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

Rodney Sparapani, Brent Logan, M.J. Maiers, Prakash Laud,
Matt Pratola and Rob McCulloch

Medical College of Wisconsin
The Ohio State University
National Marrow Donor Program
Arizona State University

# Personalized Hematopoietic Stem Cell Transplant (HSCT)

- HSCT is a common treatment for blood/bone marrow cancers
- Here we are concerned with unrelated donors that are human leukocyte antigen (HLA) 8/8 matched to the recipients transplanted from 2016:2018
- Goal: optimal donor matching for better recipient outcomes
- The outcome here is time to an event, i.e., event-free survival with both right and left censoring
- Events include death, relapse, graft failure/rejection or moderate/severe chronic graft vs. host disease (GVHD): whichever comes first

- There are $P = 45$ covariates that may have an impact
- 6 are donor-related characteristics: age, gender/childbearing, DPB1 match/unknown, DQB1 match and CMV match
- We wanted to *learn* the (likely complex) functional relationship between these covariates and the outcome with BART
- The cohort has 8567 patients: $N = 7157$ for training
- A bit too large for our current Discrete Time BART
- For this application, we developed NFT BART methodology

Given the data, we want to estimate things like

$$S(t,x) = 1 - P(T < t \,|\, x)$$

the *survival function*.

Given the information about the donor and recipient in $x$, what is the probability the time to event is greater than $t$.

All the events are bad, so $S(t,x)$ is the probability of *surviving* time $t$ or longer.

We could use this to match donors with recipients.

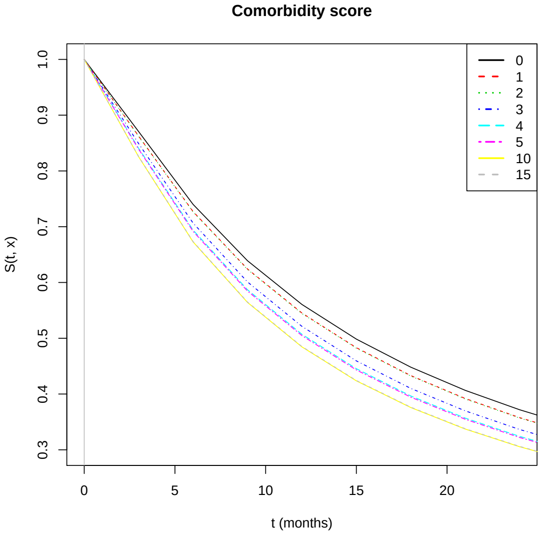
# Event-free Survival: Recipient Comorbidity Score
# Friedman's partial dependence function

How does survival depend on Comorbidity Score?

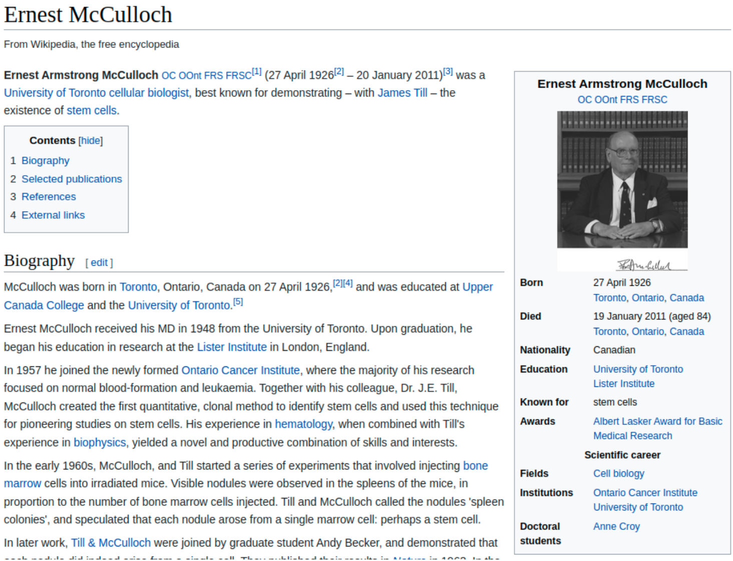
We estimate the
$S(t,x)$
as Comorbidity varies.

Friedman:
average out other $x$'s.

60% chance of living 10 months or longer.



# Kind of a funny application for Rob !!!



# Ernest McCulloch

From Wikipedia, the free encyclopedia

**Ernest Armstrong McCulloch** OC OOnt FRS FRSC[1] (27 April 1926[2] – 20 January 2011)[3] was a University of Toronto cellular biologist, best known for demonstrating – with James Till – the existence of stem cells.

**Contents** [hide]

## Biography [ edit ]

McCulloch was born in Toronto, Ontario, Canada on 27 April 1926,[2][4] and was educated at Upper Canada College and the University of Toronto.[5]

Ernest McCulloch received his MD in 1948 from the University of Toronto. Upon graduation, he began his education in research at the Lister Institute in London, England.

In 1957 he joined the newly formed Ontario Cancer Institute, where the majority of his research focused on normal blood-formation and leukaemia. Together with his colleague, Dr. J.E. Till, McCulloch created the first quantitative, clonal method to identify stem cells and used this technique for pioneering studies on stem cells. His experience in hematology, when combined with Till's experience in biophysics, yielded a novel and productive combination of skills and interests.

In the early 1960s, McCulloch, and Till started a series of experiments that involved injecting bone marrow cells into irradiated mice. Visible nodules were observed in the spleens of the mice, in proportion to the number of bone marrow cells injected. Till and McCulloch called the nodules 'spleen colonies', and speculated that each nodule arose from a single marrow cell: perhaps a stem cell.

In later work, Till & McCulloch were joined by graduate student Andy Becker, and demonstrated that

**Ernest Armstrong McCulloch**
OC OOnt FRS FRSC

| | |
|---|---|
| **Born** | 27 April 1926<br>Toronto, Ontario, Canada |
| **Died** | 19 January 2011 (aged 84)<br>Toronto, Ontario, Canada |
| **Nationality** | Canadian |
| **Education** | University of Toronto<br>Lister Institute |
| **Known for** | stem cells |
| **Awards** | Albert Lasker Award for Basic Medical Research |
| | **Scientific career** |
| **Fields** | Cell biology |
| **Institutions** | Ontario Cancer Institute<br>University of Toronto |
| **Doctoral students** | Anne Croy |

Goal:

We want to estimate these survivial functions flexibly.

A lot of survival analysis we starts by working in terms of the harzard function.

We will start from the *Acclerated Failure Time* model (**AFT**), which basically means you model the log of time.

# Basic AFT model

We have data $(t_i, \delta_i)$ where $i = 1, \ldots, N$ subjects
if $\delta_i = 0$, then $t_i$ is a right censoring time
else if $\delta_i = 1$, then a failure time
else if $\delta_i = 2$, then left censoring

Take logarithms $y_i = \log t_i$ and use a linear model.

$$y_i = \mu + x_i'\beta + \epsilon_i$$

with $\epsilon_i$ *iid* $F_\epsilon(\mu_\epsilon = 0, \sigma_\epsilon^2 = \sigma^2)$
where $\beta$ and $\sigma$ are unknown parameters to be estimated.
$F_\epsilon$ is typically parametric (normal, extreme value).

If there is censoring, $y$ is latent.
if $\delta_i = 0$, then $y_i \sim F_\epsilon(\mu + x_i'\beta, \sigma^2) I\{\log t_i, \infty\}$
else if $\delta_i = 1$, then $y_i = \log t_i$
else if $\delta_i = 2$, then $y_i \sim F_\epsilon(\mu + x_i'\beta, \sigma^2) I\{-\infty, \log t_i\}$

Or, putting aside the complexity due to the censoring,

$$y_i = \mu + x_i' \beta + \sigma Z_i$$

where $Z_i$ are iid from some parametric family.

Very simple except for the censoring.

*We don't like linear*.

*We don't like parametric errors*.

Note:

The modeling problem looks like a standard “regression” type thing.

*But*, the emphasis on the survival curve means we really need to get the probabilities right, which means we really want to get the error distribution right.

We need need the whole distribution, not just $E(Y \mid X)$.

# AFT BART

Henderson et al. (2020).

$$y_i = \mu + \mu_i + f(x_i) + \sigma Z_i$$

- $f$ is BART.
- $\{\mu_i\}$, DPM, constrained to sum to 1.
- $Z_i$ are iid N(0,1).

Note:

“Fully nonparametric Bayesian additive regression trees”, (2019), Edward George, Purushottam Laud, Brent Logan, Robert McCulloch, Rodney Sparapani

$$y_i = \mu_i + f(x_i) + \sigma_i Z_i$$

- $f$ is BART.
- $\{(\mu_i, \sigma_i)\}$, DPM.

Can you get a DPM prior that works as “easily” as the standard BART prior?

*worked pretty good, need to get the R package out !!!*.

# NFT BART

“Nonparametric Failure Time BART”.

$$y_i = \mu + \mu_i + f(x_i) + \sigma_i \, s(x_i) \, Z_i$$

- $Z_i \sim N(0,1)$, iid.
- $f$ is BART, $f$ is a sum of trees.
- $s$ is HBART, $s^2(x)$ is a product of trees.
- $(\mu_i, \sigma_i)$, DPM.
- $\sum \mu_i = 0$.
- $\frac{1}{N} \sum \sigma_i^2 = 1$.

*remember how DPM works !!*

At each MCMC draw we will have $(\mu_i, \sigma_i)$, $i = 1, 2, \ldots, N$.

But, there will be a bunch of repeats.

$g(i) \in \{1, 2, \ldots, G\}$.

$G << N$.

$$(\mu_i, \sigma_i) = (\mu^*_{g(i)}, \sigma^*_{g(i)}),$$

where

$$\{(\mu^*_g, \sigma^*_g)\}, g = 1, 2, \ldots, G$$

are the unique values.

## Survival Function

$$S(t,x) \approx 1 - \frac{1}{N} \sum_{i=1}^{N} \Phi(\frac{\log(t) - \mu - \mu_i - f(x)}{\sigma_i \, s(x)}).$$

Let $w_g$ = percent of $(\mu_i, \sigma_i) = (\mu_g^*, \sigma_g^*)$.

$$S(t,x) \approx 1 - \sum_{g=1}^{G} w_g \, \Phi(\frac{\log(t) - \mu - \mu_g^* - f(x)}{\sigma_g^* \, s(x)}).$$

Intuitively related to a mixture model.

Clearly, NFT BART is much more flexible than AFT BART.
But *is this crazy ???*

2019
Low Information Omnibus (LIO) Priors for Dirichlet Process Mixture Models Yushu Shi, Michael Martens, Anjishnu Banerjee, Purushottam Laud

2020.
Heteroscedastic BART Using Multiplicative Regression Trees Matthew Pratola, Hugh Chipman, Edward George, Robert McCulloch.

But not at all obvious you can or should get it all to work together.
Rodney did the heavy lifting.

## LIO

LIO, like BART/HBART, was designed to have robust prior default settings that should work well for most data situations without needing manual intervention except for perhaps altering the relative number of desired clusters vi the $\alpha$ prior.

Let $\tau_i = \frac{1}{\sigma_i^2}$.

DPM:

$(\mu_i, \tau_i) \sim G, \text{iid}$

$G \sim DP(\alpha, F)$

- $G$ is a discrete distribution, whose atoms are draws from $F$.

- $\alpha$ controls how many atoms effectively get weight.

So $F$ is a prior on $(\mu, \tau)$.

$F$ has the standard conjugate form:

$$\tau | a_0, b_0 \sim \text{Gamma}(a_0, b_0), \;\; \mu | \tau, k_0 \sim N(0, \tau^{-1}\, k_0^{-1}).$$

Priors on Priors:

$$k_0 \sim \text{Gamma}(a_{k_0}, b_{k_0}), \;\; b_0 \sim \text{Gamma}(a_{b_0}, b_{b_0}).$$

Have to choose: $a_0, a_{k_0}, b_{k_0}, a_{b_0}, b_{b_0}$.

Looks a bit daunting to choose all 5 parameters, but because we are "standardizing" the $(\mu_i, \sigma_i)$ with the identifiability constraints

$\sum \mu_i = 0.$
$\frac{1}{N} \sum \sigma_i^2 = 1.$

We can reasonably impose ball park constraints such as

$$E(\tau_i^{-1}) = \frac{a_{b_0}}{b_{b_0}(a_0 - 1)} = 1$$

*Same philosophy as BART !!!*

Ball park a prior which is spread out, but not too spread out and in the ball park of something reasonable.

## Simulated Examples

### NFT

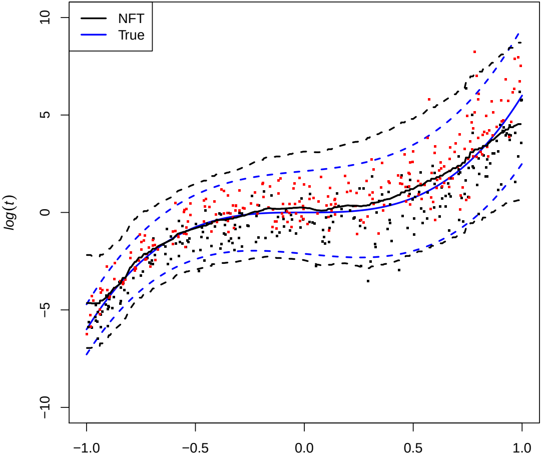
$$y = \log t = f(x) + s(x)\,\epsilon,\ \epsilon \sim t_{16}$$
$$f(x) = 6x^3,\ s(x) = \exp 0.5x.$$

$p = 1$.

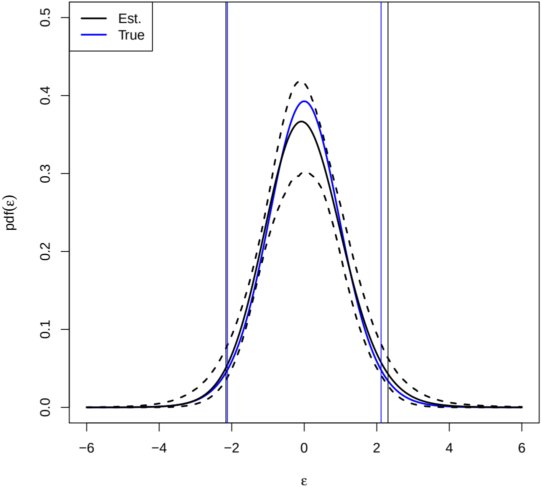
$N = 500$
50% censoring.
Red points right censored.

$x \sim U(-1, 1)$

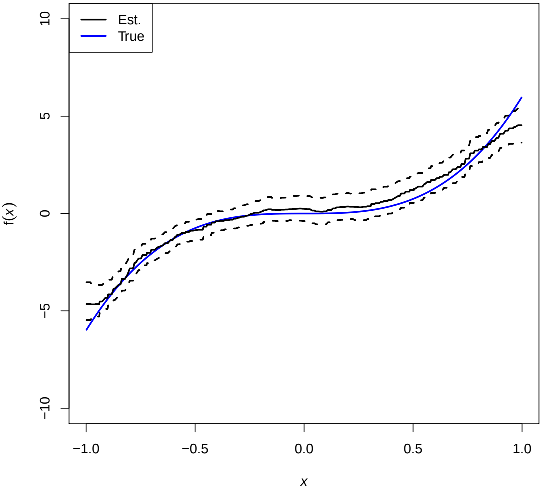
NFT estimates are posterior means.



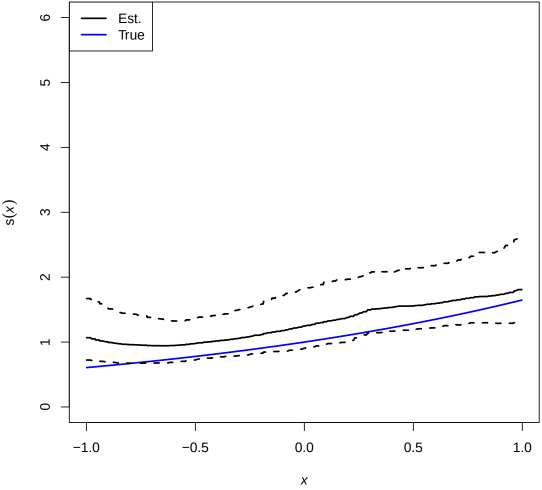
# Error density estimation with posterior intervals.



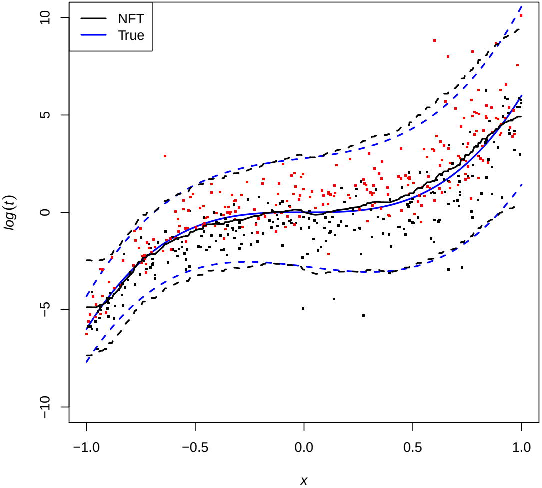
## $f$ estimation with posterior intervals.



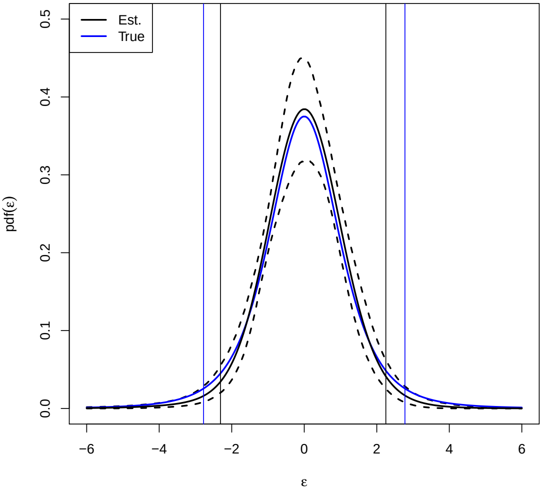
*s* estimation with posterior intervals.



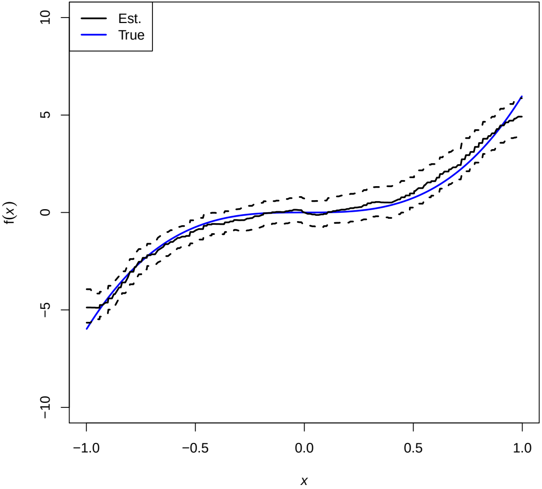
Everything the same as in the previous example except $\epsilon \sim t_4$. Posterior means.



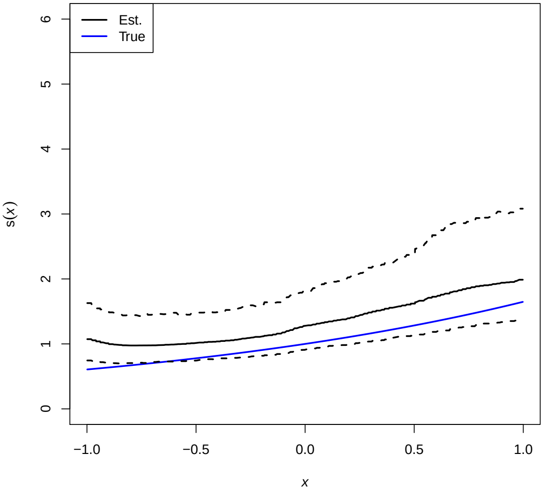
# Error density estimation with posterior intervals.



*f* estimation with posterior intervals.



*s* estimation with posterior intervals.



# Neither AFT nor NFT scenario: AFT failure!

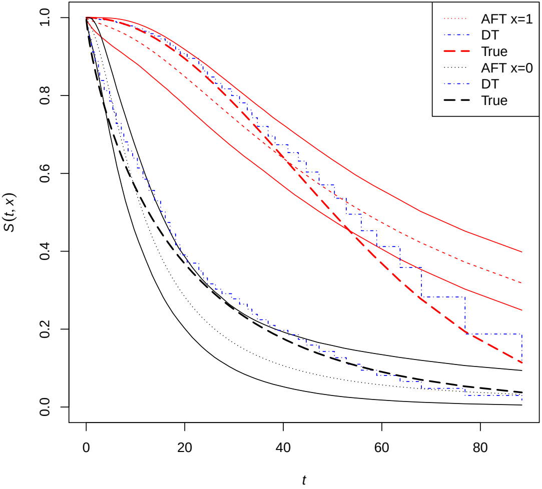
$N = 500$ with 50% censoring

$\text{Weibull}(0.8 + 1.2x, 20 + 40x),\ x \sim \textit{Bernoulli}(0.5)$

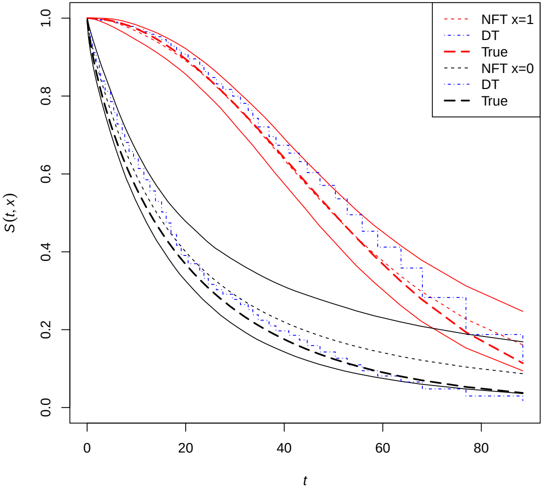
Two sets of curves.
One for x=0,
one for x=1.

DT:
Step function fit:
binarize $\delta t$.
Sparapani et. al. 2016.

Solid curves are
95% intervals.



# *NFT works !!!*



## More serious simulation comparison

Compare AFT BART with NFT BART.

- simulated data consistent with AFT BART and data consistent with NFT BART.
- N $=$ 500 and 2,000 training data size.
- for data sets of size 500 (2000) we simulated 200 (100) replicates.
- out of sample validation data set of size 500.
- 0% censoring or 50% (right) censoring.
- $p = 20$ covariates,
  $x_{2(j-1)+1}$ are Bernoulli p=.5,
  $x_{2j}$ are uniform (0,1), $j = 1, 2, \ldots, 10$..

## AFT BART:

$$y = \log(t) = 2 + 1.6\,x_1 + .8\,x_2 - 2.4\,x_2\,x_3 + \epsilon$$

$$\epsilon \sim N(0, \exp(-2))$$

## NFT BART:

$$y = \log(t) = 2 - 1.5\,x_4 + .5\,x_5 + 2\,x_5\,x_6 + \epsilon$$

$$\epsilon \sim N(0, [\exp(-2 + 1.6\,x_1 + .8\,x_2 - 2.4\,x_2\,x_3)]^2)$$

simple nonlinearity.
20 $x$, but not all come in.

## Comparison of Survival Curves

Presenting results is slightly complicated by the goal of estimating a survival function.

We will compare survival curves at the grid of *survival* values

$$S_j = .9 - .2(j-1), j = 1, 2, 3, 4, 5, \;\; = [.9, .7, .5, .3, .1].$$

Then for subject $i$ with covariates $x_i$ we obtain $t_{ij}$ such that

$$S(t_{ij}, x_i) = S_j.$$

where $S$ is the true survival function.

$x_i$ will be an $x$ for a simulated observation in the validation data set.

We then have, for example,

$\hat{S}_k(t_{ij}, x_i)$:

- $\hat{S}_k$ is the survival function obtained as the posterior mean from the $k^{th}$ training data set.
- $\hat{S}_k(t_{ij}, x_i)$, $i^{th}$ observation in the validation data, $j^{th}$ $t$ value corresponding to $S_j$.

Then for example, we have,

$$r_{ij} = \sqrt{\frac{1}{K} \sum_{k=1}^{K} (\hat{S}_k(t_{ij}, x_i) - S_j)^2}$$

$r_{ij}$: RMSE for the $i^{th}$ subject in the validation data, at the $j^{th}$ survival value, obained by averaging over the $K$ training data sets.

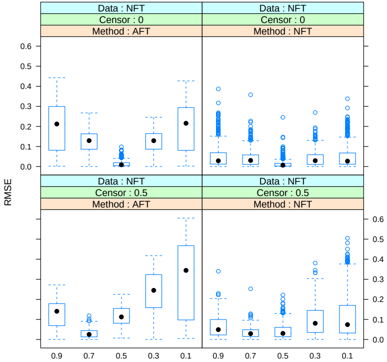
paper has 8 sets of boxplots.

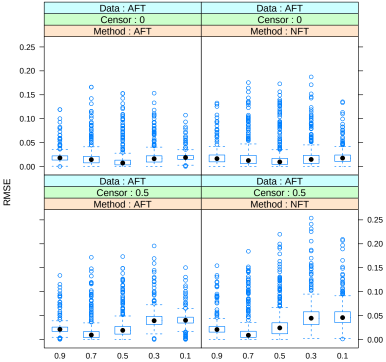
Each boxplot presents the values over observations in the validation data.

Each boxplot corresponds to a $S_j$ value.

- AFT or NFT data.
- 500 or 2000 train observations.
- rmse or 95% interval coverage.

Let's compare two sets of boxplot.

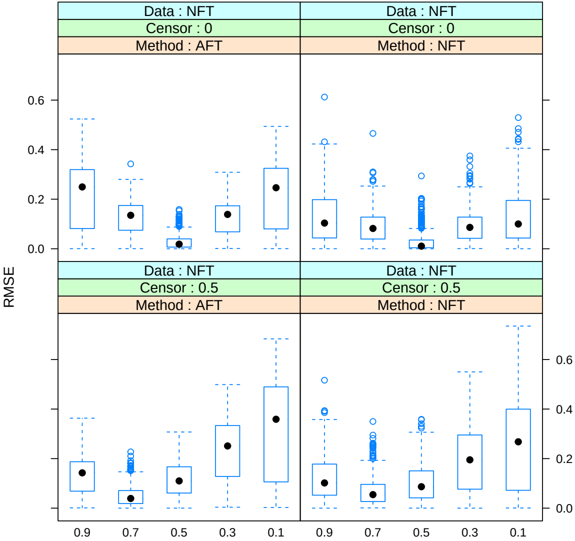




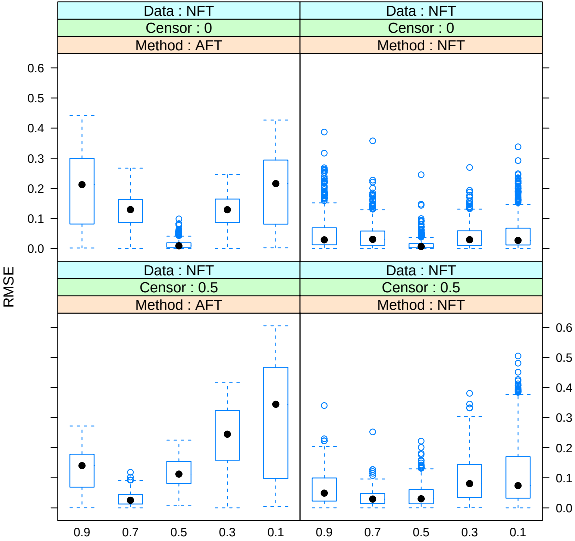
If we compare the bottom rows which correspond to 50% censoring,

- left figure: NFT is quite a bit better when data is NFT.
- right figure: NFT is not too much worse when data if AFT.

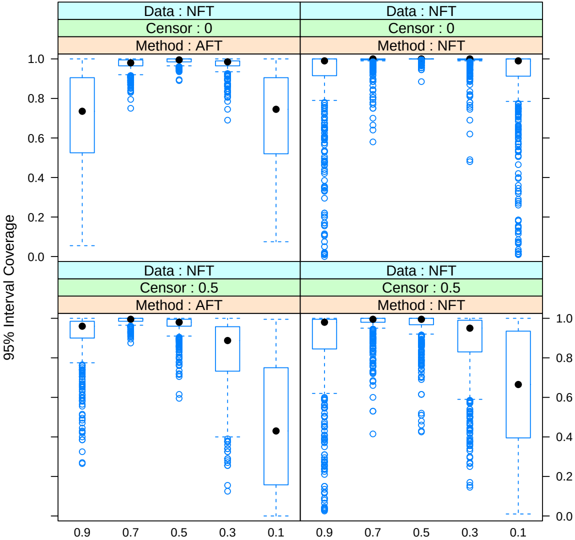
# Simulation study: NFT data generation $N = 500$



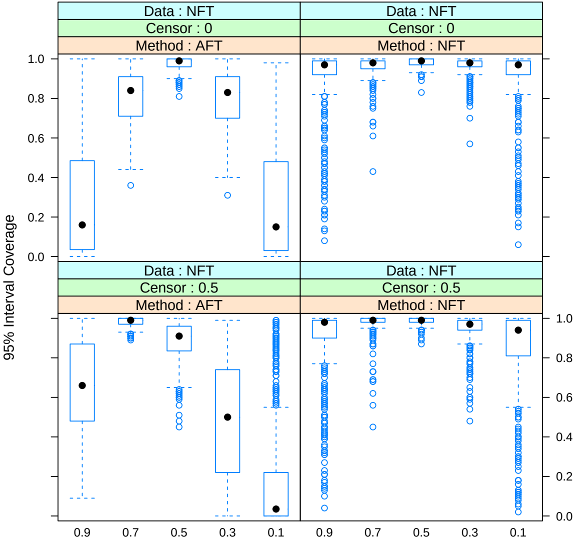
# Simulation study: NFT data generation $N = 2000$



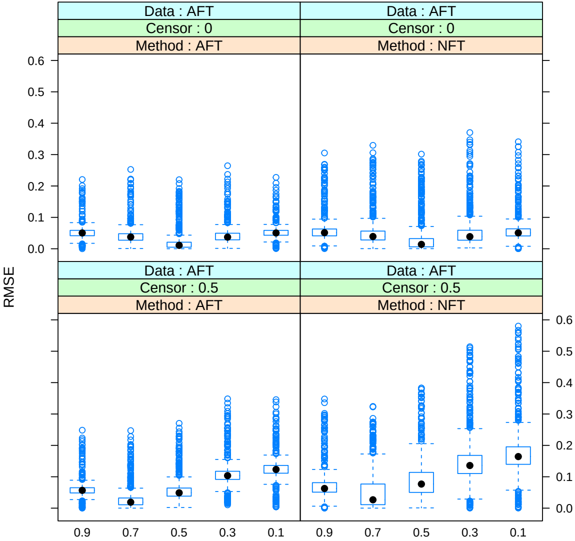
# Simulation study: NFT data generation $N = 500$



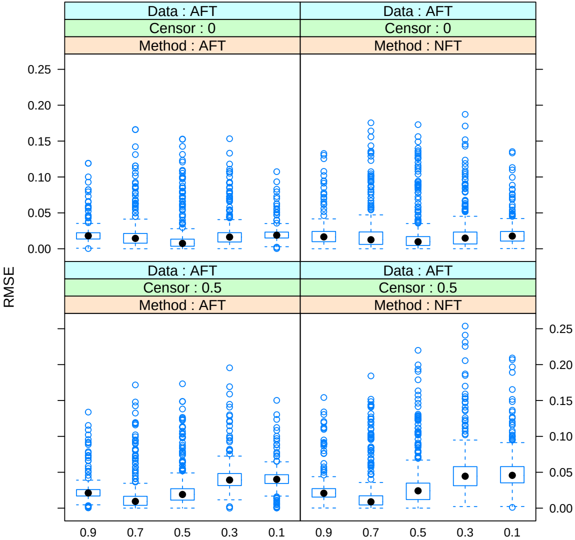
# Simulation study: NFT data generation $N = 2000$



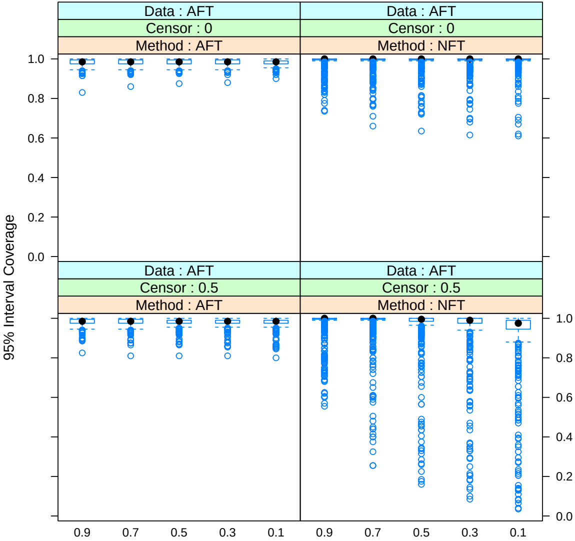
# Simulation study: AFT data generation $N = 500$



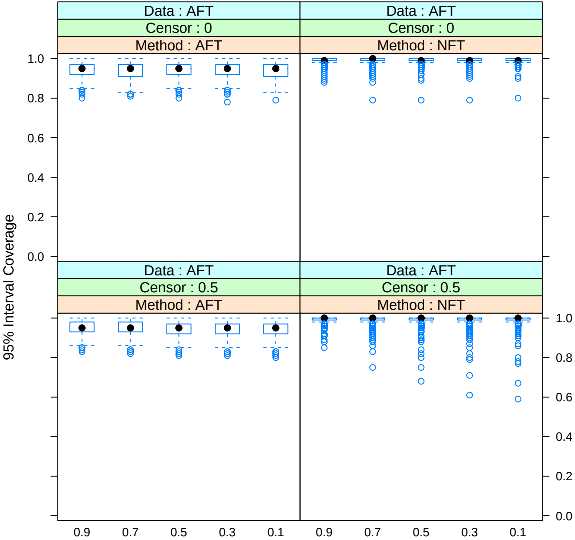
# Simulation study: AFT data generation $N = 2000$



# Simulation study: AFT data generation $N = 500$



# Simulation study: AFT data generation $N = 2000$



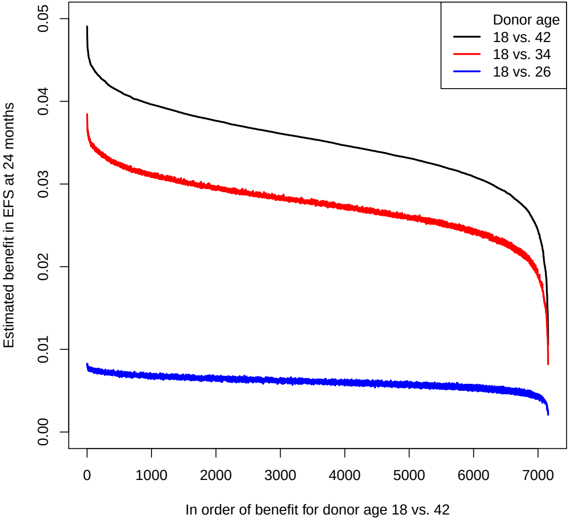
# Event-free Survival at 24 months: Waterfall plot

Estimated difference in surviving 24 months (or more).

Compare two ages
at a time.
for each train subject.

Sort by by 18 vs. 42 difference.

Estimate on train,
sort by 18 vs. 42
effect size.



# Conclusions: part 1

- We constructed our new Nonparametric Failure Time (NFT) approach from robust Bayesian Nonparametric building blocks
  - Bayesian Additive Regression Trees (BART) and Heteroskedastic BART
  - Dirichlet Process Mixtures (DPM),
    Constrained DPM and
    DPM Low Information Omnibus (LIO) prior hierarchy

# Conclusions: part 2

- NFT has desirable properties
  - computationally friendly via MCMC
  - very flexible model which does not resort to precarious restrictive assumptions
  - default prior parameter settings that work well without computationally expensive cross-validation
  - naturally extends to variable selection
- Personalized Hematopoietic Stem Cell Transplant (HSCT)
  - For Event-free Survival of HSCT recipients
    younger donors likely result in better outcomes

$>$ **install.packages('nftbart')**