Biostatistics (2018), **0**, 0, *pp.* 1–45
doi:10.1093/biostatistics/supplement-r4

Nonparametric recurrent events analysis with BART and an application to the hospital admissions of patients with diabetes

RODNEY A. SPARAPANI*, LISA E. REIN, SERGEY S. TARIMA,

TOURETTE A. JACKSON, JOHN R. MEURER
Institute for Health and Society, Medical College of Wisconsin,

8701 Watertown Plank Road, Milwaukee, Wisconsin 53226, USA

*rsparapa@mcw.edu

Supplementary Materials

APPENDIX

A. Introduction

Herein contains the Appendices. In Appendix B, we review Cox proportional intensity models commonly employed for recurrent events. In Appendix C, we describe the Sequential BART missing imputation method that was employed in our motivating example. In Appendix D, we describe the simulation study we employed to compare our new BART method vs. counting process Cox models for recurrent events. In Appendix E, we provide an introduction to the **BART** R package. In Appendix F, we discuss how to handle dependent censoring due to death.

*To whom correspondence should be addressed.

In Appendix G, additional details on the motivating example are provided.

B. Recurrent events and Cox proportional intensity models

Recurrent events are often analyzed via Cox proportional intensity models (Kalbfleisch and Prentice, 2002; Hosmer Jr *and others*, 2008). We will briefly outline four Cox models commonly employed. We adopt the following notation: $(s_i, \delta_i, \boldsymbol{t}_i, \boldsymbol{x}_i(t))$ where $i = 1, \ldots, m$ indexes subjects; s_i is the length of the observation period; $\delta_i = 0$ represents a censored event and $\delta_i = 1$ is a death; N_i is the number of events experienced during the observation period; $\boldsymbol{t}_i = [t_{i1}, \ldots, t_{iN_i}]'$ is a vector of the event times; and $\boldsymbol{x}_i(t)$ is vector of covariates that may be time-dependent. For single event survival analysis, the notation collapses, $s_i = t_i$ and $\delta_i = N_i$, and the general form of the Cox proportional intensity model is the following: $\lambda(t, \boldsymbol{x}_i(t)) = \lambda_0(t) \exp(\boldsymbol{\beta}' \boldsymbol{x}_i(t))$ where $\lambda(t, \boldsymbol{x}_i(t))$ is the intensity, $\lambda_0(t)$ is a nonparametric baseline intensity and $\exp(\boldsymbol{\beta}' \boldsymbol{x}_i(t))$ is a parametric multiplier that we call linear proportionality. The cumulative intensity is $\Lambda(t, \boldsymbol{x}_i(t)) = \int_0^t \lambda(s, \boldsymbol{x}_i(s)) \mathrm{d}s$ and the survival probability is $S(t, \boldsymbol{x}_i(t)) = \Pr(s > t | \boldsymbol{x}_i(t)) = \exp(-\Lambda(t, \boldsymbol{x}_i(t)))$. The likelihood contribution for each subject is $\lambda(t_i, \boldsymbol{x}_i(t_i))^{\delta_i} S(t_i, \boldsymbol{x}_i(t_i))$. The four Cox models we present differ in how they adapt the single event Cox model to the recurrent events setting.

B.1 *Counting process Cox model of time (CPC)*

In the counting process model of time, each subject's experience is broken up into independent observation time intervals: $(0, t_{i1}], \ldots (t_{iN_i - 1}, t_{iN_i}]$, $(t_{iN_i}, s_i]$ (which collapses to a single interval $(0, s_i]$ for a subject having no events). Each of these intervals is associated with a corresponding event indicator, δ_{ij}. Note that only the first interval starts at time zero; therefore, the remaining intervals are left truncated, i.e., delayed entry. This arrangement results in the following likelihood contribution for each subject:
$\lambda(t_{i1}, \boldsymbol{x}_i(t_{i1}))^{\delta_{i1}} S(t_{i1}, \boldsymbol{x}_i(t_{i1})) \left[\prod_{j=2}^{N_i} \lambda(t_{ij}, \boldsymbol{x}_i(t_{ij}))^{\delta_{ij}} \frac{S(t_{ij}, \boldsymbol{x}_i(t_{ij}))}{S(t_{ij-1}, \boldsymbol{x}_i(t_{ij-1}))} \right] \frac{S(s_i, \boldsymbol{x}_i(s_i))}{S(t_{ij}, \boldsymbol{x}_i(t_{ij}))}$ that simplifies to

$S(s_i, \boldsymbol{x}_i(s_i)) \prod_{j=1}^{N_i} \lambda(t_{ij}, \boldsymbol{x}_i(t_{ij}))^{\delta_{ij}}$. This model employs a robust variance that goes by various names such as the Huber sandwich estimator. So, this is one way to adapt the recurrent events data into something akin to a single event survival analysis.

B.2 *Counting process Cox model of time stratified by prior events*

The counting process model of time stratified by prior events is a simple extension of the previous model. The only difference is that the baseline intensity varies by the number of prior events, i.e., redefine $\lambda(t, \boldsymbol{x}_i(t)) = \lambda_{0j}(t) \exp(\boldsymbol{\beta}' \boldsymbol{x}_i(t))$ where $j = N_i(t-)$ and $N_i(t-)$ is the counting process of prior events for subject i just prior to time t. This model employs a robust variance via the Huber sandwich estimator.

B.3 *Counting process Cox model of sojourn time stratified by prior events*

The counting process model of sojourn time stratified by prior events is similar to the previous model. The only difference is in how the baseline intensity is parameterized with respect to time. The baseline intensity is constructed as a function of the sojourn time rather than time, i.e., redefine $\lambda(t, \boldsymbol{x}_i(t)) = \lambda_{0j}(v_i(t)) \exp(\boldsymbol{\beta}' \boldsymbol{x}_i(t))$ where the sojourn time is $v_i(t) = t - t_{ij}$ and $j = N_i(t-)$. This model employs a robust variance via the Huber sandwich estimator.

B.4 *Marginal Cox model of time*

The marginal model is a departure from the previous Cox models. We represent the maximum number of events experienced by any subject as $\kappa = \max_i N_i$. We assume that every subject is followed from time zero and has $\kappa + 1$ observation periods as follows: an event for the each of intervals $(0, t_{i1}], \ldots, (0, t_{iN_i}]$ with respective strata $j = 0, \ldots, N_i - 1$; and $\kappa + 1 - N_i$ repeated non-events for the interval $(0, s_i]$ for strata $h = N_i, \ldots, \kappa$. This model employs a robust variance via the Huber sandwich estimator.

B.5 *Cox model summary*

By no means is this an exhaustive list of the types of Cox models that might be considered for recurrent events. In our view, this is one of the issues with using Cox models for recurrent events, i.e., how should we decide which model to use with real data?

C. Handling missing data with BART

BART can handle missing data (Kapelner and Bleich, 2016; Xu *and others*, 2016). We utilize the missing data framework developed by Xu *and others* (2016) which they call Sequential BART (coincidentally, they applied it to a study of hyperglycemia with EHR). Sequential BART assumes that the missing covariates are *missing at random*, i.e., missingness only depends on what has been observed. Specifically, Sequential BART assumes that a missing covariate can be imputed by BART from the rest of the covariates, and so on, sequentially for all missing covariates.

A brief description of this method follows where we assume that all missing covariates are continuous (which is adequate for this study, although, Sequential BART can be extended to binary and categorical covariates as well).

Suppose the covariates that are always observed for all subjects are denoted by $\boldsymbol{x}_i^{\text{obs}}$ and the covariates that may be missing by $\boldsymbol{x}_i^{\text{mis}} = (x_{i1}^{\text{mis}}, \ldots, x_{iK}^{\text{mis}})$ where x_{ik}^{mis} are ordered from the least missing overall, $k = 1$, to the most missing, $k = K$, for computational convenience. If for subject i the covariate x_{ik}^{mis} is missing, then its value can be imputed by Metropolis-Hastings sampling (Hastings, 1970) at the l^{th} step as follows.

$$x_{ik}^{*}|\boldsymbol{x}_i^{\text{obs}}, x_{i1}^{\text{mis}}, \ldots, x_{i(k-1)}^{\text{mis}}, f_k, \sigma_k^2 \sim \text{N}\Big(f_k(\boldsymbol{x}_i^{\text{obs}}, x_{i1}^{\text{mis}}, \ldots, x_{i(k-1)}^{\text{mis}}), \sigma_k^2\Big) \tag{C.1}$$

$$\text{where } \boldsymbol{\theta}_k = (f_k, \sigma_k^2) \sim \text{BART}(\psi, \mu, \kappa, \alpha, \beta; \nu, \lambda, q)$$

$$\alpha_{k(l)} = \frac{\Big[x_{i(k+1)}^{\text{mis}}|\boldsymbol{x}_i^{\text{obs}}, x_{i1}^{\text{mis}}, \ldots, x_{ik}^{*}, \boldsymbol{\theta}_{k+1}\Big] \cdots \Big[x_{iK}^{\text{mis}}|\boldsymbol{x}_i^{\text{obs}}, \boldsymbol{x}_{i(-k,-K)}^{\text{mis}}, x_{ik}^{*}, \boldsymbol{\theta}_K\Big]}{\Big[x_{i(k+1)}^{\text{mis}}|\boldsymbol{x}_i^{\text{obs}}, x_{i1}^{\text{mis}}, \ldots, x_{ik}^{(l-1)}, \boldsymbol{\theta}_{k+1}\Big] \cdots \Big[x_{iK}^{\text{mis}}|\boldsymbol{x}_i^{\text{obs}}, \boldsymbol{x}_{i(-k,-K)}^{\text{mis}}, x_{ik}^{(l-1)}, \boldsymbol{\theta}_K\Big]}$$

Sample x_{ik}^{*} from the proposal density, (C.1), and accept the proposal with the probability

$\min(\alpha_{k(l)}, 1)$. For more details, see (Xu *and others*, 2016).

D. Simulated data set scenarios

To demonstrate the effectiveness of BART in recurrent events, we have developed simulated data set scenarios of known provenance. To perform BART, we use the **BART** R package (McCulloch *and others*, 2018). For Cox models, we use the **survival** R package (Therneau, 2017). Based on these scenarios, we will evaluate BART's performance; and, in the proportional case, we will compare BART's performance to that of Cox models. Since our BART method is based on discrete-time survival analysis (Fahrmeir, 1998), these data set scenarios are also based on discrete-time survival analysis where tied event times are allowed. However, this creates a challenge to find a reasonable comparison.

In general, Cox models are based on continuous-time survival analysis that assume tied event times are non-existent. For discrete-time survival analysis where there are ties by design, the Cox partial likelihood is equivalent to matched logistic regression; however, for even the moderate sample sizes envisioned here, $N = 250$, this approach is computationally infeasible. Paraphrasing Therneau (2017) on the computational challenges: "Suppose 6 of 250 subjects had an event at month 9, then the calculation needs to compute sums over all $\binom{250}{6}$ possible subsets that number more than 300 billion! Although, there is an efficient recursive algorithm, with counting process data, it is much worse since the recursion needs to start anew for each unique interval start time."

Therefore, we are forced to abandon discrete-time Cox models for comparison. Instead, we will use continuous-time, counting-process Cox models and employ a tied event time correction due to Efron (1977) that is a compromise between the discrete-time and continuous-time partial likelihoods. Efron's method will be at a slight disadvantage because of a loss of efficiency due to tied discrete event times, but it is a reasonable compromise.

D.1 *Settings*

These scenarios bear some resemblance to the motivating data example. As mentioned above, we set the sample size at $N = 250$ that is roughly the size of the validation and training cohorts. There is no censoring due to death; all patients are followed for 60 30-day months and discrete event times correspond to this monthly grid. Many patients do not experience any events while some patients will be responsible for the majority of events. There are 20 baseline covariates (which are known at time zero), 10 binary and 10 continuous, but only one binary and one continuous covariate have any bearing on the outcome. In addition, the number of previous events, $N(t-)$, is an active covariate. We simulate 400 training data sets from both a proportional and a nonproportional setting; of course, the nonproportional setting depends on time as well. In addition, to the 400 in-sample training data sets, we simulate one out-of-sample validation data set to evaluate out-of-sample prediction based on all 400 trained fits. Performance is assessed based on estimates of the cumulative intensity function for the monthly grid in comparison to the known true values. These appear to be simple settings with only 3 or 4 active covariates. However, note that the cumulative intensity is substantially non-linear with respect to time in both the proportional and nonproportional settings; therefore, BART's nonparametric flexibility is essential to fitting the cumulative intensity via time and the covariates.

In both settings, we hand-picked moderate to strong coefficients to arrive at admission scenarios that are similar to our motivating example; especially, with respect to censoring: the subjects of the example experienced 63.0% whereas the proportional (nonproportional) censoring is 50.9% (65.2%) on average; see Table 2 for a comparison. Although, we did not pick these settings based on the subsequent results of the simulations themselves, we also did not explore a wider space of parameter settings. Therefore, this is a limitation of the simulation study.

D.1.1 *Proportional Setting* In the proportional setting, the relative intensity is constant with respect to time given the covariates, $\tilde{\boldsymbol{x}}(t_{(j)})$, which are constant in the time interval $(t_{(j-1)}, t_{(j)}]$. There are two factors that contribute to the probability of an event in a given time interval: the length of the interval and the strength of the intensity itself. As mentioned above, without loss of generality, we fix the length of the intervals in our time grid at 30 days. This leads us to the following proportional intensity for the Exponential distribution with rate $\alpha_P(\tilde{\boldsymbol{x}}(t_{(j)}))$ where the only active covariates are $N(t-),\ x_1$ and x_{11} among $\tilde{\boldsymbol{x}}(t_{(j)}) = \left[\sqrt{N(t_{(j-1)})}, v(t_{(j)}), x_1, \ldots, x_{20}\right]'$ (N.B. we have replaced $N(t-)$ with $\sqrt{N(t-)}$ so that the intensity grows more slowly).

$$\alpha_P(\tilde{\boldsymbol{x}}(t_{(j)})) = 0.0001 \exp\left(\tilde{\boldsymbol{x}}(t_{(j)})'\boldsymbol{\beta}\right)$$
$$\text{where } x_h \overset{\text{iid}}{\sim} \text{U}(0,1)\,, h = 1, \ldots, 10,\ x_l \overset{\text{iid}}{\sim} \text{B}(0.5)\,, l = 11, \ldots, 20$$
$$\text{and } \boldsymbol{\beta} = [1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1.5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]'$$

The intensity unit of time is days so the baseline intensity (where all covariates are set at zero) is 0.0001 that corresponds to one hospital admission in 27 years, i.e., a relatively low admission rate with only a 0.003 probability of experiencing a hospitalization within one month or 30 days. The relative intensities are moderate to strong as they are in our motivating example, i.e., $\text{e}^{1.5} = 4.48$ for x_1 and $\text{e}^1 = 2.72$ for x_{11} and $\sqrt{N(t-)}$. So, for each 30 day interval, the probability of an event is $p_j(\tilde{\boldsymbol{x}}(t_{(j)})) = \Pr\left(t = t_{(j)}\right) = \left[1 - \exp\left(-30\alpha_P(\tilde{\boldsymbol{x}}(t_{(j)}))\right)\right]$. Now, we can calculate the true cumulative intensity as follows.

$$\Lambda(t, \tilde{\boldsymbol{x}}(t)) = \sum_{j=1}^{k} w_j(t) p_j(\tilde{\boldsymbol{x}}(t_{(j)})) \text{ where } k = \arg\min_j (t \leqslant t_{(j)})$$
$$\text{and } w_j(t) = \frac{\min(t, t_{(j)}) - t_{(j-1)}}{t_{(j)} - t_{(j-1)}}$$

See Figure 1 where we display the cumulative intensity for a simulated data set and the corresponding estimates from our model.

D.1.2 *Nonproportional Setting* In the nonproportional setting, the relative intensity varies with respect to time, and, as before, the covariates, $\tilde{\boldsymbol{x}}(t_{(j)})$, are constant in the time interval $(t_{(j-1)}, t_{(j)}]$. As mentioned above, without loss of generality, we fix the length of the intervals in our time grid at 30 days. Therefore, we will simulate from the following nonproportional intensity for the Exponential distribution with rate $\alpha_N(t_{(j)}, \tilde{\boldsymbol{x}}(t_{(j)}))$.

$$\alpha_N(t_{(j)}, \tilde{\boldsymbol{x}}(t_{(j)})) = 0.0001 \exp\left(\tilde{\boldsymbol{x}}(t_{(j)})'\boldsymbol{\beta} \times 2\frac{N(t_{(j-1)}) + 1}{\sqrt{j}}\right)$$
$$\text{where } x_h \overset{\text{iid}}{\sim} \text{U}(0,1)\,, h = 1, \ldots, 10,\ x_l \overset{\text{iid}}{\sim} \text{B}(0.5)\,, l = 11, \ldots, 20$$
$$\text{and } \boldsymbol{\beta} = [0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1.5, 0, 0, 0, 0, 0, 0, 0, 0, 0]'$$

These settings lead to a more complex relationship than the proportional setting. Mainly, there is a departure from proportionality via the intensity's dependence on time through the $\sqrt{j}$ term, i.e., $\sqrt{t_{(j)}/30} = \sqrt{j}$. And, most importantly, the probability of an event for each 30 day interval is $p_j(\tilde{\boldsymbol{x}}(t_{(j)})) = \Pr\big(t = t_{(j)}\big) = \big[1 - \exp\big(-30\alpha_N(t_{(j)}, \tilde{\boldsymbol{x}}(t_{(j)}))\big)\big]$. So, we can calculate the true cumulative intensity as above. See Figure 2 where we display the cumulative intensity for a simulated data set and the corresponding estimates from our model.

D.2 *Comparisons*

We propose two main comparisons. First, we compare our new BART model to a Counting Process Cox (CPC) model with Efron's correction for ties via simulated data sets for the proportional setting. For the proportional setting, the CPC is the correct model since it assumes proportionality so this is a fair comparison; although, the CPC is at somewhat of a disadvantage because Efron's correction is not meant for discrete-time events. Furthermore, the CPC model receives the advantage of the true covariates that generated the data: $\big(\sqrt{N(t_{(j-1)})}, x_1, \ldots, x_{20}\big)$. Meanwhile, the BART model receives no such favoritism; rather, it is provided the covariates per the conditional independence assumption, i.e., $\big(t_{(j)}, N(t_{(j-1)}), v(t_{(j)}), x_1, \ldots, x_{20}\big)$. This puts

BART at the added disadvantage of two extra noise variables, $t_{(j)}$ and $v(t_{(j)})$, as well as having to discern the correct functional form of $N(t_{(j-1)})$. Also, we use BART with its default prior settings, i.e., no attempt is made to pick optimal settings via cross-validation. This first comparison is restricted to in-sample performance. Second, we compare the BART model's in-sample and out-of-sample performance for both the proportional and nonproportional settings. Since BART is in the class of ensemble predictive models, theoretically, BART's performance on in-sample vs. out-of-sample predictions should be quite similar.

The focus of these comparisons is the cumulative intensity that is central to the recurrent events framework. For each method, we estimate the cumulative intensity and compare it to the known true cumulative intensity. Our metrics of choice are root mean square error (RMSE), bias and 95% interval coverage. We also measure the 95% interval length, but it is mainly descriptive since a shorter or longer interval is not necessarily of interest except under equivalent coverage. N.B. technically, comparing the frequentist 95% confidence intervals from the CPC model with the Bayesian 95% credible intervals from the BART model is not an apples to apples comparison. However, as a practical matter, these comparisons are often done, and, we believe, they provide useful insights into the two methods. Furthermore, with respect to interval coverage, this comparison is warranted.

Since the cumulative intensity is a function of time, for each subject, we compare it at a set of discrete-time grid points, i.e., a monthly grid of 60 months each 30 days apart for $250 \times 60 = 15000$ values. Note that the variance of the cumulative intensity increases with time. This presents a challenge to summarize our findings since the scale at month 12 can be vastly different than at month 48; similarly, at 12 months, a low risk vs. a high risk subject can have very different profiles. So, we divide the results into 6 realms per the quantiles of the true cumulative intensity: [0.00, 0.10); [0.10, 0.25); [0.25, 0.50); [0.50, 0.75); [0.75, 0.90) and [0.90, 1.00]. And, to make comparisons of bias between realms, we also present the bias divided by the corresponding RMSE.

D.3 *Results of comparisons*

Here, we provide a brief summary. We examined single data sets seeking convergence with BART. Based on these diagnostics (further described in Section E), the thinning parameter is set accordingly for all data sets. In the proportional setting, BART and CPC performance are generally consistent with a few exceptions in BART's favor; the main point is that BART's 95% interval coverage attains nominal levels, see Table 3 for a numerical summary. We summarize the performance via graphical summaries: RMSE in Figure 3, bias in Figure 4, bias/RMSE in Figure 5, interval coverage in Figure 6 and interval length in Figure 7. Also in the proportional setting, BART's in-sample vs. out-of-sample performance was comparable; the main point being that BART's 95% interval coverage for both attains nominal levels, see Table 3 for a numerical summary. The graphical summaries are provided below: RMSE in Figure 8, bias in Figure 9, bias/RMSE in Figure 10, interval coverage in Figure 11 and interval length in Figure 13. In the nonproportional setting, BART's in-sample vs. out-of-sample performance was comparable with a slight edge to in-sample that is understandable; the main point is that BART's 95% interval coverage for both attains nominal levels with the exception of the last realm, see Table 3 for a numerical summary. The graphical summaries are provided below: RMSE in Figure 15, bias in Figure 16, bias/RMSE in Figure 17, interval coverage in Figure 18 and interval length in Figure 19.

E. The **BART** R package

Along with our partners in BART computing, we have created an R package called **BART** for continuous, dichotomous, categorical and time-to-event outcomes including survival analysis, competing risks and recurrent events. **BART** is open source, free software now available from the Comprehensive R Archive Network (CRAN) at `https://cran.r-project.org/package=BART` (McCulloch *and others*, 2018). However, due to CRAN policy limiting the file archive size, the

BART package only contains a roughly 20% random sample of our motivating example data: 50 patients from the training set and 50 patients form the validation set. The complete data set can be found online at
http://www.mcw.edu/FileLibrary/Groups/Biostatistics/TechReports/TechReports5175/tr064.tar .
To install the **BART** package, follow these steps.

```
> options(repos=c(CRAN="https://cran.r-project.org"))
> install.packages("BART", dependencies=TRUE) ## depends on the Rcpp package
```

The **BART** package contains several examples that are useful in understanding how BART and BART with recurrent events works. These examples can be found in the **BART** package `demo` directory and you can locate the recurrent events examples with this snippet of R code.

```
> ## Data construction for recurrent events with BART
> system.file('demo/data.recur.pre.bart.R', package='BART')
> ## Proportional intensity for recurrent events with BART
> system.file('demo/exp.recur.bart.R', package='BART')
> ## Nonproportional intensity for recurrent events with BART
> system.file('demo/np.recur.bart.R', package='BART')
> ## Interval coverage calibration for recurrent events with BART
> system.file('demo/cal.recur.bart.R', package='BART')
> ## Geweke diagnostics for recurrent events with BART
> system.file('demo/geweke.recur.bart.R', package='BART')
> ## Diabetes and hospital admission example
> system.file('demo/dm.recur.bart.R', package='BART')
> ## Bladder cancer example for recurrent events with BART
> system.file('demo/bladder.recur.bart.R', package='BART')
```

As described, in Section 2.1, BART for dichotomous outcomes relies on either the probit BART model with Normal latents (Albert and Chib, 1993; Robert, 1995); or the logistic BART model with Logistic latents (Holmes and Held, 2006; Gramacy and Polson, 2012). BART for time-to-event outcomes takes the discrete-time approach and, therefore, recasts the problem as dichotomous outcomes. By default, BART for recurrent events utilizes Normal latents for computational efficiency. However, BART with Normal latents may have more difficulty than Logistic latents in estimating probabilities vary close to zero or one since the Normal distribution has relatively thinner tails. Therefore, BART with Logistic latents is available as an option by specifying `type='lbart'`.

Whether you are using Normal or Logistic latents, you may have a data set where the calculations involved may be time-consuming especially large data sets (such as the motivating example complete data set). Therefore, we provide both serial and parallel versions of some functions that utilize the **parallel** R package function `mcparallel`. However, note that the `mcparallel` function depends on operating system support for forking to perform parallel processing, e.g., this support is available on macOS and UNIX/Linux, but not available on Windows. The serial (parallel) version of the function for BART with recurrent events is `recur.bart` (`mc.recur.bart`).

As mentioned in Section 2.1, BART is a Bayesian nonparametric method that relies on MCMC to generate samples of f from the posterior. BART with recurrent events often requires large data sets that can present challenges for convergence. The convergence diagnostics example explores the first data set in the simulation study to determine the settings necessary to achieve convergence. But, how do you perform convergence diagnostics for BART? For continuous outcomes, convergence can easily be determined from the trace plots of the the error variance, σ^2. However, for probit BART with Normal latents, the error variance is fixed at 1 so this is not an option. Therefore, we adapt traditional MCMC diagnostic approaches to BART. We perform graphical checks via auto-correlation, trace plots and an approach due to Geweke (1992).

Geweke diagnostics is based on earlier work that characterizes MCMC as a time series (Hastings, 1970). Once this transition is made, auto-regressive, moving-average (ARMA) process theory is employed (Silverman, 1986). Generally, we define our Bayesian estimator as $\hat{\theta}_M = M^{-1}\sum_{m=1}^{M}\theta_m$. We represent the asymptotic variance of the estimator by $\sigma^2_{\hat{\theta}} = \lim_{M\to\infty}\mathrm{Var}\Big(\hat{\theta}_M\Big)$. If we suppose that θ_m is an ARMA(p,q) process, then the spectral density of the estimator is defined as $\gamma(w) = (2\pi)^{-1}\sum_{m=-\infty}^{\infty}\mathrm{Var}(\theta_0,\theta_m)\,\mathrm{e}^{\mathrm{i}mw}$ where $\mathrm{e}^{\mathrm{i}tw} = \cos(tw) + \mathrm{i}\sin(tw)$. This leads us to an estimator of the asymptotic variance that is $\hat{\sigma}^2_{\hat{\theta}} = \hat{\gamma}^2(0)$. We divide our chain into two segments, A and B, as follows: $m \in A = \{1,\ldots,M_A\}$ where $M_A = aM$; and $m \in B = \{M - M_B + 1,\ldots,M\}$ where $M_B = bM$. Note that $a + b < 1$. Geweke suggests $a = 0.1,\ b = 0.5$ and recommends the following Normal test for convergence.

$$\hat{\theta}_A = M_A^{-1}\sum_{m\in A}\theta_m \qquad\qquad \hat{\theta}_B = M_B^{-1}\sum_{m\in B}\theta_m$$

$$\hat{\sigma}^2_{\hat{\theta}_A} = \hat{\gamma}^2_{m\in A}(0) \qquad\qquad \hat{\sigma}^2_{\hat{\theta}_B} = \hat{\gamma}^2_{m\in B}(0)$$

$$Z_{AB} = \frac{\sqrt{M}(\hat{\theta}_A - \hat{\theta}_B)}{\sqrt{a^{-1}\hat{\sigma}^2_{\hat{\theta}_A} + b^{-1}\hat{\sigma}^2_{\hat{\theta}_B}}} \qquad\qquad \sim \mathrm{N}(0,1)$$

In our **BART** package, we supply R functions adapted from the **coda** R package (Plummer *and others*, 2006) to perform Geweke diagnostics: `spectrum0ar` and `gewekediag`. But, how do we apply Geweke's diagnostic to BART? We can check convergence for any estimator of the form $\theta = h(f(\boldsymbol{x}))$, but often setting h to the identify function will suffice, i.e., $\theta = f(\boldsymbol{x})$. However, BART being a Bayesian nonparametric technique means that we have many potential estimators to check, i.e., one estimator for every possible choice of $\boldsymbol{x}$.

We have supplied Figure 20 generated by the example `geweke.recur.bart.R` for the first data set from the proportional setting of the simulation study. Based on reviewing figures like these, we chose a thinning parameter of 100 that is depicted here. In the upper left quadrant, we

have plotted Friedman's partial dependence function for $f(x_1)$ vs. x_1 for 10 values of x_1. This is a check that can't be performed for real data, but it is informative in this case. Notice that $f(x_1)$ vs. x_1 is directly proportional as expected. In the upper right quadrant, we plot the auto-correlations of $f(t_{(j)}, \boldsymbol{x}_i)$ for 10 randomly selected $t_{(j)}$ and $\boldsymbol{x}_i$ combinations where i (j) indexes subjects (time points). Notice that there is a combination that has fairly high auto-correlation, but the rest are quite reasonable. In the lower left quadrant, we display the corresponding trace plots for these same combinations. The traces demonstrate that samples of $f(t_{(j)}, \boldsymbol{x}_i)$ appear to adequately traverse the sample space. In the lower right quadrant, we have selected 10 subjects and we plot their corresponding Geweke Z_{AB} statistics over the time points. Notice that only 2 or 3 subjects ever reach the 95% boundaries and only rarely; given the number of comparisons, 600, this seems reasonable as well.

Now, we explore this single data set with respect to coverage calibration. You can find this example in the file `exp.recur.bart.R`. As we have seen in the simulation study, often the out-of-sample interval coverage is slightly higher than the nominal level; and, furthermore, these interval lengths are wider than the corresponding in-sample intervals. With real data, we can perform out-of-sample interval calibration via cross-validation. In this case, we use five-fold cross-validation, i.e., divide our data set into five roughly equal blocks. Then, perform five fits each time holding out one of the blocks for the out-of-sample validation. Based on these five fits, determine the equal-tail quantiles required to arrive at a 1-α level credible interval. With this data set, we determine that a roughly 95% out-of-sample coverage is obtained via a 90% interval constructed from the 5% and 95% quantiles (rather than 2.5% and 97.5% for a 95% interval). You can compare these two settings with respect to interval coverage in Figures 11 and 12; and interval length in Figures 13 and 14.

Let's return to the vignette (see Figure 22 and Section 2.2, display (2.3) which is copied below)

that we re-iterate for convenience. Suppose that we have two subjects with the following values:

$$N_1 = 2, s_1 = 9, t_{11} = 3, u_{11} = 7, t_{12} = 8, u_{12} = 8 \Rightarrow y_{11} = 1, y_{12} = y_{13} = 0, y_{14} = 1, y_{15} = 0 \quad (2.3)$$

$$N_2 = 1, s_2 = 12, t_{21} = 4, u_{21} = 7 \Rightarrow y_{21} = 0, y_{22} = 1, y_{23} = y_{24} = y_{25} = y_{26} = 0$$

which creates the grid of times $(3, 4, 7, 8, 9, 12)$. For subject 1 (2), notice that $y_{12} = y_{13} = 0$ ($y_{23} = 0$) as it should be since no event occurred at times 4 or 7 (7). However, no events could occur since their first event had not ended yet, i.e., these time points do not contribute to the likelihood since these subjects are not currently at risk for an event. The **BART** package provides the `recur.pre.bart` function that you can use to construct the corresponding data set. Here is a short demonstration of its capabilities applied to the vignette data (adapted from `data.recur.pre.bart.R`).

```
> library(BART)
> times <- matrix(c(3, 8, 9, 4, 12, 12), nrow=2, ncol=3, byrow=TRUE)
> tstop <- matrix(c(7, 8, 0, 7,  0,  0), nrow=2, ncol=3, byrow=TRUE)
> delta <- matrix(c(1, 1, 0, 1,  0,  0), nrow=2, ncol=3, byrow=TRUE)
> recur.pre.bart(times=times, delta=delta, tstop=tstop)
$y.train                $tx.train     $tx.test
[1] 1 1 0 0 1 0 0 0           t v N         t v N
$times                  [1,]  3 3 0   [1,]  3 3 0
[1]  3  4  7  8  9 12   [2,]  8 5 1   [2,]  4 1 1
$K                      [3,]  9 1 2   [3,]  7 4 1
[1] 6                   [4,]  3 3 0   [4,]  8 5 1
                        [5,]  4 4 0   [5,]  9 1 2
                        [6,]  8 4 1   [6,] 12 4 2
                        [7,]  9 5 1   [7,]  3 3 0
```

```
[8,] 12 8 1   [8,]  4 4 0
              [9,]  7 3 1
             [10,]  8 4 1
             [11,]  9 5 1
             [12,] 12 8 1
```

Notice that `$tx.test` is not limited to the same time points as `$tx.train`, i.e., we often want/need to estimate f at counter-factual values not observed in the data.

F. Recurrent events, dependent censoring, competing risks and BART

As has been described, dealing with dependent censoring in the recurrent events framework is challenging (Cook and Lawless, 1997; Ghosh and Lin, 2000; Wang *and others*, 2001; Ghosh and Lin, 2003). Our approach is to combine the recurrent events and competing risks paradigms. Typically, competing risks (Fine and Gray, 1999; Kalbfleisch and Prentice, 2002) deal with events that are mutually exclusive, say, death from cardiovascular disease vs. death from other causes, i.e., a patient experiencing one of the events is prevented from experiencing another. Our application is slightly different in that we have two events, death and hospital admission, which are not mutually exclusive. Suffering death prevents a patient from experiencing a future hospital admission, but the converse is not true, i.e., a hospital admission does not prevent a future death. So, technically, we have what are termed semi-competing events, yet the competing events framework is sufficient for our needs. For the simplicity of this exposition, we assume that the hospital admissions are instantaneous, i.e., hospital stays are of length zero.

We create a single grid of time points for the ordered distinct times based on either type of event or censoring. To accommodate competing risks, we adapt our notation slightly: (s_i, δ_i) are death times, $\delta_i = 1$, or censoring times, $\delta_i = 0$. We model the probability of event 1 that is death: $p_1(t_{(j)}, \boldsymbol{x}_i)$. Next, we model event 2, hospital admission, which is necessarily conditioned

on patient i being alive at time $t_{(j)}$: $p_2(t_{(j)}, \boldsymbol{x}_i)$. Now, we can estimate the survival function and the cumulative incidence functions as follows.

$$S(t, \boldsymbol{x}_i) = 1 - F(t, \boldsymbol{x}_i) = \prod_{j=1}^{k} (1 - p_1(t_{(j)}, \boldsymbol{x}_i))(1 - p_2(t_{(j)}, \boldsymbol{x}_i)) \text{ where } k = \arg\max_j \left[t_{(j)} \leqslant t\right]$$

$$F_1(t, \boldsymbol{x}_i) = \int_0^t S(u-, \boldsymbol{x}_i)\lambda_1(u, \boldsymbol{x}_i)\mathrm{d}u = \sum_{j=1}^{k} S(t_{(j-1)}, \boldsymbol{x}_i)p_1(t_{(j)}, \boldsymbol{x}_i)$$

$$F_2(t, \boldsymbol{x}_i) = \int_0^t S(u-, \boldsymbol{x}_i)\lambda_2(u, \boldsymbol{x}_i)\mathrm{d}u = \sum_{j=1}^{k} S(t_{(j-1)}, \boldsymbol{x}_i)(1 - p_1(t_{(j)}, \boldsymbol{x}_i))p_2(t_{(j)}, \boldsymbol{x}_i)$$

G. The motivating example and the EHR

In this section, additional details of the motivating example are provided related to the source of data from the EHR. Specifically, we discuss anti-diabetic therapy; health care charges and relative value units (RVU); handling of missing data; and the coded conditions considered that are either comorbidities or procedures/surgeries.

For insulin, metformin and sulfonylurea, we only had access to prescription orders (rather than prescription fills) and self-reported current status of prescription therapy during clinic office visits. Generally, orders are only required after every three fills, and each fill can be for up to 90 days, so we define insulin, metformin and sulfonylurea as binary indicators that are one if there exists an order or current status indication within the prior 270 days; otherwise zero.

Health care charges and relative value units (RVU) are measures related to the services and procedures delivered. RVUs (Federal Register, 2010) are dictated by the US Medicare national health insurance program for reimbursement purposes. An RVU represents the relative clinical input (time, intensity, training, etc.) necessary to provide a given service; a service with a higher RVU is reimbursed at a higher rate. Since we are interested in preventive opportunities, very recent charges/RVUs are too closely related to services and procedures to be practically useful.

Therefore, we investigate chronologically distant previous charges/RVUs that are the sum total of the following moving windows of days prior to any given date: 31 to 90, 91 to 180, 181 to 300.

For some patients, their signs were not available on a given date so they were set to missing; similarly, if a sign was not observed within the last 180 days, then it was set to missing (except height never expires, weight extended to 365 days and body mass index is a deterministic function of the two). We used the Sequential BART missing imputation method as described in Appendix C of the Supplement. However, instead of creating several imputed data sets, we imputed a new sign at each date when it was missing, i.e., in order to address uncertainty with one data set, a new value was imputed for each date that it was missing and never carried forward.

Conditions are binary indicators that are zero until the date of the first coding and then they are one from then on (see Table 1 of the Supplement for the codes utilized). Based on clinical rationale, we identified 26 conditions (23 comorbidities and 3 procedures/surgeries) that are potential risk factors for a hospital admission and/or possible complications of diabetes; besides clinical merit, these conditions are chosen since they are present in more than just a few subjects so that they may be informative. Similarly, we employed 15 general conditions known as the Charlson diagnoses (Charlson *and others*, 1987; Quan *and others*, 2005) and 18 general conditions from the RxRisk adult diagnoses defined by prescription orders (Fishman *and others*, 2003; Johnson *and others*, 2006). Seven conditions are a composite of diagnosis codes and prescription orders: these codes are denoted by an asterisk in Table 1 of the Supplement. Notice that the conditions are not independent; for example, renal disease is a superset of chronic kidney disease that is a superset of kidney failure. These hierarchical definitions are intentional since we would like to identify the narrowest risk factor definition wherever possible. And, the following conditions are mutually exclusive so necessarily dependent: mild liver disease vs. moderate/severe liver disease; and malignancy vs. metastatic solid tumor.

Supplementary References

Albert, J and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88**, 669–79.

Charlson, ME, Pompei, P, Ales, KL and MacKenzie, CR. (1987). A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of Chronic Diseases* **40**, 373–83.

Cook, RJ and Lawless, JF. (1997). Marginal analysis of recurrent events and a terminating event. *Statistics in Medicine* **16**, 911–24.

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association* **72**, 557–65.

Fahrmeir, L. (1998). Discrete survival-time models. In: *Encyclopedia of biostatistics*. Chichester: Wiley, pp. 1163–1168.

Federal Register. (2010, November). Medicare program: payment policies under the physician fee schedule and other revisions to Part B for CY 2011. US Government Publishing Office. [https://www.gpo.gov/fdsys/pkg/FR-2010-11-29/pdf/2010-27969.pdf].

Fine, Jason P and Gray, Robert J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association* **94**, 496–509.

Fishman, PA, Goodman, MJ, Hornbrook, MC, Meenan, RT, Bachman, DJ and O'Keefe Rosetti, MC. (2003). Risk adjustment using automated ambulatory pharmacy data: the RxRisk model. *Medical Care* **41**, 84–99.

Geweke, J. (1992). *Bayesian Statistics*, fourth edition., Chapter Evaluating the accuracy of sampling-based approaches to calculating posterior moments. Oxford, UK: Clarendon Press.

Ghosh, D and Lin, DY. (2000). Nonparametric analysis of recurrent events and death. *Biometrics* **56**, 554–62.

Ghosh, D and Lin, DY. (2003). Semiparametric analysis of recurrent events data in the presence of dependent censoring. *Biometrics* **59**, 877–85.

Gramacy, Robert B and Polson, Nicholas G. (2012). Simulation-based regularized logistic regression. *Bayesian Analysis* **7**(3), 567–590.

Hastings, WK. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

Holmes, C and Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis* **1**, 145–68.

Hosmer Jr, DW, Lemeshow, S and May, S. (2008). *Applied survival analysis: regression modeling of time to event data*, second edition. Chichester: Wiley.

Johnson, ML, El-Serag, HB, Tran, TT, Hartman, C, Richardson, P and Abraham, NS. (2006). Adapting the Rx-Risk-V for mortality prediction in outpatient populations. *Medical Care* **44**, 793–7.

Kalbfleisch, JD and Prentice, RL. (2002). *The statistical analysis of failure time data*, second edition. Chichester: Wiley.

Kapelner, A and Bleich, J. (2016). **bartMachine**: machine learning with Bayesian Additive Regression Trees. *Journal of Statistical Software* **70**, 1–40.

McCulloch, RE, Sparapani, RA, Gramacy, R, Spanbauer, C and Pratola, M. (2018). BART: Bayesian Additive Regression Trees. [https://cran.r-project.org/package=BART].

Plummer, Martyn, Best, Nicky, Cowles, Kate and Vines, Karen. (2006). CODA: Convergence Diagnosis and Output Analysis for MCMC. *R News* **6**(1), 7–11. [https://journal.r-project.org/archive].

Quan, H, Sundararajan, V, Halfon, P, Fong, A, Burnand, B, Luthi, JC, Saunders, LD, Beck, CA, Feasby, TE and Ghali, WA. (2005). Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Medical Care* **43**, 1130–9.

Robert, Christian P. (1995). Simulation of truncated normal variables. *Statistics and computing* **5**(2), 121–125.

Silverman, BW. (1986). *Density estimation for statistics and data analysis*. London: Chapman and Hall.

Therneau, TM. (2017). A package for survival analysis in S. [https://cran.r-project.org/web/packages/survival].

Wang, MC, Qin, J and Chiang, CT. (2001). Analyzing recurrent event data with informative censoring. *Journal of the American Statistical Association* **96**, 1057–65.

Xu, D, Daniels, MJ and Winterstein, AG. (2016). Sequential BART for imputation of missing covariates. *Biostatistics* **17**, 589–602.

Table 1. *Codes for conditions (* diagnostic codes and prescription orders combined)*

Category	Conditions (ICD-9-CM/HCPCS Code(s))
Acute metabolic	Hyperinsulinism (962.3), Hypoglycemia (250.8x), Ketoacidosis (250.1x)
Circulatory	Atrial fibrillation (427.31), Cardiomyopathy (425.4), Coronary artery disease (414.0x), Gangrene (785.4), Hypertension* (401.x)
Eye	Blindness (369.xx), Retinopathy (362.0x)
Kidney	Chronic kidney disease (585.x, E11.22), Dialysis (V45.1, V56.x, 90935:90937), Kidney failure (585.5, 585.6), Kidney transplant (V42.0, 50360, 50365), Nephropathy (583.81)
Neurologic	Depression* (300.4, 311), Diabetic foot (713.5), Encephalopathy (348.30), Neuropathy (250.6x, 354.x, 355.x, 337.1, 353.5, 357.2, E11.40)
Procedure/Surgery	Bariatric surgery (V45.86, 43775, 46344, 43846), CABG (V45.81, 33503:33505, 33510:33516, 4110F), PTCA (V45.82, 92982:92984, 92920, 92921)
Other	Diabetic bone changes (731.8), Diabetic ulceration (707.1x, 707.8, 707.9), Medical nutrition therapy (97802, 97803), Sleep apnea (327.2x, 770.81, 770.82, 780.51, 780.53, 780.57, 786.03)
Charlson Diagnosis Conditions	**ICD-9-CM Code(s)**
Cerebrovascular disease	362.34, 430.xx:438.xx
Chronic pulmonary disease	416.8, 416.9, 490.x:496.x, 500.x:505.x, 506.4, 508.1, 508.8
Congestive heart failure	398.91, 402.01, 402.11, 402.91, 404.01, 404.03, 404.11, 404.13, 404.91, 404.93, 425.4x, 425.5x, 425.7x:425.9x, 428.xx
Dementia	290.xx, 294.1x, 331.2x
Diabetic chronic complications	250.4x:250.6x
Hemiplegia or paraplegia	334.1, 342.x, 343.x, 344.0:344.6, 344.9
Malignancy*	140.x:172.x, 174.x:195.8, 200.x:208.x, 238.6
Metastatic solid tumor	196.xx:199.xx
Mild liver disease	070.22, 070.23, 070.32, 070.33, 070.44, 070.54, 070.6, 070.9, 570.x, 571.x, 573.3, 573.4, 573.8, 573.9, V42.7
Moderate/severe liver disease	456.0:456.2, 572.2:572.8
Myocardial infarction	410.xx, 412
Peptic ulcer disease*	531.x:534.x
Peripheral vascular disease*	093.0x, 437.3x, 440.xx, 441.xx, 443.1x:443.9x, 447.1x, 557.1x, 557.9x, V43.4
Renal disease*	403.01, 403.11, 403.91, 404.02, 404.03, 404.12, 404.13, 404.92, 404.93, 582.x, 583.0:583.7, 585.x, 586.x, 588.0, V42.0, V45.1, V56.x
Rheumatic disease*	446.5, 710.0:710.4, 714.0:714.2, 714.8, 725.x
Adult RxRisk	**Conditions (Medi-Span pharmaceutical class)**
	Anxiety and tension (57), Asthma (44), Cardiac disease (35), Peripheral vascular disease* (83), Depression* (58), Epilepsy (72), Peptic ulcer disease* (49), Gout (68), Heart disease (32, 33, 34), Hyperlipidemia (39), Hypertension* (36), Malignancy* (21, 50), Parkinson's disease (73), Psychosis (59), Renal disease* (82), Rheumatic disease* (66), Thyroid disorder (28), Tuberculosis (9)

CABG: Coronary Artery Bypass Grafting; HCPCS: Healthcare Common Procedure Coding System; ICD-9-CM: International Classification of Diseases, Ninth Revision, Clinical Modification; PTCA: Percutaneous Transluminal Coronary Angioplasty.

We identify a cohort of patients suffering diabetes via their collected Electronic Health Records (EHR) to determine which covariates are related to the risk of hospital admission. Herein, we provide the codes utilized to define covariates related to comorbidities and procedures/surgeries.

Table 2. *Actual values compared to simulation settings summarized over 400 data sets*

Hospital Admissions	Cohort %	Proportional %	Nonproportional %
0	63.0	50.9	65.2
1	16.2	15.5	18.4
2-3	10.3	13.4	5.8
4-6	7.6	7.9	0.9
7+	2.9	12.3	9.7

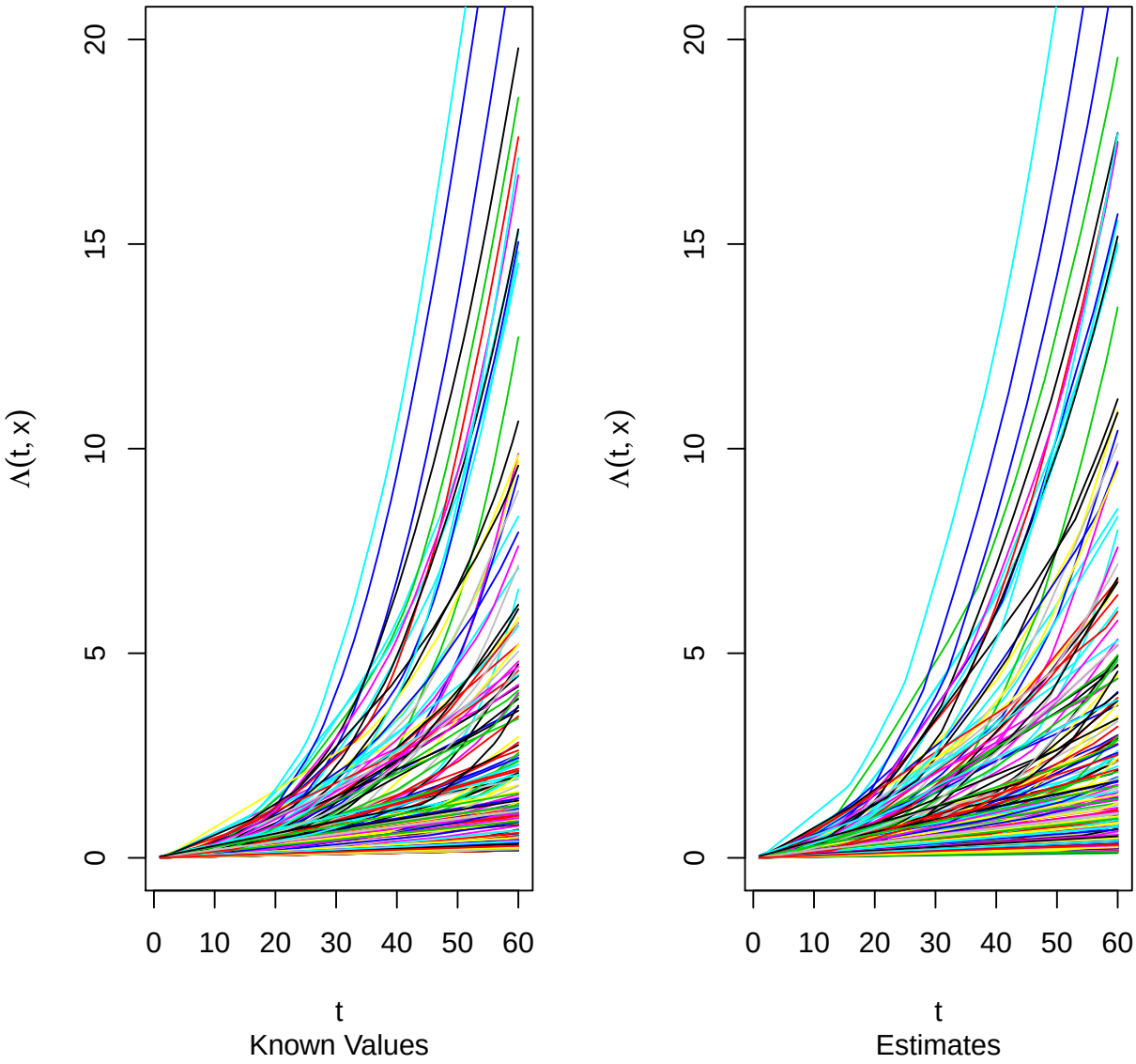
We identify a cohort of patients suffering diabetes via their collected Electronic Health Records (EHR) to determine which covariates are related to the risk of hospital admission. Based on this cohort, we have developed a study of simulated data sets where the hospital admission profile bears a resemblance to the cohort. These simulated data sets are constructed under two scenarios: a proportional setting and a nonproportional setting. Herein, we summarize hospital admissions in the cohort as well as in the proportional and the nonproportional settings.

Table 3. *Summary of 95% Interval Coverage*

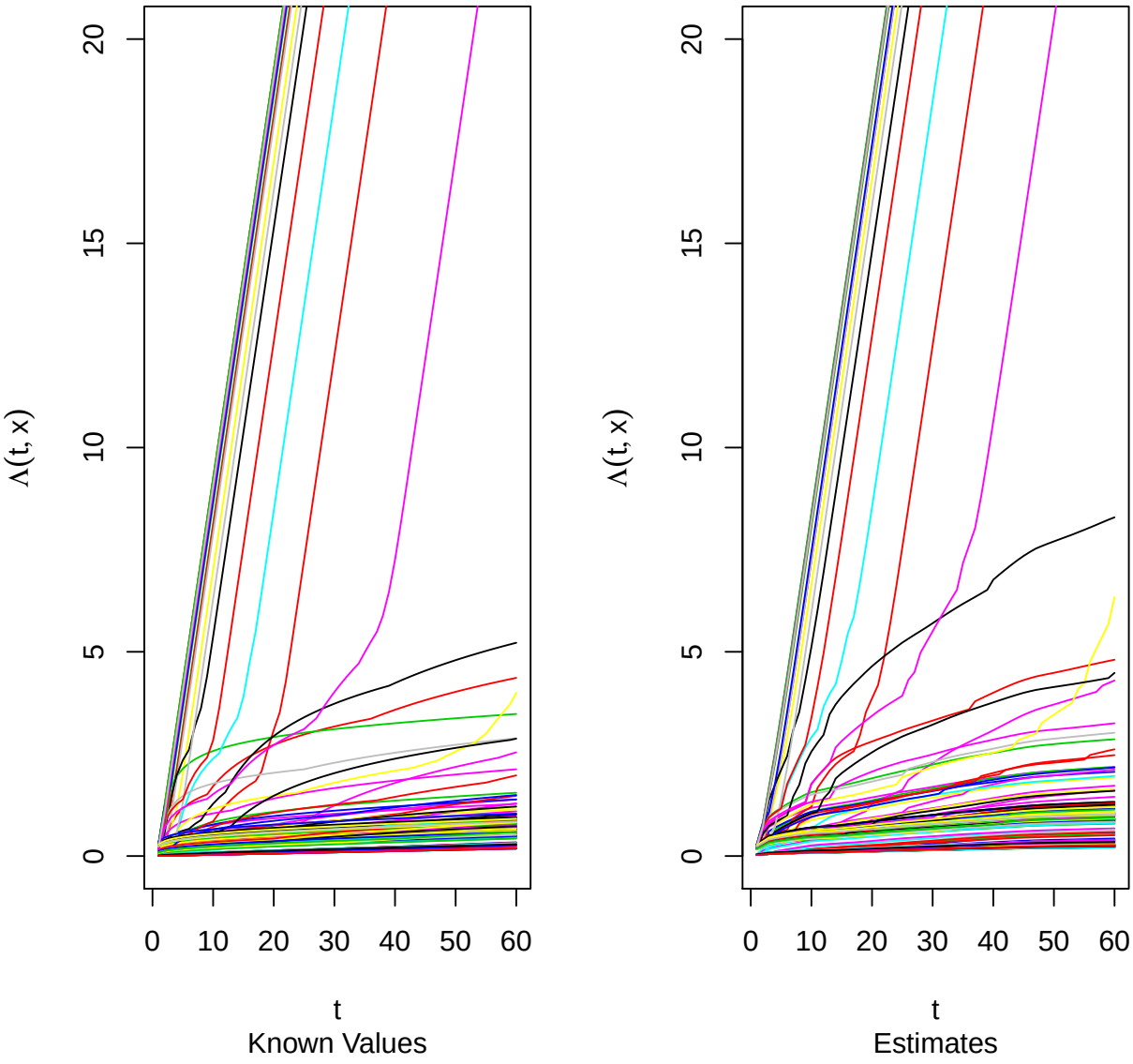
Method	CPC	BART				
Setting	P	P			NP	
Prediction	I 95%	I 95%	O 95%	O 90%	I 95%	O 95%
Overall	0.714	0.965	0.978	0.944	0.886	0.911
a	0.837	0.962	0.979	0.946	0.948	0.904
b	0.855	0.966	0.977	0.944	0.973	0.911
c	0.876	0.972	0.978	0.944	0.974	0.913
d	0.837	0.970	0.978	0.943	0.962	0.913
e	0.624	0.964	0.978	0.943	0.939	0.915
f	0.258	0.958	0.978	0.942	0.516	0.911

CPC:Counting Process Cox model.
P:Proportional, NP:Nonproportional.
I:In-sample, O:Out-of-sample.
a:[0.00, 0.10); b:[0.10, 0.25); c:[0.25, 0.50);
d:[0.50, 0.75); e:[0.75, 0.90) and f:[0.90, 1.00].
a-f are realms for the quantiles of the true cumulative intensity.

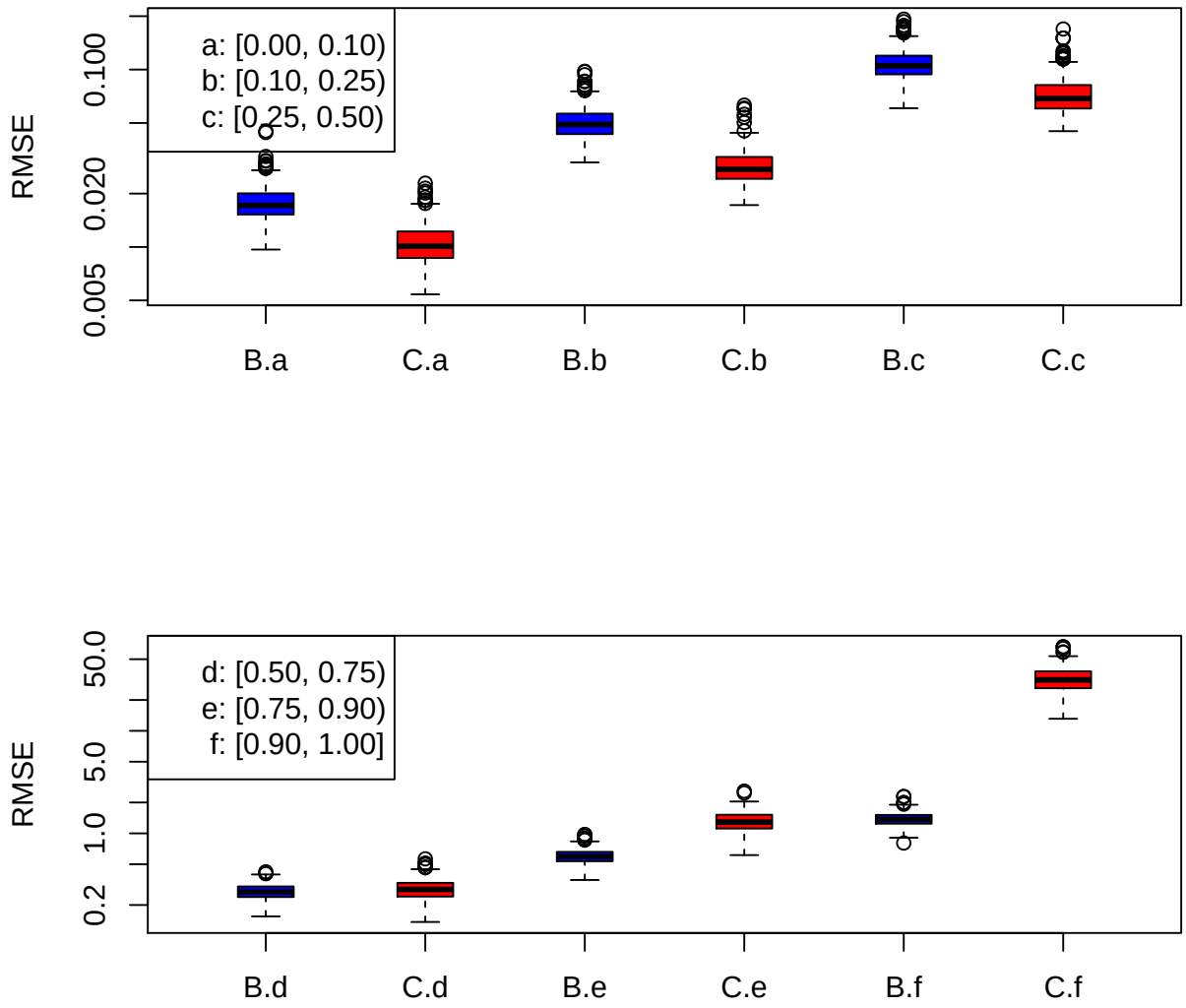
We identify a cohort of patients suffering diabetes via their collected Electronic Health Records (EHR) to determine which covariates are related to the risk of hospital admission. Based on this cohort, we have developed a study of simulated data sets where the hospital admission profile bears a resemblance to the cohort. These simulated data sets are constructed under two scenarios: a proportional setting and a nonproportional setting. We analyzed the proportional setting data sets with counting process Cox models and our recurrent events BART model. However, discrete-time tied events are incompletely controlled for diminishing the efficiency of the Cox model (given the sample sizes considered, an adequate treatment is computationally infeasible). Therefore, we only provide in-sample results for the Cox model while we provide both in-sample and out-of-sample for BART. Furthermore, we analyzed the nonproportional setting data sets only with BART since Cox assumes proportionality and, given its deficiency with discrete-time tied events, this comparison is unnecessary.



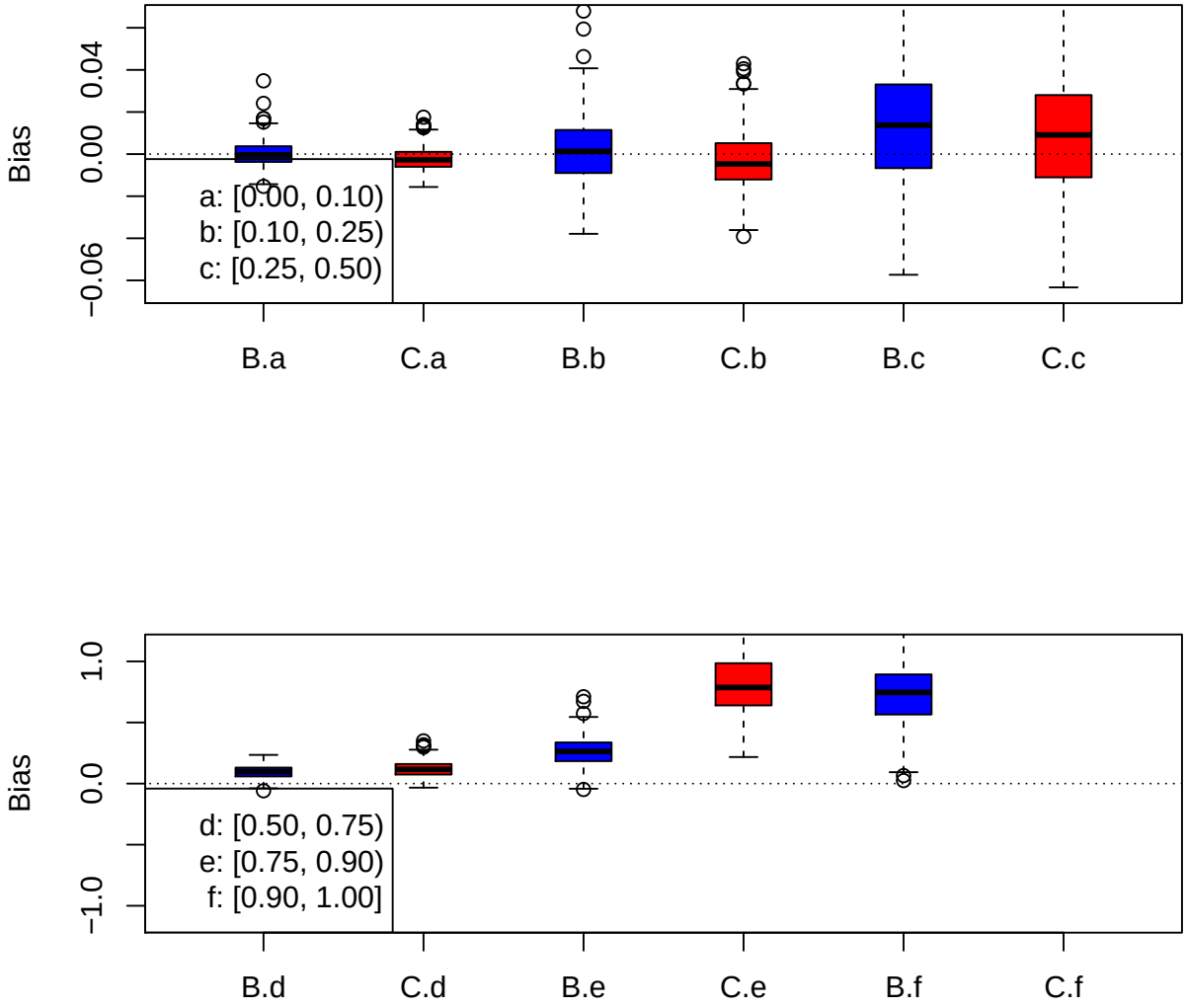
Supplementary Figure 1. Proportional setting: we simulated a data set conforming to proportionality of the covariates and then we estimated the corresponding cumulative intensity by recurrent events with BART. In the left (right) figure, we display the known values (estimates) of the cumulative intensity. The R^2 between known values and estimates is 0.984.



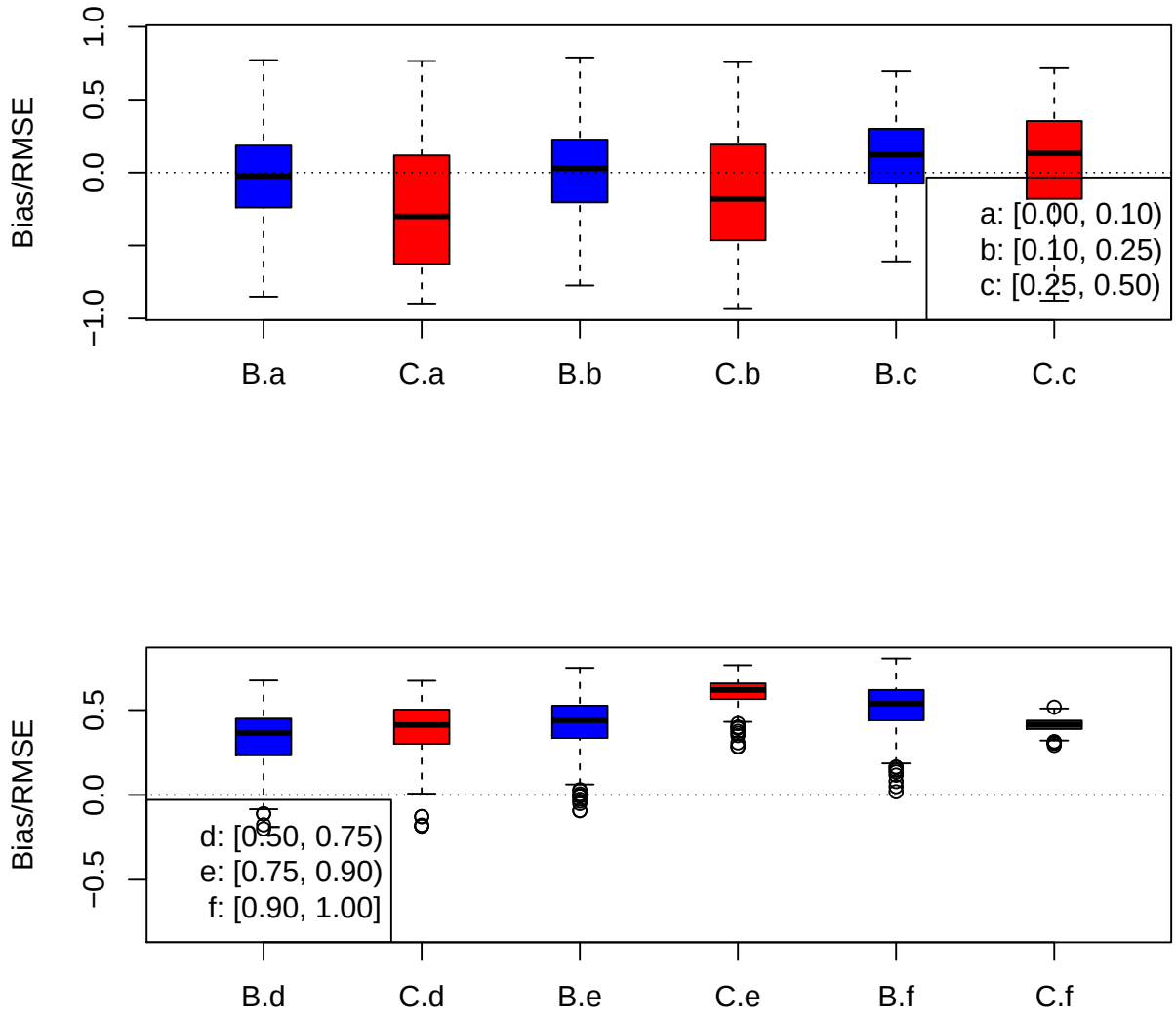
Supplementary Figure 2. Nonproportional setting: we simulated a data set not conforming to proportionality of the covariates and then we estimated the corresponding cumulative intensity by recurrent events with BART. In the left (right) figure, we display the known values (estimates) of the cumulative intensity. The R^2 between known values and estimates is 0.999.



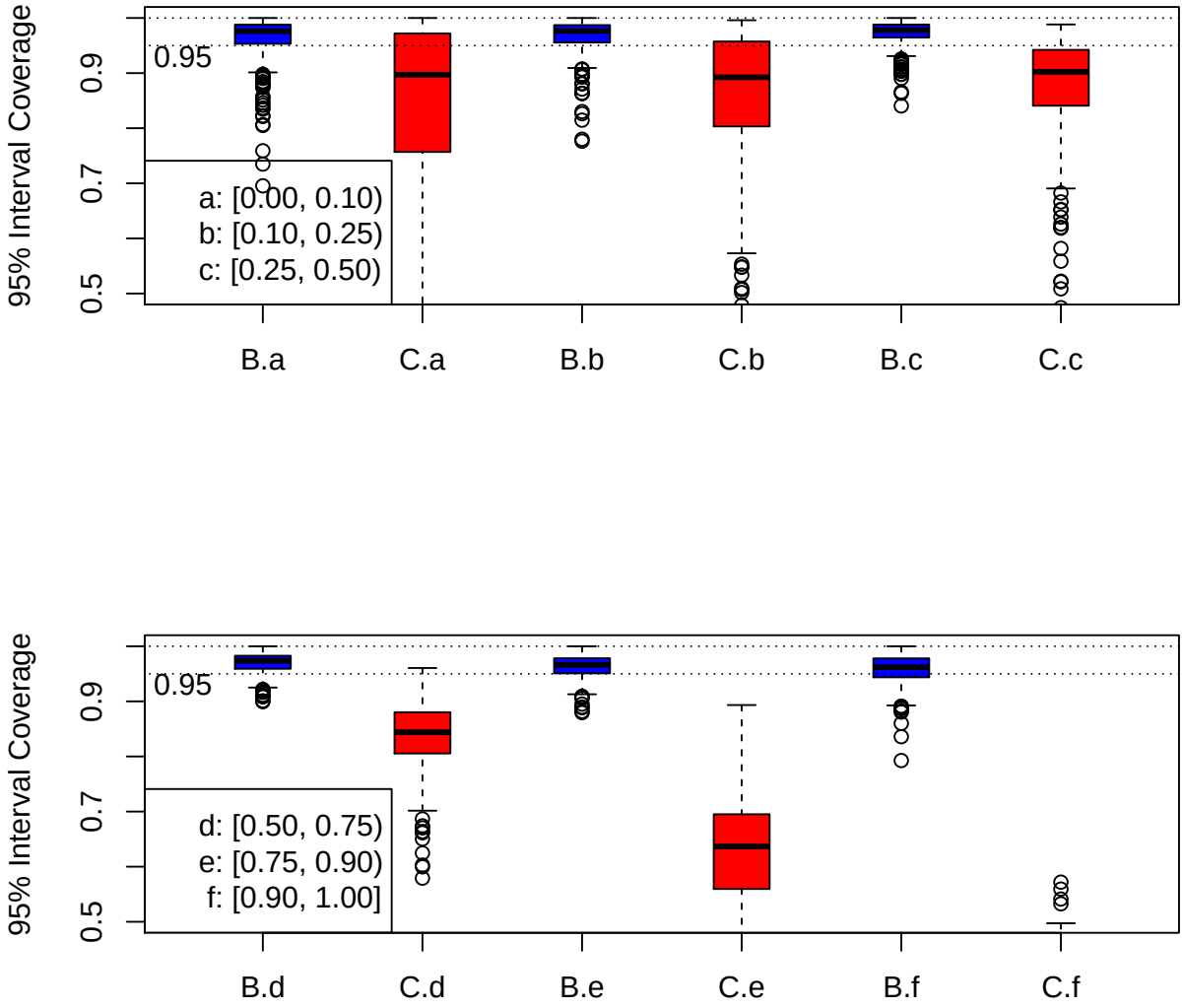
Supplementary Figure 3. Proportional setting RMSE: we performed a simulated data study conforming to proportionality of the covariates and then we estimated the corresponding cumulative intensity with a counting process Cox (CPC) model and recurrent events with BART. Here, we summarize the results for the root mean square error (RMSE): BART (B in blue) vs. CPC (C in red). RMSE is summarized over realms for the quantiles of the true cumulative intensity labeled a-f. Generally, BART and CPC are equivalent. The exception is the last realm where BART is superior.



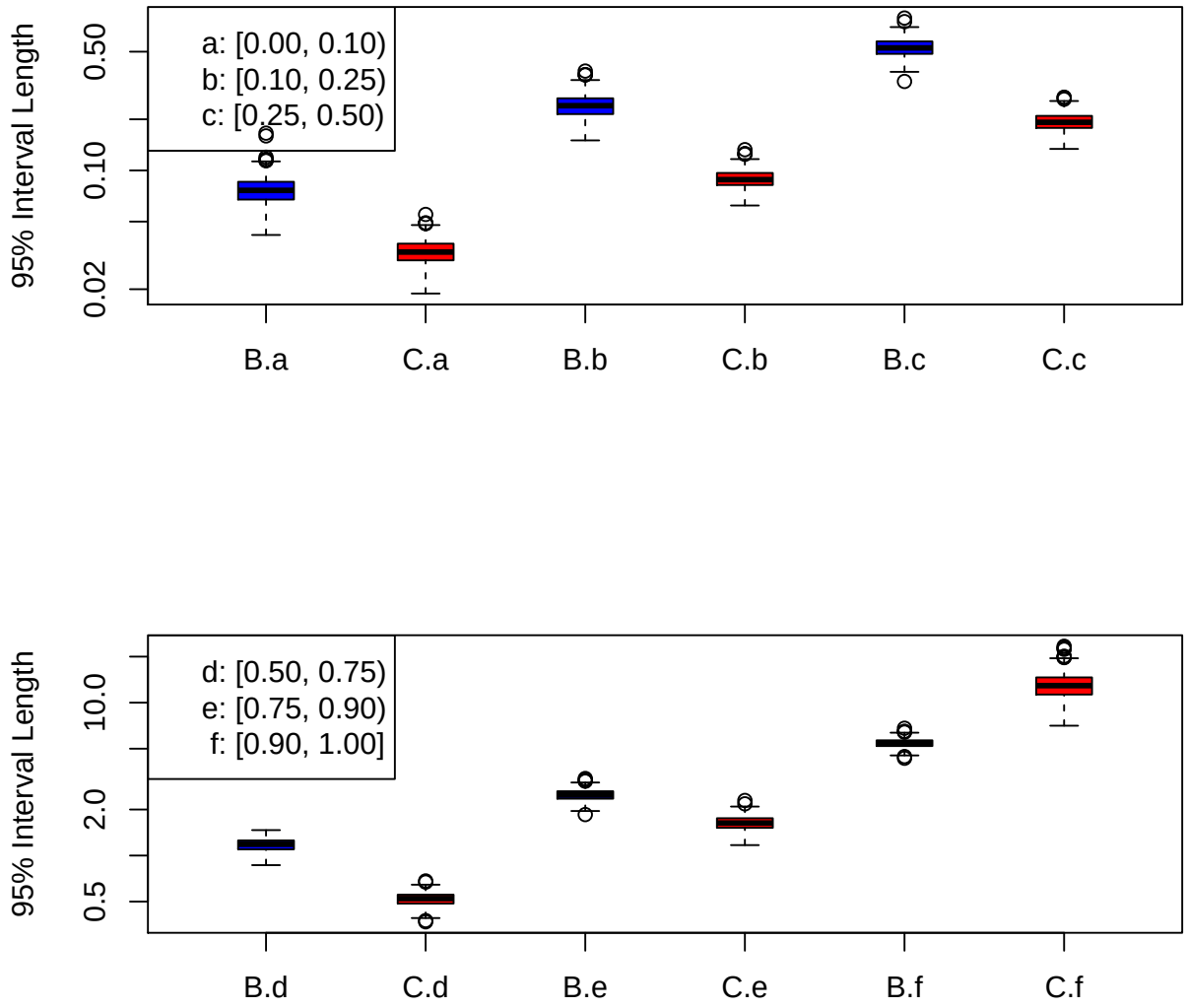
Supplementary Figure 4. Proportional setting Bias: we performed a simulated data study conforming to proportionality of the covariates and then we estimated the corresponding cumulative intensity with a counting process Cox (CPC) model and recurrent events with BART. Here, we summarize the results for the Bias: BART (B in blue) vs. CPC (C in red). Bias is summarized over realms for the quantiles of the true cumulative intensity labeled a-f. Generally, BART and CPC are equivalent. The exceptions are in the latter realms that are more challenging: [0.75, 0.90) and [0.90, 1.00]. In fact, the CPC performance in the last realm is so poor that its boxplot is out of range of the figure so it is not shown.



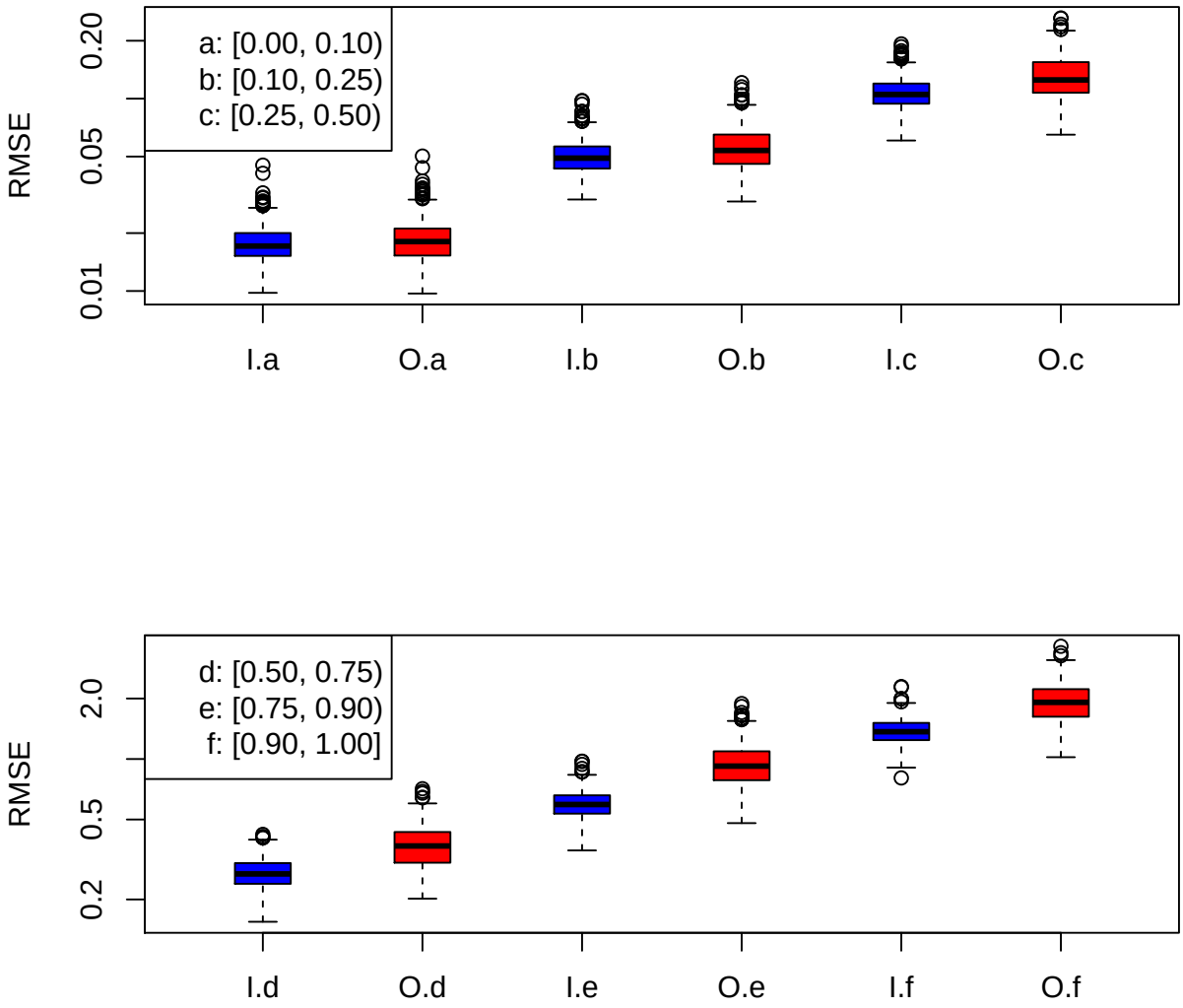
Supplementary Figure 5. Proportional setting Bias/RMSE: we performed a simulated data study conforming to proportionality of the covariates and then we estimated the corresponding cumulative intensity with a counting process Cox (CPC) model and recurrent events with BART. Here, we summarize the results for the Bias/RMSE: BART (B in blue) vs. CPC (C in red). Bias/RMSE is summarized over realms for the quantiles of the true cumulative intensity labeled a-f. Generally, BART and CPC are equivalent.



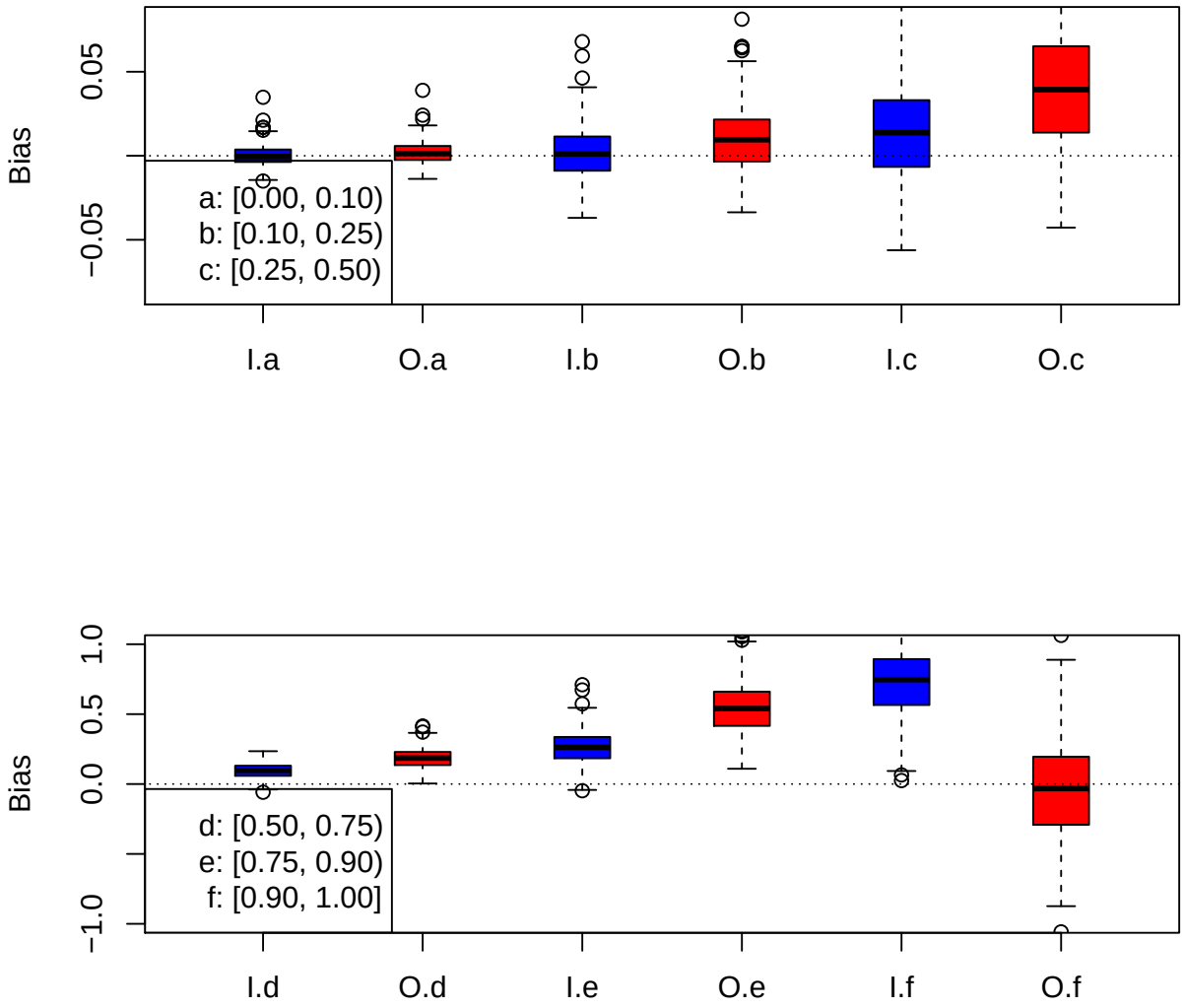
Supplementary Figure 6. Proportional setting 95% Interval Coverage: we performed a simulated data study conforming to proportionality of the covariates and then we estimated the corresponding cumulative intensity with a counting process Cox (CPC) model and recurrent events with BART. Here, we summarize the results for the 95% Interval Coverage: BART (B in blue) vs. CPC (C in red). 95% Interval Coverage is summarized over realms for the quantiles of the true cumulative intensity labeled a-f. Generally, BART coverage is nominal. However, surprisingly, CPC coverage does not approach nominal levels in the latter realms: [0.50, 0.75); [0.75, 0.90) nor [0.90, 1.00]. Perhaps, the poor CPC coverage performance is the result of the discrete-time nature of the data and the computational infeasibility of performing a proper correction for tied event times.



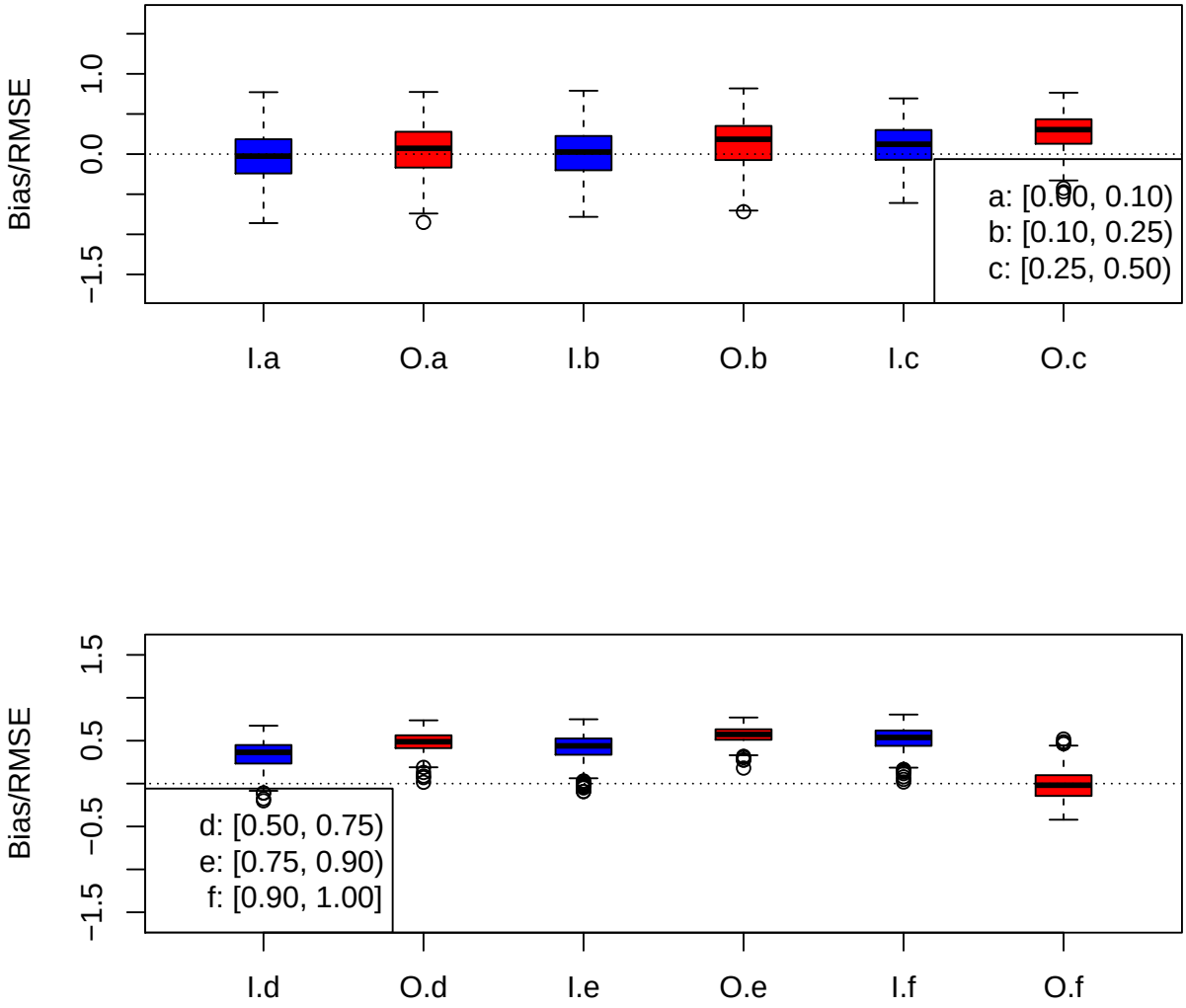
Supplementary Figure 7. Proportional setting 95% Interval Length: we performed a simulated data study conforming to proportionality of the covariates and then we estimated the corresponding cumulative intensity with a counting process Cox (CPC) model and recurrent events with BART. Here, we summarize the results for the 95% Interval Length: BART (B in blue) vs. CPC (C in red). 95% Interval Length is summarized over realms for the quantiles of the true cumulative intensity labeled a-f. Generally, BART interval length is longer that partially explains its better interval coverage. The only exception is in the last realm, [0.90, 1.00], where the CPC results are poor across all metrics.



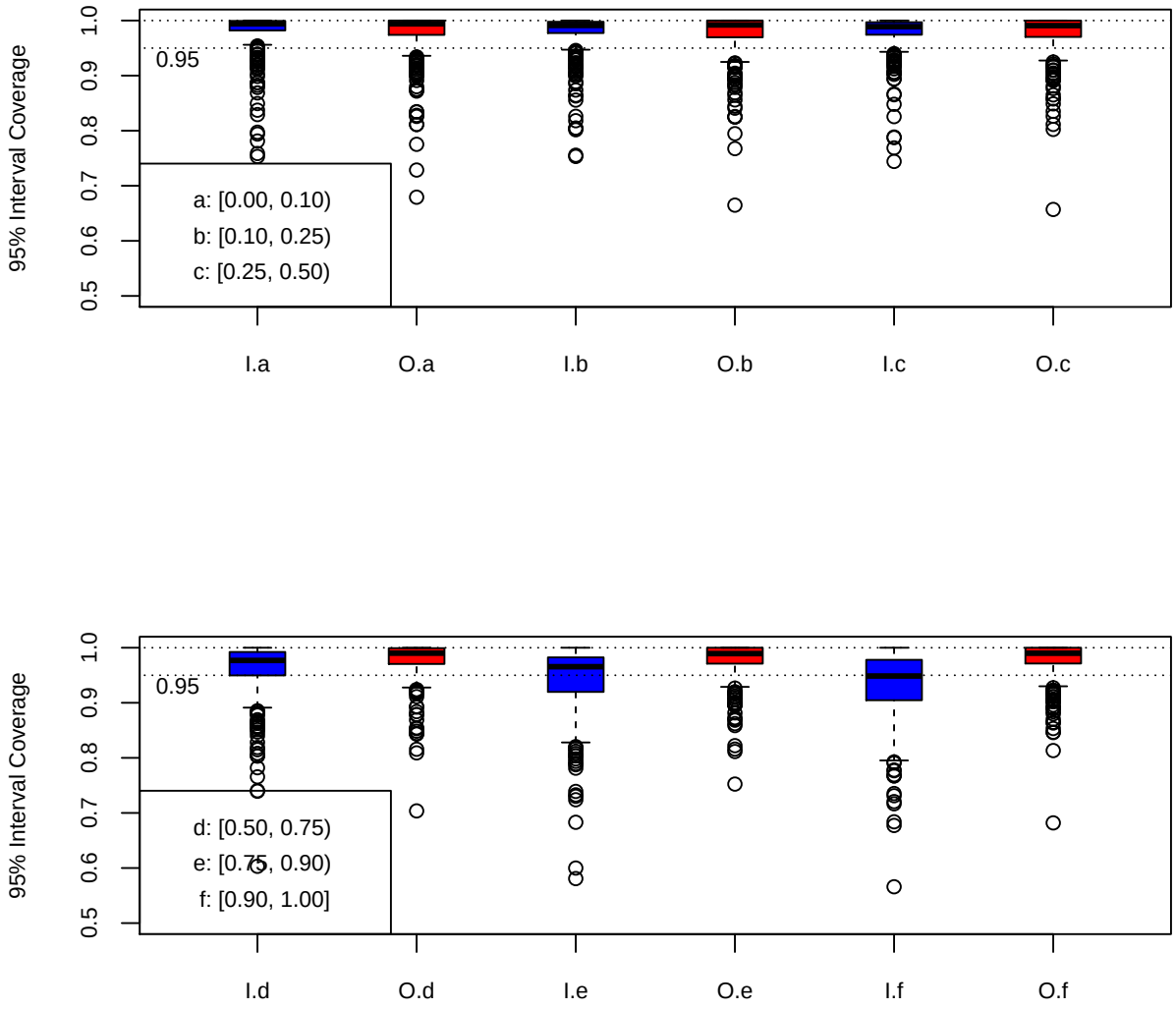
Supplementary Figure 8. Proportional setting RMSE: we performed a simulated data study conforming to proportionality of the covariates and then we estimated the corresponding cumulative intensity by recurrent events with BART. Here, we summarize the results for the root mean square error (RMSE): BART In-sample (I in blue) vs. Out-of-sample (O in red). RMSE is summarized over realms for the quantiles of the true cumulative intensity labeled a-f. In-sample and Out-of-sample performance are generally consistent in the lower half of the realms where Out-of-sample is only slightly larger as anticipated. However, in the upper half of the realms, Out-of-sample is noticeably larger.



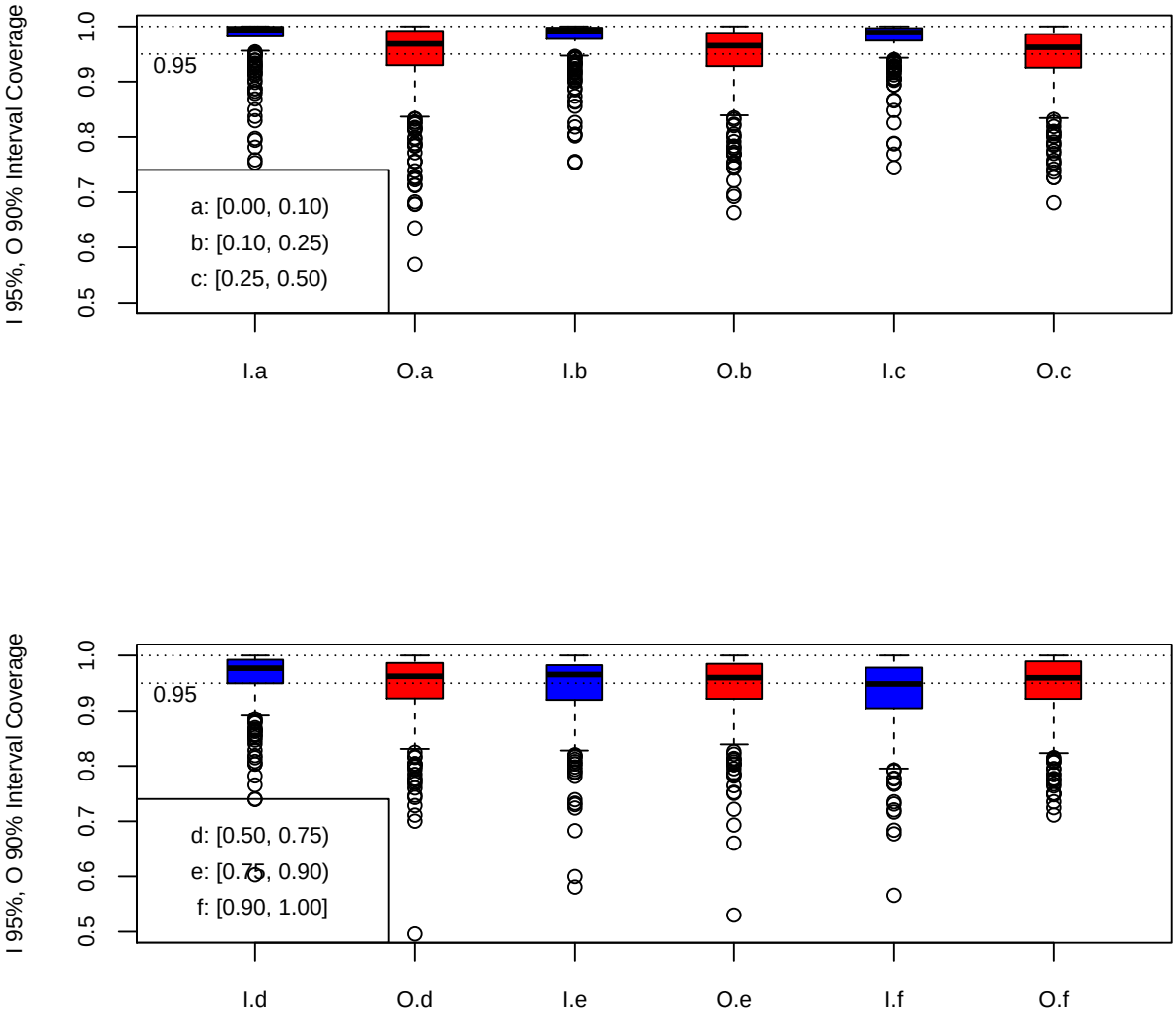
Supplementary Figure 9. Proportional setting Bias: we performed a simulated data study conforming to proportionality of the covariates and then we estimated the corresponding cumulative intensity by recurrent events with BART. Here, we summarize the results for the Bias: BART In-sample (I in blue) vs. Out-of-sample (O in red). Bias is summarized over realms for the quantiles of the true cumulative intensity labeled a-f. In-sample and Out-of-sample performance are generally consistent in the lower half of the realms where Out-of-sample is only slightly worse as anticipated. However, in realms, [0.50, 0.75) and [0.75, 0.90), Out-of-sample is noticeably worse while in the last realm, [0.90, 1.00], the opposite is the case.



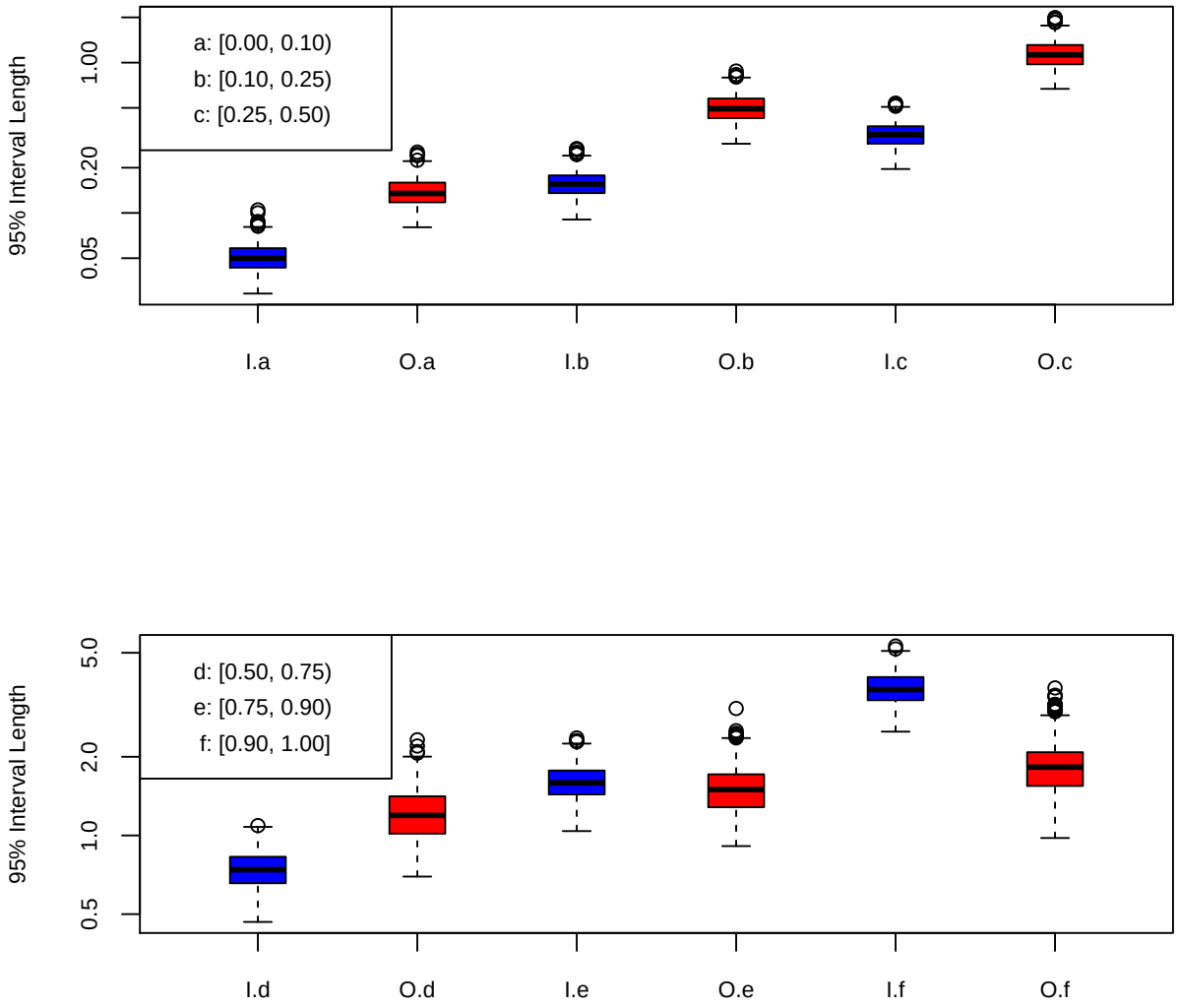
Supplementary Figure 10. Proportional setting Bias/RMSE: we performed a simulated data study conforming to proportionality of the covariates and then we estimated the corresponding cumulative intensity by recurrent events with BART. Here, we summarize the results for the Bias/RMSE: BART In-sample (I in blue) vs. Out-of-sample (O in red). Bias/RMSE is summarized over realms for the quantiles of the true cumulative intensity labeled a-f. In-sample and Out-of-sample performance are generally consistent in the lower half of the realms where Out-of-sample is only slightly worse as anticipated. However, in realms, [0.50, 0.75) and [0.75, 0.90), Out-of-sample is more noticeably worse while in the last realm, [0.90, 1.00], the opposite is the case.



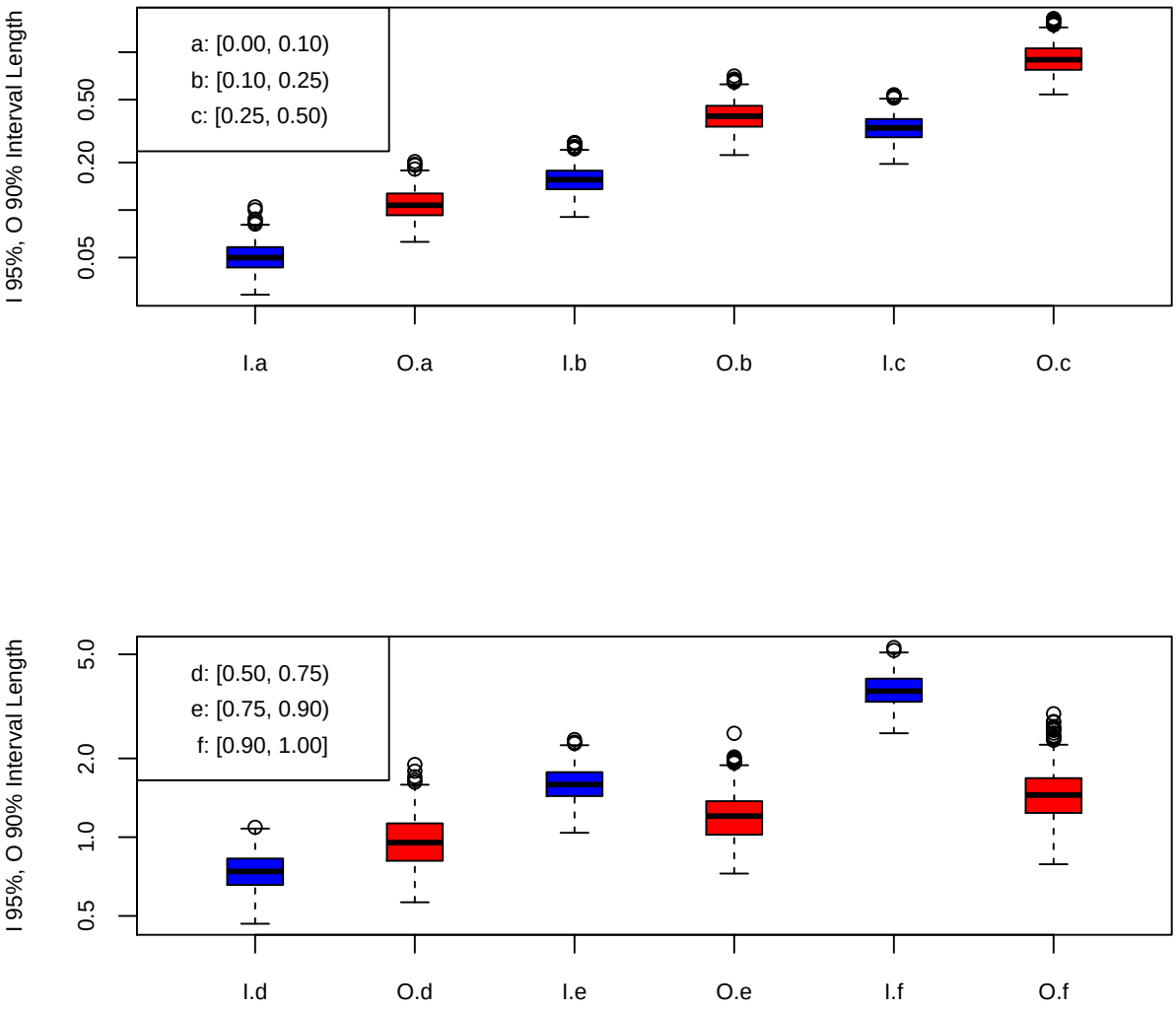
Supplementary Figure 11. Proportional setting 95% Interval Coverage: we performed a simulated data study conforming to proportionality of the covariates and then we estimated the corresponding cumulative intensity by recurrent events with BART. Here, we summarize the results for the 95% Interval Coverage: BART In-sample (I in blue) vs. Out-of-sample (O in red). 95% Interval Coverage is summarized over realms for the quantiles of the true cumulative intensity labeled a-f. In-sample performance is generally nominal while Out-of-sample exceeds nominal levels.



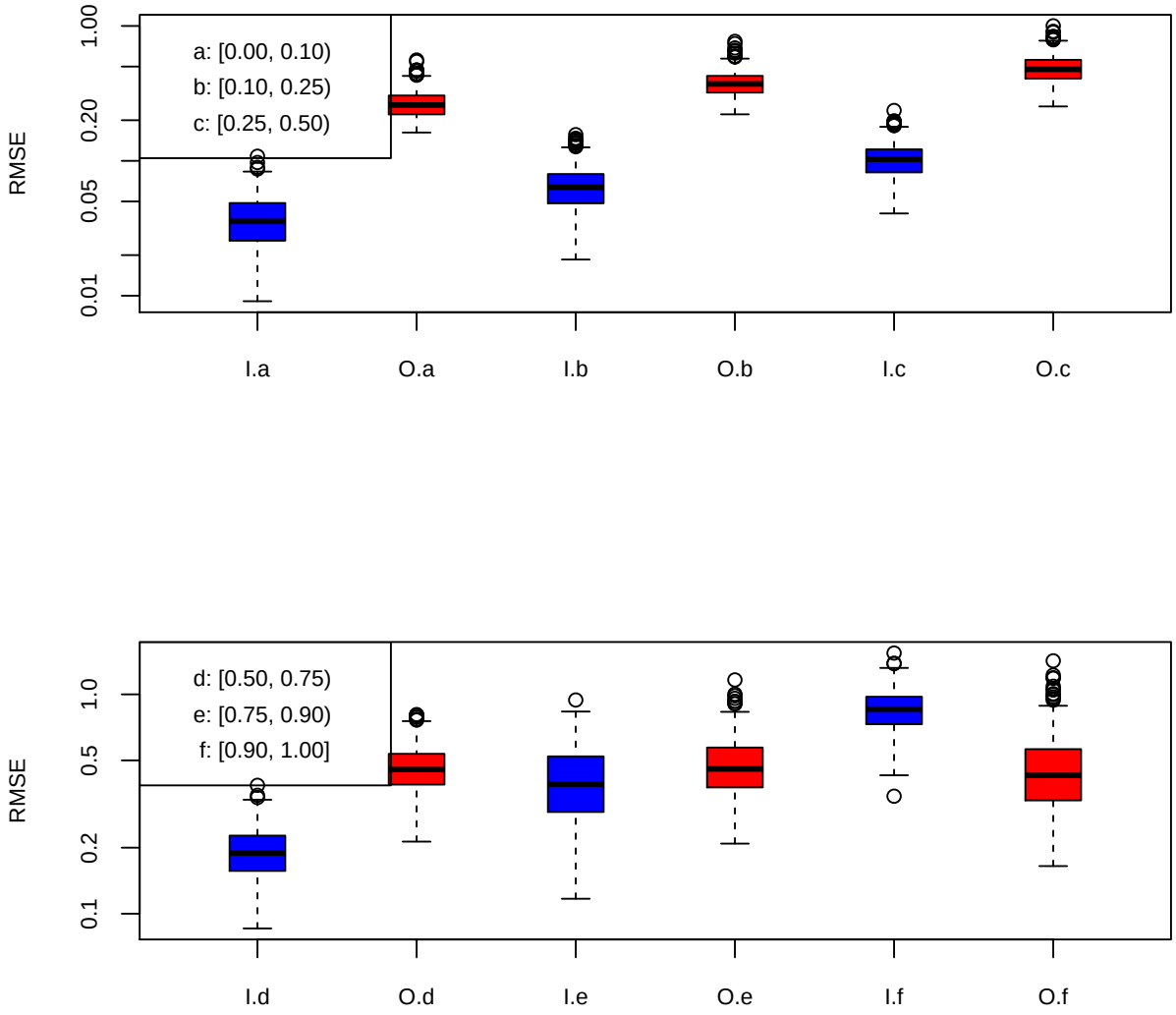
Supplementary Figure 12. Proportional setting 95% (90%) Interval Coverage for In-sample (Out-of-sample): we performed a simulated data study conforming to proportionality of the covariates and then we estimated the corresponding cumulative intensity by recurrent events with BART. Here, we summarize the results for the In-sample 95% vs. Out-of-sample 90% Interval Coverage: BART In-sample (I in blue) vs. Out-of-sample (O in red). Interval Coverage is summarized over realms for the quantiles of the true cumulative intensity labeled a-f. In-sample 95% and Out-of-sample 90% Interval Coverage are roughly nominal throughout. N.B. Out-of-sample 90% Credible Intervals were chosen as 95% nominal by cross-validation calibration based on one data set; see Section E for more details.



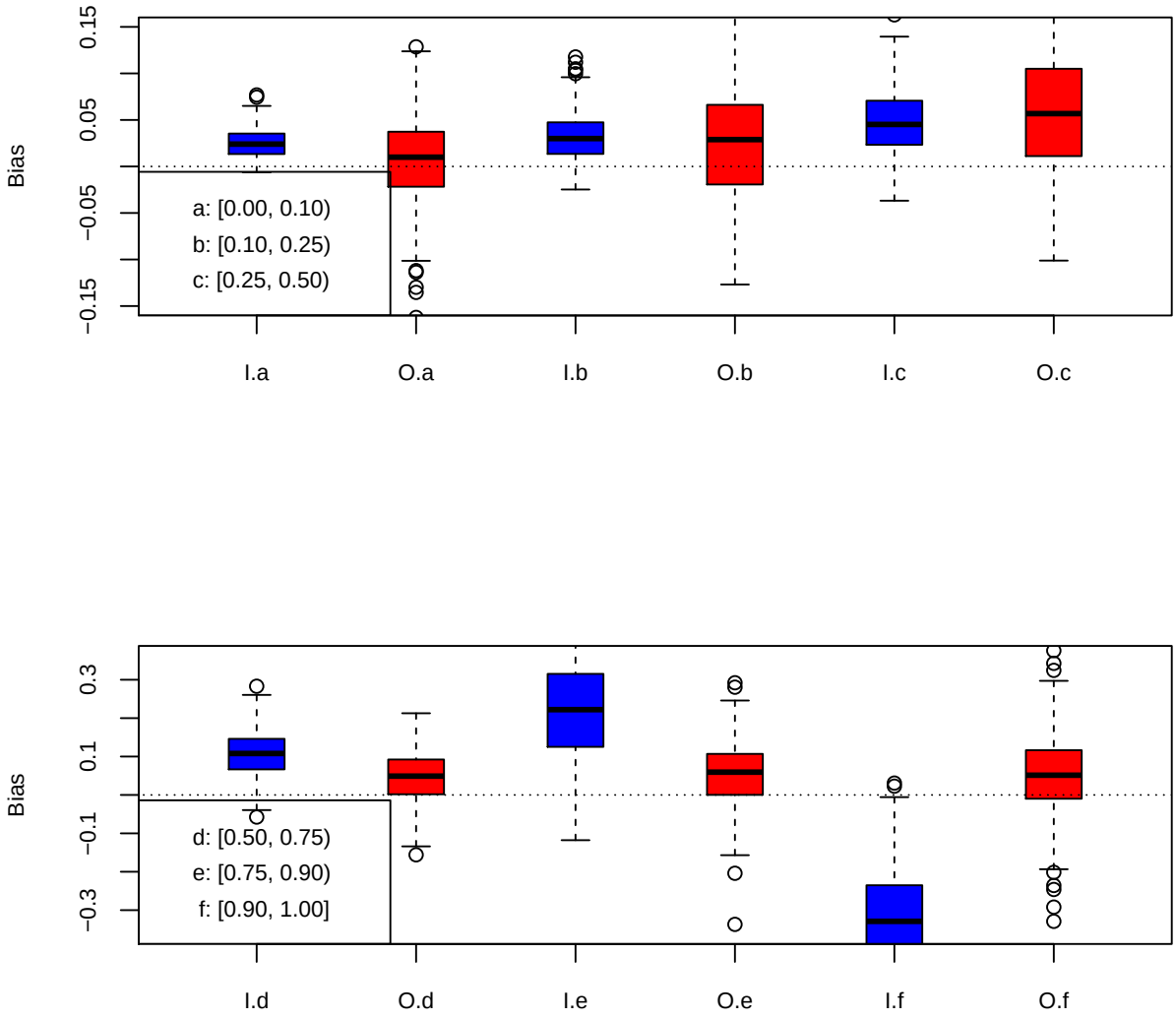
Supplementary Figure 13. Proportional setting 95% Interval Length: we performed a simulated data study conforming to proportionality of the covariates and then we estimated the corresponding cumulative intensity by recurrent events with BART. Here, we summarize the results for the 95% Interval Length: BART In-sample (I in blue) vs. Out-of-sample (O in red). 95% Interval Length is summarized over realms for the quantiles of the true cumulative intensity labeled a-f. In-sample length is generally shorter than Out-of-sample.



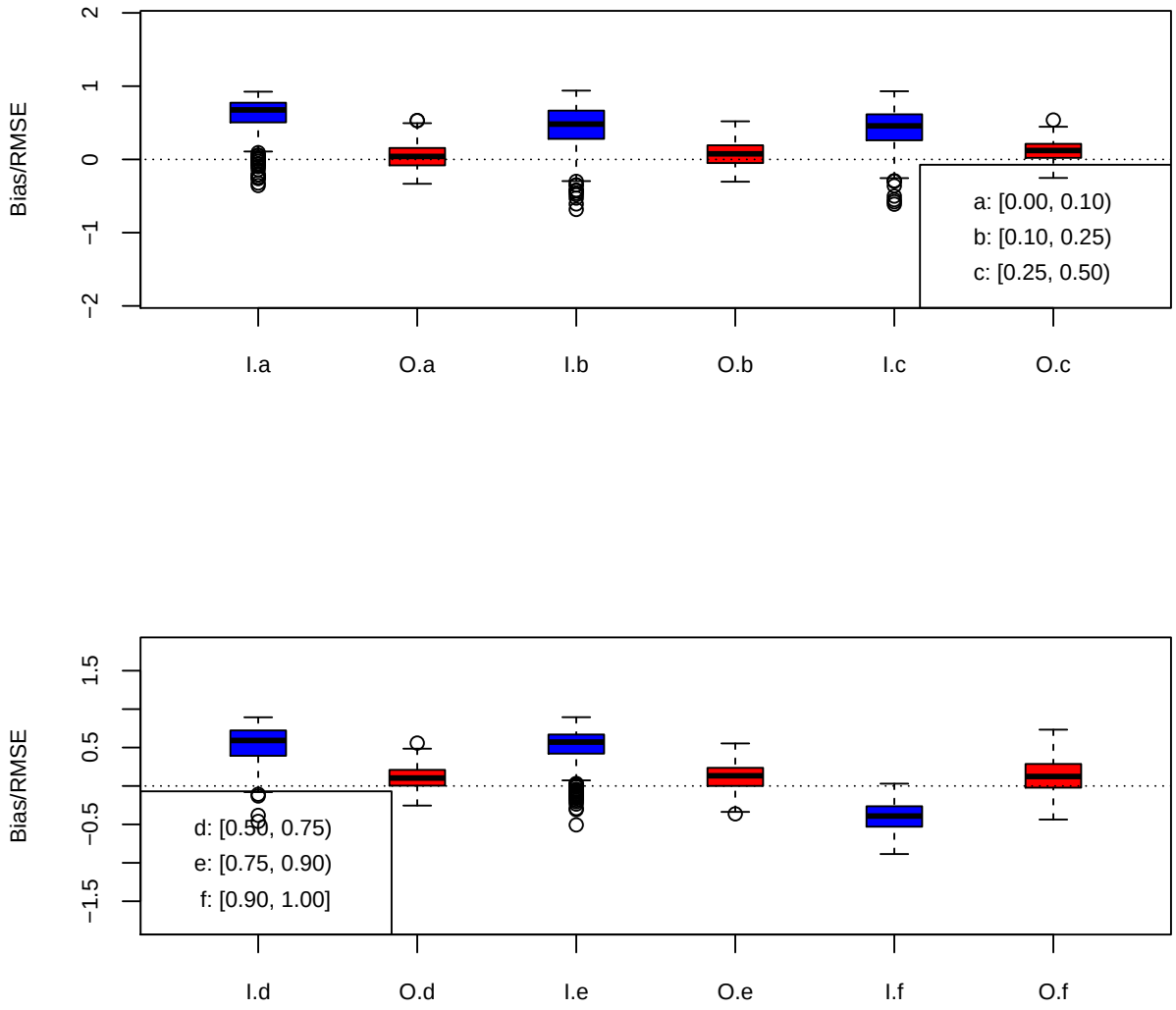
Supplementary Figure 14. Proportional setting 95% (90%) Interval Length for In-sample (Out-of-sample): we performed a simulated data study conforming to proportionality of the covariates and then we estimated the corresponding cumulative intensity by recurrent events with BART. Here, we summarize the results for the In-sample 95% vs. Out-of-sample 90% Interval Length: BART In-sample (I in blue) vs. Out-of-sample (O in red). Interval Length is summarized over realms for the quantiles of the true cumulative intensity labeled a-f. In-sample 95% Interval Length is closer to Out-of-sample 90% than to Out-of-sample 95%.



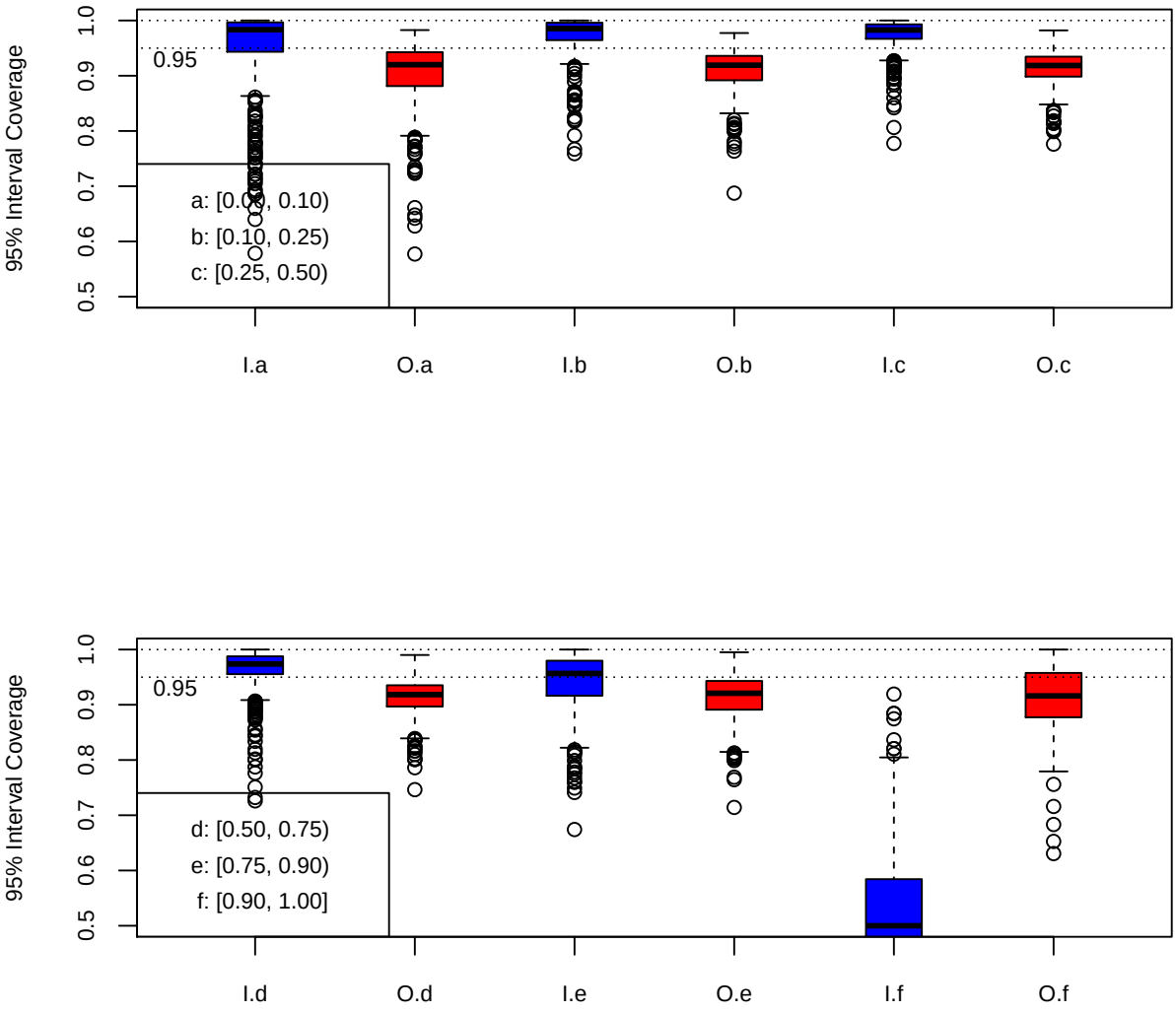
Supplementary Figure 15. Nonproportional setting RMSE: we performed a simulated data study not conforming to proportionality of the covariates and then we estimated the corresponding cumulative intensity by recurrent events with BART. Here, we summarize the results for the root mean square error (RMSE): BART In-sample (I in blue) vs. Out-of-sample (O in red). RMSE is summarized over realms for the quantiles of the true cumulative intensity labeled a-f. In-sample performance is generally better than Out-of-sample except in the last realm, [0.90, 1.00].



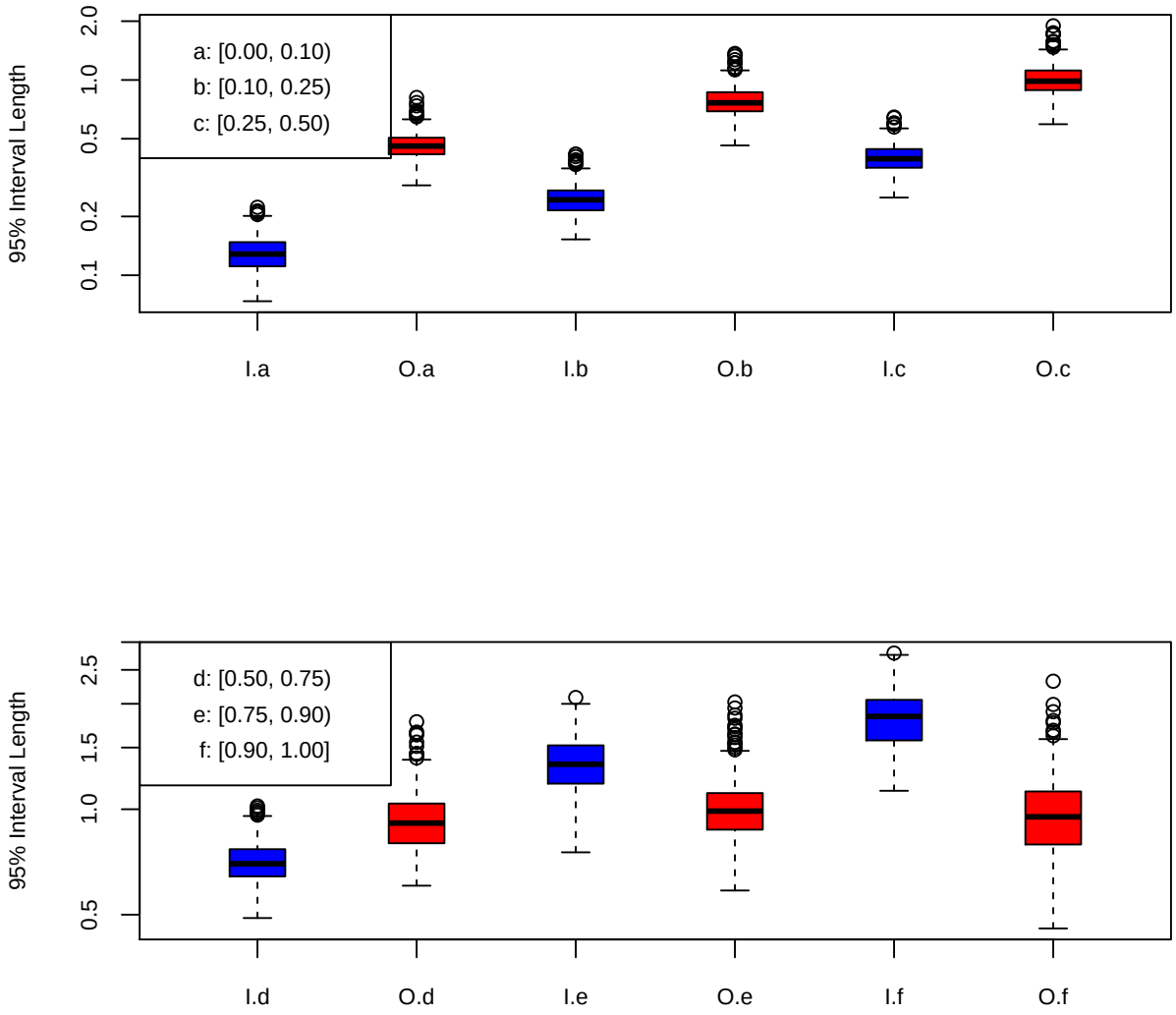
Supplementary Figure 16. Nonproportional setting Bias: we performed a simulated data study not conforming to proportionality of the covariates and then we estimated the corresponding cumulative intensity by recurrent events with BART. Here, we summarize the results for the Bias: BART In-sample (I in blue) vs. Out-of-sample (O in red). Bias is summarized over realms for the quantiles of the true cumulative intensity labeled a-f. In-sample and Out-of-sample performance are generally consistent in the lower half of the realms. However, in the upper half of realms, Out-of-sample is noticeably better.



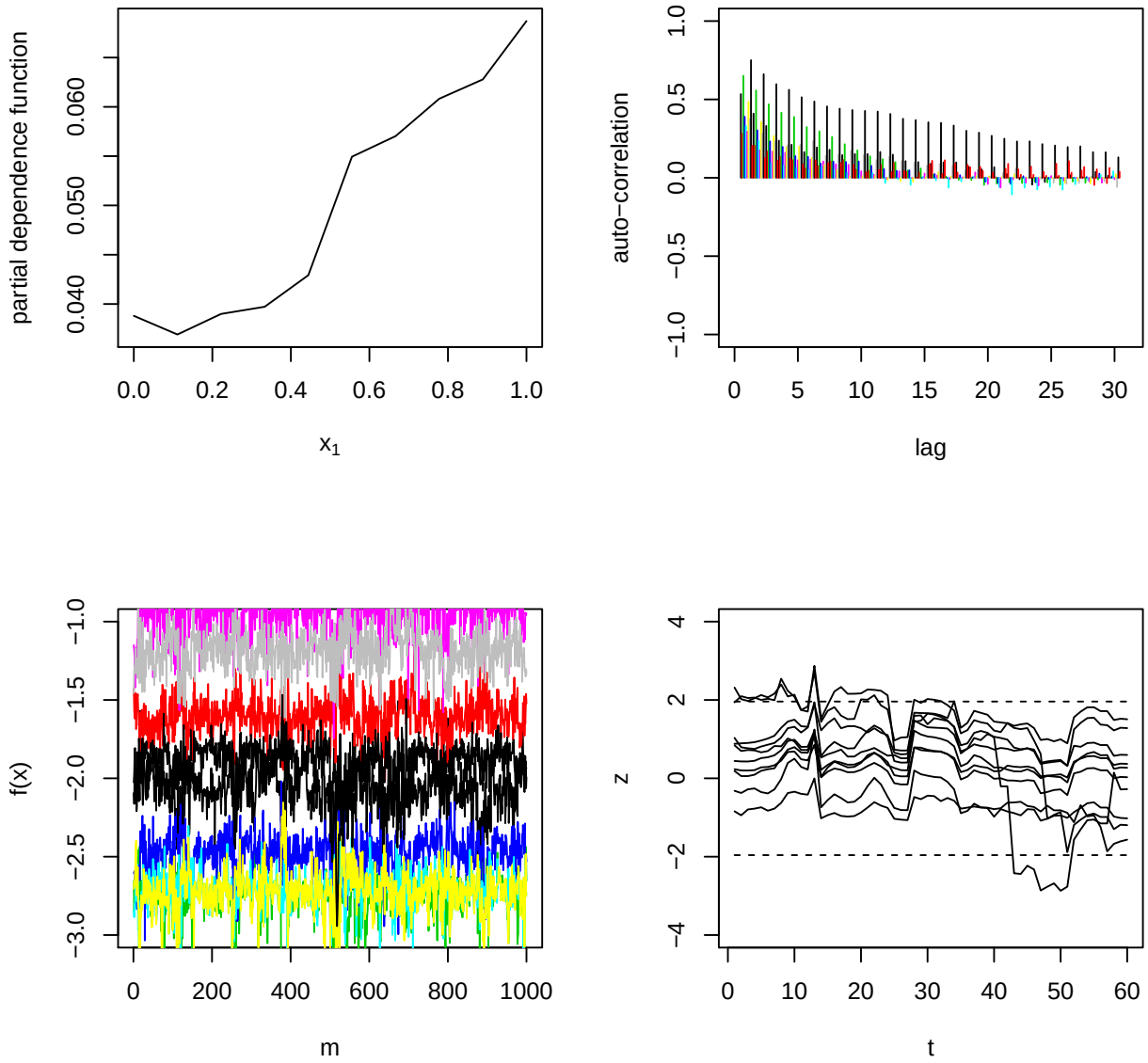
Supplementary Figure 17. Nonproportional setting Bias/RMSE: we performed a simulated data study not conforming to proportionality of the covariates and then we estimated the corresponding cumulative intensity by recurrent events with BART. Here, we summarize the results for the Bias/RMSE: BART In-sample (I in blue) vs. Out-of-sample (O in red). Bias/RMSE is summarized over realms for the quantiles of the true cumulative intensity labeled a-f. Out-of-sample performance is generally better, but this is an artifact of the Bias/RMSE metric since RMSE is worse for Out-of-sample in the lower half of the realms.



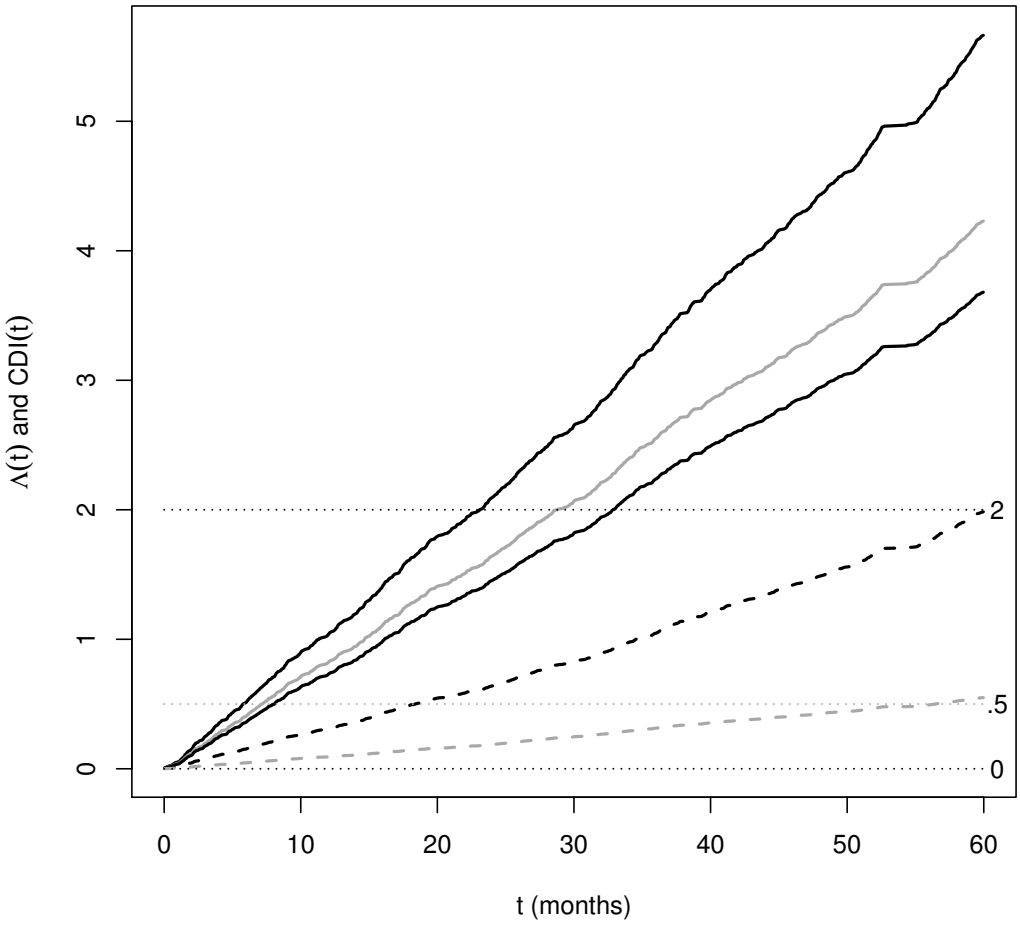
Supplementary Figure 18. Nonproportional setting 95% Interval Coverage: we performed a simulated data study not conforming to proportionality of the covariates and then we estimated the corresponding cumulative intensity by recurrent events with BART. Here, we summarize the results for the 95% Interval Coverage: BART In-sample (I in blue) vs. Out-of-sample (O in red). 95% Interval Coverage is summarized over realms for the quantiles of the true cumulative intensity labeled a-f. In-sample (Out-of-sample) performance are generally nominal (near nominal) throughout with the exception of In-sample in the last realm, [0.90, 1.00].



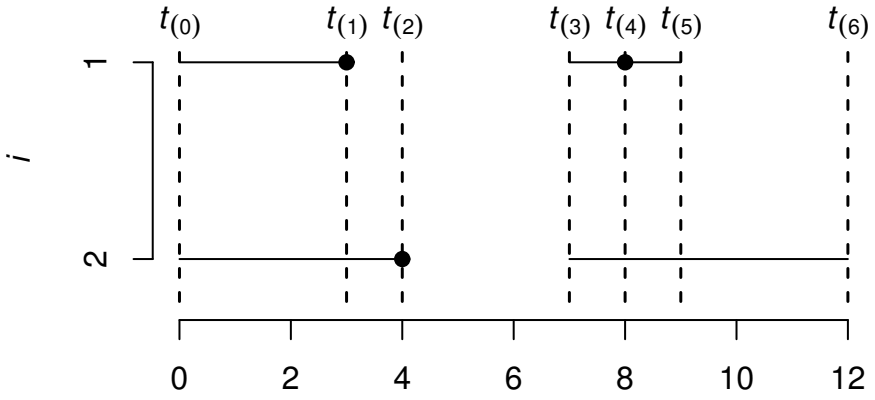
Supplementary Figure 19. Nonproportional setting 95% Interval Length we performed a simulated data study not conforming to proportionality of the covariates and then we estimated the corresponding cumulative intensity by recurrent events with BART. Here, we summarize the results for the 95% Interval Length: BART In-sample (I in blue) vs. Out-of-sample (O in red). 95% Interval Length is summarized over realms for the quantiles of the true cumulative intensity labeled a-f. Out-of-sample length is generally longer (shorter) in the first four (last two) realms.



Supplementary Figure 20. We applied convergence diagnostics to the first data set from the proportional setting of the simulation study. Based on reviewing figures like these, we chose a thinning parameter of 100 that is depicted here. In the upper left quadrant, we have plotted Friedman's partial dependence function for $f(x_1)$ vs. x_1 for 10 values of x_1. This is a check that can't be performed for real data, but it is informative in this case. Notice that $f(x_1)$ vs. x_1 is directly proportional as expected. In the upper right quadrant, we plot the auto-correlations of $f(t_{(j)}, \boldsymbol{x}_i)$ for 10 randomly selected $t_{(j)}$ and $\boldsymbol{x}_i$ combinations where i (j) indexes subjects (time points). Notice that there is a combination that has fairly high auto-correlation, but the rest are quite reasonable. In the lower left quadrant, we display the corresponding trace plots for these same combinations. The traces demonstrate that $f(t_{(j)}, \boldsymbol{x}_i)$ appear to adequately traverse the sample space. In the lower right quadrant, we have selected 10 subjects and we plot their corresponding Geweke Z_{AB} statistics over the time points. Notice that only 2 or 3 subjects ever reach the 95% boundaries and only rarely; given the number of comparisons, 600, this seems reasonable as well.



Supplementary Figure 21. We studied a cohort of patients suffering diabetes to determine the covariates related to the risk of hospital admissions. Based on a recurrent events analysis with BART, we determined that there are three important risk agonists for a new hospital admission: peripheral vascular disease (PVD), receiving insulin treatment and the number of previous hospital admissions. The effect of binary covariates like PVD and insulin are relatively easily summarized. However, the number of previous hospital admissions is more difficult because it is time-dependent. To explore these risk factors, we present the estimated cumulative intensities, $\Lambda(t)$, for three risk profiles: low (lower solid black), medium (middle solid gray) and high risk (upper solid black). In these profiles, PVD and insulin are set to either present or absent throughout the five year observation period. For low risk subjects, PVD and insulin are absent and there are no hospital admissions. For medium risk subjects, PVD is absent, insulin is present and there is one hospital admission at 24 months. For high risk subjects, PVD and insulin are present and they are admitted to the hospital at 12, 24, 36 and 48 months. The estimated cumulative intensities displayed are the effects of the risk profiles marginalizing over all other covariates with Friedman's partial dependence function. Notice that the cumulative intensities fall in the predetermined order from low, to medium, to high risk. The estimated cumulative differential intensities, $CDI(t)$, are also displayed for these profiles. The medium vs. low (dashed gray) is $CDI_{ML}(t) = \Lambda(t, x_M) - \Lambda(t, x_L)$. Similarly, the high vs. low (dashed black) is $CDI_{HL}(t) = \Lambda(t, x_H) - \Lambda(t, x_L)$. The estimated cumulative differential intensities displayed are the effects of the risk profiles marginalizing over all other covariates with Friedman's partial dependence function. Medium risk subjects will likely have 0.5 more hospital admissions over 5 years than low risk subjects. Meanwhile, high risk subjects will likely have 2 more hospital admissions than low risk. The dotted horizontal lines of 0 (black), 0.5 (gray) and 2 (black) are plotted for reference.



Supplementary Figure 22. Risk set diagram. Time is on the horizontal axis; subjects, i, are on the vertical axis. Suppose that we have two subjects with the following values:
$N_1 = 2, s_1 = 9, t_{11} = 3, u_{11} = 7, t_{12} = 8, u_{12} = 8$
$N_2 = 1, s_2 = 12, t_{21} = 4, u_{21} = 7.$
The grid of time points are dashed vertical lines; events are solid black dots; and the risk set, R_i, for each subject is a solid black line while at risk, otherwise absent. Notice for subject 1 that they are not at risk in the interval $(t_{(1)}, t_{(3)})$; no events could occur in this interval since their first event had not ended yet, i.e., these time points do not contribute to the likelihood since these subjects are not chronologically at risk for an event.