Nonparametric Failure Time: Time-to-event Machine Learning with Heteroskedastic Bayesian Additive Regression Trees and Low Information Omnibus Dirichlet Process Mixtures

R.A. Sparapani*

Medical College of Wisconsin, Division of Biostatistcs,
8701 Watertown Plank Road, Milwaukee, WI 53226, USA
**email:* rsparapa@mcw.edu

and

B.R. Logan*

Medical College of Wisconsin, Division of Biostatistcs,
8701 Watertown Plank Road, Milwaukee, WI 53226, USA
**email:* blogan@mcw.edu

and

P.W. Laud*

Medical College of Wisconsin, Division of Biostatistcs,
8701 Watertown Plank Road, Milwaukee, WI 53226, USA
**email:* laud@mcw.edu

and

Biometrics V, 1–23 DOI: 10.1111/DOI
MM YYYY

R.E. McCulloch*

Arizona State University, School of Mathematical and Statistical Sciences,

528 Wexler Hall, Tempe, AZ 85281, USA

**email:* robert.mcculloch@asu.edu

and

Others

Summary:

Many popular survival models rely on restrictive parametric or semiparametric assumptions that may lead to incorrect inference or survival predictions when the effects of covariates are complex. Modern advances in computational hardware has led to increasing interest in flexible Bayesian nonparametric methods for time-to-event data such as that provided by Bayesian additive regression trees (BART). We propose a novel approach, called nonparametric failure time (NFT) BART, that incorporates flexibility in Accelerated Failure Time (AFT) models in three ways: 1) a BART component for the mean function of the natural logarithm with time's distribution; 2) a heteroskedastic BART component to handle a covariate dependent variance function; and 3) a flexible nonparametric error distribution using Dirichlet Process Mixtures (DPM). Our proposed approach can be scaled up to large sample sizes, and can be seamlessly employed for variable selection. We provide convenient, user-friendly, computer software that is freely available as a reference implementation. Simulations demonstrate that NFT BART maintains excellent performance even when AFT assumptions are violated. We illustrate the proposed model on a real data example of hematopoietic stem cell transplantation as a treatment for blood-borne cancers.

Key words: accelerated failure time, AFT, BART, blood-borne cancer treatment, constrained DPM, hematopoietic stem cell transplant, LIO prior hierarchy, pseudo-Bayes factor, survival analysis, Thompson sampling variable selection

1. Introduction

Many popular survival models rely on restrictive parametric or semiparametric assumptions that may lead to incorrect inference or survival predictions when the effects of covariates are complex. These include the proportional hazards model (Cox, 1972), and the accelerated failure time (AFT) model (Miller, 1976; Buckley and James, 1979; Aitkin, 1981; Koul et al., 1981; Miller and Halpern, 1982). Ensemble models such as gradient boosting (Freund and Schapire, 1997; Friedman, 2001) and random survival forests (Ishwaran et al., 2008) relax these assumptions, and have been shown to have excellent out-of-sample predictive performance (Baldi and Brunak, 2001; Kuhn and Johnson, 2013). Recently, due to modern advances in computational hardware, there has been increasing interest in Bayesian nonparametric methods such as Bayesian additive regression trees (BART) (Chipman et al., 2010). BART is a Bayesian nonparametric machine learning prior methodology that possesses attractive properties for continuous, categorical and time-to-event outcomes. As the sum of a *large* number of trees, BART falls within the class of ensemble models. High-dimensional data set extensions can be naturally incorporated into BART via a sparse Dirichlet prior (Linero, 2018).

Several authors have incorporated BART within survival analysis models. Bonato et al. (2011) implement methods for proportional hazards, AFT and Weibull regression; although, these extensions include restrictive assumptions. Sparapani et al. (2016) take the discrete time approach (Fahrmeir, 2014) which is relatively assumption-free; however, due to the expansion of the data along a grid of time points, this method will struggle with increasingly larger sample sizes. AFT BART was proposed by Henderson et al. (2020) taking an AFT approach extended by Dirichlet Process mixtures (DPM) (Escobar and West, 1995) for a nonparametric random error distribution; however, this approach still relies on a restrictive AFT assumption. Most recently, Modulated BART is a nonparametric model of the failure time as the first occurrence of a non-homogeneous Poisson process (Linero et al., 2021); however, this method may struggle with increasingly larger sample sizes due to its stochastic process generation of a grid of time points.

In this research, we propose a novel time-to-event approach that we call nonparametric failure time (NFT) BART incorporating more flexibility into Accelerated Failure Time (AFT) models in three ways: 1) a BART component for the mean function of the natural logarithm with time's distribution; 2) a heteroskedastic BART (Pratola et al., 2020) component to handle a covariate dependent variance function; and 3) a flexible nonparametric error distribution using Dirichlet Process Mixtures (DPM). Our proposed approach can be scaled up to large sample sizes, and can be seamlessly employed for variable selection. We provide convenient, user-friendly, computer software that is freely available as a reference implementation.

For many patients suffering from blood-borne cancers, hematopoietic stem cell transplant (HSCT) is the therapy that gives them the best chance of survival. In this case, these transplant recipients are matched to their donors at four human leukocyte antigen (HLA) loci which is called an *eight out of eight* match which we denote by 8/8. Yet, some recipients have many choices of donor with an 8/8 match. Previous research has focused on several recipient-donor match criteria beyond 8/8: age of donor, sex/child-birth parity of donor vs. recipient, HLA loci DPB1 and DQB1 (two additional loci beyond the four) and cytomegalovirus exposure of donor vs. recipient. For example, our previous work has shown that younger donors are preferable for an early binary composite endpoint of acute graft-versus-host disease or death within 180 days, but we did not model longer-term time-to-event outcomes (Logan et al., 2021). The training data available to us consists of about 7000 patients which is at the upper limits of what can be routinely evaluated with discrete time-to-event BART (Sparapani et al., 2016); hence, the impetus for this research.

This article is organized as follows. Section 2 describing the methodology of this article has several parts. First, we introduce binary regression trees and BART in Section 2.1. Next, we describe heteroskedastic BART (HBART) in Section 2.2. We introduce the AFT model and the AFT BART extension in Section 2.4. AFT BART and NFT BART are based on DPM and constrained DPM which is introduced in Section 2.6. We introduce the novel NFT BART model in Section 2.5. In Section 2.7, we discuss posterior inference. And, in Section 2.8, we discuss model performance and comparison with Pseudo-Bayes factors and Thompson sampling variable selection. A simulation study comparison of AFT BART with NFT BART appears in Section 3.1. We describe the result of application of NFT BART to a real-world data example identifying the optimal donor characteristics for HSCT recipients in Section 4. We put this research into perspective with a discussion in Section 5. And, finally, we demonstrate the capabilities of the freely available reference software in the Appendix via an example along with a brief description of the Gibbs conditionals.

2. Methods

2.1 *Binary tree regression models and Bayesian additive regression trees (BART)*

BART (Chipman et al., 2010) is a sum of binary trees nonparametric machine learning regression model where the relationship between the outcome, y_i, and the covariates, $\boldsymbol{x}_i$, is learned from the data itself. We first describe this model for a continuous outcome. Let y_i be a continuous outcome with $i = 1, \ldots, N$ indexing subjects and $\boldsymbol{x}_i$ is a vector of covariates. The BART model has the following form.

$$\begin{array}{ll} y_i \;=\; \mu + f(\boldsymbol{x}_i) + \epsilon_i & \qquad \epsilon_i|\sigma^2 \overset{\text{iid}}{\sim} \mathrm{N}\big(0,\ \sigma^2\big) \\ f \overset{\text{prior}}{\sim} \mathrm{BART}(a, b, k, H) & \qquad \sigma^2 \overset{\text{prior}}{\sim} \nu\lambda\chi^{-2}\left(\nu\right) \end{array}$$

$$f(\boldsymbol{x}_i) \equiv \sum_{h=1}^{H} g(\boldsymbol{x}_i; \mathcal{T}_h, \mathcal{M}_h)$$

Here μ is a constant that centers the data (a typical choice is $\mu = \bar{y}$), while $g(\boldsymbol{x}_i; \mathcal{T}, \mathcal{M})$ is a regression tree function with $\mathcal{T}$ denoting the tree structure and branch decision rules and $\mathcal{M} \equiv \{\mu_1, \mu_2, \ldots, \mu_L\}$ denoting the L leaf values. For a detailed discussion of the BART priors, please refer to the following work (Chipman et al., 2010; Sparapani et al., 2021); these rely on Bayesian binary tree priors (Chipman et al., 1998; Denison et al., 1998; Wu et al., 2007; Pratola, 2016). In brief, where possible, prior default argument settings are employed that often provide adequate fitting in most settings: $a = 0.05$, $b = 2$ and $k = 2$. The number of trees, H, is *large* with typical settings of 50, 100 or 200 where 50 is a common choice (Bleich et al., 2014).

2.2 *Heteroskedastic BART*

Heteroskedastic BART (Pratola et al., 2020) is an extension to BART where we have both a mean function, f, and a variance function, s^2, to fit as a flexible nonparametric function of $\boldsymbol{x}$. This model can be written as

$$\begin{aligned} y_i &= \mu + f(\boldsymbol{x}_i) + \epsilon_i & \epsilon_i|s^2 &\overset{\text{ind}}{\sim} \mathrm{N}\big(0,\ w_i^2 s^2(\boldsymbol{x}_i)\big) \\ f &\overset{\text{prior}}{\sim} \mathrm{BART}(a, b, k, H) & s^2 &\overset{\text{prior}}{\sim} \mathrm{HBART}(\nu, \lambda, \widetilde{H}) \end{aligned} \tag{1}$$

$$s^2(\boldsymbol{x}_i) \equiv \prod_{h=1}^{\widetilde{H}} g(\boldsymbol{x}_i; \widetilde{\mathcal{T}}_h, \widetilde{\mathcal{M}}_h)$$

Here w_i^2 are known constants, w_i^2, that are multiples of the variance; these can be set to $w_i \equiv 1$ if not needed. For f and s^2, in concert, prior default argument settings are employed that often provide adequate fitting in most settings: $\nu = 10$, $\lambda = s_y^2$, $a = 0.05$, $b = 2$ and $k = 5$. For s^2 the number of trees, $\widetilde{H}$, is typically about one-fifth that of H since previous experience has shown that the data contains less information about the variance with respect to the covariates than the mean so fewer trees are necessary, i.e., the default setting is $\widetilde{H} \approx H/5$. For a more detailed discussion of the HBART prior specification, please see Pratola et al. (2020).

2.3 *Accelerated failure time (AFT)*

Suppose that we have time-to-event data of the following form: (t_i, δ_i) where t_i is time; and δ_i is the event status: 0 for right-censoring or 1 for an event. The AFT model can be parametrized in the form of a linear model on the log time scale as

$$y_i = \log t_i = \beta_0 + \boldsymbol{x}_i'\boldsymbol{\beta} + \epsilon_i \tag{2}$$

where ϵ_i has a parametric error distribution; see Kalbfleisch and Prentice (2002) or Klein and Moeschberger (2003) for more information.

2.4 *AFT BART*

Henderson et al. (2020) proposed an AFT BART model that replaces the linear regression component with a BART model and also replaces the parametric error distribution of ϵ_i with a nonparametric random error term using Dirichlet Process Mixtures (Escobar and West, 1995). The model is written as follows.

$$\begin{aligned} y_i &= \mu + f(\boldsymbol{x}_i) + \epsilon_i & \qquad \epsilon_i|(\mu_i, \sigma^2) &\overset{\text{ind}}{\sim} \mathrm{N}\big(\mu_i,\ \sigma^2\big) \\ f &\overset{\text{prior}}{\sim} \mathrm{BART}(a, b, k, H) & \qquad \sigma^2 &\overset{\text{prior}}{\sim} \nu\lambda\chi^{-2}(\nu) \end{aligned} \tag{3}$$

Here the μ_i's in the DPM are subject to a constraint $N^{-1}\sum_i \mu_i = 0$ for identifiability; we defer the description of the prior for μ_i until Section 2.6. Since some of the y_i are unobserved due to censoring, we set the centering value $\mu = \hat{\beta}_0$ from an AFT model with no covariates (2). Note that when we have a censored time, the method utilizes data augmentation by random draws from the truncated distribution. The AFT BART model has nonparametric flexibility allowing it to adapt to the distribution of random error; however, the covariates are only capable of explaining a location-shift on the log time scale, which is a result of the restrictive AFT assumption.

2.5 *NFT BART*

Our proposed method, called NFT BART, enhances the AFT model in two ways. First, it allows the covariates to flexibly explain both a location shift and a scale change. Furthermore, it boosts the flexibility of the nonparametric error distribution of ϵ_i. To facilitate explaining the model, we move fluidly between a parameterization by the precision, τ_i, and by the variance, $\sigma_i^2 = \tau_i^{-1}$, whenever it is more convenient notationally since it is often arbitrary except where noted otherwise. The NFT BART model is as follows subject to the constraints $N^{-1}\sum_i \mu_i = 0$ and $N^{-1}\sum_i \sigma_i^2 = 1$ for identifiability.

$$\begin{aligned} y_i &= \mu + f(\boldsymbol{x}_i) + \epsilon_i & \epsilon_i|(\mu_i, \tau_i, s^2) &\overset{\text{ind}}{\sim} \mathrm{N}\big(\mu_i,\ \sigma_i^2 s^2(\boldsymbol{x}_i)\big) \\ f &\overset{\text{prior}}{\sim} \mathrm{BART}(a, b, k, H) & s^2 &\overset{\text{prior}}{\sim} \mathrm{HBART}(\nu, \lambda, \widetilde{H}) \end{aligned} \tag{4}$$

As with AFT BART, we set the value $\mu = \hat{\beta}_0$ from an AFT model with no covariates (2). Also, censored times are handled using data augmentation, where

$$y_i \begin{cases} \sim \mathrm{N}(\mu + \mu_i + f(\boldsymbol{x}_i),\ \sigma_i^2 s^2(\boldsymbol{x}_i))\, \mathrm{I}(\log t_i, \infty) & \text{if } \delta = 0\text{, right-censoring} \\ = \ \log t_i & \text{if } \delta = 1\text{, an event time} \end{cases}$$

2.6 *DPM, constrained DPM and LIO DPM*

The error distribution in the NFT BART model is based on Dirichlet Process Mixtures (DPM). MCMC sampling of the posterior for Bayesian nonparametric DPM, with both conjugate and non-conjugate priors, can be performed efficiently (Neal, 2000; Ishwaran and James, 2002; Jain and Neal, 2007; Kalli et al., 2011).

The DPM shared atom clusters are random *figments* in the sense that they don't represent meaningful clusters of the data set (to detect data-derived DPM-like interpretable clusters, see Geng et al. (2019)). Rather, DPM clusters are employed here to nonparametrically adapt to the unknown distribution of random error. If we index the MCMC draws by $m = 1, \ldots, M$, then the number of clusters for draw m is the random quantity K_m that expands and contracts as needed where $K \propto \alpha \log N$ (within context, we are suppressing the m subscript for convenience). The number of subects sharing each atom is n_j for $j = 1, \ldots, K$ with corresponding weights $w_j = n_j/N$ that obviously sum to one.

$$\begin{aligned}
(\mu_i, \tau_i)|G &\overset{\text{prior}}{\sim} G & G|\alpha &\overset{\text{prior}}{\sim} \text{DP}\left(\alpha,\ F_{(\mu_0, \tau_0|k_0, b_0)}\right) \\
k_0 &\overset{\text{prior}}{\sim} \text{Gamma}\,(1.5,\ 7.5) & \mu_0|(\tau_0, k_0) &\underset{F}{\overset{\text{prior}}{\sim}} \text{N}\big(0,\ \tau_0^{-1} k_0^{-1}\big) \\
b_0 &\overset{\text{prior}}{\sim} \text{Gamma}\,(2,\ 1) & \tau_0|b_0 &\underset{F}{\overset{\text{prior}}{\sim}} \text{Gamma}\,(3,\ b_0)
\end{aligned} \tag{5}$$

$$\alpha \overset{\text{prior}}{\sim} \text{Gamma}\,(1,\ 0.1)$$

Note that the DPM structure for NFT BART differ from that of AFT BART, which only models μ_i; our strategy provides more flexibility in the nonparametric error distribution. The prior parameter default settings for (μ_i, τ_i) used in this specification follow the Low Information Omnibus (LIO) prior hierarchy for DPM (Shi et al., 2019). Note that LIO, like BART/HBART, was designed to have robust prior parameter default settings that should work well for most data situations without needing manual intervention except for perhaps altering the relative number of desired clusters via the α prior.

It is important to note that NFT BART is over-parameterized such that (f, s^2), is not identifiable as the models have been described up to this point. Therefore, we employ what is known as constrained DPM (Yang et al., 2010) to ensure identifiability. First, consider the constraint $\bar{\mu}_{.} = N^{-1} \sum_i \mu_i = 0$. Constrained DPM is relatively simple to implement by drawing $(\mu_i, \tau_i)|G$ (or $\mu_i|G$ as in AFT BART) without constraint, defining $\tilde{\mu}_i \equiv \mu_i - \bar{\mu}_{.}$ and then re-defining $\mu_i = \tilde{\mu}_i$. Similarly, for the constraint $\overline{\sigma^2} = N^{-1} \sum_i \sigma_i^2 = 1$, we can define $\tilde{\tau}_i \equiv \tau_i \overline{\sigma^2}$ and then re-define $\tau_i = \tilde{\tau}_i$.

2.7 *Posterior inference with AFT BART and NFT BART*

Our primary interest with respect to statistical inference here is the distribution of the time-to-event in relation to the corresponding impact of the covariates. In particular, the survival function, $S(t, \boldsymbol{x})$, plays a central role with respect to inference. The nonparametric estimation of survival is arrived at by aggregating over the DPM clusters (Escobar and West, 1995). So, for NFT BART, we arrive at the following calculation where $\Phi(.)$ is the standard Normal distribution function and $m = 1, \dots, M$ indexes draws from the posterior.

$$S_m(t, \boldsymbol{x}) = 1 - \sum_{j=1}^{K_m} w_{jm} \Phi \left(\frac{\log t - \mu - \mu^*_{jm} - f_m(\boldsymbol{x})}{\sigma^*_{jm} s_m(\boldsymbol{x})} \right) \tag{6}$$

From the above, we calculate our survival function estimate by the mean with respect to the posterior as $\hat{S}(t, \boldsymbol{x}) = M^{-1} \sum_m S_m(t, \boldsymbol{x})$. We can create $1-2\pi$ level credible intervals via the π and $1-\pi$ quantiles of the posterior, $(\hat{S}_\pi(t, \boldsymbol{x}), \hat{S}_{1-\pi}(t, \boldsymbol{x}))$, such that $\hat{S}_p(t, \boldsymbol{x}) = S_{m_p}(t, \boldsymbol{x})$ where m_p is the posterior draw corresponding to the $p = \pi$, or $p = 1-\pi$, quantile respectively.

However, notice that these are inferences for all covariates at once. Often, we are interested in the marginal distribution of a subset of the covariates which are arrived at via an aggregation technique similar to that employed for DPM inference. For marginal effects, we employ Friedman's partial dependence function (Friedman, 2001) that is a common choice for nonparametric regression and/or machine learning applications. We divide the covariates into a subset of interest, A, and their complement, B, where all covariates are $A \cup B$. The covariates of interest are fixed at settings of interest, a single setting denoted $\boldsymbol{x}_{jA}$. The complement take on the observed values found in the training data set, denoted $\boldsymbol{x}_{iB}$ for subject i, with the corresponding setting for all covariates denoted as $(\boldsymbol{x}_{jA}, \boldsymbol{x}_{iB})$. Therefore, we arrive at the marginal effect for setting $\boldsymbol{x}_{jA}$ for NFT BART as follows.

$$\hat{S}_A(t, \boldsymbol{x}_{jA}) = 1 - M^{-1} N^{-1} \sum_m \sum_{i=1}^{N} \Phi \left(\frac{\log t - \mu - \mu_{im} - f_m(\boldsymbol{x}_{jA}, \boldsymbol{x}_{iB})}{\sigma_{im} s_m(\boldsymbol{x}_{jA}, \boldsymbol{x}_{iB})} \right) \tag{7}$$

And, finally, credible intervals for the marginal effects are provided by the posterior quantiles as shown above.

2.8 *Model performance comparison and selection*

Assessment of model performance is important for contrasting competing models and for identifying important variables. Here, we discuss procedures for model performance comparison and selection with NFT BART that leverage the Bayesian methodology.

2.8.1 *Model comparison with pseudo-Bayes factors.*

Comparison between models is often performed with Bayes factors (BF) (Kass and Raftery, 1995). For example, suppose that we want to compare model 2 (denoted by ω_2) vs. model 1 (ω_1) with respect to the data's evidence. The BF is a ratio of marginal likelihoods: $\psi = \frac{[y|\omega_2]}{[y|\omega_1]}$ where $[y|\omega] = \int_{\theta_\omega} [y|\omega, \theta_\omega] [\theta_\omega] \, \mathrm{d}\theta_\omega$ with θ_ω denoting the parameters for model ω (and $[\theta]$ is *generic bracket notation* (Gelfand and Smith, 1990) denoting the distribution of θ, e.g., the prior for θ). A BF substantially larger than one would imply that there is more evidentiary support in favor of model 2 found within the data as opposed to model 1. However, for models with a nonparametric BART prior, the marginal distribution $[y|\omega]$ is not computable.

A proposed alternative to the marginal likelihood is the pseudo-marginal likelihood (PML) from the predictive distribution: $\widetilde{[y|\omega]} = \prod_i [y_i|y_{-i}, \omega]$ (Geisser and Eddy, 1979) where the term $[y_i|y_{-i}, \omega]$ is called the conditional predictive ordinate (CPO). The CPO can be approximated conveniently from the posterior samples by $[y_i|y_{-i}, \omega] \approx \left\{M^{-1} \sum_m [y_i|\theta_{\omega m}, \omega]^{-1}\right\}^{-1}$ (Gelfand and Dey, 1994). For the CPO calculation with NFT BART and right-censoring, we replace the term $[y_i|\theta_{\omega m}, \omega]$ with $\phi(z_{im})^{\delta_i} [1 - \Phi(z_{im})]^{1-\delta_i}$ where $z_{im} = \frac{\log t_i - \mu - \mu_{im} - f_m(\boldsymbol{x}_i)}{\sigma_{im} s_m(\boldsymbol{x}_i)}$ and $\phi(.)$ is the standard Normal density function. Therefore, we can conduct model comparisons via the so-called pseudo-Bayes factor (PBF) as the ratio of PML from each model analogously to the BF. However, these calculations may underflow so taking the natural logarithm is warranted, i.e., the log PML, or LPML, is $\log \widetilde{[y|\omega]}$ and the PBF is $\exp\left(\log \widetilde{[y|\omega_2]} - \log \widetilde{[y|\omega_1]}\right)$. N.B. Jeffreys (1961) has suggested thresholds for BF inference which are applicable to PBF as well.

2.8.2 *Variable selection with Thompson sampling.* A variety of methods for variable selection with BART have been proposed: variable importance (Chipman et al., 2010); permutation-based (Bleich et al., 2014); decoupling, shrinkage and selection (Hahn and Carvalho, 2015; Sparapani et al., 2020); sparse Dirichlet priors (Linero, 2018); and Thompson sampling (Liu and Ročková, 2021). Here we describe the application of Thompson sampling variable selection (TSVS) to the proposed NFT BART model that we will employ in our real data example. TSVS can be performed with, or without, the assistance of sparse Dirichlet priors; however, their pairing together is likely to be more effective.

TSVS relies on Thompson sampling as the name implies (for a tutorial of Thompson sampling, see Russo et al. (2018)). Briefly, Thompson sampling is a heuristic algorithm for decision problems where actions are taken sequentially counter-balancing the optimization of current performance based on what has been *learned* in favor of stochastically exploring the problem space to accumulate new knowledge benefiting future performance. The algorithm addresses a broad range of problems in a computationally efficient manner.

In the TSVS algorithm, multiple variables are randomly chosen based on posterior samples of their reward probabilities. TSVS with BART extends the reach of variable selection to nonparametric models for large data sets with many predictors (big P), or many observations (big N). Unlike deterministic optimization methods for spike-and-slab variable selection, the stochastic nature of TSVS makes it less prone to sub-optimal convergence and, hence, more robust.

Here, we give a concise adaptation of TSVS for NFT BART with big P. TSVS requires a *small* number of trees such that the BART/HBART prior is poised to select only those variables of the greatest import; therefore, we set $H + \tilde{H}$ for a total that is small such as 10, 20 or 40 where smaller numbers engender more sparsity. As shown below, TSVS is an iterative process where $k = 1, \ldots, K$ are the number of steps taken with prior parameters $a_{j0} = a_0$ and $b_{j0} = b_0$.

a. For $j = 1, ..., P$: draw $\theta_{jk} \sim \text{Beta}\left(a_{j,k-1},\ b_{j,k-1}\right)$.

b. Set $B_k = \{j : \theta_{jk} > 0.5\}$: the subset of covariates selected at step k.

c. Fit an NFT BART model with covariates x_{ij} where $j \in B_k$.

d. For $j = 1, ..., P$: do each sub-step.

   (i) If $j \notin B_k$, then $\gamma_{jk} = 0$, else $\gamma_{jk} = \text{I}(U_{jkM} + V_{jkM} > 0)$ where U_{jkM} (V_{jkM}) are the number of branch decision rules for variable x_{ij} at step k from $f\ (s^2)$ with draw M.

   (ii) Update based on the reward: $a_{jk} = a_{j,k-1} + \gamma_{jk}$ and $b_{jk} = b_{j,k-1} + 1 - \gamma_{jk}$.

   (iii) Calculate inclusion probabilities: $\pi_{jk} = \frac{a_{jk}}{a_{jk}+b_{jk}}$.

e. If $k < K$, then return to a. and increment k.

Variables are deemed to be important that have trajectories for π_{jk} exceeding 0.5 by K.

3. Simulation Study

3.1 *Simulation settings*

We conducted a simulation study to compare the AFT BART and NFT BART models. Data sets were simulated from AFT BART and NFT BART while subsequently analyzed by both models. The simulated training data sets were created with two sample sizes: 500 and 2000. For training data sets of size 500 (2000), we simulated 200 (100) data set replicates. The out-of-sample validation data set was simulated at a sample size of 500. Two cases were considered for censoring: 0% (no censoring) and 50%. For each data set, we simulated $P = 20$ covariates: $x_{2j+1} \overset{\text{iid}}{\sim} \text{B}(0.5)$ and $x_{2j} \overset{\text{iid}}{\sim} \text{U}(0, 1)$ where $j = 1, \ldots, 10$. We considered two data generation scenarios: homoskedastic AFT and heteroskedastic NFT. AFT data was generated by $\log t{\sim}\text{N}(\mu(x),\ \exp(-4))$ where $\mu(x) = 2 + 1.6x_1 + 0.8x_2 - 2.4x_2x_3$, i.e., only three covariates have an impact on the outcome and the rest are noise. NFT data was generated by $\log t \sim \mathrm{N}(\mu(x),\ \sigma^2(x))$ where $\mu(x) = 2 - 1.5x_1 + 0.5x_2 + 2x_2x_3$ and $\sigma(x) = \exp(-2 + 1.6x_4 + 0.8x_5 - 2.4x_5x_6)$, i.e., only six covariates have an impact on the outcome and the rest are noise.

Model comparisons were performed with the following metrics at a grid of times corresponding to survival probabilities of 0.9, 0.7, 0.5, 0.3 and 0.1: root mean square error (RMSE), bias, 95% interval coverage and 95% interval length. We define these metrics as follows. Suppose that $j = 1, \ldots, 5$ indexes the known survival probability at a grid of time-points chosen such that $S(t_{ij}, \boldsymbol{x}_i) = S_j = 0.9 - 0.2(j-1)$ for subject i in the validation data set. Now, we can calculate the bias for subject i at survival S_j as $b_{ij} = K^{-1} \sum_k \left[\hat{S}_k(t_{ij}, \boldsymbol{x}_i) - S_j\right]$ where $k = 1, \ldots, K$ indexes the simulated data sets. Similarly, the RMSE is $r_{ij} = \sqrt{K^{-1} \sum_k (\hat{S}_k(t_{ij}, \boldsymbol{x}_i) - S_j)^2}$. We calculate 95% interval coverage as $c_{ij} = K^{-1} \sum_k \mathrm{I}\left(\hat{S}_{k,0.025}(t_{ij}, \boldsymbol{x}_i) < S_j < \hat{S}_{k,0.975}(t_{ij}, \boldsymbol{x}_i)\right)$. And 95% interval length is $l_{ij} = K^{-1} \sum_k \left[\hat{S}_{k,0.975}(t_{ij}, \boldsymbol{x}_i) - \hat{S}_{k,0.025}(t_{ij}, \boldsymbol{x}_i)\right]$. All of these metrics are summarized via box-plots for the 500 subjects in the validation data set.

3.1.1 *Results for sample size of 2000.* Here we restrict our attention to the larger sample size of 2000 (for 500, see below). Consider the data generated from the AFT scenario In Figure 1, we summarized RMSE and their was a slight advantage in favor of AFT BART as might be expected. In Figure 2, we summarized interval coverage and there was a slight advantage in favor of AFT BART being closer to the 95% level. In Web Figure 1, we summarized bias and there was a slight advantage in favor of AFT BART as might be expected. In Web Figure 2, we summarized the 95% interval length and there was an advantage in favor of AFT BART as might be expected.

Consider data generated from the NFT scenario for the larger sample size of 2000. In Figure 3, we summarized RMSE and their was a considerable improvement in favor of NFT BART as we anticipated. In Figure 4, we summarized interval coverage and there was a considerable advantage in favor of NFT BART being closer to the 95% level at virtually all survival settings. In Web Figure 3, we summarized bias and there was a considerable advantage in favor of NFT BART as we anticipated. In Web Figure 4, we summarized the 95% interval length and there was an advantage in favor of NFT BART as we anticipated.

3.1.2 *Results for sample size of 500.* Here we restrict our attention to the smaller sample size of 500 Consider the data generated from the AFT scenario. In Web Figure 5, we summarized RMSE and their was a slight advantage in favor of AFT BART as might be expected. In Web Figure 6, we summarized interval coverage and there was a slight advantage in favor of AFT BART being closer to the 95% level. In Web Figure 7, we summarized bias and there was a slight advantage in favor of AFT BART as might be expected. In Web Figure 8, we summarized the 95% interval length and there was an advantage in favor of AFT BART for 0% censoring while NFT BART had an advantage for 50% censoring.

Consider data generated from the NFT scenario for the smaller sample size of 500. In Web Figure 9, we summarized RMSE and their was a considerable improvement in favor of NFT BART as we anticipated. In Web Figure 10, we summarized interval coverage and there was a considerable advantage in favor of NFT BART being closer to the 95% level. In Web Figure 11, we summarized bias and there was a considerable advantage in favor of NFT BART as we anticipated. In Web Figure 12, we summarized the 95% interval length and there was an advantage in favor of NFT BART as we anticipated.

4. Personalized donor matching for HSCT recipients

We illustrate the proposed NFT BART model to build a prediction model for overall survival outcomes after a hematopoietic stem cell transplant (HSCT) used to treat hematologic malignancies and non-malignant blood disorders. The data set consists of 8830 patients undergoing their first HSCT in the US between 2016 and 2018 from an Human Leukocyte Antigen (HLA) matched unrelated donor (8/8 high-resolution matched at A, B, C and DRB1 loci), with data reported to the Center for International Blood and Marrow Transplant Research (CIBMTR). A total of 7373 patients were used for the training set, with the remainder analyzed in the validation set. A variety of patient, donor, and disease factors were examined in building the prediction model. Patient factors included gender, age, race/ethnicity, performance score, Hematopoietic Cell Transplant Comorbidity Index (HCT-CI), Cytomegalovirus (CMV) status, history of mechanical ventilation, history of invasive fungal infection, history of malignancy, prior autologous transplant, median family income by ZIP code, interval from diagnosis to transplant, disease, disease status, disease related molecular markers, and other disease specific risk factors. Transplant characteristics included conditioning regimen intensity, graft type, Graft-versus-host disease prophylaxis, use of serotherapy. Donor factors included CMV matching, gender and child-bearing parity, and HLA matching information at additional loci of DPB1 and DQB1. First, we describe the impact for one of the most prognostic recipient factors: the HCT-CI comorbidity index. Based on TSVS, comorbidity is among the top five most important covariates for survival. As we can see in Figure 8 of the marginal effect (as computed by (7)), increasing comorbidity leads to a drop in survival until a value of 6, with only a small drop in survival for greater values.

In addition to building a survival prediction model, there is substantial clinical interest in understanding which donor factors are the most important with respect to survival. This can help to inform optimal donor selection for each patient, as we have explored in prior work using BART with an early binary composite endpoint of acute graft-versus-host disease or death within 180 days (Logan et al., 2021). Because of the strong effect of recipient factors, we include all 39 recipient factors in our models while searching for the most important donor factors. We employed TSVS to identify which donor factors are of paramount import: donor age and donor sex/child-bearing parity were ranked in that order; for a graphical depiction, see Figure 6. Furthermore, via PBF, we found that there was decisive evidence (Jeffreys, 1961) in favor of the model with only these two donor factors (donor age and donor sex/child-bearing parity) as opposed to all donor factors: $\exp(-7133 + 7187) = \exp(54) \approx 3 \times 10^{23}$. This suggests that there may be limited to no benefit to further optimizing donors over other factors besides donor age and sex/child-bearing parity.

5. Discussion

Our proposed NFT BART model implements an extremely flexible time-to-event Bayesian ensemble model. It avoids many restrictive assumptions such as linearity, proportional hazards and/or AFT structure by utilizing BART components for both the log time mean function and the variance function, combined with a nonparametric error distribution. This flexibility shows substantially improved prediction performance in simulation studies when the AFT model assumption does not hold, with minimal loss of performance compared to competing methods when it does. The procedure is scalable, in contrast to the nonparametric discrete time survival BART model (Sparapani et al., 2016). NFT BART can be seamlessly employed in the tasks of model comparison and variable selection with modern Bayesian/pseudo-Bayesian techniques. While NFT BART has distinct advantages, it is not immediately clear if NFT can easily be extended to other types of survival analysis outcomes such as recurrent events (Sparapani et al., 2020) and/or competing risks (Sparapani et al., 2020). This is an important area for future research. Nevertheless, NFT BART is a flexible Bayesian nonparametric time-to-event inference methodology that has attractive properties.

Acknowledgements

Supported by a research grant funded by the US Office of Naval Research N00014-18-1-2888. This research was completed in part with computational resources and technical support provided by the Research Computing Center at the Medical College of Wisconsin.

References

Aitkin, M. (1981). A note on the regression analysis of censored data. *Technometrics* **23,** 161–163.

Baldi, P. and Brunak, S. (2001). *Bioinformatics: The Machine Learning Approach.* MIT Press, Cambridge, MA, 2nd edition.

Bleich, J., Kapelner, A., George, E. I., and Jensen, S. T. (2014). Variable selection for BART: An application to gene regulation. *The Annals of Applied Statistics* **8,** 1750–1781.

Bonato, V., Baladandayuthapani, V., Broom, B. M., Sulman, E. P., Aldape, K. D., and Do, K.-A. (2011). Bayesian ensemble methods for survival prediction in gene expression data. *Bioinformatics* **27,** 359–367.

Buckley, J. and James, I. (1979). Linear regression with censored data. *Biometrika* **66,** 429–436.

Chipman, H. A., George, E. I., and McCulloch, R. E. (1998). Bayesian CART model search. *Journal of the American Statistical Association* **93,** 935–948.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). BART: Bayesian Additive Regression Trees. *The Annals of Applied Statistics* **4,** 266–298.

Cox, D. R. (1972). Regression models and life-tables (with discussions). *Journal of the Royal Statistical Society B* **34,** 187–220.

de Waal, T., Pannekoek, J., and Scholtus, S. (2011). *Handbook of Statistical Data Editing and Imputation.* John Wiley & Sons, Hoboken, NJ.

Denison, D. G., Mallick, B. K., and Smith, A. F. (1998). A Bayesian CART Algorithm. *Biometrika* **85,** 363–377.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American statistical association* **90,** 577–588.

Fahrmeir, L. (2014). Discrete survival-time models. *Wiley StatsRef: Statistics Reference Online* .

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences* **55,** 119–139.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* **29,** 1189–1232.

Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association* **74,** 153–160.

Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society: Series B (Methodological)* **56,** 501–514.

Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85,** 398–409.

Geng, J., Bhattacharya, A., and Pati, D. (2019). Probabilistic community detection with unknown number of communities. *Journal of the American Statistical Association* **114,** 893–905.

Hahn, P. and Carvalho, C. (2015). Decoupling shrinkage and selection in Bayesian linear models: a posterior summary perspective. *Journal of the American Statistical Association* **110,** 435–448.

Henderson, N. C., Louis, T. A., Rosner, G. L., and Varadhan, R. (2020). Individualized treatment effects with censored data via fully nonparametric Bayesian accelerated failure time models. *Biostatistics* **21,** 50–68.

Ishwaran, H. and James, L. F. (2002). Approximate Dirichlet process computing in finite normal mixtures: smoothing and prior information. *Journal of Computational and Graphical statistics* **11,** 508–532.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008). Random survival forests. *The annals of applied statistics* **2,** 841–860.

Jain, S. and Neal, R. M. (2007). Splitting and merging components of a nonconjugate Dirichlet process mixture model. *Bayesian Analysis* **2,** 445–472.

Jeffreys, H. (1961). *The theory of probability.* OUP Oxford.

Kalbfleisch, J. and Prentice, R. (2002). *The Statistical Analysis of Failure Time Data.* John Wiley & Sons, Hoboken, NJ, 2nd edition.

Kalli, M., Griffin, J. E., and Walker, S. G. (2011). Slice sampling mixture models. *Statistics and computing* **21,** 93–105.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American statistical association* **90,** 773–795.

Klein, J. P. and Moeschberger, M. L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data.* Springer-Verlag, New York, NY, 2nd edition.

Koul, H., Susarla, V., and Van Ryzin, J. (1981). Regression analysis with randomly right-censored data. *The Annals of statistics* **9,** 1276–1288.

Kuhn, M. and Johnson, K. (2013). *Applied Predictive Modeling.* Springer-Verlag, New York, NY.

Linero, A. (2018). Bayesian regression trees for high dimensional prediction and variable selection. *Journal of the American Statistical Association* **113,** 626–636.

Linero, A. R., Basak, P., Li, Y., and Sinha, D. (2021). Bayesian survival tree ensembles with submodel shrinkage. *Bayesian Analysis* **(ahead of print),** 1–24.

Liu, Y. and Ročková, V. (2021). Variable selection via Thompson sampling. *Journal of the American Statistical Association* **(ahead of print),** 1–41.

Logan, B. R., Maiers, M. J., Sparapani, R. A., Laud, P. W., Spellman, S. R., McCulloch, R. E., and Shaw, B. E. (2021). Optimal Donor Selection for Hematopoietic Cell Transplantation Using Bayesian Machine Learning. *JCO clinical cancer informatics* **5,** 494–507.

Loprinzi, C. L., Laurie, J. A., Wieand, H. S., Krook, J. E., Novotny, P. J., Kugler, J. W., Bartel, J., Law, M., Bateman, M., and Klatt, N. E. (1994). Prospective Evaluation of Prognostic Variables from Patient-Completed Questionnaires. North Central Cancer Treatment Group. *Journal of Clinical Oncology* **12,** 601–607.

Miller, R. and Halpern, J. (1982). Regression with censored data. *Biometrika* **69,** 521–531.

Miller, R. G. (1976). Least squares regression with censored data. *Biometrika* **63,** 449–464.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics* **9,** 249–265.

Pratola, M. T. (2016). Efficient Metropolis–Hastings Proposal Mechanisms for Bayesian Regression Tree Models. *Bayesian Analysis* **11,** 885–911.

Pratola, M. T., Chipman, H. A., George, E. I., and McCulloch, R. E. (2020). Heteroscedastic BART via multiplicative regression trees. *Journal of Computational and Graphical Statistics* **29,** 405–417.

Russo, D. J., Van Roy, B., Kazerouni, A., Osband, I., and Wen, Z. (2018). et almbox. 2018. *A tutorial on Thompson sampling. Foundations and Trends in Machine Learning* **11,** 1.

Shi, Y., Martens, M., Banerjee, A., and Laud, P. (2019). Low information omnibus (LIO) priors for Dirichlet process mixture models. *Bayesian Analysis* **14,** 677–702.

Sparapani, R., Logan, B. R., McCulloch, R. E., and Laud, P. W. (2020). Nonparametric competing risks analysis using Bayesian Additive Regression Trees (BART). *Statistical*

Methods in Medical Research **29,** 57–77.

Sparapani, R. and McCulloch, R. (2021). *Nonparametric Failure Time: Heteroskedastic Bayesian Additive Regression Trees and Low Information Omnibus Dirichlet Process Mixtures*. `https://cran.r-project.org/package=nftbart`.

Sparapani, R., Rein, L., Tarima, S., Jackson, T., and Meurer, J. (2020). Non-parametric recurrent events analysis with BART and an application to the hospital admissions of patients with diabetes. *Biostatistics* **21,** 69–85.

Sparapani, R., Spanbauer, C., and McCulloch, R. (2021). Nonparametric machine learning and efficient computation with Bayesian Additive Regression Trees: the BART R package. *Journal of Statistical Software* **97,** 1–66.

Sparapani, R. A., Logan, B. R., McCulloch, R. E., and Laud, P. W. (2016). Nonparametric survival analysis using Bayesian Additive Regression Trees (BART). *Statistics in Medicine* **35,** 2741–2753.

Tan, Y. V. and Roy, J. (2019). Bayesian additive regression trees and the General BART model. *Statistics in medicine* **38,** 5048–5069.

Wu, Y., Tjelmeland, H., and West, M. (2007). Bayesian CART: Prior Specification and Posterior Simulation. *Journal of Computational and Graphical Statistics* **16,** 44–66.

Xu, D., Daniels, M. J., and Winterstein, A. G. (2016). Sequential BART for imputation of missing covariates. *Biostatistics* **17,** 589–602.

Yang, M., Dunson, D. B., and Baird, D. (2010). Semiparametric Bayes hierarchical models with mean and variance constraints. *Computational statistics & data analysis* **54,** 2172–2186.

Supporting Information

The Web Appendix referenced in Section is available with this paper at the Biometrics website on Wiley Online Library.

Appendix Material: Software Implementation

The software necessary to implement the methodology explored in this article is not trivial to implement. For NFT BART, we created the *nftbart* R package that is freely available online hosted on the Comprehensive R Archive Network (CRAN) (Sparapani and McCulloch, 2021). The *nftbart* package relied on several key computational methods some of which were explored in this article. The next section demonstrates an example discussing missing data imputation and the marginal effects methodology employed here. Further, the Gibbs conditionals necessary for NFT BART are shown in the last section of the Appendix. Other computational methods employed include BART (Chipman et al., 2010), HBART (Pratola et al., 2020), efficient BART/HBART posterior sampling (Pratola, 2016), efficient DPM sampling (Neal, 2000), constrained DPM (Yang et al., 2010), DPM LIO (Shi et al., 2019) and data augmentation for left-/right-censoring (Henderson et al., 2020). For AFT BART (Henderson et al., 2020), we relied on the *AFTrees* R package freely available online at `https://github.com/nchenderson/AFTrees`.

Advanced lung cancer example

With the *nftbart* R package, we present a real data example of an advanced lung cancer study (Loprinzi et al., 1994). Two-hundred and twenty-eight patients with lung cancer were followed by the North Central Cancer Treatment Group for a median of roughly one year. Several covariates of interest were collected including age, sex, daily activity performance scores, diet and weight-loss information. All of these variables were largely non-missing with the exception of the calories consumed at meals for which missingness was 20.6%.

For this limited amount of missing data, we utilized record-level *cold-decking imputation* that is biased towards the null. The name reflects its similarity to hot-decking (de Waal et al., 2011) except that no attempt is made to locate a nearby/hot neighbor based on the outcome nor any other covariate criteria (near/hot vs. further/cold distances like in the children's game hide'n'seek), i.e., cold-decking is a simple random selection of a non-missing subject's record to replace the missing values with. For subject's with multiple missing values, the joint relationships between covariates are maintained by replacing all of the missing values from the non-missing subject randomly chosen. This simple missing data imputation method is sufficient for data sets with relatively few missing values; for more prevalent missingness we recommend the *sequential* BART algorithm (Xu et al., 2016).

For this example, sex was determined to be the most important covariate by TSVS with 138 male and 90 female participants. To demonstrate a common computation with *nftbart*, we will compare the survival experience of males vs. females by their marginal effects with Friedman's partial dependence function (Friedman, 2001) as shown in (7). As we can see in Figure 9, females generally have longer survival; however, for advanced lung cancer the prognosis is dire in the era of the collected data since the survival probability declines precipitously for both sexes. This demonstration is included with the *nftbart* package. You can install the *nftbart* R package and run this example as follows (use a nearby CRAN mirror for best results installing; see `http://cran.r-project.org/mirrors.html`).

```
> options(repos=c(CRAN="http://cran.r-project.org"))
> install.packages("nftbart", dependencies=TRUE)
> ## system.file() shows you where lung.R is installed to see its contents
> system.file("demo/lung.R", package="nftbart")
> source(system.file("demo/lung.R", package="nftbart"))
> ## demo("lung", package="nftbart") ## via the demo() facility
```

N.B. there is also a demonstration of TSVS for this example `"demo/TSVSlung.R"`.

Appendix Derivations for NFT BART: Gibbs conditionals

In order to perform Markov chain Monte Carlo (MCMC) posterior sampling, we need to derive the Gibbs conditionals. Derivations like these are fairly standard in the BART literature; what Tan and Roy have coined a term for: the "General BART" model (Tan and Roy, 2019).

First, we isolate the impact of f from the other parameters by $r_i \equiv y_i - \mu - \mu_i = f(\boldsymbol{x}_i) + s(\boldsymbol{x}_i)\sigma_i\epsilon_i$ where $r_i|(f, s^2, \mu_i, \tau_i) \sim \mathrm{N}(f(\boldsymbol{x}_i),\ s^2(\boldsymbol{x}_i)\sigma_i^2)$. So, let $r_i \equiv y_i - \mu - \mu_i$ be the outcome (with $w_i^2 = s^2(\boldsymbol{x}_i)\sigma_i^2$ as in (1)), then draw $f|(r, s^2, \mu_i, \tau_i)$ from its Gibbs conditional. Next, we draw s similarly: $u_i \equiv \frac{r_i - f(\boldsymbol{x}_i)}{\sigma_i} = s(\boldsymbol{x}_i)\epsilon_i$ where $u_i|(f, s^2, \mu_i, \tau_i) \sim \mathrm{N}(0,\ s^2(\boldsymbol{x}_i))$. So, with $u_i \equiv \frac{r_i - f(\boldsymbol{x}_i)}{\sigma_i}$ as the outcome, then draw $s^2|(u, f, \mu_i, \tau_i)$ as in (1). And, finally, we draw (μ_i, τ_i) with $v_i \equiv \frac{y_i - \mu - f(\boldsymbol{x}_i)}{s(\boldsymbol{x}_i)} = \frac{\mu_i}{s(\boldsymbol{x}_i)} + \sigma_i\epsilon_i$ where $v_i|(f, s^2, \mu_i, \tau_i) \sim \mathrm{N}\left(\frac{\mu_i}{s(\boldsymbol{x}_i)},\ \sigma_i^2\right)$. Here, $v_i \equiv \frac{y_i - \mu - f(\boldsymbol{x}_i)}{s(\boldsymbol{x}_i)}$ is the outcome and we draw $(\mu_i, \tau_i)|(v, f, s^2, \alpha)$ as in (5). However, notice that we are actually drawing $\theta_i = \mathrm{E}\left[v_i\right] = \frac{\mu_i}{s(\boldsymbol{x}_i)}$ rather than μ_i. Therefore, we define $\mu_i \equiv s(\boldsymbol{x}_i)\theta_i$ in the training cohort. And, since μ_i is random, we define it by analogy $\mu_j^* \equiv s(\boldsymbol{x})\theta_j^*$ in other calculations such as that shown in (6).

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

[Figure 4 about here.]

[Figure 5 about here.]

[Figure 6 about here.]

[Figure 7 about here.]

[Figure 8 about here.]

[Figure 9 about here.]

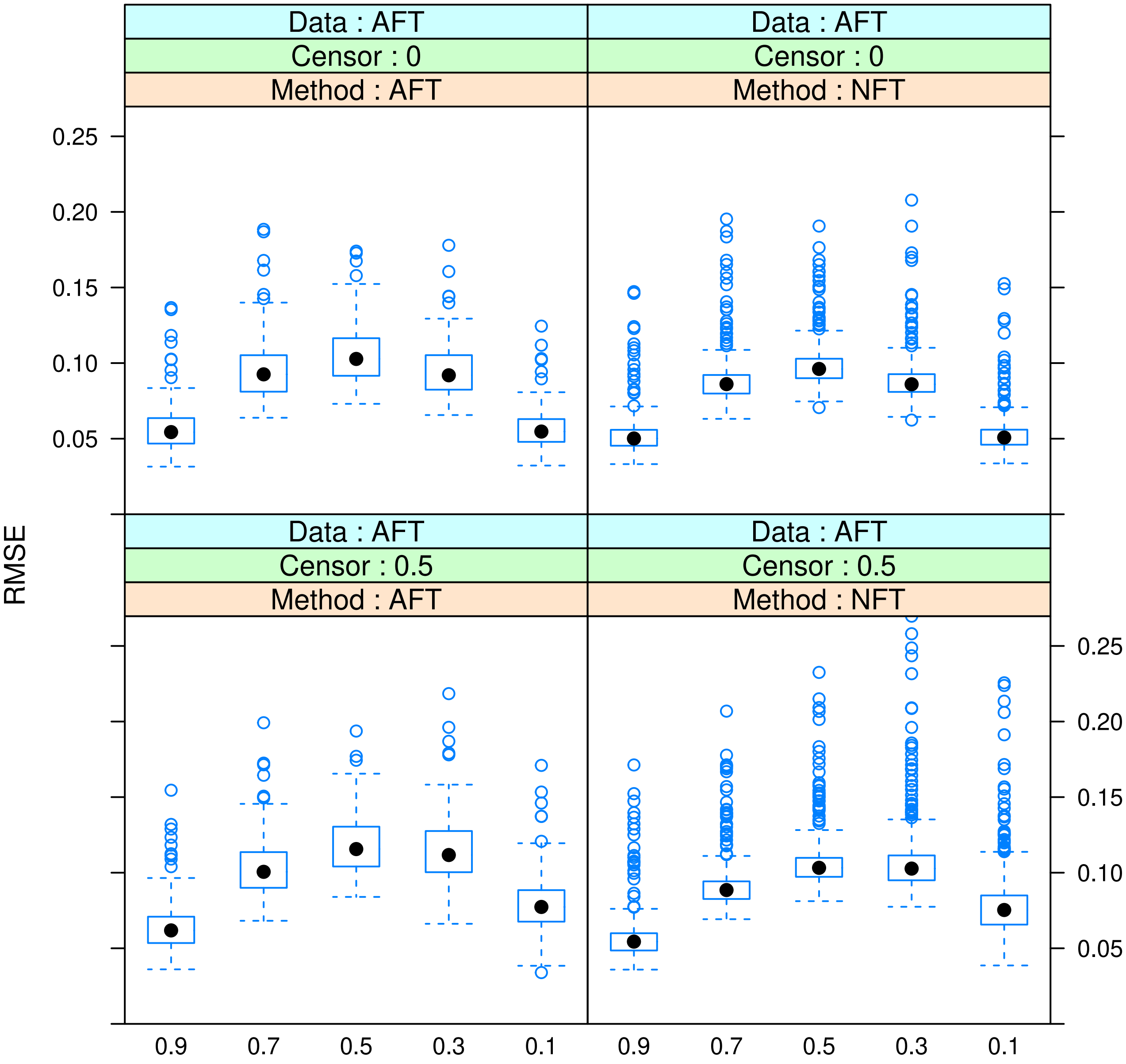

Figure 1. Results of a simulation study comparing AFT BART to NFT BART with sample size 2000. RMSE is on the vertical axis and survival settings are on the horizontal axis. This figure consists of data generated from the AFT scenario. The left (right) column are the results for AFT (NFT) BART. The top (bottom) row are for data generated with 0% (50%) censoring.

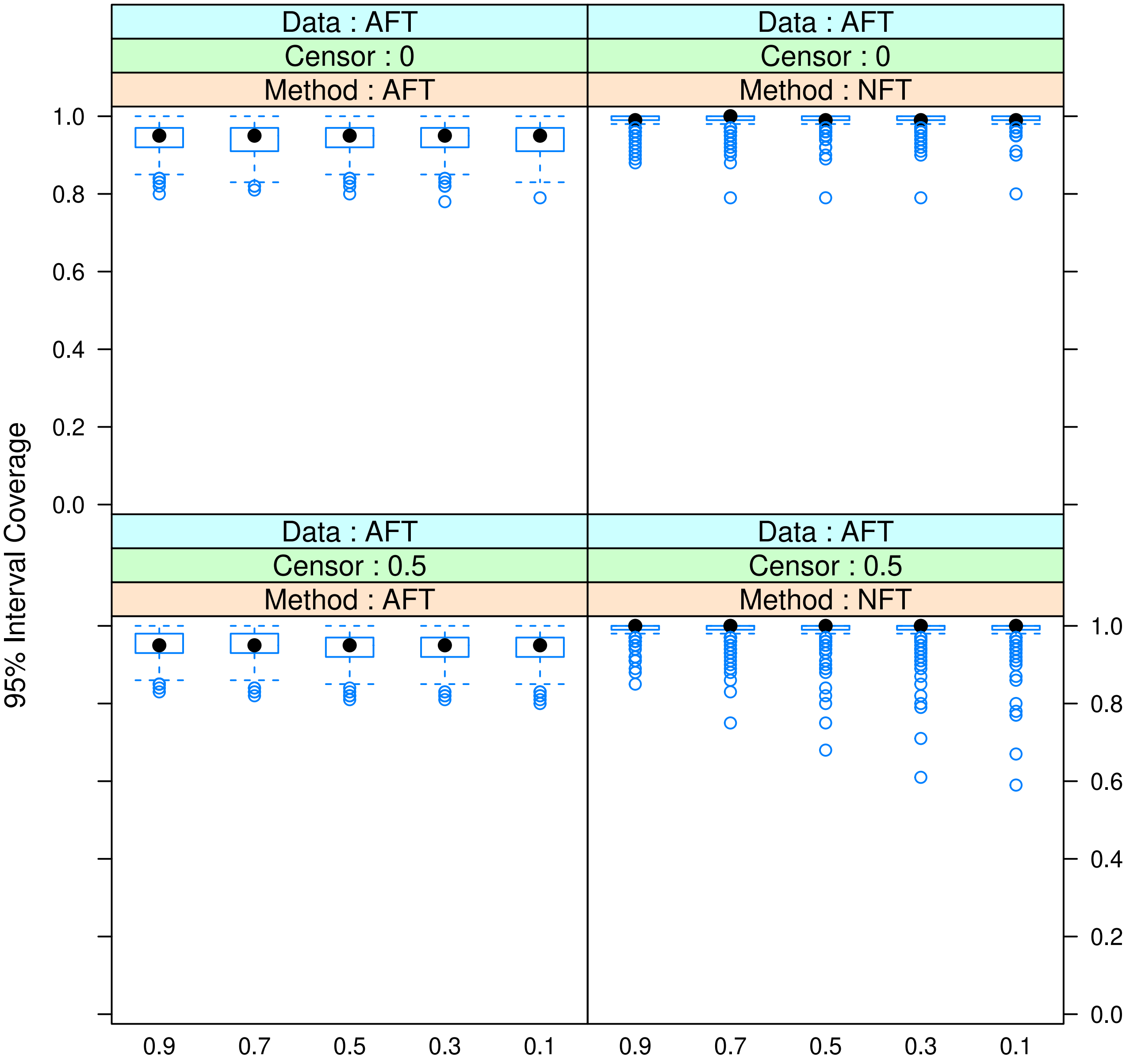

Figure 2. Results of a simulation study comparing AFT BART to NFT BART with sample size 2000. 95% interval coverage is on the vertical axis and survival settings are on the horizontal axis. This figure consists of data generated from the AFT scenario. The left (right) column are the results for AFT (NFT) BART. The top (bottom) row are for data generated with 0% (50%) censoring.

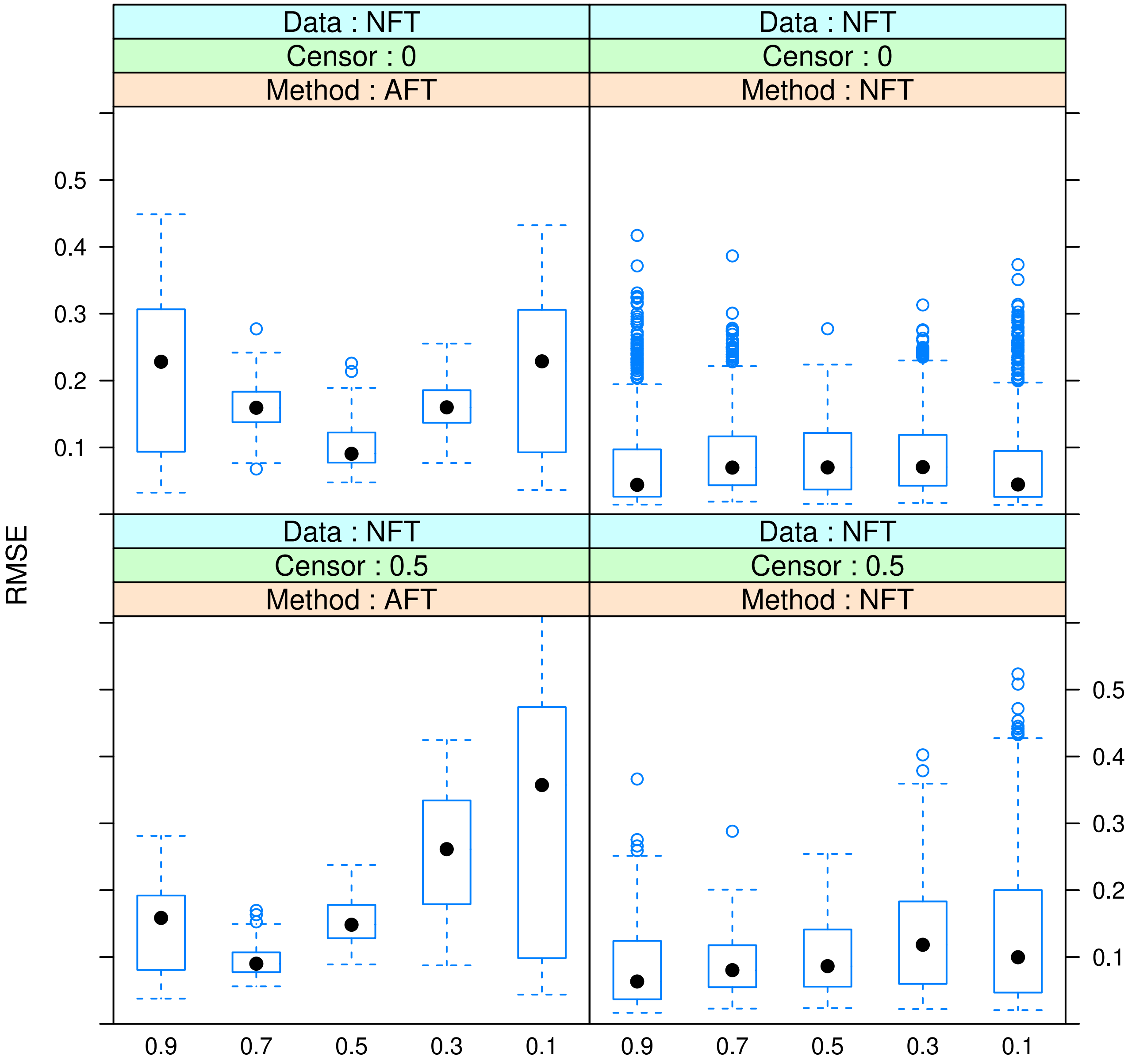

Figure 3. Results of a simulation study comparing AFT BART to NFT BART with sample size 2000. RMSE is on the vertical axis and survival settings are on the horizontal axis. This figure consists of data generated from the NFT scenario. The left (right) column are the results for AFT (NFT) BART. The top (bottom) row are for data generated with 0% (50%) censoring.

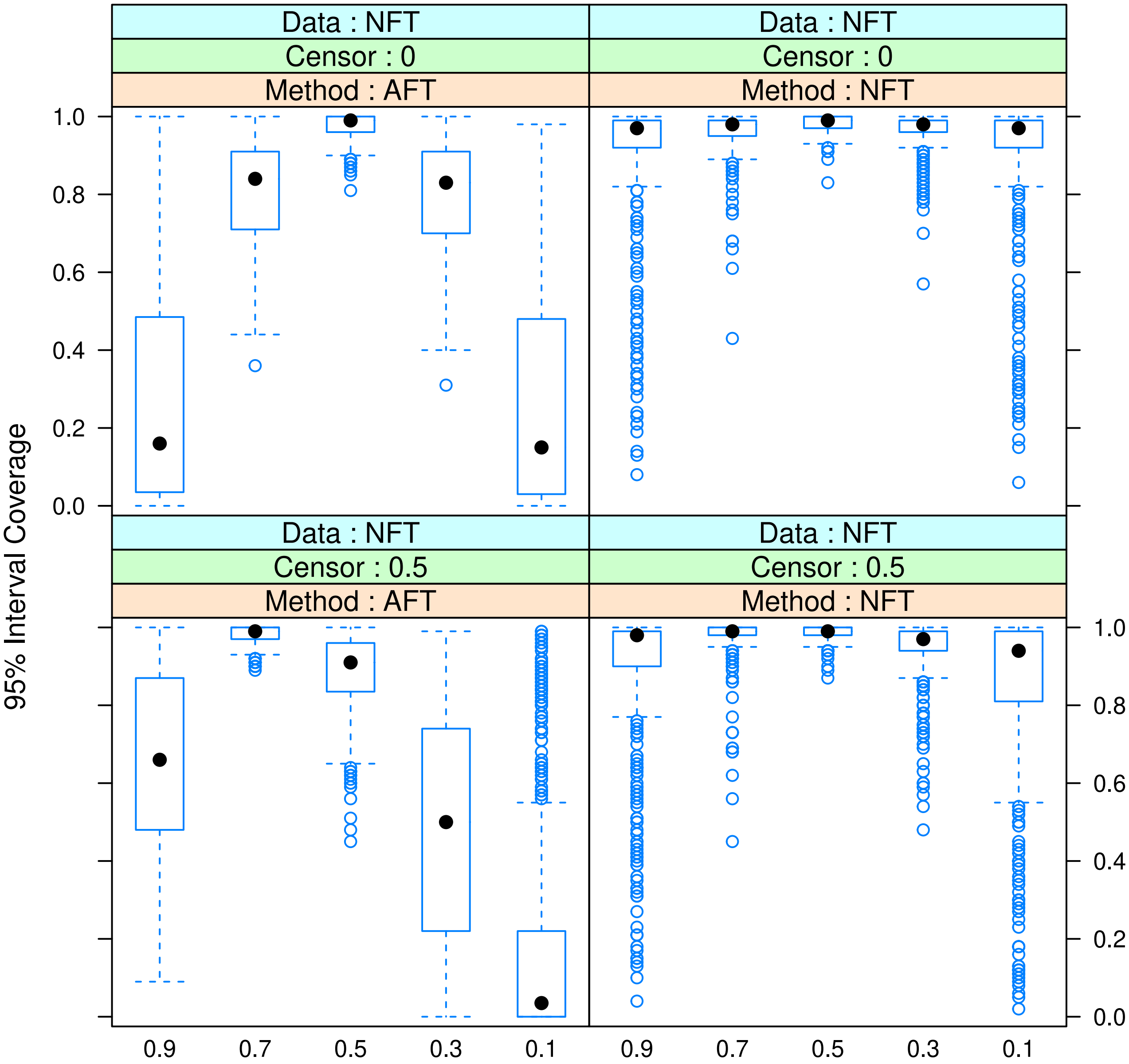

Figure 4. Results of a simulation study comparing AFT BART to NFT BART with sample size 2000. 95% interval coverage is on the vertical axis and survival settings are on the horizontal axis. This figure consists of data generated from the NFT scenario. The left (right) column are the results for AFT (NFT) BART. The top (bottom) row are for data generated with 0% (50%) censoring.

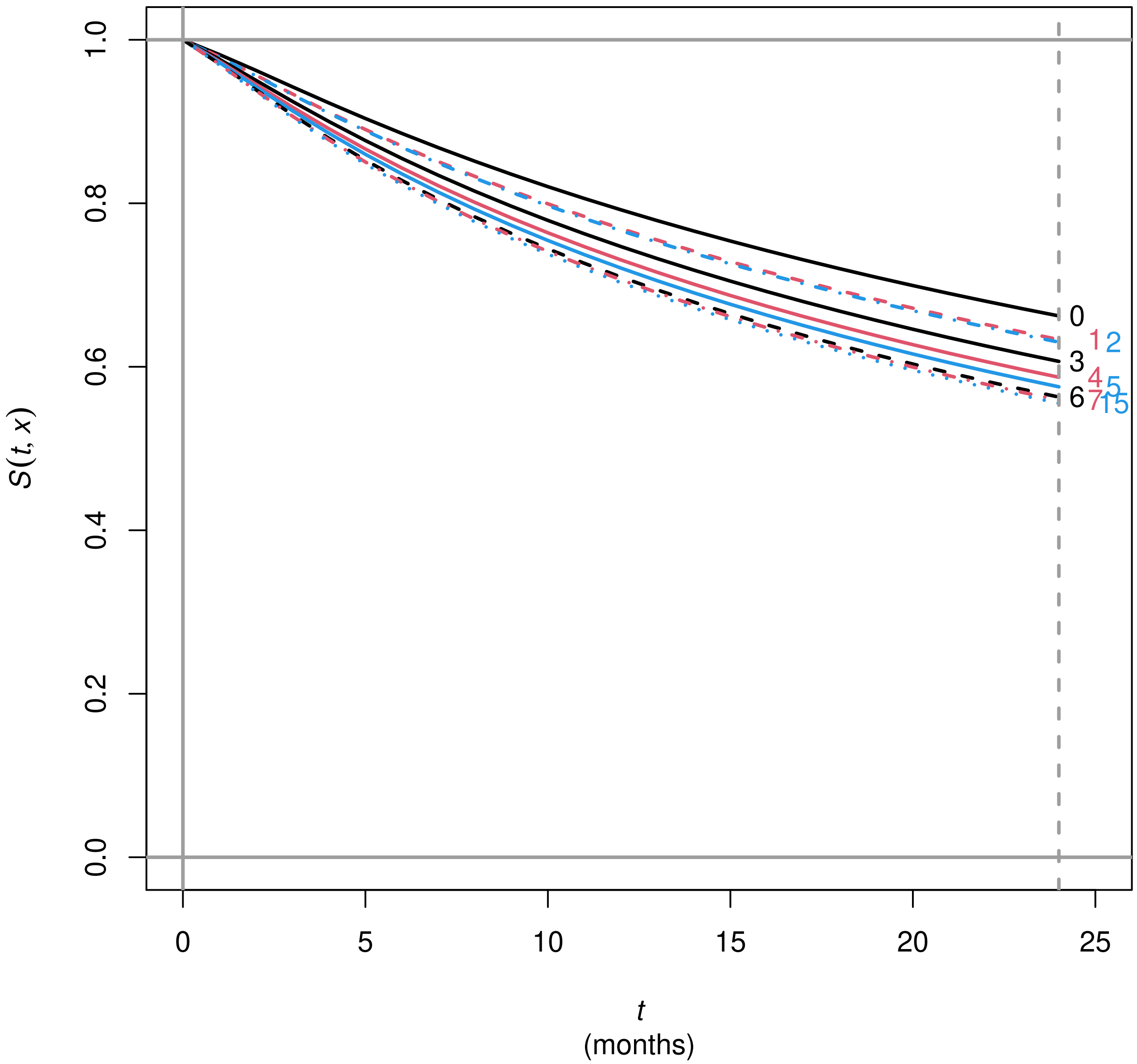

Figure 5. Hematopoietic stem cell transplant treatment for blood-borne cancer. The marginal effect due to the comorbidity index is inversely proportional to survival until approaching an asymptote at the value of 6 (as calculated by the NFT BART model with Friedman's partial dependence function). Survival is on the y-axis and months on the x-axis.

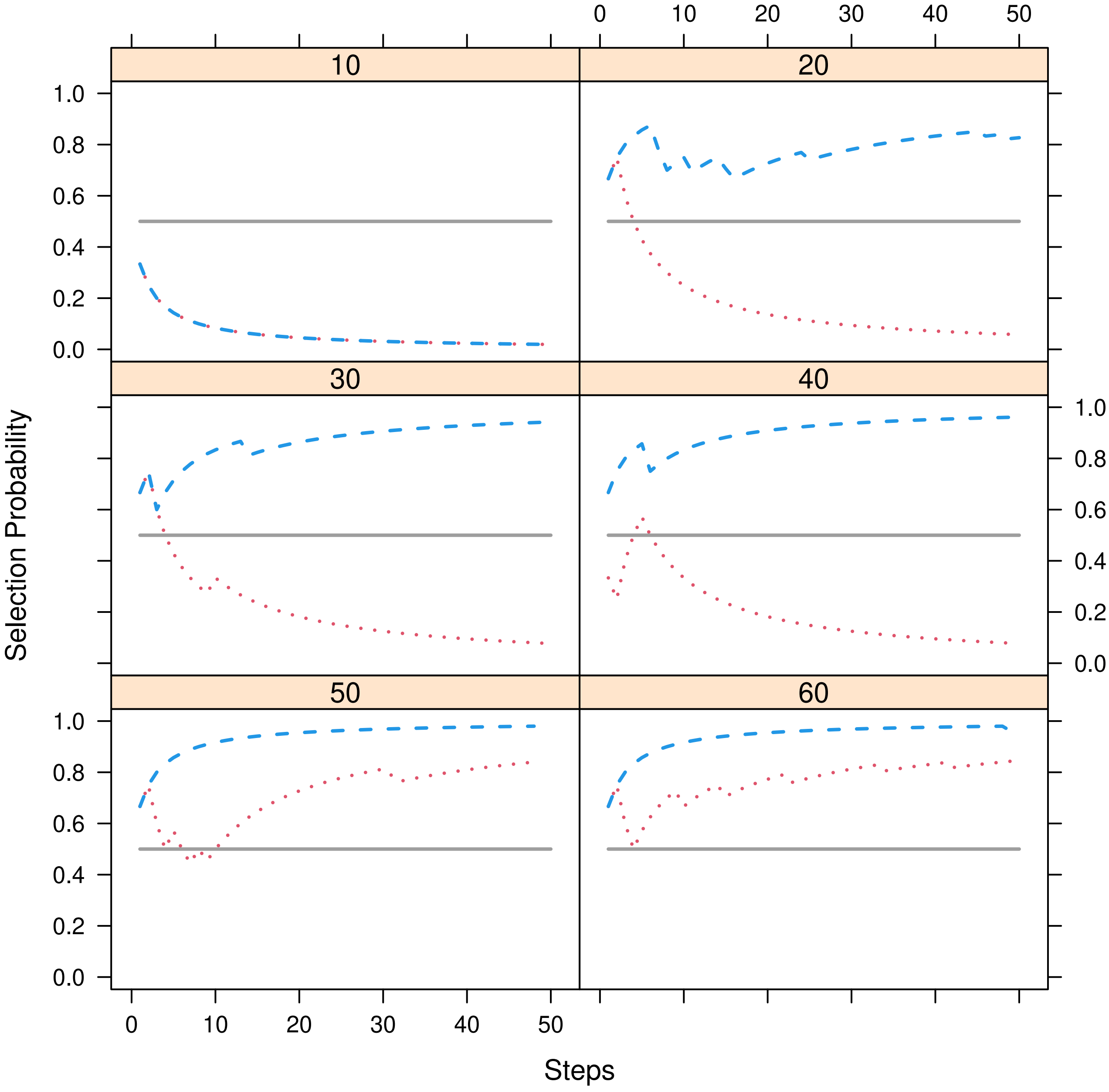

Figure 6. Hematopoietic stem cell transplant treatment for blood-borne cancer. Thompson sampling variable selection (TSVS) was performed on all of the 45 covariates. Here we restrict our attention to the donor characteristics. We performed a series of TSVS inferences by varying the number of BART trees: H = 10, 20, 30, 40, 50 and 60 that are depicted in the six cells above. Variable selection probability is on the y-axis and the TSVS steps are on the x-axis. Only donor age (blue dashed line) and donor sex/child-bearing parity (red dotted line) exceed 0.5 (solid gray line) that is the TSVS decision threshold by the last step of the algorithm. Furthermore, donor age exceeds 0.5 starting with 20 trees while donor sex/parity doesn't achieve that until 50 trees are considered, i.e., donor age is the more important covariate among these two.

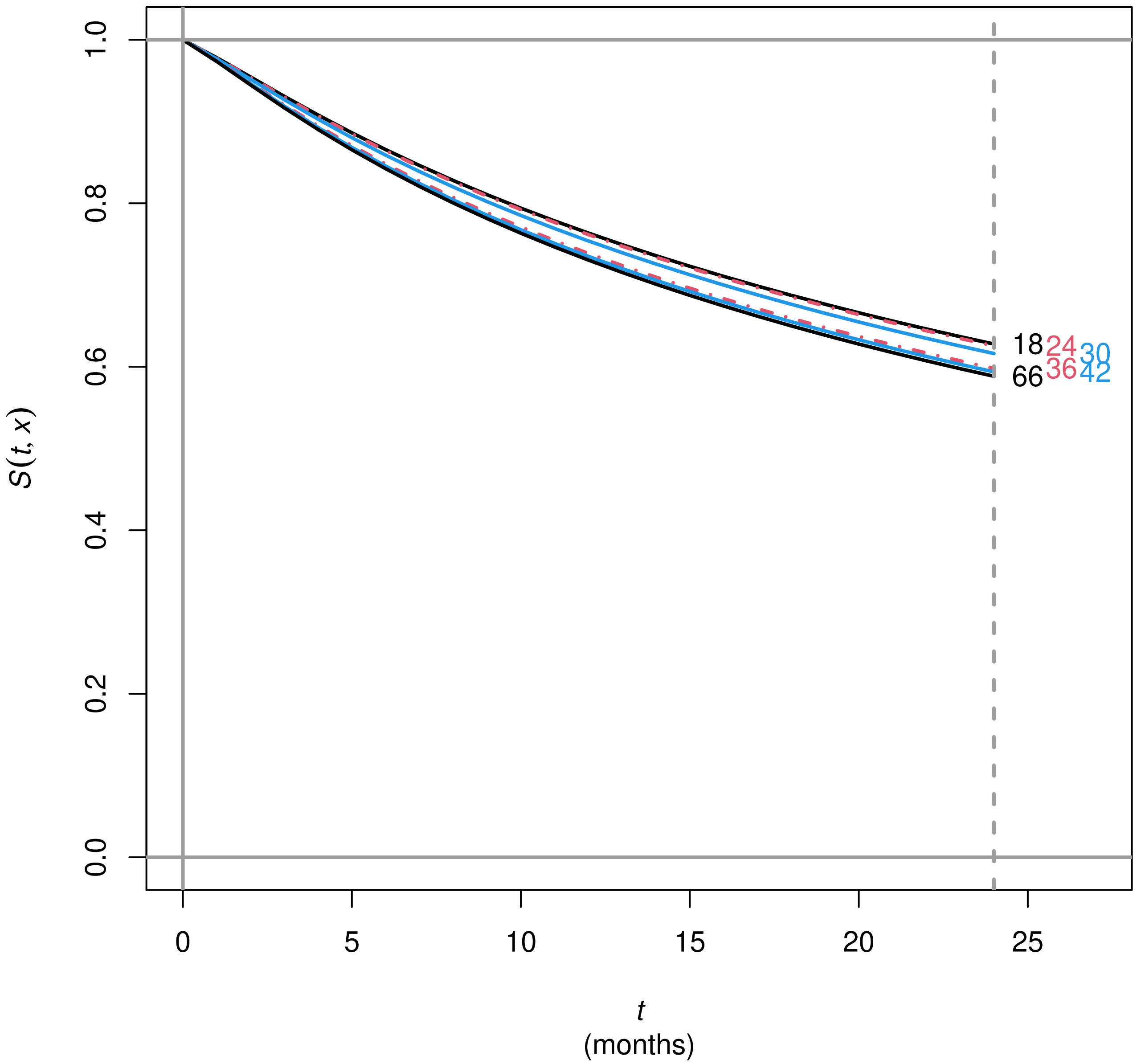

Figure 7. Hematopoietic stem cell transplant treatment for blood-borne cancer. The marginal effect due to the donor's age is inversely proportional to survival until approaching an asymptote at the value of 42 (as calculated by the NFT BART model with Friedman's partial dependence function). Survival is on the y-axis and months on the x-axis.

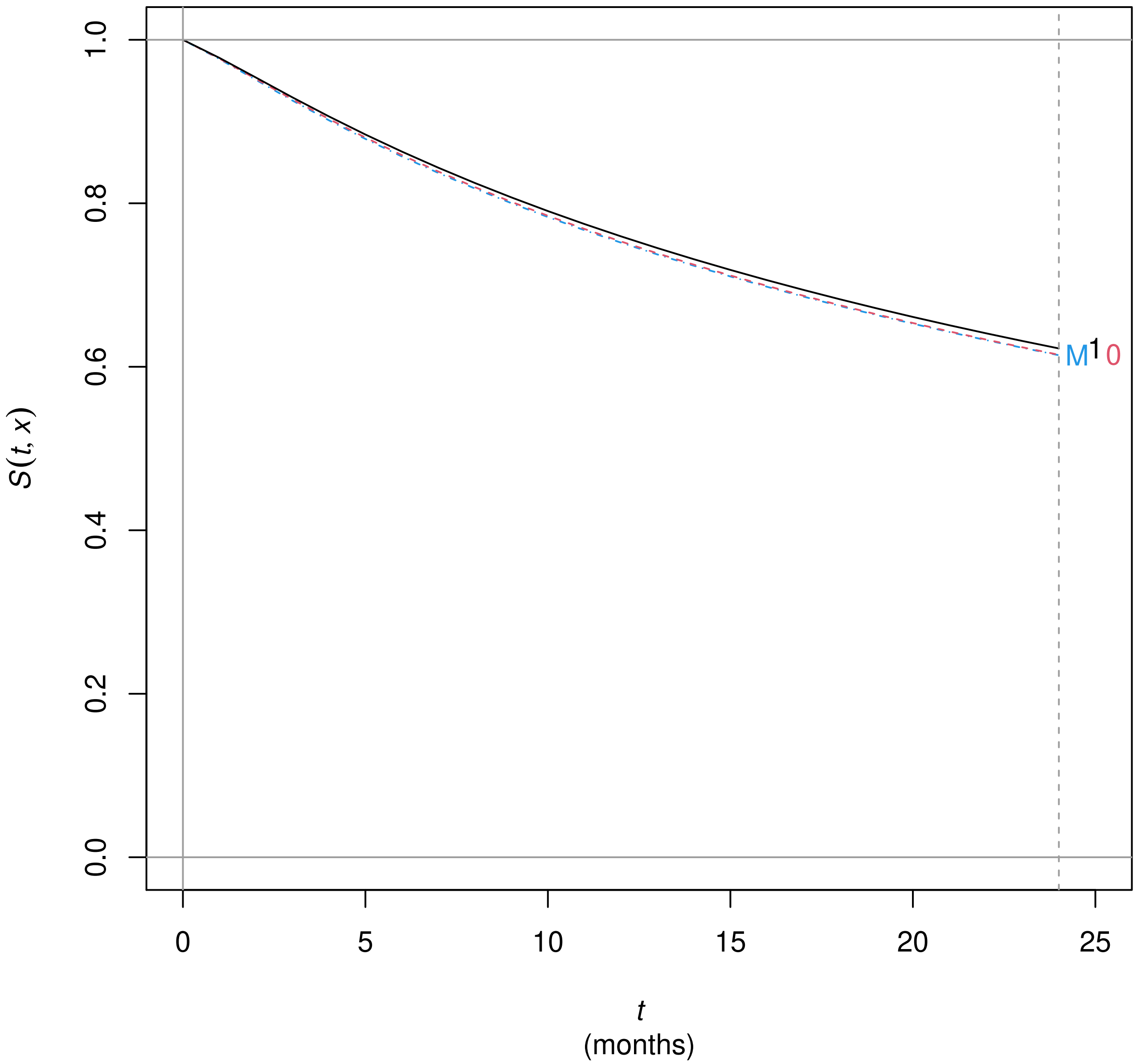

Figure 8. Hematopoietic stem cell transplant treatment for blood-borne cancer. The marginal effect due to the sex/child-birth parity of the donor for male recipients: the marginal effect for female recipients is roughly equivalent but not shown (as calculated by the NFT BART model with Friedman's partial dependence function). Male donors are denoted by "M" with a dashed blue line. Female donors who are nulliparous (parous) are denoted by "0" ("1") with a dashed red (solid black) line. Survival is on the y-axis and months on the x-axis.

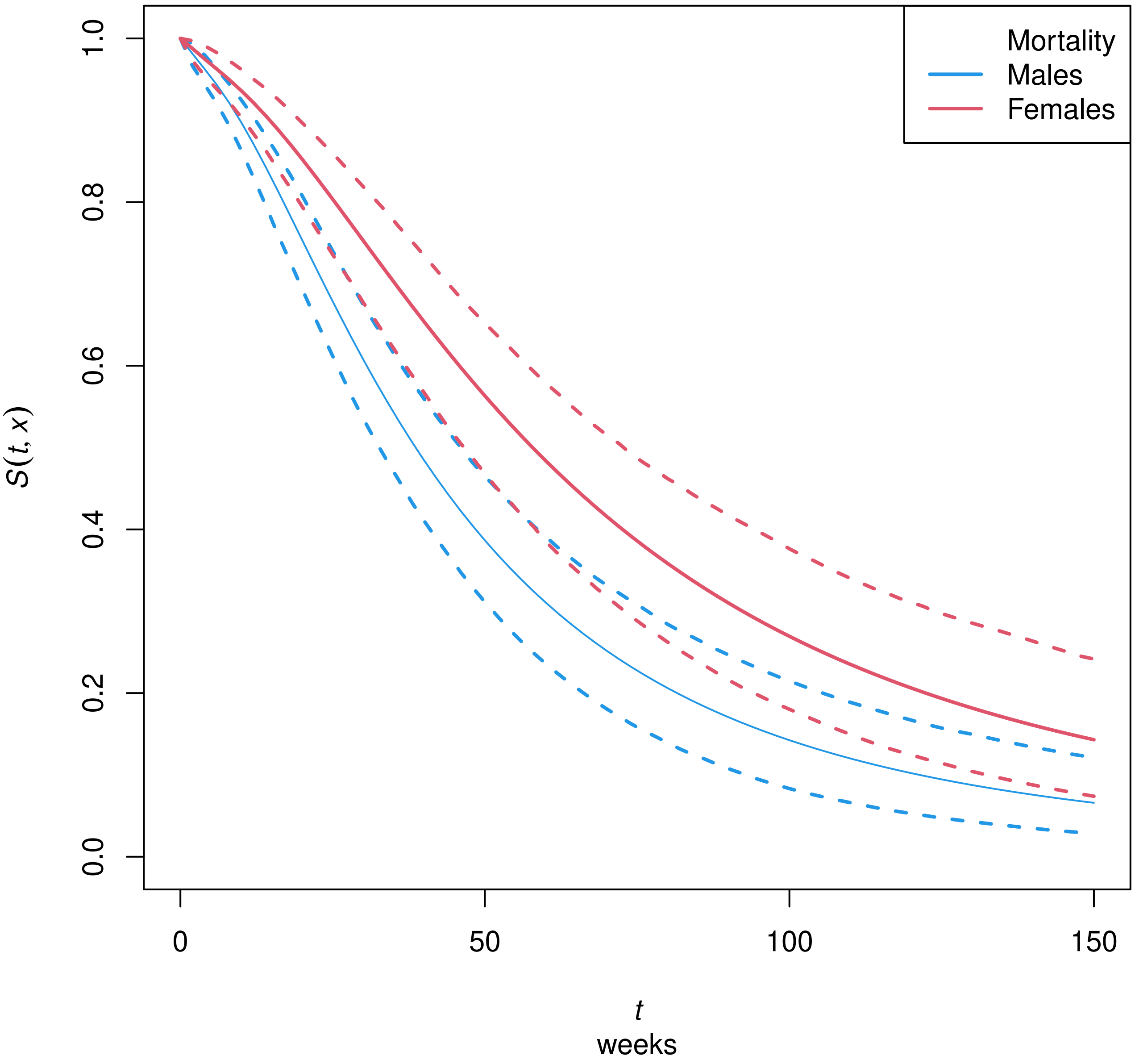

Figure 9. Advanced lung cancer study example: males vs. females. Two-hundred and twenty-eight patients with lung cancer were followed by the North Central Cancer Treatment Group for a median of roughly one year: 138 male and 90 female participants. For this data set, statistical inference was performed with NFT BART for the collected covariates including age, sex, daily activity performance scores, diet and weight-loss information. The solid lines summarize the survival marginal effect for males (blue) and females (red) where the dashed lines are 95% credible intervals.