Article

Survival Augmented Patient Preference Incorporated Reinforcement Learning to Evaluate Tailoring Variables for Personalized Healthcare

Yingchao Zhong , Chang Wang and Lu Wang *

Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA; zhongych@umich.edu (Y.Z.); wangchan@umich.edu (C.W.)

* Correspondence: luwang@umich.edu

Abstract: In this paper, we consider personalized treatment decision strategies in the management of chronic diseases, such as chronic kidney disease, which typically consists of sequential and adaptive treatment decision making. We investigate a two-stage treatment setting with a survival outcome that could be right censored. This can be formulated through a dynamic treatment regime (DTR) framework, where the goal is to tailor treatment to each individual based on their own medical history in order to maximize a desirable health outcome. We develop a new method, Survival Augmented Patient Preference incorporated reinforcement Q-Learning (SAPP-Q-Learning) to decide between quality of life and survival restricted at maximal follow-up. Our method incorporates the latent patient preference into a weighted utility function that balances between quality of life and survival time, in a Q-learning model framework. We further propose a corresponding m-out-of-n Bootstrap procedure to accurately make statistical inferences and construct confidence intervals on the effects of tailoring variables, whose values can guide personalized treatment strategies.

Citation: Zhong, Y.; Wang, C.; Wang, L. Survival Augmented Patient Preference Incorporated Reinforcement Learning to Evaluate Tailoring Variables for Personalized Healthcare. *Stats* **2021**, *4*, 776–792. <https://doi.org/10.3390/stats4040046>

Academic Editor: Haiming Zhou

Received: 3 August 2021

Accepted: 22 September 2021

Published: 27 September 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: dynamic treatment regime; m-out-of-n Bootstrap; personalized treatment decision; Q-learning; right censored survival

1. Introduction

For chronic illnesses, patients often have to navigate a series of treatment decisions. It has been increasingly recognized that due to patient heterogeneity based on genetics, environmental factors, various other factors and the interplay between the factors, a good treatment plan needs to be both personalized and adaptive to a patient's changing clinical course. Dynamic treatment regimes (DTRs) are algorithmic solutions to this clinical problem, where a DTR consists of a sequence of treatment decision rules that adapt over time in response to an individual's clinical response and health outcome trajectory.

A large number of methods have been proposed for the evaluation of the optimal DTR. Some of the earlier and foundation works include the marginal structural model with inverse probability weighting (IPW) [1], G-estimation of structural nested mean models [2], Q-learning ([3–5]), and A-learning [6]. More recently, along with the development of data science, machine learning flavored methods were also developed for DTR estimation, including tree-based and list-based methods ([7–10]), classification type methods ([11,12]), and stochastic tree search methods [13].

Despite the large number of methods that can be used to evaluate the optimal DTR, the majority of methods rely on a single pre-specified reward endpoint of interest. Often times, in practice, multiple competing priorities need to be balanced when considering a clinical decision and in fact these outcomes can be affected in opposing directions of desirability. A classic example in this case is toxicity vs. efficacy. A newly introduced drug that is highly efficacious might come with a larger burden of undesirable side effects. In recent years, a few proposed methods have tackled the delicate balance between multiple outcomes of a

proposed treatment. Butler et al. [14] balanced treatment efficacy and toxicity using patient derived preference using a Q-learning approach, while [15] assigned different rewards based on survival status, wellness (a measure of toxicity) and tumor size (a measure of drug efficacy) at each stage.

The particular scenario that we would like to address in this work is a delicate balance of quality vs. quantity, a dilemma commonly encountered for patients facing severe illness. For many chronic diseases, patients often receive a first treatment and are followed up after a short time to determine if the treatment needs to be adjusted. The patient is then followed until an adverse event or a certain maximal follow-up time. Adverse events could include hospitalization, organ failure, impending surgery, or other undesired medical consequence, but for simplicity, we look at mortality as our event of interest [16]. Our primary goal then is to estimate the optimal treatment regime that would maximize a patient preference weighted combination of quality of life and survival time. Secondly, we would like to provide an inference framework for more confident decision making. Although similar in flavor to the abovementioned works, our scenario brings with it a unique set of distinct technical challenges.

The main challenge in this scenario is the presence of censored data. Because of the long tail of survival distributions, and because of other logistical reasons (patients move, dropout due to deteriorating health, etc.), it is common to not observe the outcomes of a significant fraction of the population. However, partially observed information from censored subjects can still contribute important information and give power to the analysis if analyzed correctly. An enormous body of work has been developed to estimate the optimal rule or regime in the presence of censoring. A non-exhaustive and overlapping list includes methods for a single stage ([17–19]), methods for multiple stages ([20–23]), inverse probability of censoring weighted (IPCW) adjustment methods ([19,20,22,24]), Q-learning-based methods ([20,22]), tree-based methods ([17,18]), survival probability-based methods ([23]), accelerated failure time (AFT)-based models ([25,26]), and doubly robust methods ([21,27,28]), just to name some of the numerous ways by which these methods differ in scope and direction.

A secondary but equally important goal of our method is to provide inference on stage-specific parameters, particularly for tailoring variables. Inference in DTR methods is challenging due to the known issue of nonregularity caused by non-smooth functions that get carried forward through backward induction [29]. When the degree of nonregularity is large (in other words, a larger fraction of covariate space does not have a treatment effect), the asymptotic distribution of the true coefficient oscillates between two asymptotic distributions, resulting in asymptotic bias and poor Wald-type confidence intervals. In the same paper, Chakraborty et al. [29] proposed a hard-threshold estimator and soft-threshold estimator to adjust for this poor coverage. Laber et al. 2014 [30] proposed an adaptive confidence interval for first stage parameters by utilizing regular, uniformly convergent lower and upper bounds for the asymptotic distribution of interest, and later bootstrapping for the confidence set. Chakraborty et al. 2013 [31] proposed an adaptive bootstrap-based method that adjusts for the bias and coverage by adjusting the bootstrap sample size.

In this work, we explore, synthesize, and adapt existing methods in the literature to create a new method for estimating the optimal treatment regime and constructing stage-specific confidence intervals that fit our scenario. We utilize IPCW to enable a complete data analysis and use the Q-learning framework for the modeling approach. We further adapt the m-out-of-n bootstrap to accommodate censoring in order to obtain covariate-specific confidence intervals for inference. We illustrate the numerical performance of our method through simulation studies.

2. Notation and Framework

As shown in Figure 1, we look at a two-stage setting where a patient, upon diagnosis, receives a stage 1 treatment. Shortly thereafter, at a scheduled follow-up time, the patient will be assessed for stage 2 treatment. Following stage 2 treatment, the patient is at risk of

an adverse event, i.e., death, denoted by time D_i . Although in some cases adverse events might happen early, our assumption that the death does not happen in Stage 1 is fairly natural in the setting of many chronic illnesses such as diabetes, hypertension, chronic kidney disease, or many autoimmune diseases, where significant clinical decline is not acutely expected. Finally, in this setting, patient information might be lost to follow-up, either due to administrative censoring (surpassed maximal follow-up time) or due to patient factors.

Let T_{i1}, T_{i2} denote the times of treatment for stages 1, 2, respectively, for a patient i . Let τ be the maximal administrative follow-up time. Let S_{i1} and S_{i2} denote the amount of time survived in each stage, i.e., $S_{i1} = T_{i2} - T_{i1}$, and $S_{i2} = D_i - T_{i2}$. We assume that per protocol, everyone's S_{i1} should be the same (i.e., routine assessment following stage 1 treatment at a prespecified time interval). Let K_j be the number of treatment options in the j th stage, $j = 1, 2$. Let A_{ij} denote the treatment indicator for the i th patient in the j th stage, with a_{ij} denoting the observed treatment. Let $\mathbf{X}_{ij}$ denote patient characteristics and other historical information such as outcomes and prior treatments $A_1, \dots, A_{j-1}$ prior to treatment assignment at stage j . In addition, we assume that each patient will have an evolving preference h_{ij} , which can be derived from answers $\mathbf{W}_{ij}$ to a questionnaire at stage j .

At each stage, we assume that each patient i will have two observed outcomes, q_{ij} (with the range $[0, 1]$) for the average quality of life during stage j , and S_{ij} , the amount of time spent in stage j . q_{ij} allows us to calculate $Qu_{ij} = q_{ij} * S_{ij}$, the quality adjusted life years during stage j . However, since Qu_{ij} is dependent on S_{ij} , it is also subject to censoring at stage 2.

Figure 1. Schematic of clinical progression timeline.

We assume a utility function of the form $Qu + \{1 - \Phi(h)\}\{S - Qu\}$, where $\Phi(\cdot)$ denotes the cumulative distribution function of a normal random variable. Intuitively, this utility function is a sliding scale between S and Qu , and a patient's preference would dictate where he/she would fall. Let the overall outcome of interest that we would like to optimize be $R_{i1} + R_{i2}$, where $R_{ij} = Qu_{ij} + \{1 - \Phi(h_{ij})\}\{S_{ij} - Qu_{ij}\}$. This is a cumulative preference of adjusted quality of life years experienced on a given regime. If one's preference is such that $\Phi(h_i) = 0$ for both stages, then $R_{i1} + R_{i2}$ is the total survival time. Similarly, if a patient has the preference $\Phi(h_i) = 1$ for both stages, then $R_{i1} + R_{i2}$ would be the total quality adjusted life years. In general, we denote all history, or history up to stage K for a given variable with an overhead bar (i.e., $\bar{\mathbf{W}}_i$ and $\bar{\mathbf{W}}_{iK}$, respectively).

Let $g_j(\bar{\mathbf{X}}_{ij}, \bar{\mathbf{W}}_{ij})$ be a function that maps from covariate and survey history to the domain of treatment assignment A_{ij} . At stage j , the expected potential reward of the following decision rule $g_j(\bar{\mathbf{X}}_{ij}, \bar{\mathbf{W}}_{ij})$ for patient i is defined as $\mathbb{E}[\sum_{a_j=1}^{K_j} [Qu_{ij}^*(a_j) + \{1 - \Phi(h_{ij})\}\{S_{ij}^*(a_j) - Qu_{ij}^*(a_j)\}] I\{g_j(\bar{\mathbf{X}}_{ij}, \bar{\mathbf{W}}_{ij}) = a_j\}]$, where $S_{ij}^*(a_j) = S_{ij}^*(A_{i1}, \dots, A_{i,j-1}, a_j)$ and $Qu_{ij}^*(a_j) = Qu_{ij}^*(A_{i1}, \dots, A_{i,j-1}, a_j)$ denotes the counterfactual survival outcome and quality adjusted survival, respectively, where the patient is assumed to have taken treatment a_j at stage j , conditional on previous treatment decisions $A_{i1}, \dots, A_{i,j-1}$ in the sense that previous treatments need to be fixed to those actually given in previous stages. In our case, our primary goal is to find a sequence of individualized decision rules, $\mathbf{g}(\bar{\mathbf{X}}_i, \bar{\mathbf{W}}_i) = (g_1(\bar{\mathbf{X}}_{i1}, \bar{\mathbf{W}}_{i1}), g_2(\bar{\mathbf{X}}_{i2}, \bar{\mathbf{W}}_{i2}))$, that optimize the potential outcome of $R_{i1} + R_{i2}$.

A second but equally important objective is to conduct inference on coefficients, with particular emphasis on tailoring variables (variables that interact with treatment selection). Inference on tailoring variables is important because it obviates the need to collect data for covariates that have no evidence of significant deviation from zero. Furthermore, inference allows us to know when there is insufficient evidence to support one treatment over another so that treatment decisions could be made using other factors important to the patient. In this work, along with stage-specific decision rules, we present a censoring adapted method of obtaining confidence intervals for the covariates in both stages of the model. For the sake of brevity going forward, let us abbreviate $g_j(\bar{\mathbf{X}}_{ij}, \bar{\mathbf{W}}_{ij})$ with g_j and drop the patient index i when there is no confusion.

3. Methods: Censoring Adapted Q-Learning

3.1. Traditional Q-Learning

First, we introduce traditional Q-learning, a form of approximate dynamic programming originally proposed by [3]. Q-learning estimates the optimal DTR by postulating regression models for Q-functions and subsequently taking solutions that would yield the largest rewards. The Q-functions for the two stages are defined as:

$$Q_2(\bar{\mathbf{X}}_2, A_2) = E[R_2 | \bar{\mathbf{X}}_2, A_2]$$

and

$$Q_1(\bar{\mathbf{X}}_1, A_1) = E[R_1 + \max_{a_2} Q_2(\bar{\mathbf{X}}_2, a_2) | \bar{\mathbf{X}}_1, A_1]$$

The optimal decision rule at stage j can be expressed as $d_j(\bar{\mathbf{X}}_j) = \arg \max_{a_j} Q_j(\bar{\mathbf{X}}_j, a_j)$. Generally, we do not know the true Q-functions and so we consider linear working models for Q-functions of the form $Q_j(\bar{\mathbf{X}}_j, A_j; \beta_j, \psi_j) = \beta_j^T \bar{\mathbf{Z}}_{j,0} + (\psi_j^T \bar{\mathbf{Z}}_{j,1}) A_j$, where $\bar{\mathbf{Z}}_{j,0}$ and $\bar{\mathbf{Z}}_{j,1}$ possibly contain different components of the history $\bar{\mathbf{X}}_j$.

The two-stage Q-learning algorithm works as follows:

1. Stage 2 regression is obtained by $(\hat{\beta}_2, \hat{\psi}_2) = \arg \min_{\beta_2, \psi_2} \mathbb{E}_n (R_2 - Q_2(\bar{\mathbf{X}}_2, A_2; \beta_2, \psi_2))^2$, where $\mathbb{E}_n$ denotes the empirical mean;
2. The stage 1 pseudo-outcome is given by $\widetilde{PO}_1 = R_1 + \max_{a_2} Q_2(\bar{\mathbf{X}}_2, a_2; \hat{\beta}_2, \hat{\psi}_2)$
3. Stage 1 regression: $(\hat{\beta}_1, \hat{\psi}_1) = \arg \min_{\beta_1, \psi_1} \mathbb{E}_n (\widetilde{PO}_1 - Q_1(\bar{\mathbf{X}}_1, A_1; \beta_1, \psi_1))^2$.

The optimal decision rules can further be written as $d_j(\bar{\mathbf{X}}_j) = \arg \max_{a_j} Q_j(\bar{\mathbf{X}}_j, a_j; \hat{\beta}_j, \hat{\psi}_j) = \text{sign}(\hat{\psi}_j^T \bar{\mathbf{Z}}_{j,1})$ when we have the particular case that $A_j \in \{-1, 1\}$. We will assume that we have binary treatment options for both stages for convenience, although this can be relaxed with further assumptions.

3.2. Censoring Adapted Q-Learning

Our stage 2 optimization objective is complicated by the fact that some S_2 may be unobserved due to censoring. Let C denote the time of censoring, which happens after T_2 , the time of stage 2 treatment. We assume that $S_2 \perp C | \mathbf{A}_2, \bar{\mathbf{X}}_2, \bar{\mathbf{W}}_2$ (conditional independence).

3.2.1. Stage 2

Let $S_2^*(a_2)$ be the counterfactual outcome of survival starting from the stage 2 treatment conditional on the previous treatment A_1 . Correspondingly, $S_2^*(g_2)$ is the counterfactual outcome under decision rule g_2 , i.e., $S_2^*(g_2) = \sum_{a_2=1}^{K_2} S_2^*(a_2) I\{g_2(\bar{\mathbf{X}}_2, \bar{\mathbf{W}}_2) = a_2\}$. Similarly, we can obtain $Qu_2^*(a_2)$ through $q_2^*(a_2)$ and $S_2^*(a_2)$.

Using the linear utility function defined in Section 2 and conditional on the previous treatment A_1 , $R_2^*(a_2) = Qu_2^*(a_2) + \{1 - \Phi(H_2)\} \{S_2^*(a_2) - Qu_2^*(a_2)\}$.

Correspondingly,

$$R_2^*(g_2) = Qu_2^*(g_2) + \{1 - \Phi(H_2)\} \{S_2^*(g_2) - Qu_2^*(g_2)\}$$

is the counterfactual utility, conditional on the previous treatments A_1 under decision rule g_2 .

The optimal regime, g_2^{opt} , satisfies $E\{R_2^*(g_2^{opt})\} \geq E\{R_2^*(g_2)\} \forall g_2 \in \mathcal{G}_2$, where $\mathcal{G}_2$ is the class of all potential decision rules for stage 2.

We make the following assumptions to connect the mean of counterfactual outcomes with the observed data:

1. Consistency:

$$S_2 = \sum_{a_2=1}^{K_2} S_2^*(a_2) I\{A_2 = a_2\} \text{ and } q_2 = \sum_{a_2=1}^{K_2} q_2^*(a_2) I\{A_2 = a_2\};$$

2. No unmeasured confounding:

Treatment A_2 is randomly assigned with probability possibly dependent on $\bar{X}_2, \bar{W}_2$, i.e., $\{S_2^*(1), \dots, S_2^*(K_2)\} \perp A_2 | \bar{X}_2, \bar{W}_2$ and $\{q_2^*(1), \dots, q_2^*(K_2)\} \perp A_2 | \bar{X}_2, \bar{W}_2$;

3. Positivity:

There exist constants $0 < c_0 < c_1$ such that, with probability 1, the propensity score $\pi_{a_2}(\bar{X}_2, \bar{W}_2) = Pr(A_2 = a_2 | \bar{X}_2, \bar{W}_2) \in (c_0, c_1)$;

4. Latent variable independence:

$$H_2 \perp (A_2, Qu_2^*(a_2), S_2^*(a_2)) | \bar{X}_2, \bar{W}_2.$$

The first three assumptions are standard assumptions in causal inference. The last assumption facilitates separate modeling of outcomes and preferences and can be weakened at the expense of more complicated models [14].

We denote the marginal expectation with respect to $\bar{X}_t, \bar{W}_t$ as $E_{\bar{X}_t, \bar{W}_t}$, abbreviated as $\mathbb{E}_t$ throughout. Furthermore, we denote $\mu_{2,a_2}^S(\bar{X}_2, \bar{W}_2) \equiv E\{S_2 | A_2 = a_2, \bar{X}_2, \bar{W}_2\}$, $\mu_{2,a_2}^q(\bar{X}_2, \bar{W}_2) \equiv E\{q_2 | A_2 = a_2, \bar{X}_2, \bar{W}_2\}$, and assume that $S_2^*(a_2)$ and $q_2^*(a_2)$ are conditionally independent given $\bar{X}_2, \bar{W}_2$. As in traditional Q-learning, we assume linear working models for each of our outcomes of interest (i.e., q_2, S_2 can be generated through underlying models of predictive and tailoring variables $\beta_2^T \mathbf{Z}_{20} + (\psi_2^T \mathbf{Z}_{21}) A_2$, where $\mathbf{Z}_{20}$ and $\mathbf{Z}_{21}$ are some possibly different components of $\bar{X}_2$ and $\bar{W}_2$).

Using the causal assumptions above, we bridge the observed data to the counterfactual outcome means:

$$E[R_2^*(g_2)] = \mathbb{E}_2 \left(\sum_{a_2 \in \{-1, 1\}} [\mu_{2,a_2}^q(\bar{X}_2, \bar{W}_2) \mu_{2,a_2}^S(\bar{X}_2, \bar{W}_2) + E\{1 - \Phi(H_2) | \bar{X}_2, \bar{W}_2\} \{\mu_{2,a_2}^S(\bar{X}_2, \bar{W}_2) - \mu_{2,a_2}^q(\bar{X}_2, \bar{W}_2) \mu_{2,a_2}^S(\bar{X}_2, \bar{W}_2)\}] I\{g_2 = a_2\} \right), \quad (1)$$

where the separate modeling of preference and outcomes is allowed by the fourth assumption.

With censoring, it is unlikely that all S_2 are observed. We propose the following estimator that re-weights observed complete data using inverse probability of censoring weights (IPCW):

$$\operatorname{argmin}_{\beta_2, \psi_2} \mathbb{P}_n \left(\left[R_2 - \hat{E}(R_2(\bar{X}_2, \bar{W}_2, A_2; \beta_2, \psi_2)) \right]^2 \frac{\Delta}{\widehat{Pr}\{\Delta = 1 | \bar{X}_2, \bar{W}_2, A_2\}} \right),$$

where $\hat{E}(R_2(\bar{X}_2, \bar{W}_2, A_2; \beta_2, \psi_2))$ denotes the model estimate for R_2 using observed data and covariates, $\Delta = I(S_2 < C)$ is the event indicator that the patient's data are not censored, and $\widehat{Pr}\{\Delta = 1 | \bar{X}_2, \bar{W}_2, A_2\}$ is a working estimator of the probability that the individual has not been censored by their event time.

Denote our Q-function here to be $Q_2(\bar{X}_2, \bar{W}_2, A_2) = \hat{E}[R_2 | \bar{X}_2, \bar{W}_2, A_2]$. The derived treatment rule is $\hat{g}_2^{opt}(\bar{X}_2, \bar{W}_2) = \operatorname{argmax}_{a_2} Q_2(\bar{W}_2, \bar{X}_2, A_2; \hat{\beta}_2, \hat{\psi}_2) = \operatorname{sign}(\hat{\psi}_2^T \mathbf{Z}_{21})$, where $\mathbf{Z}_{21}$ represents the tailoring variables in the stage 2 model.

3.2.2. Stage 1

For stages prior to the last stage, we use backward induction, and then the optimal $g_1^{opt}(\bar{\mathbf{X}}_1, \bar{\mathbf{W}}_1)$ at stage 1 can be derived similarly. Assuming that stage-specific rewards have been maximized after stage 1, we define the following stage 1 reward:

$$R_1^*(a_1) = q_1^*(a_1)S_1^*(a_1) + \{1 - \Phi(\mathbf{H}_1)\}\{S_1^*(a_1) - q_1^*(a_1)S_1^*(a_1)\} + \\ q_2^*(a_1)S_2^*(a_1) + \{1 - \Phi(\mathbf{H}_2)\}\{S_2^*(a_1) - q_2^*(a_1)S_2^*(a_1)\},$$

where $S_1^*(a_1)$ is as defined previously and $S_2^*(a_1) = S_2^*(a_1, g_2^{opt})$ denotes a counterfactual outcome given future optimized treatments and taking treatment a_1 at stage 1 (similarly for $q_1^*(a_1)$ and $q_2^*(a_1)$). Using this reward, we define the optimal regime at stage 1 as the one that satisfies

$$E\{R_1^*(g_1^{opt})\} \geq E\{R_1^*(g_1)\}$$

for all $g_1 \in \mathcal{G}_1$, where $\mathcal{G}_1$ is the class of all potential regimes at stage 1.

We make the following assumptions. Note that in our set-up, S_1 is predetermined and fixed for every patient, but this assumption does not apply to q_1 .

1. Consistency:

$$\sum_{a_1=1}^{K_1} S_k^*(a_1)I(A_1 = a_1) = \begin{cases} S_1 & \text{when } k = 1 \\ S_k(A_1, g_2^{opt}) & \text{when } k = 2 \end{cases} \\ \sum_{a_1=1}^{K_1} q_k^*(a_1)I(A_1 = a_1) = \begin{cases} q_1 & \text{when } k = 1 \\ q_k(A_1, g_2^{opt}) & \text{when } k = 2; \end{cases}$$

2. No unmeasured confounding:

$\{S_1^*(1), \dots, S_1^*(K_1)\} \perp A_1 | \bar{\mathbf{X}}_1, \bar{\mathbf{W}}_1$. It holds in our setting since $S_1^*(1), \dots, S_1^*(K_1) = S_1$ for $\forall a_1$. In addition, $\{q_1^*(1), \dots, q_1^*(K_1)\} \perp A_1 | \bar{\mathbf{X}}_1, \bar{\mathbf{W}}_1$;

3. Positivity: $\pi_{a_1}(\bar{\mathbf{X}}_1, \bar{\mathbf{W}}_1) = Pr(A_1 = a_1 | \bar{\mathbf{X}}_1, \bar{\mathbf{W}}_1)$ is bounded away from zero and one;

4. Latent variable independence: $H_1 \perp (A_1, S_k^*(a_1), Qu_k^*(a_1) | \bar{\mathbf{X}}_k, \bar{\mathbf{W}}_k$ where $k = 1, 2$.

Then we have

$$E[R_1^*(g_1)] = \mathbb{E}_1\left(\sum_{a_1 \in \{-1, 1\}} [\mu_{1,a_1}^q(\bar{\mathbf{X}}_1, \bar{\mathbf{W}}_1) \mu_{1,a_1}^S(\bar{\mathbf{X}}_1, \bar{\mathbf{W}}_1) + E\{1 - \Phi(H_1) | \bar{\mathbf{X}}_1, \bar{\mathbf{W}}_1\} \times \right. \\ \left. \{\mu_{1,a_1}^S(\bar{\mathbf{X}}_1, \bar{\mathbf{W}}_1) - \mu_{1,a_1}^q(\bar{\mathbf{X}}_1, \bar{\mathbf{W}}_1) \mu_{1,a_1}^S(\bar{\mathbf{X}}_1, \bar{\mathbf{W}}_1)\} I\{g_1 = a_1\}\right) \\ + \mathbb{E}_2\left(\sum_{a_1 \in \{-1, 1\}} [\mu_{2,a_1}^q(\bar{\mathbf{X}}_1, \bar{\mathbf{W}}_1) \mu_{2,a_1}^S(\bar{\mathbf{X}}_1, \bar{\mathbf{W}}_1) + E\{1 - \Phi(H_2) | \bar{\mathbf{X}}_2, \bar{\mathbf{W}}_2\} \times \right. \\ \left. \{\mu_{2,a_1}^S(\bar{\mathbf{X}}_1, \bar{\mathbf{W}}_1) - \mu_{2,a_1}^q(\bar{\mathbf{X}}_1, \bar{\mathbf{W}}_1) \mu_{2,a_1}^S(\bar{\mathbf{X}}_1, \bar{\mathbf{W}}_1)\} I\{g_1 = a_1\}\right), \quad (2)$$

where $\mu_{1,a_1}^S(\bar{\mathbf{X}}_1, \bar{\mathbf{W}}_1) = E(S_1 | A_1 = a_1, \bar{\mathbf{X}}_1, \bar{\mathbf{W}}_1)$, $\mu_{2,a_1}^S(\bar{\mathbf{X}}_1, \bar{\mathbf{W}}_1) = E[S_2(A_1, g_2^{opt}) | A_1 = a_1, \bar{\mathbf{X}}_1, \bar{\mathbf{W}}_1]$, and similarly for the equivalents of q . Similarly to stage 2, we further assume that $q_1^*(a_1) \perp S_1^*(a_1) | \bar{\mathbf{X}}_1, \bar{\mathbf{W}}_1$ and $q_2^*(a_1) \perp S_2^*(a_1) | \bar{\mathbf{X}}_1, \bar{\mathbf{W}}_1$. Notice again that the right-hand side (RHS) can be completely estimated from observed data. Under these assumptions, the optimization problem at stage 1, among all potential regimes $\mathcal{G}_1$, can be written as $g_1^{opt} = \arg \max_{g_1 \in \mathcal{G}_1}$ RHS of Equation (2).

We maximize the stage 1 outcome through a pseudo-outcome defined as:

$$\widetilde{PO}_1 = q_1 S_1 + \{1 - \Phi(H_1)\}\{S_1 - q_1 S_1\} + \max_{a_2} Q_2(\bar{\mathbf{W}}_2, \bar{\mathbf{X}}_2, a_2; \hat{\beta}_2, \hat{\psi}_2)$$

Our proposed estimator for $(\hat{\beta}_1, \hat{\psi}_1)$ is $\arg \min_{\beta_1, \psi_1} \mathbb{P}_n \left(\widetilde{PO}_1 - Q_1(\bar{\mathbf{X}}_1, \bar{\mathbf{W}}_1, A_1; \beta_1, \psi_1) \right)^2$, where $Q_1(\bar{\mathbf{X}}_1, \bar{\mathbf{W}}_1, A_1; \beta_1, \psi_1)$ can be modeled with $\beta_1^T \mathbf{Z}_{10} + (\psi_1^T \mathbf{Z}_{11}) A_1$. The first stage

estimated optimal rule is given by $\hat{g}_1^{opt} = \operatorname{argmax}_{a_1} Q_1(\bar{\mathbf{X}}_1, \bar{\mathbf{W}}_1, a_1; \hat{\beta}_1, \hat{\psi}_1) = \operatorname{sign}(\hat{\psi}_1^T \mathbf{Z}_{11})$, where $\mathbf{Z}_{11}$ represent the tailoring variables in the stage 1 model.

3.3. Inference

In addition to obtaining the optimal stage-specific decision rules, our second goal is to draw statistical inference on the effect of each stage's covariates on the decision. Particular emphasis was placed on tailoring variables. To that end, we propose using a censoring-adjusted version of the m -out-of- n method presented by [31]. In stage 1 optimization, the pseudo-outcome $\widetilde{PO}_1 = q_1 S_1 + \{1 - \Phi(H_1)\} \{S_1 - q_1 S_1\} + \hat{\beta}_2^T \mathbf{Z}_{20} + |\hat{\psi}_2^T \mathbf{Z}_{21}|$ is a nonsmooth function of $\hat{\psi}_2$. In particular, if $P[\mathbf{Z}_{21} : \hat{\psi}_2^T \mathbf{Z}_{21} = 0] = 0$, then first stage covariates will converge to a normal distribution. However, if $P[\mathbf{Z}_{21} : \hat{\psi}_2^T \mathbf{Z}_{21} = 0] > 0$, the estimator $\hat{\psi}_1$ oscillates between the two asymptotic distributions across samples, which reflects the typical challenging problem of nonregularity in DTR literature [29]. Hence, direct estimation results in an asymptotically biased estimator and poor performance of usual Wald-type confidence intervals. Even bootstrap-based approaches suffer from underlying nonsmoothness. The m -out-of- n bootstrap was developed to address the bootstrap inconsistency due to nonsmoothness [32]. Although conceptually very similar to the original bootstrap, the resample size m (which needs to depend on n , tends to infinity with n , and $m = o(n)$ is selected to be a smaller order than n). Chakraborty et al. 2013 [31] showed through simulation studies that the m -out-of- n approach obtained desirable coverage probabilities in the two-stage DTR problem for first stage tailoring variables. Because censoring reduces the size of observed stage 2 data in our scenario, we further modified the m -out-of- n algorithm to accommodate censoring. Our algorithm works as follows.

We adopted the functional form of m as presented in [31]:

$$m = n^{\frac{1+\xi(1-\hat{p})}{1+\xi}}. \quad (3)$$

Let n be the total number of subjects in the dataset (including those who were censored). For stage 2, we create a bootstrap sample of size n and fit a regression model using the complete data weighted by IPC-weights within the bootstrapped sample to obtain stage 2 coefficient estimates. Stage 2 95% confidence intervals are obtained after getting $\hat{l}_2$ and $\hat{u}_2$, the $\alpha/2 \times 100$ and $(1 - \alpha/2) \times 100$ percentiles of $\sqrt{n}(\hat{\theta}_{2,n}^{(b)} - \hat{\theta}_{2,n})$, where α is the desired significance level, $\hat{\theta}_{2,n}^{(b)}$ is the bootstrap estimate of stage 2 coefficients with bootstrap-specific re-estimated censoring weights, and $\hat{\theta}_{2,n}$ is the plug-in estimator obtained using weighted regression from the empirical dataset. The confidence interval is given by $(\hat{\theta}_{2,n} - \hat{u}_2/\sqrt{n}, \hat{\theta}_{2,n} - \hat{l}_2/\sqrt{n})$. For stage 1, we first generate bootstrap samples of size m , which is calculated using Equation (3) after calculating a sample specific $\hat{p}$. We further use each bootstrap sample to re-estimate IPC weights, and fit a weighted lm model to obtain a bootstrap-specific stage 2 estimate. The stage 2 coefficients from each bootstrap sample are then used to calculate pseudo-outcomes, which are then used to fit a stage 1 model to obtain $\hat{\theta}_{1,\hat{m}}^{(b)}$. Then we obtain $\hat{l}_1$ and $\hat{u}_1$, the $\alpha/2 \times 100$ and $(1 - \alpha/2) \times 100$ percentiles of $\sqrt{\hat{m}}(\hat{\theta}_{1,\hat{m}}^{(b)} - \hat{\theta}_{1,n})$, where $\hat{\theta}_{1,n}$ is the plug-in estimator obtained using the complete empirical dataset, while $\hat{\theta}_{1,\hat{m}}^{(b)}$ is the estimate obtained from each bootstrap sample of size m . The confidence set is given by $(\hat{\theta}_{1,n} - \hat{u}_1/\sqrt{n}, \hat{\theta}_{1,n} - \hat{l}_1/\sqrt{n})$.

We further selected the value of ξ to be 0.01, which provided stable coverage in simulations with complete data. The calculation of $\hat{p} = \mathbb{P}\{n[\mathbf{Z}_{21}^T \hat{\psi}_{2,n}]^2 \leq \tau_n(\mathbf{Z}_{21})\}$ relies on a selection of $\tau_n(\mathbf{Z}_{21})$. We opted to use the plug-in estimator for $\tau_n(\mathbf{Z}_{21}) = (\mathbf{Z}_{21}^T \hat{\Sigma}_{21} \mathbf{Z}_{21}) \cdot \chi_{1,1-\nu}^2$ just as in [31], where $\hat{\Sigma}_{21}$ is the plug-in sandwich estimator of $n \operatorname{Cov}(\hat{\psi}_{2,n}, \hat{\psi}_{2,n})$, and $\nu = 0.01$.

In practice, however, we need to tune the two hyper parameters ξ and ν to obtain an appropriate m using double bootstrap, and tune the bootstrap sample number m straightforwardly and automatically. Such a double bootstrap algorithm for choosing m

is data-driven. Suppose we are interested in $c^T \theta$, i.e., the stage 1 variable effect, and its estimate from the original data is $c^T \hat{\theta}_n$. Consider a grid of candidate values for m :

(1) Draw B_1 n -out-of- n first-stage bootstrap samples from the data and calculate the bootstrap estimate $c^T \hat{\theta}_n^{b_1}$, ($b_1 = 1, 2, \dots, B_1$). Fix m at the smallest value in the grid.

(2) Conditional on each first-stage bootstrap sample, draw B_2 m -out-of- n second-stage (nested) bootstrap samples and calculate the double bootstrap versions of the estimate $c^T \hat{\theta}_{n,m}^{b_1 b_2}$, $b_1 = 1, 2, \dots, B_1$, $b_2 = 1, 2, \dots, B_2$.

(3) For $b_1 = 1, 2, \dots, B_1$, compute the $(\alpha/2) \times 100$ and $(1 - \alpha/2) \times 100$ percentiles of $c^T \sqrt{m}(\hat{\theta}_{n,m}^{b_1 b_2} - \hat{\theta}_n^{b_1})$, $b_2 = 1, \dots, B_2$; say $\hat{l}_{DB}^{(b_1)}$ and $\hat{u}_{DB}^{(b_1)}$ respectively. Construct the double-centered percentile bootstrap from the b_1 -th first-stage bootstrap data as $(c^T \hat{\theta}_n^{b_1} - \hat{u}_{DB}^{(b_1)} / \sqrt{n}, c^T \hat{\theta}_n^{b_1} - \hat{l}_{DB}^{(b_1)} / \sqrt{n})$, $b_1 = 1, \dots, B_1$.

(4) Estimate the coverage rate of the double bootstrap CI from all of the first-stage bootstrap datasets as:

$$\frac{1}{B_1} \sum_{b_1=1}^{B_1} I\{c^T \hat{\theta}_n^{b_1} - \hat{u}_{DB}^{(b_1)} / \sqrt{n} \leq c^T \hat{\theta}_n \leq c^T \hat{\theta}_n^{b_1} - \hat{l}_{DB}^{(b_1)} / \sqrt{n}\}.$$

(5) If the current coverage rate is located in $[(1 - \alpha) - 1.96 \times \alpha(1 - \alpha) / \sqrt{B_1}, (1 - \alpha) + 1.96 \times \alpha(1 - \alpha) / \sqrt{B_1}]$, then it is not significantly biased from $1 - \alpha$, and we will pick the current value of m as the final value. Otherwise, increase m to the next value in the grid.

(6) Repeat steps (1)–(5), until (5) is satisfied or the grid is exhausted.

3.4. Survival Augmented Patient Preference Weights (SAPP-Weights)

Going forward, we assume that information in the latest stage survey will override information from previous stages as well as other covariate information (i.e., $\mathbf{W}_j$ will override $\mathbf{W}_{j-1}$ and $\bar{\mathbf{X}}_j$). To model survey information as a function of latent preference, we assume a latent traits model [33]. We further assume that the latent preferences are related to survey responses through a modified Rasch model [34,35].

For our scenario, we assume that there are $numQ$ questions on a survey, each soliciting binary answer choices from the patient. For each binary response, we assume that the underlying generating mechanism is of the form $\text{logit}\{P(W_{jk} = 1 | H_j = h_j)\} = \alpha_{0,k} + \alpha_{1,k} h_j$, where j indicates the stage and k the question number.

Algorithm 1 outlines the algorithm for estimating patient preference $\hat{h}_j$. Then the survival augmented patient preference weights (SAPP-weights) are the transformed version of $\hat{h}_j$, which we denote as $\Phi(\hat{h}_{ij})$. Essentially, we use the Expectation-Maximization algorithm [36] to iterate between estimates of α , the questionnaire coefficients, and h_j , individual patient preferences at stage j . We use Gauss–Hermite quadrature to numerically approximate the integral, and estimate $P(h_j | \mathbf{W}_j) \propto P(\mathbf{W}_j | h_j) P(h_j)$ using the Metropolis Hastings algorithm.

Algorithm 1: EM algorithm for estimating patient preference $\hat{h}_j$.

Result: Obtain $p(h_j | \mathbf{w}_j)$ for patient i

Set an initial value of h_j for all subjects to estimate an initial guess of α_0, α_1 ;

while not reached convergence **do**

 Using MH, get an updated estimate of $P(h_j | \mathbf{w}_j)$;

 Approximate likelihood integral using Gauss–Hermite quadrature with k abscissae h_t and weight $p(h_t)$;

 Solve likelihood equations using Newton–Raphson to obtain updated estimates of α_0, α_1 ;

end

4. Numerical Results

We conducted simulation studies to investigate the performance of our proposed method. We looked at two scenarios, differing by degree of nonregularity (the estimated probability that stage 2 treatment does not provide a significant difference), p . The first scenario is an example of low nonregularity, where about 25% of people obtain similar results with both treatments. Scenario 2 is an example of higher nonregularity, where approximately 75% of patients could obtain similar results with either of the stage 2 treatments. The true value of p was estimated through simulated complete data (assuming no censoring) and true preferences.

For both scenarios, we generated baseline covariates $X_1, X_2 \sim N(0, 1)$, censoring time C , quality of life q_1, q_2 , and survival times S_1, S_2 . Preferences of both stages were generated from $N(0, SD = 0.5)$ distribution, and 10 binary preference derived questionnaire responses W_1, W_2 were generated according to our latent model with coefficients, as shown in Table 1.

Table 1. Coefficients for the latent model used to solicit preferences. α_{10} and α_{11} are the coefficients for stage 1 questionnaires, while α_{20} and α_{21} are for stage 2 coefficients.

	α_{10}	α_{11}	α_{20}	α_{21}
Q1	1.00	1.00	0.90	0.72
Q2	0.00	−1.00	−0.19	1.82
Q3	−1.64	2.35	1.47	−1.63
Q4	0.54	−2.35	−0.50	2.11
Q5	−0.88	−1.10	1.30	1.69
Q6	0.75	1.25	−0.40	−1.69
Q7	1.27	2.96	0.04	2.37
Q8	−1.50	2.00	−1.27	1.60
Q9	0.09	−1.50	0.62	−1.17
Q10	−0.55	1.35	−0.23	2.00

The two scenarios differ in terms of stage 2 parameters but share common stage 1 settings. Stage 1 treatment assignment A_1 was randomly assigned with probability 0.5. The stage 1 quality of life outcome $q_1 \in [0, 1]$ was generated from $N(\alpha_0 + \alpha_1 X_1 + \alpha_2 A_1 + \alpha_3 X_1 A_1, \sigma^2)$, where $\alpha = [0.55, 0.03, 0.06, -0.09]$, and $\sigma = 0.03$. As mentioned previously, S_1 for everyone indicates a routine follow-up time of 30 days. The outcome of stage 1 is a weighted combination of q_1 and S_1 through the equation $R_1 = q_1 S_1 + (1 - \Phi(h_1))(S_1 - q_1 S_1)$, where $\Phi(h)$ represents the cumulative distributive function of a standard normal. R_1 can be interpreted as an quality of life weighted survival during the initial follow-up time. The treatment assignment A_2 depends on R_1 : $Bernoulli(\{\exp(-3 + R_1/3)\} / \{1 + \exp(-3 + R_1/3)\})$, and thus those patients with higher R_1 are more likely to remain on the same treatment as A_1 .

Using these simulated datasets, we estimated patient preferences, the optimal dynamic treatment regime, and the confidence intervals of estimated DTR coefficients from these responses and outcomes. We evaluated each scenario through measures of bias, coverage probability, optimal mean response, and the percent of subjects correctly classified to their true optimal treatment %opt.

4.1. Scenario 1: Low Nonregularity ($p = 0.25$)

We generated stage 2 outcome $q_2 \sim N(\beta_0 + \beta_1 X_2 + \beta_2 X_2 A_2, SD = 0.03)$, where $\beta = [0.5, 0.07, 0.06]$. Similarly, we generated $S_2 \sim N(\gamma_0 + \gamma_1 R_1 + \gamma_2 (R_1 - c) A_2, SD = 5)$, where $\gamma = [50, 10, -2.5]$ and $c = 20$. q_2 and S_2 disagreed on the optimal stage 2 treatment approximately half the time, indicating that half of all patients had to make a choice between quality and quantity of life. In a randomly generated dataset, treatment 1 gave 73.4% of patients a better stage 1 reward, while treatment 1 gave 31% of patients a better stage 2 reward. The range of S_2 varied from 169 to 361 days.

Unlike stage 1, stage 2 survival times could be subject to censoring time C . We generated censoring $C \sim \exp(\log \lambda_0^C + X_2 \beta_C)$, where $\beta_C = 0.01$ and $\lambda_0^C = 0.00058$ for 15% censoring and $\lambda_0^C = 0.0013$ for 30% censoring. τ was set to be a year after initiation of stage 2 treatment.

Because S_2 and q_2 are functions of X_2 , A_2 , and R_1 , the reward combination can be rearranged as:

$$\begin{aligned} R_2 &= q_2 S_2 + \{1 - \Phi(h_2)\}(S_2 - q_2 S_2) \\ &= \gamma_0 + (\beta_0 \gamma_0 - \gamma_0) \Phi(h_2) \\ &\quad + \{\gamma_1 + (\beta_0 \gamma_1 - \gamma_1) \Phi(h_2)\} R_1 \\ &\quad + \{(\beta_1 \gamma_0 - \beta_2 \gamma_2 c) \Phi(h_2)\} X_2 \\ &\quad + \{(\beta_1 \gamma_1 + \beta_2 \gamma_2) \Phi(h_2)\} X_2 R_1 \\ &\quad + \{(\gamma_2 c - \beta_0 \gamma_2 c) \Phi(h_2) - \gamma_2 c\} A_2 \\ &\quad + \{\gamma_2 + (\beta_0 \gamma_2 - \gamma_2) \Phi(h_2)\} R_1 A_2 \\ &\quad + \{(\beta_1 \gamma_2 + \beta_2 \gamma_1) \Phi(h_2)\} R_1 X_2 A_2 \\ &\quad + \{(\beta_2 \gamma_0 - \beta_1 \gamma_2 c) \Phi(h_2)\} X_2 A_2. \end{aligned}$$

On the other hand, stage 1 coefficients are obtained by regressing a pseudo-outcome, $R_1 + \hat{\beta}_2^T \mathbf{Z}_{20} + |\hat{\psi}_2^T \mathbf{Z}_{21}|$ on X_1 , A_1 , and $X_1 A_1$, so stage 1 regression has four covariate terms (including intercept). We obtained both true stage 1 and stage 2 coefficients by performing Monte Carlo sampling regressions on a sample size of 10 million.

Tables 2 and 3 summarize the simulation results of the two stages for scenario 1, including bias estimates, the empirical standard deviations, the mean bootstrap standard deviations, mean widths of confidence interval, and the coverage probabilities of our proposed method. The coverage probabilities ranged from 86% to 96%, with the majority between 92% to 96%. Furthermore, we observed a general agreement between the empirical SD and the average bootstrap SD. In terms of trends, we saw slightly larger ESD, mean bootstrap SD, and mean width when censoring was increased from 15% to 30%, while a reduction in all three was observed when we increased n from 500 to 1000. Using covariate A_2 as an example, the ESD for $n = 500$ and 15% censoring was 11.72, which increased to 13.03 when censoring reached 30% but decreased to 8.38 when n was increased to 1000. Similar patterns were observed with the average bootstrap SD and mean width of the confidence interval.

Table 2. Stage 2 simulation results for scenario 1 ($p = 0.25$). True parameters are [7.49, 3.23, 37.58, 0.28, −1.88, 3.21, 0.21].

n	% Censor	Parameter	Bias	ESD	AvgBootSD	MeanWidth	CP
500	15	R_1	−0.32	0.48	0.50	1.94	0.92
		X_2	0.16	11.94	12.18	47.75	0.96
		A_2	−4.93	11.72	11.57	45.17	0.92
		$R_1 : X_2$	−0.01	0.52	0.52	2.04	0.96
		$R_1 : A_2$	0.22	0.50	0.50	1.94	0.92
		$X_2 : A_2$	0.38	12.51	12.20	47.68	0.93
		$R_1 : X_2 : A_2$	−0.01	0.54	0.52	2.04	0.94
	30	R_1	−0.31	0.54	0.55	2.15	0.93
		X_2	0.38	13.39	13.57	53.19	0.95
		A_2	−4.83	13.03	12.80	50.05	0.92
		$R_1 : X_2$	−0.02	0.58	0.58	2.28	0.94
		$R_1 : A_2$	0.21	0.56	0.55	2.15	0.93
		$X_2 : A_2$	0.06	13.56	13.57	53.12	0.95
		$R_1 : X_2 : A_2$	0.00	0.58	0.58	2.28	0.95

Table 2. Cont.

n	% Censor	Parameter	Bias	ESD	AvgBootSD	MeanWidth	CP
1000	15	R_1	−0.29	0.36	0.34	1.34	0.86
		X_2	−0.14	8.23	8.34	32.64	0.95
		A_2	−4.73	8.38	8.01	31.26	0.88
		$R_1 : X_2$	0.01	0.35	0.36	1.40	0.95
		$R_1 : A_2$	0.21	0.36	0.34	1.34	0.88
		$X_2 : A_2$	0.10	8.55	8.34	32.59	0.94
		$R_1 : X_2 : A_2$	0.00	0.37	0.36	1.39	0.94
	30	R_1	−0.28	0.40	0.38	1.49	0.87
		X_2	0.04	9.38	9.25	36.17	0.96
		A_2	−4.94	8.95	8.85	34.58	0.90
		$R_1 : X_2$	0.00	0.40	0.40	1.55	0.96
		$R_1 : A_2$	0.22	0.39	0.38	1.48	0.89
		$X_2 : A_2$	0.44	9.67	9.24	36.11	0.94
		$R_1 : X_2 : A_2$	−0.02	0.41	0.40	1.55	0.94

Table 3. Stage 1 simulation results for scenario 1 ($p = 0.25$). True parameters are $[3.77, 8.17, -2.37]$.

n	% Censor	Parameter	Bias	ESD	AvgBootSD	MeanWidth	CP
500	15	X_1	−0.10	0.93	0.98	3.84	0.95
		A_1	−0.30	1.05	1.07	4.18	0.93
		$X_1 : A_1$	0.41	1.21	1.23	4.78	0.94
	30	X_1	−0.10	0.93	1.00	3.93	0.96
		A_1	−0.29	1.07	1.11	4.34	0.94
		$X_1 : A_1$	0.40	1.28	1.29	5.05	0.94
1000	15	X_1	−0.04	0.67	0.68	2.66	0.96
		A_1	−0.20	0.73	0.74	2.90	0.94
		$X_1 : A_1$	0.31	0.85	0.85	3.32	0.92
	30	X_1	−0.04	0.68	0.69	2.71	0.95
		A_1	−0.21	0.75	0.77	2.99	0.94
		$X_1 : A_1$	0.31	0.88	0.90	3.49	0.94

We further investigated the distributions of the observed total reward, as well as the predicted optimal reward of one randomly selected simulation, which we illustrate in Figure 2. Aggregating means across the four sub-scenarios (based on sample size and censoring), the average observed reward was 233.77, while the average predicted optimal reward was 254.19, indicating an expected increase of 20.42 reward when everyone follows the regime to receive the treatments by our algorithm. It is also evident in the figure that the variability of the observed rewards is significantly larger than the variability of the predicted optimal reward. The aggregated SD of all observed rewards in this scenario is 39.2, while the aggregated SD of all predicted rewards is 5.66.

Figure 2. Comparison of observed reward and predicted optimal reward for scenario 1 ($p = 0.25$).

4.2. Scenario 2: High Nonregularity ($p = 0.75$)

In this scenario, we evaluated a case where high nonregularity exists. We modified the scenario 1 simulation by the following: (1) We generated the stage 2 outcome $q_2 \sim N(\beta_0 + \beta_1 X_2 + \beta_2 X_2 A_2, SD = 0.03)$, where $\beta = [0.5, 0.07, 0.021]$, and $S_2 \sim N(\gamma_0 + \gamma_1 R_1 + \gamma_2 (R_1 - c) A_2, SD = 5)$, where $\gamma = [135, 8, -0.5]$ and $c = 20$. (2) We adjusted the coefficients such that the magnitude of γ_2 and β_2 (which influence the effect of A_2) are smaller compared to scenario 1. (3) q_2 and S_2 also disagreed on the optimal stage 2 treatment approximately half the time.

In a randomly generated dataset, treatment 1 gave 74.8% of patients a better stage 1 reward, while treatment 1 gave 41.6% of patients a better stage 2 reward. The range of S_2 varied from 235 to 368 days.

Stage 2 survival times are similarly subject to censoring time C . Censoring was generated with model $C \sim \exp(\log \lambda_0^C + X_2 \beta_C)$, where $\beta_C = 0.01$ and $\lambda_0^C = 0.00048$ for 15% censoring and $\lambda_0^C = 0.0011$ for 30% censoring. As before, τ was set to be a year after initiation of stage 2 treatment.

Tables 4 and 5 present our simulation results for scenario 2. As before, the coverage probabilities range from 86% to 97%, with most covering around 95%, and the mean width, SD, and mean bootstrap SD are decreasing with decreasing censoring and increasing sample size. Again, using A_2 as an example, the ESD across simulations of its coefficient was 12.38 when $n = 500$ and censoring at 15%, which increased to 13.85 when censoring went to 30%, and decreased to 8.84 when n went to 1000.

Table 4. Stage 2 simulation results for scenario 2 ($p = 0.75$). True parameters are [6.00, 4.74, 7.53, 0.28, −0.38, 1.66, 0.07].

n	% Censor	Parameter	Bias	ESD	AvgBoot SD	MeanWidth	CP
500	15	R_1	−0.33	0.51	0.52	2.05	0.92
		X_2	0.31	12.42	12.83	50.29	0.97
		A_2	−3.47	12.38	12.27	47.86	0.93
		$R_1 : X_2$	−0.01	0.54	0.55	2.14	0.96
		$R_1 : A_2$	0.16	0.53	0.52	2.05	0.94
		$X_2 : A_2$	0.17	12.99	12.84	50.18	0.95
		$R_1 : X_2 : A_2$	−0.01	0.55	0.55	2.14	0.95
	30	R_1	−0.31	0.56	0.58	2.27	0.93
		X_2	0.41	13.92	14.32	56.06	0.95
		A_2	−3.35	13.85	13.60	53.13	0.92
		$R_1 : X_2$	−0.02	0.60	0.61	2.39	0.95
		$R_1 : A_2$	0.15	0.59	0.58	2.28	0.91
		$X_2 : A_2$	−0.08	14.13	14.31	55.97	0.94
		$R_1 : X_2 : A_2$	0.00	0.60	0.61	2.39	0.95
1000	15	R_1	−0.30	0.39	0.36	1.42	0.86
		X_2	−0.10	8.69	8.78	34.34	0.95
		A_2	−3.52	8.84	8.51	33.24	0.91
		$R_1 : X_2$	0.00	0.37	0.37	1.47	0.96
		$R_1 : A_2$	0.16	0.38	0.36	1.42	0.90
		$X_2 : A_2$	0.08	8.99	8.78	34.27	0.94
		$R_1 : X_2 : A_2$	0.00	0.38	0.37	1.46	0.94
	30	R_1	−0.29	0.42	0.40	1.57	0.88
		X_2	0.24	9.79	9.75	38.13	0.95
		A_2	−3.76	9.46	9.40	36.69	0.93
		$R_1 : X_2$	−0.01	0.42	0.42	1.63	0.94
		$R_1 : A_2$	0.17	0.41	0.40	1.57	0.93
		$X_2 : A_2$	0.37	10.25	9.73	38.04	0.93
		$R_1 : X_2 : A_2$	−0.02	0.44	0.42	1.62	0.94

Table 5. Stage 1 simulation results for scenario 2 ($p = 0.75$). True parameters are [3.19, 6.39, −9.61].

n	% Censor	Parameter	Bias	ESD	AvgBootSD	MeanWidth	CP
500	15	X_1	−0.33	0.83	0.91	3.54	0.94
		A_1	−0.42	0.95	1.01	3.94	0.94
		$X_1 : A_1$	0.61	1.14	1.19	4.63	0.92
	30	X_1	−0.35	0.83	0.93	3.62	0.94
		A_1	−0.42	0.99	1.05	4.11	0.94
		$X_1 : A_1$	0.60	1.20	1.26	4.91	0.93
1000	15	X_1	−0.22	0.61	0.63	2.46	0.95
		A_1	−0.33	0.68	0.70	2.75	0.93
		$X_1 : A_1$	0.48	0.80	0.83	3.25	0.92
	30	X_1	−0.23	0.62	0.64	2.50	0.94
		A_1	−0.34	0.69	0.73	2.86	0.93
		$X_1 : A_1$	0.49	0.84	0.88	3.43	0.92

As in Section 4.1, we investigated the distribution of the observed total reward and the predicted optimal reward of one random simulation in Figure 3. For aggregated means across the four sub-scenarios (based on sample size and censoring), the average observed reward was 263.62, while the average predicted reward was 275.35, indicating an increase

of 11.73 reward when everyone follows the regime to receive treatment by our algorithm. The aggregated SD of observed rewards was 37.77, while the aggregated SD of all predicted rewards was 4.39.

Figure 3. Comparison of observed reward and predicted optimal reward for scenario 2 ($p = 0.75$).

4.3. Optimality

Besides looking at performance of inference, we also looked at the number of times our algorithm chose the correct treatment for each patient at each stage, across the scenarios we have visited ($n = 500, 1000$ across two levels of censoring at 15%, 30%).

Table 6 shows the simulation results. In this table we also included the average stage 1 bootstrap resample sizes for each sub-scenario. As expected from Equation (3), increasing p indicates increasing nonregularity, and decreases m . We can see that for scenario 1 ($p = 0.25$), the algorithm chose the optimal treatment for stage 1 over 93% of the time, and over 83% of the time for stage 2. Our algorithm was able to assign the correct treatment to a patient over 78% across all sample sizes and censoring levels. This is significantly higher than the random guess approach, which would have landed us around 25%.

Table 6. Simulation results for percent optimal treatment chosen.

Scenario	n	% Censor	Stage 1 %opt (sd)	Stage 2 %opt (sd)	Overall %opt (sd)
$p = 0.25$	500	15	0.936 (0.012)	0.839 (0.020)	0.785 (0.022)
	500	30	0.936 (0.012)	0.837 (0.022)	0.783 (0.023)
	1000	15	0.939 (0.008)	0.844 (0.012)	0.793 (0.013)
	1000	30	0.939 (0.008)	0.843 (0.013)	0.792 (0.014)
$p = 0.75$	500	15	0.935 (0.014)	0.667 (0.051)	0.623 (0.050)
	500	30	0.934 (0.014)	0.662 (0.053)	0.618 (0.051)
	1000	15	0.938 (0.009)	0.687 (0.029)	0.644 (0.028)
	1000	30	0.938 (0.009)	0.682 (0.033)	0.640 (0.032)

By contrast, scenario 2's performance was slightly weaker, coming in at over 93% for stage 1, around 66–69% for stage 2, and with an overall correct regime assignment percentage of 61–65%; however, this is still much better than 25% with a random guess. In this scenario, we could also see a general increase in SD as compared to scenario 1, indicating higher uncertainty in our decision making process.

5. Discussion

We proposed a method that estimates the optimal decision rules for a two-stage treatment scenario subject to censoring in this paper. We proposed to balance between quality of life and quantity using a sliding scale function adjusted by patient preference. Through simulation studies, we have shown that our proposed method is capable of choosing the optimal treatment dynamically a majority of the time, as well as provide convincing confidence intervals for each of the coefficients in question.

The simulation results of scenarios 1 and 2 are notable in the following ways. Most importantly, we can see that the coverage probabilities mostly hover around 95%, showing that our confidence interval has the combination of adequate width and minimal bias required for a good coverage probability. The general congruence between the empirical SE and the mean bootstrap SE further supports that our method is sampling at the appropriate width. Generally, there is a slight increase in ESD and mean bootstrap SD when increasing censoring from 15% to 30%, indicating decreased certainty, but the difference is slight (e.g., for parameter A_2 in scenario 1, ESD increased from 11.72 to 13.03). The decrease in ESD is more significant between $n = 500$ and $n = 1000$, where for scenario 1 and 15% censoring, the ESD of the estimated coefficient for parameter A_2 decreased from 11.72 to 8.38. In the boxplots in Figures 2 and 3, we can see much larger variability in the observed rewards as compared to the predicted rewards. This is as expected for two reasons. First, observed rewards contain an error component that is not present in expected (predicted) rewards. Second, observed rewards include individuals who have by chance obtained their optimal reward, as well as those who did not. Variability invariably reduces when more individuals are predicted to their optimal reward. As expected, the distribution of the predicted optimal rewards can be mapped to the upper part of the observed rewards. The same general patterns were observed for scenario 2.

One main challenge in this work, as is the case in [31], is the selection of m , which is a crucial factor in determining the coverage probability and confidence interval. In our simulations, we selected parameters ν and ζ using background knowledge and reference simulation results and used Equation (3) to select m . This approach was recommended by [31] and is straightforward and easy to implement, but risks inappropriate values of m if either ν or ζ are selected inadequately. Hence, we recommend the double bootstrap procedure in practice, where we take our empirical dataset estimator as the truth, and build confidence intervals using nested bootstrap samples of size m from empirical bootstrap samples of size n . Similarly, another potential idea to further improve coverage across all covariates could be to select distinct values of m for each covariate. This might be of interest for future research. Alternatively, one may consider coupling our method with other meth-

ods that address the nonregularity issue in dynamic treatment regimes, notably [37]. Two main challenges exist for this proposed integration, namely the adaptability to censored data and derivation of asymptotic results in censored settings, and the specific process of choosing a tuning parameter λ for our proposed data structure. The integration would be an interesting and nontrivial problem for future research.

The authors are aware of two works in the literature that balance between two outcomes and would like to highlight certain differences at this time. Zhao et al. 2009 [15] looked at the dosage effect of a cancer drug on tumor size and toxicity. In their work, each reward function is separated into three parts: survival status, wellness, and tumor size effects. In simulation studies, the reward was assigned to be -60 if the patient died, 15 if the patient's tumor shrunk to zero, and 5 or -5 if tumor size/wellness improved and deteriorated. The method of assigning rewards is particular to the example they proposed, and there is no clear indication of how to perform inference using this method. The comparison between our method and [14] is a bit more direct. Their work used patient preference estimation to weigh between toxicity and efficacy, both continuous outcomes. There are definite similarities between this work and [14], especially in terms of preference estimation and the common Q-learning framework. However, most importantly, Butler et al. 2018 [14] would not accommodate censored data, nor did they provide any framework for inference, both of which are important contributions of this work.

This work can be improved in a couple of directions for increased generalizability and impact. First, the vast amount of methods for survival data is one indication of the complexity of generalizing the various time to event scenarios. One main direction for extension would be to accommodate multiple stages, with potential censoring and event times that could happen at any stage, similar to the set-up of [20]. Another direction that might benefit real applications could be allowing the subset of treatments to change depending on patient outcomes, as in [21]. As pointed out by a reviewer, there may be other ways of estimating and incorporating patient preference. Before settling on our proposed approach, we had considered hierarchical models, which condition directly on latent variables. However, this approach would lose causal interpretation in multiple stage settings. Even though hierarchical models were not suitable for multi-stage settings, exploring other models for calculating and estimating preference variables, e.g., generalizing from logit to other exponential families within the latent variable framework as in [33], would still be valuable. Finally, generalizing binary treatment options to a continuous version (i.e., dosage), or increasing the number of outcomes we balance are also meaningful directions for future extensions.

Author Contributions: Conceptualization, L.W.; Formal analysis, Y.Z., and C.W.; Investigation, Y.Z. and L.W.; Methodology, Y.Z. and L.W.; Supervision, L.W.; Validation, C.W.; Visualization, Y.Z.; Writing—original draft, Y.Z. and L.W.; Writing—review & editing, Y.Z., C.W. and L.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Robins, J.M. Marginal structural models versus structural nested models as tools for causal inference. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials*; Springer: Berlin/Heidelberg, Germany, 2000; pp. 95–133.
2. Robins, J.M. Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium in Biostatistics*; Springer: New York, NY, USA, 2004; pp. 189–326.
3. Murphy, S.A. Optimal dynamic treatment regimes. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2003**, *65*, 331–355. [[CrossRef](#)]
4. Murphy, S.A. An experimental design for the development of adaptive treatment strategies. *Stat. Med.* **2005**, *24*, 1455–1481. [[CrossRef](#)]

5. Moodie, E.E.; Chakraborty, B.; Kramer, M.S. Q-learning for estimating optimal dynamic treatment rules from observational data. *Can. J. Stat.* **2012**, *40*, 629–645. [[CrossRef](#)]
6. Schulte, P.J.; Tsiatis, A.A.; Laber, E.B.; Davidian, M. Q-and A-learning methods for estimating optimal dynamic treatment regimes. *Stat. Sci.* **2014**, *29*, 640. [[CrossRef](#)] [[PubMed](#)]
7. Laber, E.; Zhao, Y. Tree-based methods for individualized treatment regimes. *Biometrika* **2015**, *102*, 501–514. [[CrossRef](#)] [[PubMed](#)]
8. Tao, Y.; Wang, L. Adaptive contrast weighted learning for multi-stage multi-treatment decision-making. *Biometrics* **2017**, *73*, 145–155. [[CrossRef](#)] [[PubMed](#)]
9. Tao, Y.; Wang, L.; Almirall, D. Tree-based reinforcement learning for estimating optimal dynamic treatment regimes. *Ann. Appl. Stat.* **2018**, *12*, 1914. [[CrossRef](#)]
10. Zhang, Y.; Laber, E.B.; Davidian, M.; Tsiatis, A.A. Interpretable dynamic treatment regimes. *J. Am. Stat. Assoc.* **2018**, *113*, 1541–1549. [[CrossRef](#)]
11. Zhao, Y.; Zeng, D.; Rush, A.J.; Kosorok, M.R. Estimating individualized treatment rules using outcome weighted learning. *J. Am. Stat. Assoc.* **2012**, *107*, 1106–1118. [[CrossRef](#)] [[PubMed](#)]
12. Zhang, B.; Zhang, M. C-learning: A new classification framework to estimate optimal dynamic treatment regimes. *Biometrics* **2018**, *74*, 891–899. [[CrossRef](#)]
13. Sun, Y.; Wang, L. Stochastic Tree Search for Estimating Optimal Dynamic Treatment Regimes. *J. Am. Stat. Assoc.* **2021**, *116*, 421–432. [[CrossRef](#)]
14. Butler, E.L.; Laber, E.B.; Davis, S.M.; Kosorok, M.R. Incorporating patient preferences into estimation of optimal individualized treatment rules. *Biometrics* **2018**, *74*, 18–26. [[CrossRef](#)]
15. Zhao, Y.; Kosorok, M.R.; Zeng, D. Reinforcement learning design for cancer clinical trials. *Stat. Med.* **2009**, *28*, 3294–3315. [[CrossRef](#)]
16. Torrance, G.W.; Feeny, D. Utilities and quality-adjusted life years. *Int. J. Technol. Assess. Health Care* **1989**, *5*, 559–575. [[CrossRef](#)] [[PubMed](#)]
17. Cui, Y.; Zhu, R.; Kosorok, M. Tree based weighted learning for estimating individualized treatment rules with censored data. *Electron. J. Stat.* **2017**, *11*, 3927–3953. [[CrossRef](#)] [[PubMed](#)]
18. Zhu, R.; Kosorok, M.R. Recursively imputed survival trees. *J. Am. Stat. Assoc.* **2012**, *107*, 331–340. [[CrossRef](#)] [[PubMed](#)]
19. Zhao, Y.Q.; Zeng, D.; Laber, E.B.; Song, R.; Yuan, M.; Kosorok, M.R. Doubly robust learning for estimating individualized treatment with censored data. *Biometrika* **2015**, *102*, 151–168. [[CrossRef](#)] [[PubMed](#)]
20. Goldberg, Y.; Kosorok, M.R. Q-learning with censored data. *Ann. Stat.* **2012**, *40*, 529. [[CrossRef](#)]
21. Hager, R.; Tsiatis, A.A.; Davidian, M. Optimal two-stage dynamic treatment regimes from a classification perspective with censored survival data. *Biometrics* **2018**, *74*, 1180–1192. [[CrossRef](#)] [[PubMed](#)]
22. Zhao, Y.Q.; Zhu, R.; Chen, G.; Zheng, Y. Constructing Stabilized Dynamic Treatment Regimes. *arXiv* **2018**, arXiv:1808.01332.
23. Jiang, R.; Lu, W.; Song, R.; Davidian, M. On estimation of optimal treatment regimes for maximizing t-year survival probability. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2017**, *79*, 1165. [[CrossRef](#)] [[PubMed](#)]
24. Shen, Y.; Wang, Y.; Dagnault, J.; Spratt, M.; Morgan, T.; Taylor, R. Estimating the Optimal Personalized Treatment Strategy Based on Selected Variables to Prolong Survival via Random Survival Forest with Weighted Bootstrap. *J. Biopharm. Stat.* **2018**, *28*, 362–381. [[CrossRef](#)] [[PubMed](#)]
25. Huang, X.; Ning, J. Analysis of multi-stage treatments for recurrent diseases. *Stat. Med.* **2012**, *31*, 2805–2821. [[CrossRef](#)] [[PubMed](#)]
26. Huang, X.; Ning, J.; Wahed, A.S. Optimization of individualized dynamic treatment regimes for recurrent diseases. *Stat. Med.* **2014**, *33*, 2363–2378. [[CrossRef](#)]
27. Zhang, M.; Schaubel, D.E. Contrasting treatment-specific survival using double-robust estimators. *Stat. Med.* **2012**, *31*, 4255–4268. [[CrossRef](#)]
28. Jiang, R.; Lu, W.; Song, R.; Hudgens, M.G.; Napravnik, S. Doubly robust estimation of optimal treatment regimes for survival data—With application to an HIV/AIDS study. *Ann. Appl. Stat.* **2017**, *11*, 1763. [[CrossRef](#)]
29. Chakraborty, B.; Murphy, S.; Strecher, V. Inference for non-regular parameters in optimal dynamic treatment regimes. *Stat. Methods Med. Res.* **2010**, *19*, 317–343. [[CrossRef](#)]
30. Laber, E.B.; Lizotte, D.J.; Qian, M.; Pelham, W.E.; Murphy, S.A. Dynamic treatment regimes: Technical challenges and applications. *Electron. J. Stat.* **2014**, *8*, 1225.
31. Chakraborty, B.; Laber, E.B.; Zhao, Y. Inference for optimal dynamic treatment regimes using an adaptive m-out-of-n bootstrap scheme. *Biometrics* **2013**, *69*, 714–723. [[CrossRef](#)]
32. Shao, J. Bootstrap sample size in nonregular cases. *Proc. Am. Math. Soc.* **1994**, *122*, 1251–1262. [[CrossRef](#)]
33. Moustaki, I.; Knott, M. Generalized latent trait models. *Psychometrika* **2000**, *65*, 391–411. [[CrossRef](#)]
34. Rasch, G. On general laws and the meaning of measurement in psychology. In Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, USA, 1 January 1961; Volume 4, pp. 321–333.
35. Rasch, G. Studies in Mathematical Psychology: I. Probabilistic Models for Some Intelligence and Attainment Tests. 1960. Available online: <https://psycnet.apa.org/record/1962-07791-000> (accessed on 14 September 2021).
36. Moon, T.K. The expectation-maximization algorithm. *IEEE Signal Process. Mag.* **1996**, *13*, 47–60. [[CrossRef](#)]
37. Song, R.; Wang, W.; Zeng, D.; Kosorok, M.R. Penalized q-learning for dynamic treatment regimens. *Stat. Sin.* **2015**, *25*, 901. [[CrossRef](#)] [[PubMed](#)]