MCW Biostatistics Technical Report 72
Nonparametric Failure Time: Time-to-event Machine Learning with Heteroskedastic Bayesian Additive Regression Trees and Low Information Omnibus Dirichlet Process Mixtures

R. Sparapani, B. Logan and P. Laud

December 2, 2021

1 Introduction

For generations, biostatisticians have doggedly pursued increasingly assumption-free regression methodology for time-to-event outcomes. Notable achievements along the way consist of such stalwarts of modern applied frequentist methods as proportional hazards [Cox, 1972], accelerated failure time (AFT) [Miller, 1976, Buckley and James, 1979, Aitkin, 1981, Koul et al., 1981, Miller and Halpern, 1982] and random survival forests [Ishwaran et al., 2008] that span from linear models to the modern non-linear machine learning era. While we recognize these breakthroughs, herein we take an alternative approach based on Bayesian nonparametric methodology. Recently, due to modern advances in computational hardware, there has been increasing interest in Bayesian nonparametric methods such as that provided by Bayesian additive regression trees (BART) [Chipman et al., 2010] for time-to-event outcomes: survival analysis [Bonato et al., 2011, Sparapani et al., 2016, Henderson et al., 2020]. Rather than denigrating the previous work in this area, here we laud it by building upon the past and pushing the envelope beyond its present boundaries as so many researchers have done before us.

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. Ensembles are the best known for out-of-sample predictive performance [Baldi and Brunak, 2001, Kuhn and Johnson, 2013], e.g., BART will not over-fit to the training data at the expense of hampering predictive performance for the unseen validation data. And BART very naturally translates to high-dimensional data sets via the incorporation of a sparse Dirichlet prior [Linero, 2018, Liu and Ročková, 2021].

With respect to survival analysis, current BART methods all suffer from one or more issues that warrant improvement. The work of Bonato et al. [2011] implement methods for proportional hazards, AFT and Weibull regression (all of which are restrictive assumptions); furthermore, these lack the flexibility of nonparametric methods. 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, AFT is a precarious restrictive assumption that still remains. 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, a computational implementation is not readily available hampering its use.

All of the above mentioned deficiencies are largely resolved by our new, novel time-to-event approach introduced here that we call nonparametric failure time (NFT) BART. NFT BART builds on a Bayesian nonparametric foundation consisting of BART, heteroskedastic BART [Pratola et al., 2020] and DPM. In particular, the ability to scale to larger sample sizes is important for many applications. And, we provide convenient, user-friendly, computer software that is freely available as a reference implementation.

This report 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 in Section 2 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 with Harrell's c-index, model comparison via Pseudo-Bayes factors and Thompson sampling variable selection. The results of this research are provided in Section 3 which has two parts. A simulation study comparison of AFT BART with NFT BART appears in Section 3.1. We put this research into perspective with a discussion in Section 4. And, finally, we demonstrate the capabilities of the freely available reference software in the Appendix via an example.

2 Methods

2.1 Binary tree regression models and Bayesian additive regression trees

First, we introduce binary tree regression before moving on to BART. Binary tree regression for continuous and categorical outcomes is often referred to as classification and regression trees (CART) [Morgan and Sonquist, 1963, Friedman, 1977, Gordon and Olshen, 1978, Breiman et al., 2017] (we reserve the term CART for frequentist implementations of the methodology as it is commonly used). Chipman et al. [1998] introduced Bayesian binary tree regression models for continuous and categorical outcomes; however, we restrict our attention to continuous outcomes in this investigation. For this introduction, we have the following notation: y_i is a continuous outcome where i indexes subjects $i = 1, \ldots, N$; $\boldsymbol{x}_i$ is a vector of covariates; $\mathcal{T}$ denotes the tree structure and branch decision rules; $\mathcal{M} \equiv \{\mu_1, \mu_2, \ldots, \mu_L\}$ denotes the L leaf values; μ is a constant that centers the data (a typical choice is $\mu = \bar{y}$) and $g(\boldsymbol{x}_i; \mathcal{T}, \mathcal{M})$ is a regression tree function. These models have the following form: $y_i = \mu + g(x_i; \mathcal{T}, \mathcal{M}) + \epsilon_i$ where $\epsilon_i \overset{\text{iid}}{\sim} \text{N}\big(0,\ \sigma^2\big)$. However, the performance of these Bayesian models was no better than frequentist CART which was disappointing since an advance in predictive ability was naturally sought. Nevertheless, Bayesian tree priors were an active and promising area for future research [Chipman et al., 1998, Denison et al., 1998, Pratola, 2016].

During this time period, the machine learning revolution was growing. In particular, ensemble models were discovered [Krogh and Solich, 1997] that lead to machine learning methods such as gradient boosting [Freund and Schapire, 1997, Friedman, 2001] and random forests [Breiman, 2001]. So, it became clear that the logical extension of a Bayesian binary tree is an ensemble of binary trees which has come to be known as BART [Chipman et al., 2010] with improved out-of-sample predictive performance. BART is a sum of binary trees nonparametric machine learning regression model since the relationship between the outcome, y_i, and the covariates, $\boldsymbol{x}_i$, is learned from the data itself (i indexes subjects $i = 1, \ldots, N$). This framework consists of the following (where μ approximately centers the data as above; typically, $\mu = \bar{y}$).

$$y_i \ = \ \mu + f(\boldsymbol{x}_i) + \epsilon_i \qquad\qquad \epsilon_i|\sigma^2 \overset{\text{iid}}{\sim} \text{N}\big(0,\ \sigma^2\big)$$

$$f \overset{\text{prior}}{\sim} \text{BART}(a, b, k, H) \qquad\qquad \sigma^2 \overset{\text{prior}}{\sim} \nu\lambda\chi^{-2}\,(\nu)$$

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

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]. For a detailed discussion of these priors, please refer to the following work [Chipman et al., 2010, Sparapani et al., 2021].

2.2 Heteroskedastic BART

Heteroskedastic BART is an extension to BART where we have both a mean function, f, and a variance function, s^2, to fit [Pratola et al., 2020] (N.B. we assume that there are known constants, w_i^2, that are multiples of the variance; if not, then simply let $w_i \equiv 1$).

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

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

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 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. Furthermore, there is a vector of P observed covariates, $\boldsymbol{x}_i$, along with an unknown vector of regression coefficients, $\boldsymbol{\beta}$, that we intend to estimate. For the purpose of this discussion, let the event times follow a Weibull distribution: $t_i \overset{\text{ind}}{\sim} \text{Weibull}\,(\eta,\ \kappa)$ where $\kappa = \exp(\beta_0 + \boldsymbol{x}_i'\boldsymbol{\beta})$ and $\mathrm{E}\,[t_i] = \kappa^{-1}\Gamma(1 + \eta^{-1})$. Now suppose that we employ the natural logarithm transform, $y_i = \log t_i$, then $y_i \sim \text{ExtremeValue}\left(\log \kappa,\ \eta^{-1}\right)$. Or, if we re-parameterize the AFT model in the form of a linear model, then we have the following.

$$y_i = \log t_i = \beta_0 + \boldsymbol{x}_i'\boldsymbol{\beta} + \epsilon_i \qquad \epsilon_i|\eta \overset{\text{iid}}{\sim} \text{ExtremeValue}\left(0,\ \eta^{-1}\right) \tag{2}$$

The AFT analysis proceeds on the logarithm scale: AFT BART and NFT BART follow the same paradigm as we shall see (further details of the AFT model analysis are omitted for brevity; see Kalbfleisch and Prentice [2002] or Klein and Moeschberger [2003] for more information).

2.4 AFT BART

Although Bonato et al. [2011] proposed an AFT extension to BART, the framework created by Henderson et al. [2020] for AFT BART has the added flexibility of a nonparametric random error term; therefore, we restrict our attention to the latter work. AFT BART replaces the linear regression in AFT with BART and the parametric distribution of random error as follows subject to the constraint $N^{-1}\sum_i \mu_i = 0$ for identifiability.

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

Since some of the y_i are unobserved due to censoring, we set the value $\mu = \hat{\beta}_0$ from an AFT model with no covariates (2). We defer the description of the prior for μ_i until Section 2.6; in particular, as shown in (5), however, the details are not needed for the following points to be made. First, when we have a censored time, then we resort to data augmentation by random draws from the truncated distribution (N.B. left-/interval-censoring can be handled similarly, but we only show draws with respect to right-censoring here merely for notational convenience).

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

AFT BART has nonparametric flexibility allowing it to adapt to the distribution of random error; however, the covariates are only capable of explaining a location-shift which is a result of the restrictive AFT assumption.

2.5 NFT BART

Here we provide an extension, which we call NFT BART, that allows the covariates more nonparametric freedom to explain the time-to-event distribution. From here on, we move fluidly between a parameterization by the precision, τ_i, and by that of 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 subject to the constraints $N^{-1}\sum_i \mu_i = 0$ and $N^{-1}\sum_i \sigma_i^2 = 1$ for identifiability is as follows.

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

As with AFT BART, we set the value $\mu = \hat{\beta}_0$ from an AFT model with no covariates (2). We defer the description of the priors for μ_i and τ_i until Section 2.6; in particular, as shown in (6), but those details are unnecessary to make the following points. First, for censored times, we use data augmentation.

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

NFT BART has more nonparametric flexibility, (μ_i, τ_i), than AFT BART, μ_i, allowing it to better adapt to the distribution of the random error. Furthermore, with NFT BART, the covariates are capable of explaining both a location-shift and a scale change alleviating the restriction of the AFT assumption.

2.6 DPM, constrained DPM and LIO DPM

Both AFT BART and NFT BART are 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]. For AFT BART, the following DPM prior parameter default settings related to μ_i complete the model description [Henderson et al., 2020] (where λ below is the same prior parameter as that shown in (3)).

$$\begin{aligned} \mu_i|G &\overset{\text{prior}}{\sim} G & G|\alpha &\overset{\text{prior}}{\sim} \mathrm{DP}\,(\alpha,\ F_{\mu_0}) \\ \alpha &\overset{\text{prior}}{\sim} \mathrm{Gamma}\,(2,\ 0.1) & \mu_0 &\overset{\text{prior}}{\underset{F}{\sim}} \mathrm{N}(0,\ \lambda) \end{aligned} \tag{5}$$

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 clusters for the purpose of

interpretation, 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 sum to one.

NFT BART follows the Low Information Omnibus (LIO) prior hierarchy for DPM [Shi et al., 2019] to complete the model description for the prior parameter default settings related to (μ_i, τ_i) as follows (N.B. 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).

$$
\begin{aligned}
(\mu_i, \tau_i)|G &\overset{\text{prior}}{\sim} G & G|\alpha &\overset{\text{prior}}{\sim} \text{DP}\left(\alpha,\ F_{(\mu_0,\tau_0|k_0,b_0)}\right) \\
k_0 &\overset{\text{prior}}{\sim} \text{Gamma}\left(1.5,\ 7.5\right) & \mu_0|(\tau_0, k_0) &\overset{\text{prior}}{\underset{F}{\sim}} \text{N}\left(0,\ \tau_0^{-1}k_0^{-1}\right) \\
b_0 &\overset{\text{prior}}{\sim} \text{Gamma}\left(2,\ 1\right) & \tau_0|b_0 &\overset{\text{prior}}{\underset{F}{\sim}} \text{Gamma}\left(3,\ b_0\right)
\end{aligned}
\tag{6}
$$

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

It is important to note that both AFT BART and NFT BART are over-parameterized such that f and (f, s^2), respectively, are 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 μ_i for both AFT BART and NFT BART. We require the constraint $\bar{\mu}_\cdot = N^{-1}\sum_i \mu_i = 0$. Constrained DPM is relatively simple to implement as follows. For NFT BART (or AFT BART), simply draw $(\mu_i, \tau_i)|G$ (or $\mu_i|G$) without constraint defining $\tilde{\mu}_i \equiv \mu_i - \bar{\mu}_\cdot$ and then re-defining $\mu_i = \tilde{\mu}_i$. Similarly, for NFT BART, we require the constraint $\overline{\sigma^2} = N^{-1}\sum_i \sigma_i^2 = 1$ so we 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 AFT BART, we arrive at the following calculation where $\Phi(.)$ is the standard Normal distribution function and $m = 1, \ldots, M$ indexes draws from the posterior.

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

Similarly, for NFT BART we have the following calculation.

$$S_m(t, \boldsymbol{x}) = 1 - \sum_{j=1}^{K_m} w_{jm}\Phi\left(\frac{\log t - \mu - \mu^*_{jm} - f_m(\boldsymbol{x})}{\sigma^*_{jm}s_m(\boldsymbol{x})}\right) \tag{7}$$

From the above, we calculate our survival function estimate by the mean with respect to the posterior as $\hat{S}(t, \boldsymbol{x}) = M^{-1}\sum_m S_m(t, \boldsymbol{x})$ such that both AFT BART and NFT BART estimates are derived accordingly. And, 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; and, by serendipity, both operations are readily accomplished simultaneously as we will see. 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 (inference for AFT BART is calculated analogously).

$$\hat{S}_A(t, \boldsymbol{x}_{jA}) = 1 - M^{-1}N^{-1} \sum_m \sum_{i=1}^{N} \Phi\left(\frac{\log t - \mu - \mu_{im} - f_m(\boldsymbol{x}_{jA}, \boldsymbol{x}_{iB})}{\sigma_{im} s_m(\boldsymbol{x}_{jA}, \boldsymbol{x}_{iB})}\right) \tag{8}$$

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

2.8 Model performance comparison and selection

A challenge in survival analysis is that we do not observe the outcome of each subject; therefore, we are limited to considering the distribution of the outcome with respect to the covariates considered. This limits the approaches avaliable to assess model performance and comparisons along with the relative importance of the covariates. Here, we briefly review the approaches that are applicable to AFT BART and NFT BART.

2.8.1 Harrell's c-index

With respect to a validation data set, model performance can be assessed by the probability of concordance between pairs of event times to their corresponding survival estimates. Such an approach is what has come to be known as Harrell's c-index [Harrell Jr. et al., 1984] which compares the survival probability of all possible pairs of subjects (although, it does not have a Bayesian basis). Let's consider any two patients with the potential for right-censoring: (t_i, δ_i) and (t_j, δ_j) where $i \neq j$ with the minimum time for subject i, i.e., $t_i \leq t_j$. Theoretically, there are no ties for event times; however, in typical studies of cancer treatement, events are often recorded in days (rather than finer increments) so ties do rarely occur. If two patients suffer an event at the same time, then they are non-informative with respect to model concordance since the ordering of their survival probabilities is not evident (these pairs indicated by $a_{ij} = \delta_i \delta_i \mathrm{I}(t_i = t_j)$). In untied circumstances, when $t_i < t_j$ and $\delta_i = 0$, then the pair is also non-informative. The limitation of non-informativeness is a restriction that requires that this comparison be made with respect to a particular data set such as that held out for validation.

For all other pairings, the comparison is informative: tied or untied. For tied, $t_i = t_j$ where $\delta_i = 1$ and $\delta_j = 0$, we denote concordance by $\zeta_{ij} = \mathrm{I}(S(t_i, \boldsymbol{x}_i) < S(t_i, \boldsymbol{x}_j))$, the pair is indicated by $b_{ij} = \delta_i \mathrm{I}(t_i = t_j)$ and the total of such pairs: $B = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (1 - a_{ij})(b_{ij} + b_{ji})$. For untied, $t_i < t_j$ where $\delta_i = 1$, we denote concordance as before, the pair is indicated by $c_{ij} = \delta_i \mathrm{I}(t_i < t_j)$ and the total of such pairs: $C = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (1 - a_{ij})(1 - b_{ij} - b_{ji})(c_{ij} + c_{ji})$. So, Harrell's c-index is provided by the following.

$$(B + C)^{-1} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (1 - a_{ij})(b_{ij}\zeta_{ij} + b_{ji}\zeta_{ji} + (1 - b_{ij} - b_{ji})(c_{ij}\zeta_{ij} + c_{ji}\zeta_{ji}))$$

2.8.2 Model comparison with pseudo-Bayes factors

Model comparison and variable selection go hand-in-hand. 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. So, consider the BF as 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 what is known as 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). Furthermore, the CPO can be approximated conveniently from the posterior samples by $[y_i|y_{-i}, \omega] = \left\{M^{-1}\sum_m [y_i|\theta_{\omega m}, \omega]^{-1}\right\}^{-1}$ [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. N.B. Jeffreys [1961] has suggested thresholds for BF inference which are applicable to PBF as well.

2.8.3 Variable selection with Thompson sampling

BART and variable selection were natural partners from the very beginning [Chipman et al., 2010]. And over time, ever more powerful variable selection techniques have been proposed: 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 give a brief introduction to Thompson sampling variable selection (TSVS) 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.

TSVS builds upon a multi-armed bandit foundation. The algorithm has a Bayesian flavor; although, it is not entirely Bayesian. By randomly choosing arms/variables based on posterior samples of their reward probabilities, TSVS is an amalgam of combinatorial bandits with spike-and-slab variable selection. Bringing together Bayesian reinforcement learning 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. TSVS is an iterative process, as follows, where $k = 1, \ldots, K$ are the number of steps taken.

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

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 $\gamma_{jk} = \mathrm{I}(U_{jkM} + V_{jkM} > 0)$ 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: $\pi_{jk} = \frac{a_{jk}}{a_{jk}+b_{jk}}$.

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

3 Results

3.1 Simulated data sets of known progeny

Via a simulation study, we conducted a comparison between 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: $x_{2j+1} \overset{\text{iid}}{\sim} \mathrm{B}(0.5)$ and $x_{2j} \overset{\text{iid}}{\sim} \mathrm{U}(0, 1)$ where $j = 1, \ldots, 10$. We considered two data generation scenarios: homoskedasitc AFT and heteroskedastic NFT. AFT data was generated by $\log t \sim \mathrm{N}(\mu(x),\ \exp(-4))$ 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 $\log t \sim \mathrm{N}\big(\mu(x),\ \sigma^2(x)\big)$ 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 $c_{ij} = K^{-1} \sum_k \mathrm{I}\Big(\hat{S}_{k,0.025}(t_{ij}, \boldsymbol{x}_i) < S_j < \hat{S}_{k,0.975}(t_{ij}, \boldsymbol{x}_i)\Big)$. 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 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 2, we summarized RMSE and their was a slight advantage in favor of AFT BART as might be expected. In Figure 3, we summarized interval coverage and there was a slight advantage in favor of AFT BART being closer to the 95% level. In Figure 4, we summarized bias and there was a slight advantage in favor of AFT BART as might be expected. In Figure 5, 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 6, we summarized RMSE and their was a considerable improvement in favor of NFT BART as we anticipated. In Figure 7, 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 Figure 8, we summarized bias and there was a considerable advantage in favor of NFT BART as we anticipated. In Figure 9, we summarized the 95% interval length and there was an advantage in favor of NFT BART as we anticipated.

3.1.2 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 Figure 10, we summarized RMSE and their was a slight advantage in favor of AFT BART as might be expected. In Figure 11, we summarized interval coverage and there was a slight advantage in favor of AFT BART being closer to the 95% level. In Figure 12, we summarized bias and there was a slight advantage in favor of AFT BART as might be expected. In Figure 13, 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 Figure 14, we summarized RMSE and their was a considerable improvement in favor of NFT BART as we anticipated. In Figure 15, we summarized interval coverage and there was a considerable advantage in favor of NFT BART being closer to the 95% level. In Figure 16, we summarized bias and there was a considerable advantage in favor of NFT BART as we anticipated. In Figure 17, we summarized the 95% interval length and there was an advantage in favor of NFT BART as we anticipated.

4 Discussion

An intent of this research was to address perceived short-comings in modern time-to-event methodology. For example, parametric survival analysis has been criticized by Bayesians and frequentists alike. Therefore, we build upon a solid Bayesian nonparametric foundation of the DPM LIO prior hierarchy. Furthermore, we avoid precarious restrictive assumptions such as linearity, proportionality and/or AFT by employing heteroskedastic BART.

In this research, we have shown that NFT BART has advantages beyond that of other BART time-to-event methodology in a number of areas; particularly, for data sets of increasingly larger sample sizes. Furthermore, 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 advanced survival analysis outcomes such as recurrent events [Sparapani et al., 2020] and/or competing risks [Sparapani et al., 2020]. 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

Murray Aitkin. A note on the regression analysis of censored data. *Technometrics*, 23(2):161–163, 1981.

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

Justin Bleich, Adam Kapelner, Edward I George, and Shane T Jensen. Variable selection for BART: An application to gene regulation. *The Annals of Applied Statistics*, 8(3):1750–1781, 2014. URL `https://www.jstor.org/stable/24522283`.

Vinicius Bonato, Veerabhadran Baladandayuthapani, Bradley M Broom, Erik P Sulman, Kenneth D Aldape, and Kim-Anh Do. Bayesian ensemble methods for survival prediction in gene expression data. *Bioinformatics*, 27(3):359–367, 2011.

Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001. doi: 10.1023/a:1010933404324.

Leo Breiman, Jerome H Friedman, Richard A Olshen, and Charles J Stone. *Classification and regression trees*. Routledge, 2017.

Jonathan Buckley and Ian James. Linear regression with censored data. *Biometrika*, 66(3):429–436, 1979.

Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. Bayesian CART model search. *Journal of the American Statistical Association*, 93(443):935–948, 1998. doi: 10.1080/01621459.1998.10473750.

Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. BART: Bayesian Additive Regression Trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010. doi: 10.1214/09-aoas285.

D. R. Cox. Regression models and life-tables (with discussions). *Journal of the Royal Statistical Society B*, 34(2):187–220, 1972. URL `https://www.jstor.org/stable/2985181`.

Ton de Waal, Jeroen Pannekoek, and Sander Scholtus. *Handbook of Statistical Data Editing and Imputation*. John Wiley & Sons, Hoboken, NJ, 2011. doi: 10.1002/9780470904848.

David GT Denison, Bani K Mallick, and Adrian FM Smith. A Bayesian CART Algorithm. *Biometrika*, 85(2):363–377, 1998. doi: 10.1093/biomet/85.2.363.

Michael D Escobar and Mike West. Bayesian density estimation and inference using mixtures. *Journal of the American statistical association*, 90(430):577–588, 1995.

Ludwig Fahrmeir. Discrete survival-time models. *Wiley StatsRef: Statistics Reference Online*, 2014. doi: 10.1002/9781118445112.stat06012.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997. doi: 10.1006/jcss.1997.1504.

Jerome H Friedman. A recursive partitioning decision rule for nonparametric classification. *IEEE Transactions on Computers*, 26(4):404–408, 1977.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. URL `http://www.jstor.org/stable/2699986`.

Seymour Geisser and William F Eddy. A predictive approach to model selection. *Journal of the American Statistical Association*, 74(365):153–160, 1979.

Alan E Gelfand and Dipak K Dey. Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(3):501–514, 1994.

Alan E Gelfand and Adrian FM Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409, 1990. doi: 10.1080/01621459.1990.10476213.

Junxian Geng, Anirban Bhattacharya, and Debdeep Pati. Probabilistic community detection with unknown number of communities. *Journal of the American Statistical Association*, 114(526): 893–905, 2019.

Louis Gordon and Richard A Olshen. Asymptotically efficient solutions to the classification problem. *The Annals of Statistics*, 6(3):515–533, 1978.

PR Hahn and CM Carvalho. Decoupling shrinkage and selection in Bayesian linear models: a posterior summary perspective. *Journal of the American Statistical Association*, 110(509):435–448, 2015. doi: 10.1080/01621459.2014.993077.

Frank E Harrell Jr., Kerry L Lee, Robert M Califf, David B Pryor, and Robert A Rosati. Regression modelling strategies for improved prognostic prediction. *Statistics in medicine*, 3(2):143–152, 1984.

Nicholas C Henderson, Thomas A Louis, Gary L Rosner, and Ravi Varadhan. Individualized treatment effects with censored data via fully nonparametric Bayesian accelerated failure time models. *Biostatistics*, 21(1):50–68, 2020.

Hemant Ishwaran and Lancelot F James. Approximate Dirichlet process computing in finite normal mixtures: smoothing and prior information. *Journal of Computational and Graphical statistics*, 11(3):508–532, 2002.

Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. Random survival forests. *The annals of applied statistics*, 2(3):841–860, 2008.

Sonia Jain and Radford M Neal. Splitting and merging components of a nonconjugate Dirichlet process mixture model. *Bayesian Analysis*, 2(3):445–472, 2007.

Harold Jeffreys. *The theory of probability*. OUP Oxford, 1961.

JD Kalbfleisch and RL Prentice. *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, Hoboken, NJ, 2nd edition, 2002. doi: 10.1002/9781118032985.

Maria Kalli, Jim E Griffin, and Stephen G Walker. Slice sampling mixture models. *Statistics and computing*, 21(1):93–105, 2011.

Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the American statistical association*, 90(430):773–795, 1995.

John P Klein and Melvin L Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer-Verlag, New York, NY, 2nd edition, 2003. doi: 10.1007/b97377.

H Koul, Vyaghreswarudu Susarla, and John Van Ryzin. Regression analysis with randomly right-censored data. *The Annals of statistics*, 9(6):1276–1288, 1981.

A Krogh and P Solich. Statistical mechanics of ensemble learning. *Physical Review E*, 55(1):811–825, 1997. doi: 10.1103/physreve.55.811.

M Kuhn and K Johnson. *Applied Predictive Modeling*. Springer-Verlag, New York, NY, 2013. doi: 10.1007/978-1-4614-6849-3.

A. Linero. Bayesian regression trees for high dimensional prediction and variable selection. *Journal of the American Statistical Association*, 113(522):626–636, 2018. doi: 10.1080/01621459.2016.1264957.

Antonio R Linero, Piyali Basak, Yinpu Li, and Debajyoti Sinha. Bayesian survival tree ensembles with submodel shrinkage. *Bayesian Analysis*, (ahead of print)(1):1–24, 2021.

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

Charles Lawrence Loprinzi, John A Laurie, H Sam Wieand, James E Krook, Paul J Novotny, John W Kugler, Joan Bartel, Marlys Law, Marilyn Bateman, and Nancy E Klatt. Prospective Evaluation of Prognostic Variables from Patient-Completed Questionnaires. North Central Cancer Treatment Group. *Journal of Clinical Oncology*, 12(3):601–607, 1994.

Rupert Miller and Jerry Halpern. Regression with censored data. *Biometrika*, 69(3):521–531, 1982.

Rupert G Miller. Least squares regression with censored data. *Biometrika*, 63(3):449–464, 1976.

James N Morgan and John A Sonquist. Problems in the analysis of survey data, and a proposal. *Journal of the American statistical association*, 58(302):415–434, 1963.

Radford M Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265, 2000.

Matthew T Pratola. Efficient Metropolis–Hastings Proposal Mechanisms for Bayesian Regression Tree Models. *Bayesian Analysis*, 11(3):885–911, 2016. doi: 10.1214/16-ba999.

Matthew T Pratola, Hugh A Chipman, Edward I George, and Robert E McCulloch. Heteroscedastic BART via multiplicative regression trees. *Journal of Computational and Graphical Statistics*, 29 (2):405–417, 2020.

Daniel J Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. et almbox. 2018. *A tutorial on Thompson sampling. Foundations and Trends in Machine Learning*, 11:1, 2018.

Yushu Shi, Michael Martens, Anjishnu Banerjee, and Purushottam Laud. Low information omnibus (LIO) priors for Dirichlet process mixture models. *Bayesian Analysis*, 14(3):677–702, 2019.

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

R Sparapani, L Rein, S Tarima, T Jackson, and J Meurer. Non-parametric recurrent events analysis with BART and an application to the hospital admissions of patients with diabetes. *Biostatistics*, 21(1):69–85, 2020. doi: 10.1093/biostatistics/kxy032.

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

Rodney A Sparapani, Brent R Logan, Robert E McCulloch, and Purushottam W Laud. Nonparametric survival analysis using Bayesian Additive Regression Trees (BART). *Statistics in Medicine*, 35(16):2741–2753, 2016. doi: 10.1002/sim.6893.

Yaoyuan Vincent Tan and Jason Roy. Bayesian additive regression trees and the General BART model. *Statistics in medicine*, 38(25):5048–5069, 2019.

Dandan Xu, Michael J Daniels, and Almut G Winterstein. Sequential BART for imputation of missing covariates. *Biostatistics*, 17(3):589–602, 2016. doi: 10.1093/biostatistics/kxw009.

Mingan Yang, David B Dunson, and Donna Baird. Semiparametric Bayes hierarchical models with mean and variance constraints. *Computational statistics & data analysis*, 54(9):2172–2186, 2010.

A 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`.

A.1 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 (8). As we can see in Figure 1, 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
```

B 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 $r_i|(f, s^2, \mu_i, \tau_i) \sim \mathrm{N}\big(f(\boldsymbol{x}_i),\ s^2(\boldsymbol{x}_i)\sigma_i^2\big)$. 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 $u_i|(f, s^2, \mu_i, \tau_i) \sim \mathrm{N}\big(0,\ s^2(\boldsymbol{x}_i)\big)$. 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 $v_i|(f, s^2, \mu_i, \tau_i) \sim \mathrm{N}\Big(\frac{\mu_i}{s(\boldsymbol{x}_i)},\ \sigma_i^2\Big)$. 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 (6). However, notice that we are actually drawing $\theta_i = \mathrm{E}[v_i] = \frac{\mu_i}{s(\boldsymbol{x}_i)}$ 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 (7).

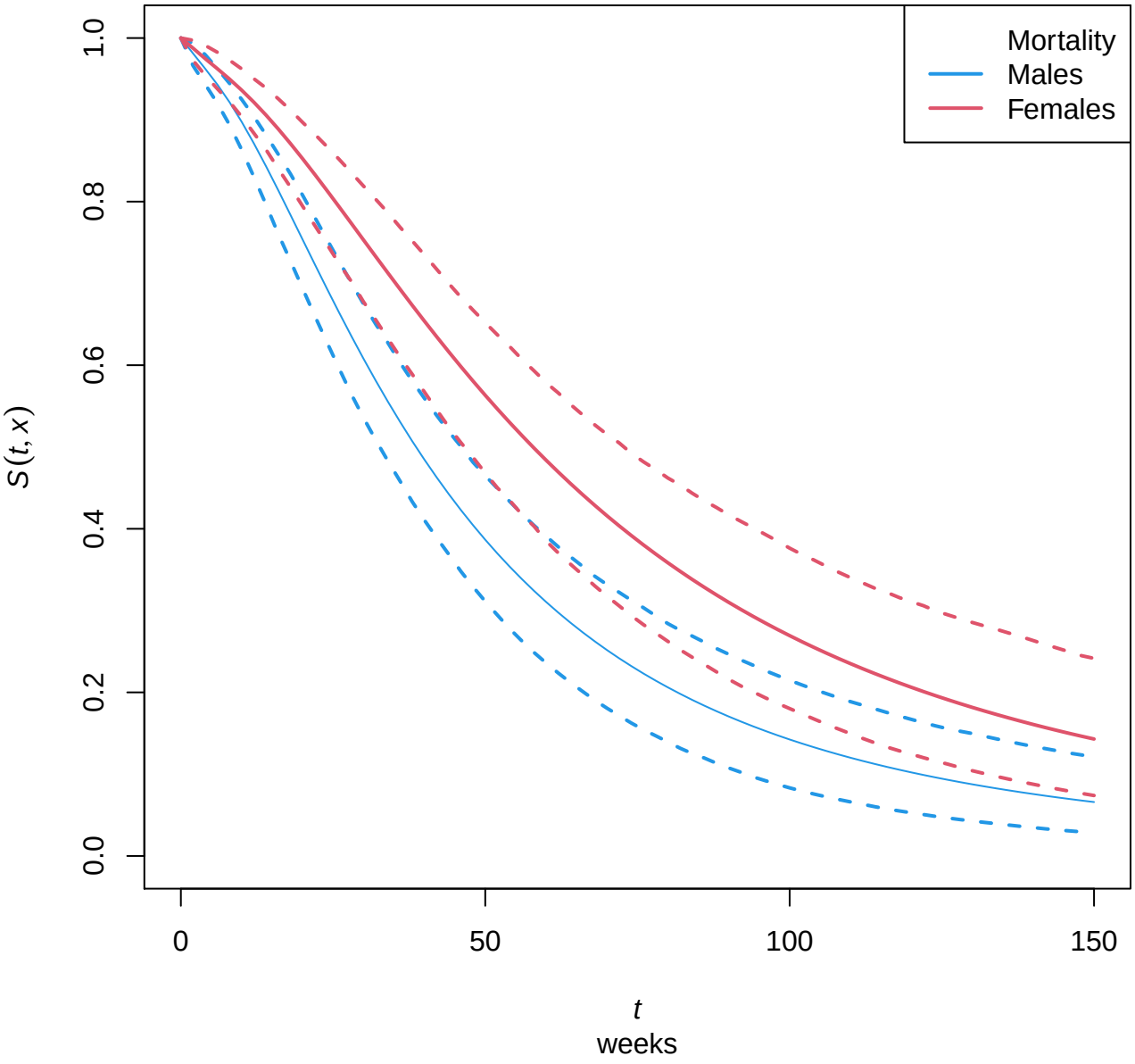

Figure 1: Advanced lung cancer study example: males vs. females. Two-hundred and twenty-eight patients with lung cancer were followed by the North Central Cancer Treatment Group for a median of roughly one year: 138 male and 90 female participants. For this data set, statistical inference was performed with NFT BART for the collected covariates including age, gender, 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.

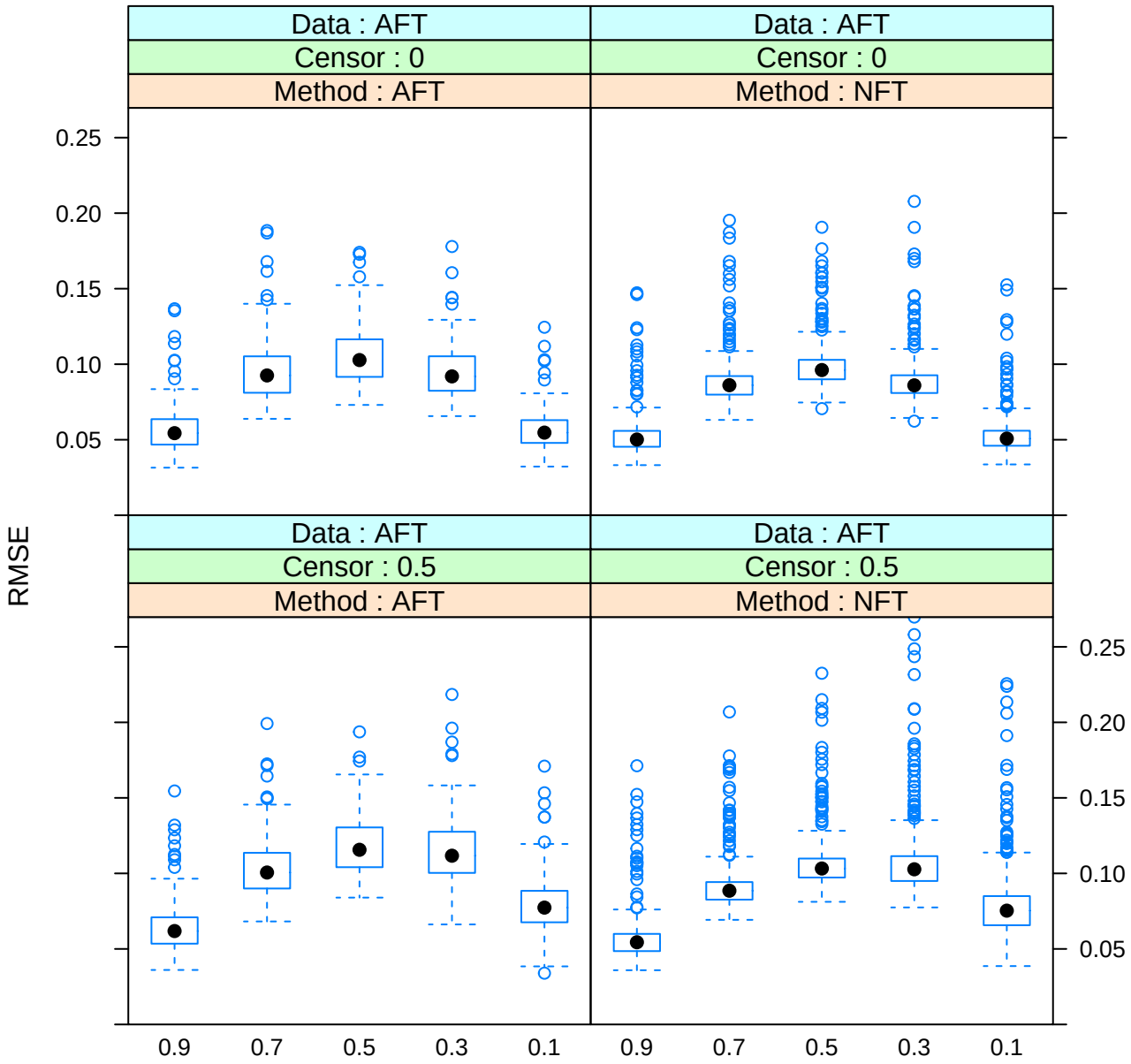

Figure 2: Results of a simulation study comparing AFT BART to NFT BART with sample size 2000. RMSE is on the vertical axis and survival settings are on the horizontal axis. This figure consists of data generated from the AFT scenario. The left (right) column are the results for AFT (NFT) BART. The top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

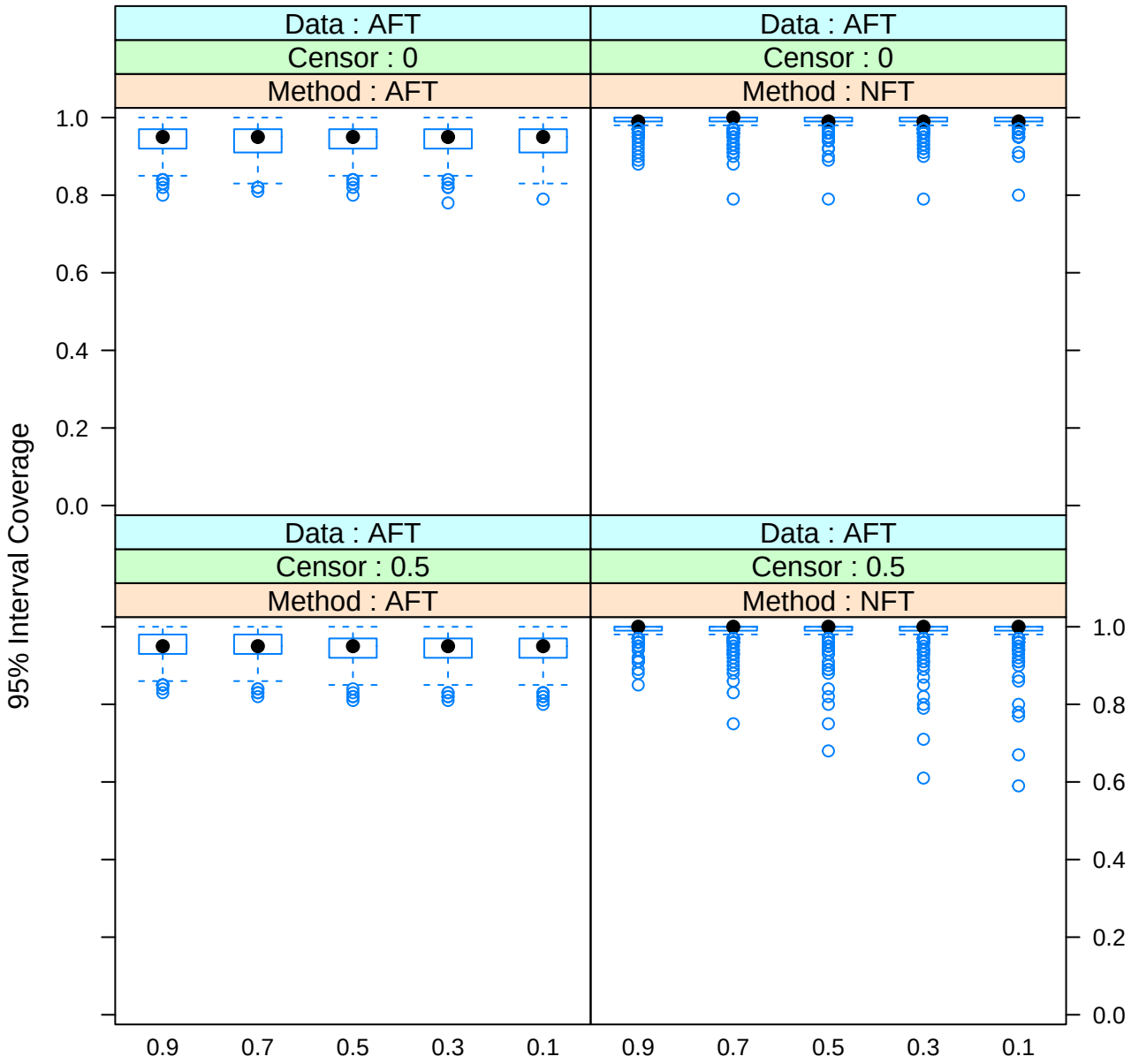

Figure 3: Results of a simulation study comparing AFT BART to NFT BART with sample size 2000. 95% interval coverage is on the vertical axis and survival settings are on the horizontal axis. This figure consists of data generated from the AFT scenario. The left (right) column are the results for AFT (NFT) BART. The top (bottom) row are for data generated with 0% (50%) censoring.

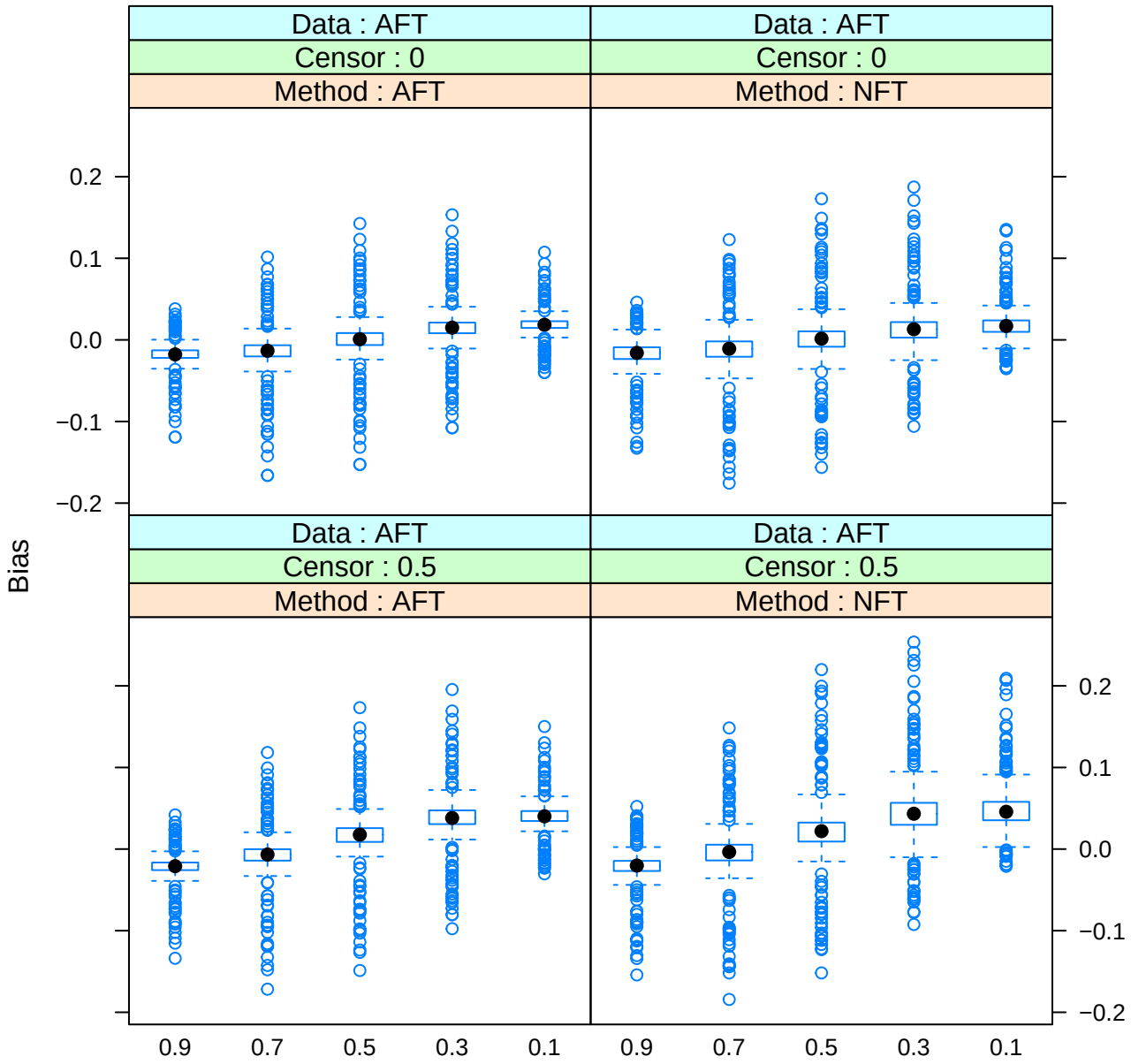

Figure 4: Results of a simulation study comparing AFT BART to NFT BART with sample size 2000. Bias is on the vertical axis and survival settings are on the horizontal axis. This figure consists of data generated from the AFT scenario. The left (right) column are the results for AFT (NFT) BART. The top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

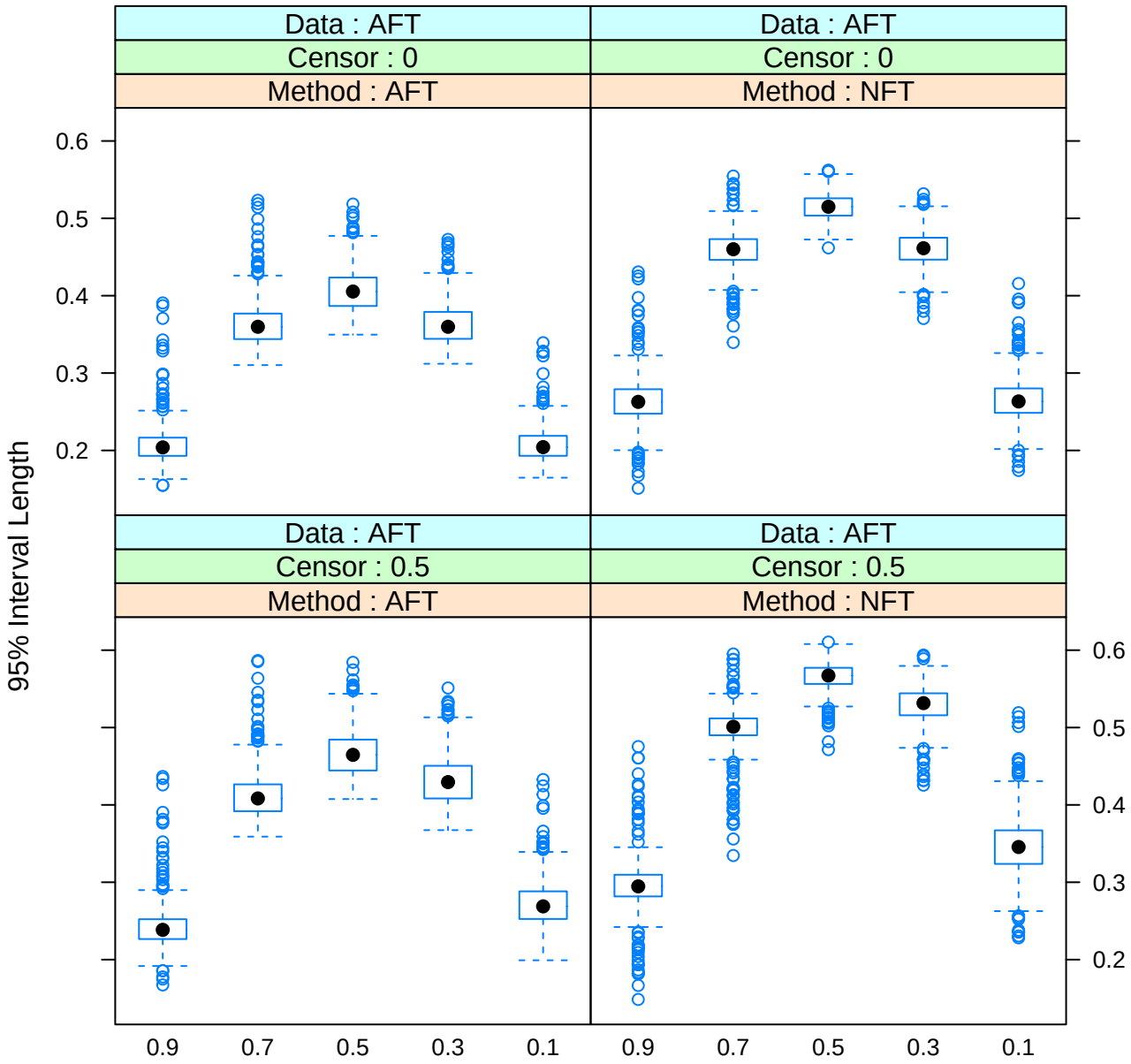

Figure 5: Results of a simulation study comparing AFT BART to NFT BART with sample size 2000. 95% interval length is on the vertical axis and survival settings are on the horizontal axis. This figure consists of data generated from the AFT scenario. The left (right) column are the results for AFT (NFT) BART. The top (bottom) row are for data generated with 0% (50%) censoring.

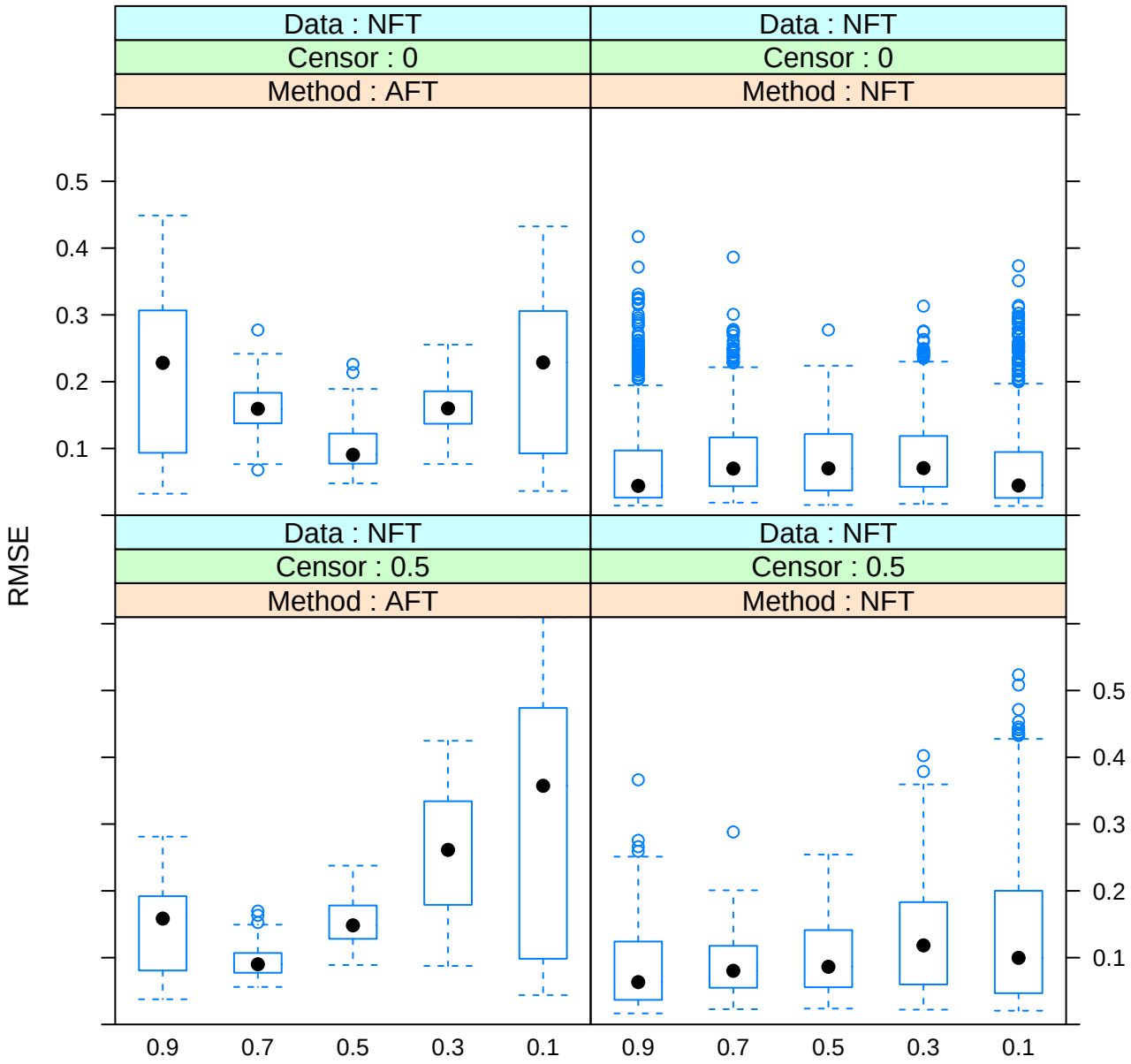

Figure 6: Results of a simulation study comparing AFT BART to NFT BART with sample size 2000. RMSE is on the vertical axis and survival settings are on the horizontal axis. This figure consists of data generated from the NFT scenario. The left (right) column are the results for AFT (NFT) BART. The top (bottom) row are for data generated with 0% (50%) censoring.

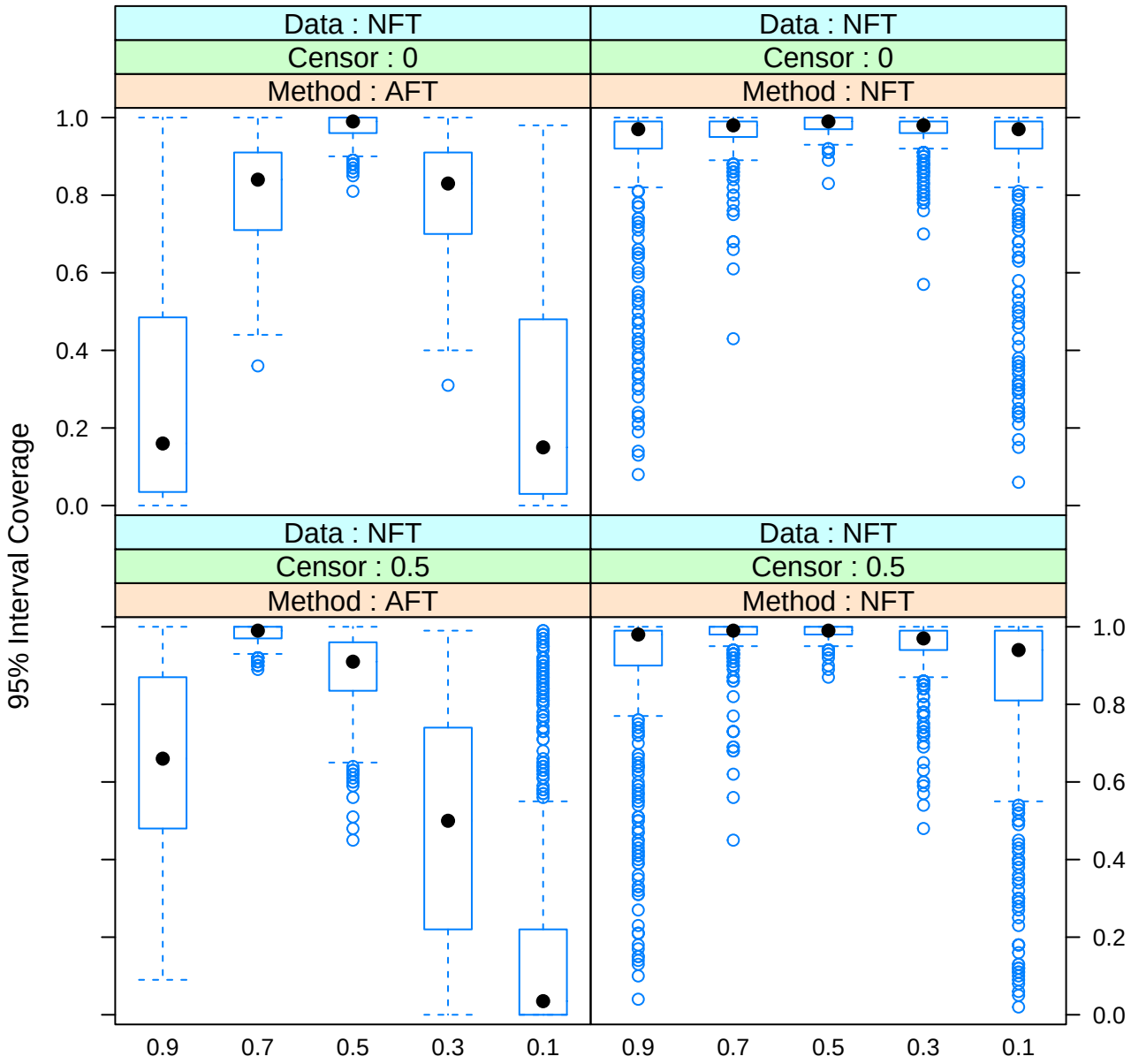

Figure 7: Results of a simulation study comparing AFT BART to NFT BART with sample size 2000. 95% interval coverage is on the vertical axis and survival settings are on the horizontal axis. This figure consists of data generated from the NFT scenario. The left (right) column are the results for AFT (NFT) BART. The top (bottom) row are for data generated with 0% (50%) censoring.

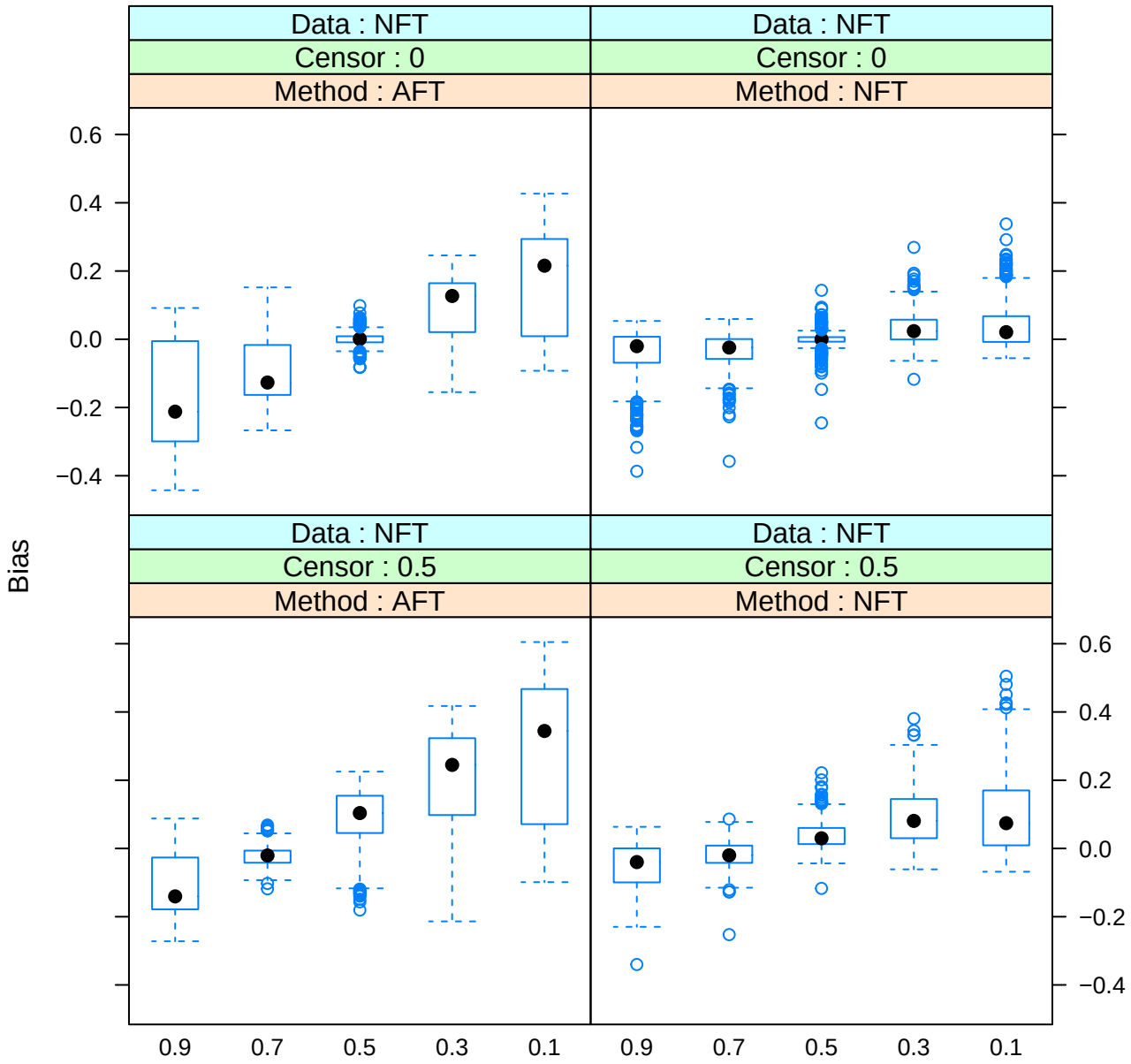

Figure 8: Results of a simulation study comparing AFT BART to NFT BART with sample size 2000. Bias is on the vertical axis and survival settings are on the horizontal axis. This figure consists of data generated from the NFT scenario. The left (right) column are the results for AFT (NFT) BART. The top (bottom) row are for data generated with 0% (50%) censoring.

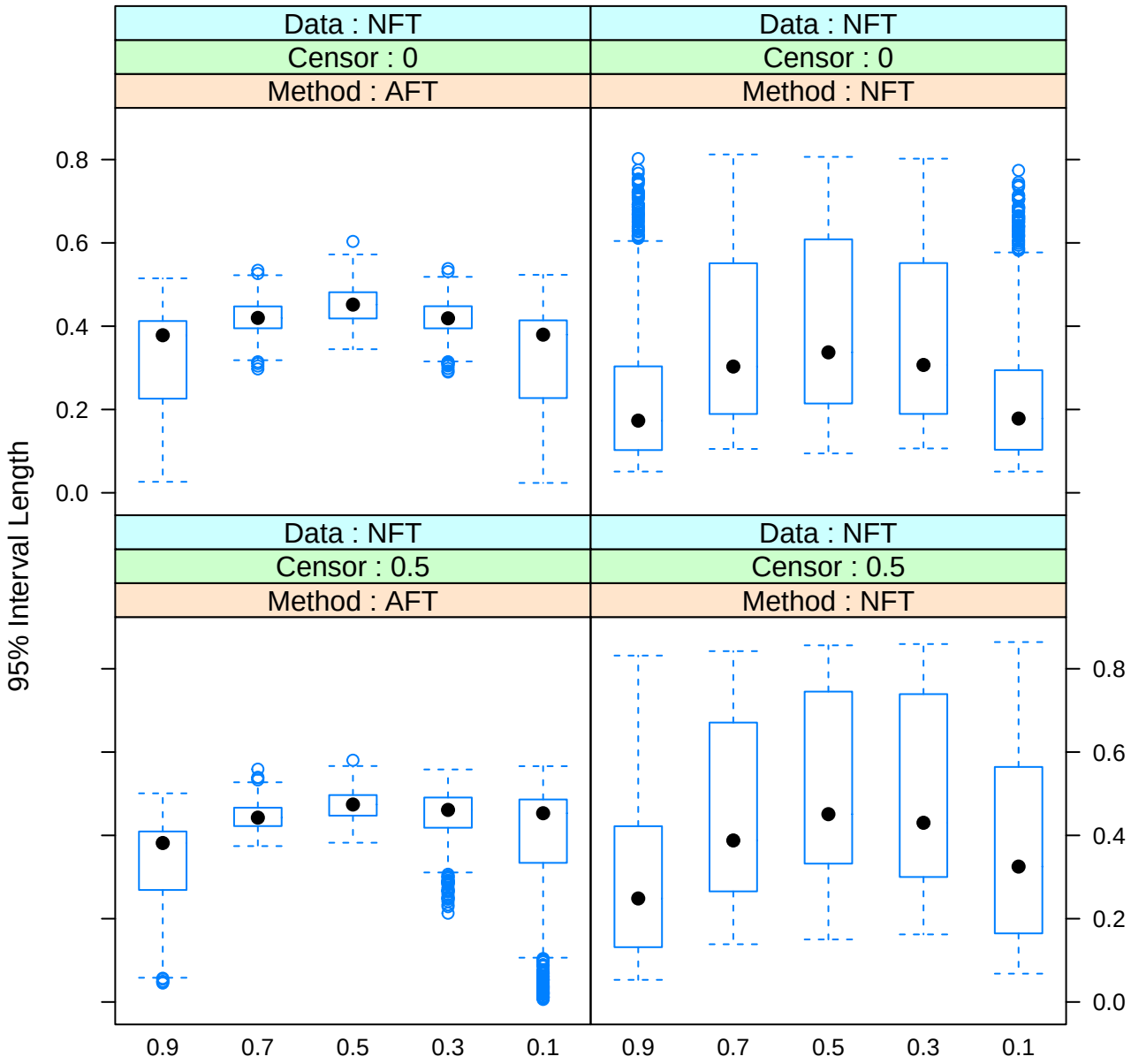

Figure 9: Results of a simulation study comparing AFT BART to NFT BART with sample size 2000. 95% interval length is on the vertical axis and survival settings are on the horizontal axis. This figure consists of data generated from the NFT scenario. The left (right) column are the results for AFT (NFT) BART. The top (bottom) row are for data generated with 0% (50%) censoring.

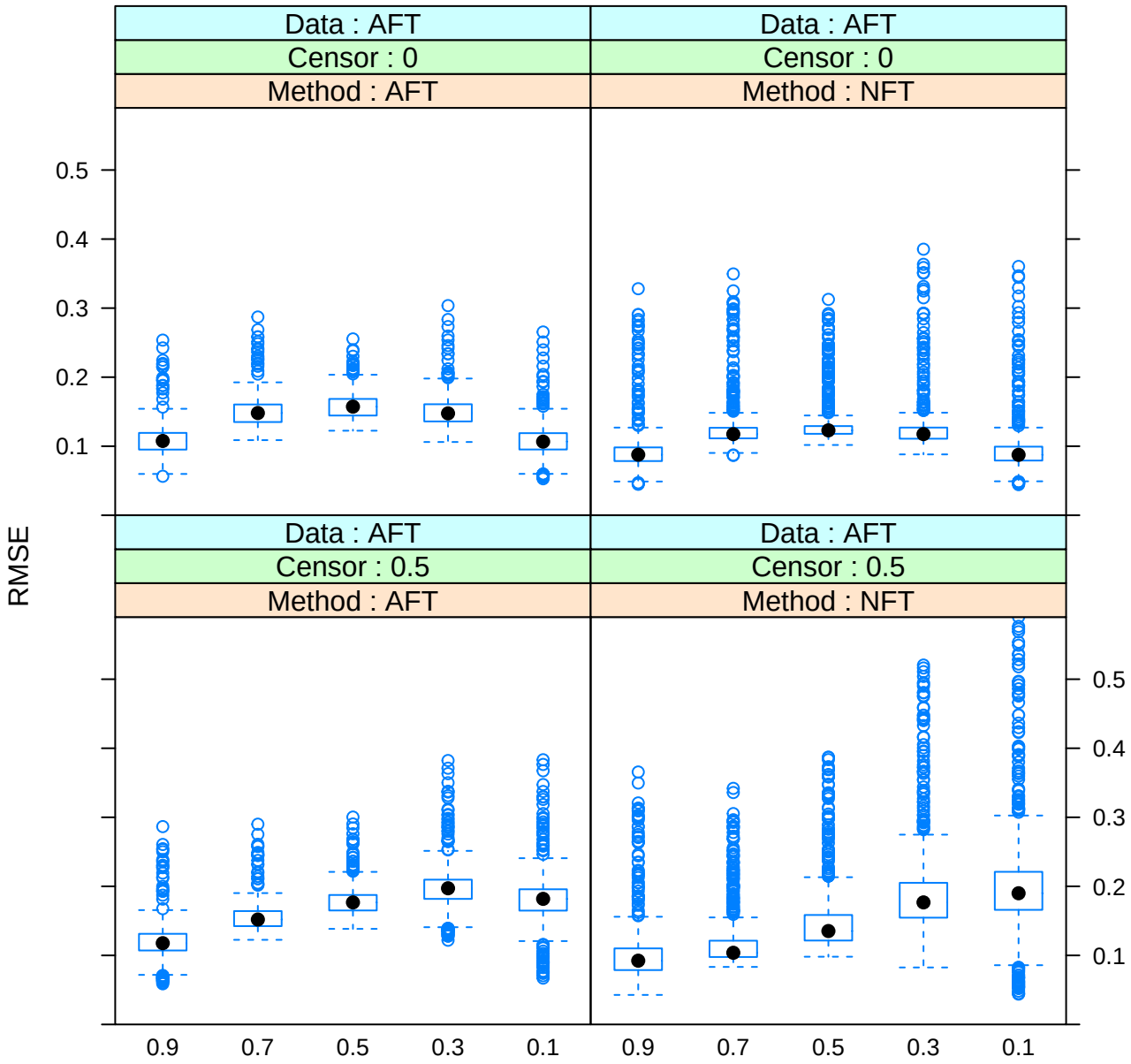

Figure 10: Results of a simulation study comparing AFT BART to NFT BART with sample size 500. RMSE is on the vertical axis and survival settings are on the horizontal axis. This figure consists of data generated from the AFT scenario. The left (right) column are the results for AFT (NFT) BART. The top (bottom) row are for data generated with 0% (50%) censoring.

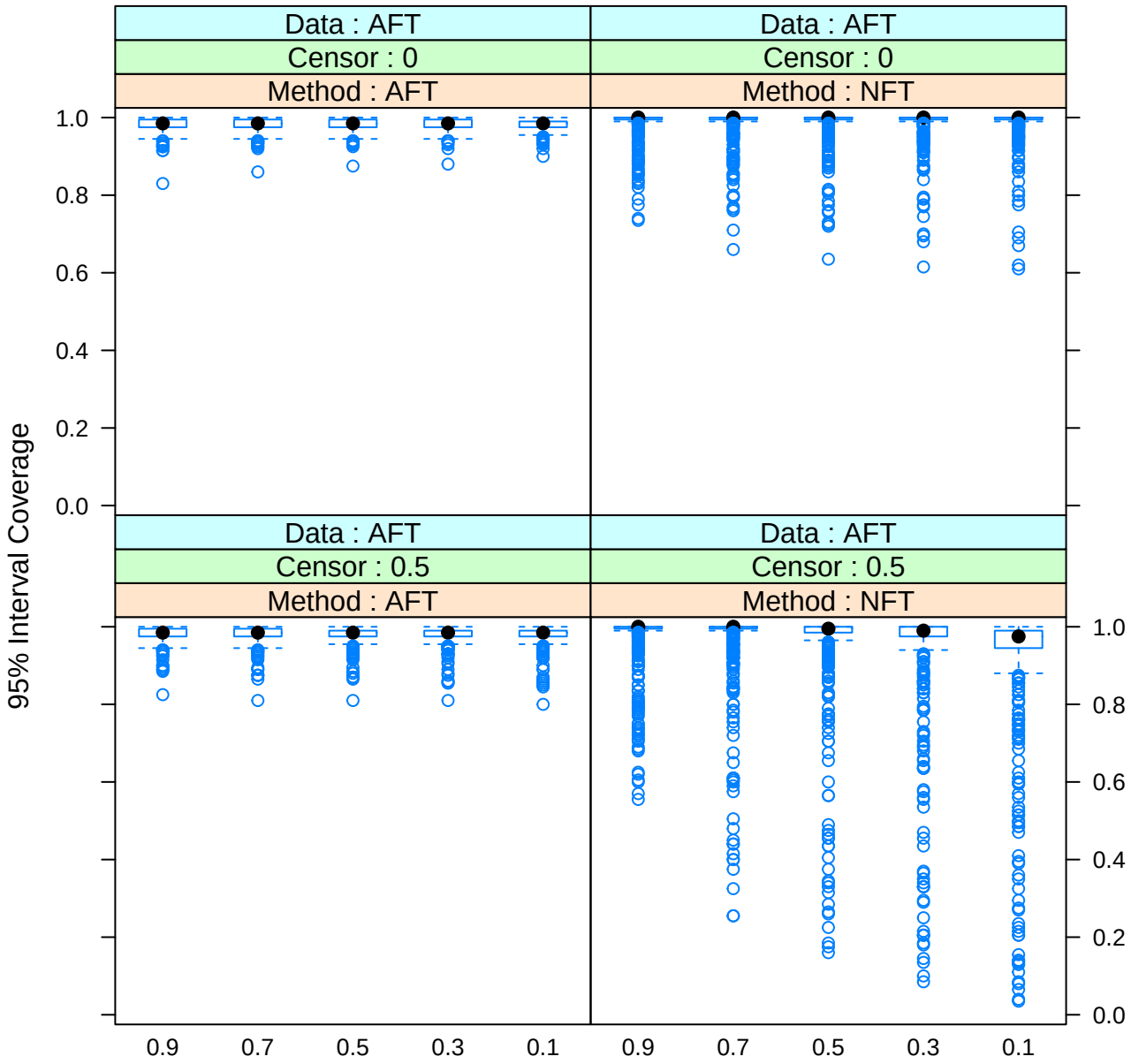

Figure 11: Results of a simulation study comparing AFT BART to NFT BART with sample size 500. 95% interval coverage is on the vertical axis and survival settings are on the horizontal axis. This figure consists of data generated from the AFT scenario. The left (right) column are the results for AFT (NFT) BART. The top (bottom) row are for data generated with 0% (50%) censoring.

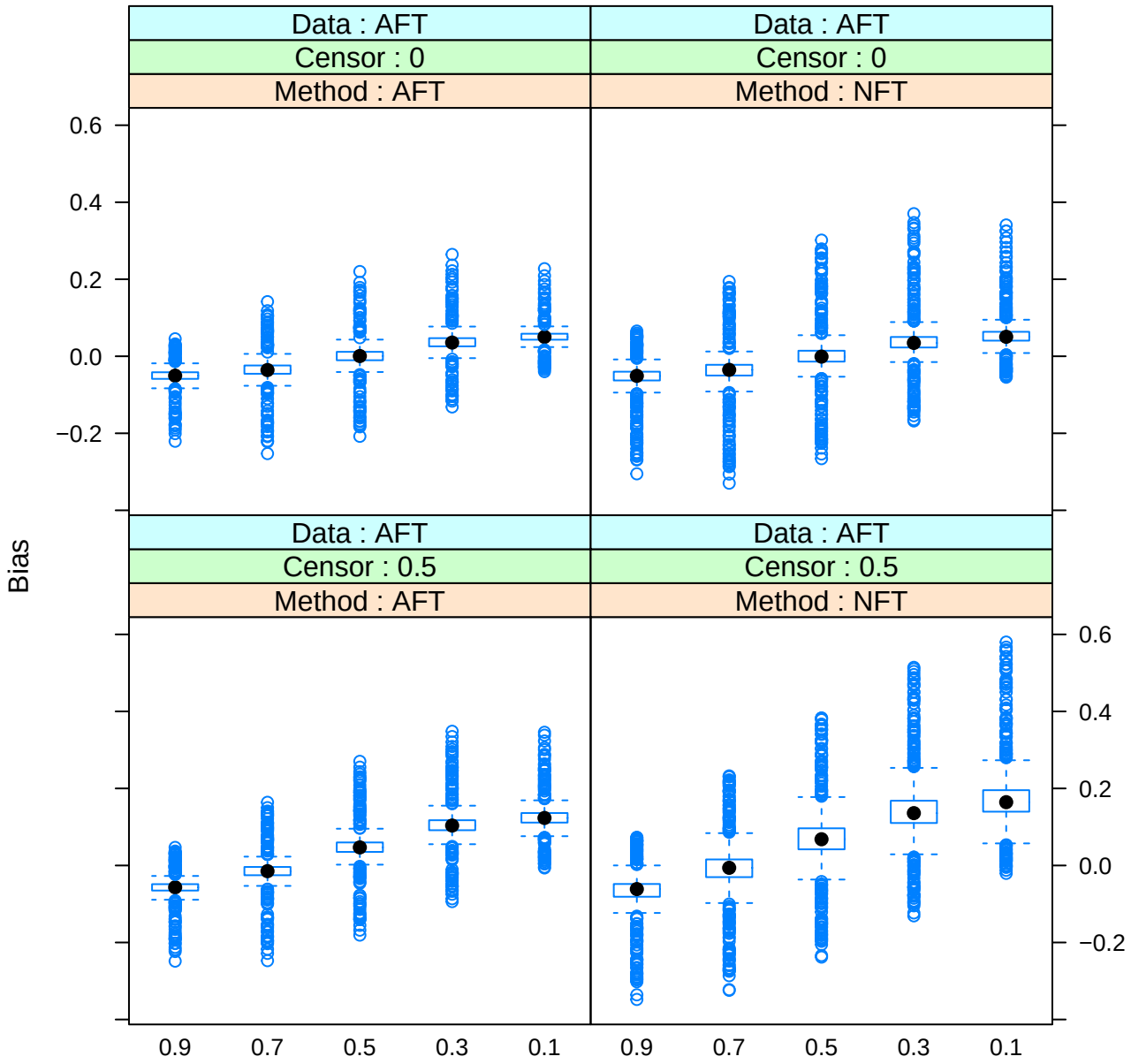

Figure 12: Results of a simulation study comparing AFT BART to NFT BART with sample size 500. Bias is on the vertical axis and survival settings are on the horizontal axis. This figure consists of data generated from the AFT scenario. The left (right) column are the results for AFT (NFT) BART. The top (bottom) row are for data generated with 0% (50%) censoring.

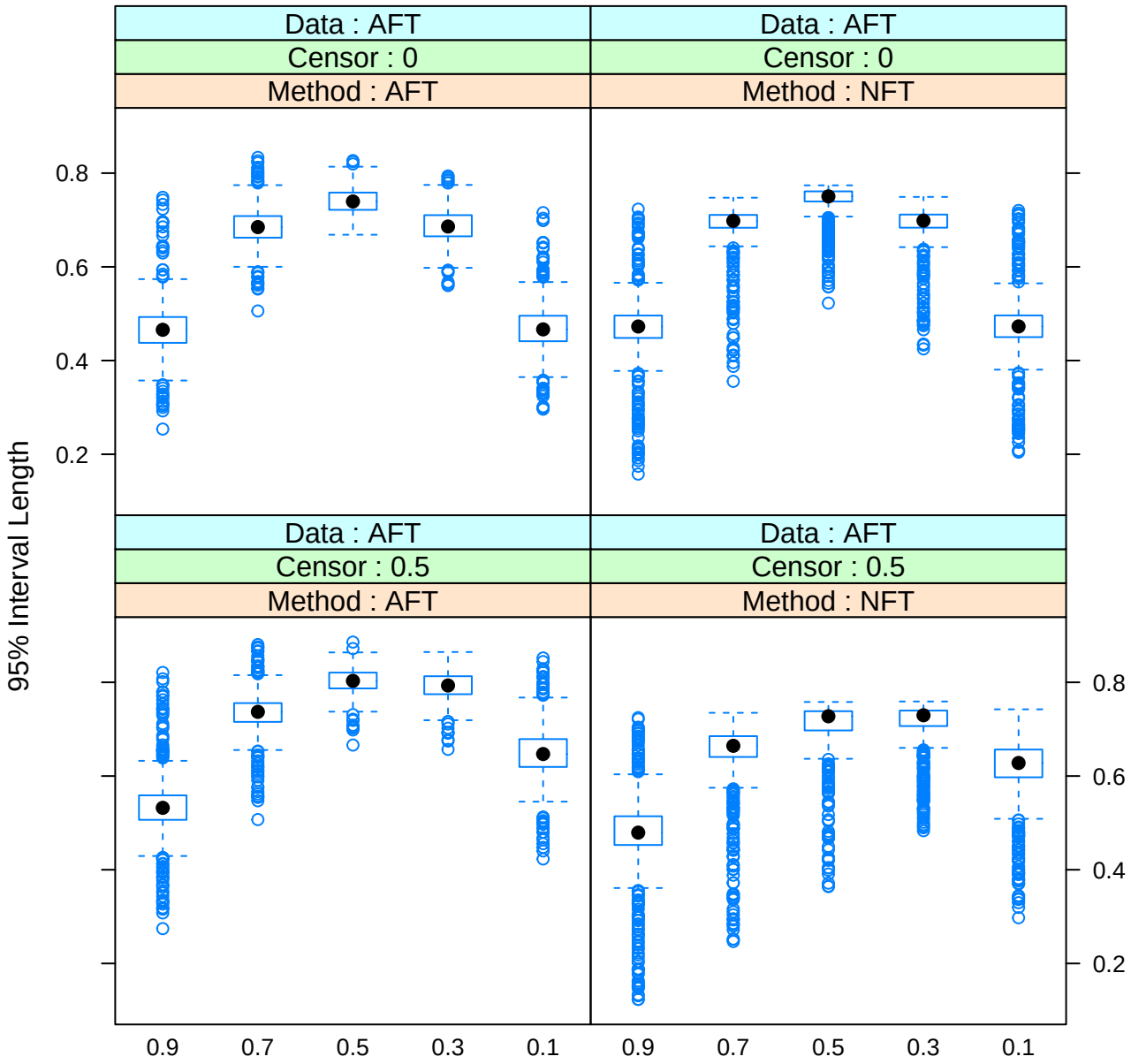

Figure 13: Results of a simulation study comparing AFT BART to NFT BART with sample size 500. 95% interval length is on the vertical axis and survival settings are on the horizontal axis. This figure consists of data generated from the AFT scenario. The left (right) column are the results for AFT (NFT) BART. The top (bottom) row are for data generated with 0% (50%) censoring.

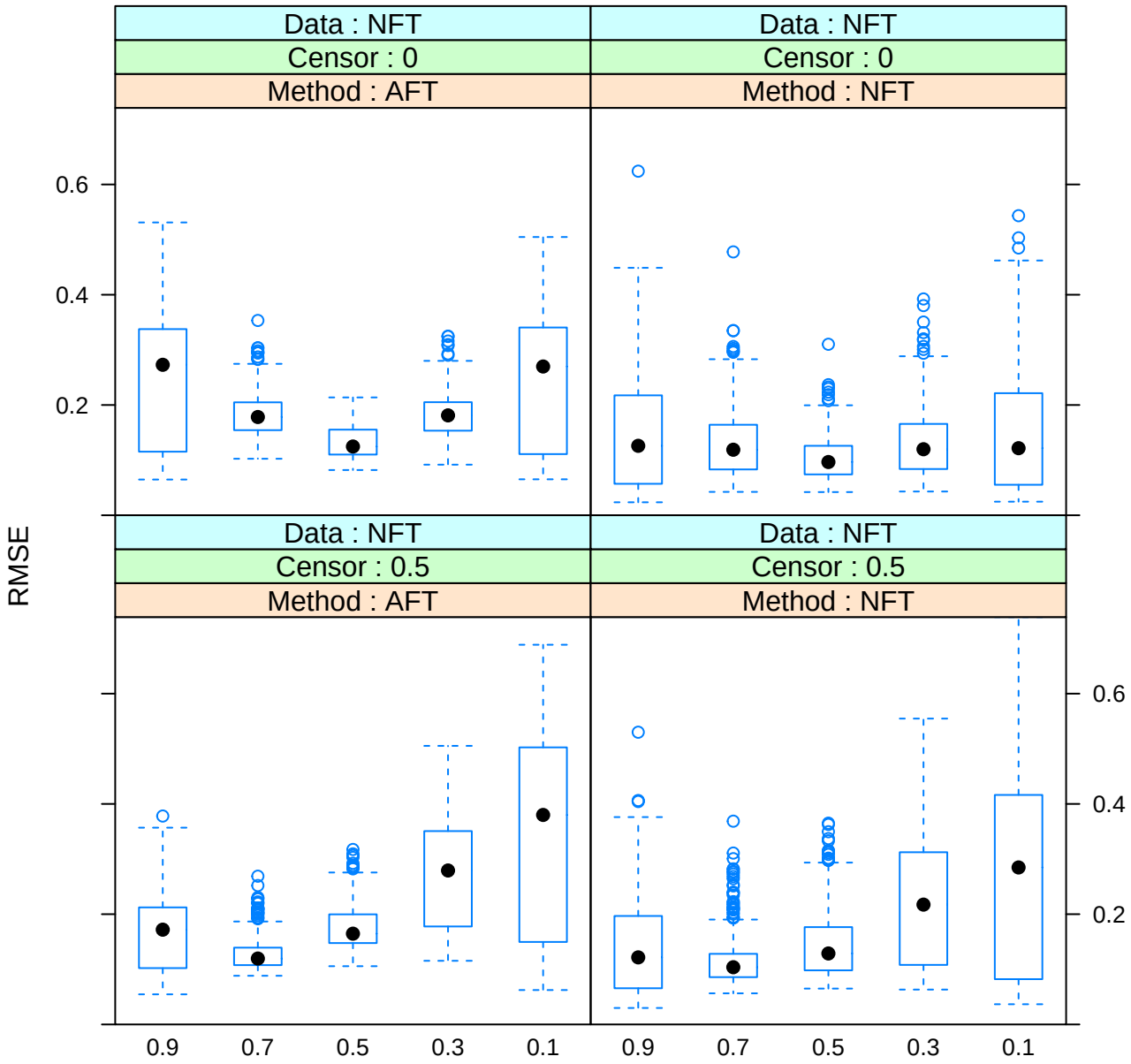

Figure 14: Results of a simulation study comparing AFT BART to NFT BART with sample size 500. RMSE is on the vertical axis and survival settings are on the horizontal axis. This figure consists of data generated from the NFT scenario. The left (right) column are the results for AFT (NFT) BART. The top (bottom) row are for data generated with 0% (50%) censoring.

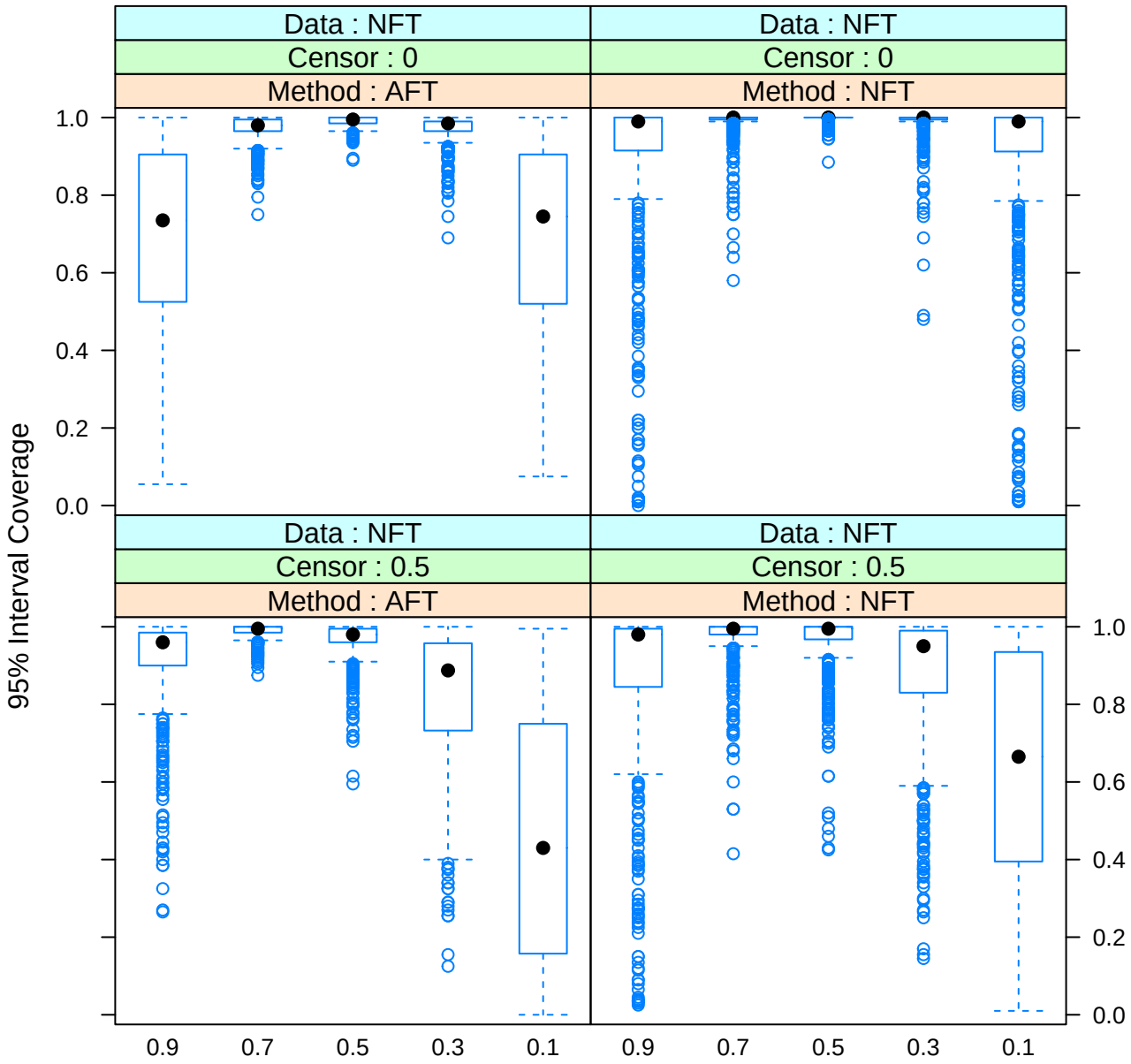

Figure 15: Results of a simulation study comparing AFT BART to NFT BART with sample size 500. 95% interval coverage is on the vertical axis and survival settings are on the horizontal axis. This figure consists of data generated from the NFT scenario. The left (right) column are the results for AFT (NFT) BART. The top (bottom) row are for data generated with 0% (50%) censoring.

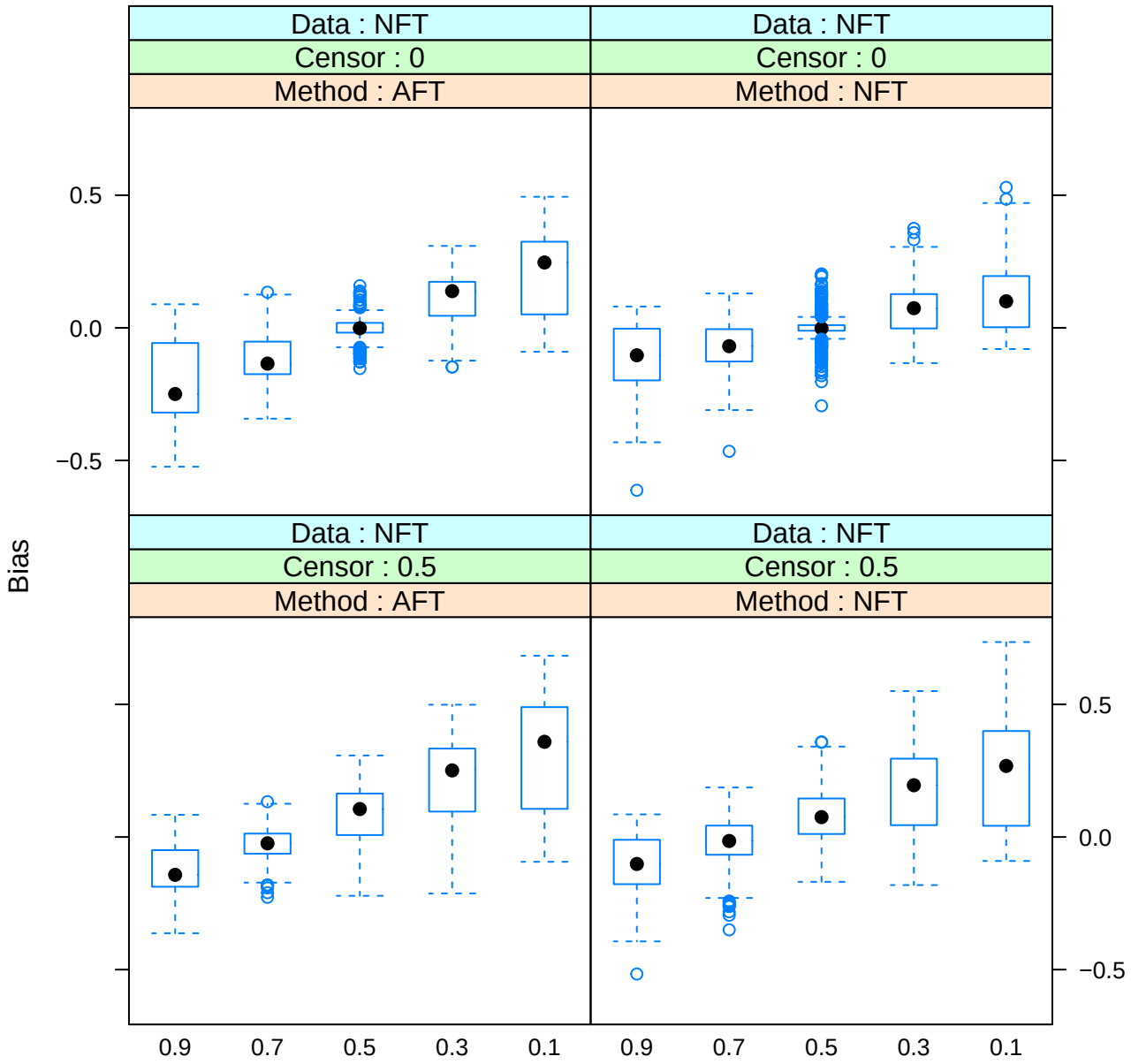

Figure 16: Results of a simulation study comparing AFT BART to NFT BART with sample size 500. Bias is on the vertical axis and survival settings are on the horizontal axis. This figure consists of data generated from the NFT scenario. The left (right) column are the results for AFT (NFT) BART. The top (bottom) row are for data generated with 0% (50%) censoring.

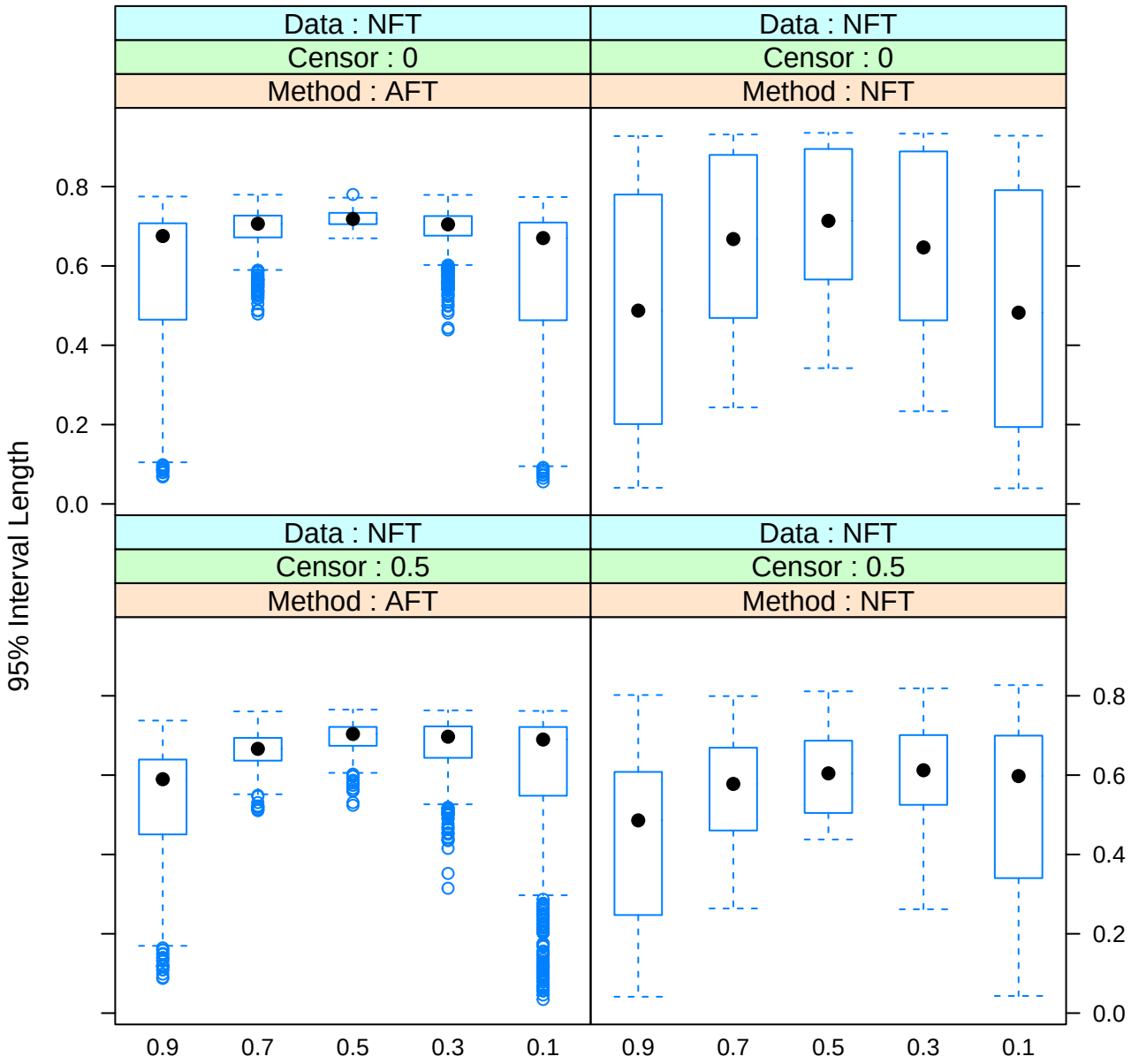

Figure 17: Results of a simulation study comparing AFT BART to NFT BART with sample size 500. 95% interval length is on the vertical axis and survival settings are on the horizontal axis. This figure consists of data generated from the NFT scenario. The left (right) column are the results for AFT (NFT) BART. The top (bottom) row are for data generated with 0% (50%) censoring.