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, Prakash Laud and
Rob McCulloch (Biometrics 2023)

RS, BL, PL and RM supported, in part, by
the US Office of Naval Research
Award Number: N00014-18-1-2888

Outline

- Motivation: a clinical application in Personalized Hematopoietic Stem Cell Transplant (HSCT) requires a new time-to-event BART methodology that scales better
- Pros and Cons of BART survival analysis methods
- BART and Heteroskedastic BART (HBART)
- Accelerated Failure Time (AFT) and AFT BART
- Nonparametric Failure Time (NFT) BART
- Dirichlet Process Mixtures (DPM), Constrained DPM and the Low Information Omnibus (LIO) DPM prior hierarchy
- Simulated data sets and methodology comparisons
- **nftbart** v1.6 R package on CRAN: **nftbart** v1.7 on my github
- Growth charts re-visited in nftbart v1.7: `demo/bmx.R`

Personalized Hematopoietic Stem Cell Transplant (HSCT)

- HSCT is a standard 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 2016:2019
- 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
- 5 are donor-related characteristics: age, sex/childbearing, HLA DPB1 match, HLA 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 10016 for training and 1802 for validation
- A bit too large for our Discrete Time BART
- For this application, we developed NFT BART methodology

Methodological and Computational Pros and Cons

	Published BART survival analysis methods				
Property	Hier.	Discrete Time (DT)	AFT	Mod.	NFT
Restrictive Assump.	Con	Pro	Con	Pro	Pro
Nonparam.	Con	Pro	Pro	Pro	Pro
Left-censor	Con	Con	Pro	Con	Pro
Comp. Complexity	Pro	Con	Pro	Con	Pro
First-author Year	Bonato 2011	Sparapani 2016	Henderson 2018	Linero 2021	Sparapani 2023

Bayesian Additive Regression Trees (BART)
NFT notation

Sparapani, Logan, Laud & McCulloch 2023 *Biometrics*

$$
\begin{aligned}
y_i &= \mu(x_i) + \epsilon_i \text{ where } \epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}\left(0,\ \sigma^2\right) \\
\mu &\overset{\text{prior}}{\sim} \text{BART}\ (a = 0.95, b = 2, H = 200, \kappa = 2, \tilde{\mu} = \bar{y}) \\
\mu(x_i) &\equiv \tilde{\mu} + \sum_h g(x_i; \mathcal{T}_h, \mathcal{M}_h)
\end{aligned}
$$

Heteroskedastic BART (HBART)
NFT notation

Pratola, Chipman, George & McCulloch 2019 *JCGS*

$$
\begin{aligned}
y_i &= \mu(x_i) + \epsilon_i \text{ where } \epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}\big(0,\ \sigma^2(x_i)\big) \\
\mu &\overset{\text{prior}}{\sim} \text{BART } (a, b, H = 200, \kappa = 5, \tilde{\mu}) \\
\sigma^2 &\overset{\text{prior}}{\sim} \text{HBART } (\tilde{a} = 0.95, \tilde{b} = 2, \widetilde{H} = 40, \tilde{\lambda}, \tilde{\nu}) \\
\sigma^2(x_i) &\equiv \prod_{h=1}^{\widetilde{H}} g(x_i; \widetilde{\mathcal{T}}_h, \widetilde{\mathcal{M}}_h) \text{ where } \widetilde{H} \approx H/5
\end{aligned}
$$

The Accelerated Failure Time (AFT) model: part 1

- Time-to-event data notation: (t_i, δ_i) $i = 1, \ldots, N$ 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
- How is failure time explained by a vector of covariates x_i?
- take logarithms $y_i = \log t_i$ and use a linear model (Con)
  $y_i = [1, x_i']\beta + \sigma\epsilon_i = \beta_0 + x_i'\beta_x + \sigma\epsilon_i$
  where β and σ are unknown coefficients to be estimated
  with $\epsilon_i \overset{\text{iid}}{\sim} F_\epsilon(\mu_\epsilon = 0, \sigma_\epsilon^2 = 1)$
  which is typically parametric (Con)

The Accelerated Failure Time (AFT) model: part 2

- Consider a *baseline* survival function for a *standard* subject where the covariates are centered, i.e., $S_0(t) = S(t|x=0)$.
- We can define the survival function for any given subject with a standard subject by accelerating, or decelerating, failure time

$$\begin{aligned} S(t|x_i) =& \mathrm{P}[s_i > t|x_i] = \mathrm{P}[y_i > \log t|x_i] \\ =& \mathrm{P}\left[\beta_0 + \sigma\epsilon_i > \log t - x_i'\beta_x|x_i\right] \\ =& S_0(t\exp\{-x_i'\beta_x\}) \end{aligned}$$

- however, AFT is a precarious restrictive assumption (Con)
  $S(t|x) = \mathrm{P}[\log s > \log t] = 1 - F_\epsilon\left(\log t; x'\beta, \sigma^2\right)$
  the covariates can only explain a log-linear location shift

Survival analysis with AFT BART
NFT notation

Henderson et al. 2018 *Biostatistics*

- $y_i = \mu(x_i) + \epsilon_i$ where $\epsilon_i | \mu_i \sim \mathrm{N}(\mu_i,\ \sigma^2)$: Pro
  $\mu \overset{\text{prior}}{\sim} \text{BART}$
- To ensure identifiability, constrain $\frac{1}{N} \sum_i \mu_i = 0$
- $\mu_i | G \sim G$
  $G|\alpha \overset{\text{prior}}{\sim} \mathrm{DP}\,(\alpha,\ F_0)$
- $S(t, x) = 1 - \frac{1}{N} \sum_i \Phi \left(\frac{\log t - \mu_i - \mu(x)}{\sigma} \right)$
  Con: the covariates still only explain a log-linear location shift

Survival analysis with NFT BART

- $y_i = \mu(x_i) + \epsilon_i$ where $\epsilon_i | (\mu_i, \sigma_i) \sim \mathrm{N}\left(\mu_i,\ \sigma_i^2 \sigma^2(x_i)\right)$: Pro
  $\mu \overset{\text{prior}}{\sim} \text{BART}$
  $\sigma^2 \overset{\text{prior}}{\sim} \text{HBART}$
- To ensure identifiability: $\frac{1}{N} \sum_i \mu_i = 0$ and $\frac{1}{N} \sum_i \sigma_i^2 = 1$
- if $\delta_i = 1$, then $y_i = \log t_i$
  else draw

$$y_i \sim \mathrm{N}\left(\mu_i + \mu(x_i),\ \sigma_i^2 \sigma^2(x_i)\right) \begin{cases} \mathrm{I}(\log t_i, \infty) & \text{if } \delta_i = 0 \\ \mathrm{I}(-\infty, \log t_i) & \text{if } \delta_i = 2 \end{cases}$$

- $(\mu_i, \sigma_i) | G \sim G$
  $G | \alpha \overset{\text{prior}}{\sim} \text{DP}\,(\alpha,\ F_0)$
- $S(t, x) = 1 - \frac{1}{N} \sum_i \Phi \left(\frac{\log t - \mu_i - \mu(x)}{\sigma_i \sigma(x)} \right)$
  Pro: the covariates can explain a location shift and rescaling!

Dirichlet Process Mixtures (DPM)

Ferguson 1973 & Antoniak 1974 *Annals of Statistics*;
Escobar & West 1995, Geng et al. 2018 *JASA*; Neal 2000 *JCGS*

$y_i \| \theta_i \sim F(\theta_i)$ where $i = 1, \ldots, N$	usual notation
$y_i \| \theta^*_{c_i} \sim F(\theta^*_{c_i})$ where $c_i \in \{1, \ldots, k\}$	ephemeral random clusters
$k \in \{1, \ldots, N\}$	k is random
$\theta_i \| G \sim G$ nonparametric (Pro)	parametric (Con): $\theta_i \overset{\text{prior}}{\sim} F_0$
$G \| \alpha \overset{\text{prior}}{\sim} \text{DP}(\alpha,\ F_0)$	G "centered" on F_0
$\alpha \overset{\text{prior}}{\sim} \text{Gamma}(a,\ b)$ $\propto k$	concentration parameter
$\theta_1 \sim F_0$	integrating over G
$\theta_2 \| \theta_1 \sim \frac{1}{1+\alpha}\delta_K(\theta_1) + \frac{\alpha}{1+\alpha}F_0$	mixture

Constrained DPM

Yang et al. 2010 *Computational Statistics and Data Analysis*

- How do we constrain $\frac{1}{N}\sum_i \mu_i = 0$?
- Simply sample $(\tilde{\mu}_i, \tilde{\sigma}_i)|G \sim G$ as usual
  Let $\tilde{\mu}_0 = \frac{1}{N}\sum_i \tilde{\mu}_i$
  And $\mu_i = \tilde{\mu}_i - \tilde{\mu}_0$
- Similarly, if we need to constrain $\frac{1}{N}\sum_i \sigma_i^2 = 1$
  Let $\tilde{\sigma}_0 = \sqrt{\frac{1}{N}\sum_i \tilde{\sigma}_i^2}$
  And $\sigma_i = \tilde{\sigma}_i / \tilde{\sigma}_0$

Low Information Omnibus (LIO) prior hierarchy

Shi, Martens, Banerjee, Laud 2018 *Bayesian Analysis*

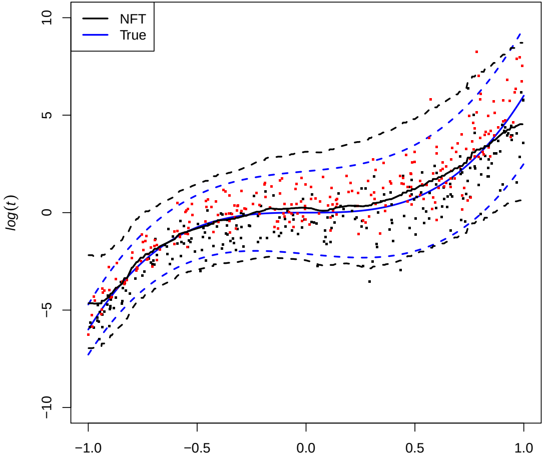
- With either DPM or Constrained DPM
- For convenience, re-parameterize in terms of $\tau_i = \sigma_i^{-2}$
  $F_0(\mu_0, k_0, a_0, b_0)$ is NoGa
  $[\mu_i, \tau_i | k_0, b_0] = [\tau_i | b_0]\,[\mu_i | \tau_i, k_0]$
  with $\tau_i | b_0 \overset{\text{prior}}{\sim} \text{Gamma}\,(a_0,\ b_0)$
  and $\mu_i | \tau_i, k_0 \overset{\text{prior}}{\sim} \text{N}\big(\mu_0,\ (\tau_i k_0)^{-1}\big)$
- LIO prior parameter settings: (standardized, unstandardized; **finite variance of errors for NFT**)
  $(0, m_0; 0)$ elicited median of the data
  $(1, s_0; 1)$ elicited $\frac{1}{2}$ distance from the median to the 95%-ile
  if no other recourse, then m_0 and s_0 can be set empirically
  $\mu_0 = (0, m_0/s_0; 0)$
  $k_0 \overset{\text{prior}}{\sim} \text{Gamma}\left(1.5,\ (7.5, 7.5/s_0^2; 7.5)\right)$
  $a_0 = (1.5, 2; 3)$
  $b_0 \overset{\text{prior}}{\sim} \text{Gamma}\left((0.5, 0.5; 2),\ (1, 1/s_0^2; 1)\right)$

NFT model: prediction intervals

- $\log t_i = y_i = \mu(x_i) + \epsilon_i$ where $\epsilon_i \sim \mathrm{N}\big(\mu_i,\ \sigma_i^2 \sigma^2(x_i)\big)$
  To ensure identifiability: $\frac{1}{N}\sum_i \mu_i = 0$ and $\frac{1}{N}\sum_i \sigma_i^2 = 1$
- $F_\epsilon = \frac{1}{N}\sum_i \mathrm{N}\big(\mu_i,\ \sigma_i^2\big)$: nonparametric mixture of Normals
- $1-\alpha$ Prediction Interval
  $(\mu(x) + c_{\alpha/2}\sigma(x),\ \mu(x) + c_{1-\alpha/2}\sigma(x))$
  where $c_\pi = F_\epsilon^{-1}(\pi)$

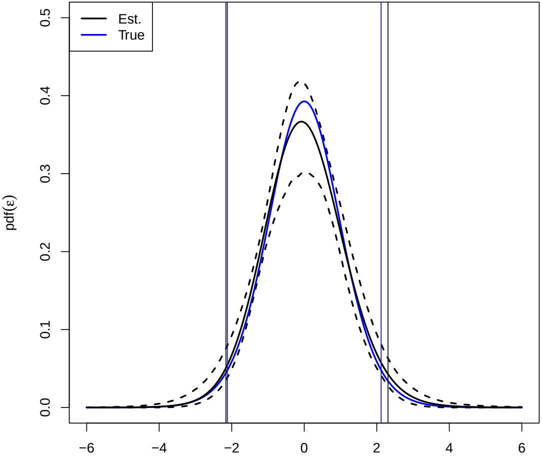
NFT scenario $t(16)$: $N = 500$ with 50% censoring

$f(x) = 6x^3$, $s(x) = \exp 0.5x$,
$\log t = f(x) + s(x)\epsilon$ where $\epsilon \sim t(16)$
and $x \sim \mathrm{U}(-1, 1)$: $R^2 = 84.8\%$ uncensored, $R^2 = 85.1\%$ censored



NFT scenario $t(16)$: $N = 500$ with 50% censoring

$f(x) = 6x^3,\ s(x) = \exp 0.5x,$
$\log t = f(x) + s(x)\epsilon$ where $\epsilon \sim t(16)$
and $x \sim \mathrm{U}(-1,1)$: $R^2 = 84.8\%$ uncensored, $R^2 = 85.1\%$ censored

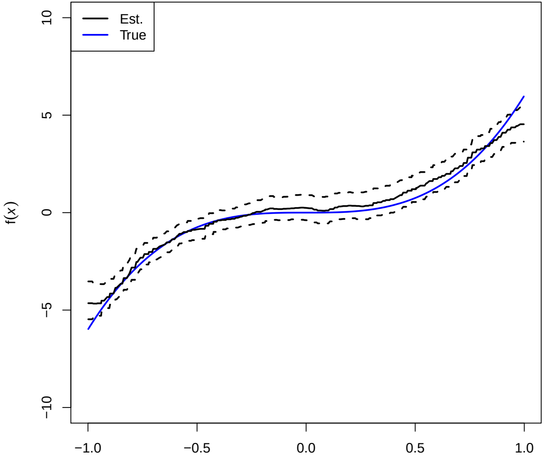


NFT scenario $t(16)$: $N = 500$ with 50% censoring

$f(x) = 6x^3$, $s(x) = \exp 0.5x$,

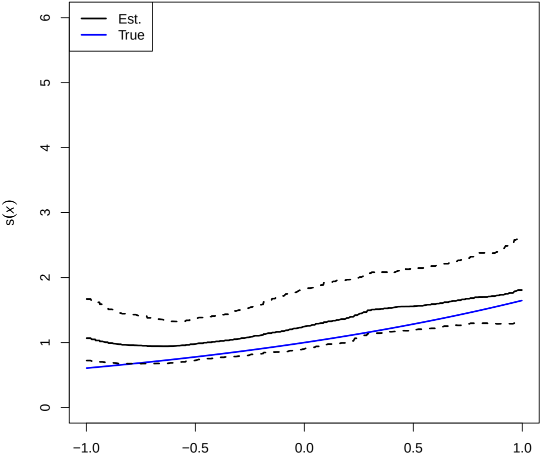
$\log t = f(x) + s(x)\epsilon$ where $\epsilon \sim t(16)$

and $x \sim \mathrm{U}(-1, 1)$: $R^2 = 84.8\%$ uncensored, $R^2 = 85.1\%$ censored



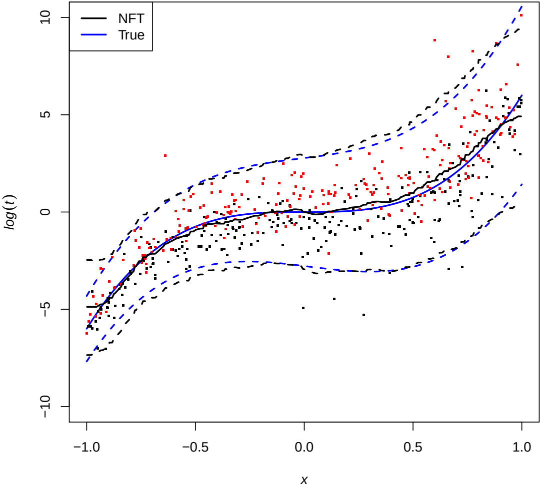
NFT scenario $t(16)$: $N = 500$ with 50% censoring

$f(x) = 6x^3,\ s(x) = \exp 0.5x,$
$\log t = f(x) + s(x)\epsilon$ where $\epsilon \sim t(16)$
and $x \sim \mathrm{U}(-1, 1)$: $R^2 = 84.8\%$ uncensored, $R^2 = 85.1\%$ censored



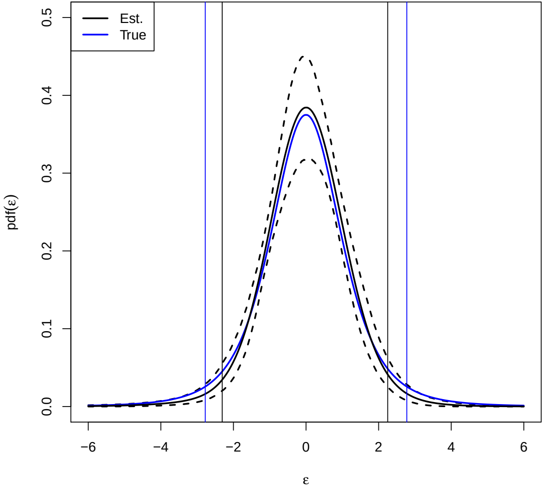
NFT scenario $t(4)$: $N = 500$ with 50% censoring

$f(x) = 6x^3$, $s(x) = \exp 0.5x$, $\log t = f(x) + s(x)\epsilon$ where $\epsilon \sim t(4)$
and $x \sim \mathrm{U}(-1, 1)$: $R^2 = 80.7\%$ uncensored, $R^2 = 78.3\%$ censored



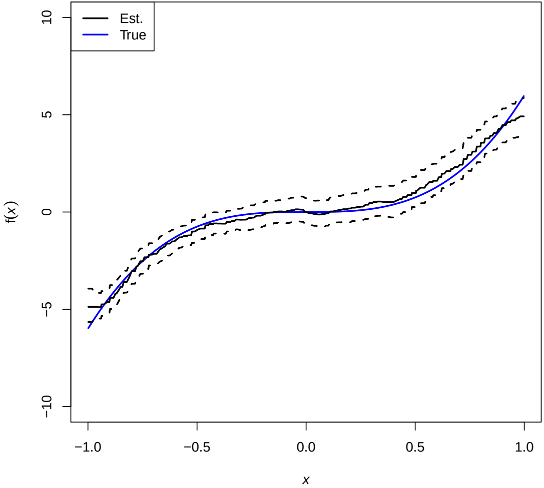
NFT scenario $t(4)$: $N = 500$ with 50% censoring

$f(x) = 6x^3$, $s(x) = \exp 0.5x$, $\log t = f(x) + s(x)\epsilon$ where $\epsilon \sim t(4)$ and $x \sim \mathrm{U}(-1, 1)$: $R^2 = 80.7\%$ uncensored, $R^2 = 78.3\%$ censored



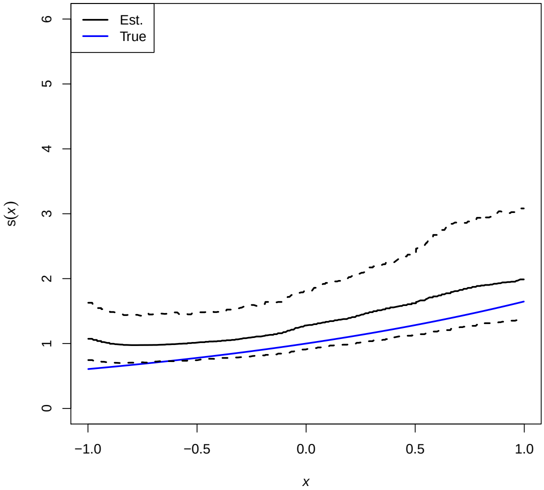
NFT scenario $t(4)$: $N = 500$ with 50% censoring

$f(x) = 6x^3$, $s(x) = \exp 0.5x$, $\log t = f(x) + s(x)\epsilon$ where $\epsilon \sim t(4)$
and $x \sim \mathrm{U}(-1, 1)$: $R^2 = 80.7\%$ uncensored, $R^2 = 78.3\%$ censored



NFT scenario $t(4)$: $N = 500$ with 50% censoring

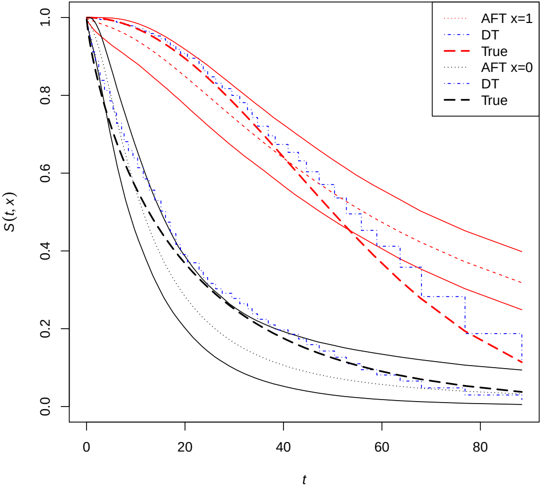
$f(x) = 6x^3$, $s(x) = \exp 0.5x$, $\log t = f(x) + s(x)\epsilon$ where $\epsilon \sim t(4)$ and $x \sim \mathrm{U}(-1, 1)$: $R^2 = 80.7\%$ uncensored, $R^2 = 78.3\%$ censored



Neither AFT nor NFT scenario: AFT failure!

$N = 500$ with 50% censoring

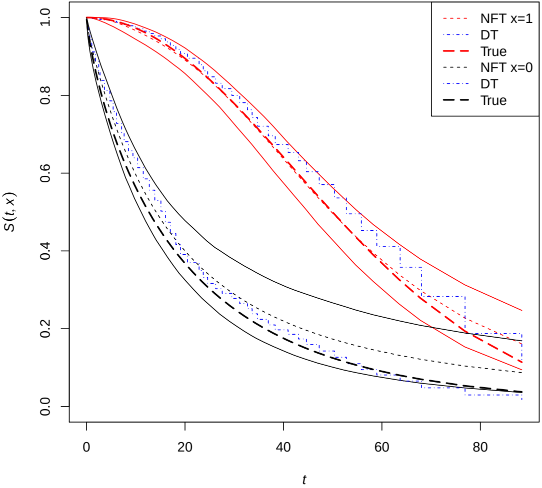
$\mathbf{Wei}\,(0.8 + 1.2x, 20 + 40x)$ where $x \sim \mathrm{B}(0.5)$



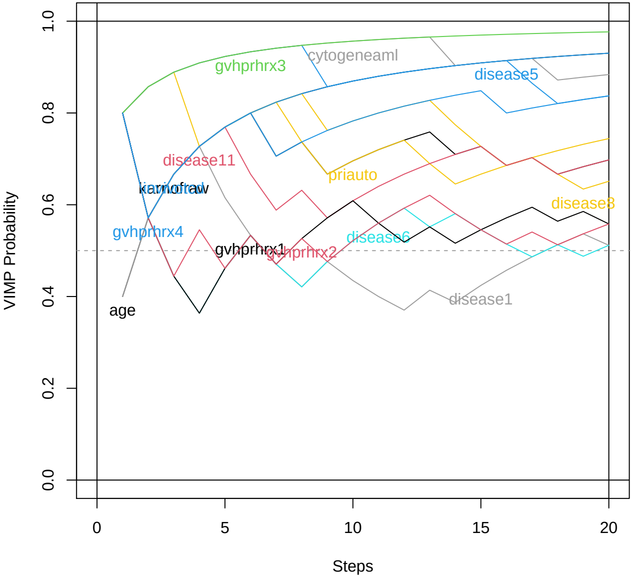
Neither AFT nor NFT scenario: NFT success!

$\boldsymbol{N = 500}$ with $\boldsymbol{50\%}$ censoring

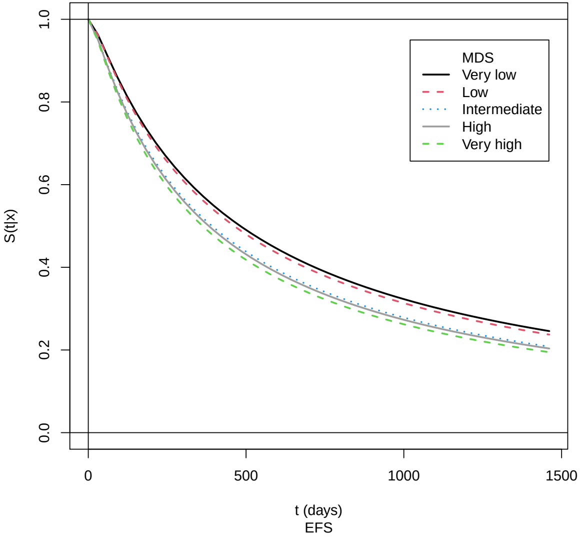
$\mathbf{Wei}\,(\mathbf{0.8 + 1.2}\boldsymbol{x}, \mathbf{20 + 40}\boldsymbol{x})$ **where** $\boldsymbol{x} \sim \mathbf{B}(\mathbf{0.5})$



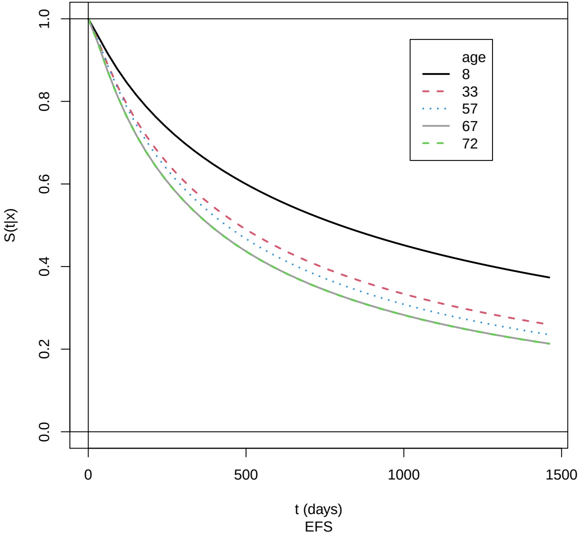
Event-free Survival: TSVS



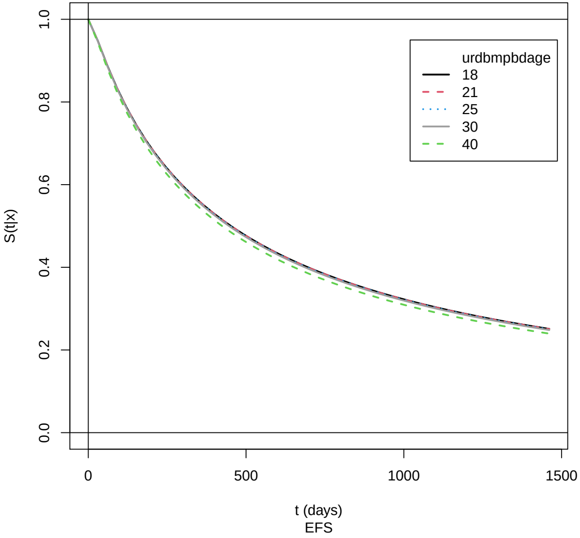
Event-free Survival: MDS `disease5`



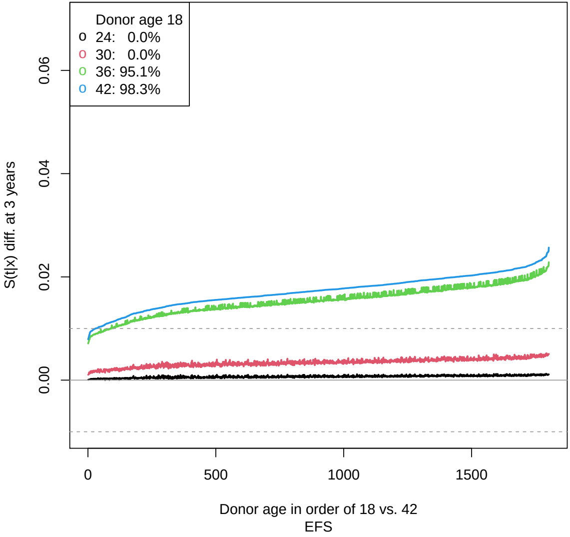
Event-free Survival: Recipient Age



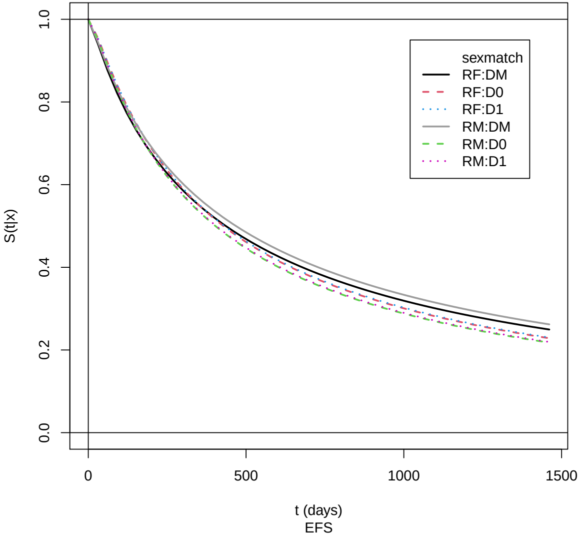
Event-free Survival: Donor Age



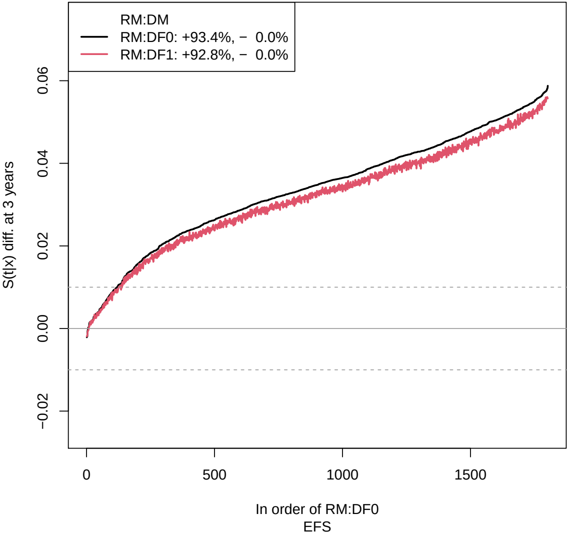
Event-free Survival: Donor Age Waterfall Plot



Event-free Survival: Donor Sex/Child-birth Parity



Event-free Survival: Donor Sex/Child-birth Parity



Event-free Survival: Donor Sex/Child-birth Parity