Published in final edited form as:

Stat Med. 2020 April 30; 39(9): 1250–1263. doi:10.1002/sim.8473.

Constructing Dynamic Treatment Regimes with Shared Parameters for Censored Data

Ying-Qi Zhao^{*,a}, Ruqing Zhu^b, Guanhua Chen^c, Yingye Zheng^a

^aPublic Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, U.S.A.

^bDepartment of Statistics, University of Illinois Urbana-Champaign, Champaign, Illinois, 61820, U.S.A.

^cDepartment of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, Wisconsin, 53792, U.S.A.

Abstract

Dynamic treatment regimes are sequential decision rules that adapt throughout disease progression according to a patient's evolving characteristics. In many clinical applications, it is desirable that the format of the decision rules remains consistent over time. Unlike the estimation of dynamic treatment regimes in regular settings, where decision rules are formed without shared parameters, the derivation of the shared decision rules requires estimating shared parameters indexing the decision rules across different decision points. Estimation of such rules becomes more complicated when the clinical outcome of interest is a survival time subject to censoring. To address these challenges, we propose two novel methods: censored shared-Q-learning and censored shared-O-learning. Both methods incorporate clinical preferences into a qualitative rule, where the parameters indexing the decision rules are shared across different decision points and estimated simultaneously. We use simulation studies to demonstrate the superior performance of the proposed methods. The methods are further applied to the Framingham Heart Study to derive treatment rules for cardiovascular disease.

Keywords

Dynamic treatment regimes; Shared parameters; Q-learning; O-learning; Censored data

1. Introduction

Dynamic treatment regimes (DTRs), also called adaptive treatment strategies (1; 2), are sequential decision rules adapting over time to the time-varying characteristics of patients. A DTR takes patient health histories as inputs and recommends the next treatment strategy at each decision point. For example, multiple lines of treatment regimens are considered for lung cancer, each based on the current health status (3). Clinicians may update treatment for

*Correspondence to: Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109. yqzhao@fredhutch.org.

major depressive disorder according to factors emerging over time, such as side-effect severity, treatment adherence, and so on (4). In these examples, the decision rules change across different decision points. Estimating an optimal DTR with time-varying parameters has been widely studied in the past few years. A well-established approach is Q-learning (5; 6; 7; 8; 9), which recursively estimates the conditional expectation of the outcomes given the current patient history, assuming that optimal decisions are made in the future. The foregoing conditional expectations are known as Q-functions. Semiparametric methods have been proposed, such as iterative minimization of regrets (1) and G-estimation (10). These methods are methodologically more complex, but could potentially provide efficiency gain. Recently, direct methods have become popular in the literature. They circumvent the modeling of conditional means of the outcome given the treatment and covariates, by directly estimating the decision rule that maximizes the expected outcomes. Examples include backward outcome weighted learning, simultaneous outcome weighted learning (11), and others (12; 13; 14). We refer readers to Chakraborty and Moodie (2013) (15), Kosorok and Moodie (2015) (16) and Kosorok and Laber (2019) (17) for detailed reviews of the current literature.

In a regular DTRs setting, decision rules are formed without shared parameters. Yet in practice, it is not unusual to have a common decision rule shared across different decision points. For example, diabetes patients are recommended medication or lifestyle intervention when Hemoglobin A1c (HbA1c) rises above a threshold, which is universal across the entire medication process. Lung transplantation may be initiated for a cystic fibrosis patient if FEV₁% (Forced Expiratory Volume in 1 second) value falls below 30%, which, again, remains the same throughout the course of disease progression. Such shared decision rules are easier to implement in practice throughout multiple decision points, in particular, when multivariate time-varying covariates are potentially involved. Unlike the DTRs in a regular setting, where there are no shared parameters among decision rules at different time points, these shared decision rules have shared parameters indexing the decision rules across decision points. They have been known as DTRs with shared parameters in the literature (18).

DTRs with shared parameters are analogous to stationary Markov decision processes with function approximations (19). Note that while the optimal DTRs need not be stationary in a finite horizon problem, it is still desirable to derive the best stationary DTRs in practice. Furthermore, in some observational studies, the times when patients are enrolled could vary, and there is no obvious time point that can be treated as a baseline. A rule with different formulations at different time points would be difficult to generalize. Our focus in this paper is to construct a DTR with shared parameters in a finite horizon problem. For solving Markov decision problems, there have been proposals to minimize the squared Bellman error, where Bellman error quantifies the difference between the estimated reward at any time point and the actual reward received (20). Chakraborty et al (2016) (18) proposed shared Q-learning to estimate the optimal shared-parameter DTR when the decision rule at each decision point is the same function of one or more time-varying covariates. In particular, they formulated the decision rules as linear functions of covariates, and the coefficients are assumed to be the same across decision points. An alternative way of identifying a DTR with shared parameters is the simultaneous G-estimation (10), which can

handle problems with shared parameters in principle. However its empirical performance is largely unknown (21). The simultaneous outcome weighted learning approach proposed in Zhao et al (2015) (11) converted the construction of DTRs into a simultaneous nonparametric classification problem, where a multi-dimensional hinge loss of the expected long-term outcome was employed. However, the method needs to be modified for DTRs with shared parameters.

Moreover, the aforementioned existing methods for shared-parameter DTRs have been limited to settings with continuous outcomes. Extension to censored failure time outcome is challenging. First, the number of decision points for each individual in the study is not fixed. This is because the event time varies by individual, and the treatment is usually stopped once the failure event happens. Second, the treatment/outcome status of a subject may be unknown when censoring occurs. In regular DTR settings, Goldberg and Kosorok (2012) (9) developed a Q-learning algorithm to adjust for censored data and allow a flexible number of decision points. Recently, Hager et al (2018) and Simoneau et al (2019) proposed methods for estimating optimal DTRs with survival endpoints subject to right-censoring, which provide insight to efficiency and double robustness properties in this setting (22; 23). Further development is needed for DTRs with shared parameters.

In this paper, we propose two methods for solving DTRs with shared parameters in the presence of censored data, termed as censored shared-Q-learning and censored shared-O-learning (abbreviated from “outcome weighted learning”). The censored shared-Q-learning method generalizes the shared-Q-learning method (18) by applying the inverse probability of censoring weights to account for the uncertainty in the outcomes of censored subjects, where the censoring weights need to be carefully constructed for each decision point due to the multistage nature of the problem. Similar to shared Q-learning, censored shared-Q-learning is an iterative procedure, which identifies the optimal decision rule for each decision point in a sequential/iterative manner. The censored shared-O-learning method uses a non-iterative approach by directly maximizing a concave relaxation of the inverse-probability-of-censoring weighted estimator of the expected survival benefit. To the best of our knowledge, this is the first article that provides a thorough solution for deriving DTRs with shared parameters in the censored data setup.

The remainder of the paper is organized as follows. We introduce the general framework of DTRs with shared parameters in Section 2. In Section 3, we introduce censored shared-Q-learning and censored shared-O-learning along with their computation algorithms. We describe numerical studies that compare the proposed methods with standard Q-learning and shared Q-learning in Section 4. Section 5 focuses on the application of the proposed methods to the Framingham Heart data. Finally, we provide a discussion of open questions in Section 6.

2. Statistical framework

In this section, we present definitions and notations used in the paper. Throughout, we use uppercase letters to denote random variables and lowercase letters to denote realizations of the random variables.

Let T be the maximal number of clinical decision points in a study. These decision points are considered as stages in a sequential randomization trial, and referred to as critical time points when a treatment decision is to be made in the observational studies. A full trajectory of an observation sequence is $\{X_1, A_1, Y_1, \dots, X_T, A_T, Y_T\}$. For any decision point $j = 1, \dots, T$, the information of a patient is represented by (X_j, A_j, Y_j) , where X_1 denotes the initial information, X_j denotes the intermediate information collected between decision points $j - 1$ and j when $j > 0$, and A_j is the treatment assigned at the j^{th} decision point subsequent to observing X_j . We assume that there are two possible treatments at each decision point, i.e., $A_j \in \mathcal{A}_j = \{-1, 1\}$. Y_j is a non-negative random variable that equals either the length of the interval between decision time point j and $j + 1$, or the length between decision time point j and failure time if the failure event occurs in that decision interval. Y_j can be viewed as the reward in the j^{th} interval, and $\sum_{k=1}^j Y_k$ is the cumulative reward up to and including decision interval j . In our context, the sum $\sum_{k=1}^j Y_k$ is the total survival time up to and including decision interval j . As the total survival times are broken down based on the number of decision points, this setup could introduce complexity to the trajectory structure. In particular, if a failure event occurs before the final decision point T , the trajectory will not be of full length (9). Subsequently, the number of decision points for different observations could be different. We denote the (random) total number of decision points for an individual by $\bar{T}$. Thus, $\sum_{j=1}^{\bar{T}} Y_j$ is the overall survival time.

In this study, the time to event is subject to censoring. Let C denote the censoring time, taking values in the segment $[0, \tau]$. We assume that the censoring and failure times are independent given the covariates of all previous decision intervals. Let Δ_j be the censoring indicator at decision point j , where $\Delta_j = I(\sum_{k=1}^j Y_k \leq C)$. If no censoring event happens before the $(j + 1)^{\text{th}}$ decision time point, then $\Delta_j = 1$ and the outcome Y_j is observed. Hence, if a censoring event occurs during decision interval $\bar{T}$, then $C \leq \sum_{j=1}^{\bar{T}} Y_j$, $\Delta_{\bar{T}-1} = 1$ and $\Delta_{\bar{T}} = 0$. In this case, we cannot observe the failure time that is censored by C . On the other hand, $\Delta_{\bar{T}} = 1$ if the failure time $\sum_{j=1}^{\bar{T}} Y_j$ is observed. While $\Delta_j = 1$ indicates that censoring has not occurred at decision point j , it does not necessarily mean $\sum_{k=1}^j Y_k$ is the time of event, which is different from the notation of a typical survival analysis. Only $\Delta_{\bar{T}}$ is the true failure time indicator.

We use an overbar to denote the collection of history information, e.g. the sequence of treatments $(A_1, A_2, \dots, A_j)$ is represented by $\bar{A}_j$, and $\bar{X}_j$ is the sequence of covariates up to j . We impose Markov assumption that the optimal treatment at decision point j depends only on the collected information from $j - k$ to j , and we use $H_j \in \mathcal{H}_j$ to denote the accrued information. Hence, the dimension of vector of tailoring covariates does not change over time. A DTR is a sequence of deterministic decision rules, $d = (d_1, \dots, d_T)$, where $d_j: \mathcal{H}_j \mapsto \mathcal{A}_j$ is a function mapping from the space of information at decision point j to the space of available treatments. Under d , a patient presenting with $H_j = h_j$ at time j is recommended to treatment $d_j(h_j)$. We use $\mathcal{D}_T$ to denote the class of all possible treatment

regimes. Our goal is to identify the optimal dynamic treatment regime $d^* \in \mathcal{D}_T$, that maximizes the expected outcome if deployed to the whole population in the future. Given that no information on survival is available beyond τ in the observed data, the outcome of interest is truncated-by- τ expected survival. Let P^d denote the distribution of a trajectory given that $A_j = d_j(H_j)$, $j = 1, \dots, \bar{T}$. The optimal DTR is the sequence of rules that maximizes

$$E^d \left\{ \min \left(\sum_{j=1}^{\bar{T}} Y_j, \tau \right) \right\}, \quad (1)$$

where E^d is the expectation with respect to P^d . We refer to (1) as the value associated with a regime d , denoted by $\mathcal{V}(d)$.

Following Goldberg and Kosorok (2012) (9), we modify the trajectories when they are not of full length or the overall survival time is greater than τ . The idea is to extend the information to full length by introducing non-informative values at decision intervals after the failure. If a failure time occurs at decision interval $j < T$, we let $X'_k = X_k$ for $1 \leq k \leq j$ and let the non-informative $X'_k = \emptyset$, for $j+1 \leq k \leq T$. Similarly, we let $A'_k = A_k$ and $Y'_k = Y_k$, for $1 \leq k \leq j$. And for $j+1 \leq k \leq T$, we set $Y'_k = 0$ and draw A'_k uniformly from $\mathcal{A}$ as the non-informative values. If for some j , $\sum_{k=1}^j Y_k \geq \tau$, then we modify Y'_j to be $Y'_j = \max(\tau - \sum_{k=1}^{j-1} Y_k, 0)$, and modify the trajectory to be non-informative at all decision intervals after j . For any DTR $d \in \mathcal{D}_T$, we define a corresponding d' for the modified trajectories, where the same action is chosen for any triplet $(x'_j, a'_{j-1}, y'_{j-1})$ if $x'_j \neq \emptyset$, and a fixed action is chosen if $x'_j = \emptyset$. It has been shown in Goldberg and Kosorok (2012) (9) that

$$E^{d'} \left(\sum_{j=1}^T Y'_j | X_1 = x_1 \right) = E^d \left\{ \min \left(\sum_{j=1}^{\bar{T}} Y_j, \tau \right) | X_1 = x_1 \right\}. \quad (2)$$

Subsequently, $\mathcal{V}(d)$ remains the same after this modification. In the following, we omit the “prime” in the modified trajectories without the risk of ambiguity. Another complication is due to censoring, where the trajectories themselves may be censored and cannot be fully observed. This needs to be carefully handled when estimating DTRs with shared parameters from the data.

3. Estimating DTRs with shared parameters for censored data

We often encounter time-varying covariates in longitudinal studies. To facilitate clinical implementation, it could be beneficial to have a shared decision rule in its functional form across multiple decision points, while allowing the covariate values to change over time. In the following, we propose two methods to estimate the DTRs with shared parameters using repeatedly measured covariate information when the outcome is subject to censoring. The proposed algorithms are based on Q-learning and O-learning in the multistage setup, respectively.

3.1. Censored shared-Q-learning

Due to delayed effects, we need to consider the entire treatment sequence in order to optimize the long-term outcome. Results from the dynamic programming literature show that $d_T^*(h_T) \arg\max_{a_T} Q_T(h_T, a_T)$, where $Q_T(h_T, a_T) = E(Y_T | H_T, A_T = a_T)$ and recursively $d_j^*(h_j) = \arg\max_{a_j} Q_j(h_j, a_j)$ where $Q_j(h_j, a_j) = E(Y_j + \max_{a_{j+1}} Q_{j+1}(H_{j+1}, a_{j+1}) | H_j = h_j, A_j = a_j)$ for $j = T-1, \dots, 1$ when the underlying generative distribution is known (24; 25). Q-learning is an approximate dynamic programming algorithm that uses regression models to estimate the Q-functions $Q_j(h_j, a_j)$, $j = 1, \dots, T$ and then estimate $d_j^*(h_j)$ recursively. Note that if a failure occurred prior to the j^{th} time point, i.e., $X_j = \emptyset$, we set Q_j as zero. A commonly used strategy is to estimate Q-functions via linear regression using working models $Q_j(H_j, A_j; \beta_j, \psi_j) = \beta_j^\top H_{j0} + (\psi_j^\top H_{j1}) A_j$, where H_{j0} and H_{j1} (including intercepts) are possibly different features of H_j .

We focus on the setting where the decision rule parameters are shared across decision points with $\psi_1 = \dots = \psi_T = \psi$, but β_j s are unshared. This means the predictive effects for taking treatments remain the same throughout, regardless of their covariates' prognostic effects on patient health status, whereas prognostic effects unaltered by treatment are allowed to be different across decision points. Let $\theta_j = (\beta_j^\top, \psi^\top)^\top$. Assuming the shared model is correctly specified, the pseudo-outcome at the j^{th} decision point is

$\tilde{Y}_j(\theta_{j+1}) = Y_j + \beta_{j+1}^\top H_{j+1,0} + |\psi^\top H_{j+1,1}|$, and $Q_j(H_j, A_j) = E\{\tilde{Y}_j(\theta_{j+1}) | H_j, A_j\}$, $j = 1, \dots, T-1$. Let $\tilde{Y}_T(\theta_{T+1}) = Y_T$. For any j , we can write $Q_j(H_j, A_j) = \beta_j^\top H_{j0} + (\psi^\top H_{j1}) A_j$. Hence, $(\hat{\beta}_1, \dots, \hat{\beta}_T, \hat{\psi})$ can be estimated (18) via

$$\arg\min_{\beta_1, \dots, \beta_T, \psi} \sum_{j=1}^T \mathbb{P}_n [\tilde{Y}_j(\theta_{j+1}) - \{\beta_j^\top H_{j0} + (\psi^\top H_{j1}) A_j\}]^2, \quad (3)$$

where $\mathbb{P}_n$ denotes the empirical averages of the data. However, Y_j in the $\tilde{Y}_j(\theta_{j+1})$ may be censored and unknown in the censoring data setup, and $\tilde{Y}_j(\theta_{j+1})$ are defined by unknown parameters.

Let $S_C(\sum_{k=1}^j Y_k | H_j, A_j) = P(C > \sum_{k=1}^j Y_k | H_j, A_j)$ be the conditional survival function for the censoring time given history information up to decision point j and treatment received at that decision point. In addition, we assume the conditional independent of censoring (i.e. $\sum_{k=1}^j Y_k \perp C | H_j, A_j$). Then

$$E \left[\frac{A_j}{S_C(\sum_{k=1}^j Y_k | H_j, A_j)} \middle| H_j, A_j \right] = 1.$$

Let $U_j = 0$ if the censoring occurs in the j^{th} decision interval, and $U_j = Y_j$ otherwise. We further define $\tilde{U}_j(\theta_{j+1}) = U_j + \beta_{j+1}^\top H_{j+1,0} + |\psi^\top H_{j+1,1}|$ for $j = T-1$ and $\tilde{U}_T(\theta_{T+1}) = U_T$. Consequently,

$$E \left[\tilde{Y}_j(\theta_{j+1}) - \{\beta_j^\top H_{j0} + (\psi^\top H_{j1}) A_j\} \right]^2 \\ = E \left[\left[\tilde{U}_j(\theta_{j+1}) - \{\beta_j^\top H_{j0} + (\psi^\top H_{j1}) A_j\} \right]^2 \frac{\Delta_j}{S_C(\sum_{k=1}^j Y_k | H_j, A_j)} \right].$$

The quantity reflected in the above equation will only involve observed data, thus instead of (3), $(\hat{\beta}_1^\top, \dots, \hat{\beta}_T^\top, \hat{\psi})^\top$ estimated via a weighted least square procedure, with

$$\underset{\beta_1, \dots, \beta_T, \psi}{\operatorname{argmin}} \sum_{j=1}^T \mathbb{P}_n \left(\left[\tilde{U}_j(\theta_{j+1}) - \{\beta_j^\top H_{j0} + (\psi^\top H_{j1}) A_j\} \right]^2 \frac{\Delta_j}{\hat{S}_C(\sum_{k=1}^j Y_k | H_j, A_j)} \right). \quad (4)$$

where $\hat{S}_C$ is an estimator for S_C .

Although the objective function (4) shares some similarity to the one presented in Goldberg and Kosorok (2012) (9), the backwards recursive procedure for the computation therein cannot be applied, due to the presence of shared parameters ψ in (4). We employ an iterative procedure for estimating θ_S (18). The censored shared-Q-learning algorithm is presented below.

1. Estimate the conditional survival function for the censoring time $\hat{S}_C(\sum_{k=1}^j Y_k | H_j, A_j)$ at decision point j , and construct the diagonal matrix

$$V = \begin{pmatrix} V_T & 0 & \dots & 0 \\ 0 & V_{T-1} & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & V_1 \end{pmatrix}$$

where $V_j = \Delta_j / \hat{S}_C(\sum_{k=1}^j Y_k | H_j, A_j)$, $j = 1, \dots, T$.

2. Set the initial value of θ , denoted by $\hat{\theta}^{(0)} = (\hat{\beta}_T^{(0)\top}, \dots, \hat{\beta}_1^{(0)\top}, \hat{\psi}^{(0)\top})^\top$.
3. At the $(l+1)^{th}$ iteration, $l = 0, 1, 2, \dots$:
 - a. Constructing the vector $\tilde{U}(\hat{\theta}^{(l)}) = (\tilde{U}_T(\hat{\theta}_{T+1}^{(l)})^\top, \dots, \tilde{U}_1(\hat{\theta}_2^{(l)})^\top)^\top$.
 - b. Solving for $\hat{\theta}^{(l+1)} = (Z^\top V Z)^{-1} Z^\top V \tilde{U}(\hat{\theta}^{(l)})$, where

$$Z = \begin{pmatrix} H_{T0}^T & 0 & \dots & 0 & H_{T1}^T A_T \\ 0 & H_{T-1,0}^T & \dots & 0 & H_{T-1,1}^T A_{T-1} \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \dots & H_{10}^T & H_{11}^T A_1 \end{pmatrix}.$$

4. Repeat steps (3a)-(3b) until $\|\hat{\theta}^{(l+1)} - \hat{\theta}^{(l)}\| \leq \epsilon$ for a prespecified value ϵ or until the maximum number of iterations is reached.

A simple choice for the initial values of θ is to set all parameters to zero,

$\hat{\theta}^{(0)} = (0^T, \dots, 0^T, 0^T)$. In this paper, we let the initial values depend on the estimates from censored Q-learning (9), where the parameters are not shared. Denote the estimates as $\hat{\beta}_1^{(0)}, \dots, \hat{\beta}_T^{(0)}, \hat{\psi}_T^{(0)}, \dots, \hat{\psi}_1^{(0)}$. We combine the distinct estimates of $\hat{\psi}_j^{(0)}$ into a single estimate via

an average. The initial values of θ can be set as $\hat{\theta}^{(A)} = (\hat{\beta}_T^{(0)T}, \dots, \hat{\beta}_1^{(0)T}, \hat{\psi}^{(A)T})^T$, where

$$\hat{\psi}^{(A)} = \sum_{j=1}^T \hat{\psi}_j^{(0)} / T.$$

3.2. Censored shared-O-learning

The censored shared-Q-learning algorithm requires correct specifications of Q-functions at each decision point. This could be unrealistic in the multistage setup since the underlying data generating mechanism is usually complicated. In this section, we propose censored shared-O-learning, which constructs the shared decision rules by directly targeting the overall benefit of the decision rule.

Define $\pi_j(a_j; h_j) = P(A_j = a_j | H_j = h_j)$ for $j = 1, \dots, T$. We assume the following conditions, including i) A_j is independent of all potential values of the outcome and future variables conditional on H_j , $j = 1, \dots, T$; and (ii) $\pi_j(1; H_j)$ is strictly between 0 and 1. Assumption (i) is true in a sequential multiple assignment randomized trial (26) but unverifiable in an observational study. As shown in Zhao et al (2015) (11),

$$\mathcal{V}(d) = E^d \left(\sum_{j=1}^T Y_j \right) = E \left[\frac{(\sum_{j=1}^T Y_j) \prod_{j=1}^T I\{A_j = d_j(H_j)\}}{\prod_{j=1}^T \pi(A_j; H_j)} \right], \quad (5)$$

where $I(\cdot)$ is the indicator function. The right-hand side indicates $\mathcal{V}(d)$ equals the weighted average of outcomes among those who received T treatments coinciding with that dictated by d , with weights $\{\prod_{j=1}^T \pi_j(A_j; H_j)\}^{-1}$. π_j s are usually known in a sequential multiple assignment randomized trial. If they are unknown, we can estimate π_j s using methods such as a logistic regression. Given the estimates $\hat{\pi}_j(a_j; h_j)$, a plug-in estimator for $\mathcal{V}(d)$ based on (5) is

$$\widehat{\mathcal{V}}(d) = \mathbb{P}_n \left[\frac{\left(\sum_{j=1}^T Y_j \right) \prod_{j=1}^T I\{A_j = d_j(H_j)\}}{\prod_{j=1}^T \hat{\pi}(A_j; H_j)} \right].$$

However, $\sum_{j=1}^T Y_j$ may not be fully observed due to censoring. Let $W_T = \min(\sum_{j=1}^T Y_j, C)$ and $S_C(\sum_{j=1}^T Y_j | H_T, A_T) = P(C > \sum_{j=1}^T Y_j | H_T, A_T)$ be the conditional survival function for the censoring time given history information up to decision point T . Denote the estimator of S_C as $\hat{S}_C$. We can estimate $\mathcal{V}(d)$ in the scenario with time-to-event outcomes using

$$\widehat{\mathcal{V}}(d) = \mathbb{P}_n \left(\frac{W_T \prod_{j=1}^T I\{A_j = d_j(H_j)\}}{\prod_{j=1}^T \hat{\pi}(A_j; H_j)} \cdot \frac{\Delta_T}{\hat{S}_C(W_T | H_T, A_T)} \right). \quad (6)$$

The decision rules are formulated as fixed linear functions of the present variables at each decision point. Mathematically, they are presented as $d_j(H_j) = \text{sgn}(\psi^\top H_{j1})$, where H_{j1} is a subset of H_j important for determining the DTRs with shared parameters similar to the one defined in Section 3.1. In addition, we define $\text{sgn}(0) = 1$. Consequently, we maximize over ψ

$$\widehat{\mathcal{V}}(\psi) = \mathbb{P}_n \left[I \left[\min_{j=1, \dots, T} A_j \psi^\top H_{j1} \geq 0 \right] \frac{W_T \Delta_T}{\prod_{j=1}^T \hat{\pi}(A_j; H_j) \hat{S}_C(W_T | H_T, A_T)} \right].$$

where we use $\widehat{\mathcal{V}}(\psi)$ to denote $\widehat{\mathcal{V}}\{\text{sgn}(\psi^\top H_{j1})\}$, and substitute $I\{A_j = \mathcal{D}_j(H_j)\}$ by $I\{A_j \psi^\top H_{j1} \geq 0\}$.

It could be challenging to optimize $\widehat{\mathcal{V}}(\psi)$ directly due to the discontinuity of the indicator functions. A computationally efficient approach is to replace the indicator by a concave surrogate. This leads to an optimization problem of

$$\max_{(f_1, \dots, f_T)} \mathbb{P}_n \left(\phi \left[\min_{j=1, \dots, T} \{A_j \psi^\top H_{j1}\} \right] \frac{W_T \Delta_T}{\prod_{j=1}^T \hat{\pi}(A_j; H_j) \hat{S}_C(W_T | H_T, A_T)} \right), \quad (7)$$

where ϕ is a concave function. In this paper, we will use $\phi(t) = -\log(1 + e^{-t})$, which is an analog of the logistic loss in the machine learning literature. We will see that later by using logistic loss, the derivatives of the objective function can be easily derived to facilitate computation. However, other choices of $\phi(t)$ are available; for example, analogs of exponential loss, hinge loss and others can also be applied (27). Specifically, if we use hinge loss, the procedure is the equivalent one generalized from the simultaneous outcome weighted learning approach proposed in Zhao et al (2015) (11). The above objective function is not differentiable in ψ . To account for the non-differentiability of the minimum function in (7), we instead consider a soft-minimum function of u and v to replace $\min(u, v)$, which equals to $-\log\{\exp(-uK) + \exp(-vK)\}/K$, with K being a positive constant. Hence,

the term $\min_{j=1, \dots, T} \{A_j \psi^\top H_{j1}\}$, in (7) can be replaced by its soft-minimum counterpart. We maximize

$$\Phi(\psi) = \mathbb{P}_n \left(\frac{W_T A_T}{\prod_{j=1}^T \hat{\pi}(A_j; H_j) \hat{S}_C(W_T | H_T, A_T)} \log \left[1 + K^{-1} \cdot \sum_{j=1}^T \exp \{ -K(A_j \psi^\top H_{j1}) \} \right] \right).$$

The derivative with respect to shared parameters of the decision rules ψ can be written as

$$\frac{\partial \Phi(\psi)}{\partial \psi} = \mathbb{P}_n \left(\frac{W_T A_T}{\prod_{j=1}^T \hat{\pi}(A_j; H_j) \hat{S}_C(W_T | H_T, A_T)} \frac{-\sum_{j=1}^T \exp \{ -K(A_j \psi^\top H_{j1}) \} \cdot (A_j H_{j1})}{1 + K^{-1} \cdot \sum_{j=1}^T \exp \{ -K(A_j \psi^\top H_{j1}) \}} \right).$$

We can employ the orthant-wise limited-memory quasi-Newton algorithm proposed by Andrew and Gao (2007) (28). We also note that it is possible to write out the Hessian matrix for our proposed objective function. However, we find it does not benefit the numerical performance since the calculation of the second derivative is rather complicated and less efficient than a numerical approximation using the Sherman–Morrison updating formula, as implemented in the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm.

Censored shared-O-learning maximizes the estimated mean outcome of a DTR over the prespecified class of DTRs with shared parameters. Hence, compared with censored shared-Q-learning, it circumvents the need for estimating Q-functions at each decision point. Furthermore, it does not require modeling the censoring distribution at each decision point but only needs to model the censoring distribution at the final decision point. However, censored shared-O-learning involves an unknown parameter K , which controls the approximation of the soft minimum function to the minimum function. In practice, we can use cross-validation to select the best K by grid search over a prespecified set of candidate values.

Remark 1 In practice, interpretable and simple rules are preferable. Sparse penalty such as LASSO can be applied in both censored shared-Q- and O-learning, where the coefficients for unimportant variables will shrink to zero. For censored shared-Q-learning, we can solve for $\hat{\theta}^{(l+1)}$ in step 3(b) using penalized weighted least squares. For censored shared-O-learning, we can maximize the penalized objective

$$\Phi(\psi) - \tau_n \|\psi\|_1,$$

where $\|\psi\|_1$ is the L_1 norm of ψ and τ_n is a tuning parameter controlling the amount of penalization. The orthant-wise limited-memory quasi-Newton algorithm, which is a limited-memory BFGS algorithm that incorporates the ℓ_1 regulation, can still be applied in this case.

4. Simulation Studies

One of the motivations for the current work derives from the long-term care of patients with diabetes. Patients are routinely examined for glycosylated hemoglobin (A1c) level every

three months, and treatments are recommended for tightly controlling A1c to prevent adverse events such as hospitalization due to the disease. Our simulation mimics such a setting, using the generative model similar to that of Timbie et al (2010) (29) and Ertefaie and Strawderman (2018) (30). We treat each check-up time as a decision point for determining treatment in the next three months. Our study consists of 10 decision points. The treatments include metformin, sulfonylurea, glitazone, and insulin. Patients start with metformin and augment with treatments sulfonylurea, glitazone, and insulin during the study period. At each decision point, patients can either continue the current treatment or augment the treatment. A binary discontinuation indicator is generated to represent patients' intolerance to treatment due to side effects, and patients who discontinue a treatment will take the next available treatment. N_j is the number of augmented treatments by the end of interval j where $N_j \in \{0, 1, 2, 3, 4\}$, and the number of augmented treatments N increases by one if a treatment is augmented. The outcome of interest is time to hospitalization, and the trajectory for each individual continues until a patient is hospitalized or has gone through 10 decision points. A censoring variable is uniformly drawn from $[0, 25]$. When an event is censored, the trajectory ends up to the time of censoring and the censoring times are given. Here are the steps we take to generate the dataset:

- Baseline variables: Variables (BP_0 , $weight_0$, $A1c_0$) are generated from a multivariate normal distribution with mean $(12, 140, 7.7)$ and the covariance matrix $\text{diag}(1, 1, 1)$, where BP is the systolic blood pressure. Also, $N_0 = L_0 = 0$, where L_j is the discontinuation indicator at decision point j .
- Treatments: Given N_j , the sets of available treatments are $A_{N_j=0} = \{0, \text{metformin}\}$, $A_{N_j=1} = \{0, \text{sulfonylurea}\}$, and $A_{N_j=2} = \{0, \text{glitazone}\}$, $A_{N_j=3} = \{0, \text{insulin}\}$, and $A_{N_j=4} = \{0\}$ where 0 means continue with the current treatment. The treatment is given as follows,
 - if $A1c_j < 7$, continue with the current treatment and $N_j = N_{j-1}$.
 - if $A1c_j > 8$ and $N_{j-1} < 4$, augment the current treatment and $N_j = N_{j-1} + 1$.
 - if $7 < A1c_j < 8$ and $N_{j-1} < 4$, then a binary variable Z_j is generated with probability

$$P(Z_j = 1 | A1c_{j-1}, N_{j-1}, L_{j-1}) = \frac{\exp(-0.2A1c_{j-1} + 0.5N_{j-1} + 0.5L_{j-1})}{1 + \exp(-0.2A1c_{j-1} + 0.5N_{j-1} + 0.5L_{j-1})},$$

where L_j is the discontinuation indicator. If $Z_j = 1$, the patient continues with the current treatment, and we set $A_j = -1$, $N_j = N_{j-1}$. If $Z_j = 0$, the treatment is augmented, and we set $A_j = 1$, $N_j = N_{j-1} + 1$.

- Treatment discontinuation indicator: A binary variable L_j is generated from a Bernoulli distribution given the last augmented treatment. The treatment discontinuation rates are $P(L_j | A_{j-1} = \text{metformin}) = P(L_j | A_{j-1} = \text{sulfonylurea}) =$

$P(L_j/A_{j-1} = \text{glitazone}) = 0.20$, and $P(L_j/A_{j-1} = \text{insulin}) = 0.35$. We assume that $P(L_j = 1 | A_{j-1} = 0) = 0$.

- A1c, BP and weight at time j : we use the following generative model for A1c,

$$A1c_j = \frac{A1c_{j-1} - \mu_{j-1} + \epsilon}{\sqrt{1 + \sigma_\epsilon^2 + \mu_j}},$$

where $s \sim N(0, 0.25)$, $\sigma_\epsilon^2 = 0.25$ and

$$\mu_j = \begin{cases} \mu_{j-1}(1 - \tau_{A_j}) & \text{if } A1c_{j-1} > 7, N_{j-1} < 4, A_j \neq 0 \text{ and } L_j \neq 1, \\ \mu_{j-1} & \text{o.w.} \end{cases}$$

τ_{A_j} is the treatment effect of A_j , where the treatment effects of metformin, sulfonylurea, glitazone, and insulin are 0.14, 0.20, 0.12, and 0.14, respectively. For the other time-varying variables at time j : $BP_j = (BP_{j-1} + \epsilon_1)/\sqrt{1 + \sigma_{\epsilon_1}^2}$ and $weight_j = (weight_{j-1} + \epsilon_2)/\sqrt{1 + \sigma_{\epsilon_2}^2}$, $\epsilon_1, \epsilon_2 \sim N(0, 0.25)$.

- Time to hospitalization: two generative mechanisms are considered. In Scenario 1, the survival time at decision interval j , i.e., time to hospitalization, starting from the beginning of decision interval j , is generated by

$$\log(Y_j) = 2.5 - 0.5 \times |A1c_j + 0.5N_{j-1} - 10| \{I(A1c_j > 0) - I(A1c_j + 0.5N_{j-1} > 10)\}^2 + \epsilon,$$

where ϵ follows a standard normal distribution. In Scenario 2, the survival time at decision interval j is generated by

$$\log(Y_j) = 2.5 - 0.5 \times |A1c_j - 7| \{I(A1c_j > 0) - I(A1c_j + 0.5N_{j-1} > 10)\}^2 + \epsilon.$$

The regret at each decision interval, i.e., the loss of reward incurred by not following the optimal treatment regime at each decision interval, for Scenario 1 is $0.5 \times |A1c_j + 0.5N_{j-1} - 10| \{I(A1c_j > 0) - I(A1c_j + 0.5N_{j-1} > 10)\}^2$, and the regret for Scenario 2 is $0.5 \times |A1c_j - 7| \{I(A1c_j > 0) - I(A1c_j + 0.5N_{j-1} > 10)\}^2$. Then the underlying optimal rule is the rule that yields zero regret for all decision intervals. Hence, in both scenarios, the optimal DTR is shared across decision points, and $d_j^*(H_j) = \text{sgn}(A1c_j + 0.5N_{j-1} > 10)$, where $H_j = (A1c_j, BP_j, Weight_j, L_j, N_{j-1})$ in our situation. Note that we can include intermediate outcomes such as $Y_1, \dots, Y_{j-1}$ in H_j as long as they are observable. The implementation algorithm remains the same. In the first example, the difference between the treatment effects, also known as the contrast function, can be specified as $0.5 \times (A1c_j + 0.5N_{j-1} - 10)$. Hence, the Q-contrast function is correctly specified when using a linear model for Q function. In Scenario 2, the treatment effect is characterized by $0.5 \times |A1c_j - 7| \{I(A1c_j > 0) - I(A1c_j + 0.5N_{j-1} > 10)\}^2$. The implied true contrast functions

are no longer of the form above. Hence, the Q-contrast function is incorrectly specified in a linear Q function, which will impact the performance of Q learning to a larger extent. See Zhang et al (2013) for more discussion (14).

The proposed censored shared-Q-learning and censored shared-O-learning are compared with censored Q-learning (9), which does not take into account the shared data structure. We consider sample sizes of 2000 and 5000. We generate a validation dataset with 50000 observations. The experiment is performed 500 times independently. In each replicate, we calculate the mean response for all subjects in the validation dataset had the whole population followed the estimated rule. The averaged outcome is used to compare different methods.

The censored shared-O-learning is implemented following (7) with $\phi(t) = -\log(1 + e^{-t})$, and K in the soft minimum function is set to 1. We looked at other values of K , which gave similar results. The propensity scores at each decision point are estimated using the treatment proportion. For the censored shared-Q-learning, we use the linear model for the Q-function, and parameters are estimated via weighted least squares as presented in (4). For the censored Q-learning, the linear model is also used for the Q-function, where we let $Q_j(H_j, A_j) = \beta_j^T H_{j0} + (\psi_j^T H_{j1}) A_j$. A weighted least square is utilized at each decision point to solve for $\hat{\beta}_j$ and $\hat{\psi}_j$.

We use the Cox proportional hazards model to estimate $S_C\{\min(\sum_{k=1}^j Y_k, C) | H_j, A_j\}$ and construct the weight at each decision point. Note that in censored shared-O-learning method, only $S_C\{\min(\sum_{k=1}^j Y_k, C) | H_T, A_T\}$ needs to be fitted. For all decision points, we use A1c, BP and weight at the baseline level to fit the Cox models. Let Z_C denote the regressors, and $\lambda_{Ci}(t)$ denote the hazard functions of censoring times for subject i . Then, $\lambda_{Ci}(t) = \lambda_{C0}(t) \exp(\beta_C^T Z_{Ci})$, where $\lambda_{C0}(t)$ is the baseline hazard functions for censoring time. The estimator for β_C , say $\hat{\beta}_C$, maximizes the partial likelihood

$$\prod_{i=1}^n \left\{ \frac{\exp(\hat{\beta}_C^T Z_{Ci})}{\sum_{U_j \leq U_i} \exp(\hat{\beta}_C^T Z_{Cj})} \right\}^{1 - \Delta_i}.$$

We use the Breslow estimator for the cumulative baseline hazard function $\Lambda_{C0}(t)$. An estimator of $S_C(t | H_{ji}, A_{ji})$ for subject i is $\hat{S}_C(t | H_{\bar{T}, i}, A_{\bar{T}, i}) = \exp\{-\hat{\Lambda}_{C0}(t)\}^{\exp(\hat{\beta}_C^T Z_{Ci})}$, where $\hat{\Lambda}_{C0}(t)$ is the estimator for $\Lambda_{C0}(t)$.

The means and the standard errors of the 500 values of the estimated DTRs on the validation set for both scenarios are presented in Table 1. In Scenario 1, when the regression model is close to being correctly specified, the censored shared-Q-learning leads to the best performance. The censored Q-learning approach does not account for the shared data structure. Thus, there is a large variation in the obtained results. In Scenario 2, both censored shared-Q-learning and censored Q-learning methods are sensitive to model misspecification.

Conversely, censored shared-O-learning has a robust performance, though it has slightly worse results compared to the censored shared-Q-learning method in Scenario 1. In practice, it is unknown to us whether the regression model in censored shared-Q-learning could be correctly specified. We can use a cross-validation approach to choose the one that yields a better result (e.g. larger estimated value) between censored shared-Q-learning and censored shared-O-learning methods.

5. Data Analysis

In this section, we apply the proposed methods to the Framingham Heart Study. The Framingham Heart Study, established in 1948, is the first longitudinal prospective large-scale cohort to study cardiovascular disease in the US. In the original cohort 5209 men and women are monitored prospectively for epidemiological and genetic risk factors for cardiovascular disease. There are maximum 32 examinations that occurred biannually during the 65 years of followup (31). For illustration we consider only information from the second to sixth visits. In our dataset, 2236 subjects are available with complete information on the risk factors at each measurement time and are free of cardiovascular disease at the time of examination. The long-term outcome of interest is time to onset of the first major cardiovascular disease event or death, which has an event rate of 18.7% by the end of the study. The median followup time is 25 years, and the ages at baseline range from 17 to 70 with a median of 43. Traditionally, hypertension medication is recommended based on the blood pressure level. However, the outcomes might be improved if other clinical information is also factored in. We utilize the Framingham Heart Study data to derive a decision rule that informs a patient whether a hypertension medication should be taken at each decision point, aiming to reduce the long-term risk of cardiovascular disease. Risk factors considered in our prediction at each visit include age, diastolic blood pressure, cholesterol, high-density lipoprotein, presence of diabetes, and smoking.

We first carry out a cross-validation procedure to select the method between censored shared-Q-learning and censored shared-O-learning. At each run, we partition the whole dataset into two equal parts, with one part serving as training data to estimate the DTRs with shared parameters using both methods, and the other part as the validation set for implementing the estimated DTRs with shared parameters. When estimating DTRs with shared parameters, the Kaplan–Meier method is used to fit the censoring probabilities. The probability functions $\pi(a; h)$ are fitted using logistic regression, and all the covariates used for prediction mentioned above are included in the model. We average the area under the Receiver Operating Characteristic (ROC) curve (AUC) for each data split across all cross validated samples. The average AUC was around 0.9, indicating that the model fitted data well. The constructed DTRs with shared parameters from the training set are evaluated using the mean residual survival time on the validation set, where censoring is adjusted for. Furthermore, we calculate the propensity scores in the validation set using the estimated coefficients from training data, which are subsequently used to adjust for possible confounding of treatments. Each part served as the validation subset once, and the cross-validated values of the constructed DTRs are obtained by averaging the empirical values on both validation subsets. The procedure is repeated 100 times.

Our implementation shows that both methods result in similar cross-validated values of mean residual survival time in years (censored shared-Q-learning: 16.60, and censored shared-O-learning: 17.00), indicating a potential improvement of outcome when compared with the mean residual survival time of 13.57 years under the current treatment plan. To obtain a final set of robust estimates, we carry out both methods on the whole dataset. The coefficients in the estimated DTRs with shared parameters, along with the 95% confidence intervals using a bootstrap method, are presented in Table 2. For future patients with covariates measured at any decision points, treatment decision rules can be derived using a linear combination of them based on these coefficient estimates. We find that results from the CSQL method have a larger variability, perhaps indicating a poor fit of the model in CSQL. Figure 1 shows the treatment allocation rates from the constructed dynamic treatment regimes. We observe different recommended rules from the two methods. This could happen when there are patients who don't have great differential treatment effects. In general, patients who are currently on medication are more likely to continue taking it. There is a slight increase in the proportion of subjects recommended for medication in their later years from censored shared-Q-learning recommendations, compared with the current data. Conversely, fewer patients are recommended for medication when using censored shared-O-learning rules. In either case, the survival benefit could be significantly improved under the recommended DTRs with shared parameters. Figure 2 shows the Kaplan–Meier curves of time to first cardiovascular disease event for patients whose treatments are consistent with censored shared-Q-learning recommendations (left panel) and censored shared-O-learning recommendations (right panel) versus those who are not, at the first and subsequent time points. It is clear that subjects whose medication coincided with the recommendation had on average better survival outcome. In the situation when recommended treatment actions are different but the treatment benefits are similar, we would recommend physicians select treatments while considering other factors such as treatment toxicity and monetary costs.

6. Discussion

We proposed two new methods for constructing a DTR with shared parameters with survival outcomes, which is a fixed function of time-varying covariates over time. Such a rule yields an optimal treatment strategy that can be easily implemented in practice. We provide efficient computing algorithms to obtain the solution. Our method for the decision rule is based on a linear combination of updated covariate information. It is also of interest to develop a more robust tree structured decision rule, e.g. Zhu et al (2017) (32), without the assumption of linearity, which has the additional advantage of ease of interpretation and dissemination. The censored shared-O-learning is proposed based on an inverse probability weighted estimator of the expected outcome that would be achieved under a particular DTR. However, such an estimator is potentially less efficient because it only uses outcome information from subjects whose treatment assignments coincide with those dictated by the DTR of interest. In the future, we can develop DTRs with shared parameters via censored shared-O-learning using an augmented inverse probability weighted estimator (14; 22). Such an approach can incorporate contributions from the subjects who did not receive the specified treatment assignments across all decision points by estimating their pseudo

outcomes using censored Q-learning. Hence, it will improve efficiency over the censored shared-O-learning proposed here.

It is likely that the proposed two methods recommend different treatment sequences while the expected clinical outcomes are similar, as shown in the data analysis. This could happen when patients are on the decision boundary, where there are few differential treatment effects for them. In the situation when recommended treatment actions are different but the treatment benefits are similar, we would recommend physicians select treatments while taking into account other factors such as treatment toxicity and monetary costs. In the future, we can develop procedures that formally incorporate cost consideration in constructing dynamic treatment regimes, which has been explored in the one-stage setup (33), but not yet in the DTRs setting.

In our simulation we considered a sample size of 2000 to be close to the Framingham study, however we expect the methods should be applicable to other cohorts with sample sizes smaller than 2000. With more visits, we do require a larger sample size since we need to ensure all possible treatment sequences are observed with a sufficient sample size. Otherwise, positivity assumptions are violated, and we can only make treatment recommendations for the subpopulation (instead of the full population) where the positivity assumptions holds. Laber et al (2016) developed a method for powering a single stage, two-armed randomized clinical trial when the primary aim is estimating the optimal single stage personalized treatment strategy (34). Following their development, we can develop sample size formula for constructing high-quality DTRs with shared parameters in the future.

We expect the consistency results will be similar to Chakraborty et al (18), as long as the model for censoring time is correctly specified for censored shared-Q-learning. For the censored shared-O-learning, since now we have restricted the form of decision rules, the consistent results do not follow directly from Zhao et al (2015) (11). Qiu et al (2018) develop a method to estimate a linear treatment rule in the one-stage setup that is guaranteed to achieve optimal reward among the class of all linear rules (35). We will investigate the theoretical properties of the censored shared-O-learning with possible modifications on surrogate loss.

In this paper, we only considered binary options for treatment at each decision point, but in many situations we need to consider making decisions among three or more treatments each time. It is straightforward to conduct censored shared-Q-learning in these settings, where we can create indicators for each treatment, and include these treatment indicators and covariates by treatment indicators interactions into the regression model for pseudo-outcome. An important extension to consider is to generalize the proposed censored shared-O-learning approach handle three or more treatments. The case of multicategory treatments in regular DTR settings has been studied recently (36), and a similar generalization may be possible for constructing DTRs with shared parameters when three or more treatments are involved. Another important extension to consider is to generalize the proposed methods to settings involving continuous dose finding. In particular, we could modify the Q-function to be a quadratic function (instead of a linear function) of the covariates and the treatment, then the shared-Q-learning will be applicable for this setting. On the other hand, Chen et al

(2016) had shown that O-learning for continuous dose is a weighted regression problem (37). Hence, replacing the weighted classification loss in shared-O-learning with a weighted regression loss can extend our shared-O-learning framework to continuous dose finding settings.

7. Software

Software in the form of R code, together with an example on implementation is available as web supplementary material.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

The Framingham Heart Study and the Framingham SHARe project are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University. The Framingham SHARe data used for the analyses described in this manuscript were obtained through dbGaP (access number: phs000007.v3.p2). This manuscript was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or the NHLBI. Ying-Qi Zhao was supported by grants R01DK108073, U10CA180819 and P30 CA015704 awarded by the National Institutes of Health and Colman Early Career Fellowship awarded by Hope Foundation. Ruqing Zhu was supported by the National Center for Supercomputer Applications Fellowship. Guanhua Chen was supported by the Clinical and Translational Science Award (CTSA), National Institutes for Health National Center for Advancing Translational Sciences (NCATS) (UL1TR000427), and Patient Centered Outcomes Research Institute (PCORI) Awards (ME-2018C2-13180). The views in this publication are solely the responsibility of the authors and do not necessarily represent the views of the PCORI, its Board of Governors or Methodology Committee. Yingye Zheng was supported by R01CA236558 and U24 CA086368 awarded by the National Institutes of Health.

References

- [1]. Murphy SA. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society, Series B* 2003; 65:331–366.
- [2]. Murphy SA. An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine* 2005; 24:1455–1481. [PubMed: 15586395]
- [3]. Socinski M, Stinchcombe T. Duration of first-line chemotherapy in advanced non small-cell lung cancer: less is more in the era of effective subsequent therapies. *J Clin Oncol* 2007; 25:5155–5157. [PubMed: 18024862]
- [4]. Murphy SA, Oslin DW, Rush AJ, Zhu J, MCATS. Methodological challenges in constructing effective treatment sequences for chronic psychiatric disorders. *Neuropsychopharmacology* 2007; 32:257–262. [PubMed: 17091129]
- [5]. Watkins CJCH. Learning from delayed rewards. PhD Thesis, King's College, Cambridge 1989.
- [6]. Nahum-Shani I, Qian M, Almirall D, Pelham WE, Gnagy B, Fabiano GA, Waxmonsky JG, Yu J, Murphy SA. Q-learning: A data analysis method for constructing adaptive interventions. *Psychological methods* 2012; 17(4):478. [PubMed: 23025434]
- [7]. Zhao Y, Kosorok MR, Zeng D. Reinforcement learning design for cancer clinical trials. *Statistics in medicine* 2009; 28(26):3294–3315. [PubMed: 19750510]
- [8]. Laber EB, Linn KA, Stefanski LA. Interactive model building for q-learning. *Biometrika* 2014; 101(4):831–847. [PubMed: 25541562]
- [9]. Goldberg Y, Kosorok MR. Q-learning with censored data. *Annals of statistics* 2012; 40(1):529. [PubMed: 22754029]

- [10]. Robins JM. Optimal structural nested models for optimal sequential decisions Proceedings of the second seattle Symposium in Biostatistics, Springer, 2004; 189–326.
- [11]. Zhao YQ, Zeng D, Laber EB, Kosorok MR. New statistical learning methods for estimating optimal dynamic treatment regimes. *Journal of the American Statistical Association* 2015; 110(510):583–598. [PubMed: 26236062]
- [12]. Robins J, Orellana L, Rotnitzky A. Estimation and extrapolation of optimal treatment and testing strategies. *Statistics in medicine* 2008; 27(23):4678–4721. [PubMed: 18646286]
- [13]. Orellana L, Rotnitzky A, Robins JM. Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, part i: main content. *The international journal of biostatistics* 2010; 6(2).
- [14]. Zhang B, Tsiatis AA, Laber EB, Davidian M. Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika* 2013; 100:681–695.
- [15]. Chakraborty B, Moodie E. *Statistical methods for dynamic treatment regimes*. Springer, 2013.
- [16]. Kosorok MR, Moodie EE. *Adaptive treatment strategies in practice: planning trials and analyzing data for personalized medicine*. SIAM, 2015.
- [17]. Kosorok MR, Laber EB. Precision medicine. *Annual review of statistics and its application* 2019; 6:263–286.
- [18]. Chakraborty B, Ghosh P, Moodie EE, Rush AJ. Estimating optimal shared-parameter dynamic regimens with application to a multistage depression clinical trial. *Biometrics* 2016; 72(3):865–876. [PubMed: 26890628]
- [19]. Sutton RS, Barto AG. *Reinforcement learning: An introduction*, vol. 1 MIT press Cambridge, 1998.
- [20]. Antos A, Szepesvári C, Munos R. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning* 2008; 71(1):89–129.
- [21]. Moodie EE, Richardson TS. Estimating optimal dynamic regimes: Correcting bias under the null. *Scandinavian Journal of Statistics* 2010; 37(1):126–146.
- [22]. Hager R, Tsiatis AA, Davidian M. Optimal two-stage dynamic treatment regimes from a classification perspective with censored survival data. *Biometrics* 2018; 74(4):1180–1192. [PubMed: 29775203]
- [23]. Simoneau G, Moodie EE, Nijjar JS, Platt RW, Investigators SERAIC. Estimating optimal dynamic treatment regimes with survival outcomes. *Journal of the American Statistical Association* 2019; (just-accepted):1–24.
- [24]. Bellman R *Dynamic Programming*. Princeton University Press: Princeton, 1957.
- [25]. Sutton RS, Barto AG. *Reinforcement Learning I: Introduction*. MIT Press: Cambridge, MA, 1998.
- [26]. Murphy SA. An experimental design for the development of adaptive treatment strategies. *Statistics in medicine* 2005; 24(10):1455–1481. [PubMed: 15586395]
- [27]. Bartlett PL, Jordan MI, McAuliffe JD. Convexity, classification, and risk bounds. *Journal of the American Statistical Association* 2006; 101(473):138–156.
- [28]. Andrew G, Gao J. Scalable training of l1-regularized log-linear models. *Proceedings of the 24th international conference on Machine learning, ACM*, 2007; 33–40.
- [29]. Timbie JW, Hayward RA, Vijan S. Diminishing efficacy of combination therapy, response-heterogeneity, and treatment intolerance limit the attainability of tight risk factor control in patients with diabetes. *Health services research* 2010; 45(2):437–456. [PubMed: 20070387]
- [30]. Ertefaie A, Strawderman RL. Constructing dynamic treatment regimes over indefinite time horizons. *Biometrika* 2018; 105(4):963–977.
- [31]. Tsao CW, Vasan RS. Cohort profile: The framingham heart study (fhs): overview of milestones in cardiovascular epidemiology. *International journal of epidemiology* 2015; 44(6):1800–1813. [PubMed: 26705418]
- [32]. Zhu R, Zhao YQ, Chen G, Ma S, Zhao H. Greedy outcome weighted tree learning of optimal personalized treatment rules. *Biometrics* 2017; 73(2):391–400. [PubMed: 27704531]

- [33]. Wang Y, Fu H, Zeng D. Learning optimal personalized treatment rules in consideration of benefit and risk: with an application to treating type 2 diabetes patients with insulin therapies. *Journal of the American Statistical Association* 2018; 113(521):1–13. [PubMed: 30034060]
- [34]. Laber EB, Zhao YQ, Regh T, Davidian M, Tsiatis A, Stanford JB, Zeng D, Song R, Kosorok MR. Using pilot data to size a two-arm randomized trial to find a nearly optimal personalized treatment strategy. *Statistics in medicine* 2016; 35(8):1245–1256. [PubMed: 26506890]
- [35]. Qiu X, Zeng D, Wang Y. Estimation and evaluation of linear individualized treatment rules to guarantee performance. *Biometrics* 2018; 74(2):517–528. [PubMed: 28960239]
- [36]. Zhang C, Chenm J, Fu H, He X, Zhao Y, Liu Y. Multicategory outcome weighted margin-based learning for estimating individualized treatment rules. *Statistica Sinica* 2019; In Press.
- [37]. Chen G, Zeng D, Kosorok MR. Personalized dose finding using outcome weighted learning. *Journal of the American Statistical Association* 2016; 111(516):1509–1521. [PubMed: 28255189]

Figure 1.

Treatment allocation of the current data, recommended rules from CSQL and CSOL; CSQL: censored shared-Q-learning; CSOL: censored shared-O-learning

Figure 2. Kaplan–Meier Curves of Probability of Time to First cardiovascular disease Event (left panel: CSQ; right panel: CSOL); CSQ: censored shared-Q-learning; CSOL: censored shared-O-learning

Table 1.

Mean values (s.e.) under the estimated optimal treatment rules. Opt: optimal value; CQL: censored Q-learning; CSQL: censored shared-Q-learning; CSOL: censored shared-O-learning.

Scenario	<i>n</i>	# of Stages	Opt	CQL	CSQL	CSOL
1	2000	10	9.71	7.56 (2.43)	9.63 (0.11)	8.25 (0.53)
1	2000	20	18.83	13.51 (5.83)	18.79 (0.07)	16.37 (1.62)
1	5000	10	9.71	7.75 (2.33)	9.66 (0.05)	8.19 (0.47)
1	5000	20	18.83	14.53 (5.42)	18.81 (0.03)	16.28 (1.70)
2	2000	10	9.68	5.89 (2.02)	4.96 (0.04)	7.86 (0.18)
2	2000	20	18.76	8.27 (4.26)	5.72 (0.09)	14.70 (0.21)
2	5000	10	9.68	5.38 (1.61)	4.96 (0.03)	7.86 (0.21)
2	5000	20	18.76	7.02 (3.49)	5.73 (0.05)	14.67 (0.71)

Table 2.

Coefficients by CSQL and CSOL for the Framingham Heart Study (CSQL: censored shared-Q-learning; CSOL: censored shared-O-learning)

	CSQL	CSOL
Intercept	1 (−6.720, 3.092)	−1 (−1.239, −0.806)
Age	0.020 (−0.057, 0.045)	0.005 (0.004, 0.008)
Medication (Yes = 1)	0.414 (−0.314, 1.698)	0.269 (0.232, 0.322)
Diastolic blood pressure	−0.027 (−0.039, 0.034)	0.007 (0.005, 0.009)
Total cholesterol	−0.007 (−0.007, 0.015)	−0.000 (−0.001, 0.000)
HDL cholesterol	0.024 (−0.017, 0.044)	0.001 (0.000, 0.003)
Diabetes (Yes = 1)	−0.268 (−2.129, 1.894)	0.098 (0.026, 0.149)