# Introduction to BART with time-to-event outcomes

Rodney Sparapani
**Medical College of Wisconsin**


September 30: BART Bootcamp
Biostatistics in the Modern Computing Era
Medical College of Wisconsin, Milwaukee campus

*Funding for this research was provided, in part, by the Advancing Healthier Wisconsin Research and Education Program under awards 9520277 and 9520364.*

# Outline

- Motivation: advanced lung cancer prognosis
- Survival analysis with BART
- Motivation: diabetes and recurrent hospital admissions
- Recurrent events with BART
- Motivation: liver transplant waiting list
- Competing risks

# Survival analysis with Cox Proportional Hazards

Cox 1972 *JRSS-B*

Data: $(s_i, \delta_i, x_i)$

$\mathbf{0} = t_{(0)} < \cdots < t_{(J)} < \infty$: distinct ordered death, $s_i$, times

$$(0, t_{(1)}] \ldots (t_{(J-1)}, t_{(J)}]$$

$$\lambda(t, x_i) = \lambda_0(t) e^{\beta' x_i}$$ Linear proportionality

$$[\beta | \lambda_0(t)] = \prod_i \frac{e^{\beta' x_i}}{\sum_{j \in R(t_i)} e^{\beta' x_j}}$$ Partial Likelihood

$$\hat{S}_0(t) = e^{-\hat{\Lambda}_0(t)} \text{ where } \hat{\Lambda}_0(t) = \sum_{t_i \le t} \frac{\delta_i}{\sum_{j \in R(t_i)} e^{\hat{\beta}' x_i}}$$

$$\hat{S}(t, x_i) = \hat{S}_0(t)^{\exp(\hat{\beta}' x_i)}$$

# Survival analysis with BART

Sparapani et al. 2016 *Statistics in medicine*

$$\mathbf{0} = t_{(0)} < \cdots < t_{(K)} < \infty : \text{distinct ordered, } s_i, \text{ times}$$

$$y_{ij}|p_{ij} \overset{\text{ind}}{\sim} \mathbf{B}\left(p_{ij}\right) \text{ where } j = 1, \ldots, J_i = \arg\min_j s_i \leq t_{(j)}$$

$$y_{ij} = \delta_i \mathbf{I}\left(j = J_i\right)$$

$$p_{ij} = p\left(t_{(j)}, x_{ij}\right) \text{ where } x_{ij} = x_i\left(t_{(j)}\right)$$

$$= \Phi\left(f\left(t_{(j)}, x_{ij}\right)\right) \text{ where } f \overset{\text{prior}}{\sim} \text{BART}$$

$$[y|p] = \prod_{i=1}^{N} \prod_{j=1}^{J_i} p_{ij}^{y_{ij}} \left(1 - p_{ij}\right)^{1-y_{ij}} \quad \text{Likelihood}$$

$$S\left(t_{(j)}, x_{ij}\right) = \mathbf{P}\left[t > t_{(j)}|x_{ij}\right] = \prod_{j' \leq j} \left(1 - p_{ij'}\right)$$

Discrete time intensity model $\Rightarrow$ longitudinal binary BART

# Survival analysis with BART and inference

We generate samples of $f$ from the posterior with MCMC

$$\hat{f}(t,x) = M^{-1}\sum_m f_m(t,x) \qquad \text{Estimate } f$$

$$\hat{S}(t,x) = M^{-1}\sum_m S_m(t,x) \qquad \text{Survival function}$$

$$(S_{0.025}(t,x), S_{0.975}(t,x)) \qquad \text{95\% Credible Interval}$$

$$RI(t,x_n(t),x_d(t)) = \frac{p(t,x_n(t))}{p(t,x_d(t))} \qquad \text{Relative Risk or Intensity}$$

$$= \frac{\Phi(f(t,x_n(t)))}{\Phi(f(t,x_d(t)))}$$

# Survival analysis with BART and Friedman's partial dependence function

Friedman 2001 *AnnStat*

$f(t,x) = f(t,x_S,x_C)$ BART function where $x = [x_S, x_C]$

$$f(t,x_S) = \mathrm{E}_{x_C}\left[f(t,x_S,x_C)|x_S\right]$$

$$\approx N^{-1}\sum_i f(t,x_S,x_{iC})$$

$$f_m(t,x_S) \equiv N^{-1}\sum_i f_m(t,x_S,x_{iC})$$

$$\hat{f}(t,x_S) \equiv M^{-1}\sum_m f_m(t,x_S)$$

# surv.bart and mc.surv.bart input and output

```
post=surv.bart(x.train, times=times, delta=delta,
    ..., ndpost=M, keepevery=10) or
post=mc.surv.bart(x.train, times=times, delta=delta,
    ..., ndpost=M, keepevery=10, mc.cores=2, seed=99)
```

Input vector `times` with $K$ distinct values and `x.train`: $x_i$

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}$$

Output `post`, of type `survbart` which is essentially a list of matrices including: `post$surv.train`: $\hat{S}_m(t_{(j)}, x_i)$

$$\begin{bmatrix} \hat{S}_1(t_{(1)},x_1) & \dots & \hat{S}_1(t_{(K)},x_1) & \dots & \hat{S}_1(t_{(1)},x_N) & \dots & \hat{S}_1(t_{(K)},x_N) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \hat{S}_M(t_{(1)},x_1) & \dots & \hat{S}_M(t_{(K)},x_1) & \dots & \hat{S}_M(t_{(1)},x_N) & \dots & \hat{S}_M(t_{(K)},x_N) \end{bmatrix}$$

# `surv.pre.bart` input and output

```
post <- surv.pre.bart(times, delta, x.train,
    x.test=x.train)
```

Output a list containing the data transformed such as

matrix `pre$tx.train` and vector `pre$y.train`:

$$\begin{bmatrix} t_{(1)} & x_1 \\ \vdots & \vdots \\ t_{(J_1)} & x_1 \\ \vdots & \vdots \\ t_{(1)} & x_N \\ \vdots & \vdots \\ t_{(J_N)} & x_N \end{bmatrix} \quad \begin{bmatrix} y_{11} = 0 \\ \vdots \\ y_{1J_1} = \delta_1 \\ \vdots \\ y_{N1} = 0 \\ \vdots \\ y_{NJ_N} = \delta_N \end{bmatrix}$$

N.B. for `pre$tx.test` $J_i = K$

# `predict.survbart` input and output

```
pred <- predict(post, pre$tx.test, mc.cores=1, ...)
```

Input matrices: `x.test`: $x_i$

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_Q \end{bmatrix}$$

Output `pred` of type `survbart` with `pred$surv.test`: $\hat{S}_m(t_{(j)}, x_i)$

$$\begin{bmatrix} \hat{S}_1(t_{(1)},x_1) & \dots & \hat{S}_1(t_{(K)},x_1) & \dots & \hat{S}_1(t_{(1)},x_Q) & \dots & \hat{S}_1(t_{(K)},x_Q) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \hat{S}_M(t_{(1)},x_1) & \dots & \hat{S}_M(t_{(K)},x_1) & \dots & \hat{S}_M(t_{(1)},x_Q) & \dots & \hat{S}_M(t_{(K)},x_Q) \end{bmatrix}$$

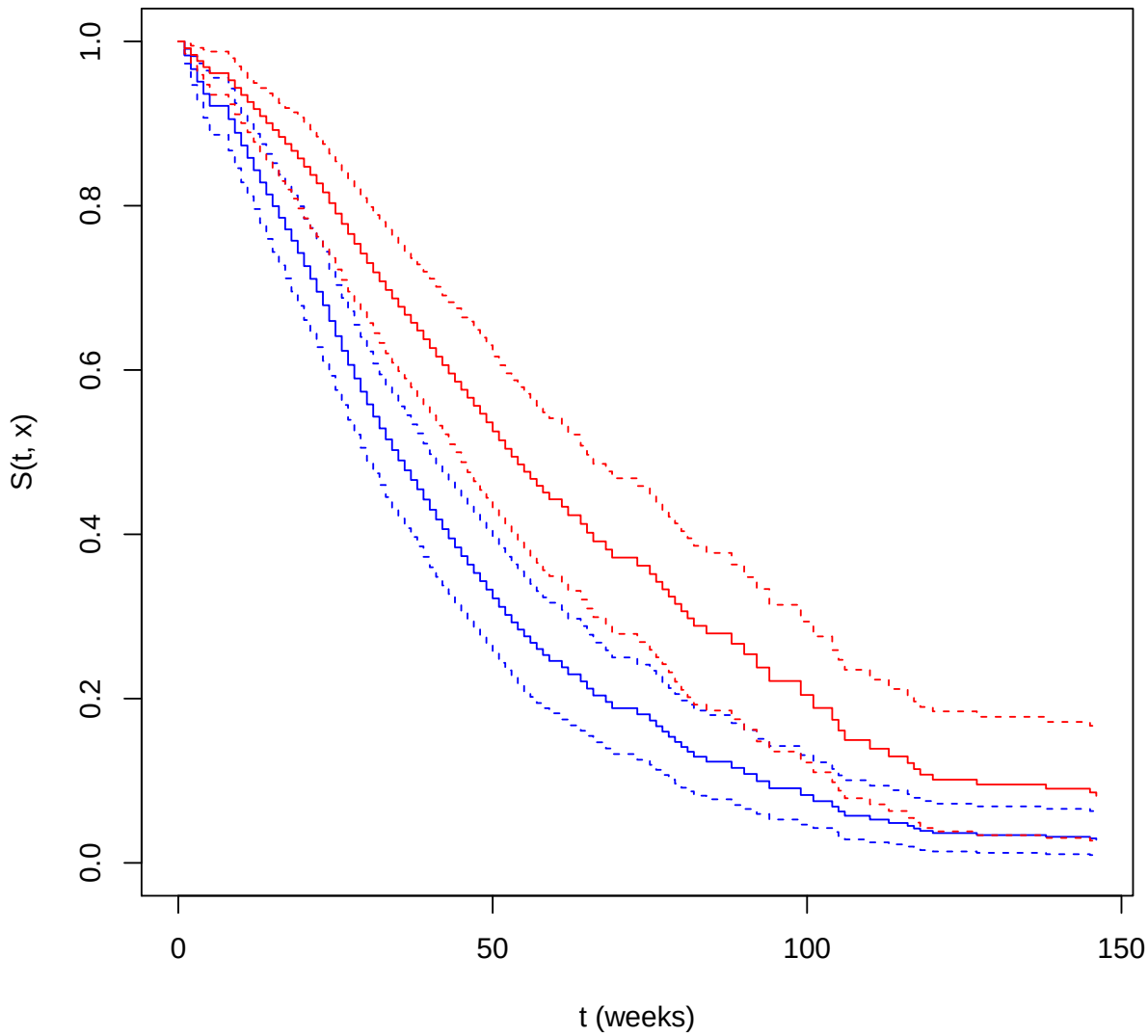
# Survival analysis: advanced lung cancer prognosis

Loprinzi et al. 1994 *JCO*

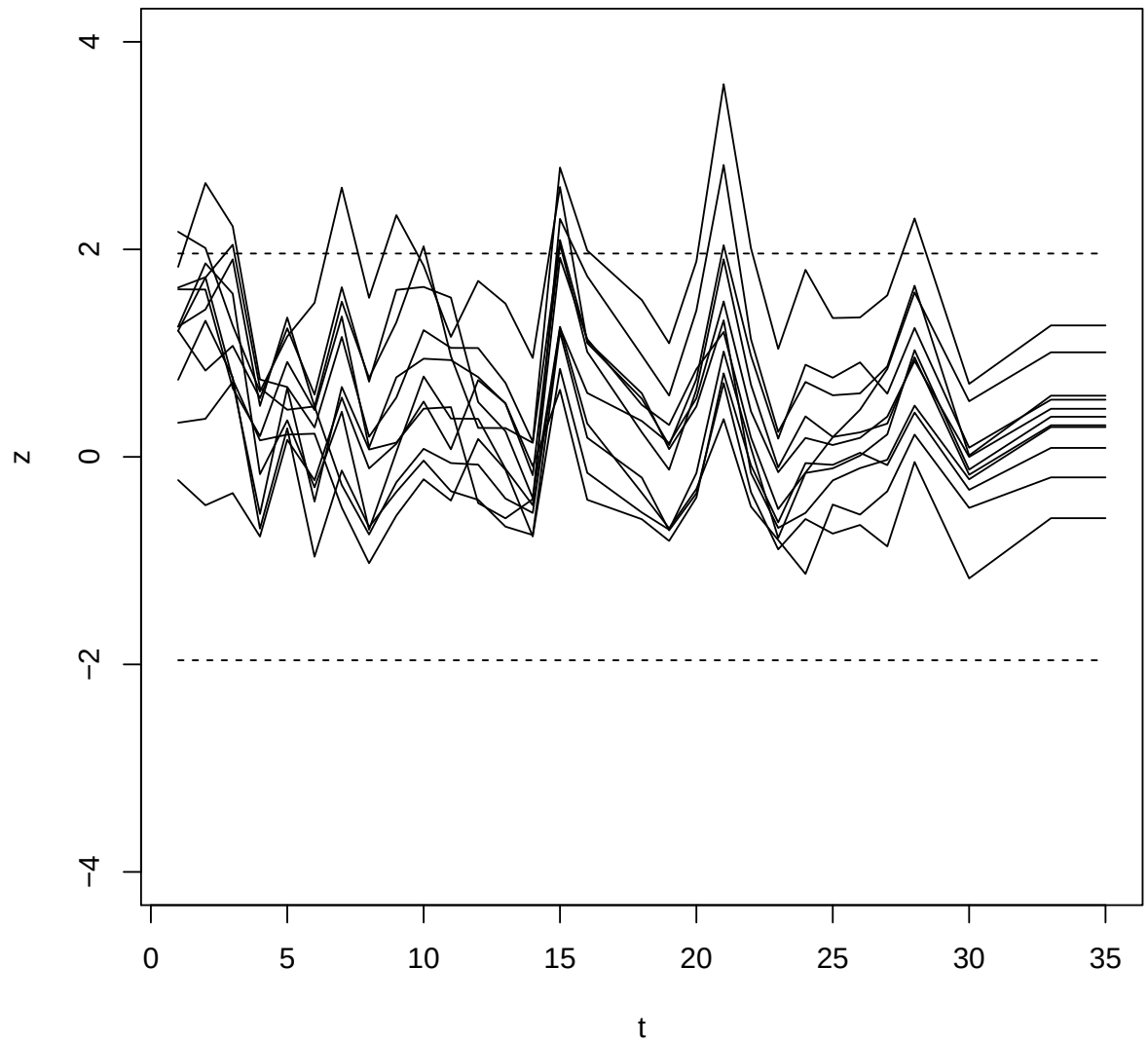
- The North Central Cancer Treatment Group surveyed 228 advanced lung cancer patients
- Study focused on prognostic variables
- Patient responses paired with some clinical variables
- We control for age, gender and Karnofsky performance score as rated by the physician
- We will compare males to females with Friedman's partial dependence function
- `lung` data set in the **BART** R package

```
system.file('demo/lung.surv.bart.R', package='BART')
system.file('demo/geweke.lung.surv.bart.R',
package='BART')
```

# Friedman's partial dependence function with 95% credible intervals: M (blue) vs. F (red)



# Geweke convergence diagnostics: Advanced lung cancer example



# Geweke convergence diagnostics: live demonstration

- `system.file('demo/geweke.surv.bart.R', package='BART')`
- Simulated data set: $N = 100, P = 20$
- $t_i \sim \mathbf{Wei}\left(2, \mathbf{e}^{f(x_i)}\right)$
- adapted from Friedman's five-dimensional test function *Annals of Statistics 1991*
- $f(x_i) = 3 + \sin(\pi x_1 x_2) - 2(x_3 - 0.5)^2 + x_4 - 0.5x_5$
- 20% censoring

# Diabetes and recurrent hospital admissions

- We have IRB approval to study a cohort of newly diagnosed diabetes patients from a single health care system
- We have the electronic health records (EHR) for these patients from 2007-2012: prior records may, or may not, be available
- EHR are an omnibus of digital health care information
- We focus on 82 covariates: patient demographics, health insurance, health care charges, diagnoses, procedures, anti-diabetic therapy, laboratory values and vital signs
- By its nature, EHR data is fundamentally time-varying
- EHR covariates are occasionally missing even when carrying the last value forward
- Imputed 15 continuous variables with Sequential BART (Xu, Daniels & Winterstein 2016 *Biostatistics*)

# Diabetes and recurrent hospital admissions

- 488 patients followed 5 years from 2008-2012
  the survival rate was high 0.939 (noninformative censoring)
  yet experienced a high rate of hospital admissions: 525 total
- For diabetes, which covariates increase the risk of admission? What about the number of previous admissions or an acutely recent admission?
- What are the functional forms of the covariates i.e. linear, quadratic, logarithm, etc.? Are the covariate effects additive or multiplicative?
- Are there interactions? Are these effects constant with respect to time, i.e., proportionality assumption?
- We want to avoid precarious restrictive assumptions hence we chose to use Bayesian Additive Regression Trees (BART)

# Recurrent event analysis with BART

Data: $(s_i, t_{i1}, \ldots, t_{iN_i}, x_i(t))$

$(0, t_{(1)}] \ldots (t_{(K-1)}, t_{(K)}]$: grid of distinct ordered times, $t_{ik}$

$$y_{ij}|p_{ij} \overset{\text{ind}}{\sim} \mathrm{B}(p_{ij}) \qquad j = 1, \ldots, J_i$$

$$y_{ij} = \max_{k=1,\ldots,N_i} \mathrm{I}\left(t_{ik} = t_{(j)}\right)$$

$$p_{ij} = \Phi(f(t_{(j)}, x_{ij})) \qquad f \overset{\text{prior}}{\sim} \mathbf{BART}$$

$$[y|p] = \prod_{i=1}^{N} \prod_{j=1}^{J_i} p_{ij}^{y_{ij}} (1 - p_{ij})^{1-y_{ij}} \qquad \text{Likelihood}$$

$$\Lambda(t_{(j)}, x_{ij}) = \int_0^{t_{(j)}} \mathrm{d}\Lambda(t, x_i(t)) = \sum_{j'=1}^{j} p_{ij'}$$

Discrete time intensity model $\Rightarrow$ longitudinal binary BART

# Semi-Markov process and conditional independence

- $(t_{i1}, \ldots, t_{iN_i})$ are not independent; rather, they are conditionally independent given $x_i(t)$ and the event history which is denoted by $N_i(t)$
- $N_i(t)$ is the counting process of events and $N_i \equiv N_i(s_i)$ When $N_i = 0$, then $t_{iN_i} = t_{i0} \equiv 0$
- Semi-Markov process, i.e., condition on summaries of the event history just prior to time $t$ which is denoted by $t-$

| | |
|---|---|
| Counting process just prior to time $t$ | $N_i(t-)$ |
| Sojourn time from the last event | $v_i(t) \equiv t - t_{iN_i(t-)}$ |

$$y_{ij}|p_{ij} \overset{\text{ind}}{\sim} \mathrm{B}\,(p_{ij})$$

$$p_{ij} = \Phi(f(t_{(j)}, \widetilde{x}_{ij}))$$

$$\text{where } \widetilde{x}_{ij} = \left[v_i(t_{(j)}), N_i(t_{(j-1)}), x_{ij}\right]$$

# Diabetes and recurrent hospital admissions

| | Patients | | Admissions | |
|---|---|---|---|---|
| Number of Admissions | 488 | | 525 | |
| 0 | 308 | (63.0) | 0 | |
| 1 | 79 | (16.2) | 79 | (15.0) |
| 2-3 | 50 | (10.3) | 115 | (21.9) |
| 4-16 | 51 | (10.5) | 331 | (63.1) |

# Diabetes and recurrent hospital admissions

- We focus on 82 covariates: patient demographics, health insurance, health care charges, diagnoses, procedures, anti-diabetic therapy, laboratory values and vital signs
- These covariates are inherently time-dependent and occasionally missing even when carrying the last value forward
- Imputed 15 continuous variables with Sequential BART
  8 lab values and 7 vital signs
  Xu, Daniels & Winterstein 2016 *Biostatistics*
- Variable selection: Decoupling Shrinkage and Selection (DSS)
  Hahn & Carvalho 2015 *JASA*; McCulloch et al. 2015 JSM
- Divided the cohort at random into training and validation sets
- Risk agonists: insulin treatment, peripheral vascular disease (PVD) and the number of previous admissions, $N_i(t-)$

# Diabetes and recurrent hospital admissions

| | Patients | | Admissions | |
|---|---|---|---|---|
| Gender | 488 | | 525 | |
| M | 216 | (44.3) | 228 | (43.4) |
| F | 272 | (55.7) | 297 | (56.6) |
| Race | 488 | | 525 | |
| Black | 174 | (35.7) | 265 | (50.5) |
| White | 314 | (64.3) | 260 | (49.5) |
| Age | 488 | | 525 | |
| Mean, SD | 60.9 | 15.0 | 60.3 | 15.7 |
| ZIP3 area | 488 | | 525 | |
| 532/urban | 378 | (77.5) | 454 | (86.5) |
| 530/suburb | 110 | (22.5) | 71 | (13.5) |
| Insurance and Age | 488 | | 525 | |
| Government 65+ | 191 | (39.1) | 224 | (42.7) |
| Government <65 | 138 | (28.3) | 208 | (39.6) |
| Commercial <65 | 143 | (29.3) | 71 | (13.5) |
| Other <65 | 16 | ( 3.3) | 22 | ( 4.2) |

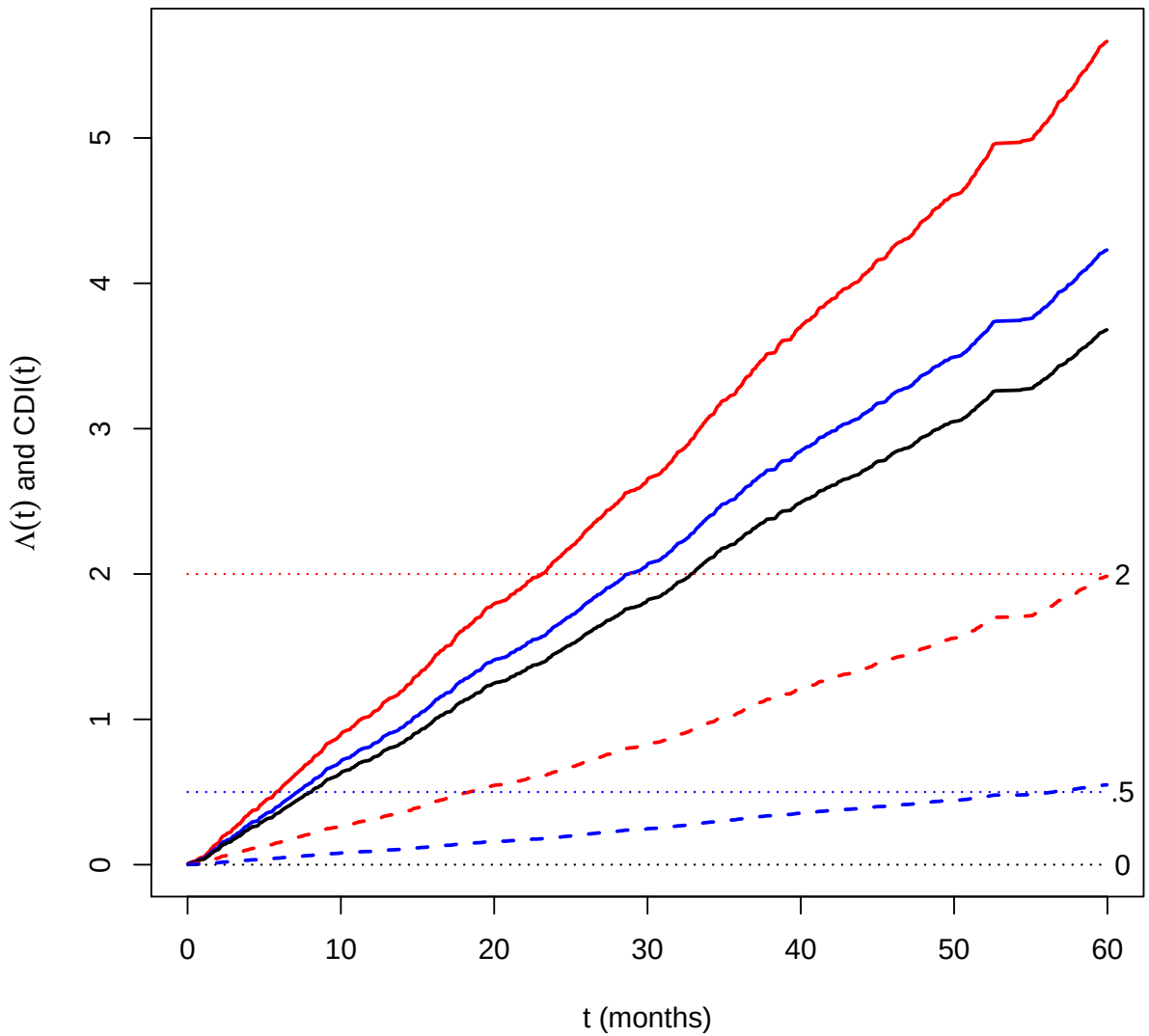
# Diabetes and recurrent hospital admissions

| | Patients | | Admissions | | Relative Intensity | 95% Credible Interval |
|---|---|---|---|---|---|---|
| Insulin | 488 | | 525 | | 2.39 | 1.56, 3.25 |
| Yes | 206 | (42.2) | 391 | (74.5) | | |
| No | 282 | (57.8) | 134 | (25.5) | | |
| PVD | 488 | | 525 | | 2.90 | 2.00, 3.89 |
| Yes | 272 | (55.7) | 488 | (93.0) | | |
| No | 216 | (44.3) | 37 | ( 7.0) | | |
| | | | | | partial dependence function | |

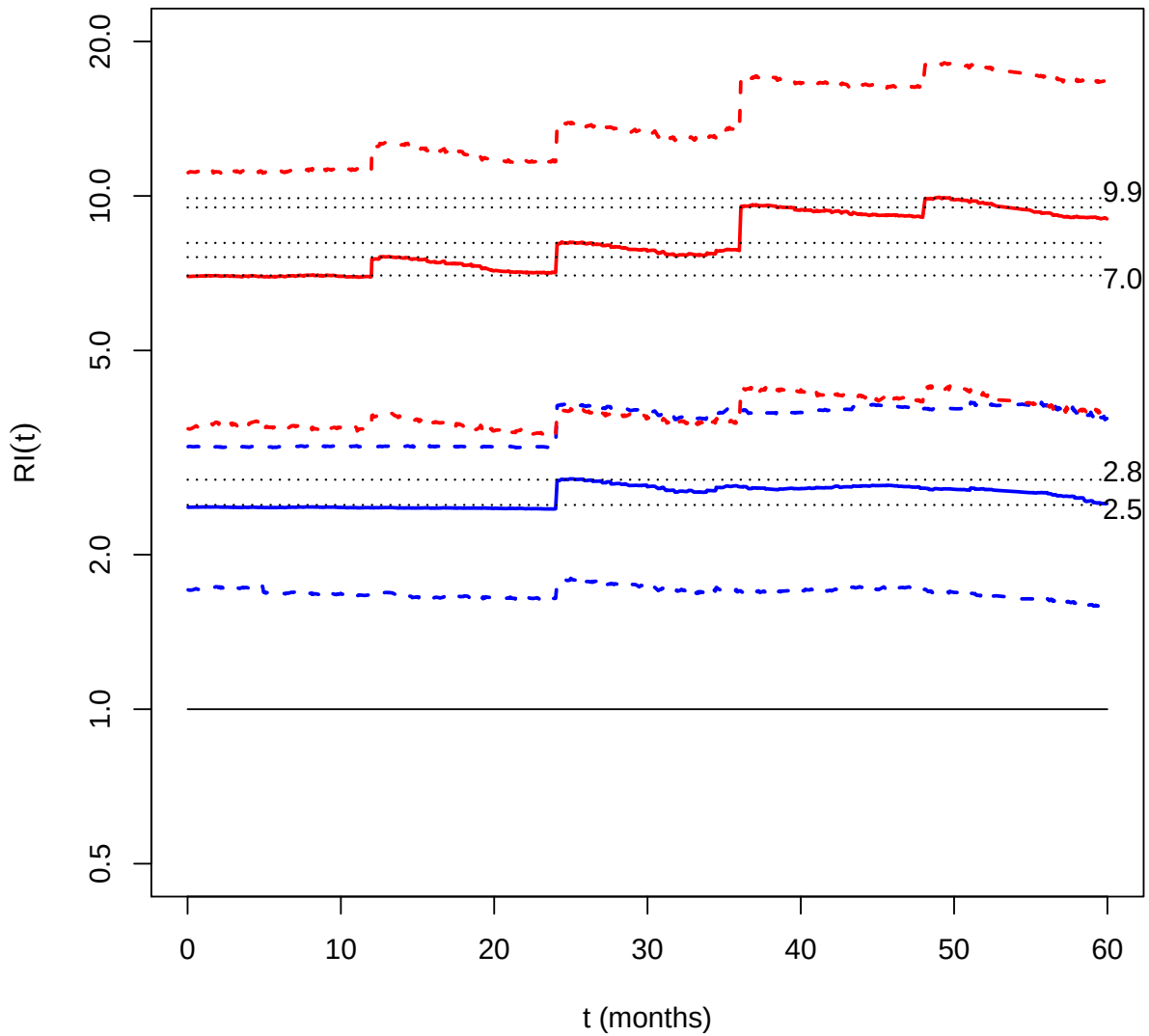
# Hospital admission risk profiles

| Risk | Insulin | PVD | $N_i(t)$ with time in months 0 | 12 | 24 | 36 | 48 | 60 |
|---|---|---|---|---|---|---|---|---|
| Low | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Medium | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| High | 1 | 1 | 0 | 1 | 2 | 3 | 4 | 4 |

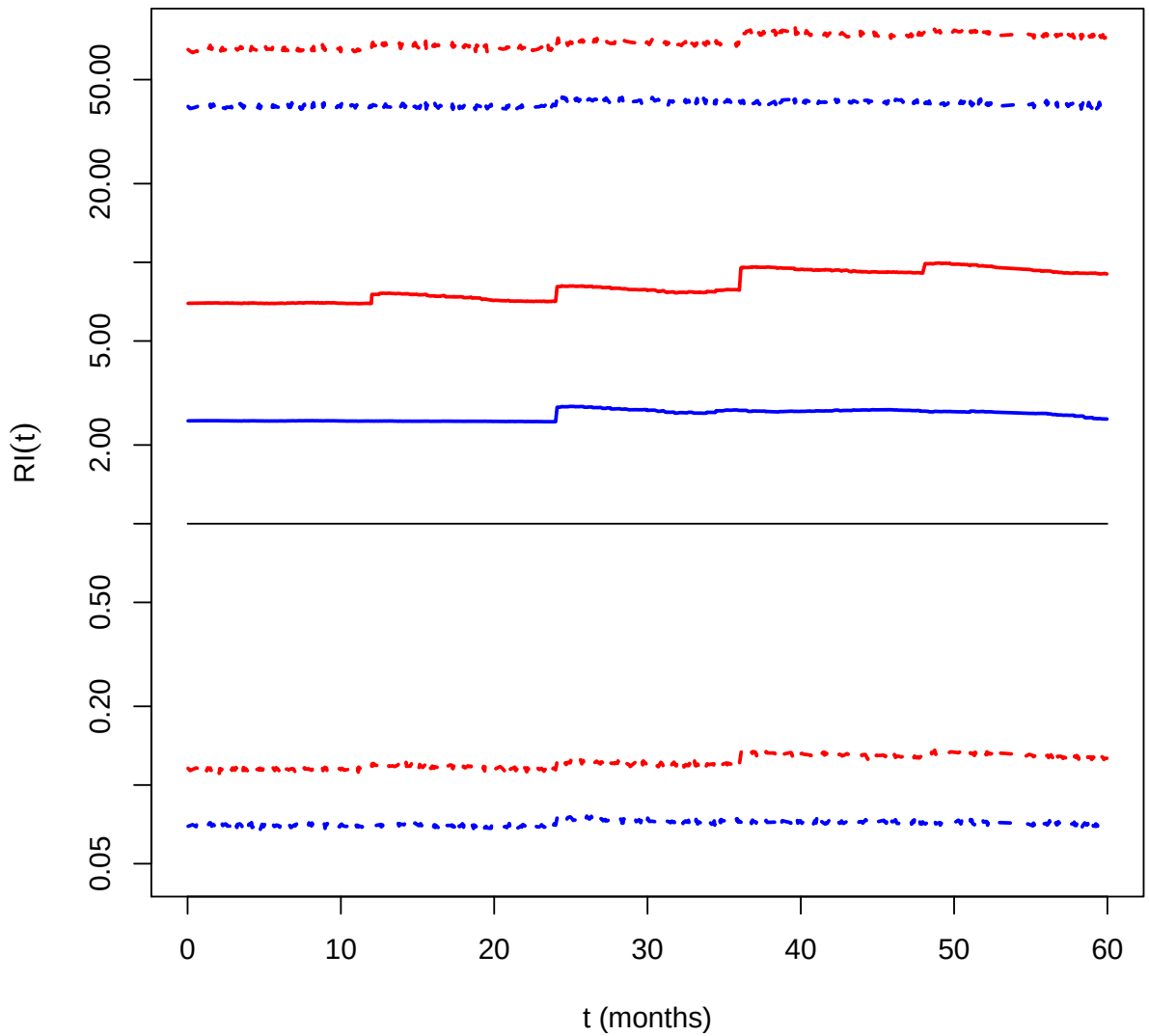
# Risk profiles: Cumulative Intensity partial dependence function



# Risk profiles: Relative Intensity and 95% Credible Intervals partial dependence function



# Risk profiles: Relative Intensity & 95% Prediction Intervals partial dependence function



# Diabetes and hospital admission risk

- Some diabetes patients are at high risk for hospital admission
  - diagnosed with PVD
  - prescribed insulin therapy
  - with a recent hospital admission
  - and/or several previous hospital admissions
- Health policy implications: Diabetic patients' health care post-discharge should be carefully orchestrated to ensure the delivery of quality clinical care which maximizes healthy outcomes while preventing adverse events and costly unnecessary hospital admissions
- **BART** package contains a roughly 20% random sample
  50 patients from training: `ydm20.train` & `xdm20.train`
  50 patients from validation: `xdm20.test`
- complete data set at `http://www.mcw.edu/FileLibrary/Groups/Biostatistics/TechReports/TechReports5175/tr064.tar`.

# Diabetes and hospital admission risk

- Acknowledgments and Disclaimers

*This research was supported, in part, by the Advancing Healthier Wisconsin Research and Education Program under award 9520277 and the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant Number UL1TR001436. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH. The data have been supplied by the Clinical and Translational Science Institute of Southeast Wisconsin's Clinical Research Data Warehouse at the Medical College of Wisconsin. The interpretation and reporting of these data are the responsibility of the author(s) and in no way should be seen as an official policy of, or interpretation by, the Medical College of Wisconsin.*

# Competing risks

Data: $(s_i, \delta_i, x_i(t))$ where $\delta_i \in \{0, 1, 2\}$

$0 = t_{(0)} < \cdots < t_{(K)} < \infty$: distinct ordered, $s_i$, times

$$
\begin{aligned}
y_{1ij} &= \mathrm{I}(\delta_i = 1)\,\mathrm{I}(j = J_i),\ j = 1, \ldots, J_i \\
y_{1ij} | p_{1ij} &\sim \mathrm{B}(p_{1ij}) \\
p_{1ij} &= \Phi(f_1(t_{(j)}, x_{ij})) \text{ where } f_1 \sim \text{BART} \\
y_{2ij} &= \mathrm{I}(\delta_i = 2)\,\mathrm{I}(j = J_i),\ j = 1, \ldots, K_i \\
&\quad \text{where } K_i = J_i - \mathrm{I}(\delta_i = 1) \\
y_{2ij} | p_{2ij} &\sim \mathrm{B}(p_{2ij}) \\
p_{2ij} &= \Phi(f_2(t_{(j)}, x_{ij})) \text{ where } f_2 \sim \text{BART}
\end{aligned}
$$

$$
[y|p] = \prod_{i=1}^{N} \left( \prod_{j=1}^{J_i} p_{1ij}^{y_{1ij}} (1 - p_{1ij})^{1 - y_{1ij}} \right)
$$

$$
\times \left( \prod_{j=1}^{K_i} p_{2ij}^{y_{2ij}} (1 - p_{2ij})^{1 - y_{2ij}} \right) \text{ Likelihood}
$$

# Competing risks

$$S(t, x_i(t)) = 1 - F(t, x_i(t)) = \prod_{j=1}^{k} (1 - p_{1ij})(1 - p_{2ij})$$

$$\text{where } k = \arg\max_j \left[ t_{(j)} \leq t \right]$$

$$F_1(t, x_i(t)) = \int_0^t S(u-, x_i(u-)) \lambda_1(u, x_i(u)) \mathrm{d}u$$

$$= \sum_{j=1}^{k} S(t_{(j-1)}, x_i(t_{(j-1)})) p_{1ij}$$

$$F_2(t, x_i(t)) = \int_0^t S(u-, x_i(u-)) \lambda_2(u, x_i(u)) \mathrm{d}u$$

$$= \sum_{j=1}^{k} S(t_{(j-1)}, x_i(t_{(j-1)})) (1 - p_{1ij}) p_{2ij}$$

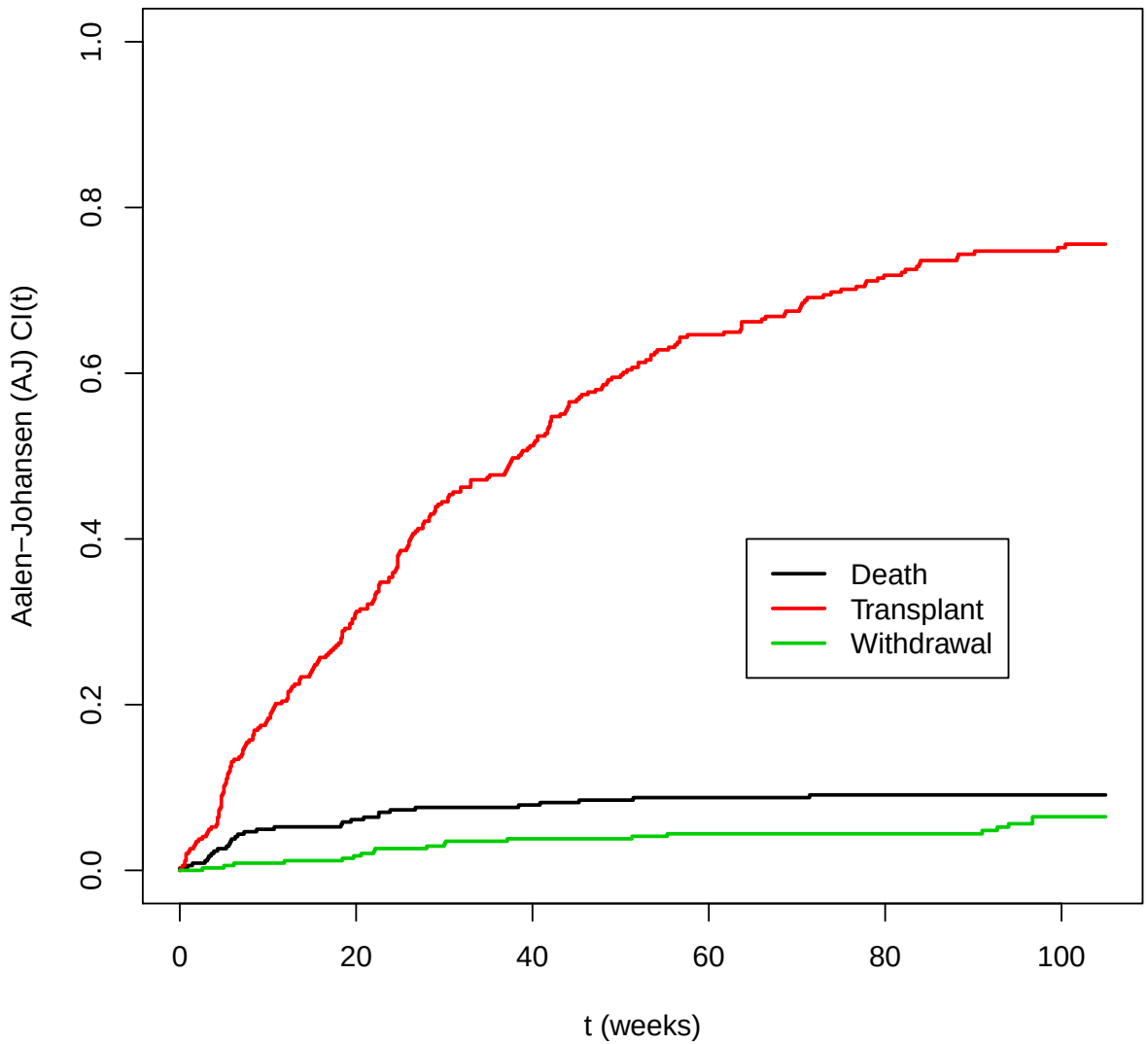
# Liver transplant

Kim et al. 2006 *Hepatology*

- Mayo Clinic Liver transplant waiting list data from 1990-1999
- During this period, liver allocation policy was flawed
- Donor livers from subjects with blood type O can be used by patients with A, B, AB or O blood types, whereas an A, B, AB liver can only be used by an A, B, AB recipient respectively
- Type O subjects on the waiting list were at a disadvantage since the pool of competitors was larger for type O donor livers
- Current policies have evolved and now depend on each individual patient's risk and need which are assessed and updated regularly while a patient is on the waiting list
- However, the overall donor liver shortage remains acute today
- `transplant` data set in **BART** R package: $N = 815$
- `system.file('demo/liver.crisk.bart.R', package='BART')`

# Liver transplant Competing Risks for Type O patients

## Aalen-Johansen estimator available in **survival** R package



# Liver transplant Competing Risks for Type O patients
# Aalen-Johansen and BART