A new method for clustered survival data: Estimation of treatment effect heterogeneity and variable selection

Liangyuan Hu

Department of Biostatistics and Epidemiology, Rutgers University, Piscataway, New Jersey, USA

Correspondence
Liangyuan Hu, Department of Biostatistics and Epidemiology, Rutgers University, 683 Hoes Lane West, Piscataway, NJ 08854.
Email: liangyuan.hu@rutgers.edu

Funding information
National Heart, Lung, and Blood Institute, Grant/Award Number: 1R01HL159077-01A1; Patient-Centered Outcomes Research Institute, Grant/Award Number: ME-2021C2-23685

Abstract
We recently developed a new method random-intercept accelerated failure time model with Bayesian additive regression trees (riAFT-BART) to draw causal inferences about population treatment effect on patient survival from clustered and censored survival data while accounting for the multilevel data structure. The practical utility of this method goes beyond the estimation of population average treatment effect. In this work, we exposit how riAFT-BART can be used to solve two important statistical questions with clustered survival data: estimating the treatment effect heterogeneity and variable selection. Leveraging the likelihood-based machine learning, we describe a way in which we can draw posterior samples of the individual survival treatment effect from riAFT-BART model runs, and use the drawn posterior samples to perform an exploratory treatment effect heterogeneity analysis to identify subpopulations who may experience differential treatment effects than population average effects. There is sparse literature on methods for variable selection among clustered and censored survival data, particularly ones using flexible modeling techniques. We propose a permutation-based approach using the predictor’s variable inclusion proportion supplied by the riAFT-BART model for variable selection. To address the missing data issue frequently encountered in health databases, we propose a strategy to combine bootstrap imputation and riAFT-BART for variable selection among incomplete clustered survival data. We conduct an expansive simulation study to examine the practical operating characteristics of our proposed methods, and provide empirical evidence that our proposed methods perform better than several existing methods across a wide range of data scenarios. Finally, we demonstrate the methods via a case study of predictors for in-hospital mortality among severe COVID-19 patients and estimating the heterogeneous treatment effects of three COVID-specific medications. The methods developed in this work are readily available in the R package riAFTBART.
1 | INTRODUCTION

Large-scale healthcare data sets garnered from multiple facilities are increasingly available, offering fertile ground for innovative investigation into emerging medical research questions. The inherent hierarchical data structure, in conjunction with the censored survival outcome, which is of strong clinical interest, poses considerable challenges for statistical analysis. Hu et al. (2022) recently developed a flexible approach, riAFT-BART to estimate the population average treatment effect on patient survival, while accounting for the clustering structure of the observations. The method riAFT-BART stands for random-intercept accelerated failure time model with Bayesian additive regression trees (BART) (Chipman et al., 2010). The utility of this new method, which leverages Bayesian machine learning, goes beyond the estimation of population average treatment effect. In this paper, we exposit how riAFT-BART can be used to solve two important statistical questions with clustered survival data: estimating the treatment effect heterogeneity and variable selection.

An immediate extension of the new method, riAFT-BART, beyond the population average treatment effect is the estimation of heterogeneous treatment effect. Prior work (Hu et al., 2021) has shown that compared to algorithm-based machine learning, leveraging the likelihood-based machine learning technique BART can provide more accurate estimation of the individual survival treatment effect (ISTE), which in turn can be used to assess the treatment effect heterogeneity. To the best of our knowledge, no work has yet discussed how BART can be used to estimate the heterogeneous effect of multiple treatments on patient survival using clustered survival observations.

Another important utility of riAFT-BART is for variable selection with clustered survival data, on which there is sparse literature. Fan and Li (2002) proposed a family of variable selection methods for the gamma frailty model based on a nonconcave penalized likelihood approach using the Newton–Raphson method. Androulakis et al. (2012) extended the penalized Cox or Gamma frailty model to a class of frailty models and proposed a generalized form of the likelihood function to allow the use of different continuous distributions for the frailty term. Utazirubanda et al. (2021) proposed group LASSO with gamma-distributed frailty for variable selection in Cox regression. However, no corresponding software is available for these methods. Bender et al. (2018) proposed a piecewise exponential additive mixed model (PEAMM) for clustered survival data, and Marra and Wood (2011) described a practical approach by which shrinkage penalties can be added to the smoothing parameters in PEAMM and variable selection can be performed by using restricted maximum likelihood estimation. Rondeau et al. (2003) proposed to conduct variable selection by maximizing the penalized likelihood using the Marquardt algorithm (Marquardt, 1963), which is a combination of the Newton–Raphson algorithm and the steepest descent algorithm. Ha et al. (2014) proposed a penalized h-likelihood estimation approach for variable selection of fixed effects using semiparametric frailty models. These methods rely on parametric assumptions about the exact relationships between covariates and survival outcomes. Misspecifying the parametric forms of covariate–outcome relationships can reverberate through the variable selection procedure and yield unsatisfactory results (Bleich et al., 2014; Hu et al., 2021).

Flexible machine learning modeling techniques can help relax the parametric assumptions and improve variable selection results. Ishwaran et al. (2010) derived the distribution of the minimal depth of a maximal subtree, which measures the predictiveness of a variable in a survival tree, and used it for high-dimensional variable selection using random survival forests. Lagani and Tsamardinos (2010) proposed an algorithm based on the theory of Bayesian networks and the Markov blanket for variable selection suitable for high-dimensional and right-censored survival data. However, neither of these two machine learning-based methods accommodates clustered survival observations.

In this work, we describe ways in which riAFT-BART can be used to (i) estimate the treatment effect heterogeneity, and to (ii) select important predictors in the presence of missing data with clustered survival observations. The rest of the paper is organized as follows. Section 2 describes the use of riAFT-BART for estimating ISTE and exploring subgroups who may experience enhanced or reduced treatment effect than population average treatment effect. Section 3 proposes a variable selection method using riAFT-BART when covariates are subject to missing data. In Section 4, we conduct an expansive simulation to assess the practical operating characteristics of riAFT-BART in estimating treatment effect heterogeneity.
and performing variable selection. Section 5 illustrates the methods using a COVID-19 data set drawn from the Mount Sinai Medical Center, and Section 6 provides a discussion.

2 | ESTIMATION OF TREATMENT EFFECT HETEROGENEITY

2.1 | Notation

We maintain notation used in Hu et al. (2022). Consider a nonexperimental study with $K$ clusters, each having treated $n_k$ individuals, indexed by $i = 1, \ldots, n_k, k = 1, \ldots, K$. The total sample size is $N = \sum_{k=1}^{K} n_k$. There are $J$ possible treatment options, denoted by $A \in \mathcal{A} = \{a_1, \ldots, a_J\}$. Denote pretreatment measured, $L$-dimensional covariates by $X_{ik} = \{X_{ik1}, \ldots, X_{ikL}\}$ for each individual $i$ in cluster $k$, whose failure time $T_{ik}$ may be right censored at $C_{ik}$. The observed outcome consists of $Y_{ik} = \min(T_{ik}, C_{ik})$ and the censoring indicator $\Delta_{ik} = I(T_{ik} < C_{ik})$. Let $V_k$ be the cluster indicators. Proceeding with the counterfactual framework, the counterfactual failure time under treatment $a_j$ for individual $i$ in cluster $k$ is defined as $T_{ik}(a_j), \forall a_j \in \mathcal{A}$, and the counterfactual censoring time under treatment $a_j$ is defined as $C_{ik}(a_j)$. Throughout, we maintained the standard assumptions (Arpino & Cannas, 2016; Chen & Tsiatis, 2001; Hu & Hogan, 2019; Hu et al., 2022, 2021): consistency, weak unconfoundedness, positivity, and covariate-dependent censoring, for drawing causal inference with observational clustered survival data. Detailed assumptions and implications of assumptions can be found in Hu et al. (2022).

2.2 | Method

Hu et al. (2022) adapted BART into a random-intercept accelerated failure time model,

$$
\log T_{ik} = f(A_{ik}, X_{ik}) + b_k + \varepsilon_{ik},
$$

$$
b_k \overset{i.i.d.}{\sim} N(0, \alpha_k \tau^2), \quad \varepsilon_{ik} \overset{i.i.d.}{\sim} N(0, \sigma^2), \quad b_k \perp \varepsilon_{ik},
$$

where $f(A_{ik}, X_{ik})$ is an unspecified function relating treatment assignment $A_{ik}$ and pretreatment confounders $X_{ik}$ to survival times $T_{ik}$. $b_k$’s are the random intercepts capturing cluster-specific main effects, and $\varepsilon_{ik}$ is the residual term. The unknown function $f$ is flexibly modeled by BART via a sum of shallow trees (Hu & Gu, 2021; Hu et al., 2020). A mean-zero normal distribution and independence are assumed for $b_k$ and $\varepsilon_{ik}$. A redundant parameter $\alpha_k$ was used for the variance of $b_k$ as a parameter expansion technique (Gelman et al., 2008) to improve computational performance. Regularizing priors are placed on the tree parameters to keep the impact of each individual tree on the overall fit small and thus prevent overfitting (Chipman et al., 2010). Sensible priors are used for parameters $\sigma^2$, $\tau^2$, and $\alpha_k$ (Henderson et al., 2020; Hu et al., 2022). To deal with right censoring, data augmentation is used. That is, the unobserved survival times are imputed from a truncated normal distribution in each Gibbs iteration. A Metropolis within Gibbs procedure (Hu et al., 2022) (Web Section 1) was proposed to draw posterior inferences about the population average treatment effect on patient survival, which is defined as

$$
\theta_{a_j, a_{j'}} = \frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{n_k} E[\log T_{ik}(a_j) - \log T_{ik}(a_{j'}) | X_{ik} = x_{ik}, V_k = v_k], \forall a_j, a_{j'} \in \mathcal{A}.
$$

Via outcome modeling, the pairwise population average treatment effect between $a_j$ and $a_{j'}$ can be estimated as

$$
\hat{\theta}_{a_j, a_{j'}} = \frac{1}{D} \sum_{d=1}^{D} \frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \{f^d(a_j, x_{ik}) - f^d(a_{j'}, x_{ik})\},
$$

where $f^d$ is the $d$th draw from the posterior distribution of $f$. Inferences can be obtained based on the $D$ posterior average treatment effects, $(1/N) \sum_{k=1}^{K} \sum_{i=1}^{n_k} \{f^d(a_j, x_{ik}) - f^d(a_{j'}, x_{ik})\}, d = 1, \ldots, D.$
Turning to the estimation of treatment effect heterogeneity. We first define the individual survival treatment effect for individual $i$ in cluster $k$ as

$$\zeta_{a_j,a_{j'}}(x_{ik}) = E\left[ \log T_{ik}(a_j) - \log T_{ik}(a_{j'}) \mid x_{ik}, v_k \right], \quad (3)$$

which will be used as the building block for estimating the heterogeneous treatment effect. From the fitted Model (1), we can estimate $\hat{\zeta}_{a_j,a_{j'}}(x_{ik})$ as

$$\hat{\zeta}_{a_j,a_{j'}}(x_{ik}) = \bar{f}(a_j, x_{ik}) - \bar{f}(a_{j'}, x_{ik}),$$

where $\bar{f}$ is the mean of $D$ draws from the posterior predictive distribution of $f$. The ISTE estimates $\{\hat{\zeta}_{a_j,a_{j'}}(x_{ik}), i = 1, ... , n_k, k = 1, ... , K\}$ can then be used to explore the degree of treatment effect heterogeneity. By the “virtual twins” approach (Foster et al., 2011), the $\{\hat{\zeta}_{a_j,a_{j'}}(x_{ik})\}$ would be regressed on the predictors via a tree diagram to identify the predictor subspaces that explain away differences in $\{\hat{\zeta}(x_{ik})\}$. In a similar spirit, the “fit-the-fit” strategy has been used to identify possible subgroups defined by combination rules of covariates that have differential treatment effects (Hu et al., 2021; Logan et al., 2019). Mathematically, we wish to find a covariate subspace $S_{a_j,a_{j'}} = \{S^1_{a_j,a_{j'}}, S^2_{a_j,a_{j'}}, ... , S^H_{a_j,a_{j'}}\}$, where $h \in \{1, ... , H\}$ is the number of combination rules of covariates, such that the average treatment effects across individuals in this region are considerably different than the population average treatment effect. Denote the measure of subgroup treatment effect between $a_j$ and $a_{j'}$ as $M(S_{a_j,a_{j'}}) = \{M(S^1_{a_j,a_{j'}}), ... , M(S^H_{a_j,a_{j'}})\}$, and

$$M(S^h_{a_j,a_{j'}}) = E \left[ \zeta_{a_j,a_{j'}}(x_{ik}) \mid x_{ik} \in S^h_{a_j,a_{j'}} \right]. \quad (4)$$

We adopt the “fit-the-fit” strategy to estimate $S_{a_j,a_{j'}}$ and $M(S_{a_j,a_{j'}})$. The algorithm proceeds with the following steps:

1. Add each candidate predictor separately to the random forests model (Breiman, 2001) using $\{\hat{\zeta}_{a_j,a_{j'}}(x_{ik})\}$ as the outcome, and record the model fit measured by $R^2$; include the predictor corresponding to the largest $R^2$ in the model;
2. Add each of the remaining candidate predictors separately to the previous step’s model and retain the predictor corresponding to the largest $R^2$ improvement;
3. Repeat step 2 until the percentage improvement in $R^2$ is less than 1%. At this point, the final model is established.

The $R^2$ is the coefficient of determination in regression analysis that determines the proportion of variance in the dependent variable that can be predicted from the independent variable(s). Different from prior work, we use random forests instead of a single tree to regress the ISTE estimates on predictors for improved modeling accuracy and stability. To interpret the tree ensembles in the final random forests model yielded in Step 3, we apply the inTrees algorithms (Deng, 2019) to translate the composite rules into a simple rule model. The inTrees algorithms extract, measure and rank rules, prune redundant rules, detect frequent variable interactions, and summarize rules into a simple tree diagram model that can be used for predicting new data. The branch decision rules in the final tree diagram model sending individuals to the terminal nodes suggest the predictor subspace $S_{a_j,a_{j'}}$; and subgroup treatment effects that are substantially different from the population effect $\theta_{a_j,a_{j'}}$ can then be estimated by averaging the ISTE estimates among individuals falling into each of the terminal nodes,

$$\hat{M}(S^h_{a_j,a_{j'}}) = \frac{1}{|\hat{S}^h_{a_j,a_{j'}}|} \sum_{i,k : x_{ik} \in S^h_{a_j,a_{j'}}} \hat{\zeta}_{a_j,a_{j'}}(x_{ik}),$$

where $|\hat{S}^h_{a_j,a_{j'}}|$ is the size of $\hat{S}^h_{a_j,a_{j'}}$.

3 | VARIABLE SELECTION

Variable selection has been a long-standing statistical problem and studied from both the classical and Bayesian perspective. In the field of causal inference, variable selection methods can aid in confounder selection (VanderWeele, 2019).
Variable selection methods are largely based on parametric models, in which the exact relationships between the response and predictor variables need to be explicitly specified. However, incorrectly specifying the parametric models may produce unsatisfactory variable selection results (Bleich et al., 2014; Hu et al., 2021; Lin et al., 2022). For example, noise variables may be selected and important variables ignored. Flexible machine learning models can better represent functional forms of arbitrary complexity, thus mitigate parametric assumptions and attendant errors in variable selection. We develop a way in which the new method riAFT-BART can be used for variable selection among clustered survival data. Notation defined in Section 2.1 is largely retained with the exception of treatment assignment $A_{ik}$, as the focus now is variable selection.

### 3.1 Random-intercept accelerated failure time model with Bayesian additive regression trees

We propose a permutation-based approach for variable selection. Following prior work, Bleich et al. (2014), Hu et al. (2021), and Lin et al. (2022), we use the variable inclusion proportion (VIP) of each predictor variable as the foundation to determine how important a predictor variable is. The VIP score is supplied by the BART model. The VIP score signifies the frequency of each predictor’s utilization in defining the splitting rules, relative to the total count of such rules appearing in the model. Each predictor’s proportion can be used to rank the variables based on their relative importance. However, these proportions should not be interpreted as prescriptive guidelines for variable selection. Instead, they serve as indicative measures of variable significance within the model (Chipman et al., 2010; Huet al., 2020). We note that in the riAFT-BART model (1), the predictor variables $X_{ik}$, along with treatment assignment $A_{ik}$, is included in the unknown function $f(A_{ik}, X_{ik})$, which is being modeled by BART. Thus, in each iteration of the riAFT-BART sampling algorithm (Hu et al., 2022), we can extract the VIP for each predictor variable in the step where a BART model is fit to regress $\log y_{i,k,cen}$ on a combination of predictor variables and treatment assignment. However, these proportions should not be interpreted as prescriptive guidelines for variable selection. Instead, they serve as indicative measures of variable significance within the model (Chipman et al., 2010; Huet al., 2020). We propose a permutation-based approach for variable selection. Following prior work, Bleich et al. (2014), Hu et al. (2021), and Lin et al. (2022), we use the variable inclusion proportion (VIP) of each predictor variable as the foundation to determine how important a predictor variable is. The VIP score is supplied by the BART model. The VIP score signifies the frequency of each predictor’s utilization in defining the splitting rules, relative to the total count of such rules appearing in the model. Each predictor’s proportion can be used to rank the variables based on their relative importance. However, these proportions should not be interpreted as prescriptive guidelines for variable selection. Instead, they serve as indicative measures of variable significance within the model (Chipman et al., 2010; Huet al., 2020). We note that in the riAFT-BART model (1), the predictor variables $X_{ik}$, along with treatment assignment $A_{ik}$, is included in the unknown function $f(A_{ik}, X_{ik})$, which is being modeled by BART. Thus, in each iteration of the riAFT-BART sampling algorithm (Hu et al., 2022), we can extract the VIP for each predictor variable in the step where a BART model is fit to regress $\log y_{i,k,cen}$ on a combination of predictor variables and treatment assignment. However, these proportions should not be interpreted as prescriptive guidelines for variable selection. Instead, they serve as indicative measures of variable significance within the model (Chipman et al., 2010; Huet al., 2020). We propose a permutation-based approach for variable selection. Following prior work, Bleich et al. (2014), Hu et al. (2021), and Lin et al. (2022), we use the variable inclusion proportion (VIP) of each predictor variable as the foundation to determine how important a predictor variable is. The VIP score is supplied by the BART model. The VIP score signifies the frequency of each predictor’s utilization in defining the splitting rules, relative to the total count of such rules appearing in the model. Each predictor’s proportion can be used to rank the variables based on their relative importance. However, these proportions should not be interpreted as prescriptive guidelines for variable selection. Instead, they serve as indicative measures of variable significance within the model (Chipman et al., 2010; Huet al., 2020). We propose a permutation-based approach for variable selection. Following prior work, Bleich et al. (2014), Hu et al. (2021), and Lin et al. (2022), we use the variable inclusion proportion (VIP) of each predictor variable as the foundation to determine how important a predictor variable is. The VIP score is supplied by the BART model. The VIP score signifies the frequency of each predictor’s utilization in defining the splitting rules, relative to the total count of such rules appearing in the model. Each predictor’s proportion can be used to rank the variables based on their relative importance. However, these proportions should not be interpreted as prescriptive guidelines for variable selection. Instead, they serve as indicative measures of variable significance within the model (Chipman et al., 2010; Huet al., 2020). We propose a permutation-based approach for variable selection. Following prior work, Bleich et al. (2014), Hu et al. (2021), and Lin et al. (2022), we use the variable inclusion proportion (VIP) of each predictor variable as the foundation to determine how important a predictor variable is. The VIP score is supplied by the BART model. The VIP score signifies the frequency of each predictor’s utilization in defining the splitting rules, relative to the total count of such rules appearing in the model. Each predictor’s proportion can be used to rank the variables based on their relative importance. However, these proportions should not be interpreted as prescriptive guidelines for variable selection. Instead, they serve as indicative measures of variable significance within the model (Chipman et al., 2010; Huet al., 2020). We propose a permutation-based approach for variable selection. Following prior work, Bleich et al. (2014), Hu et al. (2021), and Lin et al. (2022), we use the variable inclusion proportion (VIP) of each predictor variable as the foundation to determine how important a predictor variable is. The VIP score is supplied by the BART model. The VIP score signifies the frequency of each predictor’s utilization in defining the splitting rules, relative to the total count of such rules appearing in the model. Each predictor’s proportion can be used to rank the variables based on their relative importance. However, these proportions should not be interpreted as prescriptive guidelines for variable selection. Instead, they serve as indicative measures of variable significance within the model (Chipman et al., 2010; Huet al., 2020). We propose a permutation-based approach for variable selection. Following prior work, Bleich et al. (2014), Hu et al. (2021), and Lin et al. (2022), we use the variable inclusion proportion (VIP) of each predictor variable as the foundation to determine how important a predictor variable is. The VIP score is supplied by the BART model. The VIP score signifies the frequency of each predictor’s utilization in defining the splitting rules, relative to the total count of such rules appearing in the model. Each predictor’s proportion can be used to rank the variables based on their relative importance. However, these proportions should not be interpreted as prescriptive guidelines for variable selection. Instead, they serve as indicative measures of variable significance within the model (Chipman et al., 2010; Huet al., 2020).
3.2 | Comparison methods

3.2.1 | Piecewise exponential additive mixed model

For a PEAMM (Bender et al., 2018), the hazard rate at time $t$ for individual $i$ in cluster $k$ with covariates $X_{ik}$ and cluster indicator $V_k$ is given by

$$
\lambda(t \mid X_{ik}, V_k) = \exp \left\{ g_0(t_q) + \sum_{m=1}^{M} g_m(x_{ik,m}) + b_k \right\} \forall t \in (\kappa_{q-1}, \kappa_q),
$$

(5)

where $g_0(t_q)$ represents the log-baseline hazard rate, $g_m(x_{ik,m})$ is the basis functions of natural cubic splines associated with the covariates $x_{ik}$, $b_k$ is the random intercept term capturing the cluster-specific effect, and $\kappa_{q-1}$ and $\kappa_q$ are the start and stop times of the censored survival data set represented in the (start, stop] counting process form. For variable selection using PEAMM, following the practical approach by Marra and Wood (2011), a shrinkage penalty term can be added to each basis term $g_m(x_{ik,m})$ of model (5):

$$
\lambda(t \mid X_{ik}, V_k) = \exp \left\{ g_0(t_q) + \sum_{m=1}^{M} \left( g_m(x_{ik,m}) + \int g'_m(x_{ik,m})^2 dx_{ik,m} \right) + b_k \right\} \forall t \in (\kappa_{q-1}, \kappa_q),
$$

so that when the whole spline basis terms for a predictor $x_{ikl}$ are shrunk to zero, the variable $x_{ikl}$ will be selected out of the model. The shrinkage penalties, along with all other model coefficients, can be estimated using restricted maximum likelihood estimation.

3.2.2 | Frailty models by regularization methods

Given the common unobserved frailty $u_k$ for cluster $k$, the conditional hazard function at time $t$ for individual $i$ in cluster $k$ is

$$
\lambda(t \mid X_{ik}, V_k) = u_k \lambda_0(t) \exp(x^T_{ik} \beta),
$$

(6)

where $\lambda_0(t)$ is an unspecified baseline hazard function and $\beta = (\beta_1, \beta_2, ..., \beta_L)^T$ is a vector of regression parameters for $x_{ik}$. The common unobserved frailty $u_k$ is assumed to follow a gamma distribution because of its mathematical convenience. Rondeau et al. (2003) proposed to conduct variable selection by maximizing the penalized likelihood,

$$
l(\lambda(t \mid X_{ik}, V_k)) - \alpha \int_0^\infty \lambda''_0(t)dt,
$$

where $l(\lambda(t \mid X_{ik}, V_k))$ is the full marginal log-likelihood for the frailty model (6) and $\alpha$ is the smoothing parameter which controls the trade-off between the data fit and the smoothness of the baseline hazard function. The regression parameters can be estimated by the robust Marquardt algorithm (Marquardt, 1963)—a combination of the Newton–Raphson algorithm and the steepest descent algorithm. Ha et al. (2014) instead proposed to perform variable selection by maximizing the penalized profile h-likelihood,

$$
l^*(\lambda(t \mid X_{ik}, V_k)) - \sum_{l=1}^{L} B_\gamma(|\beta_l|),
$$

where $l^*(\lambda(t \mid X_{ik}, V_k))$ is the profiled h-likelihood of the frailty model, $B_\gamma(|\beta_l|)$ is a penalty function that controls model complexity using the tuning parameter $\gamma$, with the optimal value determined by the Bayesian information criterion (BIC). Three penalty functions were considered for $B_\gamma(|\beta_l|)$: least absolute shrinkage and selection operator (Tibshirani, 1996), smoothly clipped absolute deviation (Fan & Li, 2001), and h-likelihood (Lee & Oh, 2014).
3.2.3 Backward stepwise selection

We also implement a classical variable selection approach through a recursive procedure of backward stepwise selection based on statistical testing, as described in Wood et al. (2008). The backward stepwise selection starts with a random-intercept Cox regression model (riCox, Abrahantes & Burzykowski, 2005) with all potential predictors and removes from the model the predictor that has the least impact on the fit determined by the significance level. The forward selection process then checks whether removed variables should be added back into the model at \( \alpha(1 - \varepsilon) \) for \( \varepsilon \) small significance level. We use \( \alpha = 0.05 \) as suggested by Wood et al. (2008).

3.3 Variable selection with missing data

Large-scale healthcare data are susceptible to the missing data issue. Missing values in predictor variables pose challenges for variable selection. Long and Johnson (2015) proposed a general resampling approach that combines bootstrap imputation and randomized lasso-based variable selection procedure, and used simulations to demonstrate that this approach had better performance compared with several existing methods for variable selection in the presence of missing data. Recent work by Hu et al. (2021) shows that combining bootstrap imputation with flexible machine learning-based variable selection methods can substantially improve variable selection results over methods that rely on parametric modeling of the covariate–outcome relationship. In this work, we combine bootstrap imputation and variable selection, and compare each of the methods considered on the ability to select predictor variables that are truly associated with the outcome, or useful predictor variables, among incomplete clustered survival data. To ensure consistency with previous literature, such as Wood et al. (2008), Bleich et al. (2014), and Hu et al. (2021), and to facilitate comparison of our methods and results with similar studies, we adopt the four performance metrics frequently employed in the variable selection literature. A good method should have high values of precision, recall and \( F_1 \), and low Type I error.

1. Precision = \( \frac{TP}{TP + FP} \), where TP and FP are, respectively, the number of true positive and false positive selections. A true positive selection is when a useful predictor was selected; and a false positive selection is when a noise predictor was selected. The precision of a variable selection method is the proportion of truly useful predictors among all selected predictors.

2. Recall = \( \frac{TP}{TP + FN} \), where FN is the number of false negative selections. A false negative selection is when a truly useful predictor was not selected. The recall of a variable selection method is the proportion of truly useful variables selected among all useful variables. This is sometimes referred to as the power in the literature (Wood et al., 2008).

3. \( F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \). The \( F_1 \) score measures a method’s ability to avoid selecting irrelevant predictors (precision) with its ability to identify the full set of useful predictors (recall).

4. Type I error. The Type I error measures the overall probability of a method incorrectly selecting the noise predictor—the mean of the probabilities that the method incorrectly selects each of the noise predictors.

Among incomplete data, we first conduct bootstrap imputation through two steps: (i) generate \( B \) bootstrap data sets of the original data; (ii) conduct a single imputation for each bootstrap data set using an imputation method of choice. Following Hu et al. (2021), in our simulations and case study, we used \( B = 100 \) and performed standard imputation program using the \( R \) package \texttt{mice}. Variable selection via each of the methods considered was performed for all \( B \) bootstrap imputed data sets, and the final set of predictors were selected if they were selected in at least \( \pi B \) imputed data sets, where \( \pi \in (0, 1) \) is a fraction threshold for selecting a predictor.

4 SIMULATION

4.1 Treatment effect heterogeneity

We carried out a contextualized simulation to evaluate the performance of our proposed method in estimating the heterogeneous treatment effect using observational data with clustered survival outcomes and multiple treatments. For methods comparison, we adapted the popularly used inverse probability weighting method into the setting of clustered and censored survival data to form two comparison methods: inverse probability of treatment weighting with the random-
intercept Cox regression model (IPW-riCox) and doubly robust random-intercept additive hazards model (DR-riAH) (Cai & Zeng, 2011). In addition, we considered the PEAMM (Bender et al., 2018)—another outcome modeling-based method flexible at capturing nonlinear relationships. Because not all methods directly model the survival times \( T_{ik} \), to objectively compare methods, we use the individual-specific counterfactual survival curve as the basis. That is, the ISTE \( \zeta_{a_j,a_j'}(x_{ik}) \) will now be defined in a similar way as in Equation (3) but using functionals of the counterfactual survival curve—either the survival probability at a fixed point in time or the conditional restricted mean survival time (RMST) (Royston & Parmar, 2013). We will compare each method's ability to accurately estimate the ISTE \( \zeta_{a_j,a_j'}(x_{ik}) \) and the average subgroup treatment effect \( M(SH_{a_j,a_j'}(x_{ik})) \). The performance metrics used to evaluate the methods were the relative bias (or percent bias), root-mean-squared-error (RMSE) and precision in the estimation of heterogeneous effects (PEHE) (Hu et al., 2021). The relative bias evaluates the bias in the estimation of treatment effect on the relative scale with respect to the true treatment effect (which is constant across different methods). The PEHE measures the difference between the true and estimated survival probabilities or RMST across all data points. A smaller value of PEHE indicates a higher estimation accuracy. Although the methods examined in this study rely on different models, each can be utilized to estimate survival probabilities and RMST. These common outcomes will serve as a basis for an impartial comparison of these methods.

For weighting-based methods IPW-riCox and DR-riAH, we used Super Learner (Van der Laan et al., 2007) to estimate the stabilized inverse probability of treatment weights. Super Learner was implemented using the R package SuperLearner and the library argument was set to \( \text{SL.library} = \text{c("SL.xgboost", "SL.bartMachine","SL.gbm")} \). We used R package \text{coxme} to implement IPW-riCox and \text{timereg} to implement DR-riAH. For PEAMM, the \text{gam} function from R package \text{mgcv} and two helper functions, (\text{as_ped} and \text{add_surv_prob}), from R package \text{pammtools} were used. For all methods, all confounding variables available to the analyst in their original forms were included in the corresponding models.

### 4.1.1 Simulation design

We simulated data sets with structures similar to the COVID-19 data used in our case study (Section 5). We generated \( K = 10 \) clusters, each with a sample size of \( n_k = 200 \). The total sample size is \( N = 2000 \). We simulated seven confounding variables for each individual \( i \) in cluster \( k \): five continuous variables \( X_{ikl} \sim N(0,1), l \in \{1,...,5\} \); and two categorical variables \( X_{ikl} \sim \text{Multinomial}(1,.3,.3,.4), l = 6, 7 \). We generated \( J = 3 \) treatment groups with unequal sample sizes; the ratio of individuals across treatment groups was 5:3:1, which is similar to the ratio of individuals across the three treatment groups in our case study. The treatment assignment mechanism follows a random-intercept multinomial logistic regression model,

\[
\ln \frac{P(A_{ik} = 1)}{P(A_{ik} = 3)} = 1.5 + 0.1X_{ik1} + 0.1X_{ik2} + 0.1X_{ik3} + 0.5X_{ik4} + 0.4X_{ik5} + 0.2X_{ik6} + 0.3X_{ik7} + 0.4X_{ik2}^2 + 0.4X_{ik5}^2 + \tau_k,
\]

\[
\ln \frac{P(A_{ik} = 2)}{P(A_{ik} = 3)} = 0.7 + 0.1X_{ik1} + 0.3X_{ik2} + 0.2X_{ik3} + 0.2X_{ik4} + 0.1X_{ik5} + 0.4X_{ik6} + 0.5X_{ik7} - 0.3X_{ik2}X_{ik4} + 0.7X_{ik2}^2X_{ik4} + \tau_k,
\]

where the random intercept \( \tau_k \sim N(0,1^2) \). The parameter setup induces a moderate level of confounding, leading to moderate covariate overlap. The sparsity of overlap can be visualized by the distributions of true generalized propensity scores, shown in Web Figure 1. It should be noted that the overall extent of covariate overlap closely resembles that observed in the COVID-19 data set (refer to Web Figure 8). The specific one-to-one correspondences are as follows: \( A = 1 \) emulates treatment 1, dexamethasone; \( A = 3 \) emulates treatment 2, remdesivir; and \( A = 2 \) emulates treatment 3, a combination of dexamethasone and remdesivir. Following strategies (Hu et al., 2021; Lu et al., 2018) recommended for assessing the performance of methods in estimating heterogeneous treatment effects, we subclassified the simulated individuals into 40 subgroups based on the distribution of true generalized propensity scores (Feng et al., 2012), and calculated the biases and RMSE for each subgroup across 250 data replications. Note that the generalized propensity scores for each individual sum to one, \( P(A_{ik} = 1 \mid X_{ik}, \tau_k) + P(A_{ik} = 2 \mid X_{ik}, \tau_k) + P(A_{ik} = 3 \mid X_{ik}, \tau_k) = 1 \). The subclassification is based on intervals of the true generalized propensity scores for treatment 1 and treatment 2, representing a full range of assignment propensity to each treatment group. Subclasses of generalized propensity scores are presented in Web Table 1.
We simulated three sets of true counterfactual survival times from the Weibull survival distribution, which can be parameterized as both the accelerated failure time model and Cox proportional hazards model (Hu et al., 2022),

$$S_{ik}(t) = \exp \left[ \left\{ \lambda_{aj} \exp \left( q_{aj}(X_{ik}, b_k) \right) t^\eta \right\} \right],$$

where $S_{ik}(t)$ is individual-specific survival function for $i$th individual in cluster $k$, $\lambda_{aj}$ is a treatment group-specific parameter, $\eta$ is the shape parameter, and $q_{aj}(X_{ik}, b_k)$ represents a general functional form of covariates and the random-intercept. Using the inverse transform sampling, we generated counterfactual survival times by

$$T_{ik}(a_j) = \begin{cases} -\log U & \\
\lambda_{aj} \exp \left( q_{aj}(X_{ik}, b_k) \right) \end{cases} \frac{1}{\eta},$$

for $a_j \in \{1, 2, 3\}$, where $U$ is a random variable following the uniform distribution on the interval $[0, 1]$ and $\lambda_{aj} = \{5000, 800, 1200\}$ for $a_j = 1, 2, 3$. Observed and uncensored survival times are generated as

$$T_{ik} = \sum_{a_j \in \{1, 2, 3\}} T_{ik}(a_j)I(A_{ik} = a_j).$$

We considered three treatment effect heterogeneity settings with an increasing level of complexity:

(a) \( q_1(X_{ik}, b_k) = 0.1X_{ik1} + 0.3X_{ik2} + \sin(3.14X_{ik3}) + 0.6X_{ik4} + 0.5X_{ik5} + 1.2X_{ik6} + 0.4X_{ik7} + 0.3X_{ik2}^2 + 0.5X_{ik4}X_{ik5} + b_k - 1 \)
\[ q_2(X_{ik}, b_k) = 0.4X_{ik1} + 1.2\sin(3.14X_{ik2}) + 0.4X_{ik3} + 0.3X_{ik4} + 1.0X_{ik5} + 0.8X_{ik6} + 0.2X_{ik7} + 0.7X_{ik1}^2 + 0.4X_{ik1}X_{ik4} + b_k \]
\[ q_3(X_{ik}, b_k) = 0.4X_{ik1} + 0.9X_{ik2} + 0.4X_{ik3} + 0.9X_{ik4} + 0.4X_{ik5} + 0.4X_{ik6} + 0.3X_{ik7} + b_k - 2 \]

(b) \( q_1(X_{ik}, b_k) = 0.1X_{ik1} + 0.3X_{ik2} + \sin(3.14X_{ik3}) + 0.6X_{ik4} + 0.5X_{ik5} + 1.2X_{ik6} + 0.3X_{ik7} + 0.3X_{ik2}^2 + 0.5X_{ik4}X_{ik5} + b_k - 1 \)
\[ q_2(X_{ik}, b_k) = 0.4X_{ik1} + 1.2\sin(3.14X_{ik3}) + 0.4X_{ik4} + 0.3X_{ik5} + 1.0X_{ik6} + 0.8X_{ik7} + 0.7X_{ik1}^2 + 0.4X_{ik1}X_{ik4} + b_k \]
\[ q_3(X_{ik}, b_k) = 0.4\sin(3.14X_{ik1}) + 0.9X_{ik2} + 0.9X_{ik3} + 0.4X_{ik4} + 0.4X_{ik5} + 0.9X_{ik6} + 0.3X_{ik7} + 0.4X_{ik4}^2 - 0.3X_{ik2}X_{ik3} + b_k - 3 \]

(c) \( q_1(X_{ik}, b_k) = 0.1X_{ik1} + 0.3X_{ik2} + \sin(3.14X_{ik3}) + 0.6X_{ik4} + 0.5X_{ik5} + 1.2X_{ik6} + 0.3X_{ik2}^2 + 0.5X_{ik4}X_{ik5} + b_k - 1 \)
\[ q_2(X_{ik}, b_k) = 0.4X_{ik1} + 1.2\sin(3.14X_{ik3}) + 0.4X_{ik4} + 0.3X_{ik5} + 1.0X_{ik6} + 0.8X_{ik7} + 0.7X_{ik1}^2 + 0.4X_{ik1}X_{ik4} + b_k \]
\[ q_3(X_{ik}, b_k) = 0.4\sin(3.14X_{ik2}) + 0.9X_{ik3} + 0.9X_{ik4} + 0.4X_{ik5} + 0.4X_{ik6} + 0.9X_{ik7} + 0.4X_{ik4}^2 - 0.3X_{ik2}X_{ik3} + b_k - 3, \]

where the random intercept $b_k \sim N(0, 4^2)$. Across all three heterogeneity settings, $X_{ik3}, X_{ik4}, X_{ik5}, X_{ik6}$ were confounders related to both the treatment assignment mechanism and the potential outcome generating process. In scenario (a), nonlinear covariate effects are present in treatment groups 1 and 2. In scenario (b), there are nonlinear covariate effects in all three treatment groups. Scenario (c) is similar to scenario (b) except that a nonoverlapping set of covariates are included in three treatment groups. This additional complexity represents an additional source of treatment effect heterogeneity. The parameter $\eta$ was set to 2 and $\exp(0.7 + 0.5X_{ik1})$, respectively, induce proportional hazards and nonproportional hazards. We further generate the censoring time $C$ independently from an exponential distribution with the rate parameter chosen to induce a censoring proportion of 50%, which is close to the censoring proportion observed in the COVID-19 data in our case study. The Kaplan–Meier survival curves by treatment group for all three heterogeneity settings under both proportional and nonproportional hazards are presented in Web Figure 2.

### 4.1.2 Simulation results

Table 1 summarizes the overall precision measure PEHE (based on 3-week survival probability) for each of four methods under three heterogeneity settings and both proportional and nonproportional hazards. First, flexible methods...
TABLE 1 Mean and (standard deviation) of precision in the estimation of heterogeneity effects (PEHE) across 250 data replications for each of the four methods based on 3-week survival probability under six configurations: (proportional hazards [PH] vs. nonproportional hazards [nPH]) × (heterogeneity setting [HS] (a) vs. (b) vs. (c)).

| Methods         | PH          |          |          |          | nPH         |          |          |          |
|-----------------|-------------|----------|----------|----------|-------------|----------|----------|----------|
|                 | HS(a)       | HS(b)    | HS(c)    |          | HS(a)       | HS(b)    | HS(c)    |          |
| Trt 1 vs. 2     | IPW-riCox   | 0.062 (0.028) | 0.066 (0.030) | 0.072 (0.029) | 0.086 (0.028) | 0.089 (0.029) | 0.095 (0.032) |
|                 | DR-riAH     | 0.017 (0.021) | 0.017 (0.021) | 0.024 (0.023) | 0.026 (0.021) | 0.031 (0.021) | 0.036 (0.022) |
|                 | PEAMM       | 0.029 (0.022) | 0.024 (0.022) | 0.022 (0.022) | 0.040 (0.022) | 0.033 (0.022) | 0.035 (0.023) |
|                 | riAFT-BART  | 0.012 (0.010) | 0.008 (0.009) | 0.006 (0.010) | 0.017 (0.012) | 0.014 (0.012) | 0.011 (0.011) |
| Trt 1 vs. 3     | IPW-riCox   | 0.066 (0.025) | 0.068 (0.028) | 0.072 (0.030) | 0.084 (0.028) | 0.094 (0.031) | 0.095 (0.030) |
|                 | DR-riAH     | 0.021 (0.018) | 0.026 (0.020) | 0.029 (0.022) | 0.032 (0.019) | 0.036 (0.019) | 0.038 (0.021) |
|                 | PEAMM       | 0.034 (0.024) | 0.028 (0.023) | 0.027 (0.023) | 0.047 (0.026) | 0.047 (0.025) | 0.038 (0.024) |
|                 | riAFT-BART  | 0.015 (0.010) | 0.013 (0.009) | 0.010 (0.009) | 0.023 (0.023) | 0.020 (0.012) | 0.014 (0.012) |
| Trt 2 vs. 3     | IPW-riCox   | 0.091 (0.027) | 0.095 (0.031) | 0.098 (0.030) | 0.103 (0.031) | 0.113 (0.031) | 0.121 (0.035) |
|                 | DR-riAH     | 0.036 (0.023) | 0.042 (0.022) | 0.044 (0.024) | 0.045 (0.020) | 0.053 (0.023) | 0.055 (0.023) |
|                 | PEAMM       | 0.047 (0.022) | 0.044 (0.022) | 0.041 (0.022) | 0.058 (0.025) | 0.053 (0.022) | 0.050 (0.021) |
|                 | riAFT-BART  | 0.017 (0.011) | 0.014 (0.011) | 0.013 (0.010) | 0.022 (0.015) | 0.019 (0.014) | 0.016 (0.015) |

Abbreviations: DR-riAH, doubly robust random-intercept additive hazards model; IPW-riCox, inverse probability of treatment weighting with the random-intercept Cox regression model; PEAMM, piecewise exponential additive mixed model; riAFT-BART, random-intercept accelerated failure time model with Bayesian additive regression trees.

riAFT-BART and PEAMM and doubly robust weighting method DR-riAH yielded substantially smaller PEHE than the parametric single robust method IPW-riCox across all simulation scenarios. Second, as the heterogeneity setting became more complex, flexible methods riAFT-BART and PEAMM tended to have better performance evidenced by decreasing mean and variation of the PEHE. All methods show decreasing precision in estimating ISTE, demonstrated by increasing PEHE value, when the proportional hazards assumption fails to hold. This is possibly due to more complex data-generating processes, in which the rate parameter \( \eta \) depends on covariate \( x_{ik1} \). The decrease in estimation precision is the smallest for riAFT-BART. Overall, across all scenarios, the proposed method riAFT-BART had the smallest PEHE, or the highest accuracy in estimating the ISTE. Similar observations can be made from the PEHE measures based on 3-week RMST, which are provided in Web Table 2.

Figure 1 displays the relative biases and RMSE results among 40 generalized propensity score subgroups, each averaged across 250 simulation runs. Three pairwise treatment effects were estimated by averaging the ISTE \( \hat{\delta}_{ij,a,j'}(x_{ik}) \) (based on 3-week survival probability) across individuals in each subgroup. Under both settings of proportional hazards and non-proportional hazards, the proposed method riAFT-BART boasts the smallest biases and RMSE across all three treatment effects, particularly for subpopulation near the centroid of the generalized propensity score distribution. In sharp contrast, IPW-riCox, which encodes parametric and linearity assumptions and assumes proportional hazards, yielded the largest biases and RMSE. The increasing complexity level of treatment effect heterogeneity did not impact the performance of more flexible methods PEAMM and riAFT-BART, but induced larger biases and RMSE for parametric methods IPW-riCox and DR-riAH. The violation of the proportional hazards assumption, which introduced an additional data complexity (the rate parameter \( \eta \) depends on the covariate \( x_{ik1} \), had a negligible effect on the performance of riAFT-BART, but all three comparison methods had a deteriorated performance. It is worth noting that all methods produced large biases and RMSE for subpopulation at the tail regions of the generalized propensity score distribution, suggesting that causal inferences drawn about this subpopulation may not be reliable. Results based on 3-week RMST, shown in Web Figure 3, convey the same messages.

4.2 | Variable selection

We adopt the simulation settings used in Hu et al. (2021) to evaluate the performance of variable selection methods among incomplete, clustered survival data. We used the multivariate amputation approach to generate missing data scenarios with desired missingness percentages under the missing at random mechanism (Schouten et al., 2018). The multivariate
FIGURE 1 Relative biases (Panel (a)) and root-mean-squared-errors (RMSE) (Panel (b)) among 40 generalized propensity score subgroups under six data configurations: (heterogeneity settings a, b, c) × (proportional hazards (PH) and nonproportional hazards (nPH)) for each of four methods, inverse probability of treatment weighting with the random-intercept Cox regression model (IPW-riCox), doubly robust random-intercept additive hazards model (DR-riAH), piecewise exponential additive mixed model (PEAMM), and random-intercept accelerated failure time model with Bayesian additive regression trees (riAFT-BART). Three pairwise treatment effects were estimated by averaging the individual survival treatment effect (based on 3-week survival probability) across individuals in each subgroup. Each boxplot visualizes the distribution of relative biases or the distribution of RMSE for 40 subgroups, each averaged across 250 simulation runs.

The multivariate amputation procedure was implemented via the `ampute` function of the R package `mice` (Schouten et al., 2018). Imputation was performed using the R package `mice` (van Buuren & Groothuis-Oudshoorn, 2011). To implement our proposed method using riAFT-BART, we used 4500 posterior draws with the first 1000 discarded as burn-in. The
permutation distributions of VIP scores were constructed from 100 riAFT-BART model runs. To implement the variable selection procedure based on PEAMM, we used the helper function (as_ped) from R package pamftools to reformat the data and used the gam function from R package mgcv for variable selection. Variable selection by frailty models was implemented using the frailtyPenal1 function of R package frailtypack for penalized marginal likelihood; and using the function frailty.vs from R package frailtyHL for penalized profile h-likelihood. For backward stepwise selection, the riCox was fitted using the coxme function from R package coxme.

4.2.1 Simulation design

The simulation scenarios are motivated by the data structures observed in the COVID-19 dataset used in our case study. We considered \( K = 10 \) clusters, each with a sample size of \( n_k = 200 \). We generated eight useful predictors that are truly related to the survival outcomes, \( X_{ik1}, \ldots, X_{ik8} \), and 20 noise predictors, \( X_{ik9}, \ldots, X_{ik28} \). We simulated \( X_{ik1} \) and \( X_{ik2} \) independently from Bern(0.5), \( X_{ik3}, X_{ik4} \) from the standard normal distribution \( N(0,1) \), and \( X_{ik5}, X_{ik6}, X_{ik7}, X_{ik8} \) were designed to have missing values under the missing at random mechanism. For a missing at random predictor variable, we specify the true forms in which the predictor depends on the other predictors as the following:

\[
\begin{align*}
x_{ik5} | x_{ik2}, x_{ik3} & \sim N(0.3x_{ik2} - 0.2x_{ik3}, 1) \\
x_{ik6} | x_{ik3}, x_{ik4} & \sim N(-0.4x_{ik3} + 0.4x_{ik4} + 0.3x_{ik3}x_{ik4}, 1) \\
x_{ik7} | x_{ik4}, x_{ik5}, x_{ik6} & \sim N(0.1x_{ik4}(x_{ik5} - 2)^2 - 0.1x_{ik6}^2, 1) \\
x_{ik8} | x_{ik5}, x_{ik6}, x_{ik7} & \sim N(-0.3x_{ik5}^2 + 0.5x_{ik6} + 0.3x_{ik7} + 0.2x_{ik6}x_{ik7}, 1).
\end{align*}
\]

Among the 20 noise predictors, 10 were generated as continuous variables \( X_{ik9}, \ldots, X_{ik18} \iid \sim N(0,1) \), and the other 10 were generated as binary variables \( X_{ik19}, \ldots, X_{ik28} \iid \sim \text{Bern}(0.5) \).

Similar to the outcome generating process described in Section 4.1.1, we generate the observed and uncensored survival times from the Weibull survival distribution,

\[
T_{ik} = \left[ \frac{-\log U}{\lambda \exp\{q(X_{ik}, b_k)\}} \right]^{1/\eta},
\]

where \( \lambda = 3000 \), \( U \) is a random variable following the uniform distribution on the interval \([0,1]\), \( b_k \sim N(0, 4^2) \), and

\[
q(X_{ik}, b_k) = 1.8x_{ik1} + 0.5x_{ik2} + 1.1x_{ik3} - 0.4e^{x_{ik5}} + 0.4(x_{ik6} - 1.5)^2 + 0.1(x_{ik7} - 0.1)^3 - 5\sin(0.314x_{ik4}x_{ik8}) - 0.4x_{ik5}x_{ik7} + b_k.
\]  

We further generated the censoring time \( C \) independently from an exponential distribution with the rate parameter \( \eta \) selected to induce the censoring proportion of 50%. The parameter \( \eta \) was set to 2 and \( \exp(0.7 + 0.5x_{ik1}) \) to, respectively, produce proportional hazards and nonproportional hazards. This data-generating process was designed to make it difficult for any method to accurately model the survival outcome. We considered the outcome model with arbitrary data complexity that reflects common situations in health data sets: (i) discrete predictors with strong \((X_{ik1})\) and moderate \((X_{ik2})\) associations; (ii) both linear and nonlinear forms of continuous predictors \((X_{ik6} \text{ and } X_{ik7})\); (iii) nonadditive effects \((X_{ik4} X_{ik8})\).

After generating the fully observed data, we amputated the four predictors designed to have missing values \( X_{ik5}, X_{ik6}, X_{ik7}, X_{ik8} \) under the missing at random mechanism using the multivariate amputation approach. Additional technical details for the amputation procedure are available in Schouten et al. (2018); and additional amputation details for the simulation appear in Web Section 2. The resultant simulation data have an overall missingness proportion of 40%. The process of data generation was replicated 250 times.
### TABLE 2
Simulation results for each variable selection approach performed on the fully observed data, and among incomplete data with 40% overall missingness in covariates under proportional hazards and nonproportional hazards. Imputation was conducted via mice. For each of five methods, we show results corresponding to the best threshold values of $\pi$ (based on $F_1$).

|                      | Proportional hazards | Nonproportional hazards |
|----------------------|-----------------------|--------------------------|
|                      | Precision | Recall | $F_1$ | Type I error | Precision | Recall | $F_1$ | Type I error |
| **Fully observed data** | riAFT-BART | 0.95 | 0.88 | 0.91 | 0.01 | riAFT-BART | 0.93 | 0.87 | 0.90 | 0.02 |
|                      | PEAMM | 0.83 | 0.82 | 0.83 | 0.02 | PEAMM | 0.81 | 0.80 | 0.81 | 0.03 |
|                      | FrailtyPenal | 0.79 | 0.76 | 0.77 | 0.04 | FrailtyPenal | 0.75 | 0.69 | 0.72 | 0.05 |
|                      | FrailtyHL | 0.81 | 0.78 | 0.79 | 0.03 | FrailtyHL | 0.78 | 0.71 | 0.74 | 0.04 |
|                      | riCox | 0.78 | 0.72 | 0.75 | 0.05 | riCox | 0.72 | 0.67 | 0.69 | 0.05 |
| **40% overall missing** | riAFT-BART | 0.93 | 0.85 | 0.89 | 0.02 | riAFT-BART | 0.89 | 0.84 | 0.86 | 0.03 |
|                      | PEAMM | 0.88 | 0.74 | 0.80 | 0.03 | PEAMM | 0.84 | 0.72 | 0.78 | 0.04 |
|                      | FrailtyPenal | 0.76 | 0.71 | 0.73 | 0.04 | FrailtyPenal | 0.72 | 0.67 | 0.69 | 0.05 |
|                      | FrailtyHL | 0.79 | 0.73 | 0.76 | 0.05 | FrailtyHL | 0.75 | 0.70 | 0.73 | 0.04 |
|                      | riCox | 0.85 | 0.67 | 0.75 | 0.04 | riCox | 0.71 | 0.63 | 0.67 | 0.07 |
| **Complete cases** | riAFT-BART | 0.85 | 0.81 | 0.83 | 0.03 | riAFT-BART | 0.83 | 0.79 | 0.81 | 0.04 |
|                      | PEAMM | 0.77 | 0.73 | 0.75 | 0.02 | PEAMM | 0.73 | 0.71 | 0.72 | 0.03 |
|                      | FrailtyPenal | 0.71 | 0.65 | 0.68 | 0.05 | FrailtyPenal | 0.69 | 0.62 | 0.65 | 0.04 |
|                      | FrailtyHL | 0.73 | 0.68 | 0.70 | 0.04 | FrailtyHL | 0.70 | 0.65 | 0.67 | 0.05 |
|                      | riCox | 0.71 | 0.66 | 0.68 | 0.04 | riCox | 0.66 | 0.62 | 0.64 | 0.03 |

Abbreviations: FrailtyHL, frailty models with penalized profile h-likelihood; FrailtyPenal, frailty models with penalized likelihood; PEAMM, piecewise exponential additive mixed model; riAFT-BART, random-intercept accelerated failure time model with Bayesian additive regression trees, riCox, random-intercept Cox regression model.

### 4.2.2 Simulation results

Table 2 summarizes variable selection results by each method. The optimal performance in the presence of incomplete data is demonstrated by selecting the most suitable threshold value for $\pi$, which is determined based on the $F_1$ score (refer to Web Figures 4 and 5). This performance is then compared with that of fully observed data (prior to amputation) and complete cases exclusively. It is noteworthy that the optimal threshold value of $\pi$ for riAFT-BART is 0.1, corresponding to the highest $F_1$ score. In addition, the $F_1$-score declines at elevated values of $\pi$, attributable to reduced recall values.

The proposed method riAFT-BART has the best performance across all three data panels (fully observed, with missing data, and complete cases), evidenced by the highest precision, recall, and $F_1$-score, and the lowest type I error. The second best performing method is PEAMM, for its modeling flexibility over parametric methods. In the presence of missing covariate data, all methods had a slightly deteriorated performance when performing variable selection in combination with bootstrap imputation, but the performance of variable selection among only complete cases was considerably lower.

Demonstrated by Figure 2, the proposed method riAFT-BART tended to select the most useful predictors and the least noise predictors, followed by another flexible method PEAMM, under both proportional hazards and nonproportional hazards. Among parametric methods, penalized likelihood methods for frailty models had slightly better performance than riCox. All methods tended to select more noise predictors and less useful predictors under nonproportional hazards; the decrease in the performance due to elevated data complexity associated with nonproportional hazards, was only marginal for riAFT-BART but considerable for less flexible methods, particularly FrailtyHL and riCox.

A perusal of each method’s ability to select each of useful predictors is visualized in Figure 3. All methods had a good power (>0.8) for selecting a strong discrete predictor ($X_{ik1}$), but unsatisfactory power (<0.5) for detecting a discrete predictor ($X_{ik2}$) that only has a moderate relationship with the outcome. For continuous variables, all methods performed well in selecting $X_{ik3}$, which has only a linear relationship with the log survival time, but riAFT-BART produced a significantly higher power for selecting the predictor $X_{ik4}$, which has a complex nonadditive form in model (8). Among the four predictors with missing data $X_{ik5}, ..., X_{ik8}$, the advantage of riAFT-BART is demonstrated by the higher power of select-
The distribution, across 250 data replications, of the numbers of selected noise predictors and useful predictors for each of five methods: random-intercept accelerated failure time model with Bayesian additive regression trees (riAFT-BART), piecewise exponential additive mixed model (PEAMM), frailty models with penalized profile h-likelihood (FrailtyHL), frailty models with penalized likelihood (FrailtyPenal), and random-intercept Cox regression model (riCox), with clustered survival data generated under both proportional hazards (PH) and nonproportional hazards (nPH). The total number of useful predictors is 8 and the total number of noise predictors is 20. There are \( K = 10 \) clusters, each with a size of 200; the total sample size is 2000. The overall proportion of missingness is 40%.

![Figure 2](image)

### Figure 2

Power of each of five methods: random-intercept accelerated failure time model with Bayesian additive regression trees (riAFT-BART), piecewise exponential additive mixed model (PEAMM), frailty models with penalized profile h-likelihood (FrailtyHL), frailty models with penalized likelihood (FrailtyPenal), and random-intercept Cox regression model (riCox), for selecting each of eight useful predictors with clustered survival data generated under proportional hazards (PH) and nonproportional hazards (nPH), based on 250 data replications. There are \( K = 10 \) clusters, each with a size of 200; the total sample size is 2000. The overall proportion of missingness is 40%.

Filled symbols represent the PH setting, and open symbols correspond to the nPH setting.

![Figure 3](image)

### Figure 3

As our proposed method riAFT-BART pivots on the predictor’s VIP, we examined the convergence of the Markov chain by plotting 4500 posterior draws, with the first 1000 discarded as burn-in, of the VIP for four useful predictors and four noise predictors in Web Figure 6. The method converged well.

All simulation procedures were conducted in R on an iMac equipped with a 4 GHz Intel Core i7 processor. The data set under consideration was comprised of 10 clusters \( (K = 10) \), each with a sample size of 200 \( (n_k = 200) \), and included eight relevant predictors and 20 noise predictors. In terms of processing time, riAFT-BART required approximately 58 min to execute, PEAMM took around 42 min, FrailtyPenal ran for about 28 min, FrailtyHL consumed roughly 30 min, and riCox finished in about 8 min. These methods were run with the default memory size setting of 100 MB.
TABLE 3 Variable selection results by each of five methods, with the best selection threshold value of $\pi$ suggested in simulations under nonproportional hazards. Random-intercept accelerated failure time model with Bayesian additive regression trees (riAFT-BART) and random-intercept Cox regression model (riCox) both selected eight variables, piecewise exponential additive mixed model (PEAMM) selected seven variables, frailty models with penalized likelihood (FrailtyPenal) selected six variables, and frailty models with penalized profile h-likelihood (FrailtyHL) selected nine variables.

| Variables             | riAFT-BART | PEAMM | FrailtyPenal | FrailtyHL | riCox |
|-----------------------|------------|-------|--------------|-----------|-------|
| Oxygen level          | Yes        | Yes   | No           | Yes       | Yes   |
| Age                   | Yes        | Yes   | Yes          | Yes       | No    |
| Creatinine            | Yes        | Yes   | No           | Yes       | Yes   |
| White blood cell      | Yes        | Yes   | No           | No        | Yes   |
| SOFA score            | Yes        | Yes   | Yes          | Yes       | No    |
| Race                  | Yes        | No    | Yes          | Yes       | Yes   |
| D dimer               | Yes        | No    | No           | Yes       | No    |
| LDH                   | Yes        | Yes   | Yes          | Yes       | Yes   |
| C-reactive protein    | No         | Yes   | No           | Yes       | Yes   |
| Ferritin              | No         | No    | No           | Yes       | Yes   |
| Oxygen saturation     | No         | No    | Yes          | No        | No    |
| Glasgow coma scale    | No         | No    | Yes          | No        | Yes   |

Abbreviations: LDH = lactate dehydrogenase; SOFA = sequential organ failure assessment.

5 | CASE STUDY: PREDICTORS FOR IN-HOSPITAL MORTALITY AND HETEROGENEOUS SURVIVAL TREATMENT EFFECTS AMONG POST-ICU COVID PATIENTS

In this case study, we first applied the proposed variable selection methods to select important predictors for in-hospital mortality using a comprehensive COVID-19 data set, and then evaluated the treatment effect heterogeneity on patient survival using the selected predictors by implementing our proposed method riAFT-BART. The COVID-19 data set was drawn from six hospitals (clusters) of the Epic electronic medical records system at the Mount Sinai Medical Center. Included in this data set were 1955 patients who were diagnosed with COVID-19 between March 10, 2020 to February 26, 2021 and were in intensive care (ICU) due to COVID-19 during their hospital stay (Wang et al., 2020). Post ICU admission, each patient received COVID-specific medications: either dexamethasone, or remdesivir, or the combination of dexamethasone and remdesivir (dexamethasone + remdesivir). The cluster sizes and number of patients in each treatment group are provided in Web Table 4.

We used the following 28 baseline (time of ICU admission) variables: age, sex, self-reported race and ethnicity, smoking status, binary comorbidities including hypertension, coronary artery disease, cancer, diabetes, asthma, and chronic obstructive pulmonary disease, vital signs including temperature, systolic blood pressure, diastolic blood pressure, patient oxygen level (definition in Web Table 3), heart rate, the fraction of inspired oxygen, body mass index, oxygen saturation, risk score (sequential organ failure assessment score and glasgow coma scale), laboratory test results including partial pressure of oxygen, D dimer value and inflammatory markers such as lactate dehydrogenase, ferritin, c-reactive protein, creatinine and white blood cell count. Vital signs, risk score, and lab results were measured during the first 24 h of ICU admission. Among the 28 candidate predictor variables, 17 variables have missing data, and the missingness proportion ranges from 1.4% to 28.6%. Only 1187 (60.7%) patients have fully observed data for all covariates, which motivated the overall missingness proportion of 40% used in our simulation study (Section 4.2). Detailed summary statistics for baseline variables appear in Web Table 4.

The Kaplan–Meier survival curves by treatment group are displayed in Web Figure 7, with crossovers suggesting non-proportional hazards. Assessed by the distribution of the generalized propensity scores shown in Web Figure 8, there is moderate covariate overlap across three treatment groups in the COVID-19 data.

We first implemented all five variable selection methods. All 28 candidate predictors were included in the imputation models, and the optimal threshold value of $\pi$ for each method, which produced the best $F_1$ score in the simulation study, was used to select the final set of predictors. As summarized in Table 3, FrailtyHL selected the most (nine) predictors, riAFT-BART and riCox both selected eight predictors, PEAMM selected seven predictors, and FrailtyPenal selected the
least (six) predictors. Age and lactate dehydrogenase were the only two predictors selected by all five methods. Variables selected by the two flexible methods riAFT-BART and PEAMM were largely in agreement with each other, with the exception that riAFT-BART but not PEAMM selected Race and D dimer, and PEAMM but not riAFT-BART selected C-reactive protein.

To further distinguish between methods based on their ability to select the most relevant predictors, we computed the predictive performance of the five methods that modeled the outcome variable on their, respectively, selected predictor variables. For survival outcomes, we used the concordance statistic as a metric to assess the predictive performance (Harrell et al., 1996). The concordance statistic is the fraction of concordant pairs among all pairs of individuals. A pair of patients is called concordant if the patient with higher (lower) predicted survival probability at a given time point had a longer (shorter) time to death. In this case study, we chose 21 days post-ICU admission as the target time point, because it is a sufficiently long time to evaluate the effects of COVID-19 treatments on in-hospital mortality for severe patients. As the COVID-19 data set is subject to missing covariate data, to calculate the cross-validated concordance statistic, we follow the strategy described in Lin et al. (2022). We first split the data into two halves and implemented each of the five variable selection methods on half of the data. Then, we imputed the other half of the data with a single imputation and recorded the concordance statistic. Finally, we repeated the previous two steps 250 times to get the distribution of the cross-validated concordance statistics. Figure 4 shows that the proposed method riAFT-BART yielded the highest predictive performance with a median concordance statistic $>0.7$.

Supported by both the simulation study and case study, riAFT-BART provided the best variable selection results. Thus, we used the eight predictors selected by riAFT-BART as confounding variables, and estimated the heterogeneous treatment effects on patient survival using methodology described in Section 2. To address the missing data issue when drawing causal inferences about treatment effects, we fitted an riAFT-BART model (1) on each of the 100 bootstrap and imputed data sets (previously conducted for variable selection), and drew statistical inferences about the ISTE $\zeta_{a_j,a_{j'}}(x_{ik})$, defined in Equation (3), by pooling posterior samples of the ISTE, $\tilde{f}^d(a_j, x_{ik}) - \tilde{f}^d(a_{j'}, x_{ik})$ as in Equation (2), across model fits arising from the multiple data sets (Hu et al., 2022; Zhou & Reiter, 2010). Inferences about the population average treatment effects $\bar{\theta}_{a_j,a_{j'}}$ can be obtained from the posterior samples of $\tilde{f}^d(a_j, x_{ik}) - \tilde{f}^d(a_{j'}, x_{ik})$ across the sample population (Hu et al., 2022). The population average treatment effects, summarized in Web Table 5, suggest that there was no statistically significant difference in treatment effect when comparing dexamethasone with remdesivir or with a combined use of dexamethasone and remdesivir. However, there was a statistically significant treatment benefit associated with dexamethasone+remdesivir compared to using remdesivir alone. Web Figure 9 demonstrates that substantial variability in the institutional effect was captured by the riAFT-BART model: the Mount Sinai main hospital had considerably better outcomes than the Mount Sinai Queens hospital.

Turning to the treatment effect heterogeneity. Figure 5 summarizes the combination rules of covariates $S_{a_j,a_{j'}}$ that have differential subgroup treatment effect $M(S_{a_j,a_{j'}})$, defined in Equation (4), than the sample population treatment...
FIGURE 5 Final Random Forests model fit to the posterior mean of the individual survival treatment effect comparing remdesivir and dexamethasone + remdesivir. Values in each node correspond to the posterior mean, in terms of difference in log survival days, for the subgroup of individuals represented in that node. Uncertainty intervals were obtained by pooling the posterior samples arising from the multiple imputed data sets. WBC: White blood cell.

effect. When comparing remdesivir and dexamethasone + remdesivir across the sample population, there was significant treatment benefit, suggested by a longer expected survival time, for the combined use of dexamethasone and remdesivir. Our analysis suggests that combining dexamethasone and remdesivir offered enhanced treatment benefit for relatively healthier COVID-19 patients with lower oxygen saturation (<95.5%) and a normal white blood cell count (<11.4 K/μL), demonstrated by the leftmost branch of the tree diagram in Figure 5. By contrast, the rightmost branch suggests that there was no statistically significant treatment benefit associated with either treatment choice for comparatively unhealthier patients with a higher level of oxygen saturation (>95.5%).

6 | DISCUSSION

We describe two important and practical utilities of a new Bayesian machine learning method riAFT-BART for analyzing clustered and censored survival data. To address the implications of complex data structures of large-scale clinical data sets generated from multiple institutions for causal inferences about population treatment effect, riAFT-BART was developed to accurately estimate the population average treatment effect on patient survival while accounting for the variation in institutional effects (Hu et al., 2022).

Moving from broad applicability of a population study to personalized medicine requires the understanding of treatment effect heterogeneity. The treatment effect heterogeneity was traditionally identified by first enumerating potential effect modifiers with subject-matter experts and then estimating the average treatment effect within each subgroup (Kent et al., 2020). This approach is particularly suitable to confirmatory treatment effect heterogeneity analysis in randomized clinical trials. In observational data with a large number of pretreatment covariates, such a priori specification that separates the issues of confounding and treatment effect heterogeneity is often practically infeasible. We described a way in which the new method riAFT-BART can be used to perform an exploratory treatment effect heterogeneity analysis to generate scientifically meaningful hypotheses and treatment effect discovery. Through a comprehensive simulation representative of complex confounding and heterogeneity settings with right-censored survival data, we provided new empirical evidence for the better performance of our proposed method in comparison with three existing methods. In the application to a large COVID-19 data set, we found that oxygen saturation and white blood cell count were two key patient factors that modulated the comparative treatment effect between remdesivir and dexamethasone + remdesivir. We further exploited the posterior samples of the ISTE from riAFT-BART and the random forests model to explore clinically meaningful relationships between COVID-19 medications, patient profile, and expected survival time. The results could facilitate treatment effect discovery in subpopulations and may inform personalized treatment strategy and the planning of future confirmatory randomized trials. Based on the simulation results, caution is warranted when drawing causal inferences about subpopulations in the tail regions of the generalized propensity score distribution, as the reliability of these inferences may be compromised.
Drawing on prior work (Hu et al., 2021) that combines bootstrap imputation and BART for variable selection among incomplete data, we develop a way in which riAFT-BART can be used to select important predictors among incomplete, clustered, and censored survival observations. This strategy is general enough to accommodate any missing data pattern and is highly flexible as it leverages BART to model the relationship between survival times and covariates. As demonstrated by both simulations and the case study, this elevated modeling flexibility gave rise to substantially better variable selection results, particularly in terms of identifying predictors of complex functional forms, compared to existing variable selection methods suitable for clustered survival data. In general, we recommend using smaller values for the selection threshold, such as \( \pi = 0.1 \) or \( \pi = 0.2 \), when employing riAFT-BART for variable selection. Furthermore, the determination of the optimal selection threshold value can be achieved through cross-validation. Specifically, the most suitable threshold value ought to yield a set of selected variables that result in better cross-validated predictive performance.

There are several important avenues for further research. First, the riAFT-BART model could be extended to include the random slopes and accommodate the cluster-level covariates. Correspondingly, a variable selection method can be developed to identify important predictors at both the individual level and the cluster level. Second, using flexible modeling techniques alone will not address violations of the no unmeasured confounding assumption required for causal inference. Developing a formal sensitivity analysis approach (Hu et al., 2022) for unmeasured confounding would be a worthwhile and important contribution. Finally, our bootstrap imputation-based variable selection procedure is computationally demanding on a large data set. Although the bootstrap resampling can be computed in parallel on multiple cores when such resources are available, it would be worthwhile to develop strategies that could offer substantial computational savings.

ACKNOWLEDGMENTS
This work was supported in part by award ME-2021C2-23685 from the Patient-Centered Outcomes Research Institute, and by grants R21CA245855 and 1R01HL159077-01A1 from the National Institute of Health.

CONFLICT OF INTEREST STATEMENT
The author declares no conflicts of interest.

DATA AVAILABILITY STATEMENT
R codes to implement our proposed methods and the comparison methods, and to replicate our simulation studies are provided as a supplement to this paper and in the R package riAFTBART. Access to the COVID-19 data used in the case study needs to be requested and approved by the Icahn School of Medicine at Mount Sinai. The “Data & Code” section in the supplemental materials provides instructions to reproduce all manuscript tables and figures using our R scripts.

OPEN RESEARCH BADGES
This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge “Reproducible Research” for making publicly available the code necessary to reproduce the reported results. The results reported in this article were reproduced partially due to computational complexity and data confidentiality issues.

ORCID
Liangyuan Hu https://orcid.org/0000-0002-4067-892X

REFERENCES
Abrahantes, J. C., & Burzykowski, T. (2005). A version of the EM algorithm for proportional hazard model with random effects. *Biometrical Journal, 47*, 847–862.
Androulakis, E., Koukouvinos, C., & Vonta, F. (2012). Estimation and variable selection via frailty models with penalized likelihood. *Statistics in Medicine, 31*, 2223–2239.
Arpino, B., & Cannas, M. (2016). Propensity score matching with clustered data. An application to the estimation of the impact of caesarean section on the Apgar score. *Statistics in Medicine, 35*, 2074–2091.
Bender, A., Groll, A., & Scheipl, F. (2018). A generalized additive model approach to time-to-event analysis. *Statistical Modelling, 18*, 299–321.
Bleich, J., Kapelner, A., George, E. I., & Jensen, S. T. (2014). Variable selection for BART: An application to gene regulation. *Annals of Applied Statistics, 8*, 1750–1781.
Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32.
Cai, J., & Zeng, D. (2011). Additive mixed effect model for clustered failure time data. *Biometrics*, 67, 1340–1351.
Chen, P.-Y., & Tsibris, A. A. (2001). Causal inference on the difference of the restricted mean lifetime between two groups. *Biometrics*, 57, 1030–1038.
Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *Annals of Applied Statistics*, 4, 266–298.
Deng, H. (2019). Interpreting tree ensembles with inTrees. *International Journal of Data Science and Analytics*, 7, 277–287.
Fan, J., & Li, R. (2001). Variable selection via concave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96, 1348–1360.
Fan, J., & Li, R. (2002). Variable selection for Cox’s proportional hazards model and frailty model. *Annals of Statistics*, 30, 74–99.
Feng, P., Zhou, X.-H., Zou, Q.-M., Fan, M.-Y., & Li, X.-S. (2012). Generalized propensity score for estimating the average treatment effect of multiple treatments. *Statistics in Medicine*, 31, 681–697.
Foster, J. C., Taylor, J. M., & Ruberg, S. I. (2011). Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, 30, 2867–2880.
Gelman, A., van Dyk, D. A., Huang, Z., & Boscardin, J. W. (2006). Using redundant parameterizations to fit hierarchical models. *Journal of Computational and Graphical Statistics*, 17, 95–122.
Ha, I. D., Pan, J., Oh, S., & Lee, Y. (2014). Variable selection in general frailty models using penalized h-likelihood. *Journal of Computational and Graphical Statistics*, 23, 1044–1060.
Harrell, F. E., Jr, Lee, K. L., & Mark, D. B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15, 361–387.
Henderson, N. C., Louis, T. A., Rosner, G. L., & Varadhan, R. (2020). Individualized treatment effects with censored data via fully nonparametric Bayesian accelerated failure time models. *Biostatistics*, 21, 50–68.
Hu, L., & Gu, C. (2021). Estimation of causal effects of multiple treatments in healthcare database studies with rare outcomes. *Health Services and Outcomes Research Methodology*, 21, 287–308.
Hu, L., Gu, C., Lopez, M., Ji, J., & Winsivesky, J. (2020). Estimation of causal effects of multiple treatments in observational studies with a binary outcome. *Statistical Methods in Medical Research*, 29, 3218–3234.
Hu, L., & Hogan, J. W. (2019). Causal comparative effectiveness analysis of dynamic continuous-time treatment initiation rules with sparsely measured outcomes and death. *Biometrics*, 75, 695–707.
Hu, L., Ji, J., Ennis, R. D., & Hogan, J. W. (2022). A flexible approach for causal inference with multiple treatments and clustered survival outcomes. *Statistics in Medicine*, 41, 4982–4999.
Hu, L., Ji, J., & Li, F. (2021). Estimating heterogeneous survival treatment effect in observational data using machine learning. *Statistics in Medicine*, 40, 4691–4713.
Hu, L., Joyce Lin, J.-Y., & Ji, J. (2021). Variable selection with missing data in both covariates and outcomes: Imputation and machine learning. *Statistical Methods in Medical Research*, 30, 2651–2671.
Hu, L., Lin, J.-Y. J., Sigel, K., & Kale, M. (2021). Estimating heterogeneous survival treatment effects of lung cancer screening approaches: A causal machine learning analysis. *Annals of Epidemiology*, 62, 36–42.
Hu, L., Liu, B., Ji, J., & Li, Y. (2020). Tree-based machine learning to identify and understand major determinants for stroke at the neighborhood level. *Journal of the American Heart Association*, 9, e016745.
Hu, L., Liu, B., & Li, Y. (2020). Ranking sociodemographic, health behavior, prevention, and environmental factors in predicting neighborhood cardiovascular health: a Bayesian machine learning approach. *Preventive Medicine*, 141, 106240.
Hu, L., Zou, J., Gu, C., Ji, J., Lopez, M., & Kale, M. (2022). A flexible sensitivity analysis approach for unmeasured confounding with multiple treatments and a binary outcome with application to SEER-Medicare lung cancer data. *Annals of Applied Statistics*, 16, 1014–1037.
Huang, J., & Ma, S. (2010). Variable selection in the accelerated failure time model via the bridge method. *Lifetime Data Analysis*, 16, 176–195.
Ishwaran, H., Kogalur, U. B., Gorodeski, E. Z., Minn, A. J., & Lauer, M. S. (2010). High-dimensional variable selection for survival data. *Biostatistics*, 105, 205–217.
Kent, D. M., Paulus, J. K., van Klaveren, D., D'Agostino, R., Goodman, S., Hayward, R., Ioannidis, J. P., Patrick-Lake, B., Morton, S., Pencina, M., Raman, G., Ross, J. S., Selker, H. P., Varadhan, R., Vickers, A., Wong, J. B., & Steyerberg, E. W. (2020). The predictive approaches to treatment effect heterogeneity (PATH) statement. *Annals of Internal Medicine*, 172, 35–45.
Lagani, V., & Tsamardinos, I. (2010). Structure-based variable selection for survival data. *Bioinformatics*, 26, 1887–1894.
Lee, Y., & Oh, H.-S. (2014). A new sparse variable selection via random-effect model. *Journal of Multivariate Analysis*, 125, 89–99.
Lin, J.-Y. J., Hu, L., Huang, C., Lawrence, S., & Govindarajulu, U. (2022). A flexible approach for variable selection in large-scale healthcare database studies with missing covariate and outcome data. *BMC Medical Research Methodology*, 22, 132.
Logan, B. R., Sparapani, R., McCulloch, R. E., & Laud, P. W. (2019). Decision making and uncertainty quantification for individualized treatments using Bayesian Additive Regression Trees. *Statistical Methods in Medical Research*, 28, 1079–1093.
Long, Q., & Johnson, B. A. (2015). Variable selection in the presence of missing data: Resampling and imputation. *Biostatistics*, 16, 596–610.
Lu, M., Sadiq, S., Feaster, D. J., & Ishwaran, H. (2018). Estimating individual treatment effect in observational data using random forest methods. *Journal of Computational and Graphical Statistics*, 27, 209–219.
Marquardt, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, II, 431–441.
Marra, G., & Wood, S. N. (2011). Practical variable selection for generalized additive models. *Computational Statistics & Data Analysis*, 55, 2372–2387.
Rondeau, V., Commenges, D., & Joly, P. (2003). Maximum penalized likelihood estimation in a gamma-frailty model. *Lifetime Data Analysis*, 9, 139–153.

Royston, P., & Parmar, M. K. (2013). Restricted mean survival time: An alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Medical Research Methodology*, 13, 1–15.

Schouten, R. M., Lugtig, P., & Vink, G. (2018). Generating missing values for simulation purposes: A multivariate amputation procedure. *Journal of Statistical Computation and Simulation*, 88, 2909–2930.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58, 267–288.

Utazirubanda, J. C., M León, T., & Ngom, P. (2021). Variable selection with Group LASSO approach: Application to Cox regression with frailty model. *Communications in Statistics-Simulation and Computation*, 50, 881–901.

van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). Chapman & Hall.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). *mice*: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45, 1–67.

van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6, 1–23.

VanderWeele, T. J. (2019). Principles of confounder selection. *European Journal of Epidemiology*, 34, 211–219.

Wang, Z., Zheutlin, A., Kao, Y.-H., Ayers, K., Gross, S., Kovatch, P., Nirenberg, S., Charney, A., Nadkarni, G., De Freitas, J. K., O’Reilly, P., Just, A., Horowitz, C., Martin, G., Branch, A., Glicksberg, B. S., Charney, D., Reich, D., Oh, W. K., … Li, L. (2020). Hospitalised COVID-19 patients of the Mount Sinai Health System: a retrospective observational study using the electronic medical records. *BMJ Open*, 10, e040441.

Wood, A. M., White, I. R., & Royston, P. (2008). How should variable selection be performed with multiply imputed data? *Statistics in Medicine*, 27, 3227–3246.

Zhou, X., & Reiter, J. P. (2010). A note on Bayesian inference after multiple imputation. *American Statistician*, 64, 159–163.

**SUPPORTING INFORMATION**

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Hu, L. (2024). A new method for clustered survival data: Estimation of treatment effect heterogeneity and variable selection. *Biometrical Journal*, 66, 2200178. [https://doi.org/10.1002/bimj.202200178](https://doi.org/10.1002/bimj.202200178)