_{t-1}, A_{t-1}, Y_{t-1}, \mathbf{X}_t)$; that is, from t to $t+1$, we update the history with the treatment and response collected at t, and the patient covariates collected between t and $t+1$. An important technical note is that not all treatments in $\mathscr{A}_t$, $t = 1, \ldots, T$, may be feasible for all patients. This may be because of drug interactions, past history of drug intolerability, a previous therapy that precludes a future therapy (e.g., amputation of a limb precludes future amputation of the same limb), or other contraindications to a particular treatment or course of therapy. Using the history notation we formalize feasible treatments; letting $\mathscr{H}_t$ be the space of all possible histories, we denote the set of feasible treatments for a patient presenting with $\mathbf{H}_t = \boldsymbol{h}_t$ at time t as $\psi_t(\boldsymbol{h}_t) \in \mathscr{A}_t$.

2.2 DTRs and optimal DTRs

With some notation established, we can now formally define a dynamic treatment regime (DTR). A DTR is a sequence of mappings $\boldsymbol{d} = (d_1, \ldots, d_T)$ such that for $t = 1, \ldots, T$, $d_t : \mathscr{H}_t \to \mathscr{A}_t$, and $d_t(\boldsymbol{h}_t) \in \psi_t(\boldsymbol{h}_t)$ for all $\boldsymbol{h}_t$. That is, at each decision point, a DTR maps from known patient features and history up to that point to a treatment. Our interest is in finding the optimal DTR, the DTR that, if followed, would lead to the greatest expected value of the outcome (assuming that larger outcomes are better). DTRs may take many forms, ranging from simple decision lists to complicated non-linear functions of features and treatments. We will denote the class of regimes under consideration as $\mathscr{D}$. The optimal DTR is how the precision medicine idea of "leveraging heterogeneity" is operationalized. Patient outcomes may be improved by assigning treatments tailored to patients' unique features and disease histories, and the optimal DTR is the rule that tells us how to make optimal tailored treatment assignments.

We formalize the concepts of treatment regime and interim information using the example of a two-stage setting. In this setting, treatments are assigned according to a

treatment regime $\boldsymbol{d} = (d_1, d_2)$, and we denote the corresponding interim information and final outcome as follows.

$$\mathbf{X}_2^*(d_1) = \sum_{a_1 \in \mathscr{A}_1} \mathbf{X}_2^*(a_1) I\{d_1(\mathbf{X}_1) = a_1\},$$
$$Y^*(\boldsymbol{d}) = \sum_{(a_1,a_2) \in \mathscr{A}_1 \times \mathscr{A}_2} I\{d_1(\mathbf{X}_1) = a_1, d_2(\mathbf{X}_1, a_1, \mathbf{X}_2^*(a_1)) = a_2\} Y^*(a_1, a_2), \quad (1)$$

where $I(\cdot)$ is an indicator function. For an arbitrary T, the potential interim information and potential outcome can be defined analogously.

Define the (marginal) value function of a regime $\boldsymbol{d}$ as

$$V(\boldsymbol{d}) = \mathbb{E}\{Y^*(\boldsymbol{d})\}, \quad (2)$$

the expected value of the outcome if all individuals received treatment assignment according to $\boldsymbol{d}$. Finally, an optimal treatment regime $\boldsymbol{d}^{\text{opt}}$ maximizes the expectation of a prespecified cumulative outcome measure Y and satisfies:

(a) for $t = 1, \ldots, T$, $\forall \boldsymbol{h}_t \in \mathscr{H}_t$, $d_t^{\text{opt}}(\boldsymbol{h}_t) \in \psi_t(\boldsymbol{h}_t)$;
(b) for $t = 1, \ldots, T$, $\forall \boldsymbol{h}_t \in \mathscr{H}_t$, for $\boldsymbol{d}$ satisfying $d_t(\boldsymbol{h}_t) \in \psi_t(\boldsymbol{h}_t)$, $\mathbb{E}Y^*(\boldsymbol{d}^{\text{opt}}) \geq \mathbb{E}Y^*(\boldsymbol{d})$.

That is, an optimal dynamic treatment regime is a regime that, if followed, would yield the greatest expected outcome when compared to other regimes in $\mathscr{D}$.

2.3 Identification

In precision medicine, our goal is to use the observed data to learn or estimate $\boldsymbol{d}^{\text{opt}}$. Thus far, however, we have described the decision-making framework and the value of DTRs in terms of potential outcomes. Because patients can receive only one treatment at each decision point, only the potential outcomes corresponding to realized treatment histories are observed, and the rest are unobserved. Without further assumptions, $V(\boldsymbol{d})$ cannot be identified from the observed data. Fortunately, the application of the g-computation algorithm [30] from the causal inference literature shows how $V(\boldsymbol{d})$ can be identified from the data under certain assumptions.

Suppose the observed data consists of n i.i.d. trajectories of the longitudinal data $\{(\mathbf{X}_t, A_t, Y_t)\})_{t=1}^T$, presented in the form $\{(\mathbf{X}_{1,i}, A_{1,i}, Y_{1,i}, \ldots, \mathbf{X}_{T,i}, A_{T,i}, Y_{T,i})\}_{i=1}^n$. To identify the value of a regime using observed data, we will assume that the following identifying assumptions hold:

1. Stable unit treatment value assumption (SUTVA): also referred to as the consistency assumption [35], which can be formalized as

$$\mathbf{X}_t = \mathbf{X}_t^*\left(\overline{\mathbf{A}}_{t-1}\right) = \sum_{\overline{\boldsymbol{a}}_{t-1} \in \mathscr{A}_1 \times \cdots \times \mathscr{A}_{t-1}} \mathbf{X}_t^*\left(\overline{\boldsymbol{a}}_{t-1}\right) I\left(\overline{\mathbf{A}}_{t-1} = \overline{\boldsymbol{a}}_{t-1}\right), \ t = 2, \ldots, T,$$
$$Y_t = Y_t^*\left(\overline{\mathbf{A}}_t\right) = \sum_{\overline{\boldsymbol{a}}_t \in \mathscr{A}_1 \times \cdots \times \mathscr{A}_t} Y_t^*\left(\overline{\boldsymbol{a}}_t\right) I\left(\overline{\mathbf{A}}_t = \overline{\boldsymbol{a}}_t\right), \ t = 1, \ldots, T.$$

In other words, the potential outcome is consistent with the observed outcome under the treatment option, and SUTVA also implies no interference between subjects and treatment variation irrelevance.

2. Sequential randomization assumption (SRA): also known as the no unmeasured confounders [30, 33], sequential ignorability, or exchangeability assumption in the multiple decision setting, which can be formalized as for any regime $\overline{\boldsymbol{a}}_t$, $t = 1, \ldots, T$,

$$\{\mathbf{X}^*_{t+1}(\overline{\boldsymbol{a}}_t), \ldots, \mathbf{X}^*_T(\overline{\boldsymbol{a}}_{T-1}), Y^*_t(\overline{\boldsymbol{a}}_t), \cdots, Y^*(\overline{\boldsymbol{a}}_T)\} \perp A_t | \mathbf{H}_t.$$

In other words, for any possible regime $\overline{\boldsymbol{a}}_t$, the current treatment A_t one receives is independent of all future potential covariates or outcomes given the observed histories $\mathbf{H}_t, t = 1, \ldots, T$, respectively. In SMART, the treatments are assigned at random; thus, SRA is automatically satisfied. However, in observational studies, SRA is untestable because it is impossible to verify from the observed data, and to tell from the data at hand if there are additional associated variables not recorded. SRA can be formalized in different ways, ranging from weak to strong versions. For further discussion, see [40].

3. Positivity: Positivity assumption guarantees the existence of observations for every treatment and each possible realized history given the covariates; otherwise, the effect of the treatment regime might not be estimable. The positivity assumption can be formalized as: for all treatments a_t that are feasible for the history $\boldsymbol{h}_t$,

$$P(A_t = a_t | \mathbf{H}_t = \boldsymbol{h}_t) > 0, \;\; t = 1, \ldots, T.$$

With these identifiability assumptions, we can introduce the conditional expectation for the outcome given covariates and treatments, that is, Q-function, which we focus on and give a detailed introduction through the following of the chapter.

3 Q-learning for Precision Medicine

In this section, we formalize Q-learning in precision medicine and contrast with reinforcement learning. We provide a conceptual overview of reinforcement learning, and illustrate how it can be linked with Q-learning as well as the distinctions between the two.

3.1 An incomplete conceptual overview of reinforcement learning to Q-learning

In computer science, reinforcement learning (RL) is a type of machine learning characterized by a learning agent and an environment that the agent wants to learn about, as illustrated in Figure 2. At each decision point, the agent observes the state of their

current environment and chooses an action in the action set. Then, the environment responds to the action by transitioning to a new state, and the agent observes a reward that corresponds to the immediate desirability of the action the agent chose. RL problems can be characterized in many different ways. One key descriptor is the time horizon or the number of stages (decision points) and thereby actions an agent takes. When the number of stages is finite, it is called a finite horizon, and when the number of stages is infinite, it is called an infinite horizon. Much of the RL literature has traditionally focused on online learning, where the agent interacts with the environment and uses experiences from those interactions to improve their decision-making. However, in many medical contexts, this is not feasible, and optimal decisions must be learned from a sample of previously collected data. When optimal policies are learned solely from previously collected data, the RL problem is described as offline.

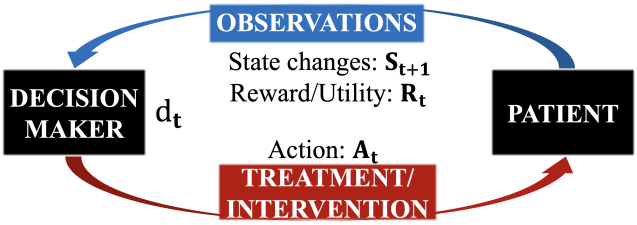

Fig. 2 Reinforcement learning flowchart.

We will mostly use the notation previously established in Section 2 throughout our formal development of RL. As the language of RL differs from that of statistics and precision medicine, before continuing, we will comment on some of these differences. In the parlance of RL, the environment in which the agent is making decisions is referred to as the set of states, usually denoted as $\mathbf{S}_t$; in precision medicine, these states correspond to patient features $\mathbf{X}_t$. Similarly, the action the agent takes at time t in the RL setting corresponds to treatment decisions A_t in the precision medicine setting, and rewards (usually denoted as R_t) correspond to proximal outcomes Y_t along with the terminal outcome $Y \equiv Y_T$. The functions mapping from the history of previous states, previous actions, and the current state to the next action are called policies in RL and correspond to DTRs d. However, it is worth noting that the term DTR is commonly used to refer to finite time horizon policies exclusively. In Table 1, we summarize the corresponding terminology between RL and statistical precision medicine, which was originally introduced in [6].

In RL, rewards are conceptualized as known functions y of the history, current action, and the next state; that is,

$$Y_t = y_t(\mathbf{H}_t, A_t, \mathbf{X}_{t+1}) = y_t(\overline{\mathbf{X}}_t, \overline{\mathbf{A}}_t, \mathbf{X}_{t+1}), \tag{3}$$

Table 1 Corresponding terminologies between the dynamic treatment regimes and reinforcement learning literature (the original table was provided in [6]).

Generic	Reinforcement learning	Dynamic treatment regimes	Symbol
Observation unit index	Trajectory	Patient	i
Time index	Time (or time point)	Stage	t
Utility	Reward (R_t)	Outcome	Y, y
Context	State ($\mathbf{S}_t$)	Covariate	X, x
Decision	Action	Treatment	A, a
Decision strategy	Policy	(Dynamic) treatment regime	d

and as before, we will assume that larger outcomes are better. The goal of RL is to learn the map, i.e., policy, from state space $\mathscr{H}_t$ to action space $\mathscr{A}_t$ to maximize the total (discounted) reward. For ease of exposition, we let $T = \infty$ for the remainder of this subsection as we develop the main ideas from RL, and we will return to the finite horizon case in Section 3.2. For time t, $1 \leq t \leq \infty$, the goal can be written as $\max \sum_{k \geq 0} \gamma^k Y_{t+k}$, where $\gamma \in [0,1]$ is a discount factor. When $\gamma = 0$, the agent is short-sighted and evaluates the action only with the current award; when γ approaches 1, the agent learns the action in each iteration more based on the long-term reward. In the infinite time horizon setting, it is required that $\gamma < 1$ to guarantee the convergence of cumulative reward.

An important function in RL is the state-value function, the total expected future rewards of an agent in a particular state if the agent were to follow the given policy thereafter. Letting d be a policy from the class of all policies $\mathscr{D}$, we can write the value function at time t for state $\boldsymbol{h}_t$ as

$$V_t^d(\boldsymbol{h}_t) = \mathbb{E}_d \left(\sum_{k \geq 0} \gamma^k Y_{t+k} | \mathbf{H}_t = \boldsymbol{h}_t \right), \tag{4}$$

where $\mathbb{E}_d$ denotes that the expectation is taken with respect to the trajectories that would be generated under policy d. Immediately, it can be seen that one could find the optimal policy, the policy that yields the highest (discounted) reward, starting at time t in state $\boldsymbol{h}_t$ by taking the policy that yields the highest value of (4). Except for the simplest of settings, using (4) in this way is impractical. Instead, in RL, one typically uses a recursive formulation of the state-value function to learn optimal policies:

$$\begin{aligned}
V_t^d(\boldsymbol{h}_t) =& \mathbb{E}_d\left(\sum_{k\geq 0}\gamma^k Y_{t+k}\middle|\mathbf{H}_t=\boldsymbol{h}_t\right)\\
=& \mathbb{E}_d\left(Y_t\middle|\mathbf{H}_t=\boldsymbol{h}_t\right)+\mathbb{E}_d\left(\sum_{k\geq 1}\gamma^k Y_{t+k}\middle|\mathbf{H}_t=\boldsymbol{h}_t\right)\\
=& \mathbb{E}_d\left(Y_t\middle|\mathbf{H}_t=\boldsymbol{h}_t\right)+\mathbb{E}_d\left\{\mathbb{E}\left(\gamma\sum_{k\geq 0}\gamma^k Y_{(t+1)+k}\middle|\mathbf{H}_{t+1}\right)\middle|\mathbf{H}_t=\boldsymbol{h}_t\right\}\\
=& \mathbb{E}_d\left(Y_t\middle|\mathbf{H}_t=\boldsymbol{h}_t\right)+\gamma\mathbb{E}_d\left\{V_{t+1}^d\left(\mathbf{H}_{t+1}\middle|\mathbf{H}_t=\boldsymbol{h}_t\right)\right\}\\
=& \mathbb{E}_d\left\{Y_t+\gamma V_{t+1}^d\left(\mathbf{H}_{t+1}\right)\middle|\mathbf{H}_t=\boldsymbol{h}_t\right\}. \qquad (5)
\end{aligned}$$

As we will see, this recursive formulation of the state-value function, and later the action-value function, will be essential for learning optimal policies in the infinite time horizon setting and gives us a general strategy for learning optimal policies.

In terms of the time t state-value function, we can express our goal of learning an optimal policy as

$$V_t^{\text{opt}}(\boldsymbol{h}_t)=\max_{d\in\mathscr{D}}V_t^d(\boldsymbol{h}_t). \qquad (6)$$

Put otherwise, an optimal policy is the policy in $\mathscr{D}$ that if we start in state h_t maximizes the total (discounted) future reward, and we let V^{opt} denote the optimal state-value. We say "an" optimal policy as opposed to "the" optimal policy since more than one policy in $\mathscr{D}$ may optimize the state-value function. Optimal polices are related to the state-value function via the Bellman optimality equation:

$$\begin{aligned}
V_t^{\text{opt}}(h_t) &= \max_{a_t\in\mathscr{A}_t}\mathbb{E}_{d^{\text{opt}}}\left(\sum_{k\geq 0}\gamma^k Y_t\middle|\mathbf{H}_t=\boldsymbol{h}_t,A_t=a_t\right)\\
&= \max_{a_t\in\mathscr{A}_t}\mathbb{E}\left\{Y_t+\gamma\sum_{k\geq 1}\gamma^k Y_{(t+1)+k}\middle|\mathbf{H}_t=\boldsymbol{h}_t,A_t=a_t\right\}\\
&= \max_{a_t\in\mathscr{A}_t}\mathbb{E}_{d^{\text{opt}}}\left\{Y_t+\gamma V_{t+1}^{\text{opt}}(\mathbf{H}_{t+1})\middle|\mathbf{H}_t=\boldsymbol{h}_t,A_t=a_t\right\}. \qquad (7)
\end{aligned}$$

One can see that the optimal policy satisfies

$$d_t^{\text{opt}}(h_t)\in\arg\max_{a_t\in\mathscr{A}_t}\mathbb{E}\left\{Y_t+\gamma V_{t+1}^{\text{opt}}(\mathbf{H}_{t+1})\middle|\mathbf{H}_t=\boldsymbol{h}_t,A_t=a_t\right\}. \qquad (8)$$

In precision medicine, we will often be interested in the marginal value function

$$V^d=\mathbb{E}_{\mathbf{H}_1}\left\{V^d(\boldsymbol{h}_1)\right\}, \qquad (9)$$

where $\mathbb{E}_{\mathbf{H}_1}$ denotes expectation with respect to the baseline distribution of patient features. Averaging over the baseline features can be thought of as taking the average

value of a policy over the population to which it is applied. The marginal value function corresponds to the value function from the precision medicine framework defined in Section 2.2.

Before going further, it is worth mentioning that the first class of methods developed to solve multistage decision problems is called dynamic programming and was introduced by [2]. Dynamic programming is not the most practical strategy because it depends on fully knowing the system dynamics of the learning environment, although work by [3, 9, 29] have adapted dynamic programming to environments with unknown transition dynamics. A breakthrough in RL came in 1989 when Watkins developed Q-learning, a way to learn optimal policies based on sample data trajectories. This approach is known as off-policy learning as it allows for learning policies outside of the policy that generates the data trajectory. Broadly speaking, in Q-learning, one seeks to find the action that yields the high expected reward given the current state rather than determining the expected value of a given state. Indeed, Q-learning methods focus on the Q-function, which returns the quality of an action. The Q-function, also referred to as the action-value function, at time t, is defined as

$$Q_t^d(\boldsymbol{h}_t, a_t) = \mathbb{E}_d\left(\sum_{k\geq 0}\gamma^k Y_{t+k}\Big|\mathbf{H}_t = \boldsymbol{h}_t, A_t = a_t\right). \tag{10}$$

The Q-function at time t yields the total (discounted) expected return if one were in state $\boldsymbol{h}_t$, took action a_t, and then followed policy d thereafter. Like the value function, the Q-function can be defined recursively

$$Q_t^d(\boldsymbol{h}_t, a_t) = \mathbb{E}_d\left\{Y_t + \gamma Q_{t+1}^d(\mathbf{H}_{t+1}, A_{t+1})\Big|\mathbf{H}_t = \boldsymbol{h}_t, A_t = a_t\right\}. \tag{11}$$

Similar to the development of the optimal state-value function, the optimal Q-function relates the recursive formulation of the Q-function to an optimal policy

$$Q^{\text{opt}}(\boldsymbol{h}_t, a_t) = \mathbb{E}_{d^{\text{opt}}}\left\{Y_t + \gamma \max_{a_{t+1}\in\mathscr{A}_{t+1}} Q^{\text{opt}}(\mathbf{H}_{t+1}, a_{t+1})\Big|\mathbf{H}_t = \boldsymbol{h}_t, A_t = a_t\right\}, \tag{12}$$

and the optimal decision at time t can be derived directly from the Q-function as

$$d_t^{\text{opt}} = \arg\max_{a_t\in\mathscr{A}_t} Q_t^{\text{opt}}(\boldsymbol{h}_t, a_t). \tag{13}$$

Before connecting RL to the precision medicine context, we mention that the value function is closely related to the Q-function, i.e., $V_t^d(\mathbf{h}_t) = \max_{a_t\in\mathscr{A}_t} Q_t(\mathbf{h}_t, a_t)$.

3.2 Q-learning in the context of precision medicine

So far, we have described RL generally. While the framework echoes medical decision-making, some aspects differ from the precision medicine framework de-

scribed in Section 2. For example, we described medical decision-making in terms of potential outcomes which are generally not part of the RL setting. Moreover, we have purposefully, to this point, avoided making any explicit assumptions about the transition probabilities between states, and some readers familiar with the RL literature will have certainly noticed that we have not (yet) invoked the Markov Decision Process (MDP) structure (although we will shortly). In fact, we implicitly allowed for non-Markovian transition probabilities as we wrote treatment regimes as functions of the entire accrued history of patients. Susan Murphy's seminal works [21, 22] united the RL literature with the potential outcomes framework, commonly used to describe treatment regimes, and regression modeling to develop what is referred to as Q-learning in the statistical literature. Her work provided an algorithm for the finite horizon precision medicine setting to learn the optimal dynamic treatment by recursively estimating Q-functions, one for each decision point. This method avoids the need to know or estimate the transition dynamics, making it possible to use Q-learning for optimal policy guidance without assuming the overall MDP structure or knowing the specific transition probability from one stage to another.

For the remainder of this section, we will describe the optimal Q-functions for learning optimal policies in the precision medicine context, first in the finite horizon setting and then in the infinite horizon setting. Learning the optimal Q-functions is essential for learning optimal policies using Q-learning, and we will discuss some of the nuances of applying Q-learning in the precision medicine framework. Later, Sections 4 and 5 will focus on the implementation of Q-learning in the finite horizon and infinite horizon settings.

For precision medicine problems in the finite horizon setting, the optimal Q-functions are defined as follows:

$$\begin{aligned} Q_T^{\text{opt}}(\boldsymbol{h}_T, a_T) =& \mathbb{E}\left(Y_T \middle| \mathbf{H}_t = \boldsymbol{h}_t, A_T = a_T\right), \text{ and} \\ Q_t^{\text{opt}}(\boldsymbol{h}_t, a_t) =& \mathbb{E}\left\{ Y_t + \gamma \max_{a \in \psi(\mathbf{H}_{t+1})} Q_{t+1}(\mathbf{H}_{t+1}, a) \middle| \mathbf{H}_t = \boldsymbol{h}_t, A_t = a_t \right\} \end{aligned} \tag{14}$$

for $t = 1, \ldots, T-1$. We observe that once Q_T^{opt} is defined, the $t = 1, \ldots, T-1$ optimal Q-functions are defined recursively as in (12). The maximization in (14) is taken over the feasible treatments $\psi(\mathbf{H}_{t+1}) \in \mathscr{A}_{t+1}$ rather than $\mathscr{A}_{t+1}$ as in (12). This is to reflect the realities of medical decision-making where not all treatments may be feasible for all patients, e.g., due to an allergy or a known counterindication for a particular treatment, a situation that is generally not encountered in general reinforcement learning problems. Often, in the finite horizon precision medicine setting, γ is taken to be 1 so that the reward is the undiscounted sum of the proximal and terminal outcomes. Futhermore, it is often the case that there is only one terminal outcome of interest Y_T, and we take $Y_t = 0$ for $t = 1, \ldots, T$. As we describe in Section 4, the algorithm for learning the optimal DTR using (14) proceeds iteratively via backwards induction.

To estimate optimal DTRs in the infinite time horizon setting in precision medicine, additional modeling assumptions are required. A common strategy is to

model decision-making as an MDP as in the traditional RL literature. An MDP assumes constant states that describe an individual's health status and constant actions available at each time point $t = 1, \ldots, T$, denoted as $\mathscr{X}$ and $\mathscr{A}$, respectively. It is also assumed that the reward function $y : \mathscr{X} \times \mathscr{A} \times \mathscr{X} \to \mathbb{R}$ is constant, for $t = 1, \ldots, T$. That means, for an individual with health status $\boldsymbol{x}$, receiving treatment a, and transition to health status $\boldsymbol{x}'$, the proximal reward, $y(\boldsymbol{x}, a, \boldsymbol{x}')$, is constant. Moreover, an MDP assumes that the transitions between states are time-homogeneous and Markovian for all t so that formally $P(\cdot|\mathbf{H}_t, A_t) = P(\cdot|\mathbf{X}_t, A_t)$ [38]. In this setting, the optimal infinite-horizon Q-function is defined as

$$Q^{\text{opt}}(\boldsymbol{x}, a) = \mathbb{E}_{d^{\text{opt}}} \left(\sum_{t=1}^{\infty} \gamma^{t-1} Y_t \Big| \mathbf{X}_1 = \boldsymbol{x}, A_1 = a \right), \tag{15}$$

where as before $\gamma \in [0, 1)$ is a discount factor. Because the optimal infinite-horizon Q-function is the Q-function that corresponds to the optimal policy d^{opt}, a general strategy of Q-learning in this setting is to estimate Q^{opt} and thereby learn the optimal DTR. Rather than using the optimal infinite-horizon Q-function directly, a recursive formulation is used to estimate Q^{opt}. Observe that

$$\begin{aligned} Q^{\text{opt}}(\boldsymbol{x}, a) =& \mathbb{E}_{d^{\text{opt}}} \left(\sum_{t=1}^{\infty} \gamma^{t-1} Y_t \Big| \mathbf{X}_1 = \boldsymbol{x}, A_1 = a \right) \\ =& \mathbb{E} \left\{ Y_1 + \gamma \max_{a'} \mathbb{E}_{d^{\text{opt}}} \left(\sum_{t=1}^{\infty} \gamma^{t-1} Y_{t+1} | \mathbf{X}_{t+1}, A_1 = a' \right) \Big| \mathbf{X}_1 = \boldsymbol{x}, A_1 = a \right\} \\ =& \mathbb{E}_{d^{\text{opt}}} \left\{ Y_1 + \gamma \max_{a'} Q^{\text{opt}}(\mathbf{X}_2, a') \big| \mathbf{X}_1 = \boldsymbol{x}, A_1 = a \right\}, \end{aligned}$$

so that in general

$$Q^{\text{opt}}(\boldsymbol{x}, a) = \mathbb{E} \left\{ Y_t + \gamma \max_{a'} Q^{\text{opt}}(\mathbf{X}_{t+1}, a) \big| \mathbf{X}_t = \boldsymbol{x}, A_t = a \right\},$$

which is precisely the MDP analogue of (12). We let B denote the Bellman optimality operator defined as

$$(Bf)(\boldsymbol{x}, a) = \mathbb{E} \left\{ Y_t + \gamma \max_{a' \in \mathscr{A}} f(\mathbf{X}_{t+1}, a') | \mathbf{X}_t = \boldsymbol{x}_t, A_t = a \right\} \tag{16}$$

for a function f, then the optimal infinite-horizon Q-function can be written as $Q^{\text{opt}}(\boldsymbol{x}, a) = (BQ^{\text{opt}})(\boldsymbol{x}, a)$. From this representation, we observe that Q^{opt} is a fixed point and motivates what is called the value-iteration algorithm: $Q_{t+1} \to BQ_t$ for $T = 1, 2, \ldots$, until convergence. We will see in Section 5 how this recursive expression and value iteration are used to estimate the optimal infinite-horizon Q-function.

4 Q-Learning for Optimal DTRs in the Finite Time Horizon Setting

To describe how Q-learning is used in the finite-horizon settings, we start with the single-decision setting ($T = 1$), which is highly instructive before generalizing to the multistage decision setting ($2 \leq T < \infty$). Although we only have one decision point, we still refer to the learned regime as "dynamic" because it still tailors treatments to patient features and patient history, unlike a "static" regime in which the treatment(s) is fixed in advance with a priori.

4.1 Q-learning for optimal single-stage DTRs

For the single-decision setting, a DTR is a function $d : \mathscr{X} \to \mathscr{A}$ that satisfies $d(\boldsymbol{x}) \in \psi(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathscr{X}$. As in (1), we denote the potential outcome that would be observed if treatments were assigned according to $d \in \mathscr{D}$ as

$$Y^*(d) = \sum_{a \in \mathscr{A}} Y^*(a) I\{d(\mathbf{X}) = a\}.$$

The value of a DTR and the definition of the optimal DTR are given analogously in Section 2.2.

Under the requisite identifying assumptions, the optimal Q-function in the single-decision setting is defined as in (14). Since the first stage is the terminal stage, we simply have

$$Q(\boldsymbol{x}, a) = \mathbb{E}(Y | \mathbf{X} = \boldsymbol{x}, A = a),$$

which suggests $d^{\text{opt}} = \arg\max_{a \in \psi(\boldsymbol{x})} Q(\boldsymbol{x}, a)$ and that the value of regime d^{opt} is

$$V(d^{\text{opt}}) = \mathbb{E}\left\{ \max_{a \in \mathscr{A}} Q(\mathbf{X}, a) \right\}.$$

Because the Q-function in this case is a conditional mean model, it is reasonable to propose to estimate the Q-function using regression modeling [40]. That is, if we posit a regression model for $Q(\boldsymbol{x}, a)$ and suppose the model is correctly specified, we can construct a regression estimator $\widehat{Q}_n(\boldsymbol{x}, a)$ of $Q(\boldsymbol{x}, a)$, and plug it in to get $\widehat{d}^{\text{opt}}(\boldsymbol{x}) = \arg\max_{a \in \psi(\boldsymbol{x})} \widehat{Q}_n(\boldsymbol{x}, a)$.

As a simple illustration, suppose $\mathscr{X} \subset \mathbb{R}$, $\mathscr{A} = \{0, 1\}$, $\psi(x) = \mathscr{A}$ for all $x \in \mathscr{X}$, and Y is continuous. We posit the following parametric regression model for the Q-function:

$$Q(x, a; \boldsymbol{\beta}) = \beta_0 + x\beta_1 + a\beta_2 + ax\beta_3, \tag{17}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^\top$ are regression parameters. We let $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3)^\top$ denote estimates of the regression parameters that can be obtained via least-squares regression or other M-estimation methods. Then $\widehat{Q}_n(x, a; \widehat{\boldsymbol{\beta}})$ is our estimator for the

Q-function, and the optimal DTR is

$$\begin{aligned}\widehat{d}^{\text{opt}}(x) &= \underset{a\in\psi(x)}{\arg\max}\, \widehat{Q}_n(x,a;\widehat{\boldsymbol{\beta}}) \\ &= \underset{a\in\{0,1\}}{\arg\max}\, \widehat{\beta}_0 + x\widehat{\beta}_1 + a\widehat{\beta}_2 + ax\widehat{\beta}_3 \\ &= \underset{a\in\{0,1\}}{\arg\max}\, \widehat{\beta}_0 + x\widehat{\beta}_1 + a\left(\widehat{\beta}_2 + x\widehat{\beta}_3\right) \\ &= I\left(\widehat{\beta}_2 + x\widehat{\beta}_3 > 0\right).\end{aligned}$$

Furthermore, $V(d^{\text{opt}})$ can be estimated by

$$\begin{aligned}\widehat{V}(d^{\text{opt}}) &= \mathbb{P}_n \max_{a\in\psi(x)} Q(x_i,a;\widehat{\boldsymbol{\beta}}) \\ &= \mathbb{P}_n\left\{\widehat{\beta}_0 + x_i\widehat{\beta}_1 + I\left(\widehat{\beta}_2 + x_i\widehat{\beta}_3 > 0\right)\left(\widehat{\beta}_2 + x_i\widehat{\beta}_3\right)\right\}.\end{aligned}$$

Of note, in Q-learning, the class of treatment regimes under consideration is determined by the Q-function. This is illustrated in (17); the class of regimes under consideration are those of the form $I(\beta_2 + x\beta_3 > 0)$ and indexed by the parameters β_2 and β_3.

4.2 Additional considerations for the single-stage setting

Before proceeding, it is worth mentioning the relationship between the DTRs and optimal DTRs in the single-decision setting and the conditional average treatment effect (CATE) often encountered in the causal inference and other related literature. To illustrate this relationship, suppose momentarily that $\mathscr{A} = \{0,1\}$. In this scenario, the CATE is defined as

$$\Delta(\boldsymbol{x}) = \mathbb{E}\{Y^*(1) - Y^*(0)|\mathbf{X} = \boldsymbol{x}\}.$$

The CATE represents the difference in the average outcome that would be achieved if individuals with baseline features $\boldsymbol{x}$ were treated with $A = 1$ and the average outcome that would be achieved if individuals with baseline features $\boldsymbol{x}$ were treated with $A = 0$. In the parlance of DTRs, treating everyone with $A = 1$ is equivalent to the particular DTR $d(\mathbf{X}) \equiv 1$, and the value of this DTR is $\mathbb{E}\{Y^*(1)\}$; similarly, treating everyone in the population with $A = 0$ is equivalent to the particular DTR $d(\mathbf{X}) \equiv 0$ and the value of this particular DTR is $\mathbb{E}\{Y^*(0)\}$. Thus the expected value of the CATE taken over the covariates is the difference of the value of the regime $d(\mathbf{X}) \equiv 1$ and $d(\mathbf{X}) \equiv 0$, i.e., $\mathbb{E}\{\Delta(\mathbf{X})\} = V\{d(\mathbf{X}) \equiv 1\} - V\{d(\mathbf{X}) \equiv 0\}$. Generally, if the optimal treatment regime is an indicator function of the linear combination of

the covariates, CATE will give a good estimate for d^{opt}. See more discussions and examples in [15].

Statistical inference and large sample properties of DTRs and related estimators are of great statistical interest. Consider the single-stage decision setting. In this case, estimating the optimal DTR using Q-learning is relatively straightforward since regression-based analysis may be employed. Asymptotic consistency and asymptotic normality of the regression coefficients are achievable under standard regularity conditions. However, the conventional asymptotic theory cannot be applied directly to the estimators for the maximum value of the regime, even in a simple case. Let's revisit the simple example in (17). It is evident that the asymptotic distributions of $\widehat{\beta}_0$, $\widehat{\beta}_1$, $\widehat{\beta}_2$ and $\widehat{\beta}_3$ are jointly normally distributed; however, $\widehat{V}(d^{\text{opt}})$ is not asymptotically normally distributed due to the non-differentiability of the indicator function. In general, because of the involvement of the non-smooth maximum operator in definition, it is difficult to provide inference for $V(d^{\text{opt}})$, and the difficulty extends to the multistage decision setting; see [40] for more discussion.

4.3 Q-learning for optimal finite-stage DTRs

In the multistage decision setting ($2 \leq T < \infty$), where there are multiple times at which treatment decisions need to be made, it is crucial to account for both the immediate and long-term impacts of treatment [15]. As we will see in this section, the backward induction algorithm can help us do so when there are multiple decisions and thus sequences of treatments at hand.

The characterization of the optimal regime in the multistage decision setting is more complicated than that in the single-decision setting, and the backward induction algorithm is usually employed to define the optimal regime. The backward inductive argument starts at the end and moves backwards through the stages. To demonstrate the reasoning, we first consider a simple two-stage case, that is, $T = 2$. In this case, the Q-functions are defined as

$$\begin{aligned} Q_2(\boldsymbol{h}_2, a_2) &= \mathbb{E}\left(Y | \mathbf{H}_2 = \boldsymbol{h}_2, A_2 = a_2\right), \\ Q_1(\boldsymbol{h}_1, a_1) &= \mathbb{E}\left\{ \max_{a_2} Q_2(\boldsymbol{h}_2, a_2) | \mathbf{H}_1 = \boldsymbol{h}_1, A_1 = a_1 \right\}, \end{aligned} \tag{18}$$

or equivalently,

$$\begin{aligned} Q_2(\boldsymbol{x}_1, a_1, \boldsymbol{x}_2, a_2) &= \mathbb{E}\left(Y | \mathbf{X}_1 = \boldsymbol{x}_1, A_1 = a_1, \mathbf{X}_2 = \boldsymbol{x}_2, A_2 = a_2\right), \\ Q_1(\boldsymbol{x}_1, a_1) &= \mathbb{E}\left\{ \max_{a_2} Q_2(\mathbf{X}_1, A_1, \boldsymbol{x}_2, a_2) | \mathbf{X}_1 = \boldsymbol{x}_1, A_1 = a_1 \right\}. \end{aligned}$$

It follows from dynamic programming [2] that

$$d_2^{\text{opt}}(\boldsymbol{h}_2) = \underset{a_2 \in \psi_2(\boldsymbol{h}_2)}{\arg\max}\, Q_2(\boldsymbol{h}_2, a_2), \quad d_1^{\text{opt}}(\boldsymbol{h}_1) = \underset{a_1 \in \psi_1(\boldsymbol{h}_1)}{\arg\max}\, Q_1(\boldsymbol{h}_1, a_1).$$

Similar to the illustration in the single-decision setting, we consider the simple example that $\mathscr{X} \subset \mathbb{R}$, $\mathscr{A} = \{0,1\}$, Y is continuous, and the Q-functions are regression models with

$$Q_2(\boldsymbol{h}_2, a_2; \boldsymbol{\beta}_{21}, \boldsymbol{\beta}_{22}) = \boldsymbol{h}_2^\top \boldsymbol{\beta}_{21} + a_2 \left(\boldsymbol{h}_2^\top \boldsymbol{\beta}_{22}\right),$$
$$Q_1(h_1, a_1; \beta_{11}, \beta_{12}) = h_1\beta_{11} + a_1 (h_1\beta_{12}).$$

By solving M-estimating equations, the following steps are involved to obtain the coefficient estimators:

Step I. Stage-2 regression coefficients

$$(\widehat{\boldsymbol{\beta}}_{21}, \widehat{\boldsymbol{\beta}}_{22}) = \underset{\boldsymbol{\beta}_{21}, \boldsymbol{\beta}_{22}}{\arg\min}\, \mathbb{P}_n \left\{ Y - \boldsymbol{h}_2^\top \boldsymbol{\beta}_{21} - a_2 \left(\boldsymbol{h}_2^\top \boldsymbol{\beta}_{22}\right) \right\}^2.$$

Step II. Stage-1 regression coefficients

$$(\widehat{\beta}_{11}, \widehat{\beta}_{12}) = \underset{\beta_{11}, \beta_{12}}{\arg\min}\, \mathbb{P}_n \left[\max_{a_2} \left\{ \boldsymbol{h}_2^\top \widehat{\boldsymbol{\beta}}_{21} + a_2 \left(\boldsymbol{h}_2^\top \widehat{\boldsymbol{\beta}}_{22}\right) \right\} - h_1\beta_{11} - a_1 (h_1\beta_{12}) \right]^2.$$

Once the Q-functions have been estimated, the optimal rule can be estimated by

$$\begin{aligned}
\widehat{d}_2^{\text{opt}}(\boldsymbol{h}_2) &= \underset{a_2}{\arg\max}\, Q_2 \left(\boldsymbol{h}_2, a_2; \widehat{\boldsymbol{\beta}}_{21}, \widehat{\boldsymbol{\beta}}_{22}\right) \\
&= I\left\{ Q_2(\boldsymbol{h}_2, 1; \widehat{\boldsymbol{\beta}}_{21}, \widehat{\boldsymbol{\beta}}_{22}) > Q_2(\boldsymbol{h}_2, 0; \widehat{\boldsymbol{\beta}}_{21}, \widehat{\boldsymbol{\beta}}_{22}) \right\} = I\left(\boldsymbol{h}_2^\top \widehat{\boldsymbol{\beta}}_{22} > 0\right), \\
\widehat{d}_1^{\text{opt}}(h_1) &= \underset{a_1}{\arg\max}\, Q_1 \left(h_1, a_1; \widehat{\beta}_{11}, \widehat{\beta}_{12}\right) \\
&= I\left\{ Q_1(h_1, 1; \widehat{\beta}_{11}, \widehat{\beta}_{12}) > Q_1(h_1, 0; \widehat{\beta}_{11}, \widehat{\beta}_{12}) \right\} = I\left(h_1 \widehat{\beta}_{12} > 0\right).
\end{aligned}$$

The procedure for determining the optimal regime can be extended to the case where there are more than two decisions, $T > 2$, and (18) can be generalized as

$$\begin{aligned}
Q_T(\boldsymbol{h}_T, a_T) &= \mathbb{E}\left(Y | \mathbf{H}_T = \boldsymbol{h}_T, A_T = a_T\right), \\
Q_t(\boldsymbol{h}_t, a_t) &= \mathbb{E}\left\{ \max_{a_{t+1}} Q_{t+1}(\boldsymbol{h}_{t+1}, a_{t+1}) | \mathbf{H}_t = \boldsymbol{h}_t, A_t = a_t \right\}, \quad t = 1, \ldots, T-1.
\end{aligned}$$

The estimated optimal DTR is

$$\widehat{d}_t^{\text{opt}} = \underset{a_t}{\arg\max}\, \widehat{Q}_t(\boldsymbol{h}_t, a_t), \quad t = 1, \ldots, T.$$

4.4 Additional considerations for the general finite-stage setting

Obviously, the success of this approach relies on the modeling assumption and the high-quality estimator of the Q-function. An alternative choice to Q-learning is A-learning or advantage learning [21], which also can be viewed as a special case of g-estimation [32, 37]. Both Q-learning and A-learning rely on the Bellman equation to learn the optimal policy as opposed to directly optimizing an objective function, and are sometimes referred to as indirect methods. A-learning does not require full knowledge of the Q-function, i.e., the conditional mean of the outcome; instead, A-learning models contrast functions to construct the estimate. Under the model misspecification for Q-function, A-learning is more robust; and when the model of Q-function is correctly specified, Q-learning is more efficient. One can refer to [36] for more details on comparisons, discussions and implementations for these two popular methods. Finally, we note that as in the single-decision setting, inference for the value of a multistage decision rule is complicated and in general non-regular.

5 Q-Learning for Optimal DTRs in the Infinite Time Horizon Setting

For some diseases, decision support for disease management and prevention requires many similar decisions to be made over time, as illustrated in Figure 2. For example, individuals living with diabetes make decisions about exercise, food, and insulin dosage multiple times every day to manage their blood glucose. In these cases, an infinite time horizon formulation of the decision problem is used. As previously described in Section 3, infinite-horizon problems are modeled as MDPs. Letting $\psi(\boldsymbol{x}_t) \in \mathscr{A}$ denote the feasible treatments for an individual with health status $\boldsymbol{x}_t$ as in the finite time horizon setting, a dynamic treatment regime is a map $d : \mathscr{X} \to \mathscr{A}$ such that $d(\boldsymbol{x}) \in \psi(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathscr{X}$. Note that by construction, d is stationary, and thus can be applied for all t including $t > T$. As in the finite time horizon setting, the potential outcomes framework is used to characterize the decision-making problem. Then we can write the potential health status if regime d had been followed as

$$\mathbf{X}_t^*(d) = \sum_{\overline{\boldsymbol{a}}_{t-1} \in \mathscr{A}_1 \times \cdots \times \mathscr{A}_{t-1}} \mathbf{X}_t^*(\overline{\boldsymbol{a}}_{t-1}) \prod_{k=1}^{t-1} I\left[d\{\mathbf{X}_k^*(\overline{\boldsymbol{a}}_{k-1})\} = a_k\right],$$

and the potential momentary outcome for that would be observed under d as $Y_t^*(d) = y(\mathbf{X}_t^*(d), d(\mathbf{X}_t^*(d)), \mathbf{X}_{t+1}^*(d))$, where $y(\cdot,\cdot,\cdot)$ is the reward function defined in Section 3.2.

We assume that the observed data used to learn d^{opt} are n i.i.d. trajectories $\{(\mathbf{X}_{1,i}, A_{1,i}, \mathbf{X}_{2,i}, \ldots, \mathbf{X}_{T-1,i}, A_{T-1,i}, \mathbf{X}_{T,i})\}_{i=1}^n$. Although data are collected over a finite time horizon T, the Markovian structure of the MDP enables us to learn an optimal DTR that can be applied repeatedly over time. Under the same identifying

assumptions as in the finite horizon setting, the optimal infinite-horizon Q-function is identified, and we repeat its Bellman equation representation here:

$$Q^{d^{\text{opt}}}(\boldsymbol{x},a)=\mathbb{E}_{d^{\text{opt}}}\left(\sum_{t=1}^{\infty}\gamma^{t-1}Y_t|\mathbf{X}_1=\boldsymbol{x},A_1=a\right)=(BQ^{\text{opt}})(\boldsymbol{x},a).$$

With the increasing prevalence of long-term chronic disease conditions requiring ongoing management, there is growing interest in the infinite-horizon precision medicine problem, and it is an open area of research. Following [6], we describe two such methods that have been proposed using Q-functions. The first is an estimating equation approach proposed by [10]. They model the optimal infinite-horizon Q-function with a differentiable parametric model $Q(\cdot;\boldsymbol{\theta})$. Substituting this into the Bellman equation yields

$$Q(\boldsymbol{x},a;\boldsymbol{\theta})=\mathbb{E}\left\{Y_t+\gamma\max_{a'}Q(\mathbf{X}_{t+1},a';\boldsymbol{\theta})|\mathbf{X}_t=\boldsymbol{x},A_t=a\right\}.$$

Then it follows that

$$\begin{aligned}
&\mathbb{E}\left\{Y_t+\gamma\max_{a'}Q(\mathbf{X}_{t+1},a';\boldsymbol{\theta})-Q(\boldsymbol{x},a;\boldsymbol{\theta})|\mathbf{X}_t=\boldsymbol{x},A_t=a\right\}=0\\
\Rightarrow&\mathbb{E}\left[\left\{Y_t+\gamma\max_{a'}Q(\mathbf{X}_{t+1},a';\boldsymbol{\theta})-Q(\boldsymbol{x},a;\boldsymbol{\theta})\right\}\nabla_{\boldsymbol{\theta}}Q(\boldsymbol{x},a;\boldsymbol{\theta})|\mathbf{X}_t=\boldsymbol{x},A_t=a\right]=\mathbf{0}\\
\Rightarrow&\mathbb{E}\left[\sum_{t=1}^{T-1}\left\{Y_t+\gamma\max_{a'}Q(\mathbf{X}_{t+1},a';\boldsymbol{\theta})-Q(\boldsymbol{x},a;\boldsymbol{\theta})\right\}\nabla_{\boldsymbol{\theta}}Q(\boldsymbol{x},a;\boldsymbol{\theta})|\mathbf{X}_t=\boldsymbol{x},A_t=a\right]=\mathbf{0},
\end{aligned}$$

yielding an estimating equation for $\boldsymbol{\theta}$. The empirical estimating equation is given by

$$\mathbb{P}_n\left(\sum_{t=1}^{T}\left[\left\{Y_t+\gamma\max_{a'}Q(\mathbf{X}_{t+1},a';\boldsymbol{\theta})-Q(\boldsymbol{x},a;\boldsymbol{\theta})\right\}\nabla_{\boldsymbol{\theta}}Q(\boldsymbol{x},a;\boldsymbol{\theta})\right]\right)=\mathbf{0}.$$

For a linear model for Q^{opt}, [10] showed that the solution to the estimating equation is consistent and asymptotically normal under some regularity conditions.

In [9], an alternative approach to learning the optimal dynamic treatment regime in the infinite-horizon setting using Q-learning was proposed, called fitted Q-iteration (FQI). Similar in spirit to the Q-learning in the finite-horizon setting, FQI fits a sequence of candidate Q-functions. Let $\mathscr{Q}_0,\mathscr{Q}_1,\ldots$ be a sequence of candidate Q-functions, then let $\widehat{Q}_0\leftarrow\arg\min_{Q_0\in\mathscr{Q}_0}\mathbb{P}_n[\sum_{t=1}^{T}\{Y_t-Q_0(\mathbf{X}_t,A_t)\}^2]$. The algorithm proceeds as follows:

for $k = 0, 1, \ldots,$ until convergence,

$$Q_{k+1} \leftarrow \underset{Q_{k+1} \in \mathcal{Q}_{k+1}}{\arg\min}\ \mathbb{P}_n \left[\sum_{t=1}^{T} \left\{ Y_t + \gamma \max_a \widehat{Q}_k(\mathbf{X}_t, a) - Q_{k+1}(\mathbf{X}_t, A_t) \right\}^2 \right].$$

FQI allows for highly flexible models to be used for the Q-functions; however, it suffers from bias and nonregularity. To address these limitations, [5] offered a variant of FQI with linear function approximations, which helps to reduce bias.

6 An Illustrative Example

To concretize finite-horizon Q-learning and the backward induction approach, we present a simple example of Q-learning implementation for the two-stage setting in the R statistical programming language. We will use the `bmiData` dataset from the `DynTxRegime` package [13], a package that features a number of functions and methods for implementing Q-learning and other precision medicine methods. Although the `DynTxRegime` package includes a function for Q-learning, `qLearn`, we will use primitive R functions for tractability and ease of exposition in the following example.

The `bmiData` dataset mimics a two-stage randomized clinical trial that studied obesity in adolescents. At the first randomization point, adolescents in the study are randomized to meal replacement shakes "MR" or their conventional diet "CD". The trial collected data to study the effect of meal replacement shakes on adolescent obesity, with the primary endpoint being body mass index (BMI) defined as body weight (kg) divided by squared height (m^2). Throughout, we assume that there is no missing data, there were no participants lost to follow-up, and all study participants were compliant with their assigned treatment. We begin by loading the `DynTxRegime` package. Once loaded, we load the `bmiData` dataset and use the function `head` to take a look at the first five observations in the dataset.

```
# Load the DynTxRegime package
library(DynTxRegime)

# Load the bmiData dataset
data("bmiData")

# Take a peek at the bmiData
head(bmiData, 5)
```

```
  gender race parentBMI baselineBMI month4BMI month12BMI A1 A2
1      0    1  31.59683    35.84005  34.22717   34.27263 CD MR
2      1    0  30.17564    37.30396  36.38014   36.38401 CD MR
3      1    0  30.27918    36.83889  34.42168   34.41447 MR CD
4      1    0  27.49256    36.70679  32.52011   32.52397 CD CD
5      1    1  26.42350    34.84207  33.72922   33.73546 CD CD
```

We see that this dataset contains the baseline features `gender`, `race`, `parentBMI`, and `baselineBMI`. Variables `A1` and `A2` record the treatments assigned at the first and second randomization point, respectively, `month4BMI` is a BMI measurement after the first treatment assignment and before the second treatment assignment, and `month12BMI` is the final BMI measurement at the end of the study period. Our analysis will focus on the percent change in BMI between the baseline BMI measurement and the measurement taken after 12 months from the initial randomization. We add the variable `pctBMIchange` to `bmiData` in the following. Additionally, we negate the percent change in BMI change so that larger percent changes are better.

```
dplyr::mutate(bmiData,
  pctBMIchange = -100 * (month12BMI - baselineBMI)/baselineBMI)
```

To learn an optimal DTR using Q-learning, we proceed using the backward induction. This is a two-stage problem, so we begin by positing a model for the Stage 2 Q-function $Q_2(\boldsymbol{h}_2, a_2)$ given in (18):

$$\begin{aligned} Q_2(\boldsymbol{h}_2, a_2; \boldsymbol{\beta}_2) =& \mathbb{E}(pctBMIchange | \boldsymbol{h}_2, a_2; \boldsymbol{\beta}_2) \\ =& \beta_{2,0} + \beta_{2,1} gender + \beta_{2,2} race + \beta_{2,3} parentBMI + \beta_{2,4} baselineBMI \\ &+ \beta_{2,5} A1 + \beta_{2,6} month4BMI + \beta_{2,7} A2 + \beta_{2,8} A2 \times baselineBMI \\ &+ \beta_{2,9} A2 \times month4BMI + \beta_{2,10} A2 \times parentBMI + \beta_{2,11} A2 \times A1. \end{aligned} \tag{19}$$

The model specified in (19) is a linear model. As specified, `gender`, `race`, `parentBMI`, `baselineBMI`, first-stage treatment (`A1`), and `month4BMI` are specified as treatment effect modifiers; that is, they affect the conditional mean of the percent change in BMI. Additionally, `baselineBMI`, `month4BMI`, `parentBMI`, and `A1` are specified as prescriptive variables meaning that the derived decision rule will recommend treatment based on those four variables. In particular, the form of the decision rules under consideration is

$$I(\beta_{2,7} + \beta_{2,8} baselineBMI + \beta_{2,9} month4BMI + \beta_{2,10} parentBMI + \beta_{2,11} A1 > 0),$$

where $\beta_{2,7}, \ldots, \beta_{2,11}$ are real-valued parameters, and if the treatment rule is valued as 1, then meal replacement therapy is recommended at Stage 2. In R, we use the `lm` function to estimate the parameters in (19).

```
# Fit the second stage regression model (Q2)
stage2mod <- lm(pctBMIchange ~ gender + race + parentBMI +
                baselineBMI + A1 + month4BMI + A2 +
                A2:baselineBMI + A2:month4BMI + A2:parentBMI +
                A2:A1, data = bmiData)
```

Using the Stage 2 model, we can use the predicted values to estimate the Stage 2 optimal treatment rule $\widehat{d}_2^{\text{opt}}(\boldsymbol{h}_2)$. In particular, we set `A2` to "MR" for all study participants and compute the predicted values for this modified dataset, and then we repeat the procedure by setting `A2` to "CD". Next, we compare the predicted values. For those whose predicted response is greater when the Stage 2 treatment is

set to "MR" than to "CD", the optimal Stage 2 treatment recommendation is "MR". Similarly for those whose predicted response is greater when the Stage 2 treatment is set to "CD", the optimal Stage 2 treatment is "CD".

```
# Fitted values when A2 == "MR"
stage2MRfitted <- predict(stage2mod,
                          data = dplyr::mutate(bmiData, A2 = "MR"))

# Fitted values when A2 == "CD"
stage2CDfitted <- predict(stage2mod,
                          data = dplyr::mutate(bmiData, A2 = "CD"))

# Optimal Stage 2 treatment recommendations
d2opt <- dplyr::if_else(stage2MRfitted > stage2CDfitted,
                        "MR", "CD")

# Allocation of Stage 2 treatments under the optimal rule
table(d2opt)
```

```
CD  MR
109 101
```

Thus, at the second stage decision, the optimal rule recommends meal replacement for 101 study participants and conventional diet for 109 study participants. Using the estimated coefficients for the Stage 2 Q-model, we can write down the estimated stage-decision rule as

$$\widehat{d_2}(\boldsymbol{h}_2;\widehat{\boldsymbol{\beta}}_2) = I(-0.68 - 0.13 \times baselineBMI + 0.13 \times month4BMI \\ + 0.04 \times parentBMI - 0.30 \times I(A1 = \text{“}MR\text{”}) > 0),$$

where $\widehat{d_2}(\boldsymbol{h}_2;\widehat{\boldsymbol{\beta}}_2) = 1$ recommends treatment with meal replacement at stage 2. With the Stage 2 decision rule now learned, we can now work backwards to learn the optimal decision rule for the Stage 1 decision. As with second-stage analysis, we posit a model for the Stage 2 Q-function $Q_1(\boldsymbol{h}_1, a_1)$ given in (18). One notable difference between the Stage 2 and Stage 1 Q-functions is that, for the Stage 2 Q-function, we modeled the conditional mean of the response variable, whereas, for the Stage 1 Q-function, we will model the conditional mean of the predicted value estimated from the model posited for Q_2 assuming that optimal treatment was assigned in Stage 2. Denoting this fitted value as $\widehat{Y}_1$, we posit the following model:

$$\begin{aligned} Q_1(\boldsymbol{h}_1, a_1; \boldsymbol{\beta}_1) =& \mathbb{E}[\widehat{Y}_1 | \boldsymbol{h}_1, a_1; \boldsymbol{\beta}_1] \\ =& \beta_{1,0} gender + \beta_{1,1} race + \beta_{1,2} parentBMI + \beta_{1,3} baselineBMI \\ &+ \beta_{1,4} A1 + \beta_{1,5} A1 \times baselineBMI + \beta_{1,6} A1 \times parentBMI. \end{aligned} \quad (20)$$

For simplicity, we have again chosen a linear Q model though more diverse models could be considered. Gender, race, parental BMI, and baseline BMI are specified as modifiers of the first-stage treatment. By interacting baseline BMI and parental BMI with the first-stage treatment, they are the variables by which the first-stage

treatment rule will tailor treatments. The form of the decision rules under consideration is $I(\beta_{1,4} + \beta_{1,5} baselineBMI + \beta_{1,6} parentBMI)$, where $\beta_{1,4}$, $\beta_{1,5}$, and $\beta_{1,6}$ are real-valued parameters, and if the value of the function is 1, then the recommended treatment is meal replacement. As before, we use the `lm` function in R to estimate the parameters in (20).

```
# Compute yhat
yhat <- dplyr::if_else(stage2MRfitted > stage2CDfitted,
                       stage2MRfitted, stage2CDfitted)

# Append yhat to bmiData
bmiData <- cbind(bmiData, yhat)

# Fit the first stage regression model (Q1)
stage1mod <- lm(yhat ~ gender + race + parentBMI + baselineBMI +
                A1 + A1:baselineBMI + A1:parentBMI,
                data = bmiData)
```

As we did with the Stage 2 analysis, we can use the predicted values from the Stage 1 estimated Q-function to estimate the Stage 2 optimal treatment rule $\widehat{d}_1^{\text{opt}}(\boldsymbol{h}_1)$. As before, we compare the predicted values generated by the estimated Stage 1 model with Stage 1 treatment set to "MR" to Stage 1 treatment set to "CD".

```
# Fitted values when A1 == "MR"
stage1MRfitted <- predict(stage1mod,
                          data = dplyr::mutate(bmiData, A1 = "MR"))

# Fitted values when A1 == "CD"
stage1CDfitted <- predict(stage1mod,
                          data = dplyr::mutate(bmiData, A1 = "CD"))

# Optimal Stage 1 treatment recommendations
d1opt <- dplyr::if_else(stage1MRfitted > stage1CDfitted,
                        "MR", "CD")

# Allocation of Stage 1 treatments under the optimal rule
table(d1opt)
```

```
  CD  MR
 109 101
```

Coincidentally, the same number of participants (but not necessarily the same participants) are recommended to receive meal replacement at the first stage as the second stage under the optimal dynamic treatment regime. Using the regression coefficient estimates, we can write down the estimated optimal Stage 1 decision rule

$$\widehat{d}_1^{\text{opt}}(\boldsymbol{h}_1; \boldsymbol{\beta}_1) = I(11.11 + 0.74 \times baselineBMI - 1.23 \times parentBMI),$$

where $\widehat{d}_1^{\text{opt}}(\boldsymbol{h}_1; \boldsymbol{\beta}_1) = 1$ recommends meal replacement and $\widehat{d}_1^{\text{opt}}(\boldsymbol{h}_1; \boldsymbol{\beta}_1) = 0$ recommends conventional diet. Together, $\widehat{d}_1^{\text{opt}}$ and $\widehat{d}_2^{\text{opt}}$ form the estimated optimal dynamic treatment regime.

We have illustrated the two-stage treatment regime using basic R functions (and a few functions from `dplyr` just to keep things tidy), but the `DynTxRegime` package offers functions that automate the steps we have outlined, as well as automatically calculate the value of the estimated optimal DTRs on the training data and test data, as shown below. This is important as it allows for the assessment of the generalizability of the estimated DTRs, and whether it can effectively predict outcomes for new patient populations.

```
# Q-learning to learn the optimal dynamic treatment regime
# - using DynTxRegime package
y12 <- -100*(bmiData[,6L] - bmiData[,4L])/bmiData[,4L]
moMain <- buildModelObj(model = ~parentBMI+month4BMI +gender +
                race + baselineBMI +A1, solver.method = 'lm')
moCont <- buildModelObj(model = ~baselineBMI + parentBMI +
                month4BMI + A1, solver.method = 'lm')
fitSS <- qLearn(moMain = moMain, moCont = moCont,
                data = bmiData, response = y12,  txName = 'A2')
```

```
  First step of the Q-Learning Algorithm.

  Outcome regression.
  Combined outcome regression model: ~ parentBMI + month4BMI +
      gender + race + baselineBMI + A1 + A2 +
      A2:(baselineBMI + parentBMI + month4BMI + A1).
  Regression analysis for Combined:

  Call:
  lm(formula = YinternalY ~ parentBMI + month4BMI +
      gender + race + baselineBMI + A1 + A2 +
      baselineBMI:A2 + parentBMI:A2 + month4BMI:A2 +
      A1:A2, data = data)

  Coefficients:
       (Intercept)       parentBMI       month4BMI          gender
           7.94094        -0.04681         -2.66904        -0.10585
              race     baselineBMI             A1MR           A2MR
           0.19576         2.49094          0.16042        -0.67686
  baselineBMI:A2MR  parentBMI:A2MR  month4BMI:A2MR     A1MR:A2MR
          -0.13449         0.04189          0.12747        -0.30170


  Recommended Treatments:
   CD  MR
  109 101

  Estimated value: 6.682276
```

```
moMain <- buildModelObj(model = ~parentBMI + baselineBMI +
                race + gender, solver.method = 'lm')
moCont <- buildModelObj(model = ~parentBMI+baselineBMI,
                solver.method = 'lm')
fitFS <- qLearn(moMain = moMain, moCont = moCont,
                data = bmiData, response = fitSS,  txName = 'A1')
```

```
Step 2 of the Q-Learning Algorithm.

Outcome regression.
Combined outcome regression model: ~ parentBMI + baselineBMI +
    race + gender + A1 + A1:(parentBMI + baselineBMI).
Regression analysis for Combined:

Call:
lm(formula = YinternalY ~ parentBMI + baselineBMI + race +
    gender + A1 + parentBMI:A1 + baselineBMI:A1, data = data)

Coefficients:
  (Intercept)     parentBMI       baselineBMI              race
      -2.66451      -0.06937           0.34328          -0.46095
       gender          A1MR    parentBMI:A1MR  baselineBMI:A1MR
       0.26356      11.10540          -1.22588           0.73594


Recommended Treatments:
 CD  MR
109 101

Estimated value: 9.416704
```

The R code above is available for access in the specified GitHub repository: https://github.com/FIRST-Data-Lab/Q-learning.

7 Summary/Conclusions

In this chapter, we have described how Q-learning is used in precision medicine to estimate optimal dynamic treatment regimes in both finite-horizon and infinite-horizon settings. Given the breadth of the precision medicine literature, there are a number of Q-learning extensions that were not included. For example, we have mainly focused on the case where $Y \in \mathbb{R}$, but statistical Q-learning strategies have been developed for more complicated outcomes such as right-censored data [11].

Q-learning is an attractive method for learning optimal DTRs since in many cases Q-functions can be modeled using off-the-shelf software routines. Nonetheless, the method depends on the model of the Q-function, so it is important to have high-quality estimators for the Q-function. Despite these challenges, there is still much exciting work to be done with Q-learning in the precision medicine setting. Inference for DTRs and the value function continues to be an active area of research, and methods for infinite-horizon medical decision-making are being developed for mHealth applications. Overall, Q-learning is an effective strategy for learning optimal treatment recommendations and is likely to be an important strategy in the future.

Acknowledgements We express our sincere gratitude to the reviewers and the editors for their insightful comments and suggestions that greatly improved the quality of this chapter. Research reported in this publication was partially supported by the National Institute Of General Medical Sciences of the National Institutes of Health under Award Number P20GM139769 (Xinyi Li), National Science Foundation awards DMS-2210658 (Xinyi Li) and DMS-2203207 (Li Wang). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Ashley, E.: The precision medicine initiative: a new national effort. J. Am. Med. Assoc. **313**, 2119–2120 (2015)
2. Bellman, R.: Dynamic Programming. Princeton Univ. Press, Princeton, NJ (1957)
3. Bertsekas, D.P., Tsitsiklis, J.: Neuro-Dynamic Programming. Athena Sci, Nashua, NH (1996)
4. Cain, L.E., Robins, J.M., Lanoy, E., Logan, R., Costagliola, D., Hernán, M.A.: When to start treatment? A systematic approach to the comparison of dynamic regimes using observational data. Int. J. Biostat. **6**, 18 (2010)
5. Chakraborty, B., Strecher, V., Murphy, S.: Bias correction and confidence intervals for fitted Q-iteration. In Workshop on Model Uncertainty and Risk in Reinforcement Learning (NIPS 2008). https://cs.uwaterloo.ca/∼ppoupart/nips08-workshop/accepted-papers/nips08paper01-final.pdf
6. Clifton, J., Laber, E.B.: Q-learning: theory and applications. Annu. Rev. Stat. Appl. **7**, 279–301 (2020)
7. Collins, F., Varmus, H.: A new initiative on precision medicine. N Engl J Med. **372**, 793–795 (2015)
8. Collins, L.M., Murphy, S.A., Bierman., K.: A conceptual framework for adaptive preventive interventions. Prev. Sci. **5**, 185–196 (2004)
9. Ernst, D., Geurts, P., Wehenkel, L.: Tree-based batch mode reinforcement learning. J. Mach. Learn. Res. **6**, 503–556 (2005)
10. Ertefaie, A., Strawderman, R.L.:Constructing dynamic treatment regimes over indefinite time horizons. Biometrika. **105**, 963–977 (2018)
11. Goldberg, Y., Kosorok, M. R.: Q-learning with censored data. Ann. Statist. **40**, 529–560 (2012)
12. Hernán, M.A., Robins, J.M.: Causal Inference: What if. Chapman & Hall/CRC, Boca Raton (2020)
13. Holloway, S.T., Laber, E.B., Linn, K.A, Zhang, B., Davidian, M., Tsiatis, A.A.: DynTxRegime: Methods for Estimating Optimal Dynamic Treatment Regimes. (2020) https://CRAN.R-project.org/package=DynTxRegime
14. Kidwell, K.M.: DTRs and SMARTs: definitions, designs, and applications. See Kosorok & Moodie 2016, pp. 7–24 (2016)
15. Kosorok, M.R., Laber, E.B.: Precision medicine. Annu. Rev. Stat. Appl. **6**, 263–286 (2019)
16. Kosorok, M.R., Moodie, E.E.M.: Adaptive Treatment Strategies in Practice: Planning Trials and Analyzing Data for Personalized Medicine. SIAM, Philadelphia (2016)
17. Lavori, P.W., Dawson, R., Rush, A.J.: Flexible treatment strategies in chronic disease: Clinical and research implications. Biol. Psychiatry. **48**, 605–614 (2000)
18. Liao, P., Klasnja, P., Murphy, S.: Off-policy estimation of long-term average outcomes with applications to mobile health. J. Am. Stat. Assoc. **116**, 382–391 (2021)
19. Luckett, D.J., Laber, E.B., Kahkoska, A.R., Maahs, D.M., Mayer-Davis, E., Kosorok, M.R.: Estimating dynamic treatment regimes in mobile health using V-learning. J. Am. Stat. Assoc. **115**, 692–706 (2020)
20. Lunceford, J.K., Davidian, M., Tsiatis, A.A.: Estimation of survival distributions of treatment policies in two-stage randomization designs in clinical trials. Biometrics, **58**, 48–57 (2002)

21. Murphy, S.A.: Optimal dynamic treatment regimes (with Discussion). J. R. Statist. Soc. B. **66**, 331–366 (2003)
22. Murphy, S.A.: A generalization error for Q-learning. J. Mach. Learn. Res. **6**, 1073–1097 (2005a)
23. Murphy, S.A.: An experimental design for the development of adaptive treatment strategies. Stat. Med. **24**, 1455–1481 (2005b)
24. Nahum-Shani, I., Qian, M., Almiral, D., Pelham, W., Gnagy, B., Fabiano, G., Waxmonsky, J., Yu, J., Murphy., S.A.: Experimental design and primary data analysis methods for comparing adaptive interventions. Psychol. Methods. **17**, 457–477 (2012)
25. Neyman, J.: On the application of probability theory to agricultural experiments. Essay in principles. Section 9 (translation published in 1990). Statist. Sci. **5**, 472–480 (1923)
26. Pearl, J.: Causal inference in statistics: an overview. Stat. Surv. **3**, 96–146 (2009)
27. Pelham, W.E., Fabiano, G.A.: Evidence-based psychosocial treatments for attention-deficit/hyperactivity disorder. J. Clin. Child. Adolesc. Psychol. **37**, 184–214, (2008).
28. Petersen, M.L., Deeks, S.G., van der Laan, M.J.: Individualized treatment rules: Generating candidate clinical trials. Stat. Med. **26**, 4578–4601 (2007)
29. Powell, W.B.: Approximate Dynamic Programming: Solving the Curses of Dimensionality. Wiley, New York (2007)
30. Robins, J.M.: A new approach to causal inference in mortality studies with a sustained exposure period to control of the healthy worker survivor effect. Math. Model. **7**, 1393–1512 (1986)
31. Robins, J.M.: Addendum to "a new approach to causal inference in mortality studies with a sustained exposure period to control of the healthy worker survivor effect". Comput. Math. with Appl. **14**, 923-945 (1987)
32. Robins, J.M.: Optimal structural nested models for optimal sequential decisions. In Lin, D.Y., Heagerty, P.J., editors, Proceedings of the Second Seattle Symposium in Biostatistics. pp. 189–326. Springer, New York: (2004)
33. Robins, J. M: Causal inference from complex longitudinal data. See M. Berkane (Ed.), Latent variable modeling and applications to causality: Lecture notes in statistics (pp. 69–117). New York: Springer (1997)
34. Rubin, D.B.: Estimating causal effects of treatments in randomized and nonrandomized studies. J. Educ. Psychol. **66**, 688–701 (1974)
35. Rubin, D.B.: Bayesian inference for causal effects: The role of randomization. Ann. Statist. **6**, 34–58 (1978)
36. Schulte, P.J., Tsiatis, A.A., Laber, E.B., Davidian, M.: Q- and A-learning methods for estimating optimal dynamic treatment regimes. Statist. Sci. **29**, 640–661 (2014)
37. Stephens, D.A.: G-estimation for dynamic treatment regimes in the longitudinal setting. See Kosorok & Moodie 2016, pp. 89–117 (2016)
38. Sutton, R.S., Barto, A.G.: Reinforcement learning: an introduction. MIT Press (2018)
39. Thall, P.F., Wooten, L.H., Logothetis, C.J., Millikan, R.E., Tannir, N.M.: Bayesian and frequentist two-stage treatment strategies based on sequential failure times subject to interval censoring. Stat. Med. **26**, 4687–4702 (2007)
40. Tsiatis, A. A., Davidian, M., Holloway, S.T., Laber, E.B.: Dynamic Treatment Regimes: Statistical Methods for Precision Medicine. CRC press (2019)
41. van der Laan, M.J., Petersen, M.L.: Causal effect models for realistic individualized treatment and intention to treat rules. Int. J. Biostat. Article 3, vol. 3(1), 2007.
42. Wahed, A.S., Tsiatis, A.A.: Optimal estimator for the survival distribution and related quantities for treatment policies in two-stage randomized designs in clinical trials. Biometrics. **60**, 124–133 (2004)
43. Zhang, B., Tsiatis, A.A., Laber, E.B., Davidian, M.: A robust method for estimating optimal treatment regimes. Biometrics. **68**, 1010–1018 (2012)
44. Zhang, B., Tsiatis, A.A., Laber, E.B., Davidian, M.: Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. Biometrika. **100**, 681–694 (2013)
45. Zhao, Y.Q., Zeng, D., Laber, E.B., Kosorok, M.R.: New statistical learning methods for estimating optimal dynamic treatment regimes. J. Am. Stat. Assoc. **110**, 583–598 (2015)

46. Zhao, Y.Q., Zeng, D., Rush, A.J., Kosorok, M.R.: Estimating individualized treatment rules using outcome weighted learning. J. Am. Stat. Assoc. **107**, 1106–1118 (2012)
47. Zhao, Y.Q., Zeng, D., Socinski, M.A., Kosorok, M.R.: Reinforcement learning strategies for clinical trials in non-small cell lung cancer. Biometrics. **67**, 1422–1433 (2011)