Received: 4 May 2020 | Revised: 10 March 2021 | Accepted: 1 April 2021

DOI: 10.1111/biom.13478

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

BIOMETRIC METHODOLOGY

Semiparametric analysis of clustered interval-censored survival data using soft Bayesian additive regression trees (SBART)

Piyali Basak[1] | Antonio Linero[2] | Debajyoti Sinha[1] | Stuart Lipsitz[3]

[1] Florida State University, FL, USA

[2] University of Texas at Austin, TX, USA

[3] Brigham and Women's Hospital, MA, USA

Correspondence Piyali Basak, Florida State University.
Email: pb15d@my.fsu.edu

Funding information
National Science Foundation, Grant/Award Number: DMS-1712870; Center for Scientific Review, Grant/Award Number: R03CA205018-01

Abstract

Popular parametric and semiparametric hazards regression models for clustered survival data are inappropriate and inadequate when the unknown effects of different covariates and clustering are complex. This calls for a flexible modeling framework to yield efficient survival prediction. Moreover, for some survival studies involving time to occurrence of some asymptomatic events, survival times are typically interval censored between consecutive clinical inspections. In this article, we propose a robust semiparametric model for clustered interval-censored survival data under a paradigm of Bayesian ensemble learning, called soft Bayesian additive regression trees or SBART (Linero and Yang, 2018), which combines multiple sparse (soft) decision trees to attain excellent predictive accuracy. We develop a novel semiparametric hazards regression model by modeling the hazard function as a product of a parametric baseline hazard function and a nonparametric component that uses SBART to incorporate clustering, unknown functional forms of the main effects, and interaction effects of various covariates. In addition to being applicable for left-censored, right-censored, and interval-censored survival data, our methodology is implemented using a data augmentation scheme which allows for existing Bayesian backfitting algorithms to be used. We illustrate the practical implementation and advantages of our method via simulation studies and an analysis of a prostate cancer surgery study where dependence on the experience and skill level of the physicians leads to clustering of survival times. We conclude by discussing our method's applicability in studies involving high-dimensional data with complex underlying associations.

KEYWORDS

Bayesian additive regression trees, machine learning, nonproportional hazards, semiparametric, survival analysis

1 | INTRODUCTION

Interval-censored survival data arise frequently in asymptomatic diseases that have no immediate outward symptoms (Sun, 2006) and the event of interest, such as device failure or relapse of a disease after initial treatment, is known to occur only between two consecutive inspection times. For instance, recurrence of biomarkers are often detected either during scheduled clinic visits or via some invasive procedures/diagnostic tests causing the survival times to such events to be interval censored within consecutive inspection times with opposite diagnosis/test results. Our motivation behind this article is to compare two competing surgical techniques, robotic versus non-robotic radical retropubic prostatectomy (RRP), for prostate removal in terms of the patients' time to recurrence (survival time of interest) of prostate specific antigen (PSA) after the surgery (Barbash and Glied, 2010). The PSA level in the blood after prostate removal surgery is 0.0 ng/ml, and a PSA recurrence is defined as the time after surgery at which the PSA level exceeds 0.2 ng/ml. Genitourinary surgeons and oncologists are particularly interested in whether a surgery using a robotic device improves the time to PSA recurrence compared to non-robotic surgery for removing the cancerous prostate. Because continuous monitoring of PSA is not feasible, the time to PSA recurrence for each patient is interval censored between consecutive blood tests. Further, any assumed parametric model for the regression effects of various baseline covariates on risk/hazard of time to PSA recurrence is restrictive. In practice, it is difficult to justify the proportional hazards assumption strictly based on subject-matter knowledge, much less the more restrictive parametric assumptions which underlie methods such as Weibull regression. Hence, using a flexible modeling approach which accommodates possible nonlinearity and multi-way interactions among the variables is prudent. Moreover, we expect these survival times to be clustered due to the possible presence of unobserved surgeon effects, such as the surgeon's experience, skill, and access/familiarity with modern medical technology. Hence, we need to account for an appropriate within-cluster association (in this case, due to surgeon effects) to obtain efficient and appropriate models for survival prediction.

Most semiparametric models for interval-censored survival data use either an accelerated failure time (AFT) model (Zhou et al, 2017) or a proportional hazards model (Sinha *et al.*, 1999; Sun, 2006; Hanson and Yang, 2007). However, AFT models do not straightaway give an often-desired interpretation of instantaneous risk and dynamic comparison between two subjects over time. On the other hand, survival data can often violate the restrictive assumption of proportional hazards in practice. In addition, dependence on factors such as experience and skill level of physicians and clinics often lead to clustering of survival data due to random clinician/clinical site effects. Some examples of clustered interval-censored survival data include the National Aeronautics and Space Administration's hypobaric decompression sickness study (HDSD) (Conkin *et al.*, 1992; Conkin and Powell, 2001) and the lymphatic filariasis study (Dreyer *et al.*, 2006). However, failure to account for clustering effects may produce misleading results and, in particular, inaccurate estimation of the precision of the estimated parameters. Most methods for analyzing clustered survival data use frailty random effects (Oakes, 1982) to accommodate within-cluster association as well as between-cluster heterogeneity. Goethals *et al.* (2009) used a proportional hazards model with a Gamma frailty to analyze clustered interval-censored survival data.

However, main regression effects and effects of interaction among different covariates in survival data are often time dependent and are potentially more complex than what can be envisioned by a pre-specified parametric model of covariate effects on either the hazard, the median, or the mean. This calls for a flexible modeling framework for covariate effects. Some of the proposed approaches to accommodate such complex regression relationships include the boosting proportional hazards model of Li and Luan (2005), bagging (Hothorn *et al.*, 2004), and random survival forests (Ishwaran *et al.*, 2008; Umlauf *et al.*, 2015). Using a Gaussian process to incorporate nonparametric covariate effects, Fernandez *et al.* (2016) presented a Bayesian semiparametric survival model centered on a parametric baseline hazard model in a fashion similar to ours. Their approach makes use of a random feature expansion to approximate the kernel of the Gaussian process for computational purposes; however, the kernel they use for their simulations and data analysis incorporates predictors into the kernel in a restrictive fashion, and essentially restricts the form of their model to a time-varying coefficient model.

Boosting is an example of ensemble learning, which combines multiple weak learners to attain both high stability and a flexible model. The Bayesian additive regression trees (BART) framework proposed by Chipman *et al.* (2010) is a Bayesian framework for tree-based ensemble models. BART is a computationally efficient and flexible technique, and can accommodate complex nonlinear regression relationships. Since its introduction, the BART framework has been extended to account for many different types of model specifications, including semiparametric regression under heteroscedasticity (Pratola *et al.*, 2017), log-linear models (Murray, 2017), multivariate and multi-scale data (Linero *et al.*, 2020), causal inference (Hahn *et al.*, 2020), and varying coefficient models (Deshpande *et al.*, 2020). For systematic reviews of Bayesian tree-based methods and BART, see Linero (2017) and Hill *et al.* (2020). A recent surge of interest in BART lies in its applications for analyzing survival data. By modeling nonparametric effects of covariates on the survival times through the tree ensembles, Bonato *et al.* (2011) employed BART for survival prediction in high-dimensional genetic studies under three specific models - the proportional hazards regression model, the AFT regression model, and the Weibull regression models where the last one is the special parametric case of former two models. Sparapani *et al.* (2016) addressed discrete-time survival data with the discrete-time hazard being modeled nonparametrically via BART. However, this approach is aimed essentially at discrete-time survival data and requires the specification of a finite number of possible failure times in order to be used for continuous survival data, and the computational cost increases rapidly for even a moderate increase in the number of possible failure times. Despite this stream of research, to the best of our knowledge, there are no applications of BART for analysis of interval-censored survival data. Additionally, two potential shortcomings of the usual BART framework include the lack of smoothness of the estimated regression function and sensitivity to the curse of dimensionality (Linero, 2018). Linero and Yang (2018) recently addressed these drawbacks with an extension of BART called SBART, which provides a considerable improvement over BART by employing an ensemble of "soft" decision trees that can adapt to the unknown smoothness level of the true regression function, and can also remove irrelevant predictors.

Other strategies for taking a Bayesian nonparametric approach to this problem have been considered in the literature. For example, the ANOVA-DDP approach of De Iorio *et al.* (2009) models the survival function using a Dirichlet process mixture whose atoms depend on covariates. The structure of their model effectively assumes that the log survival time density can be expressed as a mixture of normal distributions, with locations of the mixtures being a function of the covariates, but with fixed mixing weights.

As mentioned by a reviewer, the proposal of this paper is closely related to *hazard regression*, or HARE, approach (Kooperberg et al., 1995). Kooperberg and Clarkson (1997) adapt HARE to address interval-censored data, and models the hazard $\log \lambda(t \mid \mathbf{x})$ using a linear spline basis function expansion. A similar approach is taken by Mallick *et al.* (1999), who use a Bayesian MARS model (which itself is related to the BART models we consider) to perform hazard regression, using log-normal frailties to address clustering. Hazard regression models can be fit in `R` using the `polspline` package. Other tree-based methods exist in the literature for addressing censored survival data with frailties. Su and Tsai (2005) introduce a *tree augmented Cox proportional hazards model*, which uses an auxiliary decision tree to allow for departures from the Cox proportional hazards model. Su *et al.* (2008) propose the use of *interaction trees* to perform subgroup analysis with censored survival outcomes. Calhoun *et al.* (2018) implement *multivariate survival trees* (MSTs) to analyze censored survival data, and allow clustering using frailties as we do here. Each of these methods invokes a *single* tree which acts as a strong learner, unlike the BART-based approach we take which uses many decision trees which act as weak learners.

In this paper, we present a flexible semiparametric model for interval-censored survival data using BART. Our approach models the hazard function of the survival times as a product of (a) a parametric baseline hazard which represents the "guess/center" of the actual hazard and (b) a nonparametric component modeled using SBART to account for the possible deviation from the guess/center model at (a) as well as complex time-dependent effects of covariates and cluster-specific random effects on hazard. We use a data augmentation scheme based on "thinning" a Poisson process (Adams *et al.*, 2009) to construct an efficient Markov chain Monte Carlo algorithm for sampling from the posterior distribution. The proposed strategy of using a generative model based on sampling from a model defined in terms of rejection sampling can also be used to construct models for conditional distribution estimation; see Li et al. (2020), who construct a BART-based conditional density estimation model using a model defined through rejection sampling from a baseline density. Our proposed model is applicable for left-censored, right-censored, and interval-censored survival data and our computational algorithms are highly scalable to obtain posterior survival prediction. Our simulation studies provide a comprehensive comparison of our proposed model with some existing models for interval-censored survival data. The rest of this paper is organized as follows. In Section 2, we provide a brief review of BART and SBART. In Section 3, we describe our proposed semiparametric model for survival analysis and its extension to censored data, and also provide the data augmentation algorithm used to carry out inference. In Section 4, we present an extensive simulation study to illustrate the performance of the proposed model and compare the results with those obtained from existing methods for survival prediction in interval-censored data. In Section 5, we illustrate our methods via analyzing the motivating study of the PSA data. We finally conclude by discussing our key findings in Section 6.

2 | A BRIEF REVIEW OF BART AND SBART

The BART framework, proposed by Chipman *et al.* (2010), is a popular Bayesian ensemble method which combines multiple "weak" decision trees into a single "strong" learner with high predictive accuracy. When the response surface is smooth, as is often the case in practice, BART can be improved theoretically and practically by using the SBART framework introduced by Linero and Yang (2018). SBART makes use of "soft" decision trees, and is capable of adapting to the unknown smoothness level of the response surface. We present here a brief review of BART and SBART.

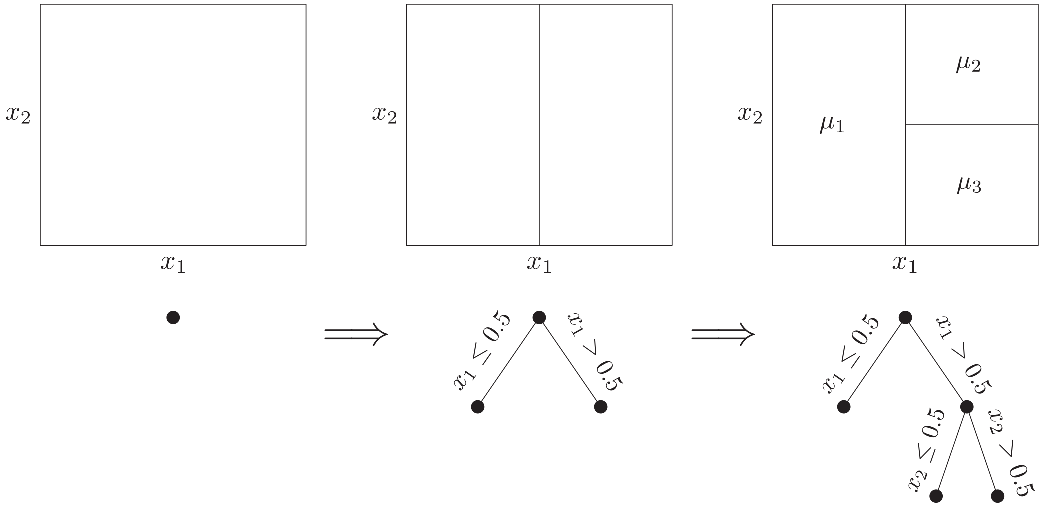

FIGURE 1 Schematic illustrating how to draw from the prior on $(\tau_t, \mathcal{M}_t)$. For this tree, we first determine that the root node will be a branch, which occurs with probability γ; we then sample the splitting coordinate $j = 1$ and the cutpoint $C = 0.5$. This process then iterates; the left child node is set to be a leaf node with probability $1 - \gamma/2^\beta$, and the right child is made a leaf with probability $\gamma/2^\beta$. Eventually this process terminates, and we sample a mean parameter μ for each leaf node.

Consider a nonparametric regression model $Y = f_0(\mathbf{x}) + \epsilon$ where $\mathbf{x}$ is a p-dimensional covariate vector, $f_0 : \mathbb{R}^p \to \mathbb{R}$ is an unknown regression function, and $\epsilon \sim N(0, \sigma^2)$ is a random error. BART models the unknown function $f_0(\mathbf{x})$ as a sum of T regression trees given by $f(\mathbf{x}) = \sum_{t=1}^T g(\mathbf{x}; \tau_t, \mathcal{M}_t)$, where τ_t denotes the topology and splitting rules of tree t and $\mathcal{M}_t = (\mu_{t1}, \dots, \mu_{tL_t})$ denotes the set of predictions associated with the L_t terminal nodes of the tth decision tree. Following Chipman *et al.* (2010), tree τ_t is assigned a branching process prior with each node at depth $k = 0, 1, 2, \dots$ being non-terminal with probability $q(k) = \gamma(1 + k)^{-\beta}$, where $\gamma > 0$ and $\beta > 0$ are positive hyperparameters that control the shape of the tree. For every branch node b, a splitting rule of the form $[x_j \leq C_b]$ is assigned with $\mathbf{x}$ going down left of the tree if the condition is satisfied and right down the tree otherwise. The splitting variable x_j is chosen uniformly from all the p available variables and C_b is assigned a Uniform prior on the set of the available splitting values. Independent Gaussian priors are designated to the terminal node parameters, with $\mu_{tl} \overset{iid}{\sim} N(0, \sigma_\mu^2)$. A schematic showing how the branching process prior generates a sample of a decision tree, and its associated partition, is given in Figure 1.

While BART is highly flexible and capable of capturing complex regression structures, estimates from BART essentially behave as a sum of step functions, and hence lack smoothness. Note that the function $g(\mathbf{x}; \tau_t, \mathcal{M}_t)$ can be written as the step function

$$g(\mathbf{x}; \tau_t, \mathcal{M}_t) = \sum_{\ell=1}^{L_t} \varphi_\ell(\mathbf{x}; \tau_t)\, \mu_{t\ell}, \tag{1}$$

where $\varphi_\ell(\mathbf{x}; \tau_t)$ is the indicator that $\mathbf{x}$ is associated to leaf ℓ of τ_t. Using BART to estimate $f(\mathbf{x})$ in this fashion would lead to non-smooth estimates, which is not desirable if $f(\mathbf{x})$ is believed to be smooth. SBART addresses this drawback by smoothing the weights assigned to each leaf, replacing the indicator function with

$$\varphi_\ell(\mathbf{x}; \tau_t) = \prod_{b \in A_t(\ell)} \psi(x_{j_b}; C_b, \alpha_b)^{1-R_b} \{1 - \psi(x_{j_b}; C_b, \alpha_b)\}^{R_b},$$

where $A_t(\ell)$ denotes the collection of *ancestor nodes* of leaf ℓ and R_b is the indicator that the path from the root to ℓ at branch b goes right. Here, $\psi(x; c, \alpha)$ is a continuous distribution function of a location-scale family with location c and scale α; in this paper, we set $\psi(x; c, \alpha)$ to be the inverse-logit $[1 + \exp\{-(x - c)/\alpha\}]^{-1}$. The parameter C_b plays the role of a cutpoint, while α functions as a bandwidth parameter, with $\alpha \to 0$ corresponding to the usual BART model. When the underlying true $f(\mathbf{x})$ is smooth, we expect that SBART will provide lower variance than BART while introducing negligible bias. Additionally, Linero and Yang (2018) show that SBART priors are capable of automatically detecting an appropriate amount of smoothness to use. In the next section, we describe the specifications of our flexible semiparametric model for survival prediction while utilizing this improved prediction performance of SBART to model underlying regression structures. SBART also allows for incorporating categorical covariates into the model. As with other BART implementations, SBART binarizes categorical predictors and groups the associated splitting probabilities (with the smoothing of SBART not having a strong consequence for binary variables).

3 | A SEMIPARAMETRIC MODEL FOR CLUSTERED SURVIVAL DATA

Let T_{ij} denote the survival time and $\mathbf{x}_{ij} = (x_{ij1}, \dots, x_{ijp}) \in \mathbb{R}^p$ be a vector of p covariates for subject $j = 1, \dots, n_i$ in cluster $i = 1, \dots, N$. Let f_{ij} and F_{ij} be the density and the cumulative distribution function (respectively) of the survival times T_{ij}, conditional on the parameters of the model and the random effects. We denote the associated survival function as $S_{ij}(t) = 1 - F_{ij}(t)$ and the hazard function as $\lambda_{ij}(t) = \frac{f_{ij}(t)}{S_{ij}(t)}$. In this section, we propose a flexible semiparametric model for survival prediction which is well equipped to capture the complex underlying associations among the different variables in the survival studies, while incorporating unobserved heterogeneity in the population due to random cluster effects. We model the conditional hazard function $\lambda_{ij}(t \mid W_i; \mathbf{x}_{ij})$ of the survival times T_{ij} given cluster-specific random frailty W_i as

$$\lambda_{ij}(t \mid W_i; \mathbf{x}_{ij}) = \lambda_0(t)\, W_i\, \Phi(l(t, \mathbf{x}_{ij})), \tag{2}$$

where $\Phi(\cdot)$ is the distribution function of a standard normal random variable and $\lambda_0(\cdot)$ is a parametric baseline hazard on which the model is centered; note, for example, that if $l(t, \mathbf{x}_{ij}) = 0$, then $\lambda_{ij}(t \mid W_i; \mathbf{x}_{ij}) = \lambda_0(t)\, W_i/2$, which is proportional to the baseline hazard $\lambda_0(t)$. The Gamma frailty $W_i \overset{ind}{\sim} \text{Gamma}(\eta, \eta)$ (Hugaard, 1995) accommodates within-cluster association, and the covariate effects are modeled in a time-varying nonparametric fashion using the SBART prior on $l(t, \mathbf{x}_{ij})$. Note that, due to how SBART is constructed, we have $\sup_{t,\mathbf{x}} |l(t, \mathbf{x})| < \infty$ almost-surely so that $\lambda(t \mid W_i; \mathbf{x}) \asymp W_i \lambda_0(t)$ almost-surely; this implies that $\int \lambda(t \mid W_i, \mathbf{x})\, dt = \infty$ if-and-only-if $\int \lambda_0(t)\, dt = \infty$, so that our model almost-surely defined a valid hazard function. Our nonparametric model for $l(t, \mathbf{x}_{ij})$ allows nonlinear effects and interactions among the covariates $\mathbf{x}_{ij}$ and the survival times t. The specification that the shape and rate of the frailty distribution are the same is made to ensure that $E(W_i) = 1$, which is required for identifiability (Hougaard, 1995). The amount of unknown heterogeneity among clusters (surgeons for the PSA study) due to random cluster effects is quantified by the variance $\text{Var}(W_i) = \eta^{-1}$ of the frailties, where a small value of η would indicate a large variability among different clusters in the population. As long as $\Lambda_0(t) = \int_0^t \lambda_0(s) ds \to \infty$ as $t \to \infty$, the marginal cumulative hazard function $H_{ij}(t|\mathbf{x}_{ij}) = \log S_{ij}(t|\mathbf{x}_{ij})$ resulting from the Gamma frailty model in (2) also converges to ∞ as $t \to \infty$ because the sample path of $l(s, \mathbf{x}_{ij})$ is bounded with probability 1. The appeal of our model in Equation (2) is that it is more semiparametric than usual Cox's proportional hazards model. Because of its semiparametric nature, we do not get the covariate effects as the time-constant hazards ratio of the Cox model that is more parametrically specified. However, we can still nonparametrically estimate the quantities of interest for our clinical collaborators such as the survival curves, median survival times at given values of covariates and time, and restricted mean survival time (RMST) and even relative risk ratio at a pre-specified time point of interest for given covariate values.

With interval-censored survival data we do not observe T_{ij}, but instead, we only record that T_{ij} is in the interval $(A_{ij}, B_{ij}]$, where $A_{ij} \leq B_{ij}$ are the two consecutive inspection times for the unit (i, j). We denote the observed interval-censored survival data as $\mathcal{D} = \{(\mathbf{x}_{ij}, A_{ij}, B_{ij}) : j = 1, \dots, n_i; i = 1, \dots, N\}$ where, $A_{ij} = B_{ij}$ when T_{ij} is uncensored, and $B_{ij} = \infty$ when T_{ij} is right censored at A_{ij}. Under the hazard function model in (2), the likelihood contribution from the unit (i, j), conditional on the unobserved random effect W_i, is

$$\begin{aligned}
&\Pr[T_{ij} \in (A_{ij}, B_{ij}] \mid W_i; \mathbf{x}_{ij}] \\
&\quad = S_{ij}(A_{ij} \mid W_i;\, \mathbf{x}_{ij}) - S_{ij}(B_{ij} \mid W_i;\, \mathbf{x}_{ij}), \\
&\text{where } S_{ij}(t \mid W_i;\, \mathbf{x}_{ij}) \\
&\quad = \exp\left\{ -W_i \int_0^t \lambda_0(s)\, \Phi(l(s, \mathbf{x}_{ij}))\, ds \right\}
\end{aligned} \tag{3}$$

is the conditional survival function of T_{ij} given frailty W_i. Using the conditional independence of the within-cluster survival times given W_i and across cluster independence of all survival times, the overall likelihood based on all N clusters is then derived as

$$\prod_{i=1}^{N} \int_0^\infty \left\{ \prod_{j=1}^{n_i} \Pr[T_{ij} \in (A_{ij}, B_{ij}] \mid W_i; \mathbf{x}_{ij}] \right\} g(W_i \mid \eta)\, dW_i, \tag{4}$$

after integrating out W_i with respect to its Gamma(η, η) density denoted by $g(\cdot \mid \eta)$ in (4). To fully specify the hierarchical Bayesian formulation of our model, we specify prior distributions of our model parameters as $p(\eta, \Omega, l, \Psi) \propto p(\eta) \times p(\Omega) \times p(l \mid \Psi) \times p(\Psi)$, where $p(\eta)$ is the marginal prior of the shape (as well as the scale) parameter η of the Gamma(η, η) frailty density, $p(\Omega)$ is the prior on the set of unknown parameters Ω associated with the parametric baseline hazard $\lambda_0(\cdot)$, and $p(l|\Psi)$ and $p(\Psi)$ are the priors associated with the SBART model and its hyperparameters.

In all of our examples, we use a default prior for $\Psi = (\sigma_\mu, \gamma, \beta, r_\alpha)$ which takes $\sigma_\mu = 3/(k\sqrt{m})$, $(\gamma, \beta) = (0.95, 2)$, and $r_\alpha = 10$ where $\alpha_t \sim \text{Gamma}(1, r_\alpha)$ in the SBART prior, with the same bandwidth α_t being used for all branches within a single tree. We use a Gamma(a, b) prior for η, with $a, b > 0$ carefully chosen to reflect our prior opinion about the extent to which the cluster effects W_i can affect the underlying hazard. By default we set $k = 2$.

We fit this model using Markov chain Monte Carlo; the primary challenge in implementing the model (4) with the model specified above is computational, as even evaluating (4) requires numerically evaluating integrals of the form $\int_0^t \lambda_0(s)\,\Phi(l((s, \mathbf{x}))\,ds$. As such, it is difficult to construct efficient Markov chain Monte Carlo algorithms for sampling from the posterior distribution by working directly with (4). We address this by extending the method of Adams et al. (2009) and Fernandez *et al.* (2016) to augment the T_{ij}'s by viewing them as the first "accepted" point of a thinned Poisson process; this will allow us to obtain simple updates for the distributions for all of the model parameters. We elaborate on the implementation of this data augmentation (DA) scheme for uncensored survival times in Section 3.1 and extend it to left-censored, right-censored, and interval-censored survival times in Section 3.2.

3.1 | Data augmentation scheme for no censoring

From the conditional nonparametric hazard function in (2) and the corresponding survival function (3), it is apparent that the evaluation of the likelihood contribution

$$L_{ij} = W_i\,\lambda_0(T_{ij})\,\Phi(l(T_{ij}, \mathbf{x}_{ij})) \times \exp\left[-\int_0^{T_{ij}} W_i\,\lambda_0(s)\,\Phi(l(s, \mathbf{x}_{ij}))\,ds\right], \qquad (5)$$

for each uncensored T_{ij} case given W_i requires integrating a time-dependent function $\lambda_{ij}(s \mid W_i; \mathbf{x}_{ij}) = \lambda_0(s)\,W_i\,\Phi(l(s, \mathbf{x}_{ij}))$; this involves a nonparametrically modeled function $l(s, \mathbf{x}_{ij})$, which varies by the subjects (i, j). Having this time-varying $l(s, \mathbf{x}_{ij})$ inside the integral in (5) also prevents us from applying the usual Bayesian backfitting algorithm of Chipman *et al.* (2010) to update $l(s, \mathbf{x}_{ij})$ given the rest of the parameters. To address these challenges, we use a data augmentation procedure, which introduces additional latent variables $\{G_{ijk}$ for $k = 1, \dots, m_{ij}\}$ to ease the posterior computation (that is, by assuring either closed form or standard updates for the conditional posteriors of the parameters). These random latent variables are generated from a non-homogeneous Poisson process (NHPP) with intensity

$$\lambda_0(s)\,W_i\,\{1 - \Phi(l(s, \mathbf{x}_{ij}))\} \qquad (6)$$

in the interval $(0, T_{ij})$ (Ibrahim *et al.*, 2012). This process can be sampled by first simulating a Poisson process with intensity function $\lambda_0(s)\,W_i$ on the interval $(0, T_{ij})$ and then "thinning" each point with probability $\Phi(l(s, \mathbf{x}_{ij}))$. Note that the likelihood of the augmented $\mathbf{G}_{ij}$'s simulated via this NHPP is given by

$$\begin{aligned}&\Pr(\text{events at } G_{ijk} \text{ for } k\\ &\quad = 1, \dots, m_{ij}, \text{ and no other events in } (0, T_{ij}))\\ &\quad = \left[\prod_{k=1}^{m_{ij}} \lambda_0(G_{ijk})\,W_i\,(1 - \Phi(l(G_{ijk}, \mathbf{x}_{ij}))\,dG_{ijk}\right]\\ &\quad\quad \times e^{-W_i \int_0^{T_{ij}} \lambda_0(s)(1-\Phi(l(s,\mathbf{x}_{ij})))ds}. \end{aligned} \qquad (7)$$

Combining this conditional likelihood given T_{ij} with the likelihood of T_{ij}, we get

$$\begin{aligned}&\lambda_0(T_{ij})\,W_i\,\Phi(l(T_{ij}, \mathbf{x}_{ij}))\\ &\quad \times \prod_{k=1}^{m_{ij}} \lambda_0(g_{ijk})\,W_i\,\Phi(Z_{ijk} \mid l(G_{ijk}, \mathbf{x}_{ij})) \times e^{-W_i \Lambda_0(T_{ij})},\end{aligned} \qquad (8)$$

where $\Lambda_0(t) = \int_0^t \lambda_0(s)\,ds$ denotes the baseline cumulative hazard function of the survival times. This likelihood is greatly simplified because it removes the integral $\int_0^{T_{ij}} \lambda_0(s)\,\Phi(l(s, \mathbf{x}_{ij}))\,ds$ by adding to it $\int_0^{T_{ij}} \lambda_0(s)\{1 - \Phi(l(s, \mathbf{x}_{ij}))\}\,ds$, cancelling the $\Phi(l(s, \mathbf{x}_{ij}))$ term. While we have phrased this data augmentation directly in terms of simulating $\mathbf{G}_{ij}$ from a distribution which simplifies the integral, it is also possible to view this as a traditional data augmentation algorithm by viewing T_{ij} as the first "accepted" point from a "thinned" Poisson process with intensity $\lambda_0(s)\,W_i$ and acceptance probability $\Phi(l(s, \mathbf{x}_{ij}))$, with $\mathbf{G}_{ij}$ being the collection of "rejected" points. From this perspective, we are simulating $\mathbf{G}_{ij}$ from its full conditional distribution, and (8) gives the joint likelihood of the accepted and the rejected points (Fernandez *et al.*, 2016).

The augmented likelihood in (8) is still not ideal because the Bayesian backfitting algorithm of Chipman *et al.* (2010) requires that the likelihood of $l(t, \mathbf{x})$ takes the form of a semiparametric Gaussian regression

$$Z = l(t, \mathbf{x}) + \epsilon \qquad (9)$$

with $\epsilon \sim N(0, \sigma^2)$; instead, $l(t, \mathbf{x})$ enters the likelihood through the probit terms $\Phi(l(T_{ij}, \mathbf{x}_{ij}))$ and $1 - \Phi(l(G_{ijk}, \mathbf{x}_{ij}))$. To accommodate this, we perform another layer of data augmentation using the strategy of Albert and Chib (1993). We introduce truncated normal latent variables $\{Z_{ijk}; k = 1, \ldots, m_{ij} + 1\}$ such that

$$Z_{ijk} \sim \begin{cases} N(l(G_{ijk}, \mathbf{x}_{ij}), 1) I(-\infty, 0), & \text{if } k = 1, \ldots, m_{ij} \\ N(l(T_{ij}, \mathbf{x}_{ij}), 1) I(0, \infty), & \text{if } k = m_{ij} + 1. \end{cases}$$

After augmenting the Z_{ijk}'s to the model, we will use the joint likelihood

$$\prod_{i=1}^{N} \prod_{j=1}^{n_i} \left\{ \lambda_0(T_{ij}) W_i \, N(Z_{ij(m_{ij}+1)} \mid l(T_{ij}, \mathbf{x}_{ij}), 1) \right.$$
$$\left. \times \prod_{k=1}^{m_{ij}} \lambda_0(G_{ijk}) W_i \, N(Z_{ijk} \mid l(g_{ijk}, \mathbf{x}_{ij}), 1) \times e^{-W_i \Lambda_0(T_{ij})} \right\},$$

where $N(x \mid \mu, \sigma^2)$ denotes the density of a normal distribution with mean μ and variance σ^2. This allows us to utilize the continuous outcome SBART model described by Linero and Yang (2018) by treating the Z_{ijk}'s as Gaussian responses with variance $\sigma^2 = 1$. That is, we apply the Bayesian backfitting approach described by Linero and Yang (2018) to the model $Z_{ijk} = l(G_{ijk}, \mathbf{x}_{ij}) + \epsilon_{ijk}$ and $\epsilon_{ijk} \sim N(0, 1)$, for $k = 1, \ldots, m_{ij}$, and similarly with $Z_{ijk} = l(T_{ij}, \mathbf{x}_{ij}) + \epsilon_{ijk}$ for $k = m_{ij} + 1$.

In addition to updating the function $l(t, \mathbf{x}_{ij})$ using this data augmentation approach, we must also update the frailties W_i, the parameter of the frailty density η, the parameters Ω of the baseline hazard $\lambda_0(\cdot)$, and the hyperparameters Ψ of the SBART model. For the Gamma(η, η) distributed frailties W_i, the likelihood is proportional to

$$W_i^{n_i + \sum_j m_{ij}} \, e^{-W_i \sum_j \Lambda_0(T_{ij})} \tag{10}$$

and the corresponding conjugate conditional posterior is Gamma$(\eta + n_i + \sum_j m_{ij}, \ \eta + \sum_j \Lambda_0(T_{ij}))$ where n_i is the size of the ith cluster. For the remaining parameters, we use slice sampling (Neal, 2003). In summary, our data augmentation scheme is based on the following hierarchical specification of the the model, where each line is conditioned on all of the lines above it and all terms within each line are conditionally independent:

$$(\eta, \Psi, \Omega) \sim p(\eta)\, p(\Psi)\, p(\Omega)$$

$$l \sim \text{SBART}(\Psi)$$

$$W_i \sim \text{Gamma}(\eta, \eta)$$

$$\text{Data: } T_{ij} \sim \lambda_{ij}(t \mid \mathbf{x}_{ij}, l, W_i, \Omega)$$

$$\exp\left\{ -\int_0^t \lambda_{ij}(s \mid \mathbf{x}_{ij}, l, W_i, \Omega)\, ds \right\}$$

$$\{G_{ijk} \ : \ k = 1, \ldots, m_{ij}\} \sim \text{NHPP}\{\lambda_0(t)\, W_i\, (1 - \Phi(l(t, \mathbf{x}_{ij})\}$$

ALGORITHM 1 Inference algorithm with clustered and uncensored survival times T_{ij}

```
Input: Observed survival times $T_{ij}$, and initiated values of $\Omega$
       (baseline hazard parameter), $l$ (SBART), $\Psi$
       (parameters of leaf $l$), $W_i$ (frailty variables) and $\eta$
       (shape parameter of the frailty density)
for $m = 1, \ldots, S$ do
  for $i = 1, \ldots, n$ do
    for $j = 1, \ldots, n_i$ do
      $q_{ij} \sim$ Poisson$(1; \Lambda_0(T_{ij} W_i)$
      $C_{ij} \sim$ Uniform$(q_{ij}; 0, \Lambda_0(T_{ij}) W_i)$
      $\tilde{G}_{ij} \leftarrow \Lambda_0^{-1}(\frac{C_{ij}}{W_i})$
      $\mathbf{U}_{ij} \sim$ Uniform$(q_{ij}; 0, 1)$
      $\mathbf{G}_{ij} \leftarrow \{\tilde{G}_{ij} : U_{ij} \leq 1 - \Phi(l(\tilde{G}_{ij}, \mathbf{x}_{ij}))\} =$
        $\{G_{ij1}, \ldots, G_{ijm_{ij}}\}$
      $Z_{ijk} \sim$
        $\begin{cases} N(l(G_{ijk}, \mathbf{x}_{ij}), 1) I(-\infty, 0), & \text{if } k = 1, \ldots, m_{ij}, \\ N(l(T_{ij}, \mathbf{x}_{ij}), 1) I(0, \infty), & \text{if } k = m_{ij} + 1. \end{cases}$
    end for
    Update $\Omega$ by slice sampling
    Update $l$ using the Bayesian backfitting algorithm
      of Linero and Yang (2018)
    Update $\Psi$ by slice sampling
    Update $\eta$ by slice sampling
    Update $W_i : i = 1, \ldots, n$ by sampling
      $W_i \sim$ Gamma$(\eta + n_i + \sum_j m_{ij}, \eta + \sum_j \Lambda_0(T_{ij}))$
  end for
end for
```

for $t \in (0, T_{ij})$

$$Z_{ijk} \sim \begin{cases} N(l(G_{ijk}, \mathbf{x}_{ij}), 1) I(-\infty, 0), & \text{if } k = 1, \ldots, m_{ij} \\ N(l(T_{ij}, \mathbf{x}_{ij}), 1) I(0, \infty), & \text{if } k = m_{ij} + 1. \end{cases}$$

The steps of the data augmentation algorithm are summarized in Algorithm 1, which also shows how to sample from the NHPP distribution for $\mathbf{G}_{ij}$.

While Algorithm 1 provides the steps of inference and parameter updates with clustered and uncensored survival times T_{ij}, it does not account for the possibility of censoring. In the next sub-section, we discuss the extension of our model for clustered and (left, right, or interval) censored survival times.

3.2 | Augmenting T_{ij} for interval censoring

We now consider the case where T_{ij} is either right or interval censored. Suppose first that T_{ij} is censored at A_{ij}, so that we observe the event $[T_{ij} \geq A_{ij}]$. Assuming that the censoring mechanism is non-informative (Kalbfleisch and Prentice, 2002) the targeted likelihood contribution of the event $[T_{ij} \geq A_{ij}]$ is

$$L^*_{ij} = \exp\left[-\int_0^{A_{ij}} \lambda_0(s)\, W_i\, \Phi(l(s, \mathbf{x}_{ij}))\, ds\right]. \tag{11}$$

If we augment $\mathbf{G}_{ij}$ as before from the NHPP with intensity $\lambda_0(s)\, W_i\, \{1 - \Phi(l(s, \mathbf{x}_{ij}))\}$ then the corresponding augmented data likelihood contribution is not given by (8), but instead by

$$\begin{aligned} &L^*_{ij} \times \Pr[\text{“thinned}'' \text{events at } \mathbf{G}_{ij};\ \text{no other}'' \\ &\text{thinned}'' \text{events in} (0, A_{ij}]\}] \\ &= \prod_{k=1}^{m_{ij}} \{\lambda_0(G_{ijk})\, W_i\, (1 - \Phi(l(G_{ijk}, \mathbf{x}_{ij}))\, dG_{ijk}\} \\ &\times e^{-W_i \Lambda_0(A_{ij})}. \end{aligned} \tag{12}$$

Hence, we can apply the same data augmentation algorithm as before, where the censored observations contribute (12) to the likelihood. To remove the probit terms from the likelihood, we now only simulate $Z_{ijk} \sim N(l(G_{ijk}, \mathbf{x}_{ij}), 1) I(-\infty, 0)$ and do not simulate a latent variable for $k = m_{ij} + 1$. For the right-censored setting, let $Y_{ij} = T_{ij}$ if T_{ij} is observed and A_{ij} otherwise. Then the resulting likelihood after performing both layers of data augmentation is

$$\begin{aligned} &\prod_{i=1}^{N} \prod_{j=1}^{n_i} \left\{\lambda_0(Y_{ij})\, W_i\, N(Z_{ij(m_{ij}+1)} \mid l(Y_{ij}, \mathbf{x}_{ij}), 1)\right\}^{I(Y_{ij} = T_{ij})} \\ &\times \prod_{k=1}^{m_{ij}} \lambda_0(G_{ijk})\, W_i\, N(Z_{ijk} \mid l(G_{ijk}, \mathbf{x}_{ij}), 1) \\ &\times e^{-W_i \Lambda_0(Y_{ij})}. \end{aligned}$$

We address interval censoring between $(A_{ij}, B_{ij}]$ by augmenting the true survival time T_{ij}; this also addresses left censoring by setting $A_{ij} = 0$. After augmenting survival time T_{ij}, we augment $\mathbf{G}_{ij}$ using the same approach as for when T_{ij} is observed. The augmented uncensored survival time T_{ij} can be generated as the smallest event time of a NHPP with intensity $\lambda_0(t)\, W_i\, \Phi(l(t, \mathbf{x}_{ij}))$ on the interval $(A_{ij}, B_{ij}]$ *conditional on there being at least one event*. An algorithm for augmenting T_{ij} is given by Algorithm 2. The time T_{ij} is sampled by first running a Poisson process with intensity $\lambda_0(t)\, W_i$ on $(A_{ij}, B_{ij}]$ and then accepting each point u with probability $\Phi(l(u, \mathbf{x}_{ij}))$; this process is then repeated until we have at least one acceptance. To make the algorithm more efficient, we condition the initial draw from $\lambda_0(t)\, W_i$ on $(A_{ij}, B_{ij}]$ to have at least one point. It is important to note here, that the imputation step, as described in Algorithm 2, has to be repeated at the beginning of each iteration, for each subject in the dataset, with left-censored or interval-censored survival times.

ALGORITHM 2 Imputation of the survival time for subject j in cluster i with interval-censored observations $(A_{ij}, B_{ij}]$

```
$m^*_{ij} \leftarrow 0$
while $m^*_{ij} = 0$ do
    $q^*_{ij} \sim \text{Poisson}(1; W_i(\Lambda_0(B_{ij}) - \Lambda_0(A_{ij})))I(q^*_{ij} > 0)$
    $\mathbf{C}^*_{ij} \sim \text{Uniform}(q^*_{ij}; \Lambda_0(A_{ij})W_i, \Lambda_0(B_{ij})W_i)$
    $\tilde{\mathbf{G}}_{ij} = \Lambda_0^{-1}(\frac{\mathbf{C}^*_{ij}}{W_i})$
    $\mathbf{U}^*_{ij} \sim \text{Uniform}(q^*_{ij}; 0, 1)$
    $\mathbf{G}^*_{ij} = \{\tilde{\mathbf{G}}_{ij} : U^*_{ij} \leq \Phi(l(\tilde{\mathbf{G}}_{ij}, \mathbf{x}_{ij}))\}$
end while
Impute $T^*_{ij}$ as $t^*_{ij} = \min\{v : v \in \mathbf{G}^*_{ij}\}$
```

The algorithms discussed in this section are computationally fast. For example, in our simulation study in Section 4 with clustered interval-censored survival data (under setting D) with 10 clusters and fixed cluster size 10, the algorithms took 0.16 s for each iteration using unoptimized `R` code to perform the data augmentation. Similarly, for the PSA data analysis example in Section 5 with 597 total observations and 9 clusters, the algorithm took only 0.22 s for each iteration (all computations were performed on a laptop with Intel(R) core i7 processor and 8GB RAM).

3.3 | Computational considerations

In addition to giving very strong performance across a variety of different problems, methods based on BART are usually more computationally efficient than competing Bayesian nonparametric methods. Several extremely fast implementations of BART exist, including the `dbarts` and `bartMachine` packages available on `CRAN`. A single iteration of the Bayesian backfitting algorithm takes $O(NT)$ computations to compute a Metropolis–Hastings ratio (each observation is dropped down each tree). This is a substantial improvement computationally over the method of Fernandez *et al.* (2016), who use a crude/restrictive Gaussian process approximation to avoid an $O(N^3)$ matrix inversion.

Relative to standard BART models, the main computational difficulty of our proposed approach lies in the size of the augmented dataset, as the $O(NT)$ complexity per iteration becomes $O(N_{\text{aug}} T)$ where N_{aug} is $\sum_i n_i + \sum_j m_{ij}$

(assuming no censoring). Hence, the more data that we augment, the slower the computations. While this is not pursued here, this gives additional motivation for choosing a possibly covariate-dependent choice of baseline hazard $\lambda_0(t \mid x)$, as if this is a good approximation of the true baseline hazard then we will not need to augment large datasets. An additional computational bottleneck for our approach occurs due to our use of soft trees, which do not permit the same computational shortcuts that standard BART models do. In future work, we hope to develop faster SBART algorithms.

3.4 | The choice of the Baseline hazard

To this point, we have considered using only an exponential or Weibull specification for the baseline hazard $\lambda_0(t)$. Subsequent work by the authors (Linero *et al.*, 2021) gives (among other things) an in-depth examination of the role of the baseline hazard, and develops *covariate-dependent* choices of the baseline hazard $\lambda_0(t \mid \mathbf{x})$. A particularly flexible choice of $\lambda_0(t \mid \mathbf{x})$ they study takes the baseline hazard function *itself* to be modeled using BART. This allows the user, for example, to shrink our nonparametric model toward a semiparametric Weibull proportional hazards model, or toward a semiparametric Cox proportional hazards model with a flexibly modeled baseline hazard.

A poor choice of $\lambda_0(t \mid \mathbf{x})$ has two consequences for our methods. First, if the shape of $\lambda_0(t \mid \mathbf{x})$ does not match the data well, then the model will compensate by having larger thinning probabilities. This has the consequence of introducing more rejected points into the MCMC scheme, causing inefficient inference. Second, a poor choice of $\lambda_0(t \mid \mathbf{x})$ results in less statistically efficient inferences than would be obtained with a good choice.

4 | SIMULATION STUDY

In this section, we illustrate the applicability of our proposed semiparametric model for survival prediction in situations when the regression relationships between the survival times and the covariates are nonlinear and complex. We compare the performance of our model with some of the existing survival regression models with readily available computational packages under four different simulation settings:

(1) Simulation A: Survival times are independent and uncensored.
(2) Simulation B: Survival times are clustered, but uncensored.
(3) Simulation C: Survival times are independent, but interval censored.
(4) Simulation D: Survival times are both clustered as well as interval censored (thus mimicking the motivating post-surgery PSA recurrence study).

For each of the above simulation settings, we generate $M = 20$ replicates of training datasets, each with $n = 100$ subjects, with a subject-specific five-dimensional covariate vector $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5)$ drawn from a unit hypercube. We first describe the simulation model we use to generate independent and uncensored survival times (for Setting A) and then proceed to describe the extension of this data generation scheme to simulate clustered or/and interval-censored survival times (for Settings B–D).

Data generation scheme: We use the common test function

$$f_0(\mathbf{x}) = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5$$

that includes nonlinearity and interaction effects among five covariates $(x_1, \ldots, x_5)$. The function f_0, proposed initially by Friedman (1991), has nonlinear dependence on the first three variables x_1, x_2 and x_3, linear dependence on x_4 and x_5, and incorporates a nonlinear interaction between x_1 and x_2. For the ith subject under simulation A, independent survival times T_i are generated from a Gamma distribution with shape parameter $\lambda_i = f_0(\mathbf{x}_i)$ and rate parameter $= 6$. The observed independent survival data under setting A is $\{\mathbf{x}_i, T_i : i = 1, \ldots, n\}$. Predicted survival probabilities are evaluated at a grid of 10 time points, for each of $n^* = 100$ independent subjects in a test dataset generated in the same fashion as the training data. Under simulation setting B, each of our $M = 20$ replicated training datasets consists of $N = 10$ clusters, with fixed cluster size $n_i = 10$. We simulate a cluster-specific random effect $W_i \sim Uniform(0, 0.2)$ and simulate the survival time for the (i, j)th subject as $T_{ij} \sim \text{Gamma}(\lambda_{ij}, 6)$, where $\lambda_{ij} = f_0(\mathbf{x}_{ij}) + W_i$. Under simulation setting B, we observe $\{\mathbf{x}_{ij}, T_{ij} : i = 1, \ldots, N, j = 1, \ldots n_i\}$.

From the independent survival times T_i (generated as in setting A), we obtain interval-censored survival times $(A_i, B_i]$ (for simulation C) as follows:

- Generate $K_i \sim \text{Poisson}(T_i)$.
- Set $A_i = I(K_i = 0) + V_i^{\frac{1}{K_i}} \times T_i \times I(K_i > 0)$, where $V_i \sim Uniform(0, 1)$ and $I(\cdot)$ is an indicator function.
- Set $B_i = T_i + R_i$, where $R_i \sim \text{Exponential}(1)$.

This ensures that the censoring mechanism is non-informative. Thus, each replicate of the simulated interval-censored survival dataset under setting C is $\{\mathbf{x}_i, A_i, B_i :$

$i = 1, \ldots, n\}$. Finally, clustered and interval-censored survival times $\{(A_{ij}, B_{ij}]; i = 1, \ldots N, j = 1, \ldots n_i\}$ (for simulation D) are obtained following the same interval censoring mechanism as above, except now, the censoring is introduced after simulating the clustered survival times T_{ij} (as was obtained for simulation setting B).

Evaluation of model parameters: The proposed semiparametric model for survival prediction is fitted to the simulated data in each setting, using a constant baseline hazard $\lambda_0(t) = \Omega$, that is, assuming that the baseline distribution of the survival times is exponential with rate Ω. We assign Ω a conjugate and relatively non-informative Gamma prior with the shape and the rate hyperparameters estimated from the data. For analyzing clustered survival times (as in simulation settings B and D), the cluster-specific random effects W_i are assumed to follow a Gamma distribution with equal shape and rate parameter η. Note that the amount of unobserved heterogeneity in the population due to the random cluster effect is quantified by the variance η^{-1} of the frailty W_i, where a smaller value of η indicates a larger cluster effect on the hazard function. However, it is reasonable to believe that the multiplicative effect of the frailties on the estimated hazard function will likely be within 25%. This is ensured by assigning η a Gamma(4, 0.01) prior, which allows a probability of less than 0.5% for the frailties to influence the hazard function by more than 25%. A brief analysis of the model sensitivity to different values of the hyperparameter η has been discussed in Section 5. The SBART component of our model (l) uses a default prior described by Linero and Yang (2018) with 50 trees; we find that this default performs reasonably for the survival setting.

We compare our proposed SBART-based semiparametric model for survival prediction with parametric AFT models, Bayesian generalized AFT model (Zhou *et al.*, 2017), frequentist proportional hazards model, random survival forests, and a semiparametric Bayesian proportional hazards model. The parametric AFT models have been fitted using Weibull distributions for the survival times. The Bayesian generalized AFT model (Zhou *et al.*, 2017) as implemented by the `frailtyGAFT` function in the R package `SpBayesSurv` allows for a non-proportional hazards model with arbitrary censoring, and the semiparametric proportional hazards model (Zhou *et al.*, 2020) as implemented by `SpBayesSurv` package in R uses the transformed Bernstein polynomial for fitting the baseline hazard functions and independently distributed Gaussian frailties to model the within-cluster association for the clustered survival data (for simulations B and D). Another related model which we could have considered for comparison is the proportional hazards model of Henschel *et al.* (2009), which can be fit in the `spatsurv`, `spatsurv`, and `R2BayesX` packages.

All Bayesian models were fitted using 2500 burn-in and 2500 sampling iterations. Monte Carlo approximation of the root mean squared error (RMSE) of survival prediction of the $j = 1, 2, \ldots, M = 20$ replicate is obtained as

$$R\hat{M}SE_j = \sqrt{\frac{\sum_{i=1}^{n^*} \sum_{g=1}^{10} \{S(t_g \mid \mathbf{x}_i) - \hat{S}(t_g \mid \mathbf{x}_i)\}^2}{n^* \times 10}},$$

where $S(t_g \mid \mathbf{x}_i)$ and $\hat{S}(t_g \mid \mathbf{x}_i)$ are, respectively, the true and the predicted survival probabilities at time point t_g for the ith subject of the test dataset, having covariate vector $\mathbf{x}_i$. For the Bayesian procedures, $\hat{S}(t_g \mid \mathbf{x}_i)$ is taken to be the posterior mean of $S(t_g \mid \mathbf{x}_i)$. Performances of the different methods are compared on the basis of the average RMSE, obtained from the 20 replicated simulations and the results are reported in Table 1.

As seen from Table 1, among the methods considered, our proposed model performs the best in predicting survival functions in terms of the average RMSE. It has also been shown that our model is easily implementable with clustered and/or interval-censored survival data. It is worth mentioning here that we found a number of computational packages in R, (e.g., `parfm`, `frailtypack`) which, although have been built for survival regression models, give frequent issues with parameter convergence and likelihood optimization, especially when the regression relationships among the survival times and the other covariates are complex and nonlinear.

5 | APPLICATION: PROSTATE SURGERY STUDY

A practical example of interval-censored clustered survival data is any medical follow-up study to to compare the survival functions of the time to prostate specification antigen (PSA) recurrence for competing surgery techniques to remove prostate (Barbash and Glied, 2010). PSA level in the blood, immediately following surgery, is 0.0 ng/ml; time to PSA recurrence is defined as the time after surgery when the PSA level in the blood exceeds 0.2 ng/ml. Since continuous monitoring of PSA is not feasible, and the level of PSA in the blood can only be determined via blood tests (typically every 3 to 6 months after surgery), the exact time to PSA recurrence is often interval censored between consecutive post-surgery visits to the clinic. Also, these survival times are typically clustered within each surgeon/clinic. The main analysis goal of this study is to compare surgery technique groups as well as possible baseline characteristics such as (i) age at surgery (continuous); (ii) number of positive cores (continuous, range 0–17); (iii) Gleason score regarding how the cancer cells are arranged

TABLE 1 Simulation results based on 20 replicates of data comparing Monte Carlo estimates of average root mean squared error (RMSE) obtained from fitting the parametric accelerated failure time (AFT) model with a Weibull distribution for the log survival times, a random survival forest, a Bayesian Generalized AFT model (Zhou *et al.*, 2017), frequentist proportional hazards (PH) model, the proposed semiparametric survival model based on SBART while assuming Gamma frailty and a constant baseline hazard function and a semiparametric Bayesian PH model with Gaussian frailties, and transformed Bernstein polynomials for the baseline hazard

Methods	Simulation A (independent, uncensored)	Simulation B (independent, interval-censored)	Simulation C (clustered uncensored)	Simulation D (clustered interval censored)
Parametric AFT	0.7785	0.8793	0.4366	—
Bayesian generalized AFT	0.1343	0.1198	0.1343	0.1054
Random survival forest	0.1798	—	—	—
Frequentist PH	0.1621	—	0.1313	—
Semiparametric survival model with SBART	0.1038	0.1106	0.1063	0.0944
Semiparametric Bayesian PH	0.1403	0.1229	0.1418	0.1211

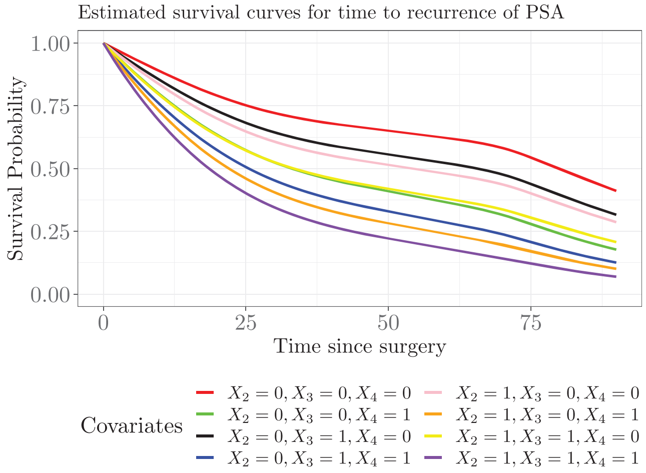

FIGURE 2 Estimated survival curves for time to PSA recurrence for a subject who had "more aggressive" prostate cancer (Gleason score = 1) with three positive cores and had undergone a prostate removal surgery at 64 years of age. Covariates shown in the legend correspond to one of two surgical techniques, surgical margin (positive or negative) and pT stage (spread or not spread). This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

in the prostate (less aggressive or more aggressive cancers); (iv) surgical margin (negative or positive); and (v) pT stage (not spread or spread). However, due to some errors in the coding of some of the confounding variables and due to the confidentiality issues, the results of the analysis of the original study cannot be presented at this point. To circumvent this problem, for illustrative purposes and to allow replication of the results using our method, we have created a simulated dataset designed to replicate many of the salient features of the original study. This simulated dataset preserves the number of patients ($n = 597$), number and average sizes of clusters (approximately 9 and 66), and number and types of covariates, as well as the performance of our approach relative to competing methods in terms of LPML (our procedure performs better on the real study relative to other methods than it does on the simulated data). However, because it is a simulated dataset and the original dataset had coding errors in it, we must emphasize that the analysis results of this paper cannot be interpreted as either the observed or estimated effects of the types of surgery. The simulated data does not preserve the effect sizes, and we have blinded all of the confounding and treatment variables with the labels $X_1, \ldots, X_6$.

Figure 2 shows the estimated survival curves of times to recurrence of PSA for patients with varying values of the covariates $X_2, \ldots, X_4$ (corresponding, in some order, to surgery technique, surgical margin, and pT stage) with $X_1 \equiv 1$ and the two continuous covariates X_5 and X_6 fixed at their median values. We see that the highest survival times are obtained when each of the is set to 0, with reasonable consistent drops in survival probability at each time as we change X_j from 0 to 1.

To assess more fully the effect of each covariate and evaluate potential interactions, we analyze the effect of surgery technique and other covariates on the time to PSA recurrence using the difference in RMST caused by varying the covariate values. RMST) at a pre-specified time point of interest, τ is defined as $\text{RMST}(\tau) = E[\min(T, \tau)] = \int_0^{\tau} S(u)\, du$, where T is a non-negative random variable representing the time to event of a particular individual and $S(\cdot)$ is the associated survival function. Inference based on RMST has recently gained popularity, especially due to its ability to summarize survival profiles with non-proportional hazards. Figure 3 shows the differences in RMST to PSA recurrence for different combinations of three binary covariates (surgery technique, surgical margin, and pT stage) for a patient who had a "more aggressive" prostate cancer (Gleason score = 1) withthree positive

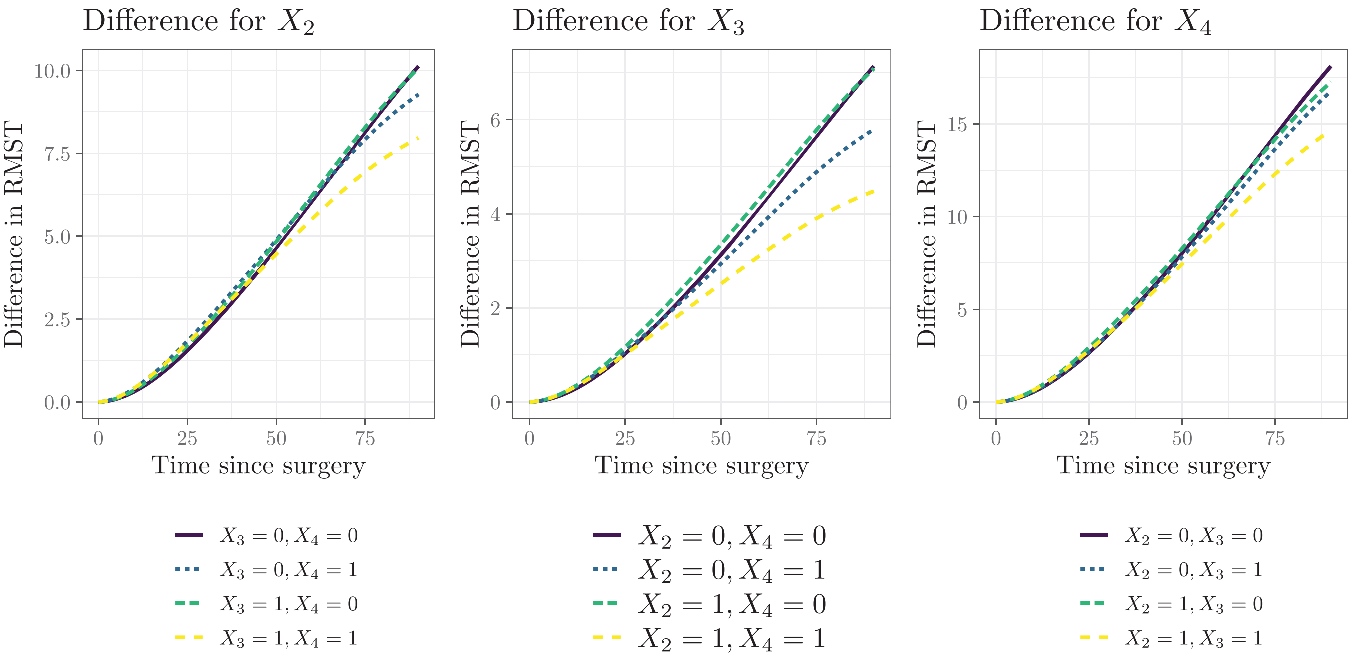

FIGURE 3 Difference in restricted mean survival time (RMST) (measured in months) for time to PSA recurrence due to surgery techniques (technique 1–technique 0), surgical margin (negative–positive), and pT stage (not spread–spread) (from left to right) for a patient who had a "more aggressive" prostate cancer (Gleason score = 1) with three positive cores and had undergone a prostate removal surgery at 64 years of age. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

cores and had undergone a prostate removal surgery at 64 years of age. The variable with the weakest effect on RMST is X_3, which is always estimated to be less than 7 months; the other variables have maximal effects ranging from 10 to 18 months. For variables X_2 and X_4, the difference in RMST profiles does not appear to interact strongly with the other variables, whereas for X_3 we see that the profile differs across levels of X_2 and X_4. It should be noted that, for the sake of brevity, here we present the estimated survival functions for only certain specific covariate combinations, however, similar analysis and detailed interpretations can be obtained for all other covariates of interest.

Our SBART-based semiparametric model, when fitted to the PSA data set yields the maximum LPML value (−461.39) as compared to Zhou et al's (2017) generalized AFT model with non-proportional hazards (LPML: −480.19), a semiparametric proportional hazards model (LPML: −471.81), a semiparametric proportional odds model (LPML: −473.35) with a transformed Bernstein polynomial for the baseline hazard function, and a semiparametric AFT model (LPML: −478.32). These competing methods were fit using the spBayesSurv package in R. This further establishes the goodness of fit of our proposed model for analyzing clustered interval-censored survival data with possible non-proportional hazards.

For the sake of judging the model sensitivity to different hyperparameters, we carried out the PSA data analysis on the original dataset while assigning prior distributions Gamma(5, 0.01), Gamma(10, 0.01), Gamma(7, 0.05), and Gamma(4, 0.05) to the hyperparameter η. Note that all the prior choices considered, allow the multiplicative cluster effects to be large (75%). We observed a maximum change of only 4.9% in the predicted survival probabilities of prostate cancer patients who had undergone prostate removal surgeries, the maximum effect of the hyperparameter choice being observed at around 80 months from the time of surgery. This might be attributable to the fact that the motivating dataset only has a few observations with survival times larger than 80 months from the date of surgery. This leads us to conclude that our SBART-based semiparametric model for clustered, interval-censored survival data is robust and will perform consistently for any reasonable choice of the hyperparameter η.

6 | CONCLUDING REMARKS

In this article, we have introduced a robust and flexible semiparametric model for clustered, interval-censored survival data under the BART framework. This was accomplished by modeling the hazard function as a product of a parametric baseline hazard and a nonparametric component that uses SBART to incorporate unknown time-dependent effects of the covariates and their interactions on the hazard function. Besides being applicable for clustered as well as interval-censored survival data, simulation results validate that our model also attains excellent accuracy in survival prediction, while also requiring minor changes to the usual Bayesian backfitting algorithm used to fit other BART models. Code for fitting the SBART interval-censored survival model with clustering will be made available on the authors' websites. Applicability of the proposed model has been demonstrated via analysis of the motivating post-surgery PSA recurrence study.

The restrictive aspect of our approach is that our frailty model (2) assumes that the cluster effects W_i are time constant and multiplicative on the hazard (similar to frailty model of Hougaard (1995)). There is some scope for generalizing our approach to allow the frailties to be incorporated more flexibly; at the most extreme end, we might incorporate the frailty into the SBART ensemble. However, this comes with additional computational challenges and the risk of overfitting. We defer determining the best way to incorporate the frailty to future work.

An interesting aspect of our methodology is its adaptability to high-dimensional survival studies involving a large number of covariates and complex underlying associations. Although this paper focuses only on the use of SBART to nonparametrically model the deviations of the hazard function from the baseline hazard, our approach can also be extended to allow for ultrahigh-dimensional predictors when the function $l(t, \mathbf{x}_{ij})$ is sparse using the sparsity-inducing Dirichlet prior introduced by Linero (2018), as well as allowing for the penalization of groups of predictors simultaneously (Du and Linero, 2019) in a similar fashion to the group Lasso. This approach might be particularly useful when the primary concern of the study is variable selection, and might provide a new direction for future research in survival prediction.

DATA AVAILABILITY STATEMENT

The data that support the findings in this paper are available from the corresponding author upon reasonable request.

ORCID

Piyali Basak https://orcid.org/0000-0002-0173-3543
Antonio Linero https://orcid.org/0000-0002-9531-5667
Stuart Lipsitz https://orcid.org/0000-0003-2619-1389

REFERENCES

Adams, R.P., Murray, I. and MacKay, D.J.C. (2009) Tractable nonparametric Bayesian inference in Poisson processes with Gaussian process intensities. *Proceedings of the 26th International Conference on Machine Learning (ICML)*.

Albert, J.H. and Chib, S. (1993) Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422), 669–679.

Barbash, G. I. and Glied, S.A. (2010) New technology and health care costs: the case of robot-assisted surgery. *New England Journal of Medicine*, 363(8), 701–704.

Bonato, V., Baladandayuthapani, V., Broom, M.B., Sulman, E.P., Aldape, K.D. and Do, K. A. (2011) Bayesian ensemble methods for survival prediction in gene expression data. *Bioinformatics*, 27(3), 359–367.

Calhoun, P., Su, X., Nunn, M. and Fan, J. (2018) Constructing multivariate survival trees: the MST package for R. *Journal of Statitical Software*, 83, 1–21.

Chipman, H.A., George, E.I. and McCulloch, R.E. (2010) BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4, 266–298.

Conkin, J., Bedahl, S. R. and van Liew H. D., (1992) A computerized data bank of decompression sickness incidence in altitude chambers. *Aviation, Space and Environmental Medicine*, 63, 819–824.

Conkin, J. and Powell, M. (2001) Lower body adynamia as a factor to reduce the risk of hypobaric decompression sickness. *Aviation, Space and Environmental Medicine*, 72, 202–214.

De Iorio, M., Johnson, W.O., Muller, P. and Rosner, G.L. (2009) Bayesian nonparametric nonproportional hazards survival modeling. *Biometrics*, 65, 762–771.

Deshpande, S.K., Bai, R., Balocchi, C. and Starling, J.E. (2020) VC-BART: Bayesian trees for varying coefficients. *arXiv:2003.06416*.

Du, J. and Linero, A.R. (2019) Interaction detection with Bayesian decision tree ensembles. *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics, PMLR 89*, pp. 108–117.

Dreyer, G., Addiss, D., Williamson, J. and Noroes, J. (2006) Efficacy of co-administered diethylcarbamazine and albendazole against adult Wuchereria bancrofti. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 100, 1118–1125.

Fernandez, T., Rivera, N. and Teh, Y.W. (2016) Gaussian processes for survival analysis. *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 16, pp. 5021–5029.

Friedman, J. (1991) Multivariate adaptive regression splines. *Annals of Statistics*, 19, 1–141.

Goethals, K., Ampe, B., Berkvens, D., Laevens, H., Janssen, P. and Duchateau, L. (2009) Modeling interval-censored, clustered cow udder quarter infection times through the shared gamma frailty model. *Journal of Agricultural, Biological and Environmental Statistics*, 14, 1–14.

Hahn, P.R., Murray, J.S. and Carvalho, C.M. (2020) Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. *Bayesian Analysis*. To appear.

Hanson, T. and Yang, M. (2007) Bayesian semiparametric proportional Oodds models. *Biometrics*, 63, 88–95.

Henschel, V., Engel, J., Hölzel, D. and Mansmann, U. (2009) A semiparametric Bayesian proportional hazards model for interval censored data with frailty effects. *BMC Medical Research Methodology*, 9, 9.

Hill, J., Linero, A.R. and Murray, J.S. (2020) Bayesian additive regression trees: a review and look forward. *Annual Review of Statistics and its Application*, 7(1), 251–278.

Hothorn, T., Lausen, B. and Benner, A. (2004) Bagging survival trees. *Statistics in Medicine*, 23(1), 77–91.

Hougaard, P. (1995) Frailty models for survival data. *Lifetime Data Analysis*, 1(3), 255–273.

Ibrahim, R., L'Ecuyer, P., Regnard, N. and Shen, H. (2012) On the modeling and forecasting of call center arrivals. *Proceedings of 2012 Winter Simulation Conference, Berlin*, 256–267.

Ishwaran, H., Kogalur, U.B., Blackstone, E.H. and Lauer, M.S. (2008) Random survival forests. *The Annals of Applied Statistics*, 2(3), 841–860.

Kalbfleisch, J.D. and Prentice, R.L. (2002) *The Statistical Analysis of Failure Time Data*, 2nd edition. Hoboken: John Wiley & Sons.

Kooperberg, C. and Clarkson, D.B. (1997) Hazard regression with interval censored data. *Biometrics*, 53, 1485–1494.

Li, Y., Linero, A.R. and Murray, J.S. Adaptive Conditional Distribution Estimation with Bayesian Decision Tree Ensembles. *arXiv e-prints*.

Li, H. and Luan, Y. (2005) Boosting proportional hazards models using smoothing splines, with applications to high-dimensional microarray data. *Bioinformatics*, 21(10), 2403–2409.

Linero, A.R. (2017) A review of tree-based Bayesian methods. *Communications for Statistical Applications and Methods*, 24(6), 543–559.

Linero, A.R. (2018) Bayesian regression tree ensembles that adapt to smoothness and sparsity. *Journal of the American Statistical Association*, 113(522), 626–636.

Linero, A.R., Basak, P., Li, Y. and Sinha, D. (2021) Bayesian Survival Tree Ensembles with Submodel Shrinkage. *arXiv e-prints*.

Linero, A.R., Sinha, D. and Lipsitz, S.R. (2020) Semiparametric mixed-scale models using shared Bayesian forests. *Biometrics*, 76(1), 131–144.

Linero, A.R. and Yang, Y. (2018) Bayesian regression tree ensembles that adapt to smoothness and sparsity. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 80(5), 1087–1110.

Mallick, B.K., Denison, D.G.T. and Smith, A.F.M. (1999) Bayesian survival analysis using a MARS model. *Biometrics*, 55, 1071–1077.

Murray, J.S. (2017) Log-linear Bayesian additive regression trees for categorical and count responses. *arXiv preprint arxiv:1701.01503*.

Neal, R.M. (2003) Slice sampling. *The Annals of Statistics*, 31, 705–767.

Oakes, D.R. (1982) A model for association in bivariate survival data. *Journal of the Royal Statistical Society, Series B*, 44, 414–428.

Pratola, M., Chipman, H., George, E. and McCulloch, R. (2017) Heteroscedastic BART using multiplicative regression trees. *arXiv preprint arXiv:1709.07542*.

Sinha, D.Chen, M-H, and Ghosh, S. (1999) Bayesian analysis and model selection for interval-censored survival data. *Biometrics*, 55, 585–590.

Sparapani, R., Logan, B.R., McCulloch, R.E. and Laud, P.W. (2016) Nonparametric survival analysis using Bayesian Additive Regression Trees (BART). *Statistics in Medicine*.

Su, X. and Tsai, C.-L. (2005) Tree-augmented Cox proportional hazards models. *Biostatistics*, 6, 486–499.

Su, X., Zhou, T., Yan, X. and Fan, J. (2008) Interaction trees with censored survival data. *International Journal of Biostatistics*, 4, 1–26.

Sun, J. (2006) *The Statistical Analysis of Interval-Censored Failure Time Data*. Berlin: Springer.

Umlauf, N., Adler, D., Kneib, T., Lang, S. and Zeileis, A. (2015) Structured additive regression models: an R interface to BayesX. *Journal of Statistical Software*, 63, 1–46.

Zhou, H., Hanson, T. and Zhang, J. (2017) Generalized accelerated failure time spatial frailty model for arbitrarily censored data. *Lifetime Data Analysis*, 23, 495–515.

Zhou, H., Hanson, T. and Zhang, J. (2020) spBayesSurv: fitting Bayesian spatial survival models using R. *Journal of Statistical Software*, 92, 1–33.

SUPPORTING INFORMATION

Tables, and Figures referenced in Sections 2, 4 and 5 are available with this paper at the Biometrics website on Wiley Online Library. Simulated PSA data used to generate figures 2 and 3 and all relevant computational codes will also be made available at the Biometrics website on Wiley Online Library and can also be obtained from the corresponding author upon reasonable request.

How to cite this article: Basak P, Linero A, Sinha D, Lipsitz S. Semiparametric analysis of clustered interval-censored survival data using soft Bayesian additive regression trees (SBART). *Biometrics*. 2021;1–14. https://doi.org/10.1111/biom.13478