



Dynamic Treatment Regimes Using Bayesian Additive Regression Trees for Censored Outcomes

Xiao Li¹ · Brent R. Logan¹ · S. M. Ferdous Hossain² · Erica E. M. Moodie² 

Received: 15 October 2022 / Accepted: 16 July 2023 / Published online: 2 September 2023
© The Author(s) 2023

Abstract

To achieve the goal of providing the best possible care to each individual under their care, physicians need to customize treatments for individuals with the same health state, especially when treating diseases that can progress further and require additional treatments, such as cancer. Making decisions at multiple stages as a disease progresses can be formalized as a dynamic treatment regime (DTR). Most of the existing optimization approaches for estimating dynamic treatment regimes including the popular method of Q-learning were developed in a frequentist context. Recently, a general Bayesian machine learning framework that facilitates using Bayesian regression modeling to optimize DTRs has been proposed. In this article, we adapt this approach to censored outcomes using Bayesian additive regression trees (BART) for each stage under the accelerated failure time modeling framework, along with simulation studies and a real data example that compare the proposed approach with Q-learning. We also develop an R wrapper function that utilizes a standard BART survival model to optimize DTRs for censored outcomes. The wrapper function can easily be extended to accommodate any type of Bayesian machine learning model.

Keywords Accelerated failure time (AFT) · Allogeneic hematopoietic cell transplantation · Precision medicine · Individualized treatment rules · Survival analysis

✉ Erica E. M. Moodie
erica.moodie@mcgill.ca

Xiao Li
xiaoli@mcw.edu

Brent R. Logan
blogan@mcw.edu

S. M. Ferdous Hossain
ferdous.hossain@mail.mcgill.ca

¹ Division of Biostatistics, Medical College of Wisconsin, Milwaukee, WI, USA

² Biostatistics, McGill University, Montreal, QC, Canada

1 Introduction

Optimizing medical therapy often requires that the treatment be tailored to the individual initially, and that the treatment be adaptive to an individual's changing characteristics over time. Since individual responses can often be heterogeneous, it is challenging for physicians to customize treatments for individuals based on traditional clinical trial results, which lack the ability to identify subgroups that have different treatment effects and rarely consider successions of treatments. For chronic diseases that can evolve, it is even more important and difficult to choose the best therapy in sequence. To give a simple example, oncologists typically choose an initial immunosuppressant regime for patients with acute myeloid leukemia (AML) who are undergoing allogeneic hematopoietic cell transplantation (AHCT), to prevent a serious potential complication called graft-versus-host disease (GVHD). At the time that such an initial regime fails, a salvage treatment is chosen based on the patient's prior treatments and responses. Such a multi-stage treatment decision has been summarized as a dynamic treatment regime (DTR) by Murphy (2003). Each decision rule in DTR takes a patient's individual characteristics, treatment history and possible intermediate outcomes observed up to a certain stage as inputs, and outputs a recommended treatment for that stage.

A number of approaches have been proposed for estimating and optimizing DTRs, including those by Robins (2004), Moodie et al. 2007; Qian and Murphy 2011; Zhao et al. 2015; Krakow et al. 2017; Murray et al. 2018, and Simoneau et al. 2020. Two textbooks and an edited volume have been published on the topic of DTRs (Chakraborty and Moodie (2013); Kosorok and Moodie (2015); Tsiatis et al. (2020)), and a recent paper surveying value-search approaches by Jiang et al. (2019) was the subject of a lively discussion. Bayesian approaches have received relatively little attention in the DTR literature, though exceptions exist (Arjas and Saarela 2010; Saarela et al. 2015, 2016; Rodriguez Duque et al. 2022). However much of the Bayesian DTR methodology is relatively parametric. An exception to this is the Bayesian machine learning (BML) method developed by Murray et al. (2018), which innovatively bridges the gap between Bayesian inferences and dynamic programming methods from machine learning. A key advantage to a Bayesian approach to estimation is the quantification of uncertainty in decision making through the resulting posterior distribution. A second benefit that arises specifically in the BML approach is the highly flexible estimation that is employed, which minimizes the risk of estimation errors due to model mis-specification.

However, the BML method has not yet been adapted to censored outcomes, which is one of the more common types of outcomes in controlling chronic diseases. Motivated by the study of optimal therapeutic choices to prevent and treat GVHD, in this paper, we extend this approach to censored outcomes under the accelerated failure time (AFT) model framework. By modifying the data augmentation step in the BML method, the censored observation times can be imputed in an informative way so that the observed censoring time is well utilized. This extension is implemented using Bayesian additive regression trees (BART);

we accomplished the implementation of the proposed AFT-BML approach by developing an R function that utilizes standard BART survival software directly without needing to modify existing (complex) BART software directly. Parallel computing was used to speed up the computational calculations. This R wrapper function can be easily adjusted to accommodate other types of Bayesian machine learning methods.

This paper is organized as follows. In Sect. 2, we briefly review related methods, algorithms, and describe the extended AFT-BML approach for optimizing DTRs for censored outcomes under the accelerated failure time framework. Section 3 presents simulation studies to demonstrate our model performance by comparing it to estimation using Q-learning. An analysis of our motivating dataset of patients diagnosed with AML is given in Sect. 4. Finally, in Sect. 5, we discuss the advantages and disadvantages of our approach and provide some suggestions for future work.

2 Methods

2.1 Dynamic treatment regimes

A dynamic treatment regime is a series of decision rules that assign treatment based on an individual’s characteristics and history at each stage. Without loss of generality, we focus on a two-stage intervention problem. Furthermore, we start by describing DTRs in the non-survival setting, before proceeding to the censored survival setting later. Following Murray’s notation (Murray et al. (2018)), as well as convention for Bayesian notation of using lower case for observable data, let $o_1 \in \mathcal{O}_1$ be the covariates observed before Stage 1, and $a_1 \in \mathcal{A}_1$ be the action taken at Stage 1. Denote y_1 as the pay-off observed after Stage 1 and before Stage 2; $\{o_2, a_2, y_2\}$ are defined similarly for Stage 2. The total pay-off (also called the reward, or outcome) is assumed to be $y = y_1 + \eta y_2$, where η is an indicator that the individual entered Stage 2. A general diagram to present the two-stage decision making problem is

$$o_1 \longrightarrow a_1 \longrightarrow y_1 \xrightarrow{\text{if } \eta=1} o_2 \longrightarrow a_2 \longrightarrow y_2.$$

Denote the accumulated history before Stage 2 treatment as $\bar{o}_2 = (o_1, a_1, y_1, o_2) \in \bar{\mathcal{O}}_2$. In this setting, a DTR consists of two decision rules, one for each stage,

$$d_1 : \mathcal{O}_1 \rightarrow \mathcal{A}_1 \quad \text{and} \quad d_2 : \bar{\mathcal{O}}_2 \rightarrow \mathcal{A}_2.$$

Optimizing the two-stage DTR (d_1, d_2) is equivalent to finding the decision rules that maximize the expected total pay-off $E(y)$.

2.2 Bayesian machine learning for DTRs

Murray et al. (2018) described a new approach called Bayesian Machine Learning to optimize DTRs; the method requires fitting a series of Bayesian regression models in reverse sequential order under the approximate dynamic programming

framework. The authors use the potential outcomes notation to describe their approach, where $y(a_1, a_2)$ denotes the pay-off observed when action a_1 is taken at Stage 1 and action a_2 is taken at Stage 2, and other potential outcomes ($y_2(a_1, a_2)$, $y_1(a_1)$, and $o_2(a_1)$) are similarly defined. Assuming causal consistency, the observed outcome corresponds to the potential outcome for the action actually followed, i.e., $y_1(a_1) = y_1$, $o_2(a_1) = o_2$, $y_2(a_1, a_2) = y_2$, and $y(a_1, a_2) = y$. In this language of potential outcomes, optimizing the two-stage DTR (d_1, d_2) can be expressed as

$$d_2^{opt}(\bar{o}_2) = \arg \max_{a_2 \in \mathcal{A}_2} E(y_2(a_1, a_2) | \bar{o}_2, a_2) \quad \forall \bar{o}_2 \in \bar{\mathcal{O}}_2,$$

$$d_1^{opt}(o_1) = \arg \max_{a_1 \in \mathcal{A}_1} E(y(a_1, d_2^{opt}) | o_1, a_1) \quad \forall o_1 \in \mathcal{O}_1,$$

where the argument \bar{o}_2 of d_2^{opt} is suppressed in the second expression to be more concise.

The approach can be summarized as follows. The Stage 2 regression model for $y_2(a_1, a_2)$ is estimated first, using the observed covariates (\bar{o}_2, a_2) and the observed response variable y_2 . Based on the assumed/postulated Stage 2 model, the optimal mapping from $\bar{\mathcal{O}}_2$ to \mathcal{A}_2 , simply denoted as d_2^{opt} , can be identified, as well as the relevant potential pay-off at Stage 2, denoted as $y_2(a_1, d_2^{opt})$. With d_2^{opt} and potential pay-off $y_2(a_1, d_2^{opt})$, the response variable for Stage 1 can be constructed as $y(a_1, d_2^{opt})$; this so-called *pseudo-outcome* is composed of the observed Stage 1 pay-off y_1 and the potential Stage 2 pay-off $y_2(a_1, d_2^{opt})$. Note that if the observed outcome a_2 matches the optimal outcome according to d_2^{opt} , then the potential pay-off is simply the observed pay-off $y = y_1 + \eta y_2$. Otherwise, the potential pay-off is unobserved and must be imputed (in this BML method, it is actually sampled from the posterior predictive distribution as described further below). Given imputed values, the Stage 1 regression model for the pseudo-outcome $y(a_1, d_2^{opt})$ then can be estimated with observed covariates (o_1, a_1) to identify d_1^{opt} . This type of backward induction strategy is used in several DTR estimation methods, including g-estimation, Q-learning, and dynamic weighted ordinary least squares (Robins 2004; Moodie et al. 2007; Nahum-Shani et al. 2012; Goldberg and Kosorok 2012; Simoneau et al. 2020). Such methods can be contrasted with more fully parametric models such as g-computation (Robins 1986), which require modelling – and correct specification of – the full joint distribution of (o_1, o_2, y) , which is potentially a complex, mixed-covariate type, high-dimensional multivariate distribution; mis-specification of this distribution can lead to bias and incorrect inference (Robins 2004).

Estimation of the terminal stage regression model is simply a typical model of outcome by predictors fit using standard Bayesian methods. The estimation of the nonterminal stage models, on the other hand, is not easily done with standard Bayesian software because of the potential pay-off under the unobserved optimal action at each subsequent stage, which is used in constructing the pseudo-outcome at the current stage. To address this problem, Murray et al. (2018) developed a backward induction Gibbs (BIG) sampler to implement the proposed BML approach in practice. It consists of three steps, repeated until convergence, using * for random variables to indicate

sampled values in an MCMC algorithm. Notationally, i indexes sampled individuals, of whom there are n in the analytic dataset. The algorithm is below:

- step 1 Draw a posterior sample of parameters θ_2^* in the Stage 2 model and set the optimal action $a_{i2}^{opt,*} = d_2^{opt,*}(\bar{o}_{i2}; \theta_2^*)$, $i = 1, \dots, n$.
- step 2 Compare the observed a_{i2} and the optimal $a_{i2}^{opt,*}$. For $i = 1, \dots, n$, if $a_{i2} = a_{i2}^{opt,*}$, then set $y_{i2}^{opt,*} = y_{i2}$; else, sample $y_{i2}^{opt,*}$ from the posterior predictive distribution of $y_2(a_{i1}, a_{i2}^{opt,*})$.
- step 3 Draw a posterior sample of parameters θ_1^* in the Stage 1 model using pseudo-outcome $y_{i1} + \eta_i y_{i2}^{opt,*}$.

The BML approach to backwards induction relies on several standard causal assumptions. Working with the potential outcomes framework requires the axiom of consistency (thus linking potential outcomes to observed data), treatment variation irrelevance and the stable unit treatment value assumption (Rubin 1980) to rule out the possibility of interference, and sequential ignorability (also known as no unmeasured confounding) (Robins 2000). Being, essentially, a Q-learning-like approach focusing on sequential regressions, BML requires that the outcome models in each stage must be correctly specified; the flexible nature of the typical BML implementation lends credibility to this assumption.

2.3 AFT-BART

Bayesian additive regression trees form a Bayesian nonparametric regression model developed by Chipman et al. (2010), which is an ensemble of trees. The accelerated failure time BART (Bonato et al. 2011) is an extension of the approach to accommodate censored outcomes assuming the event time follows a log normal distribution. Let t_i be the event time and c_i the censoring time for individual i . Then the observed survival time is $s_i = \min(t_i, c_i)$, and the event indicator is $\delta_i = I(t_i < c_i)$. Denote by $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$ the p -dimensional vector of predictors. The relationship between t_i and \mathbf{x}_i is expressed as

$$\begin{aligned} \log t_i &= \mu + f(\mathbf{x}_i) + \varepsilon_i, & \varepsilon_i &\stackrel{\text{iid}}{\sim} N(0, \sigma^2) \\ f &\stackrel{\text{prior}}{\sim} \text{BART}, & \sigma^2 &\stackrel{\text{prior}}{\sim} \nu \lambda \chi^{-2}(\nu), \end{aligned}$$

where the constant μ centers the data (a typical default is $\mu = \overline{\log t}$), $f(\mathbf{x}_i)$ is a sum of r regression trees $f(\mathbf{x}_i) \equiv \sum_{j=1}^r g(\mathbf{x}_i; \mathcal{T}_j, \mathcal{M}_j)$ with \mathcal{T}_j denoting a binary tree with a set of internal nodes and terminal nodes and $\mathcal{M}_j = \{\mu_{j1}, \dots, \mu_{jb_j}\}$ denoting the set of parameter values on the terminal nodes of tree \mathcal{T}_j . Each $g(\mathbf{x}_i; \mathcal{T}_j, \mathcal{M}_j)$ assigns a $\mu_{j\ell} \in \mathcal{M}_j$ to \mathbf{x}_i . Full details of the BART model, including prior distributions and MCMC sampling algorithm, can be found in Chipman et al. (2010), but briefly the prior specification $f \stackrel{\text{prior}}{\sim} \text{BART}$ assumes independent priors for each $(\mathcal{T}_j, \mathcal{M}_j)$, with $p(\mathcal{T}_j, \mathcal{M}_j) = p(\mathcal{T}_j)p(\mathcal{M}_j|\mathcal{T}_j)$. The prior on the tree structure, $p(\mathcal{T}_j)$, includes three aspects: 1) the probability that a node at depth \tilde{d} is nonterminal is $\alpha(1 + \tilde{d})^{-\gamma}$ where

$\alpha \in (0, 1)$ and $\gamma \geq 0$; 2) the choice of a covariate given an interior node is uniform; and 3) the choice of the decision rule branching value given the covariate for an interior node is also uniform. Finally, the prior on the terminal node parameters $p(\mathcal{M}_j | \mathcal{T}_j) = \prod_{\ell=1}^{b_j} p(\mu_{j\ell})$ where b_j is the number of terminal nodes for tree j and $\mu_{j\ell} \sim N(0, \tau^2/r)$ on the values of the terminal nodes. This gives $f(\mathbf{x}) \sim N(0, \tau^2)$ for any \mathbf{x} since the value $f(\mathbf{x})$ will be the sum of r independent $N(0, \tau^2/r)$. The prior on σ is calibrated by choosing a ν that results in an appropriate shape, and a λ so that the q th quantile of the prior on σ equals to $\hat{\sigma}$, where $\hat{\sigma}$ is the standard deviation of the residuals from a least squares linear regression of $\log t_i$ on the \mathbf{x}_i .

Since the t_i of censored observation times are not observable, an extra data augmentation step to impute t_i is needed in each iteration when drawing Markov chain Monte Carlo (MCMC) posterior samples with Gibbs sampling. In particular, the unobserved event times are randomly sampled from a truncated normal distribution as

$$\log t_i | s_i, \delta_i = 0, f(\mathbf{x}_i), \sigma^2 \sim N(\mu + f(\mathbf{x}_i), \sigma^2) \times I(t_i > s_i).$$

After data augmentation, the complete log event times are treated as continuous outcomes and the standard BART MCMC draws can be applied.

The AFT-BART model with a log normal survival distribution is implemented within the BART R package (Sparapani et al. 2021); additional details are found in the Appendix D.

2.4 Proposed AFT-BML algorithm

Since the BML approach by Murray et al. (2018) is not directly applicable to censored observations, we extended it by modifying the BIG sampler so that censoring can be accommodated. Here we are interested in the time to an event (such as death) from the start of Stage 1. The Stage 2 treatment decision initiates at an intermediate event such as disease progression. This effectively separates the pay-off or event time into two components: the time to the earliest of the event of interest and the intermediate event triggering Stage 2 (t_1), and if the individual enters Stage 2 ($\eta = 1$), the time from the start of Stage 2 to the event of interest (t_2). Observed data accounting for censoring and entry to Stage 2 are denoted (s_1, δ_1) for Stage 1 and (s_2, δ_2) for Stage 2. Continuing with the potential outcomes notation, let $t(a_1, a_2)$ denote the time to the event of interest when action a_1 is taken at Stage 1 and action a_2 is taken at Stage 2. Similarly, let $t_2(a_1, a_2)$ denote the event time in Stage 2 (starting at the entry to Stage 2) under actions (a_1, a_2) . Finally, potential time $t_1(a_1)$ is the time in Stage 1 until the first of the event of interest or entry to Stage 2. Corresponding pay-offs on the log time scale are denoted $y(a_1, a_2) = \log t(a_1, a_2)$, $y_2(a_1, a_2) = \log t_2(a_1, a_2)$, and $y_1(a_1) = \log t_1(a_1)$. Under consistency, the observed outcome corresponds to the potential outcome for the action actually followed, e.g., $t_1(a_1) = t_1$, $t_2(a_1, a_2) = t_2$, and $t(a_1, a_2) = t$, and similarly for the $y = \log t$ versions.

Murray et al. (2018) recommended using Bayesian nonparametric regression models in Stages 1 and 2 for robustness. Here we illustrated our approach with AFT-BART models in each stage. As before, we use * for random variables to

indicate sampled values in an MCMC algorithm. The Stage 2 regression model for $t_2(a_1, a_2)$ is estimated first, using the observed covariates (\bar{o}_2, a_2) and the observed time to event data (s_2, δ_2) , according to the AFT-BART model

$$\begin{aligned} \log t_{i2} &= \mu_2 + f_2(\bar{o}_{i2}, a_{i2}) + \varepsilon_i, & \varepsilon_i &\sim \text{iid } N(0, \sigma_2^2) \\ f_2 &\overset{\text{prior}}{\sim} \text{BART}, & \sigma_2^2 &\overset{\text{prior}}{\sim} \nu \lambda \chi^{-2}(\nu). \end{aligned} \tag{1}$$

We can run the Stage 2 BART model until convergence, draw M posterior samples from the model (specifically f_2^*), and then sample the optimal Stage 2 treatment rule for each MCMC sample according to

$$d_2^{opt,*}(\bar{o}_2) = \arg \max_{a_2 \in A_2} E(\log t_2 | \bar{o}_2, a_2) = \arg \max_{a_2 \in A_2} f_2^*(\bar{o}_2, a_2).$$

We also can implement a sampling procedure to generate pseudo-outcomes for the total time from Stage 1 assuming optimal Stage 2 treatment as $t^*(a_1, d_2^{opt,*}) = t_1 + \eta t_2^*(a_1, d_2^{opt,*})$, where $t_2^*(a_1, d_2^{opt,*}) = t_2(a_1, a_2)$ if $a_2 = d_2^{opt,*}$ and $\delta_2 = 1$, and is sampled from the posterior predictive distribution otherwise. Some of the potential outcomes resulting from this procedure may still be censored, and we denote the possibly censored version of these potential outcomes as (s^*, δ^*) ; details of calculation of (s^*, δ^*) are provided in the algorithm below. These event time data are then modeled as a function of covariates (o_1, a_1) using another AFT-BART model given by

$$\begin{aligned} \log t_i^*(a_1, d_2^{opt,*}) &= \mu_1 + f_1(o_{i1}, a_{i1}) + \varepsilon_i, & \varepsilon_i &\sim \text{iid } N(0, \sigma_1^2) \\ f_1 &\overset{\text{prior}}{\sim} \text{BART}, & \sigma_1^2 &\overset{\text{prior}}{\sim} \nu \lambda \chi^{-2}(\nu), \end{aligned} \tag{2}$$

For each sampled pseudo-outcomes dataset, we run the Stage 1 AFT-BART model for the pseudo-outcome composed of the sum of the observed Stage 1 outcome and the potential pay-off at Stage 2 until convergence, and then draw one posterior sample from each fitted BART model to determine a sample from the posterior of d_1^{opt} according to

$$d_1^{opt,*}(o_1) = \arg \max_{a_1 \in A_1} E(\log t^*(a_1, d_2^{opt,*}) | o_1, a_1) = \arg \max_{a_1 \in A_1} f_1^*(o_1, a_1).$$

Details of the AFT-BML algorithm are as follows:

- step 1 Run the BART model in equation (1) on the Stage 2 data until convergence and draw M samples (denoted $(f_2^{*(m)}, \sigma_2^{2,*(m)})$, $m = 1, \dots, M$) from the posterior distribution of f_2 and σ_2^2 . Use these to draw M samples from the posterior distribution of a_{i2}^{opt} for each individual at Stage 2, using $a_{i2}^{opt,*(m)} = \arg \max_{a_2} f_2^{*(m)}(\bar{o}_{i2}, a_2)$.
- step 2 Determine the pseudo-observation for the Stage 2 event time under optimal treatment in Stage 2 according to:

$$\log t_{i2}^{opt,*(m)} \begin{cases} = \log t_{i2} & \text{if } a_{i2} = a_{i2}^{opt,*(m)}, \\ & \text{and } \delta_{i2} = 1 \\ \sim N(\mu_2 + f_2^{*(m)}(\bar{o}_{i2}, a_{i2}^{opt,*(m)}), \sigma_2^{2,*(m)}) & \text{if } a_{i2} = a_{i2}^{opt,*(m)}, \\ \quad \times I(t_{i2}^{opt,*(m)} \geq s_{i2}) & \text{and } \delta_{i2} = 0 \\ \sim N(\mu_2 + f_2^{*(m)}(\bar{o}_{i2}, a_{i2}^{opt,*(m)}), \sigma_2^{2,*(m)}) & \text{if } a_{i2} \neq a_{i2}^{opt,*(m)} \end{cases}$$

- step 3 Determine the pseudo-observation data (event times and censoring indicators) for the model for Stage 1 event time under optimal treatment in stage 2. For those who reached Stage 2 ($\eta_i = 1$), set the observed data for the Stage 1 model in sample m as the Stage 1 pseudo event time under optimal Stage 2 treatment $t_{i1}^{*(m)}(a_1, d_2^{opt,*(m)})$, e.g., $s_{i1}^{*(m)} = t_{i1} + t_{i2}^{opt,*(m)}$, and set $\delta_{i1}^{*(m)} = 1$. For those who did not reach Stage 2, set the observed data for the Stage 1 model as $s_{i1}^{*(m)} = s_{i1}$ and $\delta_{i1}^{*(m)} = \delta_{i1}$.
- step 4 Run the BART model in (2) separately on each of the M augmented Stage 1 datasets $(s_1^{*(m)}, \delta_1^{*(m)})$ until convergence and draw 1 sample (denoted $f_1^{*(m)}, \sigma_1^{2,*(m)}$) from the posterior distribution of f_1 and σ_1^2 for each augmented Stage 1 dataset. Use these to draw one sample from the posterior distribution of a_{i1}^{opt} for each individual in Stage 1, using $a_{i1}^{opt,*(m)} = \arg \max_{a_1} f_1^{*(m)}(o_{i1}, a_1)$.

The original BIG sampler indicated that the sampling of the Stage 1 parameters should be updated using the values from the prior iteration. However, while that could potentially speed up implementation as it may not require a full burn-in for each new Stage 1 dataset, it is challenging to implement because most BART software does not allow for starting an update step from a specified value of the tree structure and the terminal node means. Instead, we leverage the fact that the BART chain for Stage 2 does not depend on any updates of the Stage 1 model parameters. Because of this, the BART model for Stage 2 can be run independently and used to generate the potential datasets for Stage 1. Once the M datasets for Stage 1 have been sampled, the BART analyses of each of these Stage 1 datasets in Step 3 can be done in parallel using off the shelf BART software.

Our approach for drawing event times for individuals whose outcomes were censored and who received optimal treatment in Stage 2 was to first sample exact event times for Stage 2 data from the Stage 2 model and then pass this value as an event to the Stage 1 dataset (after adding the observed time in the first stage). Alternatively, one could pass the value as censored to the Stage 1 dataset, in which case the AFT-BART model would implicitly sample event times using the Stage 1 model, instead of using the Stage 2 model as in the algorithm above. We also implemented and examined this alternative approach in our simulation studies, but found no measurable difference in the results, so we did not consider it further. Note also that one could have imputed Stage 2 observations from the predictive distribution under optimal treatment, even when the optimal treatment matches the actual treatment. Instead, we chose to use the observed rather than imputed data whenever appropriate, which we believe will be less sensitive to model assumptions.

In addition to the causal assumptions required for BML, for the context of censored outcomes, we assume that censoring times are independent of the event

times conditional on the covariates, and assume that the log survival time is normally distributed such that an AFT-BML is appropriate. We implemented the proposed method by creating a wrapper function called `dtr1` that utilizes the BART R package (Sparapani et al. 2021). Specifically, the AFT-BART function (`abart`) was called in our wrapper function. The default tuning parameters for the BART prior were adopted, including $\alpha = 0.95$, $\gamma = 2$, $\nu = 3$, $q = 0.9$ (Chipman et al. 2010). Details on the software implementation can be found in Appendix D.

Finally, note that our algorithm provides samples from the Stage k model parameters $(f_k^{*(m)}, \sigma_k^{2,*(m)})$, as well as samples from the optimal treatments for each individual i in Stage k , $a_{ik}^{opt,*(m)}$, and the optimal decision rule $a_k^{opt,*(m)}$. Since these optimal treatment decisions can fluctuate from one sample to another, a practical matter is estimating a single treatment rule from the posterior samples. This could be done for example by picking the treatment with the highest posterior mean of being optimal, i.e.,

$$\hat{a}_{ik}^{opt} = \arg \max_a M^{-1} \sum_m I(a_{ik}^{opt,*(m)} = a). \tag{3}$$

Similarly, we could use the posterior samples to estimate other parameters related to the event time distribution under optimal treatment. For example the means of the log event time for Stage 1 and 2 under optimal treatment are estimated by the posterior mean

$$\begin{aligned} \log \hat{t}_i(a_1^{opt}, d_2^{opt}) &= M^{-1} \sum_m (\mu_1 + f_1^{*(m)}(o_{i1}, a_{i1}^{opt,*(m)})), \\ \log \hat{t}_{i2}(a_2^{opt}) &= M^{-1} \sum_m (\mu_2 + f_2^{*(m)}(\bar{o}_{i2}, a_{i2}^{opt,*(m)})). \end{aligned} \tag{4}$$

Other scales of the survival distribution could also be considered. For example, the median survival time under optimal treatment can be estimated using

$$\widehat{\text{Median}}_i(a_1^{opt}, d_2^{opt}) = M^{-1} \sum_m \exp(\mu_1 + f_1^{*(m)}(o_{i1}, a_{i1}^{opt,*(m)}))$$

and

$$\widehat{\text{Median}}_{i2}(a_2^{opt}) = M^{-1} \sum_m \exp(\mu_2 + f_2^{*(m)}(\bar{o}_{i2}, a_{i2}^{opt,*(m)})).$$

Similarly, the survival probability at time t under optimal treatment can be estimated using

$$\hat{S}_i(t, a_1^{opt}, d_2^{opt}) = M^{-1} \sum_m \Phi \left(\frac{\mu_1 + f_1^{*(m)}(o_{i1}, a_{i1}^{opt,*(m)}) - \log t}{\sigma_1^{*(m)}} \right)$$

and

$$\hat{S}_{i2}(t, a_2^{opt}) = M^{-1} \sum_m \Phi \left(\frac{\mu_2 + f_2^{*(m)}(\bar{o}_{i2}, a_{i2}^{opt,*(m)}) - \log t}{\sigma_2^{*(m)}} \right),$$

where $\Phi(\cdot)$ is the standard normal CDF. Credible intervals for each target parameter of interest can be obtained using the corresponding quantiles of the posterior samples.

3 Simulations

3.1 Simulation design

We conducted simulation studies with 200 replicated training sets of sample size $N = 800$ and an independent testing set of sample size $n = 400$ for each scenario of interest to demonstrate the predictive performance of our method. An observational study with two stages of treatment setting was used with two candidate treatments at each stage. The treatment assignments were generated from a Bernoulli distribution with a probability $P(a_1 = 1 | o_1)$ and $P(a_2 = 1 | \bar{o}_2)$, respectively. Both the event time at Stage 2, as well as the overall event time assuming optimal treatment at Stage 2, were generated from AFT-BML models, assuming a log-normal distribution, similar to the approach of Simoneau et al. (2020).

We fit each training dataset with our method, and made predictions of the optimal action and the mean of the log-normal event time distribution under optimal treatment at each stage on the test dataset. Our performance was compared against Q-learning, including an oracle model along with other models that misspecified the relationship for either stage. We looked at the proportion of optimal treatment (POT), mean squared error (MSE), and 95% credible intervals coverage rate (CR) (for the BART only approach). Simulation settings, method implementation, and simulation metrics are described further in the below sections, with the results following.

3.2 Simulation settings

For individual i , a continuous baseline covariate x_{i1} was drawn from a Uniform distribution with limits 0.1 and 1.29, denoted $U(0.1, 1.29)$, and a binary baseline covariate b_{i1} was from a Bernoulli distribution with probability 0.5. Similarly, a continuous covariate x_{i2} that was measured at the beginning of Stage 2 was also generated from a $U(0.9, 2)$ distribution, and a binary covariate b_{i2} measured at the beginning of Stage 2 was randomly drawn from a $Bern(0.5)$ distribution. Additionally there were two noise covariates, $z_{i1} \sim N(10, 3^2)$, $z_{i2} \sim N(20, 4^2)$, collected at the beginning of Stage 1 and Stage 2, respectively. When fitting the data, all the stage-wise covariates were included in the models to mimic real-world settings in which there is uncertainty as to which covariates are relevant predictors of the outcomes. The Stage 1 treatment was assigned from a Bernoulli distribution with the probability of receiving treatment $P(a_{i1} = 1 | o_1) = \text{expit}(2x_{i1} - 1)$, where $\text{expit}(x) = \exp(x)/(1 + \exp(x))$ is the inverse of

the logit function. For those who entered the second stage ($\eta_i = 1$), the Stage 2 treatment was sampled from a Bernoulli distribution with $P(a_{i2} = 1|\bar{d}_2) = \text{expit}(-2x_{i2} + 2.8)$. The probability of entering Stage 2 was fixed at 0.6, i.e., $P(\eta_i = 1) = 0.6$. Treatment covariates in each stage were coded as $a_k = 1$ or 0 for treatment or control, respectively.

We considered two different scenarios for the relationship between the log event time and the covariates. In Scenario 1, we used an AFT-BML model to generate the event time at Stage 2 as

$$\log t_{i2} = 4 + 0.3x_{i2} + b_{i2} - 0.6x_{i2}b_{i2} + 0.3x_{i1} + 0.4b_{i1} - 0.5x_{i1}b_{i1} + a_{i2}(-0.7 + 0.5x_{i2} - 0.9b_{i2}) + \epsilon_{i2}, \quad \epsilon_{i2} \sim N(0, 0.3^2). \tag{5}$$

The true optimal treatment a_{i2}^{opt} , given by $I(-0.7 + 0.5x_{i2} - 0.9b_{i2} > 0)$, was plugged into equation (5) as a new a_{i2} to calculate the optimal Stage 2 event time t_{i2}^{opt} had everyone received their optimal treatment at Stage 2. The overall event time assuming optimal Stage 2 treatment was generated again from an AFT-BML model as

$$\log t_i(a_1, d_2^{opt}) = 6.3 + 0.7x_{i1} + 0.6b_{i1} - 0.8x_{i1}b_{i1} + a_{i1}(0.1 - 0.2x_{i1} + 0.6b_{i1}) + \epsilon_{i1}, \quad \epsilon_{i1} \sim N(0, 0.3^2). \tag{6}$$

For those who did not enter Stage 2, $t_i(a_1, d_2^{opt})$ was their event time. For those who entered Stage 2, the observed Stage 1 survival time was $t_{i1} = t_i(a_1, d_2^{opt}) - t_{i2}^{opt}$, and the Stage 2 event time was t_{i2} . The censoring time c_i was generated from $U(100, 2000)$ to yield an overall censoring rate of around 20%.

As a comparator, we fitted the data with parametric Q-learning models as well. Since there were two stages in our simulation data, we chose either correctly specified (T) or misspecified (F) Q-function models for each stage as

$$\begin{aligned} \text{stage 1 } Q_{1T}: & x_{i1} + b_{i1} + x_{i1}b_{i1} + a_{i1} + a_{i1}x_{i1} + a_{i1}b_{i1} \quad Q_{1F}: x_{i1} + b_{i1} + z_{i1} + a_{i1} + a_{i1}x_{i1} \\ & + a_{i1}z_{i1} \\ \text{stage 2 } Q_{2T}: & x_{i2} + b_{i2} + x_{i2}b_{i2} + x_{i1} + b_{i1} + x_{i1}b_{i1} + a_{i2} + a_{i2}x_{i2} + a_{i2}b_{i2} \quad Q_{2F}: \\ & x_{i2} + b_{i2} + z_{i2} + x_{i1} + b_{i1} + a_{i2} + a_{i2}x_{i2} + a_{i2}z_{i2} \end{aligned}$$

Combining the two stages together yields four possible modelling specifications: Q_{1T2T} , Q_{1T2F} , Q_{1F2T} , and Q_{1F2F} . Among these four Q-learning models, Q_{1T2T} correctly specifies the parametric form in both stages; we refer to this as the oracle model.

In Scenario 2, we followed a similar structure to simulate the data but with a different set of true models that include non-linear transformations of the covariates. The event time at Stage 2 was generated based on the following equation as

$$\log t_{i2} = 4 + \cos(x_{i2}^3) - 0.4(x_{i2}b_{i2} + 0.5)^2 - 0.1x_{i1} - \sin(\pi x_{i1}b_{i1}) + a_{i2}(0.7x_{i2}^2 - 1) + \epsilon_{i2}, \quad \epsilon_{i2} \sim N(0, 0.1^2). \tag{7}$$

The true optimal treatment, $a_{i2}^{opt} = I(0.7x_{i2}^2 - 1 > 0)$, was used to replace a_{i2} in equation (7) to calculate the optimal Stage 2 event time t_{i2}^{opt} . The overall event time assuming optimal Stage 2 treatment was generated as

$$\log t_i(a_1, d_2^{opt}) = 7.4 + \sin(x_{i1}^2) + x_{i1}^4 + x_{i1}b_{i1} + a_{i1}(0.1 - 0.2x_{i1}^3) + \epsilon_{i1}, \quad \epsilon_{i1} \sim N(0, 0.1^2). \tag{8}$$

The Stage 1 and Stage 2 survival times were calculated in the same way as in Scenario 1, depending on whether the individual entered the second stage. Censoring time c_i was now generated from $U(400, 5000)$ to achieve an overall censoring rate of around 30%.

Based on the underlying true nonlinear functions of covariates, we constructed two misspecified Q-learning models besides the oracle model. The first misspecified model Q_{lin} considered only linear terms in the covariates for both stages as

$$\begin{aligned} \text{Stage 1 } Q_{lin}: & x_{i1} + b_{i1} + z_{i1} + a_{i1}, \\ \text{Stage 2 } Q_{lin}: & x_{i2} + b_{i2} + z_{i2} + x_{i1} + b_{i1} + z_{i1} + a_{i2}. \end{aligned}$$

The second misspecified model Q_{int} considered all two-way interactions among covariates and all interactions between treatment and covariates in each stage in addition to the linear terms in Q_{lin} as

$$\begin{aligned} \text{Stage 1 } Q_{int}: & x_{i1} + b_{i1} + z_{i1} + x_{i1}b_{i1} + x_{i1}z_{i1} + b_{i1}z_{i1} + a_{i1} + a_{i1}x_{i1} + a_{i1}b_{i1} + a_{i1}z_{i1}, \\ \text{Stage 2 } Q_{int}: & x_{i2} + b_{i2} + z_{i2} + x_{i2}b_{i2} + x_{i2}z_{i2} + b_{i2}z_{i2} + x_{i1} + b_{i1} + z_{i1} + x_{i1}b_{i1} + x_{i1}z_{i1} \\ & + b_{i1}z_{i1} + a_{i2} + a_{i2}x_{i2} + a_{i2}b_{i2} + a_{i2}z_{i2}, \end{aligned}$$

such that these models were not correctly specified but were nonetheless richer and more flexible than their ‘only linear’ counterparts.

Two additional simulation settings were conducted using Scenario 2 as a backbone to investigate sensitivity to model assumptions. In Scenario 3, we investigate robustness of the performance of the proposed method to deviations from the log-normal time to event distribution. Here we use an extreme value (Gumbel) distribution for the error distribution of the log event times, leading to a Weibull distribution for the survival times. The parameters of the Gumbel distribution are calibrated to have the same mean and variance of Scenario 2, specifically a location parameter of -0.21 and scale parameter of 0.08 . In Scenario 4, we used a censoring mechanism that is covariate-dependent instead of an independent uniform censoring distribution. Here, the censoring times were generated according to $c_i = 400 + 3800x_{i1} + 2500b_{i1}$.

3.3 Method implementation and simulation metrics

For the proposed method, denoted as BART in the figures, we created a wrapper function `dtr1` that implemented the algorithm described in Sect. 2.4; further documentation of this implementation is available in Appendix D. For Q-learning, we first used `survreg` function from the R package `survival` (Therneau and Grambsch 2000; Therneau 2022) to fit the Stage 2 model, then made predictions of the optimal second stage treatment and corresponding optimal survival time to create Stage 1 data. The `survreg` function was called again to fit the

new augmented Stage 1 data and estimate the optimal first stage treatment with corresponding optimal overall survival time. For Scenario 3 with a Gumbel error distribution, we used the Weibull option to fit the Q-learning approaches with the correct error distribution.

The general evaluation framework is the same for the Q-learning approach as for the proposed method. Given that all the covariates at both stages were simulated for every individual, the related Stage 2 treatment and time were predicted for everyone in the test set, even those who did not actually enter Stage 2. This works for a simulation study, and was done to remove the variability in the set of patients entering Stage 2 from across the simulated test datasets. However, note that this is unrealistic in practice since some Stage 2 covariates are not available if an individual never entered Stage 2, and one can only predict Stage 2 outcomes for those who actually entered Stage 2.

The proportion of optimal treatment is defined as the ratio of the number of individuals who have the true optimal treatment correctly identified by the model and the total number of individuals in the test set, which is 400. More specifically, for the stage-wise POT, the individual is counted in the numerator if the optimal treatment matches with the truth in a specific stage, as shown in equation below for Stage k

$$POT_k = E(I\{\hat{a}_{ik}^{opt} = a_{ik}^{opt}\}), \quad k = 1, 2, \tag{9}$$

where the expectation is over the simulated datasets and the observations in the test set. For the overall or combined POT, only those who have the true optimal treatment correctly identified at both stages are included in the numerator, as in the following expression:

$$POT = E(I\{\hat{a}_{i1}^{opt} = a_{i1}^{opt}\}I\{\hat{a}_{i2}^{opt} = a_{i2}^{opt}\}).$$

It is straightforward to calculate POTs with Q-learning since that approach makes only one prediction of the optimal treatment at each stage for each observation. For the AFT-BML approach, we use the expression for \hat{a}_{ik} in equation (3). We also examined the prediction performance for the means of the log event time distribution under optimal treatment; that is, we calculated the mean squared error by comparing the estimated optimal Stage 2 and overall log event time means to the true means according to

$$MSE_1 = E[(\log \hat{t}_i(a_1^{opt}, d_2^{opt}) - \log t_i(a_1^{opt}, d_2^{opt}))^2]$$

and

$$MSE_2 = E[(\log \hat{t}_{i2}(a_2^{opt}) - \log t_{i2}(a_2^{opt}))^2].$$

For the AFT-BML approach we use the expressions for $\hat{t}_i(a_2^{opt})$ and $\hat{t}_i(a_1^{opt}, d_2^{opt})$ in equation (4). Implementation of the proposed method on a simulation dataset took approximately 15 min using 16 threads.

3.4 Simulation results

Figure 1 shows the decomposition of MSE at both stages, as well as the stage-wise POT and overall POT, for the proposed method, oracle Q-learning model, and the three other misspecified models in Scenario 1. Notice that in both equations (5) and (6), the relationship between log event time, and covariates is linear. As a parametric method that has the right structure as the underlying true model, the performance of the oracle model outperforms the others, with a very small MSE with zero bias, close to 100% stage-wise POT at both stages and a 92% overall POT. For Stage 2, Q_{1F2T} has the same MSE and stage-wise POT as the oracle model since they specified the functional form in the exact same way. Among the other three models, our method performs the best, in terms of a smaller MSE with an even smaller bias, and a higher stage-wise POT with the difference greater than 20%. For Stage 1, the MSE from Q_{1T2F} is slightly bigger but very similar to the oracle, and the stage-wise POT is almost the same as the oracle model, even though the predicted Stage 2 optimal survival time from Q_{1T2F} was based on a misspecified Stage 2 model. This is mainly due to the fact that the simulated Stage 2 event time was relatively small compared to the overall event time so that an incorrect prediction for Stage 2 has a minimal impact on the augmented overall survival time. This resulted a very similar dataset between Q_{1T2F} and the oracle when fitting the Stage 1 model. The proposed Bayesian method, as in Stage 2, has the smallest MSE and the highest stage-wise POT compared to the other two models (Q_{1F2F} and Q_{1F2T}). Our proposed method is better than all approaches except the oracle approach when taking optimal treatment for both stages into consideration using the overall POT. As expected, the variance of the AFT-BML approach has higher variance than the parametric Q-learning

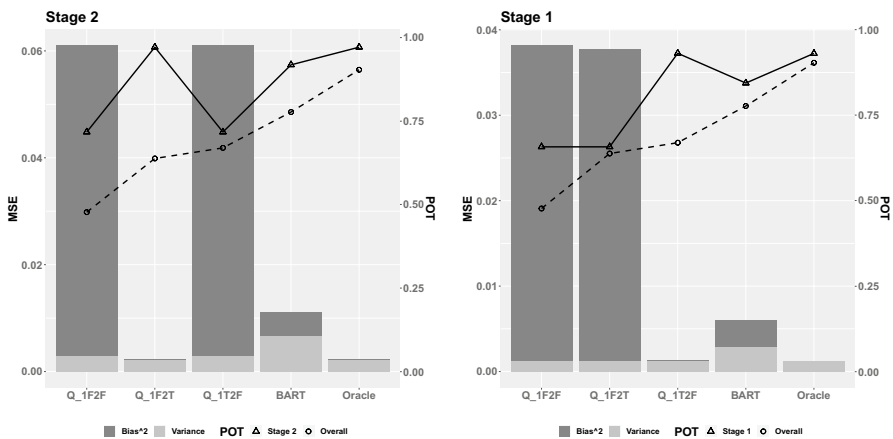


Fig. 1 Mean squared error decomposed into variance and bias² for Scenario 1, in which there is linear dependence of the outcome on covariates at Stage 1 and 2. The left y-axis corresponds to MSE (vertical bars); the secondary y-axis corresponds to stage-wise and overall POT (solid and dashed lines, respectively). Q-learning methods are denoted generically by Q_{1X2Y} , where the X indicates whether the Stage 1 model is correctly (T) or incorrectly (F) specified, and the Y indicates whether the Stage 2 model is correctly (T) or incorrectly (F) specified

approaches; this is likely because of the flexible nonparametric functional form of the relationship between the covariates and outcomes. The good performance of our method in terms of MSE is explained by the dominance of the bias term due to misspecified parametric models in the Q-learning approach.

The results from Scenario 2 are shown in Fig. 2. The relationship is nonlinear between the covariates and log event time in both Eqs. (7) and (8). As expected, the oracle model has close to zero MSE and close to 100% POTs. Our method outperforms the other two models (Q_{lin} and Q_{int}) with a much smaller MSE and higher POTs. The magnitude of the differences in Fig. 2 are larger than those in Fig. 1. The advantage of our nonparametric method becomes more obvious in exploring the nonlinear dependencies, while the other two parametric Q-learning models suffered from incorrect model structures.

Another quantity that is not easily estimable for Q-learning but comes without extra cost for BART is the measure of uncertainty in the estimated optimal values. By drawing MCMC posterior samples, the standard error of log optimal survival times can be calculated as the sample standard deviation. The credible intervals of log event time are also derivable with a collection of posterior samples. On the contrary, to obtain the standard error and confidence interval with Q-learning, bootstrap sampling must be carried out. Here, we only show the coverage rate (CR) of 95% credible intervals for our method in Fig. 3, with coverage rate calculated for the log survival time mean under optimal treatment for each individual in the test set, and boxplots representing the variability in CR across individuals in the test set. Q-learning models are not presented because of the poor fit and high biases in Figs. 1 and 2. The boxplot of 95% CR for both scenarios are almost always above the nominal 95% at Stage 1, and always cover with the lower quartiles above the nominal 95% at Stage 2. This indicates that the proposed method has good accuracy in estimating the uncertainty in the log event time mean under optimal treatment for both stages

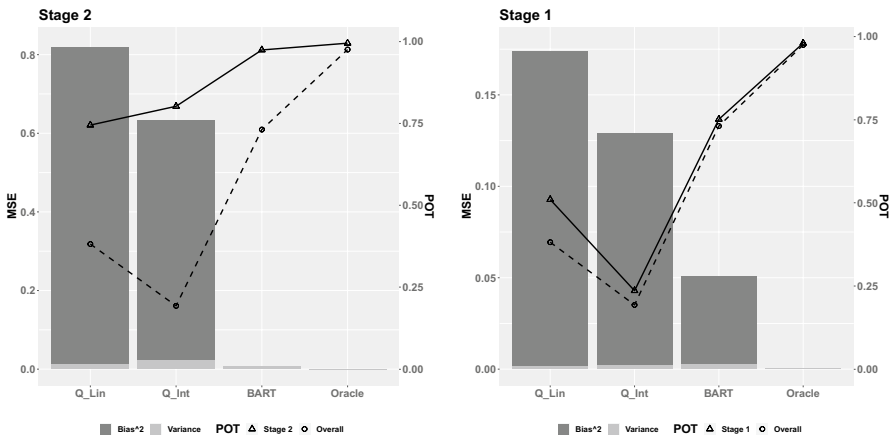


Fig. 2 Mean squared error decomposed into variance and bias² for Scenario 2, in which there is nonlinear dependence of the outcome on covariates at Stage 1 and 2. The left y-axis corresponds to MSE (vertical bars); the secondary y-axis corresponds to stage-wise and overall POT (solid and dashed lines, respectively)

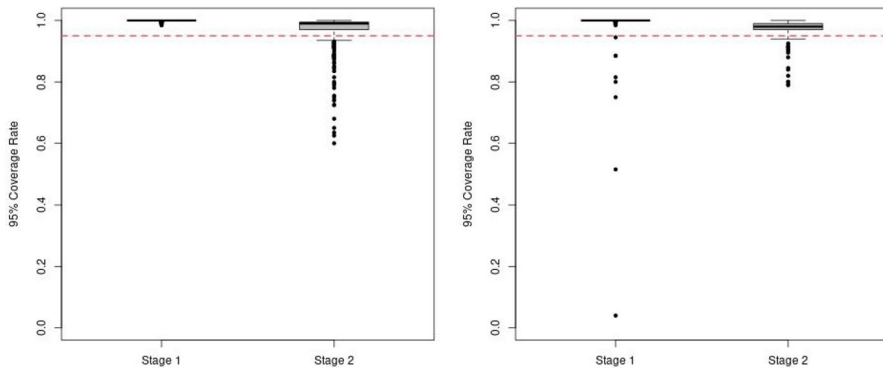


Fig. 3 The coverage rate (CR) of 95% credible intervals for the log event time mean under optimal treatment from the proposed method for Scenario 1 (Left) and Scenario 2 (Right) with the red reference line indicating the nominal 95% level

although the CRs are slightly over 95%. Note that the Bayesian intervals are 95% credible intervals, not 95% confidence intervals, and so the coverage rates won't necessarily equal to the nominal level of 95%, though we expect them to be close. If one is interested in aligning the coverage rate with the Bayesian credible intervals, it is possible to calibrate the credible intervals to target a specific coverage rate.

Finally, results from Scenario 3 and Scenario 4 are shown in Appendix B Figs. 8 and 9. Our proposed AFT-BML model performed well in Scenario 3 in terms of MSE even when the error distribution of the approach was incorrectly specified. It outperformed the Q-learning approaches which had the correct error distribution but incorrect form of the mean term, indicating that correct modeling of the mean term may be more important than the error distribution for performance. The AFT-BML approach also performed well when the censoring was dependent on covariates; this result was expected because the proposed approach relies on the assumption that censoring is independent of the event time given covariates but does not require that censoring be independent of covariates as well.

4 Motivating analysis: optimal treatment for AML patients undergoing transplant

In this section, we applied the proposed method to a retrospective cohort study using registry data collected by the Center for International Blood and Marrow Transplant Research (CIBMTR) (Krakow et al. 2017). There are 4171 patients with complete information in this data who received graft-versus-host disease prophylaxis for their allogeneic hematopoietic cell transplant, which was used to treat their myeloid leukemia, between 1995 and 2007. Some patients were subsequently given a salvage treatment after they developed GVHD and experienced unsuccessful initial treatment. The two stages considered in this study were upfront GVHD prophylaxis treatment and salvage treatment after developing GVHD

and failing initial treatment (which is consistently given as steroids). In each stage, patients were assigned one of two treatments, nonspecific highly T-cell lymphodepleting (NHTL) immunosuppressant therapy or standard prophylaxis immunosuppressant. Estimating an optimal DTR to maximize the overall disease-free survival time (DFS) for patients is our primary goal. The primary outcome is time to death, disease persistence, or relapse. The Stage 1 time is defined as the time from graft infusion to diagnosis of steroid-refractory acute GVHD (if they enter Stage 2) or to the primary outcome or last follow-up (if they do not enter Stage 2). The Stage 2 time is defined as the time from starting salvage treatment for steroid refractory GVHD to the primary outcome or last follow up. Among the 13 covariates of interest, time from graft infusion to acute GVHD onset (≥ 1 month, < 1 month) and use of ≥ 4 immunosuppressors to treat acute GVHD on index form (Yes, No) are only available for those patients who failed at the first treatment. The other covariates include recipient’s age group (< 10 years, $10 - 39$ years, ≥ 40 years), Karnofsky/Lansky performance status at time of transplant ($\geq 80\%$, $< 80\%$), disease status at time of transplant (Early, Intermediate, Advanced), donor relationship (Related, Unrelated), donor-recipient sex (female-male, other), graft source (Bone marrow, Peripheral blood, Umbilical cord), human leukocyte antigen (HLA) match (Well-matched, Partially matched, Mismatched), total-body irradiation (Yes, No), cytomegalovirus status (Negative-negative, Donor or recipient positive), conditioning intensity (Myeloblative, RIC/nonmyeloablative), and use of corticosteroids as part of GVHD prophylaxis (No, Yes). The prophylaxis assigned in Stage 1 is also used in fitting the salvage Stage 2 model. A frequency cross table of prophylaxis and salvage treatment assigned is shown in Table 1. The censoring rate in this cohort was 32%.

Both Q-learning and AFT-BML approaches were used to fit this two stage survival data DTR estimation. All the main effects and the two-way interactions between stage-wise treatment and the other covariates are included in Q-learning models, and 1000 nonparametric bootstrap resamples were generated to estimate the uncertainty of the quantities of interest. For nonparametric AFT-BML, with 1000 MCMC posterior samples, the full distribution was available for any predictions. The point estimates of parameters along with bootstrap mean and 95% confidence interval (CI) at each stage were examined for Q-learning. The waterfall plots for the mean differences in DFS on the log time scale under each treatment at each stage for each individual were created for both Q-learning and AFT-BML, as well as the 95% and 50% credible intervals (bootstrap CIs for Q-learning) presented on the same plot. The differences in the median DFS were also explored for both methods, as were the differences in the two-year DFS probabilities.

Table 1 Treatment assigned at Stage 1 and 2

		Stage 2		Not entered Stage 2
		Standard	NHTL	
Stage 1	Standard	673	219	2180
	NHTL	240	91	768

The analysis results from Q-learning, including point estimates and bootstrap mean, as well as the bootstrap 95% CI, are shown in the Appendix C in Tables 3 and 4. Inspection of the 95% CI for the interaction terms in Table 3 reveals covariate combinations that can be used to identify subgroups where the NHTL or the standard treatment is preferable. For example, assuming all the other covariates are at the reference level, an unrelated donor would benefit more from NHTL than the standard treatment at Stage 1. Similarly in Table 4, when holding the other covariates at the reference level, a patient who received NHTL at Stage 1 would be expected to have a longer DFS time if the standard treatment was given at Stage 2, since the 95% CI of $A2.NHTL * A1.NHTL$ is negative. Intuitively, this might be the case because salvage treatment that is different than the initial treatment which already failed might be expected to be more effective.

In Fig. 4, we present the estimated treatment differences on the log time scale for each stage, along with 95% CI and 50% CI. For AFT-BML, a direct posterior prediction difference for each individual can be calculated from 1000 MCMC posterior samples based on patient-level characteristics. The credible intervals are constructed using the quantiles of posterior samples of each patient. For Q-learning, a standard error can be estimated from the predictions of 1000 bootstrap resamples at an individual level. Using the estimated standard error, the CIs for each patient are calculated in a standard way. A positive difference means NHTL is the preferred treatment for a given stage. The patients are presented in descending order based on the estimated difference, separately for each method.

The results from AFT-BML (Fig. 4a, c) suggest that there is little to be gained by individualizing the treatment in Stage 2 since everyone benefits from the standard treatment, but at Stage 1 there may be significant clinical value in choosing treatment in a personalized fashion to maximize the overall DFS time. In fact, there may be four subgroups that have a distinct difference in expected log event time, in which two groups would benefit from NHTL, one group is indifferent to treatment choice, and one final group that would have longer survival time with the standard treatment. The right panels, showing Fig. 4b, d, provides a similar message using Q-learning: while the standard treatment is the preferred treatment for most (though not all) patients at Stage 2, individualizing the treatment at Stage 1 may lead to important benefits in log DFS time. However, there is a more continuous spectrum of treatment differences in Stage 1 using Q-learning, compared to AFT-BML, and the magnitude of the differences appears larger with Q-learning. Although it is impossible to know why this is the case, we suspect that Q-learning may be overfitting the Stage 1 and Stage 2 models since we have forced all the variables of interest and treatment interaction terms into the model. This can lead to a wider variability in the estimated treatment differences. It may be possible to reduce the potential for overfitting in the Q-learning approach by using a penalized variable selection strategy in the Stage 1 and Stage 2 regression models, but we did not consider this further.

In figure 5, we present the predicted treatment differences on a different time scale, specifically the estimated differences of median DFS time with 95% CI and 50% CI at each stage. Since the event time is assumed to follow a log normal distribution, the exponential of the log normal mean, which transforms the predictions back to the original time scale, is the median rather than the mean. The results on

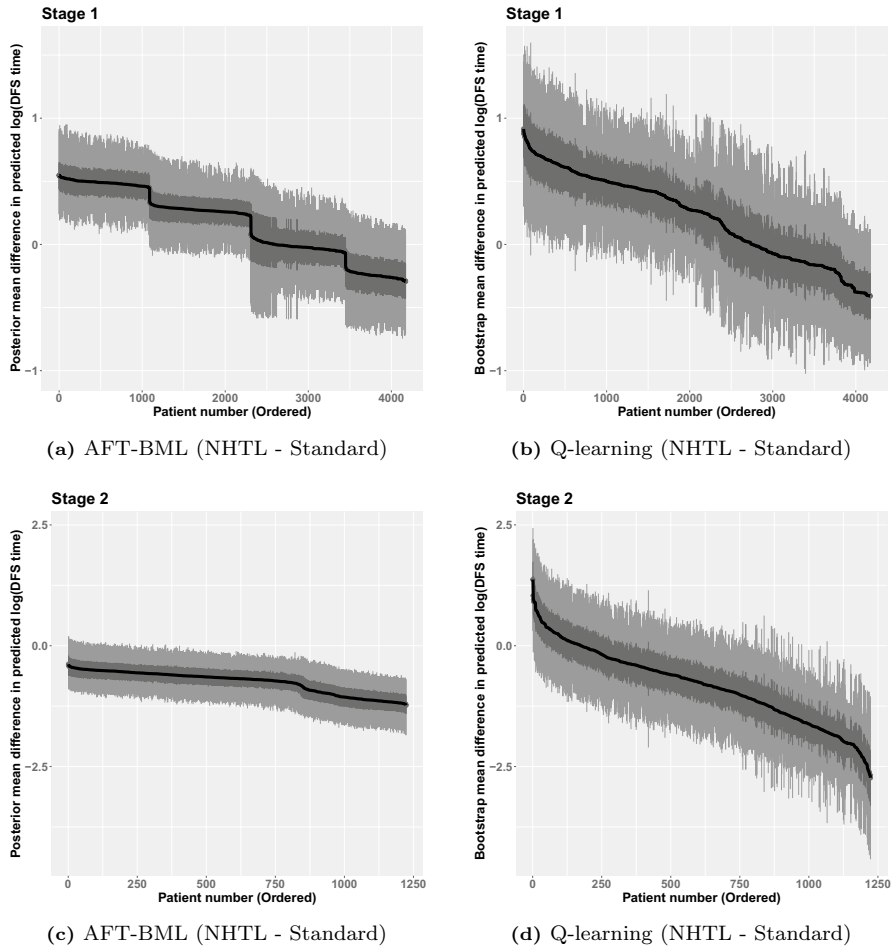


Fig. 4 Predicted mean difference of log DFS time among patients who received AHCT with NHTL versus standard treatment

this scale are generally consistent with the log time scale results, though the scale change produces some unusual results. There are some individuals in the middle and edges of Fig. 5a who have wider CIs than their neighbors. These larger CI widths are an artifact of the scale of the treatment difference and the corresponding ordering of the treatment differences. In particular, the wide intervals correspond to individuals whose median survival predictions are higher (under one or both treatments). When doing inference on the median survival using the exponential transformation of the model parameters, the variance increases with the median. As a result, for these cases where the median survival time under one or more treatments is large, the intervals are wider. Furthermore, because we are plotting in order of the median difference, these cases can occur in several places on the plot, as long as one or more of the medians is large. For example, the wide intervals in the middle of the plot occur

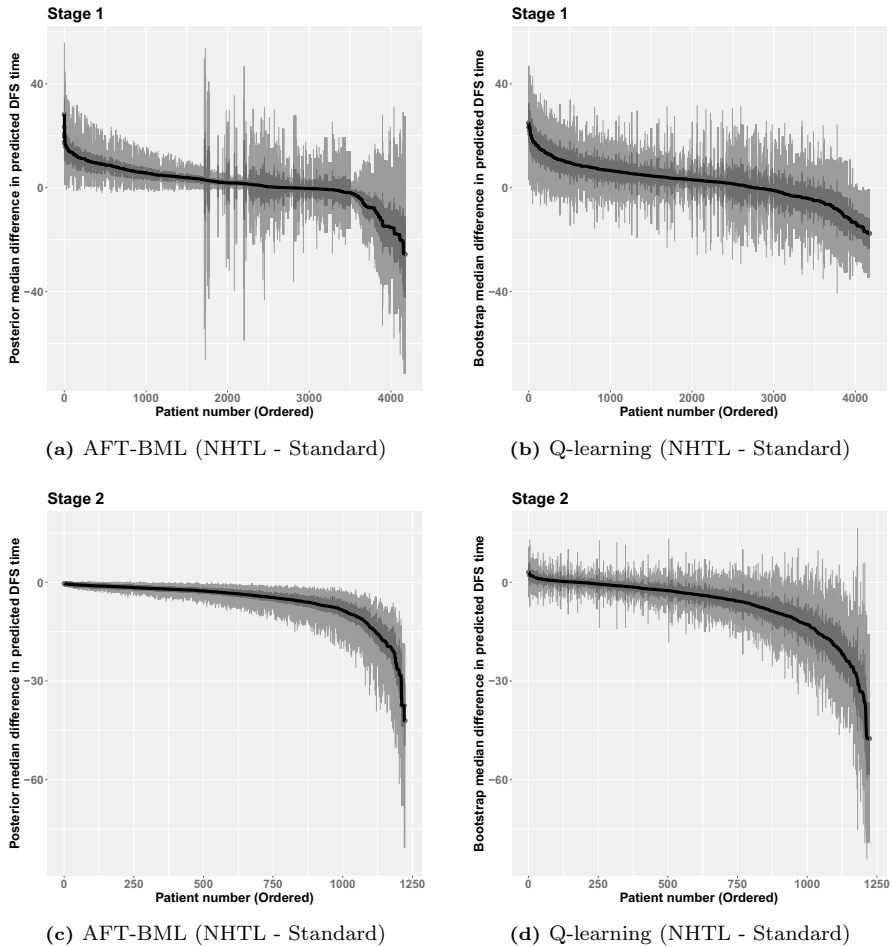


Fig. 5 Predicted difference in median DFS time among patients who received AHCT with NHTL versus standard treatment

when the medians under each treatment are both large and of similar magnitude, while the wide intervals on the edges correspond to individuals where the median under one of the treatments is large, but the other is of a different magnitude. If we had instead plotted in order of the median under one of the treatments, you would see the width of the intervals generally get larger with increasing median.

DTRs could also be defined as the optimal treatment rules to maximize each patient’s two-year DFS probability. Figure 6 shows the two-year DFS probability difference between NHTL and standard treatment for all patients sorted in descending order. For Stage 2, all patients are expected to have a higher two-year survival probability if assigned the standard treatment based on AFT-BML method. The Q-learning method agrees with the inferences from AFT-BML except for a very small proportion of patients, although the magnitude of the differences are much

bigger for Q-learning. For Stage 1, it is easy to notice that AFT-BML splits the patients into four subgroups, similar to what we saw in Fig. 4a. With Q-learning, however, it is difficult to recognize any clear cut points on the curve. As in Stage 2, the magnitude of the differences are smaller for AFT-BML. The proportions of patients who should have NHTL as the optimal treatment for Stage 1 are consistent between AFT-BML and Q-learning.

To further compare the survival probability predictive performance of AFT-BML vs. Q-learning, we calculate the time dependent area under the ROC curve (Heagerty et al. 2000) with the R package `timeROC` (Blanche et al. 2013) using the predicted survival time estimated by both AFT-BML and Q-learning as predictors. For the Stage 2 model, the observed time and event indicator can be used directly in calculating the time dependent AUC. For the Stage 1 prediction model (which assumes

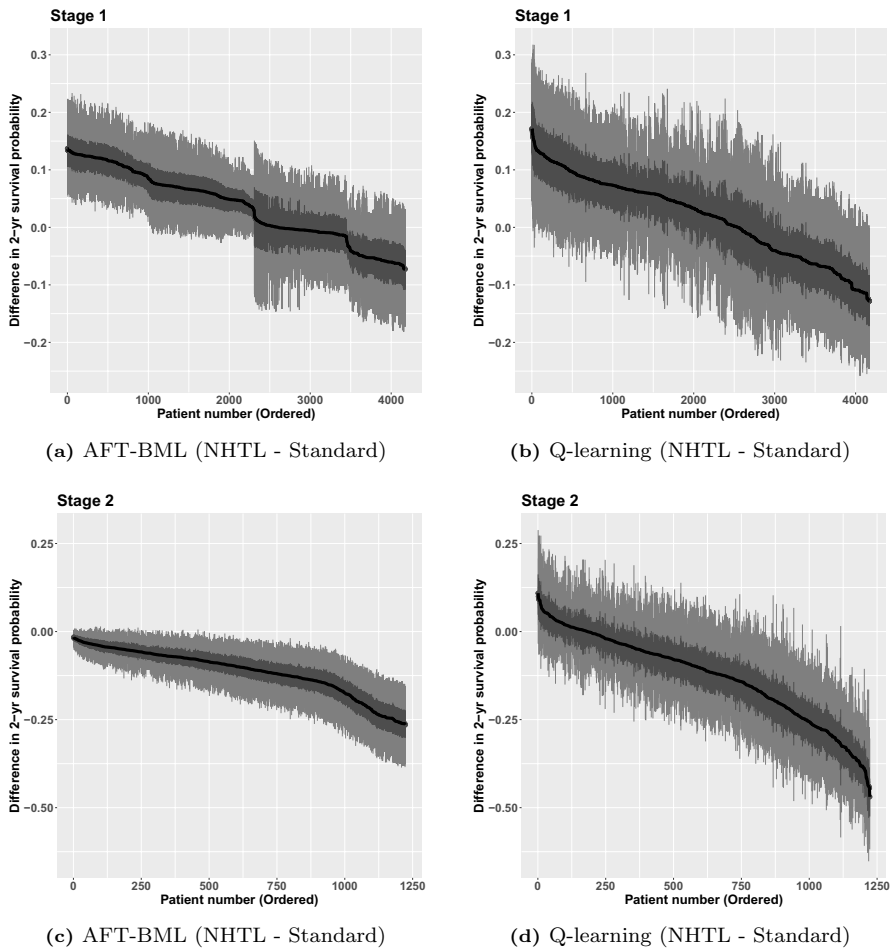


Fig. 6 Two-year DFS probability difference among patients who received AHCT with NHTL versus standard treatment

patients receive optimal treatment in Stage 2), we need to account for not all patients receiving their optimal treatment in Stage 2. To handle this, depending on whether the estimated optimal treatment at Stage 2 was observed or not, the original observation was kept as is, or was censored at the time of entering Stage 2. Since the estimated optimal Stage 2 treatment could be different from AFT-BML to Q-learning, we examined three sets of censored Stage 1 data, including optimal Stage 2 treatment identified by AFT-BML, or Q-learning, or consistently optimal under both Stage 2 models. The time points of interest for Stage 1 are one year, two years, and three years. For Stage 2, only the median and third quartile of the observed time are evaluated. The results from both stages are shown in Table 2. For Stage 2, AFT-BML improves the AUC by 0.55% at the median, and 1.87% at the third quartile, indicating that AFT-BML has a better predictive performance at Stage 2. For Stage 1, the time dependent AUC at 1 year from AFT-BML is approximately 1.7% higher than Q-learning in all three settings. This improvement increases to around 2.2% as time goes to 2 years and 3 years. It indicates that AFT-BML once again outshines Q-learning at Stage 1 in predictive performance.

To visualize the AFT-BML based DTRs, we applied the ‘fit-the-fit’ method and plotted a single tree as in Logan et al. (2019). Since there is little value in differentiating the treatment for Stage 2, we only focus on the Stage 1 model here. Here the outcome used for the single tree fit is the posterior mean treatment difference of the log survival time, although other outcomes such as median DFS or DFS probabilities at fixed timepoints could also be used as outcomes. The R^2 goodness of fit measure for using a single tree in Fig. 7 to model the Stage 1 posterior mean differences in log mean DFS time predictions is above 90%, indicating that this (highly interpretable) single tree is a reasonable representation of the original AFT-BML model at Stage 1. Values in the nodes are the posterior mean differences in mean log DFS

Table 2 Time dependent AUC for Stage 1 and Stage 2 with either AFT-BML model or Q-learning model

Stage	Time in months	Suboptimal treatment censoring rule	Time dependent AUC	
			AFT-BML (%)	Q-learning (%)
1	12	AFT-BML based	71.34	69.61
		Q-learning based	70.89	69.14
		Both agreed	70.81	69.23
	24	AFT-BML based	72.68	70.52
		Q-learning based	72.07	69.86
		Both agreed	71.99	69.99
	36	AFT-BML based	72.50	70.35
		Q-learning based	71.82	69.60
		Both agreed	71.79	69.86
2	3.2 (Median)	NA	70.33	69.78
	15 (Third quartile)	NA	76.09	74.22

For Stage 1, observations were censored at entry to Stage 2 for calculation of the time dependent AUC if they did not receive optimal treatment in Stage 2 (with optimal treatment determined using AFT-BML, using Q-learning, or when both agreed)

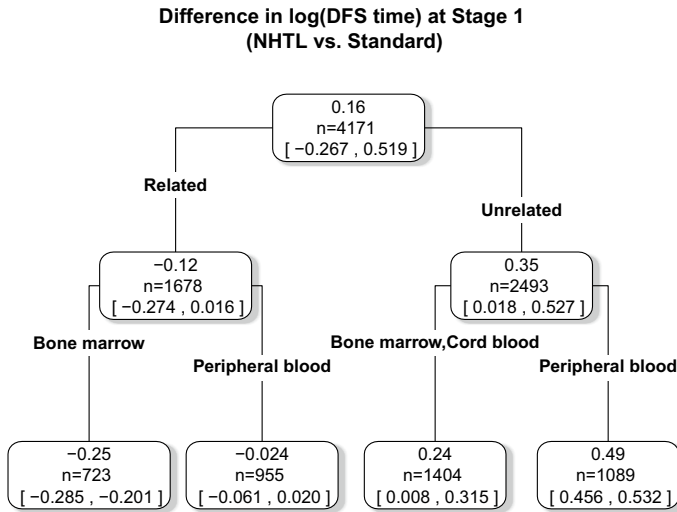


Fig. 7 Single tree fit to the posterior mean log(DFS time) of treatment differences as estimated by AFT-BML in Stage 1. The first split is on the donor relationship (related vs. unrelated), while the second split is on graft source (Bone marrow vs. Peripheral blood vs. Cord blood)

time in months between NHTL and standard treatment. The corresponding 95% CIs are also shown in the same node. The first split (on donor type) indicates that almost all patients receiving unrelated donor transplants would benefit from receiving NHTL as GVHD prophylaxis for their AHCT, while relatively few patients receiving related donor transplants should receive NHTL. As the tree grows, the patients can be divided into four subgroups. After the first split, the two bottom nodes on the left are well apart from each other, as are the two nodes on the right side. These observations agree with Fig. 4a, that identified four subgroups with a distinct posterior mean difference in log DFS times.

5 Conclusion

In medical practice, it is of important clinical value to select the optimal treatment based on an individual’s characteristics. In many settings, a sequence of optimal treatments is desired, such as for diseases like cancer that can progress. The AFT-BML approach for identifying the DTRs can assist physicians to make sound, data-supported decisions at each stage. Classical parametric approaches, such as Q-learning, are constrained by the necessity of correctly specifying functional forms, including all the interaction terms among covariates and treatment. The Bayesian machine learning approach, in contrast, avoids this restriction and allows the model to adaptively determine the relationships between the outcome and the covariates, facilitating optimal treatment identification. We have extended the Bayesian machine learning approach to censored survival data in an AFT-BML model framework, and also provide parallelizable code for implementation

of the computationally intensive algorithm. This wrapper function utilizes standard available BART software without needing to modify the complex underlying BART code. With the simulation studies, we have shown that the AFT-BML approach can achieve almost the same performance as the oracle model for censored outcomes. The results from AFT-BML not only include the optimal treatment and optimal outcomes that classical parametric approaches can provide, but also directly offers the uncertainty measurement for these targets of inference. This extra uncertainty information could be useful in practice since physicians could better assess the confidence they should have in the recommended optimal treatments.

Our comparison to Q-learning in the simulations and the example used a fixed model specification without variable selection. With a larger number of covariates, a penalized AFT model could be used for variable selection in the Q-learning approach. Furthermore, we utilized bootstrapping to estimate the uncertainty of the Q-learning predictions. Alternative approaches, such as penalized Q-learning (Song et al. 2014), could be used to directly provide the uncertainty measurements. However, we are not aware that this approach has been extended to censored data models, such as the AFT model used here.

We compared model performance between AFT-BML and Q-learning in the example by censoring the outcomes of individuals who did not receive optimal subsequent treatment. This was feasible in this dataset because most of the patients received optimal treatment in Stage 2. However, a general strategy of assessing the model performance in the dynamic treatment regimes setting warrants further investigation.

There are some limitations in our approach. We have demonstrated the BML approach for censored data using a parametric log normal AFT-BML model, which has substantial parametric model assumptions even though the functional form of the covariate effects is flexible; other types of Bayesian survival models (Sparapani et al. 2016; Henderson et al. 2020; Linero et al. 2021) could alternatively be used which may require fewer assumptions. In such cases, the methodology described here and our wrapper function can serve as a template for implementing alternative models in a BML DTR framework.

Another limitation is that our AFT-BML approach and wrapper function are currently implemented for the two-stage AFT-BART survival model. We present one possible algorithm for the general setting with more than two stages in the Appendix A, but have not yet implemented it. Modifying our current software implementation to handle more than two stages would introduce some computational challenges. The computational time would increase since we will have more layers of chain burn-in if more than two stages are present, as can be seen from the general algorithm in the appendix. This limitation could be reduced if the multi stage model fittings were embedded together, though this would lose the flexibility of applying BART functions off-the-shelf. Essentially one would need to output the tree structure and terminal node means at the end of one update, and then pass these to the next imputed dataset being analyzed as an initial tree structure and terminal node mean that is being updated.

Appendix A: Extension of AFT-BML algorithm to $K > 2$ stages

Notationally, we assume that observed data accounting for censoring and entry to subsequent stages are denoted (s_k, δ_k) for Stage $k, k = 1, \dots, K$, with $\eta_k, k = 2, \dots, K$ a binary variable indicating that the individual entered Stage k . We omit some of the details previously shown for brevity, and just describe the final algorithm. The algorithm starts at the final stage (Stage K), and then proceeds forward to previous stages, as follows:

step 1 Run the BART model on the Stage K data until convergence and draw M samples (denoted $(f_K^{*(m)}, \sigma_K^{2,*(m)}), m = 1, \dots, M$) from the posterior distribution of f_K and σ_K^2 . Use these to draw M samples from the posterior distribution of a_{iK}^{opt} for each individual at Stage K , using $a_{iK}^{opt,*(m)} = \arg \max_{a_K} \int_K^{*(m)}(\bar{\delta}_{iK}, a_K)$.

step 2 For Stage $k = K - 1, \dots, 1$ perform the following steps to draw relevant posterior samples for f_k, σ_k^2 , and a_{ik}^{opt} .

step 2a Determine the pseudo-observation for the Stage $k + 1$ event time under optimal treatment in stages $k + 1$ or greater according to:

$$\log t_{i,k+1}^{opt,*(m)} \begin{cases} = \log t_{i,k+1} & \text{if } a_{ij} = a_{ij}^{opt,*(m)}, \forall j > k \\ & \text{and } \delta_{i,k+1} = 1 \\ \sim N(\mu_{k+1} + f_{k+1}^{*(m)}(\bar{\delta}_{i,k+1}, a_{i,k+1}^{opt,*(m)}), \sigma_{k+1}^{2,*(m)}) & \text{if } a_{ij} = a_{ij}^{opt,*(m)}, \forall j > k \\ \quad \times I(t_{i,k+1}^{opt,*(m)} \geq s_{i,k+1}) & \text{and } \delta_{i,k+1} = 0 \\ \sim N(\mu_{k+1} + f_{k+1}^{*(m)}(\bar{\delta}_{i,k+1}, a_{i,k+1}^{opt,*(m)}), \sigma_{k+1}^{2,*(m)}) & \text{if } a_{ij} \neq a_{ij}^{opt,*(m)}, \exists j > k \end{cases}$$

step 2b Determine the pseudo-observation data (event times and censoring indicators) for the model for Stage k event time under optimal treatment in stages $k + 1$ or greater. For those who reached Stage $k + 1$ ($\eta_{i,k+1} = 1$), set the observed data for the Stage k model in sample m as the Stage k pseudo event time under optimal subsequent treatment $t_{ik}^{*(m)}(a_k, a_{k+1}^{opt,*(m)}, \dots, a_K^{opt,*(m)})$, e.g., $s_{ik}^{*(m)} = t_{ik} + t_{i,k+1}^{opt,*(m)}$, and set $\delta_{ik}^{*(m)} = 1$. For those who did not reach Stage $k + 1$, set the observed data for the Stage k model as $s_{ik}^{*(m)} = s_{ik}$ and $\delta_{ik}^{*(m)} = \delta_{ik}$.

step 2c Run the BART model separately on each of the M augmented Stage k datasets $(s_k^{*(m)}, \delta_k^{*(m)})$ until convergence and draw 1 sample (denoted $(f_k^{*(m)}, \sigma_k^{2,*(m)})$) from the posterior distribution of f_k and σ_k^2 for each augmented Stage k dataset. Use these to draw one sample from the posterior distribution of a_{ik}^{opt} for each individual in Stage k , using $a_{ik}^{opt,*(m)} = \arg \max_{a_k} \int_k^{*(m)}(\bar{\delta}_{ik}, a_k)$.

Appendix B: Simulation results for scenarios 3 and 4

(See Figs. 8 and 9).

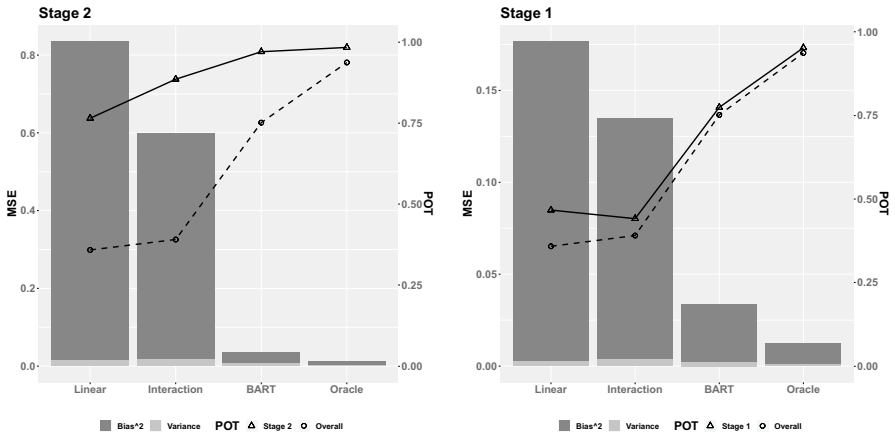


Fig. 8 Mean squared error decomposed into variance and bias² for Scenario 3, in which there is non-linear dependence of the outcome on covariates at Stage 1 and 2 and the error distribution is generated from a Gumbel distribution. The left y-axis corresponds to MSE (vertical bars); the secondary y-axis corresponds to stage-wise and overall POT (solid and dashed lines, respectively)

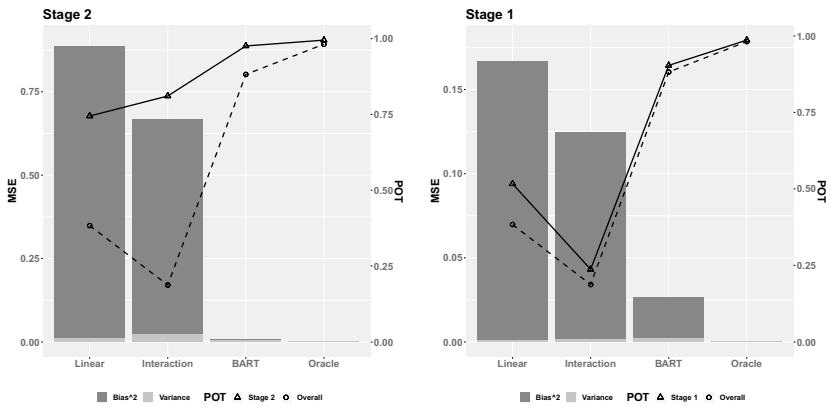


Fig. 9 Mean squared error decomposed into variance and bias² for Scenario 4, in which there is non-linear dependence of the outcome on covariates at Stage 1 and 2 and the censoring times depend on the covariates. The left y-axis corresponds to MSE (vertical bars); the secondary y-axis corresponds to stage-wise and overall POT (solid and dashed lines, respectively)

Appendix C: Analysis results from Q-learning

(See Tables 3 and 4)

Table 3 Predictors of DFS among patients who received AHCT, according to prophylactic GVHD treatment (A1.NHTL) received

Parameter	Point estimate	Bootstrap mean	95% CI
(Intercept)	2.744	2.872	[2.590, 3.256]
A1.NHTL	-0.257	-0.404	[-0.865, 0.063]
Age	-0.068	-0.092	[-0.164, -0.023]
Karnofsky score	-0.521	-0.531	[-0.700, -0.348]
Disease risk - Early	0.741	0.876	[0.738, 1.035]
Disease risk - Intermediate	0.684	0.685	[0.512, 0.872]
Unrelated donor	-0.541	-0.490	[-0.750, -0.263]
Female-to-Male	-0.049	0.014	[-0.168, 0.168]
Graftype - Cord blood	0.592	0.415	[-0.156, 1.014]
Graftype - Peripheral blood	-0.131	-0.136	[-0.357, 0.078]
rec_match	-0.087	-0.139	[-0.275, -0.025]
TBI	-0.018	-0.105	[-0.233, 0.025]
CMV pair - Negative-Negative	0.052	0.051	[-0.115, 0.218]
Conditioning - RIC_NMA	-0.187	-0.108	[-0.371, 0.151]
pgvhcor Yes	-0.196	-0.150	[-0.311, 0.013]
A1.NHTL:Age	0.019	-0.008	[-0.160, 0.143]
A1.NHTL:Karnofsky score	0.107	0.173	[-0.181, 0.549]
A1.NHTL:Disease risk - Early	0.084	-0.063	[-0.356, 0.280]
A1.NHTL:Disease risk - Intermediate	-0.112	-0.126	[-0.422, 0.167]
A1.NHTL:Unrelated donor	0.398	0.548	[0.204, 0.903]
A1.NHTL:Female-to-Male	-0.016	0.002	[-0.328, 0.301]
A1.NHTL:Graftype - Cord blood	-0.299	-0.251	[-0.952, 0.407]
A1.NHTL:Graftype - Peripheral blood	0.115	0.198	[-0.111, 0.531]
A1.NHTL:rec_match	-0.137	0.003	[-0.177, 0.203]
A1.NHTL:TBI	-0.094	0.074	[-0.211, 0.356]
A1.NHTL:CMV pair - Negative-Negative	0.313	0.221	[-0.063, 0.552]
A1.NHTL:Conditioning - RIC_NMA	0.170	-0.005	[-0.350, 0.341]
A1.NHTL:pgvhcor Yes	0.312	0.162	[-0.200, 0.507]
Log(scale)	0.466	0.510	[0.442, 0.660]

Average estimates from 1000 bootstrap samples and bootstrap confidence intervals (95% CI) are given

Table 4 Predictors of DFS among patients who received AHCT and then proceeded to acute GVHD salvage (A2.NHTL) treatment

Parameter	Point estimate	Bootstrap mean	95% CI
(Intercept)	1.898	1.727	[1.171, 2.300]
A2.NHTL	-1.036	-0.745	[-1.761, 0.290]
Age	-0.244	-0.034	[-0.205, 0.140]
Karnofsky score	-1.015	-0.858	[-1.329, -0.399]
Disease risk - Early	1.175	1.051	[0.711, 1.401]
Disease risk - Intermediate	0.661	0.842	[0.453, 1.241]
Unrelated donor	-0.577	-0.799	[-1.159, -0.419]
Female-to-Male	0.060	-0.074	[-0.407, 0.292]
Graftype - Cord blood	0.553	0.738	[0.016, 1.505]
Graftype - Peripheral blood	-0.137	-0.225	[-0.563, 0.116]
rec_match	-0.290	-0.246	[-0.488, 0.000]
TBI	0.213	0.354	[0.044, 0.674]
CMV pair - Negative-Negative	0.142	0.377	[0.046, 0.713]
Conditioning - RIC_NMA	-0.112	-0.025	[-0.393, 0.336]
pgvhcorYes	-0.466	-0.325	[-0.699, 0.055]
Four or more ISP	-0.912	-0.606	[-0.945, -0.265]
Time to acute GVHD	0.406	0.487	[0.157, 0.837]
A1.NHTL	0.670	0.557	[0.177, 0.934]
A2.NHTL:Age	0.108	-0.217	[-0.495, 0.063]
A2.NHTL:Karnofsky score	1.107	0.563	[-0.085, 1.232]
A2.NHTL:Disease risk - Early	-0.683	-0.665	[-1.216, -0.081]
A2.NHTL:Disease risk - Intermediate	-0.275	-0.353	[-1.025, 0.283]
A2.NHTL:Unrelated donor	0.694	1.085	[0.486, 1.673]
A2.NHTL:Female-to-Male	0.047	0.235	[-0.339, 0.841]
A2.NHTL:Graftype - Cord blood	-0.405	-0.988	[-2.486, 0.423]
A2.NHTL:Graftype - Peripheral blood	0.023	0.123	[-0.471, 0.693]
A2.NHTL:rec_match	-0.181	-0.089	[-0.484, 0.292]
A2.NHTL:TBI	0.205	-0.022	[-0.533, 0.506]
A2.NHTL:CMV pair - Negative-Negative	0.171	-0.269	[-0.824, 0.291]
A2.NHTL:Conditioning - RIC_NMA	0.523	0.213	[-0.485, 0.907]
A2.NHTL:pgvhcorYes	-0.167	-0.131	[-0.733, 0.422]
A2.NHTL:Four or more ISP	0.559	0.358	[-0.199, 0.911]
A2.NHTL:Time to acute GVHD	-0.169	-0.324	[-0.900, 0.276]
A2.NHTL:A1.NHTL	-0.649	-0.596	[-1.163, -0.047]
Log(scale)	0.661	0.693	[0.635, 0.749]

Average estimates from 1000 bootstrap samples and bootstrap confidence intervals (95% CI) are given

Appendix D: Software

The R wrapper function that we created for this paper can be found at <https://github.com/xiaoli-mcw/dtrBART>. Simply download and save the `wrapper.R` file. Note

that this function utilizes some functions from the `BART3` R package, which is not available on CRAN yet. The `BART3` package can be found at <https://github.com/rsparapa/bnptools>. The function `dtr1` can be called conveniently after sourcing `wrapper.R` using

```
source("myfolder/wrapper.R")
```

Remember to replace `myfolder` with the actual directory where `wrapper.R` is saved.

The data should have all the covariates including treatments observed in both stages, an indicator for entering Stage 2 defined as η_i in Sect. 2.4, an overall event indicator δ_i , regardless of entering Stage 2 or not, a survival time for Stage 1, and a survival time for Stage 2. For those who did not enter Stage 2, since their information are not used in fitting Stage 2 model, their survival time for Stage 2 can be coded in any reasonable way.

```
dtr1(x1=c("x1"), a1="a1", time1="y1", x2=c("x1", "x2"),
a2="a2", stg2ind="eta", time2="y2", delta="delta", data,
newdata=NULL, opt=TRUE, mc.cores=8)
```

In this function, `x1` is a vector of covariate names that are used in fitting the Stage 1 model. `a1` is the variable name of the action in Stage 1. `time1` is the variable name of survival time at Stage 1. `x2`, `a2`, and `time2` are similar for Stage 2. `stg2ind` is the variable name indicating whether an individual entered Stage 2 indicator. `delta` is the variable name of the overall event indicator. `data` is the dataset. If a `newdata` is provided, predictions of optimal action and optimal outcome for the new data will be returned. `opt` is `TRUE` or `FALSE`, indicating whether only the optimal action and survival time at each stage will be returned (`TRUE`) or whether additional survival times under each action option at each stage will be returned (`FALSE`). `mc.cores` specifies the number of threads to be used in the calculation. With more threads, the program will execute more quickly.

There is a demo dataset `dtrdata.csv` in the same repository on GitHub which is used to explain the results returned by our wrapper function. Here, the `dtrdata` with 1000 observations is split into training (80%) and testing (20%) data.

```
> ind<- sample(1:n, 800)
> train<- dtrdata[ind,]
> test<- dtrdata[-ind,]
> res.dtr<- dtr1(x1="x1", a1="a1", time1="t1",
x2="x2", a2="a2", time2="t2", stg2ind="eta",
delta="delta", data=train, newdata=test, opt=FALSE)
> str(res.dtr)
List of 14
 $a2.opt: int [1:476, 1:1000] 1 1 1 0
           1 0 0 1 0 0...
 $yhat2optmean: num [1:476, 1:1000] 4.82 4.88
           5.06 4.5 4.7...
 $newa2.opt :int [1:121, 1:1000] 0 0 0 1
           0 0 1 1 0 1...
 $newyhat2optmean: num [1:121, 1:1000] 4.43 4.47
           4.33 4.71 4.65...
```

```

$ sigma2: num [1:1000] 0.323 0.331
           0.321 0.326 0.333...
$ a1.opt: num [1:800, 1:1000] 0 0 1 0
           1 0 1 1 1 1...
$ yhatloptmean: num [1:800, 1:1000] 6.54 6.4
           7.26 6.38 7.32...
$ newa1.opt: num [1:200, 1:1000] 1 1 1 1
           1 1 0 0 1 1...
$ newyhatloptmean: num [1:200, 1:1000] 6.99 6.83
           7.42 6.82 7.19...
$ sigma1: num [1:1000] 0.335 0.348
           0.345 0.346 0.342...
$ a2_0: num [1:597, 1:1000] 4.64 4.72
           4.79 4.5 4.62...
$ a2_1: num [1:597, 1:1000] 4.82 4.88
           5.06 4.37 4.7...
$ a1_0: num [1:800, 1:1000] 6.54 6.4 7
           6.38 7.1...
$ a1_1: num [1:800, 1:1000] 6.51 6.37
           7.26 6.36 7.32...

```

The result `res.dtr` is a list of optimal actions (`a2.opt`, `a1.opt`) and corresponding outcomes (`yhat2optmean`, `yhatloptmean`) estimated for both stages, as well as the variances of estimated outcome (`sigma2`, `sigma1`). Since we supplied testing data to the wrapper function, we have the predicted optimal actions (`newa2.opt`, `newa1.opt`) and corresponding outcomes (`newyhat2optmean`, `newyhatloptmean`) reported. The estimated outcomes under each possible action at both stages (`a2_0`, `a2_1`, `a1_0`, `a1_1`) for the training data are also presented because we set `opt=FALSE`. The rows represent individuals, and the columns represent the MCMC samples.

Acknowledgements This work builds on the M.Sc. thesis of S M Ferdous Hossain Hossain (2019). The authors are grateful to Dr Elizabeth Krakow for permission to use the analytic dataset created by her for (Krakow et al. 2017). The dataset was provided by the Center for International Blood and Marrow Transplant Research (CIBMTR); the CIBMTR data registry is partially supported by the US National Cancer Institute, Grant/Award Number: U24CA076518 and HHS234200637015C (HRSA/DHHS).

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Arjas E, Saarela O (2010) Optimal dynamic regimes: presenting a case for predictive inference. *International Journal of Biostatistics* 3(2), Article 10
- Blanche P, Dartigues JF, Jacqmin-Gadda H (2013) Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Statistics in Medicine* 32(30):5381–5397
- Bonato V, Baladandayuthapani V, Broom BM, Sulman EP, Aldape KD, Do KA (2011) Bayesian ensemble methods for survival prediction in gene expression data. *Bioinformatics* 27(3):359–367
- Chakraborty B, Moodie EEM (2013) *Statistical Methods for Dynamic Treatment Regimes: Reinforcement Learning, Causal Inference, and Personalized Medicine*. Springer
- Chipman HA, George EI, McCulloch RE et al (2010) BART: Bayesian additive regression trees. *The Annals of Applied Statistics* 4(1):266–298
- Goldberg Y, Kosorok MR (2012) Q-learning with censored data. *The Annals of Statistics* 40:529–560
- Heagerty PJ, Lumley T, Pepe MS (2000) Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* 56(2):337–344
- Henderson NC, Louis TA, Rosner GL, Varadhan R (2020) Individualized treatment effects with censored data via fully nonparametric bayesian accelerated failure time models. *Biostatistics* 21(1):50–68
- Hossain SMF (2019) Flexible modelling of optimal dynamic treatment regimes for censored outcomes. Master's thesis, McGill University, Montreal, Canada
- Jiang B, Song R, Li J, Zeng D (2019) Entropy learning for dynamic treatment regimes. *Statistica Sinica* 29(4):1633–1655
- Kosorok MR, Moodie EEME (2015) *Adaptive Treatment Strategies in Practice: Planning Trials and Analyzing Data for Personalized Medicine*. ASA-SIAM Publishing, Philadelphia, PA
- Krakov EF, Hemmer M, Wang T, Logan B, Arora M, Spellman S, Couriel D, Alousi A, Pidala J, Last M et al (2017) Tools for the precision medicine era: how to develop highly personalized treatment recommendations from cohort and registry data using q-learning. *American Journal of Epidemiology* 186(2):160–172
- Linero AR, Basak P, Li Y, Sinha D (2021) Bayesian survival tree ensembles with submodel shrinkage. *Bayesian Analysis* 1(1):1–24
- Logan BR, Sparapani R, McCulloch RE, Laud PW (2019) Decision making and uncertainty quantification for individualized treatments using Bayesian Additive Regression Trees. *Statistical Methods in Medical Research* 28(4):1079–1093
- Moodie EE, Richardson TS, Stephens DA (2007) Demystifying optimal dynamic treatment regimes. *Biometrics* 63(2):447–455
- Murphy SA (2003) Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65(2):331–355
- Murray TA, Yuan Y, Thall PF (2018) A Bayesian machine learning approach for optimizing dynamic treatment regimes. *Journal of the American Statistical Association* 113(523):1255–1267
- Nahum-Shani I, Qian M, Almira D, Pelham W, Gnagy B, Fabiano G, Waxmonsky J, Yu J, Murphy S (2012) Q-learning: A data analysis method for constructing adaptive interventions. *Psychological Methods* 17:478–494
- Qian M, Murphy SA (2011) Performance guarantees for individualized treatment rules. *The Annals of Statistics* 39(2):1180
- Robins J (1986) A new approach to causal inference in mortality studies with sustained exposure periods - application to control of the healthy worker survivor effect. *Mathematical Modelling* 7:1393–1512
- Robins JM (2000) Robust estimation in sequentially ignorable missing data and causal inference models. In: *Proceedings of the American Statistical Association*, vol. 1999, pp. 6–10
- Robins JM (2004) Optimal structural nested models for optimal sequential decisions. In: *Proceedings of the Second Seattle Symposium in Biostatistics*, pp. 189–326. Springer
- Rodriguez Duque D, Stephens DA, Moodie EEM, Klein MB (2022) Semi-parametric Bayesian inference for optimal dynamic treatment regimes via dynamic marginal structural models. *Biostatistics* p. (in press)
- Rubin DB (1980) Randomization analysis of experimental data: The Fisher randomization test (comment). *Journal of the American Statistical Association* 75(371):591–593
- Saarela O, Arjas E, Stephens DA, Moodie EEM (2015) Predictive bayesian inference and dynamic treatment regimes. *Biometrical Journa* 57(6):941–958

- Saarela O, Belzile LR, Stephens DA (2016) A Bayesian view of doubly robust causal inference. *Biometrika* 103(3):667–681
- Simoneau G, Moodie EE, Nijjar JS, Platt RW (2020) Estimating optimal dynamic treatment regimes with survival outcomes. *Journal of the American Statistical Association* 115(531):1531–1539
- Song R, Wang W, Zeng D, Kosorok M (2014) Penalized Q-learning for dynamic treatment regimes. *Statistica Sinica* 25:901–920
- Sparapani R, Spanbauer C, McCulloch R (2021) Nonparametric machine learning and efficient computation with Bayesian additive regression trees: The BART R package. *Journal of Statistical Software* 97(1), 1–66. <https://doi.org/10.18637/jss.v097.i01>
- Sparapani RA, Logan BR, McCulloch RE, Laud PW (2016) Nonparametric survival analysis using Bayesian additive regression trees (BART). *Statistics in Medicine* 35(16):2741–2753
- Therneau Terry M, Grambsch Patricia M (2000) *Modeling Survival Data: Extending the Cox Model*. Springer, New York
- Therneau TM (2022) A Package for Survival Analysis in R. <https://CRAN.R-project.org/package=survival>. R package version 3.3-1
- Tsiatis AA, Davidian M, Holloway ST, Laber EB (2020) *Dynamic Treatment Regimes: Statistical Methods for Precision Medicine*. Chapman & Hall/CRC
- Zhao YQ, Zeng D, Laber EB, Kosorok MR (2015) New statistical learning methods for estimating optimal dynamic treatment regimes. *Journal of the American Statistical Association* 110(510):583–598

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.